LNAI 7769

Takashi Washio
Jun Luo (Eds.)

Emerging Trends in Knowledge Discovery and Data Mining

PAKDD 2012 International Workshops: DMHM, GeoDoc, 3Clust, and DSDM
Kuala Lumpur, Malaysia, May/June 2012
Revised Selected Papers

Springer

Lecture Notes in Artificial Intelligence 7769

Subseries of Lecture Notes in Computer Science

Takashi Washio Jun Luo (Eds.)

Emerging Trends in Knowledge Discovery and Data Mining

PAKDD 2012 International Workshops:
DMHM, GeoDoc, 3Clust, and DSDM
Kuala Lumpur, Malaysia, May 29 – June 1, 2012
Revised Selected Papers

Springer

Series Editors

Volume Editors

Takashi Washio
Osaka University, The Institute of Scientific and Industrial Research (ISIR)
8-1 Mihogaoka, Osaka, Ibaraki 5670047, Japan
E-mail: washio@ar.sanken.osaka-u.ac.jp

Jun Luo
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
1068 Xueyuan Boulevard, Shenzhen, Guangdong 518055, China
E-mail: jun.luo@siat.ac.cn

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-36777-9 e-ISBN 978-3-642-36778-6
DOI 10.1007/978-3-642-36778-6
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013932132

CR Subject Classification (1998): I.2.6-7, I.2.1, H.2.8, H.3.3-5, H.4.1-3, I.4.9

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Techniques of knowledge discovery and data mining (KDD) have rapidly developed along with the significant progress of computer and network technologies. In addition, the growth of industries, infrastructures, and services associated with computer networks in the Pacific Asia region have been significant in the last two decades, and these have further boosted attention on the KDD research not only in this region but all over the world. Under these circumstances, the Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD) is now one of the internationally representative conferences in the KDD area.

Past PAKDD conferences have hosted workshops in concert with the main tracks to provide an excellent opportunity for the presentations of highly potential yet early-stage research. In addition, the PAKDD conference in 2012 organized a doctoral symposium on data mining to provide a forum for early-career researchers such as PhD students and junior researchers who have just received their PhD degrees to share their latest results. These activities have contributed to fostering young and active researchers, particularly in the Pacific Asia region, and have contributed to expanding the international KDD research community. This year, PAKDD 2012 accepted four outstanding workshop proposals: Data Mining for Healthcare Management (DMHM 2012), Multi-View Data, High-Dimensionality, External Knowledge: Striving for a Unified Approach to Clustering (3Clust 2012), Geospatial Information and Documents (GeoDoc 2012), and Pacific Asia Workshop on Intelligence and Security Informatics (PAISI 2012). Moreover, we hosted the second Doctoral Symposium on Data Mining (DSDM 2012).

This proceedings volume is a collection of outstanding papers selected from these workshops and the doctoral symposium, except for PAISI 2012 which had its own proceedings. The Third Workshop on Data Mining for Healthcare Management (DMHM 2012) aimed at providing a common platform for the discussion of challenging issues and potential techniques in this emerging field of data mining for healthcare management, accepted four oral presentations from eight paper submissions and selected three papers out of the four presentations for these proceedings. Multi-view Data, High Dimensionality, External Knowledge: Striving for a Unified Approach to Clustering (3Clust 2012) focused on solving emerging problems such as clustering ensembles, semi-supervised clustering, subspace/projective clustering, co-clustering, and multi-view clustering, accepted five oral presentations from eight submissions, and selected an outstanding paper from the five presentations for the proceedings. Geospatial Information and Documents (GeoDoc 2012) highlighted the formalization of geospatial concepts and relationships with a focus on the extraction of geospatial relations in free text datasets to offer to the database community a unified framework for geodata discovery, accepted five oral presentations from eight submissions, and selected two papers out of the five presentations. The Second PAKDD Doctoral Symposium on Data Mining (DSDM 2012) received 19 paper submissions. It accepted five presentations for oral presentations, and allowed them to be included in these proceedings upon their very strict revisions.

Participants in these workshops and the symposium shared the opportunity to present and discuss their recent works on data mining foundations, techniques, and applications with senior researchers in the community. With a focus on facilitating discussion, these workshops can be a lively venue for pushing forward the envelope of research in knowledge discovery and data mining.

We expect these proceedings will contribute to the growth of the world-wide community of KDD researchers.

December 2012

Takashi Washio
Jun Luo

Organization

The paper selection of the workshops was made by the Program Committee of each organization respectively. After paper selection, the book was edited and managed by the volume editors.

Volume Editors

Takashi Washio	Osaka University, Japan
Jun Luo	Shenzhen Institutes of Advanced Technology, CAS, China
Prasanna Desikan	Center for Healthcare Innovation, Allina Hospitals and Clinics, USA
Kuo-Wei Hsu	National Chengchi University, Taiwan
Jaideep Srivastava	University of Minnesota, USA
Ee-Peng Lim	Singapore Management University, Singapore
Maguelonne Teisseire	UMR TETIS, Cemagref, France
Mathieu Roche	LIRMM, CNRS, University of Montpellier 2, France
Carlotta Domeniconi	George Mason University, USA
Francesco Gullo	Yahoo! Research, Spain
Andrea Tagarelli	University of Calabria, Italy
Hung Khoon Tan	Universiti Tunku Abdul Rahman, Malaysia
Wong, Chee Onn	Multimedia University, Malaysia

Third Workshop on Data Mining for Healthcare Management

Workshop Chairs

Prasanna Desikan	Center for Healthcare Innovation, Allina Hospitals and Clinics, USA
Kuo-Wei Hsu	National Chengchi University, Taiwan
Jaideep Srivastava	University of Minnesota, USA
Ee-Peng Lim	Singapore Management University, Singapore

Program Committee

Chi-Huang Chen	National Taiwan University, Taiwan
Po-Hsun Cheng	National Kaohsiung Normal University, Taiwan
Hong Tat Ewe	Universiti Tunku Abdul Rahman, Malaysia
Feipei Lai	National Taiwan University, Taiwan

Lam Hong Lee	Universiti Tunku Abdul Rahman, Malaysia
Vincent Shin-Mu Tseng	National Cheng Kung University, Taiwan
Chandan Reddy	Wayne State University, USA
VRK Subrahmanya Rao	Cognizant Technologies, India
Yonghong Tian	Peking University, China
P. Krishna Reddy	International Institute of Information Technology, Hyderabad, India

Geospatial Information and Documents

Workshop Chairs

Maguelonne Teisseire	UMR TETIS, Cemagref, France
Mathieu Roche	LIRMM, CNRS, University of Montpellier 2, France

Program Committee

Masanori Akiyoshi	Osaka University, Japan
Torben Bach Pedersen	Aalborg University, Denmark
Jason Baldridge	University of Texas, USA
Mete Celik	Erciyes University, Turkey
Robert Haining	University of Cambridge, UK
Tahar Kechadi	UCD School of Computer Science and Informatics, Ireland
Stan Matwin	University of Ottawa, Canada
Donato Malerba	University of Bari, Italy
Pascal Poncelet	University of Montpellier 2, France

Multi-View Data, High-Dimensionality, External Knowledge: Striving for a Unified Approach to Clustering

Workshop Chairs

Carlotta Domeniconi	George Mason University, USA
Francesco Gullo	Yahoo! Research, Spain
Andrea Tagarelli	University of Calabria, Italy

Program Committee

Ana Fred	Technical University of Lisbon, Portugal
Arthur Zimek	Ludwig-Maximilians-Universität München, Germany
Chris Ding	University of Texas at Arlington, USA
Dimitrios Gunopulos	University of Athens, Greece
Emmanuel Mller	Karlsruhe Institute of Technology (KIT), Germany

Huan Liu	University of Arizona, USA
Huzefa Rangwala	George Mason University, USA
James Bailey	University of Melbourne, Australia
Joydeep Ghosh	University of Texas - Austin, USA
Pu Wang	StumbleUpon, USA
Rosa Meo	University of Turin, Italy
Tao Li	Florida International University, USA
Thomas Seidl	RWTH Aachen University, Germany
Wei Fan	IBM Research, USA

The second PAKDD Doctoral Symposium on Data Mining

Workshop Chairs

Hung Khoon Tan	Universiti Tunku Abdul Rahman, Malaysia
Wong, Chee Onn	Multimedia University, Malaysia

Program Committee

Cao, Juan	Chinese Academy of Science, China
Chua, Linda Sook Ling	Multimedia University, Malaysia
Hwang, Kyu-Baek	Soongsil University, Korea
Jiang, Yu-gang	Fudan University, China
Jun, Sese	Tokyo Institute of Technology, Japan
Khor, Siak Wang	Universiti Tunku Abdul Rahman, Malaysia
Kida, Takuya	Hokkaido University, Japan
Kishigami, Jay	Nippon Telegraph and Telephone, Japan
Li, Shuai Cheng	City University of Hong Kong, Hong Kong
Liu, Xingwu	Chinese Academy of Science, China
Ono, Hirotaka	Kyushu University, Japan
Teoh, Andrew Beng Jin	Yonsei University, Korea
Ting, Choo Yee	Multimedia University, Malaysia
Wang, Feng	East China Normal University, China
Wei, Xiao-yong	Sichuan University, China
Wu, Xiao	Southwest Jiaotong University, China
Yang, Jialiang	Chinese Academy of Science, China
Zhao, Wan-lei	University of Kaiserslautern, Germany

Table of Contents

Modality Classification for Medical Images Using Sparse Coded Affine-Invariant Descriptors

Viktor Gál[1], Illés Solt[2], Etienne Kerre[1], and Mike Nachtegael[1]

[1] Department of Applied Mathematics and Computer Science,
Ghent University, Belgium
{viktor.gal,etienne.kerre,mike.nachtegael}@ugent.be

[2] Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary
solt@tmit.bme.hu

Abstract. Modality is a key facet in medical image retrieval, as a user is likely interested in only one of e.g. radiology images, flowcharts, and pathology photos. While assessing image modality is trivial for humans, reliable automatic methods are required to deal with large un-annotated image bases, such as figures taken from the millions of scientific publications. We present a multi-disciplinary approach to tackle the classification problem by combining image features, meta-data, textual and referential information. We test our system's accuracy on the ImageCLEF 2011 medical modality classification data set. We show that using a fully affine-invariant feature descriptor and sparse coding on these descriptors in the Bag-of-Words image representation significantly increases the classification accuracy. Our best method achieves 87.89% accuracy and outperforms the state of the art.

Keywords: image classification, image feature extraction, image modality, sparse coding, text mining.

1 Introduction

Imaging modality is an important aspect of the image for medical retrieval [1]. In user-studies, clinicians have indicated that modality is one of the most important filters that they would like to be able to limit their search by. However, this modality is typically extracted from the caption and is often not correct or present. Studies have shown that the modality can be extracted from the image itself using visual features [2,3,4]. Therefore, in this paper, we propose to use both visual and textual features for medical image representation.

Our main focus in this paper is on the impact of using a fully affine-invariant feature descriptor (ASIFT [5]) and (extension of) the Bag-of-Words (BoW) image representation [6]. In the classical BoW image representation vector quantisation is applied to encode descriptors (e.g SIFT, ASIFT) of local image patches. Prior to encoding, a codebook is learned via an unsupervised clustering algorithm (e.g. K-means), which summarizes the distribution of signals by a set of "visual words".

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 1–10, 2013.

As a result, these vector quantised codes represent the image through the frequencies of these visual words.

As vector quantisation introduces a significant error in encoding a signal, to overcome this problem, sparse coding has attracted much attention in image classification [7,8]. In this paper we use Least angle regression algorithm [9] for sparse coding the extracted feature descriptors.

We show that by using ASIFT on the images and applying sparse coding on these features we achieved better performance as the state-of-art results for modality classification of medical images.

The proposed algorithm is evaluated in the context of the ImageCLEF 2011 Modality Classification task [10], which uses a data set of 988+1024 images taken from PubMed articles.

The rest of this paper is organised as follows. In Section 2, we describe in detail our experimental setting. In Section 3, we present and discuss the different experiments and we conclude in Section 4.

2 Methods

In this section, we describe in detail our experimental setting.

2.1 Feature Extraction

Caption Text. Figures in scientific publications often have descriptive captions that provide information on the modality of the image. "Contrast-enhanced axial computed tomographic scan", "HRCT showing extensive areas of consolidation with air bronchogram" are examples of captions of images assigned to the 'CT' modality class. However, the caption may be missing or may not hint at the modality, e.g. "E. coli that satisfy the similarity threshold values." As the examples suggest, the linguistic constructs expressing modality can have a high variation.

Considering these remarks, we extract binary features from caption texts as follows. We define a set of regular expressions to be matched against the caption text, a match results in a value of 1. Regular expressions were initially created for each word having a high information gain for any of the modality classes and were later manually refined to capture linguistic variations (e.g. `f?MRI?`) and multi-word phrases (e.g. `error bars?`).

MeSH Terms. Scientific articles indexed by Medline/PubMed are tagged with MeSH terms (medical subject headings) by field experts. MeSH terms can be seen as a thesaurus for the life sciences containing entries like 'Human', 'Liver Neoplasms' and 'Magnetic Resonance Imaging', entries can be further qualified by e.g. 'methods', 'pathology'. We hypothesise that the article's MeSH terms and its figures' modality are correlated, and hence define features corresponding to individual MeSH terms and qualifiers. A unique identifier for the article (e.g. PMID or DOI) is required to retrieve its MeSH annotations, however, such identifiers can be absent. As the number of MeSH terms, qualifiers and their combinations far exceeds the number of modality labels, we perform feature selection by keeping only those that are present for at least a predefined number of articles in the training set.

Colour Histogram. Using colour histograms in content-based image retrieval system has been successfully applied in the past, for a detailed review see [11]. Based on these studies we have chosen to use HSV colour-space based histogram, and quantised the *hue* and the *saturation* to three and the *value* to four levels.

Based on this we defined $\boldsymbol{f}_{hist}$ feature vector, where each element of the vector represents the normalised number of pixels in a given histogram bin.

Mean of Pixels. Through manually supervised error analysis on the training set, we identified that the images in `Graphic` 1st-level group are mainly having a white background. Hence, we have defined a simple feature $f_{mean} = \overline{\boldsymbol{I}_j}$, that represents the mean value of the pixels in an image. By simply thresholding these values one could identify the images that belong to the `Graphic` group with a very high accuracy.

Axis Recognition. The previously mentioned mean of pixels method gave a strong support for recognising images in the `Graphic` top-level group, but as it consists of two sub-groups, `Graphs` and `Drawing`, thus a new feature was required to differentiate the images belonging to one or the other category. By manually observing the images in these two categories one can easily point out the main difference by using a simple edge detector: the images belonging to the `Graphs` category are mainly consisting of horizontal and vertical lines (i.e. the x-y axis of a graph), whereas the images in `Drawing` category are mostly diagrams, where the orientation of the lines is random.

Based on this idea we have defined the following feature. Let $L_{\boldsymbol{I}_j}$ be the set of all the detected lines and $GL_{\boldsymbol{I}_j}$ be the set of *good lines* in an arbitrary image $\boldsymbol{I}_j$, where a given line is a *good line* if its orientation is horizontal or vertical and it is within a given margin of the picture's border. The latter condition is to eliminate the borders of an image as *good lines.*

Using these two sets we defined a feature

$$f_{lines}(\boldsymbol{I}_j) = \frac{|GL_{\boldsymbol{I}_j}|}{|L_{\boldsymbol{I}_j}|} \tag{1}$$

In order to detect the lines and their orientation in an image we used a simple Hough transform [12].

Skin Detection. The images in the `Dermatology` category was one of the most difficult to recognise. As not only it was the least represented category in the whole training set, i.e. there are only seven examples (see Table 1) for this category, but the images in this set are simple photographs (of various skin abnormalities) thus they have very similar characteristics to the `general photo` labeled images.

Hence, most of the previously defined features failed to distinguish the images in `Dermatology` set from the others.

Using a simple skin detector algorithm [13] we defined a new feature $f_{skin}(\boldsymbol{I}_j)$ for an image $\boldsymbol{I}_j$

$$f_{skin}(\boldsymbol{I}_j) = \overline{SD(\boldsymbol{I}_j)} \tag{2}$$

where the function $SD(\cdot)$ calculates the skin-segmented binary image of an input image, and $\overline{\boldsymbol{I}_k}$—as previously defined—is the mean value of image $\boldsymbol{I}_k$.

Radiopaedia. Radiopaedia (`http://radiopaedia.org`) is a community wiki for radiology images and patient cases. Images are tagged by users with the body system (e.g. Heart, Musculoskeletal) depicted, but unfortunately for us, not with the type of radiology method used to create the image. Leveraging the mutual information between body systems and radiology methods, we derived features for modality classification by taking the output probabilities of a classifier trained to predict body systems shown in the image.

Bag of visual-words The state-of-the-art content based image retrieval systems has been significantly improved by the introduction of scale-invariant feature transform (SIFT) [14] features and the bag-of-words image representation [15,16,17,18].

The bag-of-visual-words image representation is based on the bag of words (BoW) model in natural language processing (NLP). BoW in NLP is a popular method for representing documents. In this model a document is simply represented by the number of different words that are in the document. The idea behind this is, that documents on the same topic have similar words with similar number of occurrences in them (see LDA [19]).

In case of an image, the basic idea of bag-of-words model is that a set of local image patches is sampled using some method—e.g. densely or using a keypoint detector—and a vector of visual descriptors is evaluated on each patch independently.

In this paper we used two variants of the well known SIFT descriptor on each patch:

– **SIFT.** The SIFT descriptor computes a gradient orientation histogram within the support region. For each of eight orientation planes, the gradient image is sampled over a four by four grid of locations, hence resulting in a 128-dimensional feature vector for each region. In order to make the descriptor less sensitive to small changes in the position of the support region and put more emphasis on the gradients that are near the centre of the region a Gaussian window function is used to assign a weight to the magnitude of each sample point.
– **Affine-SIFT.** (ASIFT) [5] The SIFT detector normalizes rotations and translations and simulates all zooms out of the query and of the search images. Because of this feature, it is the only fully scale-invariant method. ASIFT simulates with enough accuracy all distortions caused by a variation of the camera optical axis direction. Then it applies the SIFT method. In

other words, ASIFT simulates three parameters: the scale, the camera longitude angle, and the latitude angle (which is equivalent to the tilt) and normalizes the other three (translation and rotation). The mathematical proof that ASIFT is fully affine invariant is given in [5]. The key observation is that, although a tilt distortion is irreversible due to its non-commutation with the blur, it can be compensated up to a scale change by digitally simulating a tilt of same amount in the orthogonal direction. As opposed to the normalization methods that suffer from this non-commutation, ASIFT simulates and thus achieves the full affine invariance.

After acquiring the feature descriptors for all the images in the data set, first we created a visual-word dictionary $\mathbf{D}$ (analogy to a word dictionary) by performing a K-means clustering algorithm over all the vectors. This dictionary is used to map similar visual patches into one, or more visual-words of the acquired dictionary. The mapping can be done by simple vector quantisation [20], where each visual patch is mapped to the nearest visual-word in the dictionary or by using sparse coding, where the visual patch is a linear combination of a small number of the visual-words.

The sparse coding of the visual patches was achieved by using least angle regression algorithm [9] for solving the Lasso. Given a matrix of signals $\mathbf{X} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_p}] \in \Re^{m \times p}$ and a dictionary $\mathbf{D} = [\boldsymbol{d_1}, \ldots, \boldsymbol{d_n}] \in \Re^{m \times n}$, the algorithm computes a matrix $\mathbf{A} = [\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_p}] \in \Re^{n \times p}$, where for each column $\boldsymbol{x}$ of $\mathbf{X}$, it returns a coefficient vector $\boldsymbol{\alpha}$ which is a solution of

$$\min_{\alpha \in \Re^n} ||\boldsymbol{x} - \mathbf{D}\alpha||_2^2 \ \texttt{s.t.}\ ||\alpha||_1 \leq \lambda \tag{3}$$

In our bag-of-visual-words model we used the *tf-idf* weighting [21] scheme, that has proven to be a very successful approach for image retrieval as well. The *tf* part of the weighting scheme represents the number of features described by a given visual word. The frequency of a visual word in the image provides useful information about repeated structures and textures. The *idf* part captures the informativeness of visual words, the ones that appear in many different images are less informative than those that appear rarely. This weighting scheme is generally applied only to integer counts of visual-words in images. Thus, in case of sparse coding the scheme had to be modified to handle the weight vector $\boldsymbol{\alpha}$. We found the same approach to be the best solution as in [22]. I.e. for the term frequency we simply used the normalized weight value for each visual word. For the inverse document feature measure, we found that counting an occurrence of a visual word as one, no matter how small its weight, gave the best results.

2.2 Classification

Based on the numerical and binary features of the images obtained through feature extraction, we perform vector space classification to predict modality classes of unseen images. Among the classification algorithms available in Weka [23], we found the support vector machine SMO to have the best standalone performance over the full feature space in cross-validation on ImageCLEF 2011 training data set. We used SMO with default settings for the rest of the experiments.

2.3 Evaluation Setting

Our experiments are based on the ImageCLEF 2011 medical modality classification data set [10], where there are 988 images in training- and 1024 images in the testing data set, which were taken from PubMed articles. The data set defines 18 different modality classes.

Table 1 shows how imbalanced the distribution of the images within the various modality classes are.

Table 1. Modality labels at ImageCLEF 2011 and their distribution

	Modality label		Training	
Group	Code	Description	#	%
Radiology	AN	angiography	11	1.1
	CT	computed tomography	70	7.1
	MR	magnetic resonance imaging	17	1.7
	US	ultrasound	30	3.0
	XR	X-ray	59	6.0
Microscopy	FL	fluorescence	44	4.5
	EM	electronmicroscopy	16	1.6
	GL	gel	50	5.1
	HX	histopathology	208	21.1
Photograph	PX	general photo	165	16.7
	GR	gross pathology	43	4.4
	EN	endoscopic imaging	10	1.0
	RN	retinograph	5	0.5
	DM	dermatology	7	0.7
Graphic	GX	graphs	161	16.3
	DR	drawing	43	4.4
Other	3D	3D reconstruction	32	3.2
	CM	compound figure (> 1 type of image)	17	1.7
Total		18	988	100.0

3 Results and Discussion

In this section we provide the results of the four different experimental settings. Table 2 shows the correctly classified percentage for each case and we included the result of the best submitted run [24] of the challenge as well. In all of the four cases we used all the features that has been introduced in the previous section.

Table 2. Results of the runs for the medical modality classification task. For the reference we have included the best performing run of the competition.

Run	Method	Accuracy
#1	sparse coded Affine-SIFT	**87.89**
#2	hard vector quantised Affine-SIFT	84.66
#3	sparse coded SIFT	86.42
#4	hard vector quantised SIFT	86.03
Best of ImageClef '11	SLR with Fisher Vector	86.91

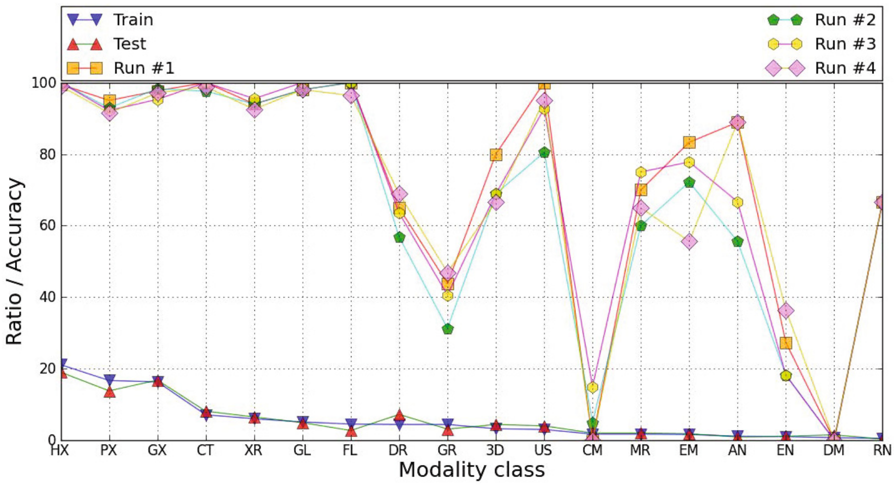


Fig. 1. Modality class distribution and the performance of different runs. Modality classes are sorted by support in descending order. For the names of modality classes, see Table 3.

For simplicity table 2 only shows the difference between the used features in the different runs.

Run #4 is our best best submitted run for the ImageCLEF 2011 medical image modality challenge. Although, sparse coding on the extracted SIFT features (run #3) does improve the classification accuracy, it is still lower than the best submitted run for the challenge.

It is important to note that using a fully affine-invariant feature descriptor will not necessarily improve the classification accuracy. On the contrary, as run #2 shows, the overall accuracy of the system significantly dropped by using ASIFT instead of SIFT. But as run #1 shows, if sparse coding is used instead of hard vector quantisation on ASIFT descriptors, the accuracy significantly improves and outperforms the state of the art. The performance of the runs broken down for the individual classes is shown in Table 3.

Table 3. Correctly classified images per category for the submitted runs. For each modality class, the result of the best performing run is typeset in bold.

	Ratio (%)		Run			
Modality class	train	test	#1	#2	#3	#4
3D : 3D render	3.2	4.4	**80.0**	68.9	68.9	66.7
AN: Angiography	1.1	0.9	**88.9**	55.6	66.7	**88.9**
CM: Compound figure	1.7	2.0	0.0	5.0	**15.0**	0.0
CT: Computed tomography	7.1	8.1	**100**	97.6	**100**	98.8
DM: Dermatology	0.7	1.5	0.0	0.0	0.0	0.0
DR: Drawing	4.4	7.2	64.9	56.8	63.5	**68.9**
EM: Electronmicroscope	1.6	1.8	**83.3**	72.2	77.8	55.6
EN: Endoscope	1.0	1.1	27.3	18.2	18.2	**36.4**
FL : Fluorescence	4.5	2.7	**100**	**100**	**100**	96.4
GL: Gel	5.1	4.9	98.0	98.0	**100**	98.0
GR: Gross pathology	4.4	3.1	43.8	31.3	40.6	**46.9**
GX: Graphics	16.3	16.8	97.7	**98.3**	95.3	97.1
HX: Histopathology	21.1	19.0	99.5	**100**	**100**	99.0
MR: MRI	1.7	2.0	70.0	60	**75.0**	65.0
PX: Photo	16.7	13.8	**95.0**	92.9	92.2	91.5
RN: Retiongraph	0.5	0.3	66.7	66.7	66.7	66.7
US : Ultrasound	3.0	4.0	**100**	80.5	92.7	95.1
XR: X-ray	6.0	6.5	94.0	94.0	95.5	92.5

4 Conclusion

In this paper, we proposed to extract different visual and textual features for medical image representation, and fusion the different extracted visual feature and textual feature for modality classification. To extract visual features from the images, we used some state-of-art methods like bag-of-visual words and some standard ones like colour histogram and introduced some heuristic representations of the images specialised for the ImageCLEF 2011 medical modality classification task.

We showed that using sparse coding instead of vector quantisation in the BoW representation for encoding the extracted affine-invariant feature descriptors with a given visual-word dictionary will increase the classification accuracy.

With the suggested feature extraction algorithms in this paper we have achieved a 87.89% accuracy that outperforms the state of the art.

Acknowledgements. Special thanks for Dr. Tikk Domonkos for his insightful comments and suggestions for the draft of the paper. Viktor Gál was supported by Marie Curie Initial Training Networks (ITN) Ref. 238819 (FP7-PEOPLE-ITN-2008).

References

1. Hersh, W.R., Müller, H., Jensen, J.R., Yang, J., Gorman, P.N., Ruch, P.: Advancing Biomedical Image Retrieval: Development and Analysis of a Test Collection. Journal of the American Medical Informatics Association 13(5), 488–496 (2006)
2. Pentland, A., Picard, R.W., Sclaroff, S.: Photobook: Content-based manipulation of image databases. International Journal of Computer Vision 18(3), 233–254 (1996)
3. Lakdashti, A., Moin, M.S.: A New Content-Based Image Retrieval Approach Based on Pattern Orientation Histogram. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2007. LNCS, vol. 4418, pp. 587–595. Springer, Heidelberg (2007)
4. Jain, A.: Image retrieval using color and shape. Pattern Recognition 29(8), 1233–1244 (1996)
5. Morel, J.-M., Yu, G.: ASIFT: A New Framework for Fully Affine Invariant Image Comparison. SIAM Journal on Imaging Sciences 2(2) (April 2009)
6. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, vol. 1, p. 22 (2004)
7. Jarrett, K., Kavukcuoglu, K., Ranzato, M.A., LeCun, Y.: What is the best multi-stage architecture for object recognition? In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2146–2153 (2009)
8. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1794–1801 (2009)
9. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least Angle Regression. ArXiv Mathematics e-prints (June 2004)
10. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A.G.S., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: CLEF 2011 Working Notes, Amsterdam, The Netherlands (2011)
11. Veltkamp, R.C.: A survey of content-based image retrieval systems. Content-based Image and Video Retrieval (2002)
12. Duda, R.O.: Use of the Hough transformation to detect lines and curves in pictures. Communications of the ACM (1972)
13. Chai, D., Ngan, K.N.: Face segmentation using skin-color map in videophone applications. IEEE Transactions on Circuits and Systems for Video Technology 9(4), 551–564 (1999)
14. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision, ICCV 1999, pp. 1150–1157. IEEE Computer Society, Washington, DC (1999)
15. Nister, D., Stewenius, H.: Scalable Recognition with a Vocabulary Tree. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2161–2168 (2006)
16. Jegou, H., Harzallah, H., Schmid, C.: A contextual dissimilarity measure for accurate and efficient image search. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), pp. 1–8 (2007)
17. Chum, O., Philbin, J., Sivic, J., Isard, M.: Total Recall: Automatic query expansion with a generative feature model for object retrieval. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8. IEEE (October 2007)
18. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008)

19. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
20. Gal, V., Solt, I., Gedeon, T., Nachtegael, M., Kerre, E.: Multi-disciplinary modality classication for medical images. In: CLEF 2011 Working Notes, Amsterdam, The Netherlands (2011)
21. Sivic, J., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. In: 9th IEEE International Conference on Computer Vision (ICCV 2003), pp. 1470–1477. IEEE Computer Society (2003)
22. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8 (2008)
23. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations 11(1), 10–18 (2009)
24. Csurka, G., Clinchant, S., Jacquet, G.: XRCE's Participation at Medical Image Modality Classification and Ad-hoc Retrieval Tasks of Image CLEF 2011. In: CLEF 2011 Working Notes, Amsterdam, The Netherlands (2011)

Mining Web Data for Epidemiological Surveillance

Didier Breton[1], Sandra Bringay[2,3], François Marques[1],
Pascal Poncelet[2], and Mathieu Roche[2]

[1] Nevantropic, France
[2] LIRMM – CNRS, Univ. Montpellier 2, France
[3] MIAp Department, AMIS Group, Univ. Montpellier 3, France

Abstract. Epidemiological surveillance is an important issue of public health policy. In this paper, we describe a method based on knowledge extraction from news and news classification to understand the epidemic evolution. Descriptive studies are useful for gathering information on the incidence and characteristics of an epidemic. New approaches, based on new modes of mass publication through the web, are developed: based on the analysis of user queries or on the echo that an epidemic may have in the media. In this study, we focus on a particular media: web news. We propose the Epimining approach, which allows the extraction of information from web news (based on pattern research) and a fine classification of these news into various classes (new cases, deaths...). The experiments conducted on a real corpora (AFP news) showed a precision greater than 94% and an F-measure above 85%. We also investigate the interest of tacking into account the data collected through social networks such as Twitter to trigger alarms.

1 Introduction

In the context of epidemiological surveillance, the analysis of relevant information is crucial to the decision-making process when an expert has to decide to trigger or not an alarm. The question we tackle in this article is the following: Can the flow of information exchanged on the Web be used to improve the daily monitoring of the epidemiological reality that epidemiologists themselves sometimes have difficulty to establish?

Health professionals can use news as new resources of information. However, they have to deal with the abundance of information. How to sort efficiently this pool of resources, to keep only the relevant information according to a specific issue?

The work presented in this paper is based on a collaboration between the Nevantropic company and the LIRMM laboratory. The company focuses on the development of operational monitoring of the environment at local and regional scales. In this collaboration framework, we are particularly interested in the automatic tracking of the evolution of H1N1 from heterogeneous resources of the Web. Our goal is to extract knowledge from news to provide new indicators

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 11–21, 2013.

for health authorities in order to assist them in the decision-making process. In this paper, we present a method for automatic detection of weak signals (task of epidemiological surveillance) from a news series. This method is based on pattern research to extract information from a news corpus and on the classification of these annotated sentences of news into topics (e.g. news cases, death...). We also investigate the interest of tacking into account the data collected through social networks such as Twitter to trigger alarms.

Our contribution comes in treefold: (1) to annotate the news according to a set of concepts; (2) to classify the news into categories, (2) to identify, count and locate the cases associated with an epidemic thanks to this classification. A brief state-of-the-art is presented in Section 2. In Section 3, we present the EPIMINING approach. The conducted experiments are described in Section 4, and discussed in Section 5. In Section 6, we discuss about the information that can be obtained from social networks and mainly focus on Tweets related to disease. We thus illustrate how such an information can be useful for improving the monitoring. Finally in Section 7, we conclude with future work.

2 Background

2.1 Context

Agencies managing the traditional systems of epidemiological surveillance (e.g. Institut National de Veille Sanitaire in France, European Influenza Surveillance Schema, US CDC Center for Diease Control and Prevention[1]) generally use virologic data, clinical information from medical reports or pharmacies in order to monitor an epidemic. For example, in France, one of the objectives of the Sentinel Network[2] composed of physicians and pharmacists is to monitor, according to the medical consultations, various diseases (e.g. asthma, diarrhea, influenza-like illness...). Even if these approaches are very effective, the proposed analyzes only focus on the events of the previous weeks and only few approaches are able to monitor outbreak in real-time [1].

Recently, Yahoo and Google have proposed systems which take advantage of the mass of information now available online for epidemiological surveillance. In 2008, [2] have examined the relationship between searches for influenza and actual influenza occurrences, using search queries from the Yahoo! search engine. The principle is based on the assumption that when a person has disease symptoms, he tends to query the Web like: "What are the symptoms of this disease?", "Which web sites deal with this disease?". Using the keywords chosen by the web users and their location, it is possible to define what are the trends of the users and consequently to predict potential outbreaks. [3] made a similar proposal by using the Google search engine to predict in advance the H1N1 epidemic peaks. The results of these two experiments showed that these

[1] http://www.invs.sante.fr/, http://www.ecdc.europa.eu/en/activities/surveillance/EISN/Pages/home.aspx, http://www.cdc.gov/

[2] http://websenti.b3e.jussieu.fr/sentiweb/

approaches predicted an increase of the epidemic up to 5 weeks in advance from the US CDC. Even if these approaches are very effective, they require to access to the content of the user's requests and also to have a sufficient number of users to define a prediction model.

2.2 State-of-the-Art

Different approaches are based on the extraction of information available in Web documents (news, reports, and so forth) in order to predict knowledge [4,5,6].

The principle generally used is the following one: From a large volume of Web documents, they extract features such as numbers and location. The collected numbers are often used to display with different colors (more or less dark) information that may be located on a map. For example, systems such as MedISys, Argus, EpiSpider, HealthMap, or BioCaster[3] support the global and real-time monitoring of a disease for a country. These systems are not intended to replace the traditional collaborative systems based on the exchange of official data, but allow to trigger a pandemic alert by integrating data from regions or countries for which official sources are limited or unavailable. However, these approaches suffer from some drawbacks. Because of the aggregate view, it is difficult to monitor an epidemic with a low granularity (time and space). For example, it could be interesting for the epidemiologist to identify which city or village develop new cases instead of having the information for a country. Moreover, most of the systems rarely support a fine result classification (e.g. difference between new cases or deaths). Knowing that in a country, there are occurrences of the H1N1 virus is relevant but, classifying the information retrieved in the news into new cases or new deaths is also informative. Finally, many methods return documents but not relevant segments in these documents. The epidemiologists have to read all the documents to find a section of interest.

In order to predict relevant information, the first stage consists in extracting relevant features in texts. For this extraction process, a lot of methods use patterns [7,4]. These ones match entity classes by using regular expressions and lists of terms from the studied domain. For instance, the lists include verbs of infection, named entity and so on [4]. To extract information and build knowledge bases of epidemiologic studies, other methods use machine learning approaches [8]. This kind of supervised method has an important limit: a lot of labeled data are necessary in order to learn a model.

Our objective in this paper is to address the limitations of the previous approaches. We are interested in the echo that may have an epidemic in the media through news that we classify automatically according to their content into very specific categories (i.e. new cases, new deaths). For this, we first use an extraction method to annotate the news based on pattern recognition and a classification algorithm that takes into account the number of patterns retrieved in the news.

---

[3] `http://medusa.jrc.it`,
`http://biodefense.georgetown.edu/projects/argus.aspx`,
`http://www.epispider.org`, `http://www.healthmap.org`,
`http://www.biocaster.org`

The classification based on an unsupervised approach is not done at the level of the document but at the level of the segments in the documents. Finally, in order to assist the decision-maker, the epidemiologist, we provide different visualizations of the results either as graphical statistics (histogram, pie chart), or as geographical representations of events using GoogleMap.

3 The EPIMINING Approach

In this section, we present the overall EPIMINING approach detailed in the Figure 1.

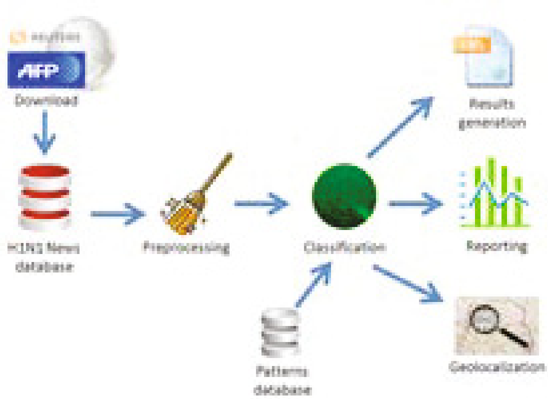


Fig. 1. EPIMINING approach

3.1 Acquisition and Pre-processing of the Corpus

To feed the News database, we queried sites such as Reuters or the French equivalent AFP. We used keywords associated with the disease (e.g. swine flu, H1N1, influenza...). We tokenize and tag words that appear in the retrieved news with the TreeTagger tool [9]. For example, let us consider the subpart of the second sentence of the news presented in figure 2:

"10 deaths had occurred in adults all under the age of 65 in England"

The associated lemmatized sentence, composed of the original form of each word (i.e. first element), the grammatical category (i.e. second element) and the lemma (i.e. third element) is:

"10/CD/Card deaths/NNS/death had/VHD/have occurred/VVN/occur in/IN/in adults/NNS/adult all/RB/all under/IN/under the/DT/the age/NN/age of/IN/of 65/CD/Card".

Ten dead as H1N1 flu returns to Britain

Recommander | Une personne recommande ça. Soyez le premier de vos amis.

Tweet 0 | Share | Share this | +1 0 | Email | Print

LONDON | Sat Dec 11, 2010 11:48am EST

(Reuters) - The H1N1 swine flu virus which swept the globe last year has returned to Britain with 10 people dying in the last six weeks, health officials said Saturday.

Britain's Health Protection Agency said the 10 deaths had occurred in adults all under the age of 65, most of whom had underlying health issues.

"Over the last few weeks, we have seen a rise in the number of cases of seasonal flu both H1N1 (2009) and flu B in the community," Professor John Watson, head of the HPA's respiratory diseases department, said in a statement.

Factbox

Factbox: Tobacco - One of the world's biggest health threats
Thu, Nov 25 2010

Related News

Doctors encouraged pregnant women to get flu shot
Thu, Dec 2 2010

Haiti's cholera part of old pandemic:

Fig. 2. An example of H1N1 news

3.2 Annotation of the News

Pre-treated news are automatically annotated thanks to a Pattern Database which enables to identify the relevant concepts. We apply an approach similar to the one described in [10] who details different Information Extraction (IE) tools in order to find specific information in free texts. Like our method, the developed tools used patterns associated to part-of-speech knowledge. Note that the EPIMINING system described in this paper is more specific to the epidemiology domain. To recognize patterns in documents, we rely on their linguistic characteristics and other syntactic rules of their arrangement. Specifically, the tagged documents are parsed in order to detect patterns. The analysis started by applying a set of syntactic rules to locate all the patterns present in different sections of the document. A filter is then applied to favour the longest pattern among several patterns sharing the same lemmatised words. For example, in the sentence of the Figure 2, we identify the concept PERSON thanks to the presence of the lemmatized word "adult". Similarly, the concept YEARS_OLD is retrieved via the pattern series: <PERSON> followed by the expression "under the age of" followed by the number 65.

The Pattern Database is composed of patterns specified by an expert. These patterns were identified after a textual analysis of the news content. They take into account the specificities of the news regarding the other types of text documents. Applying this method, we identify in the previous example the concepts: NUMBER, DEATH, PERSON, YEARS_OLD, CITY. To refine the information about the location in the news, we use a database of geographic information (Geolocalisation database). After this step, documents are labelled for an easier classification: When this was possible, each sentence is associated with a number of sick and dead people, a location, a date... Finally, we obtain the following annotations:

"<NUMBER>10</NUMBER><PERSON><DEAD>death</DEAD> <AGE> under 65</AGE></PERSON> <CITY> London </CITY>".

3.3 Classification

A news can contain information, which can be classified into various categories. For example, we can find in the same new information about sick and dead patients. Consequently, the news classification at the document level is often not relevant. To obtain a fine classification, we decide not to classify the news but sentences of these news. The classification is performed as follows: Each class is associated with a set of patterns. If patterns of a defined class are retrieved in a news, the one is associated to this class. For instance, the news of the Figure 2 is associated to the class Death because we have found the news with the concept DEAD. For each association between a sentence and a class, we calculate the Epimining score according to the following heuristic. The score equals 1 if all the elements that are expected are found in the sentence (e.g. for the association between a sequence containing a date, a number of death, a geographical location and class Death). The score is based on the reliability of extracted information. For example, if the location is not in the sentence, the search is expanded in the nearest sentence to find the missing information and the score is decreased.

4 Experiments

In order to evaluate the performance of our approach, two data sets in French were used for experiments: a database of 510 AFP news over the period of September 2009 to February 2010 and a database of 353 Reuters news over the period of January 2009 to February 2010. To analyze the quality of the returned results, 477 AFP news, and 7147 sentences have been manually annotated. The objective was to evaluate the news classification into four categories. The first two ones depends on when the cases mentioned in the news are listed: "New cases" corresponds to the description of information about new patient at a given time and "Report" corresponds to older cases. The last two categories correspond to the categorization of the patient: "Dead patient" and "Sick patient". Two types of evaluations have been conducted (1) by considering the documents as objects to be classified and (2) by considering the sentences. To evaluate the results of these two classifications, we measure the precision (ratio of relevant documents found on the total number of selected documents), recall (ratio of relevant documents found in the total number of relevant documents), and the F-measure (harmonic average between precision and recall).

In Table 1, results of the tests conducted on the news classification are reported. The best results are obtained for classes "Report" and "Dead patient". This is justified by the fact that the distinction between illness and death is not always present in the news and by the fact that the concept of novelty is more difficult to detect. Even when the analysis is conducted by an expert, the difference between the two classes is not necessarily obvious to capture.

Table 2 presents the experiments conducted on the classification of the sentences. We worked with different EPIMINING score values corresponding to the search for patterns in different sentences close to the evaluated segment. With a high confidence score (i.e. [50..100]), we obtain the best precision (83.6%).

Finally, Tables 1 and 2 show that the EPIMINING approach focuses on precision. Indeed, the patterns are often quite restrictive to return relevant elements. To increase the recall, we can consider the sentences with a large EPIMINING confidence as shown in Table 2 .

Table 1. News classification

Classes	Retrieved and relevant	Retrieved	Relevant	Precision	Recall	F-Measure
Dead	100	106	128	94.3%	78.1%	85.5%
Ill	43	55	65	78.2%	66.2%	71.7%
Report	88	103	114	85.4%	77.2%	81.1%
New	48	59	78	81.4%	61.5%	70.1%

Table 2. Sentence classification

Confidence	Retrieved and relevant	Retrieved	Relevant	Precision	Recall	F-Measure
[0..25[	20	46	280	43,5%	7,1%	12,3%
[25..50[	58	97	280	59,8%	20,7%	30,8%
[50..100]	112	134	280	83,6%	40,0%	54,1%
[0..100]	190	277	280	68,6%	67,9%	68,2%

5 Discussion

A prototype dedicated to healthcare professionals was set up by the Nevantropic company. Figure 3 shows an interface of this tool that presents various indicators that can be used for decision-making.

On the left, the evolution of the number of cases of sick and died people identified through the news dealing with H1N1 are presented over several months or years. On the right, the cases are located on a GoogleMaps at a given time.

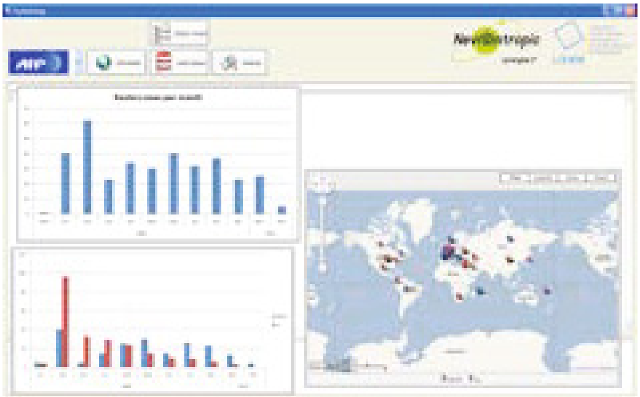


Fig. 3. EPIMINING tool

Of course, only the indicators derived from the news classification are presented on this image but of course, it is the combination of several indicators that make sense for healthcare professionals who must take a decision. For example, the tool can be used by epidemiologists who should or should not trigger an alarm or by physicians to guide their diagnosis during the visit of a patient in suspected cases of epidemic in the country where he travel back. The proposed architecture for the monitoring of H1N1 is of course adaptable to other types of epidemics.

The type of approach presented in this paper, based on the exploitation of massive data published on the Web, like the approaches proposed by Yahoo and Google, are relevant because they help to early alert health authorities. The results of these methods must be considered as new and indispensable sources of information that have to be crossed with more traditional sources of information provided by the agencies managing the traditional systems of epidemiological surveillance, either to confirm, disprove or in most cases to clarify. These methods are especially useful in geographic areas that do not have a conventional surveillance infrastructure but where the deployment of the Internet is already well advanced.

6 What's about a More Real Time Information?

In this section, we consider another kind of information that can be very useful for helping to evaluate the propagation of epidemics. In the previous sections, we focused on information available on news. That means that this information is basically evaluated by a journalist grouping and aggregating together different data or information. In an other way, the development of social and collaborative Web 2.0 underlines the central and active role of users in collaborative networks. Easy to create and manage these tools are used by Internet users to communicate about themselves. Thus, this data represents an important source of information that can be used for helping epidemiological surveillance. For instance, Twitter is a platform for microblogging, i.e. a system for sharing information where users can either follow other users who post short messages (140 characters) or can be followed. Furthermore, Tweets are associated with meta-informations such as date or location. For instance, from Tweets we can extract the following messages "*I have a huge headache...*" expressed in New York in November or "*... gasrointestinal problems are not good. go 2 a doc!*" from Los Angeles in December.

We have investigated this new kind of media. Basically by using the MeSH (*Medical Subject Headings*) National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a twelve-level hierarchy that permits the search to be carried out at various levels of specificity. At the most general level of the hierarchical structure are very broad headings such as "Anatomy" or "Mental Disorders". More specific headings are found at more narrow levels, such as "Ankle" and "Conduct Disorder". In 2011, 26,142 descriptors were available in MeSH. We conducted some experiments by focusing on the "Disease" part of the hierarchy and we queried Twitter by using all the terms of the corresponding hierarchy. We thus collected 1,801,310 tweets in English from January 2011 to February 2011.

For instance, Figure 4 reports the results of the number of occurrences of terms "Pneumonia", "Leukemia" and "Hepatitis" over the period. It is interesting to notice that, for the decision maker, two peaks are important for the "Hepatitis" (i.e. end of january 2010 and beginning of january 2011). By using the same tools as in Epimining, we can easily locate the origins of these tweets as illustrated in Figure 5.

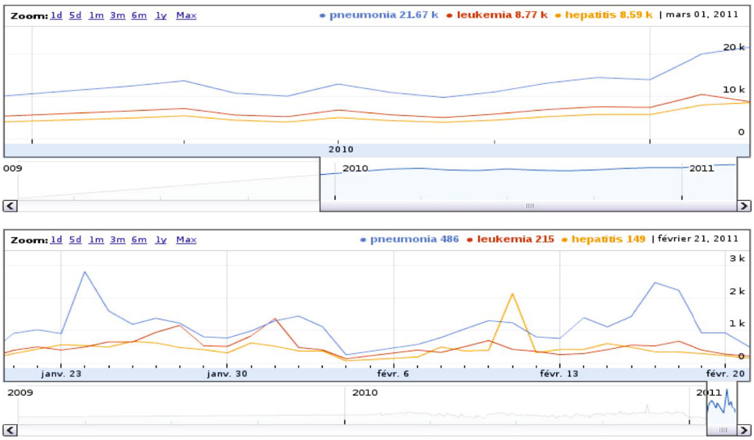


Fig. 4. Occurrences of diseases "Pneumonia", "Leukemia", 'Hepatitis" from January 2011 to February 2011

Fig. 5. Localization of Tweets for Pneumonia

Interestingly, we can notice that lots of tweets are exchanged in Ecuador or in Russia. A closer analysis highlights that one alert have been triggered by FluTrackers[4] in Ecuador and that, following the exchanges of tweets, a same kind of alert has been triggered in Russia.

7 Conclusion and Future Work

In this paper, we have proposed a new approach, called Epimining, to monitor epidemics, based on automatic knowledge extraction and news classification. Epimining have been illustrated by a prototype for monitoring indicators on the H1N1 epidemic. The advantage of our approach, by measuring the echo of an epidemic in the media, is to be complementary to traditional surveillance networks and user's queries analysis proposed by Yahoo and Google systems for instance. The perspectives associated to our proposal are numerous. We can easily improve the classification with learning methods in order to automatically extract the representative patterns of a class. In addition, we plan to extend our approach to other types of textual datasets (e.g. weblogs). We also plan to combine this method with the ones based on other types of datasets (air transport, meteorological, entomological data...). Finally, to answer to our initial question, we can say that the data issued from the web seem to be relevant variables, which can be included into the models of epidemics to better anticipate and predict their dynamics. Furthermore, as illustrated in the last section of the paper, it is more and more important to consider social network to improve the anticipation of epidemics. Knowing, for instance, that some people have fever, headache, gastrointestinal problems, muscle pain at the same time and in the same location is clearly important to better anticipate the propagation of an epidemy.

[4] `http://www.flutrackers.com/forum/showthread.php?t=158136.`

References

1. Tsui, F.C., Espino, J., Dato, V.M., Gesteland, P.H., Hutman, J., Wagner, M.: Technical description of rods: A real-time public health surveillance system. The Journal of the American Medical Informatics Association 10, 399–408 (2003)
2. Polgreen, P., Chen, Y., Pennock, D., Forrest, D.: Healthcare epidemiology: Using internet searches for influenza surveillance. Invited Article in Clinical Infectious Diseases – Infectious Diseases Society of America 47, 1443–1448 (2008)
3. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature, 1012–1015 (2009)
4. Collier, N., Doan, S., Kawazoe, A., Goodwin, R., Conway, M., Tateno, Y., Ngo, Q., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., Taniguchi, K.: Biocaster: detecting public health rumors with a web-based text mining system. Bioinformatics 24(24), 2940–2941 (2008)
5. Zant, M.E., Royauté, J., Roux, M.: Représentation événementielle des déplacements dans des dépêches épidémiologiques. In: TALN 2008, Avignon (2008)
6. Zhanga, Y., Danga, Y., Chena, H., Thurmondb, M., Larsona, C.: Automatic online news monitoring and classification for syndromic surveillance. Decision Support Systems 47(4), 508–517 (2009)
7. Turchin, A., Kolatkar, N.S., Grant, R.W., Makhni, E.C., Pendergrass, M.L., Einbinder, J.S.: Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. Journal of the American Medical Informatics Association: JAMIA 13(6), 691–695 (2006)
8. Lu, Y., Xu, H., Peterson, N.B., Dai, Q., Jiang, M., Denny, J., Liu, M.: Extracting epidemiologic exposure and outcome terms from literature using machine learning approaches. Int. J. Data Min. Bioinformatics 6(4), 447–459 (2012)
9. Schmid, H.: Probabilistic Part-of-Speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, pp. 44–49 (1994)
10. Muslea, I.: Extraction patterns for information extraction tasks: A survey. In: AAAI 1999 Workshop on Machine Learning for Information Extraction, pp. 1–6 (1999)

Getting a Grasp on Clinical Pathway Data: An Approach Based on Process Mining

Jochen De Weerdt[1], Filip Caron[1], Jan Vanthienen[1], and Bart Baesens[1,2]

[1] Department of Decision Sciences and Information Management, KU Leuven
Naamsestraat 69, B-3000 Leuven, Belgium
Jochen.DeWeerdt@kuleuven.be
[2] School of Management, University of Southampton
Highfield Southampton, SO17 1BJ, United Kingdom

Abstract. Since healthcare processes are pre-eminently heterogeneous and multi-disciplinary, information systems supporting these processes face important challenges in terms of design, implementation and diagnosis. Nonetheless, streamlining clinical pathways with the purpose of delivering high quality care while at the same time reducing costs is a promising goal. In this paper, we propose a methodology founded on process mining for intelligent analysis of clinical pathway data. Process mining can be considered a valuable approach to obtain a better understanding about the actual way of working in human-centric processes such as clinical pathways by investigating the event data as recorded in healthcare information systems. However, capturing tangible knowledge from clinical processes with their ad hoc and complex nature proves difficult. Accordingly, this paper proposes a data analysis methodology focussing on the extraction of tangible insights from clinical pathway data by adopting both a drill up and a drill down perspective.

Keywords: process mining, clinical pathways, healthcare information systems, event logs, fuzzy miner.

1 Introduction

Worldwide, the healthcare sector goes through a major reform. This has many reasons. First, the costs of healthcare are rising up to 15% in the United States (US) and close to 10% in Europe [1]. This is due to the increasing needs of a greying population, but also due to technological and pharmacological innovations that are really widening the possibilities for diagnosis and treatment. Second, there is a shift in the role of patients, going from a more passive role into a role of active consumers of care. Patients want to be informed and involved. Third, there is growing attention to quality and safety. The main drive comes from the *to err is human* report from the Institute of Medicine[1] [2]. This report indicated that as many as 44.000 to 98.000 US citizens die in hospitals each year as the result of medical errors. Even using the lower estimate, this would make medical

[1] http://www.iom.edu

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 22–35, 2013.

errors the eighth leading cause of death in the US - higher than motor vehicle accidents (43.458), breast cancer (42.297) or AIDS (16.516). The report was publicly discussed in the Senate and was the start of an overall hospital reform. Most important in the discussion is that people are not blamed - *to err is human* indeed - but that the focus should be on improving the system.

An excellent way to do this is learning from past experience. Currently, it can be observed that with the growing implementation of integrated healthcare information systems, vast amounts of data are becoming available about the actual way of working in clinical pathways. These data form the cornerstone of this study. Accordingly, the notion of a clinical pathway is a crucial element. The terminology has its origins in methodologies such as PERT (Project Evaluation and Review Technique) and CPM (Critical Path Method), but transformed into "clinical" instead of "critical" pathways because of the very specific nature of healthcare. Clinical pathways are formally defined as *a complex intervention for the mutual decision making and organization of care for a well-defined group of patients during a well-defined period* by the European Pathway Association[2].

In this study, we propose an approach for deriving useful insights from clinical pathway data by making use of process mining techniques. Process mining is a relatively young research area [3], which lies at the intersection of data mining and Business Process Management (BPM) [4]. It consists of a family of analysis techniques for analyzing event logs as recorded by the logging infrastructures of information systems. Since these techniques rely on real data, they are capable of providing insight into the actual way of working in the context of a certain business process. In this paper, we will describe a methodology based on state-of-the-art process mining techniques for the analysis of clinical pathway data. The main contribution of this study is the development of solution strategies for dealing with the extremely unstructured nature of clinical pathway data.

2 Related Work

Process Aware Healthcare Information Systems. A fair share of information systems implemented in healthcare organizations can be described as process aware. This is because process aware information systems not only encompass traditional workflow management systems, but also include systems that provide much more flexibility. Accordingly, once an information system can be described as having an explicit notion of the process it supports, it can be described as process aware [5]. A such, many healthcare information systems fit within this definition. Since clinical pathways are inherently heterogeneous and multi-disciplinary in nature, the goal of IT support for healthcare processes is not to control the course of the process entirely, but to assist healthcare professionals by reducing cognitive overload and improving the basis for their decisions [6]. In this way, process orientation can be considered as a beneficial approach towards streamlining clinical pathways with the purpose of delivering high quality care while at the same time reducing costs [7]. While business process support

[2] http://www.e-p-a.org

for structured processes (e.g. manufacturing, logistics) has always been an important research topic, the growing importance of service organizations such as healthcare has triggered the need for different approaches towards business process support [8,9]. Because of the human-centric nature of such processes, they contain much more flexibility, alternative routings, loops, human judgement and variability than traditional business processes. Accordingly, the analysis of the actual way of working by making use of the data captured by healthcare information systems is promising. However, traditional business process analysis techniques come short in realizing this goal and therefore this paper proposes a novel analysis methodology based on process mining.

Process Mining. The field of process mining can be best defined as a broad family of techniques for the analysis of event logs as recorded by the logging infrastructures of information systems. These techniques can be broadly categorized into three groups according to three commonly distinguished process mining tasks: discovery, conformance and enhancement [3]. The most important learning task is called process discovery [10] which entails the extraction of control-flow models from such event logs. In the process mining literature, a lot of attention has been paid to the development of process discovery techniques [11,12,13]. However, discovery tasks can also focus on other aspects of an event log, for instance on organizational information [14]. Conformance [15] is a second important process mining task. Hereto, a process model is compared with the data in the event log with the purpose to verify whether reality conforms to the model and vice versa. Finally, the idea of enhancement tasks is to extend or improve process models with other information about the actual process as recorded in the event log. For instance, the addition of a performance perspective enhances a process model and provides the analyst with different insights.

Process Mining in Healthcare. Process mining techniques have been applied in a healthcare context. A first study by Mans et al. [16] shows how different process mining techniques such as HeuristicsMiner, social network analysis and dotted chart analysis allow for obtaining insights into care flow data. Another study by Rebuge and Ferreira [17] also describes a methodology for the analysis of business processes in a healthcare environment. The methodology consists of seven phases with its main asset being the application of sequence clustering techniques. Further, Bose and van der Aalst [18] propose the use of fuzzy mining and trace alignment for investigating clinical pathway data. Finally, Caron et al. [19] demonstrate the applicability of various process mining techniques to healthcare data by adopting both a department and a treatment based focus. This study differs from previous studies because it shows the benefits of both a drill up and a drill down perspective on the data relying on control-flow discovery with the Fuzzy Miner and networked graph visualizations.

3 Description of the Clinical Pathway Data

The data set concerns real data of a gynecological oncology process at the AMC hospital in Amsterdam, The Netherlands. It was first used in [16], but recently the data set was made publicly available (doi:10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54) for the first Business Process Intelligence Challenge (BPIC'11). The data contains 150.291 events of 1.143 patient treatment processes related to individuals diagnosed with cancer pertaining to the cervix, vulva, uterus and/or ovary. Each case in the event log corresponds to a single patient and as such, the data presents a wide variety of care activity sequences. In the remainder of this section, the three most important dimensions of the data are outlined.

Diagnosis. Each case in the data set contains information on the type of disease the patient is diagnosed with. The related attributes are denoted *Diagnosis code* and *Diagnosis*. The data presents a total of eleven different diagnosis codes (e.g. M11, M12, 823, etc.). *Diagnosis* is a textual description which specifies the diagnosis code, taking values such as "adenocarcinoma stage Ia" or "clear cell carcinoma". It should be noted that the data contains up to 16 diagnosis code - diagnosis combinations for a single patient (denoted as *diagnosis code:1 to 16*). Accordingly, a single case might contain different codes. We observe 38 distinct diagnosis code combinations, for instance {M16, 821}. Table 2 presents an overview of the diagnosis codes detailing the region, example diagnoses and the number of cases showing this diagnosis code.

Treatment. Next to diagnosis information, one can also find details concerning the treatments of each of the patients. However, in contrast to diagnosis, the data only provides a treatment code and no further information. As such, the treatment perspective is more difficult to analyze. On top of this, there exist 46 distinct treatment codes which form 236 distinct treatment code combinations in similar fashion as the diagnosis code combinations.

Departments. The final important data dimension is organizational in nature, i.e. the departments that are involved in the clinical pathways. Each event in the log contains an attribute "org:group" which denotes the department where the corresponding activity was performed. In the data, one can find 43 distinct organizational units. The most frequently observed departments are depicted in Table 3.

A very specific feature of the data is that events pertaining to certain departments occur in bursts. For instance, sets of blood diagnosis tests performed by the *General Lab Clinical Chemistry* department are often found. Similarly, bursts of radiotherapy-, nursing-, operating room-related and many other types of events can be observed. This data characteristic will be further employed in the next section.

Table 1. Excerpt of the event log

Case ID	Event name	Dept.	Timestamp	Diagn. code	Diagnosis	Treatm. code	Age	...
0	1st polyclinic consult	Radiotherapy	03/01/2005	M13	cervical malign.	23	33	...
0	administrative reg.	Radiotherapy	03/01/2005	M13	cervical malign.	23	33	...
0	gynec. cost assign.	Nursing ward	03/01/2005	M13	cervical malign.	23	33	...
0	ultrasonography	Obstr.&gyn. clinic	03/01/2005	M13	cervical malign.	23	33	...
0	1st consult	Nursing ward	03/01/2005	M13	cervical malign.	23	33	...
...	...	...	...	...	...	...	...	...

Table 2. Diagnosis codes in the clinical pathway data

Code	Region	Example diagnoses	# cases
M11	vulva	squamous cell carcinoma, borderline malignancy	176
M12	vagina	squamous cell carcinoma, adenocarcinoma	22
M13	cervix	squamous cell carcinoma, malignant neoplasms	368
M14	corpus uteri	adenocarcinoma, clear cell carcinoma	145
M15	corpus uteri, myometrium	sarcoma	17
M16	ovary	squamous cell carcinoma, non-epithelial malignancy	235
106	cervix, vulva, corpus uteri, vagina	squamous cell carcinoma, borderline malignancy	298
821	ovary	serous and mucinous squamous cell carcinoma, neoplasms	48
822	cervix	squamous cell carcinoma, adenocarcinoma	131
823	corpus uteri, ovary, endometrium	(serous) adenocarcinoma	16
839	ovary, vulva	serous adenocarcinoma, malignant neoplasms	21

Table 3. Number of events pertaining to organizational units by frequency

Organizational unit	# events
General Lab Clinical Chemistry	94917
Nursing ward	31066
Obstetrics & Gynaecology clinic	7065
Medical Microbiology	4170
Radiology	3171
Radiotherapy	2233
Internal Specialisms clinic	2146
Pathology	1975
Operating rooms	942
Pharmacy Laboratory	498
Recovery room / high care	495
Nuclear Medicine	281
Special lab radiology	279
...	...

4 Analysis Methodology and Results

Our data analysis approach combines two important strategies for extracting tangible knowledge from clinical pathway data. First, drill up is applied in order to get insight into the general behavior of the healthcare process. In a second phase, a drill down approach is described that centers on a certain part of the data.

4.1 Complexity of Clinical Pathway Data

The crucial challenge for data analysis in the context of clinical pathways is the complexity of the data. This is because clinical pathways are inherently ad hoc, multi-disciplinary and strongly human-centric. Because of these characteristics, almost every observed clinical pathway is unique, which is also the case for the the data set employed in this study. On top of that, the original data set contains 624 different activity types. Further, these activity types as registered in the different departments are not always of the same granularity. A final element that complicates the data analysis significantly is the fact that we can only observe care activities executed within the AMC hospital. However, it is undoubtedly reasonable to assume that other care activities in peripheral hospitals, by GP's, etc. are being executed but not registered in the data. Because of these data complexities, there is a need for versatile data analysis methods. In this study, it is shown how process mining techniques can be used, both in a drill up as well as in a drill down mode. Furthermore, we demonstrate the use of networked graphs for visualizing sets of cases and their respective characteristics.

4.2 Drill Up Analysis

Both in [16] and [19], it is shown that the straightforward application of existing process discovery techniques is infeasible for the data at hand. Even the stronger generalization capabilities of Fuzzy Mining [20] prove not very helpful. Therefore, we show how abstraction applied in a data preprocessing step can be beneficial in order to obtain general, but useful insights based on the entire data set.

Data Preprocessing. Realizing abstraction in a data preprocessing phase consists of replacing the bursts of events belonging to the same organizational unit by the name of the organizational unit itself. In this way, a clinical pathway in terms of the unique activities performed by different organizational units is transformed into sequences of departments.

A General View Using Process Discovery. As stated earlier, process discovery is the most important asset of the process mining domain. Process discovery is defined as the extraction of control-flow models from event logs. Note that these techniques make use of different process modeling notations (e.g. Petri nets, heuristic nets, fuzzy nets, etc.) in order to represent the discovered model. In this case, we applied the Fuzzy Miner to the transformed clinical pathways. The resulting fuzzy net is depicted in Figure 1. Note that the nodes contain a significance value between 0 and 1. Further, the figures on the edges indicate the edge significance and correlation, also ranging between 0 and 1.

With the purpose to increase the comprehensibility, we restricted the visualization to the eleven most frequent departments. Together with the abstraction power of Fuzzy Miner, the discovered graph provides some interesting insights with respect to the gynecological oncology process under investigation:

- A majority of the patients first visit the obstetrics and gynecology department.
- From top to bottom, we can clearly observe the diagnostic-therapeutic cycle characteristic to the majority of care processes. The nursing wards have a pivotal role in between diagnostics and therapeutics.
- From a diagnostic perspective, lab analyses (majorally blood sample tests) and to a lesser extent radiology are essential elements of disease typification in the context of gynecological oncology.
- Despite the fact that the data cover patients with very comparable diagnoses (i.e. gynecological cancers), streamlined clinical pathways cannot be observed, even in terms of involved departments.

4.3 Drill Down Analysis

As described in the previous section, drill up is a valuable approach for extracting general knowledge from care process data. Nevertheless, due to the characteristics of the data, intelligent drill down into specific parts of the data is bound

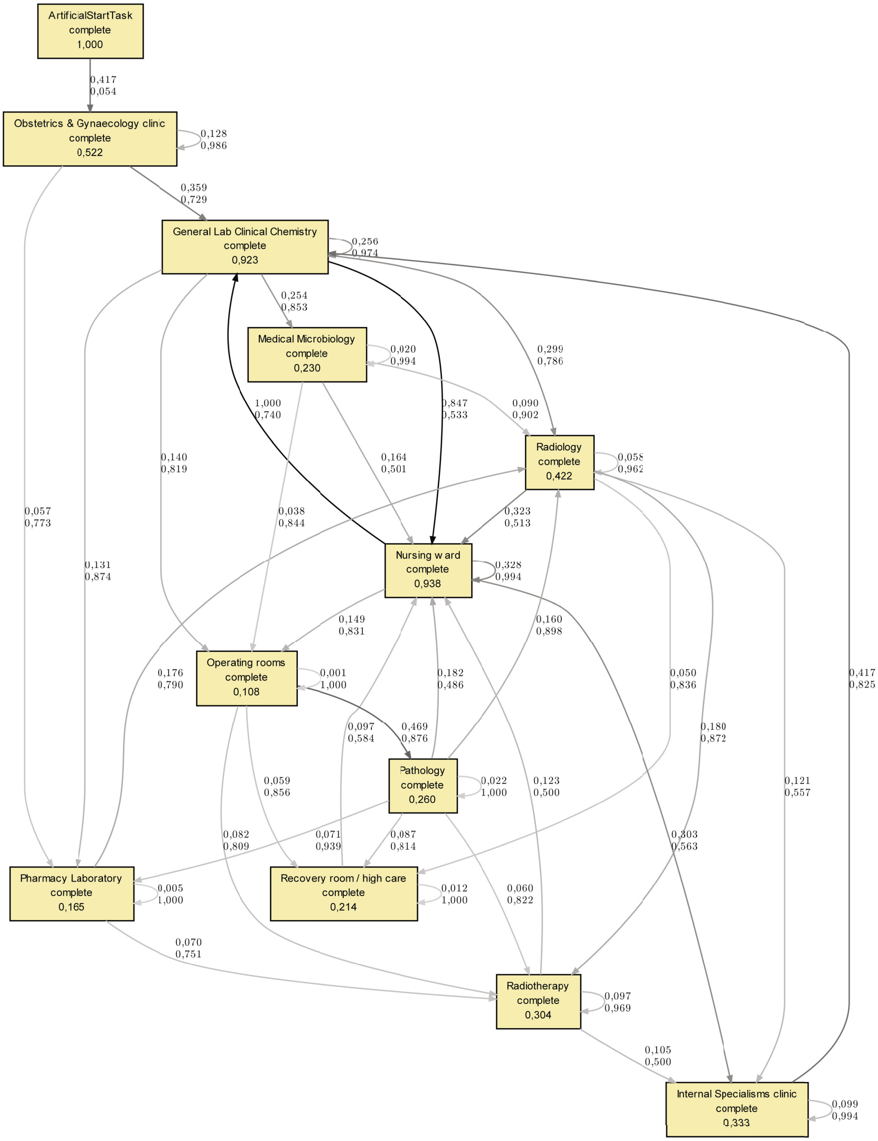


Fig. 1. Fuzzy process model showing the relations between frequently occurring organizational units

to provide even more interesting insights. Hereto, we demonstrate a particular focus on the therapeutic side of the clinical pathways. Since the data provides only limited information on specific treatments (only treatment codes, without any explanation), it is opted to investigate this perspective in more detail.

Focalizing on Therapeutic Activities. To adopt the focus on treatment, therapeutic activities need to be singled out. By inspection of the cases, it can be noticed that three different types of treatments can be identified: radiotherapy, chemotherapy and surgery. For radiotherapy, the selection of activities was rather straightforward since the granularity of the events is relatively coarse-grained. The data contains events such as *teletherapie - megavolt fotonen bestrali* and *brachytherapie - interstitieel - intensi*. Further, the radiotherapy department also carries out hyperthermia treatments, which are strictly speaking not radiotherapeutic, but often used in combination with radiation therapy. For chemotherapy, the selection of appropriate events was slightly more difficult due to the fact that this type of therapy is scattered between different organizational units. Nonetheless, we identified two important chemotherapeutic activities, viz. *paclitaxel* and *doxorubicine*. Finally, also surgical treatments should be taken into account. However, looking at the events pertaining to the Operating rooms department, there clearly exist two different types of procedures. On the one hand, the data set shows a multitude of diagnostic surgical procedures, whereas on the other hand only therapeutic operations are of interest given our current focus. Nonetheless the thin line between both, we were able to distinguish between the diagnostic or therapeutic nature of procedures by investigating the names of the events. For example, hysterectomies and vulvectomies were considered as therapeutic surgical activities, while hysteroscopies and urethrocystoscopies were not.

After singling out these therapeutic activities, 477 cases could be observed for which at least one therapeutic activity has taken place. The event log consisting of all these events was used to visualize the therapeutic activities by means of a process model. However, due to the large number of different surgical procedures, we renamed all these events to *surgery*, an abstraction which allows for more useful visualizations. The resulting process model as obtained with fuzzy mining is depicted in Figure 2. In contrast to Figure 1, the fuzzy net is slightly adapted by replacing the original figures in the nodes and on the arcs by more insightful statistics. As such, the nodes contain their frequency of occurrence, while the figure on each of the edges of the graph shows the number of times a case followed the transition from the source node to the target node.

The analysis allows for the formulation of the following findings:

– The fuzzy net shows a number of possible therapeutic choices. However, because the process model does not contain any such paths, it can be concluded that the combination of surgery or chemotherapy with radiotherapy occurs highly infrequent.
– The use of chemotherapy is rather limited despite the fact that regimens, i.e. combinations of different chemotherapy drugs, are often recommendable.

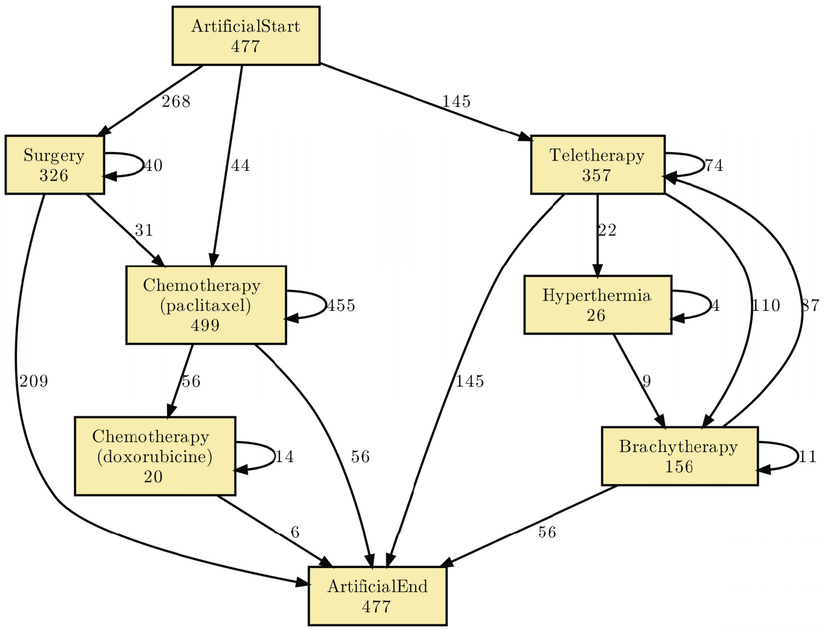


Fig. 2. Adapted fuzzy process model showing the relations between different therapeutic activities

In this way, we could only observe very few chemotherapeutic combinations of Paclitaxel with Doxorubicin. It should be further investigated why chemotherapy is underrepresented. One possible explanation might be that chemotherapeutic procedures might be carried out in peripheral hospitals, thus not captured in the data.

Visualizations Using Networked Graphs. A second drill down approach consists of visualizing a subset of cases by means of a networked graph. This methodology is useful because it allows to visualize cases from different angles by supplementing control-flow information with other perspectives. The construction of a networked graph consists of three steps. First cases are selected based on some criterion. In this case, we considered all cervical cancer cases. Secondly, a Euclidean distance matrix is constructed denoting the distance between each pair of cases. This matrix is built by making use of the MRA (Maximum Repeat Alphabet) technique as proposed in [21]. The MRA technique relies on the identification of specific patterns which characterize the traces. Notwithstanding the fact that the authors employ the method for clustering log traces, we use the underlying distance matrix to construct a networked graph. Such a network graph connects nodes which represent a case in the data. For comprehensibility reasons, sparsification is applied in order to reduce the number of connections between the nodes because otherwise a fully connected graph is obtained. In this case, we applied K-nearest-neighbors with $K = 2$.

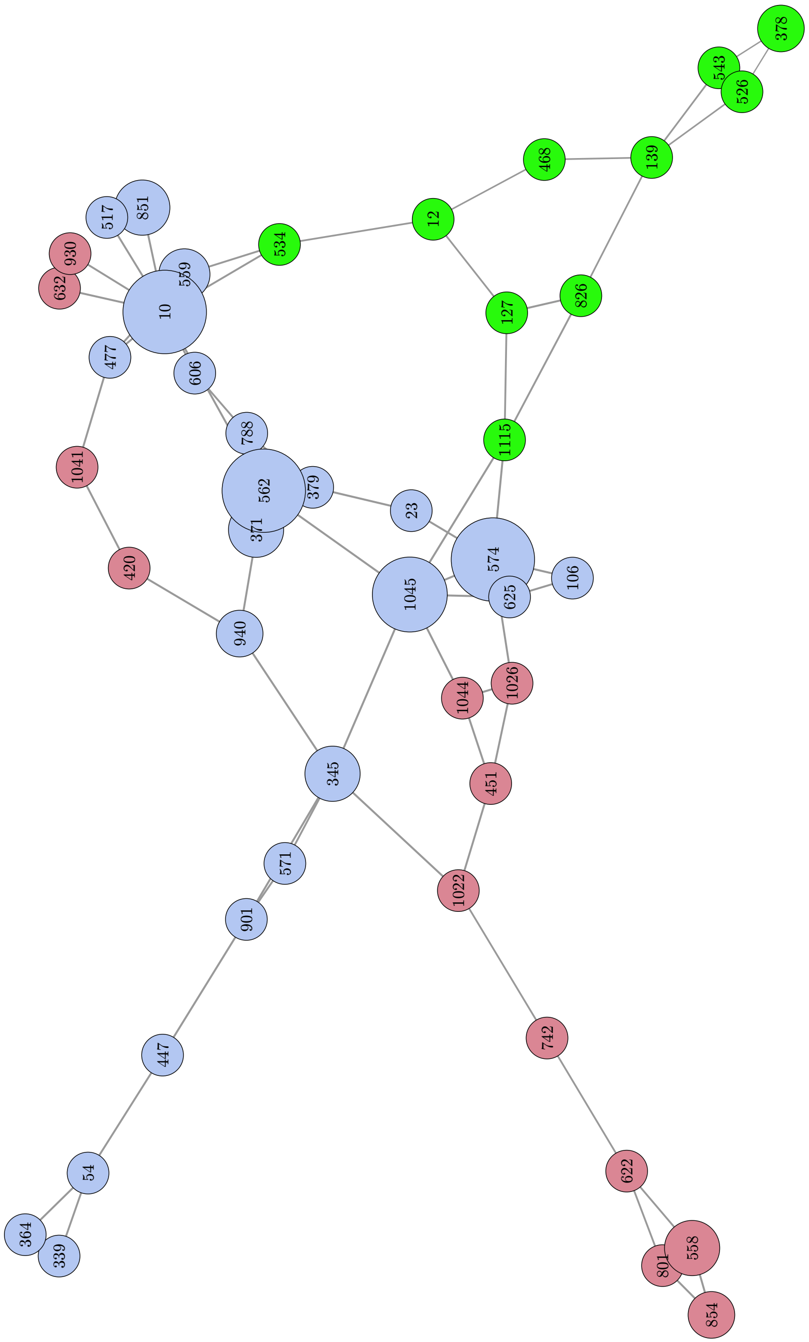


Fig. 3. Networked graph visualizing cervical cancer cases

Figure 3 shows one visualization created with this methodology. Note that for visualization of the graph, we employed the Yifan Hu algorithm [22] as implemented in Gephi[3]. The graph shows all patients diagnosed with some type of cervical cancer. The distance between the nodes is determined by the MRA distance, which entails that nodes which are closer together present similar execution paths in terms of the therapeutic events they contain. The figure in the nodes denotes one representative case ID, with the size of the nodes representing the frequency of a certain sequence of therapeutic activities. Note that it was infeasible to represent all ID's in each of the nodes of the graph. Furthermore, the node colors indicate the application of some specific therapeutic procedure. As such, the green nodes denote the occurrence of chemotherapy. In contrast, the red nodes are cases where hyperthermia treatment is applied. From this visualization, it can be seen that a vast majority of cases do not rely on hyperthermia or chemotherapy. Cervical cancer is majorally treated by either surgery or radiotherapy (teletherapy/brachytherapy).

5 Conclusion

In this study, it was shown how intelligent analysis of clinical pathway data based on process mining techniques can deliver valuable insights into the actual carrying out of a care process. In a practical case consisting of data on the clinical pathways of 1.147 gynecological oncology patients at the AMC hospital, it was demonstrated how both drill up as well as drill down approaches are useful for care flow knowledge discovery. We are convinced that data analysis based on the innovative techniques in the process mining domain is an ideal means for better streamlining and overall improvement of clinical care processes. In the future, we will focus on the development of novel methodologies for analyzing the complex data that is typically found in the logging infrastructures of healthcare information systems. As such, we will further elaborate the idea of networked graph visualizations and improve its integration with existing process discovery techniques. The major benefit of the technique is the enhancement of pure control-flow patterns with other data dimensions. Therefore, additional information, for instance on whether patients were cured or not, would instigate a wide variety of analysis possibilities.

Acknowledgments. The authors would like to thank the Flemish Research Council for financial support under Odysseus grant B.0915.09 and KU Leuven for grant OT/10/010.

References

1. OECD: OECD health data 2010: Statistics and indicators (2010), http://www.oecd.org/health/healthdata
2. Kohn, L.T., Corrigan, J.M., Donaldson, M.S.: To Err Is Human: Building a Safer Health System. The National Academies Press, Washington DC (2000); Committee on Quality of Health Care in America, Institute of Medicine

[3] www.gephi.org

3. van der Aalst, W.M.P.: Process Mining - Discovery, Conformance and Enhancement of Business Processes. Springer (2011)
4. Weske, M.: Business Process Management: Concepts, Languages, Architectures. Springer (2007)
5. Dumas, M., van der Aalst, W.M.P., ter Hofstede, A.H.M.: Process-Aware Information Systems: Bridging People and Software through Process Technology. John Wiley & Sons, Inc. (2005)
6. Lenz, R., Reichert, M.: It support for healthcare processes - premises, challenges, perspectives. Data Knowl. Eng. 61(1), 39–58 (2007)
7. Anyanwu, K., Sheth, A.P., Cardoso, J., Miller, J.A., Kochut, K.: Healthcare enterprise process development and integration. Journal of Research and Practice in Information Technology 35(2), 83–98 (2003)
8. Lenz, R., Elstner, T., Siegele, H., Kuhn, K.A.: A practical approach to process support in health information systems. Journal of the American Medical Informatics Association 9(6), 571–585 (2002)
9. Reijers, H.A., Russell, N., van der Geer, S., Krekels, G.A.M.: Workflow for healthcare: A methodology for realizing flexible medical treatment processes. In: [24], pp. 593–604
10. van der Aalst, W.M.P., Weijters, A.J.M.M., Maruster, L.: Workflow mining: Discovering process models from event logs. IEEE Trans. Knowl. Data Eng. 16(9), 1128–1142 (2004)
11. Alves de Medeiros, A.K., Weijters, A.J.M.M., van der Aalst, W.M.P.: Genetic process mining: an experimental evaluation. Data Mining and Knowledge Discovery 14(2), 245–304 (2007)
12. Weijters, A.J.M.M., van der Aalst, W.M.P., Alves de Medeiros, A.K.: Process mining with the heuristicsminer algorithm. BETA Working Paper Series 166, TU Eindhoven (2006)
13. Goedertier, S., Martens, D., Vanthienen, J., Baesens, B.: Robust process discovery with artificial negative events. Journal of Machine Learning Research 10, 1305–1340 (2009)
14. Song, M., van der Aalst, W.M.P.: Towards comprehensive support for organizational mining. Decision Support Systems 46(1), 300–317 (2008)
15. Rozinat, A., van der Aalst, W.M.P.: Conformance checking of processes based on monitoring real behavior. Information Systems 33(1), 64–95 (2008)
16. Mans, R.S., Schonenberg, H., Song, M., van der Aalst, W.M.P., Bakker, P.J.M.: Application of Process Mining in Healthcare - A Case Study in a Dutch Hospital. In: Fred, A.L.N., Filipe, J., Gamboa, H. (eds.) BIOSTEC 2008. CCIS, vol. 25, pp. 425–438. Springer, Heidelberg (2008)
17. Rebuge, Á., Ferreira, D.R.: Business process analysis in healthcare environments: A methodology based on process mining. Information Systems 37(2), 99–116 (2012)
18. Bose, R.P.J.C., van der Aalst, W.M.P.: Analysis of patient treatment procedures. In: [23], pp. 165–166
19. Caron, F., Vanthienen, J., De Weerdt, J., Baesens, B.: Advanced care-flow mining and analysis. In: [23], pp. 167–168
20. Günther, C.W.: Process Mining in Flexible Environments. PhD thesis, TU Eindhoven (2009)
21. Bose, R.P.J.C., van der Aalst, W.M.P.: Trace clustering based on conserved patterns: Towards achieving better process models. In: [24], pp. 170–181

22. Hu, Y.: Algorithms for Visualizing Large Networks. In: Naumann, U., Schenk, O. (eds.) Combinatorial Scientific Computing (to appear)
23. Daniel, F., Barkaoui, K., Dustdar, S. (eds.): BPM Workshops 2011, Part I. LNBIP, vol. 99. Springer, Heidelberg (2012)
24. Rinderle-Ma, S., Sadiq, S.W., Leymann, F. (eds.): BPM 2009. LNBIP, vol. 43. Springer, Heidelberg (2010)

ALIVE: A Multi-relational Link Prediction Environment for the Healthcare Domain

Reid A. Johnson, Yang Yang, Everaldo Aguiar, Andrew Rider, and Nitesh V. Chawla

Department of Computer Science and Engineering
University of Notre Dame, Notre Dame, IN 46556
{rjohns15,yyang1,eaguiar,arider1,nchawla}@nd.edu

Abstract. An underlying assumption of biomedical informatics is that decisions can be more informed when professionals are assisted by analytical systems. For this purpose, we propose ALIVE, a multi-relational link prediction and visualization environment for the healthcare domain. ALIVE combines novel link prediction methods with a simple user interface and intuitive visualization of data to enhance the decision-making process for healthcare professionals. It also includes a novel link prediction algorithm, MRPF, which outperforms many comparable algorithms on multiple networks in the biomedical domain. ALIVE is one of the first attempts to provide an analytical and visual framework for healthcare analytics, promoting collaboration and sharing of data through ease of use and potential extensibility. We encourage the development of similar tools, which can assist in facilitating successful sharing, collaboration, and a vibrant online community.

Keywords: Link Prediction, healthcare analytics, multi-relational networks.

1 Motivation

An idea that has taken root as the "fundamental theorem" of biomedical informatics is that a person working in partnership with an information resource is better than that same person unassisted. For this theorem to hold, however, the information resource must offer something that the person does not already have. As the people who interact with these resources often possess a high degree of knowledge relating to their domain of expertise, it can be challenging to offer people a resource that they find truly useful and informative.

Link prediction in complex networks has attracted attention from computer scientists and biologists for its ability to provide useful information. However, while most existing link prediction studies are designed for homogeneous networks, where only one type of object exists in the network [1,11–13,15], most networks are in reality heterogeneous and multi-relational [4,9], and attribute values of objects are often difficult to fully obtain. Therefore, the use of topological features between objects in a heterogeneous network is critical to predicting

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 36–46, 2013.

links in a holistic way. In multi-relational homogeneous networks, topological features have different values in different dimensions (relations), while in multi-relational heterogeneous networks the situation becomes more complicated, as the linkage types are different.

By applying link prediction to these types of networks, one can explore unknown or potential links between diseases, genes, and drugs, with findings that can lead to improved biological knowledge and clinical standards, and which can ultimately benefit the quality of healthcare. We propose an approach that uses cutting-edge link prediction algorithms to supply the accuracy needed to provide useful information and a visual environment that can assist healthcare professionals in making observations that can lead to innovation in the healthcare domain.

2 Proposed Environment

Healthcare professionals need data that is correct and informative, both of which can be challenging tasks. To address these challenges, we have developed a virtual platform called "ALIVE" (A Link Information and Visualization Environment). ALIVE is an online link-prediction environment oriented towards healthcare professionals and aimed at benefiting users from a variety of domains and with differing degrees of expertise. The environment takes advantage of the availability of health information records, facilitating data and knowledge management, network analysis, and visual analytics to pursue pioneering inter-disciplinary integration and providing tailored information. ALIVE also encompasses a novel link prediction algorithm that can improve the analysis of healthcare data. In our development of ALIVE, we have focused on developing a tool that will charter a path from data to knowledge to insight, ultimately supporting its users in making more informed healthcare decisions. Such tools can assist many efforts: an epidemiologist might learn of a potential relation between two diseases from a hunch, or a pharmaceutical researcher may discover that a particular drug is unexpectedly effective against a virulent disease for which it was not originally intended.

We foresee the potential for ALIVE to truly fulfill the concept encapsulated by the fundamental theorem of biomedical informatics. The environment has the potential to grow into a full-fledged virtual organization, serving as a data and knowledge warehouse and fostering expert collaboration. As an online platform, ALIVE has the potential to be a powerful information resource combined with a collaborative effort of informed people, which can undoubtedly achieve a vision far greater than the sum of its parts.

3 Background

The problem of predicting unknown links between diseases and genes continues to attract active interest from biological scientists, as it has proven useful in assisting research and make their work more efficient.

Despite a significant and continuous increase in medical research spending, the annual number of new drugs approved and new drug targets identified has remained almost constant for the past 20-25 years, with about twenty new drugs and about five new targets per year. At this rate it will take more than 300 years to double the number of available drugs [7,14]. However, there are several ways to address these burdens. Promising areas of drug design include: wide-range screens of existing drugs, seeking novel applications, combination therapy, (the combined use of several drugs or short DNA oligomers) and the development of multi-target drugs [2,3,6,8,16].

Currently, interactions between diseases, genes, and drugs are studied separately; researchers usually only use interaction networks of diseases and gene (Fig. 1a) to predict disease and gene interactions, or only employ interactions between drug and gene (Fig. 1b) to predict drug-target interaction. We propose to combine these two kinds of networks together to improve understanding and analysis in the medical domain. We believe that the multi-relational network approach will allow us to improve predictions made relating to drug-target interaction.

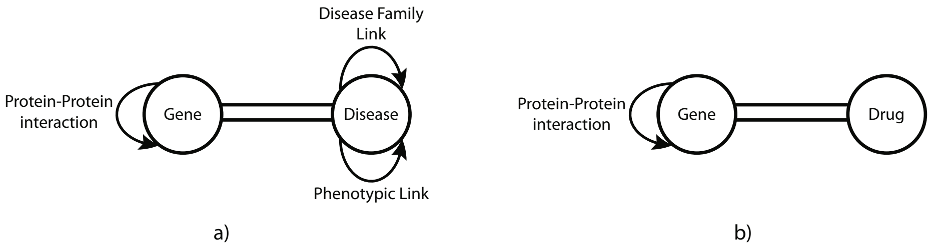


Fig. 1. A visual depiction of the types of interactions that exist in disease, gene, and drug networks. a) The links that may be present in disease-gene interactions. b) The links that may be present in disease-drug interactions.

4 Multi-relational Heterogeneous Networks

A network consists of nodes, representing some concept such as disease, and edges, representing relationships between these nodes. These relationships can encode a variety of information, such as whether diseases tend to occur in the same patient, whether they can be treated by the same drug, and whether they have the same underlying genetic causes. A typical network approach considers only a single type of edge. In contrast, a multi-relational network allows all edge types to exist in the network simultaneously, even overlapping each other. Overlapping edges contain additional information not available in the typical network approach: they may give additional support for two nodes or diseases being linked, or in combination they may specify a particular kind of relationship that was previously not understood.

An additional layer of information and complication is added in heterogenous networks when nodes can represent multiple concepts. For example, different nodes could represent either diseases or genes. An edge between the two types of nodes might represent the confidence with which a gene is related to a disease.

4.1 Link Prediction in Multi-relational Networks

Multi-relational link prediction is a new field of research in data mining. Few attempts have been made to solve this problem due to both difficulty in obtaining real data and the complications inherent in multi-relational networks.

The multi-relational link prediction problem can be described as follows: Given a multi-relational network $G = (V, E_1, E_2, \ldots, E_k)$, predict whether there is a link of type $i = (1, 2, \ldots, k)$ between pair of nodes u and v. To solve the problem of multiple edge types, one needs to know the relationships between each pair of nodes in the network. We define a parameter $\sigma(E_1, E_2)$ to represent the influence between two kinds of edges/relations in the network. $\sigma(E_1, E_2)$ is an asymmetric value, which means $\sigma(E_1, E_2) \neq \sigma(E_2, E_1)$. In other words, relation A and relation B may influence each other with differing degree. For instance, location-based data about people could greatly assist the prediction of their friendship relations, while friendships may not support the prediction of location to the same degree.

Our work builds on two previous approaches to link prediction, the Katz method and PropFlow. The Katz method is a variation on shortest path distance, directly summing over all the paths that exist between a pair of vertices. Specifically,

$$Katz(x, y) = \sum_{l=1}^{\beta_l} paths(x, y) \times l \tag{1}$$

where l is the path length and x and y are a pair of vertices, and β_l is a tuning parameter [10]. In effect, the method penalizes the contribution of longer paths in the similarity computation by exponentially reducing the contribution of a path by a factor of β_l.

PropFlow is an unsupervised path-based link predictor that models the link prediction score as being propagated radially outward from the source [13]. The algorithm uses a breadth-first search approach to propagate the probability that a restricted random walk starting at v_i ends at v_j in l steps or fewer using link weights as transition probabilities, where each score s_{ij} can serve as an estimation of the likelihood of new links. Formally, this likelihood score between nodes u and v is computed as

$$flow(u, v) = score(u) \times \frac{w(u, v)}{d(u)} \times \beta^{h-1} \tag{2}$$

where w is weight, d is degree, and h is the shortest number of hops from u to v.

5 MRPF Algorithm

In our experiments, we find that if we combine the original PropFlow method with the Katz method, we can achieve a higher area under the Receiver Operating Characteristic (AUROC) score than by using either alone. We alter PropFlow by penalizing scores by β so that the similarity between nodes u and v not only depends upon the weights of the shortest path between them, but also upon the number of hops in the path.

However, the PropFlow method as it is currently formulated cannot be directly applied to multi-relational networks; it is designed to work exclusively on single-relational networks, such as single mode homogeneous or bipartite networks. Therefore, we have developed a method to generalize PropFlow features to work on multi-relational networks, which we term multi-relational PropFlow (MRPF). The heuristics of MRPF are as follow:

1. For any two kinds of edges E_1 and E_2 in the network, the influence between E_1 and E_2 can be expressed by the correlation coefficient between their corresponding networks, denoted $\sigma(E_1, E_2)$ as previously described.
2. For node s and its neighbor t, the influence that flows from s to t in edge type E_i is as described in equation 3.

We find that using $p(E_1|E_2)$ in place of the correlation coefficient can achieve a better AUROC score for all of the datasets we have tested. Accordingly, we modify the calculation of flow from 2 to the following:

$$\begin{aligned} flow(s,t,i) = score(s) \times \beta^{h-1} \times \tfrac{w(s,t,i)}{d(s,k)} + \\ \beta^{h-1} \times p(i) \times (1-p(i)) \times \frac{\sum_{k \neq i}^{K} p(i|k) \times \frac{w(s,t,k)}{d(s,k)}}{K-1} \end{aligned} \tag{3}$$

where w is weight, d is degree, K is the number of edge types incident to source node s, and β is a tuning parameter. Generally we set $\beta = 0.05$.

Like PropFlow, our algorithm employs a breadth-first search to propagate information through the whole network, with the addition that we must compute each propagation K (number of edge types) times through an edge rather than only once. Therefore, the complexity of our algorithm is $K \cdot O(|V| \cdot |E|)$, which provides us the means of executing the algorithm in real-time for most practical datasets.

Figure 2 shows a conceptual overview of the MRPF algorithm. In the example, flow is propagated to successive nodes in relation to the degree of correlation. Starting from the source node with a score of 1, all neighboring nodes are given a weighted share of the score. The scores continue to flow outward, summing together for nodes that are reached by several paths.

6 Data

We acquired the data from one of our previous studies [5]. The disease networks were constructed based on the disease-gene associations from OMIM, Swiss-Prot,

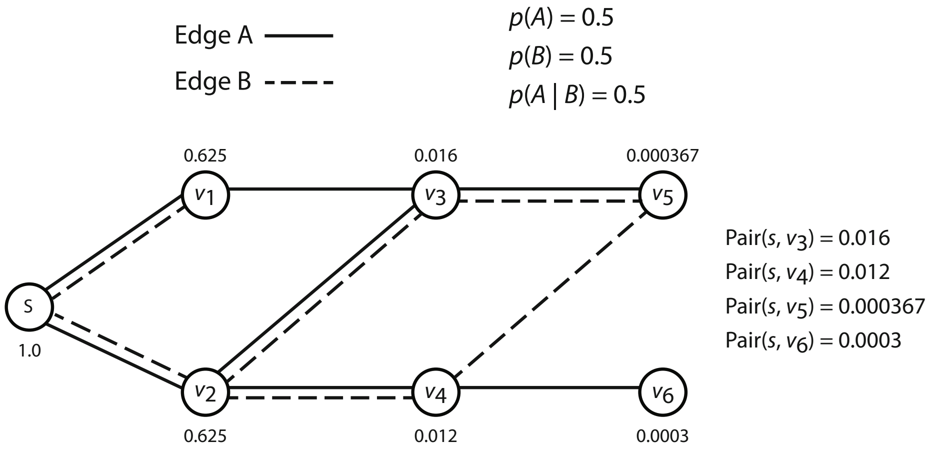


Fig. 2. A conceptual overview of our MRPF alogrithm. Flow propagates outward from the source node S.

and HPRD. The diseases are classified by Disease Ontology (DO) codes and the gene names are based on the HUGO Gene Nomenclature. We constructed a gene-disease network from this data using diseases and nodes, and establishing a link between nodes if the diseases share significantly more gene associations than randomly expected based on generality of the diseases.

Disease co-morbidity was calculated from patient medical diagnoses collected from a regional health system. Each data record is a single visit represented by an anonymized patient ID and a primary diagnosis code, as defined by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). For consistency with the first dataset, the ICD-9-CM codes have been converted to Disease Ontology codes based on mappings provided within the DO coding. The mapping is many to many, so a single ICD-9-CM code often translates to a list of DO codes, and a DO code may apply to multiple ICD-9-CM codes as well. We constructed a phenotypic disease network from patient data, where the nodes are diseases and links indicated disease co-morbidity, where co-morbidity can be broadly defined as co-occurrence in the same patients significantly more than chance.

7 Evaluation

For all experiments, we use a 10-fold cross-validation stratified edge holdout scheme. We use holdout evaluation because longitudinal data was either not available or not relevant for disease, gene, and drug networks. Link prediction is evaluated for each edge type x separately on all eligible node pairs (s, t).

We evaluate each link predication algorithm using the the receiver operating characteristic curve (ROC). The ROC curve presents achievable true positive rates with respect to all possible false positive rates by varying the decision threshold on probability scores. ROC curves can provide information about the operating range of link predictors, with the area under the ROC curve (AUROC) providing a measure of the performance over all predictive thresholds.

We discuss an example to calculate the AUROC. Assume a simple graph with five nodes, seven existing links, and three non-existent links ((1, 2), (1, 4), and (3, 4)). To test the algorithm's accuracy, we select some existing links as probe links. We may, for instance, pick (1, 3) and (4, 5) as probe links, which are presented by dashed lines in the right plot. This means that any algorithm being evaluated may only make use of the information contained in the training graph without (1, 3) and (4, 5).

Let us assume that the scores assigned by an algorithm to non-observed links are $s_{12} = 0.4, s_{13} = 0.5, s_{14} = 0.6, s_{34} = 0.5$, and $s_{45} = 0.6$. Then to calculate AUROC, we need to compare the scores of a probe link and a nonexistent link. There are in total six pairs: $s_{13} > s_{12}, s_{13} < s_{14}, s_{13} = s_{34}, s_{45} > s_{12}, s_{45} = s_{14}$ and $s_{45} > s_{34}$. Hence, the AUROC value equals 0.67.

8 Results

We applied MRPF and other link prediction methods to three biological networks, including a disease-gene network, a disease-disease-phenotype network, and a protein-protein interaction (PPI) network [5]. Table 1 shows results in terms of area under the ROC curve for several link prediction methods. The selected link predictors are among those most frequently used in the task of link prediction; included in these predictors is the latest method proposed by [4].

MRPF outperforms all of the methods on the disease network and PPI network and performs nearly as well as the best method in the phenotypic network. It is worth noting that while MRPF is outperformed on the phenotypic network by several methods, its performance is a significant improvement over that obtained by PropFlow; MRPF also demonstrates incremental improvements on the other networks. These results indictate that incorporating information on relation types into the flow algorithm can significantly improve performance.

Table 1. AUROC statistics for link prediction algorithms used on genetic, phenotypic, and protein-protein interaction networks. The highest values for each type of network are in bold.

Disease	PA	PF	JC	CN	AA	MRLP	MRPF
Genetic	0.903	0.951	0.957	0.951	0.956	0.974	**0.975**
Phenotypic	**0.943**	0.762	0.771	0.909	0.911	0.938	0.901
PPI	0.827	0.888	0.786	0.788	0.789	0.808	**0.890**

Note: PA = Preferential attachment, PF = PropFlow, JC = Jaccard's coefficient, CN = Common neighbors, AA = Adamic/Adar, MRLP = Multi-relational link predictor proposed by [4], and MRPF = Multi-relational PropFlow, proposed herein.

9 Interface Implementation

One of the goals of ALIVE is to facilitate analysis of medical data by healthcare professionals. Therefore we aspire to have an interface that allows non-computer scientists to use cutting-edge network analysis tools. The interface design is a key part of ALIVE that attempts to present analysis options and the results in an intuitive way.

Though the tools necessary to facilitate web-based visualization are not yet as advanced as those for application use, there are several libraries that provide the level of interaction that our project requires. We elected to use D3, a JavaScript library that allows one to bind arbitrary data to a Document Object Model (DOM), and to then apply data-driven transformations to the document. We use D3 to provide an interactive visualization of the relevant networks. Figure 3 provides a high-level illustration of how this interface is organized with relation to the MRLP framework.

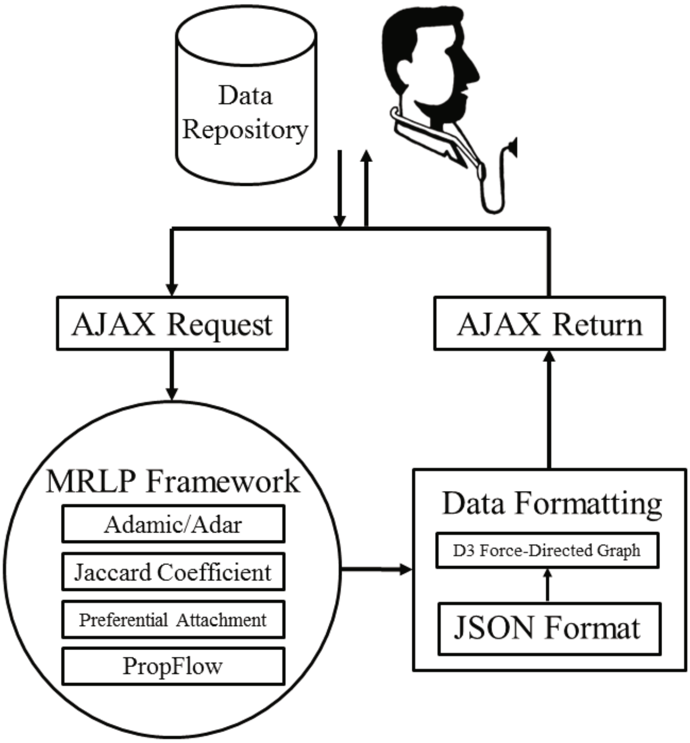


Fig. 3. A high-level component overview

ALIVE seeks to combine cutting-edge network analytics with healthcare data for use by healthcare professionals. Therefore, we designed the user interface with a clean and simple layout, consisting of two basic parts. First there is the main tab, where most of the functionality is available, and to which users are first taken upon navigating to our web-service. There, they are be able to choose between one of the several (previously uploaded) datasets. Our front-end also allows users to select which predictor or predictors they wish to run on the selected data. In order to expand the applicability of our tool, users are also permitted to upload their own datasets and run a variety of algorithms on them.

9.1 Network Visualization

Networks are a natural fit to our goal of providing an informative view of healthcare data. Networks contain emergent properties that are held in their structure and made up of the way in which nodes are related to each other. A healthcare professional may be able to notice emergent patterns in a network that can be the basis for a hypothesis. ALIVE provides missing links in a network and allows healthcare professionals to apply their domain knowledge to a more complete picture of the data, and it does so with a dynamic, interactive visualization of the network generated by user-supplied data.

This work involved several components. As the visualization feature utilizes the functionality of a JavaScript package called D3, the output of the functions that compute the network attributes-which are written in Java and provide output as comma-separated files-needed to be converted to the JavaScript Object Notation (JSON), a file format compatible with JavaScript. This conversion was accomplished via the implementation of an additional Java class and corresponding methods. The D3 package was then leveraged to create a dynamic HTML page with the JSON file as input.

The current implementation is interactive and allows for nodes to be grouped by color, link weights to be designated by line stroke width, and tag information to be displayed for each node. Conceptually, the interface allows users to upload data, which can be evaluated by the MRLP framework with relation to a current data repository. The resulting scores computed by the link prediction algorithms are output visually.

9.2 Extensibility

As developed, our framework also allows for enormous extensibility. With minor additions, users would not only be able to interact with the output via the visualization, but could also have the option of exporting that particular output and saving it for later reference. For networks containing a large number of nodes, we could adjust the graph crop factor (the cutoff for edge inclusion) and let the user zoom in and out for a more versatile visualization. Moreover, as a web-based tool, ALIVE could be expanded into a general repository of healthcare information, allowing health professionals to submit and share data.

10 Conclusions

ALIVE has a great deal of potential as a useful tool in healthcare analytics. Not only have we contributed to the science of link prediction, but we have also provided the basis for an accessible, web-based tool, that has potential to be the nucleosing agent for a healthcare data warehouse. If expanded, ALIVE could ultimately foster a vibrant online community of healthcare professionals, providing the tools necessary to facilitate successful collaboration.

References

1. Al Hasan, M., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: Workshop on Link Discovery: Issues, Approaches and Apps, Citeseer (2005)
2. Borisy, A.A., Elliott, P.J., Hurst, N.W., Lee, M.S., Lehár, J., Price, E.R., Serbedzija, G., Zimmermann, G.R., Foley, M.A., Stockwell, B.R., et al.: Systematic discovery of multicomponent therapeutics. Proceedings of the National Academy of Sciences 100(13), 7977 (2003)
3. Csermely, P., Agoston, V., Pongor, S.: The efficiency of multi-target drugs: the network approach might help drug design. Trends in Pharmacological Sciences 26(4), 178–182 (2005)
4. Davis, D., Lichtenwalter, R., Chawla, N.V.: Multi-relational link prediction in heterogeneous information networks. In: 2011 International Conference on Advances in Social Networks Analysis and Mining, pp. 281–288. IEEE (2011)
5. Davis, D.A., Chawla, N.V.: Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. PloS One 6(7), e22670 (2011)
6. Diacon, A.H., Pym, A., Grobusch, M., Patientia, R., Rustomjee, R., Page-Shipp, L., Pistorius, C., Krause, R., Bogoshi, M., Churchyard, G., et al.: The diarylquinoline tmc207 for multidrug-resistant tuberculosis. New England Journal of Medicine 360(23), 2397–2405 (2009)
7. DiMasi, J.A., Hansen, R.W., Grabowski, H.G.: The price of innovation: new estimates of drug development costs. Journal of Health Economics 22(2), 151–185 (2003)
8. Fitter, S., James, R.: Deconvolution of a complex target using dna aptamers. Journal of Biological Chemistry 280(40), 34193 (2005)
9. Han, J.: Mining Heterogeneous Information Networks by Exploring the Power of Links. In: Gama, J., Costa, V.S., Jorge, A.M., Brazdil, P.B. (eds.) DS 2009. LNCS, vol. 5808, pp. 13–30. Springer, Heidelberg (2009)
10. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18(1), 39–43 (1953)
11. Leroy, V., Cambazoglu, B.B., Bonchi, F.: Cold start link prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 393–402. ACM (2010)
12. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology 58(7), 1019–1031 (2007)
13. Lichtenwalter, R.N., Lussier, J.T., Chawla, N.V.: New perspectives and methods in link prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 243–252. ACM (2010)

14. Ma'ayan, A., Jenkins, S.L., Goldfarb, J., Iyengar, R.: Network analysis of fda approved drugs and their targets. Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine 74(1), 27–32 (2007)
15. Wang, C., Satuluri, V., Parthasarathy, S.: Local probabilistic models for link prediction. In: Seventh IEEE International Conference on Data Mining, ICDM 2007, pp. 322–331. IEEE (2007)
16. Wong, P.K., Yu, F., Shahangian, A., Cheng, G., Sun, R., Ho, C.M.: Closed-loop control of cellular functions using combinatory drugs guided by a stochastic search algorithm. Proceedings of the National Academy of Sciences 105(13), 5105 (2008)

The Relevance of Spatial Relation Terms and Geographical Feature Types

Chunju Zhang, Xueying Zhang, and Chaoli Du

Key Laboratory of Virtual Geography Environment (Nanjing Normal University), MOE, Nanjing, China
zcjtwz@sina.com, zhangsnowy@163.com

Abstract. Spatial relation terms can generally indicate spatial relations described in natural language context. Their semantic representation is closely related to geographical entities and their characteristics e.g. geometry, scale and geographical feature types. This paper proposes a quantitative approach to explore the semantic relevance of spatial relation terms and geographical feature types in text. Firstly, a classification of spatial relation terms is performed. Secondly, the "Overlap" similarity measure is introduced to define the relevance of spatial relation terms and geographical feature types based on a large scale annotation corpus. Thirdly, the relevance is expanded with the semantic distance and hierarchical relationship of the classification system of geographical feature types. Finally, a knowledge base based on protégé is developed to formally represent and visualize geographical feature types, spatial relation classifications, and the relevance of spatial relation terms and geographical feature types. This study indicates that spatial relation terms are strongly relevant to geographical feature types. The semantic representation of topological relation terms is diverse and their relevance with geographical feature types is much stronger than directional relation and distance relation terms, but the annotation quality and the classification granularity of geographical entities in the corpus have a great effect on the performance.

Keywords: spatial relation, geographical feature type, spatial relation term, relevance.

1 Introduction

Natural language describes the nature of people's internal representation of space and is the primary means for representation and exchange of geographical information, such as geographical entities, spatial relations, etc. Spatial relations are the associations or connections between different real world features, and play an important role in spatial data modeling, spatial query, spatial analysis, spatial reasoning, and map comprehension [1]. The semantic research of spatial relations is the premise and basis for the description and expression of spatial relations. Spatial relations have been in a high priority in many research fields, such as linguistics, cognitive science, GIS and spatial reasoning. The linguistics field focus on the words, lexical, syntactic and semantic structure of spatial relation expressions, and the relationship with

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 47–56, 2013.

human's spatial cognition [2][3]. In recent years, spatial relations in natural language have become a hot topic of geographical information science. Mark [4] and Lautenschütz [5] investigated the influence of geometry and scale characteristics, spatial relation types and geographical feature types on human's chosen of spatial relation terms by questionnaire method, and then Mark [6] made a further research on the mapping between spatial relation terms and GIS computational model. Shariff [7] and Xu [8] summarized the knowledge rules of different geographical feature types and spatial relation vocabularies to construct a semantic mapping model of spatial relation terms. Du and Wang [9] explored the formal expression of GIS querying sentences described in a restricted syntactic pattern of spatial relation descriptions in natural language.

Spatial relation terms can indicate spatial relations described in natural language context. Different from the early models of spatial relations which focused on the geometry, it is now widely recognized that the semantic meaning of spatial relation terms is also dependent on functional and pragmatic features in situated context [5]. Their semantic descriptions in natural language are closely related to geographical entities and their characteristics of geometry, scale and geographical feature types. Especially, some spatial relation terms can be used for several different geographical feature types, while some are just for a certain geographical feature type. For example, the spatial relation term of watershed is used to indicate the junction between mountains and waters, and cannot describe geographical entities of other geographical feature types. This paper proposes a quantitative approach to explore the relevance of spatial relation terms and geographical feature types from text corpuses and the classification system of geographical feature types. Properly understanding the semantic meaning of spatial relation terms in text will improve geographical information retrieval, GIS natural language query, extraction of spatial relations from text, and qualitative spatial reasoning.

The reminder of this paper is structured as follows: Section 2 investigates the basic categories of spatial relations in natural language, and the classification of spatial relation terms. Section 3 discusses the calculation of relevance of spatial relation terms and geographical feature types based on Corpus and geographical feature types. The semantic knowledge expression of spatial relation terms based on Ontology is in section 4. The conclusion and future work are given in Section 5.

2 Classification of Spatial Relation Terms

Spatial relations are considered to be one of the most distinctive aspects of spatial information. According to Egenhofer and Franzosa's argument, spatial relations can be grouped into three different categories of topological relations, direction relations and distance relations. Natural-language spatial relations are spatial relations described in natural language among people's daily communication, it is much closer to people's habit than GIS spatial relations [10]. For example, the description of "Yangtze River is across Nanjing city in the northwest, and is 10 kilometers from XinJieKou Shop", there are a topological and direction relation between Yangtze

River and Nanjing city, and a distance relation between Yangtze River and XinJieKou Shop. The description and expression forms of spatial relations in natural language and GIS are very different. Spatial relations in natural language are richer, but with a qualitative, fuzzy, uncertainty and unstructured characters, while spatial relations in GIS are quantitative, structured, and accurate. Topological relations have long been considered as the most important spatial relations in GIS while direction and distance relations are with the highest using frequency in people's daily life. Spatial relations in natural language are expressed through a series of spatial relation terms. In different language, those terms are with different diversity and complexity. Taking the spatial relation term of "crossing" in English for example, in Chinese it can be expressed as "穿越(chuanyue, crossing)", "交叉(jiaocha, crossing)", "横贯(hengguan, crossing)", etc. Meanwhile, some spatial relation terms in Chinese indicate more than one spatial relation type. For example, the spatial relation term "北靠(beikao, north and near)" not only expresses the north direction but also implies a topological relation of extended connection. In addition, there are some spatial relation terms in text descriptions whose semantic meanings cannot be expressed with existing calculation models of GIS spatial relations. Taking the spatial relation term of "支流(zhiliu, tributary)" for example, it may describe an including relation of the main vein and tributaries of a river, however, this semantic relation cannot be expressed in GIS spatial relation models.

Region connection calculus (RCC) model takes geographical entities in the real world as a region and describes spatial relations with the region connectedness [11]. Therefore, it is in accordance with human's cognition habit and more suitable for qualitative representation and reasoning of spatial relations. The ternary point configuration calculus (TPCC) describes directions such as front, back, left and right [12]. Distance relations specify that how far the object is away from the reference object. Based on RCC8, TPCC, and the frequency of spatial relation terms in natural language context, basic categories of spatial relations and classifications of spatial relation terms are described in Table 1.

From table 1, we can see that one spatial relation category may include multi-spatial relation terms, and one spatial relation term may correspond to more than one spatial relation categories. Also, there are some commonly used spatial relation terms which cannot be clustered into these categories, such as between, round, etc. Here it should be noted that this paper only discusses a binary instance of spatial relations between two geographical entities, not consider the composite spatial relations. For some compound spatial relation terms, the classification will be determined by the last direction word, such as "中南部 (zhongnanbu, central south)" with a direction of south. Also, there are some connected words which cannot reflect topological or direction relations but provide the connection between the source and target objects. So they play a role in auxiliary judgments of spatial relations, such as "located", "is", "as", "with", "by", etc.

Table 1. Samples of classifications of spatial relation terms

Spatial Relations	**Spatial Relation Terms**
Topological relation	
…..IN(tangential and non-tangential proper parts)	包含(baohan, including)，属于(shuyu, belong to)
…..EC(extended connection)	相接(xiangjie, touch)，流入(liuru, flow into)
…..DC(discrete connection)	相离(xiangli, discrete connection)，相距(xiangju, apart)
…..PO(Partially overlap)	贯穿(guanchuan, run through)，交叠(jiaodie, overlap)
…..EQ(equality)	相等(xiangdeng, equal)，别名(bieming, alias)
Directional relation	
Relative direction	
…..F(front)	前头(qiantou, front)，前部(qianbu, anterior)
…..BE(behind)	后端(houduan, back-end)，后面(houmian, behind)
…..L(left)	左边(zuobian, left side)，左面(zuomian, left)
…..R(right)	右边(youbian, right)，右端(youduan, right)
…..A(above)	上端(shangduan, above)，上面(shangmian, above)
…..BW(below)	下端(xianduan, below)，下(xia, below)
…..INT(inner)	内(nei, in)，内部(neibu, inner)，里面(limian, inside)
…..EXT (exterior)	外(wai，outer)，外部(waibu，exterior)，外头(waitou, outside)
Absolute direction	
…..E(east)	东方(dongfang，east)，东端(dongduan，east)，东(dong, east)
…..W(west)	西端(xiduan, west)，西部(xibu, west)，西(xi, west)
…..S(south)	南部(nanbu，south)，南(nan，south)，南方(nanfang, south)
…..N(north)	北面(beimian，north)，北方(beifang，north)，北(bei, north)
…..C(centre)	中部(zhongbu, middle)，中心(zhongxin, center)
…..NE(northeast)	东北面(dongbeimian，northeast),东北方(dongbeifang, northeast)
…..SE(southeast)	东南边(dongnanbian, southeast)，东南方(dongnanfang, southeast)
…..NW(northwest)	西北(xibei, northwest)，西北部(xibeibu, northwest)
…..SW(southwest)	西南部 (xinanbu, southwest)，西南边 (xinanbian, southwest)
Distance relation	距离(juli，distance)，相离(xiangli，distance)，相距(xiangju, apart from)

3 Calculation Method of Relevance

3.1 Calculation Based on Corpus

A binary spatial relation could be formalized as < geographical entity A, spatial relation terms, geographical entity B > in natural language context. Obviously, one spatial relation term should associate with a pair of geographical entities. For the concept characteristics of geographical entities could be defined by the type of geographical features, therefore, a single spatial relation can be further abstracted as < feature type of geographical entity A', spatial relation term, feature type of geographical entity B' >. There is an order for target and reference objects in spatial relation descriptions. To simplify the calculation, this order between A and B is not distinguished in this paper. In linguistics, a text corpus is a large and semi-structured set of texts which are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe. This paper takes the large scale annotation corpus (Geocorpus) of spatial relations of "Chinese Geography Encyclopedia" from paper [13] as an experimental data, and summarizes 600 commonly used spatial relation terms. Overlap is a classic calculation method for semantic relations, and it is based on the co-occurrence frequency of two events in a data set [14]. Therefore, the relevance of spatial relation terms and the type of geographical entities based on Geocorpus can be defined as in formula 1.

$$R(T, A', B') = \frac{|T \cap A'B'|}{\min(|T|, |A'B'|)} \tag{1}$$

In formula 1, T represents the occurrences of a spatial relation term in the Geocorpus, A' and B' denote the occurrence of two geographical feature types, R indicates the relevance degree between T and a pair of A 'and B'. Taking the spatial relation term of "流入 (liuru, flow into)" and "北部(beibu, north)" as example, the results of the relevance are just as shown in Table 2.

Table 2. Relevance of spatial relation terms and geographical feature types based on corpus

Spatial Relation Terms T	Geographical Feature Types A'	Geographical Feature Types B'	Relevance
流入 (liuru, flow into)	river	ocean elements	0.8333
	river	lake	1.0000
	river	river	0.1333
北部 (beibu, north)	resident	natural landscape	0.1428
	resident	river	0.0714
	Natural landscape	natural landscape	0.0444
	natural landscape	Lake	0.2222

The annotation and experiment result shows that spatial relation terms are strongly relevant to geographical feature types. The term of "北部(beibu, north)" is related to resident, river, natural landscape, and the other geographical feature types, while the term of "流入(liuru, flow into)" is just in co-occurrence with ocean, lake and river in the corpus. However, some of the relevance has a higher R-value, and some is lower. This is because that there is a natural phenomenon of imbalance of geographical concepts in the real world. Text is a main expression vector of people's cognition from real world, so the geographical concepts in the Geocorpus are not in an imbalance. Meanwhile, some geographical concepts in the Geocorpus have a coarse granularity. Therefore, the R value is higher when the phenomenon is more common.

3.2 Calculation Based on Geographical Feature Types

There is a level and hierarchical relationship in the classification system of geographical feature types. It could be seen as a semantic network diagram. In this diagram, each node represents geographical feature types, edges indicate their relationship, and the weights of edges represent the semantic distance of geographical feature types. With this distance, the semantic relation and relevance between geographical feature types can be analyzed. Based on the relevance from corpus calculation, a quantitative approach to expand the relevance of spatial relation terms and geographical feature types from the classification system of geographical feature types is proposed.

In theory, a pair of geographical feature types with the father-son relationship has a higher semantic relation than brotherhood or nephew relationship. Terms which describe spatial relations between the father geographical feature types cannot describe their son geographical feature types. However, spatial relation terms for the son geographical feature types can also describe their father geographical feature types. Therefore, the semantic relevance between spatial relation terms and geographical feature types should be with a consideration of the semantic distance and the inheritance direction of geographical concept. Assuming that C1 and C2 represent a pair of geographical feature types respectively and their semantic relation in the classification system is R′, and α is the semantic relation value. The specific calculation rules are as follows:

- If C1 and C2 have a father-son relationship, then R ' (C1, C2) = α (0 <α <1, the default value is 0.75); if C1 and C2 have a reverse relationship, then R' (C1, C2) = 1;
- If C1 is inherited indirectly from C2 among n concepts, then R' (C1, C2) =α^n（0<α<1，the default value is 0.75）; If C2 is inherited indirectly from C1 among n concepts, then R' (C1, C2) =1;
- If C1 and C2 have a brother relationship, then R' (C1, C2) =α（0<α<1，the default value is 0.75）;
- Other relations are defined with the above composition.

With the above calculation rules, if the spatial relation term T is in co-occurrence with geographical feature types A ' and B' in the Geocorpus, then taking A 'and B' as a starting point and R as weight value to expand the semantic relevance between T and the pair of A ' and B'. For the hierarchical relationship of geographical feature types in the classification system, the relevance value is expanded with an iterative calculation. This calculation will stop until no new semantic relevance occurs.

With the 600 commonly used spatial relation terms and classification system of geographical feature types (GB/T 13923-2006), the relevance is a large net structure chart. Taking the term “流入 (liuru, flow into)” as an example, the relevance is as in Figure 1.

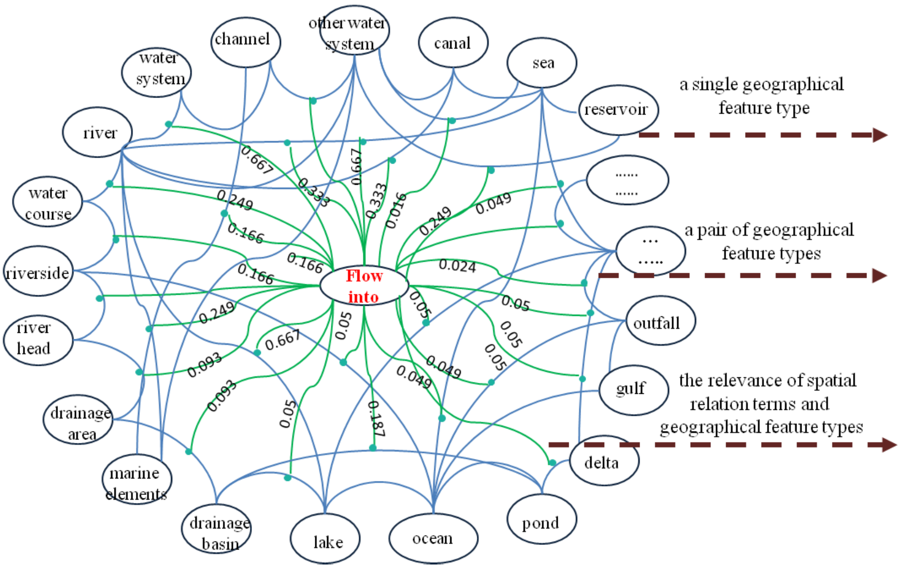


Fig. 1. The relevance of “流入 (liuru, flow into)” and geographical feature types

In the experiment, the R-value stands for the relevance degree between spatial relation terms and geographical feature types. With the classification system of geographical feature types, the relevance value is expanded. However, some of the relevance has a higher R-value, and some is lower than 0.05. In order to keep the balance of the relevance we can set and adjust a threshold to filter the uncommon relevance in a text corpus, such as 0.05 for “流入 (liuru, flow into)”. Then the term “流入 (liuru, flow into)” only describes spatial relations of geographical entities of water system, such as river, ocean and lake. As we all know, there are a lot of spatial relation terms to describe spatial relations of river, ocean, lake, etc, however, people are used to choosing “流入 (liuru, flow into)” to describe them in daily life. Therefore, this result comparatively conforms to people’s language and cognitive habit. In addition, the relevance of topological relation terms and geographical feature types are significantly stronger than directional relation and distance relation terms.

4 Semantic Knowledge Expression Based on Ontology

Ontology formally represents rich knowledge as a set of concepts and the relationships between those concepts within a domain. It can improve the consistency, accuracy, reusability and sharing features of knowledge to understand and use. In this paper, a knowledge base is developed based on protégé to formally represent and visualize geographical feature types, spatial relation classifications, spatial relation terms and their relevance with geographical feature types (see Figure 2). Geographical feature types and spatial relation classifications are expressed with a class in OWL language, and the hierarchy relationship is established by the subClassOf. The ObjectProperty expresses the semantic relations between spatial relation terms and geographical feature types, and the quantitative constraint values are organized in DatatypePoperty. Then the relevance is defined a property with "rdfs: domain" and "rdfs: range" respectively, which can restrict the application fields and scope. Finally, instances of spatial relation terms can be made according to the semantic relevance of ObjectProperty and DatatypePoperty. This knowledge base can improve GIS natural language query, extraction of spatial relations from text, geographical information retrieval, qualitative spatial reasoning, etc.

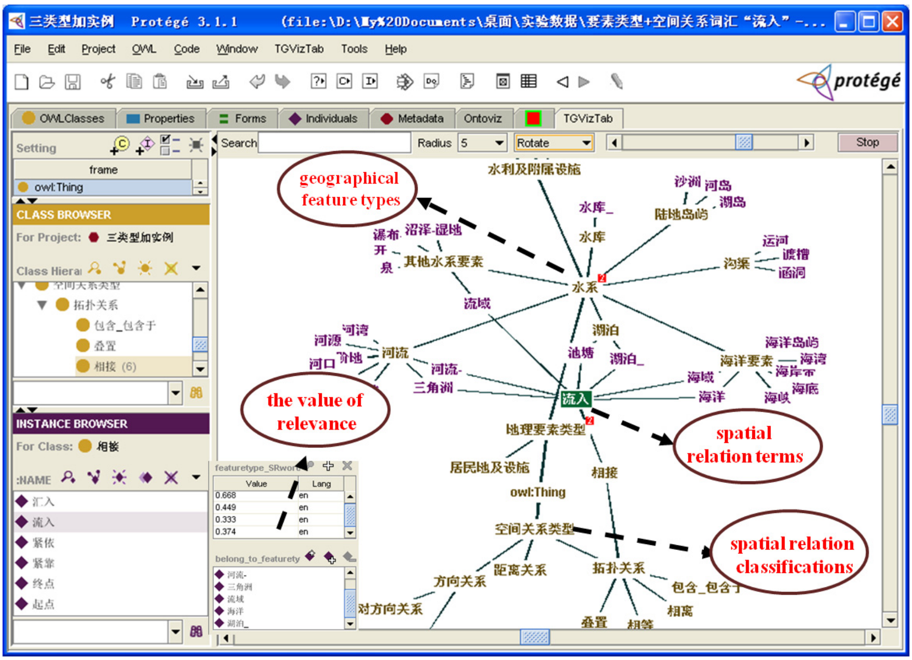


Fig. 2. Knowledge base of the relevance of spatial relation terms and geographical feature types

5 Conclusion

Based on a large scale text corpus and the geographical feature classification scheme this paper proposed a method to explore the relevance of spatial relation terms and geographical feature types. The experiment indicates that our proposed approach can effectively obtain meaningful results. However, the annotation quality of the corpus and the classification granularity of geographical entities have a great effect on the performance, especially for a general dataset. In our future work, we will start the classification on different kinds of texts describing the same kind of data (e.g. documents addressing only water, sea) in order to better extract relations specified for a particular domain. Moreover, to simplify the calculation, the order between geographical entities of A and B is notdistinguished in this paper. In addition to geographical feature types, geometric features and spatial scales of geographical entities also have an effect on spatial relations. Obviously, the semantic relevance of spatial relations and geographical feature types can be further improved with a comprehensive consideration of the description order, scale and geometric features in a further research.

Acknowledgement. This work was supported in part by the National Nature Science Foundation under grant number 40971231, and the Innovative Postgraduate Projects funded by Jiangsu province under grant number CXLX11_0874.

References

1. Egenhofer, M.J., Franzosa, R.: Point-Set Topological Spatial Relations. International Journal of Geographical Information Systems 5, 161–174 (1991)
2. Herskovits, A.: Language and Spatial Cognition: an Interdisciplinary Study of the Prepositions in English. Cambridge University Press, Cambridge (1986)
3. Levinson, S.C.: Space in Language and Cognition: Explorations in Cognitive Diversity. Cambridge University Press, Cambridge (2003)
4. Mark, M.D., Comas, D., Egenhofer, M.J., et al.: Evaluating and Refining Computational Models of Spatial Relations through Cross-linguistic Human-subjects Testing. In: Kuhn, W., Frank, A.U. (eds.) COSIT 1995. LNCS, vol. 988, pp. 553–568. Springer, Heidelberg (1995)
5. Lautenschütz, A.-K., Davies, C., Raubal, M., Schwering, A., Pederson, E.: The Influence of Scale, Context and Spatial Preposition in Linguistic Topology. In: Barkowsky, T., Knauff, M., Ligozat, G., Montello, D.R. (eds.) Spatial Cognition V. LNCS (LNAI), vol. 4387, pp. 439–452. Springer, Heidelberg (2007)
6. Mark, M.D.: Calibrating the Meaning of Spatial Predicates from Natural Language: Line-Region Relations. In: Proceedings of the Sixth International Symposiun on Spatial Data Handing, pp. 538–553 (1994)
7. Shariff, A.R., Egenhofer, M.J.: Natural-language Spatial Relations between Linear and Areal Objects: The Topology and Metric of English-language Terms. International Journal of Geographical Information Science 12(3), 215–246 (1998)
8. Xu, J.: Formalizing Natural-language Spatial Relations between Linear Objects with Topological and Metric Properties. International Journal of Geographical Information Science 21(4), 377–395 (2007)

9. Du, S.H., Wang, Q., Luo, G.: A Model for Describing and Composing Direction Relations between Overlapping and Contained Regions. Information Sciences: An International Journal Archive 178(14), 2928–2949 (2008)
10. Egenhofer, M.J., Rashid, A., Shariff, B.M.: Metric Details for Natural-Language Spatial Relations. ACM Transactions on Information Systems 16(4), 295–321 (1998)
11. Li, S.J., Wang, H.Q.: RCC8 Binary Constraint Network Can be Consistently Extended. Artificial Intelligence 170, 1–18 (2006)
12. Dylla, F., Moratz, R.: Empirical Complexity Issues of Practical Qualitative Spatial Reasoning about Relative Position. In: Proceedings of Workshop on Spatial and Temporal Reasoning at ECAI, pp. 169–175 (2004)
13. Shen, Q.J., Zhang, X.Y., Jiang, W.M.: Annotation of Spatial Relations in Natural Language. In: Proceedings of 2009 International Conference on Environmental Science and Information Application Technology, pp. 418–421 (2009)
14. Smith, E.P.: Statistical Comparison of Weighted Overlap Measures. Transactions of the American Fisheries Society 114(2), 250–257 (1985)

Applying NLP Techniques for Query Reformulation to Information Retrieval with Geographical References

José M. Perea-Ortega[1], Miguel A. García-Cumbreras[2], and L. Alfonso Ureña-López[2]

[1] Languages and Information Systems Department, University of Sevilla
E.T.S. Ingeniería Informática, Avda. Reina Mercedes s/n, 41012, Sevilla, Spain
jmperea@us.es
[2] Computer Science Department, University of Jaén
Escuela Politécnica Superior, Campus Las Lagunillas s/n, 23071, Jaén, Spain
{magc,laurena}@ujaen.es

Abstract. Geographic Information Retrieval (GIR) is an active and growing research area that focuses on the retrieval of textual documents according to a geographical criteria of relevance. However, since a GIR system can be treated as a traditional Information Retrieval (IR) system, it is important to pay attention to finding effective methods for query reformulation. In this way, the search results will improve their quality and recall. In this paper, we propose different Natural Language Processing (NLP) techniques of query reformulation related to the modification and/or expansion of both parts thematic and geospatial that are usually recognized in a geographical query. We have evaluated each of the reformulations proposed using GeoCLEF as an evaluation framework for GIR systems. The results obtained show that all proposed query reformulations retrieved relevant documents that were not retrieved using the original query.

Keywords: Geographic query reformulation, Geographic Information Retrieval, Query expansion, GeoCLEF.

1 Introduction

In the Information Retrieval (IR) field [2], the approach based on the modification of the user query to improve the quality of the IR results is known as query reformulation. The aim of such process is to satisfy the user information need, usually improving the quality and recall of the results obtained using the original user query. This feature is explicitly supported by some search engines suggesting related queries or providing different completions of the initial user query. Moreover, other search engines also support query reformulation in an implicit manner, by expanding the original query with terms related to their keywords, for example.

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 57–69, 2013.

Geographic Information Retrieval (GIR) is an active and growing research area that focuses on the retrieval of textual documents according to a geographical criteria of relevance. For this reason, GIR is considered as an extension of the field of IR. Specifically, GIR is concerned with improving the quality of geographically-specific information retrieval, focusing on access to unstructured documents [10,13]. The IR community has primarily been responsible for research in the GIR field, rather than the Geographic Information Systems (GIS) community. The type of query in a IR engine is based usually on natural language, in contrast to the more formal approach common in GIS, where specific geo-referenced objects are retrieved from a structured database. In a GIR system, a geographic query can be structured as a triplet of <*theme*><*spatial relationship*><*location*>, where <*theme*> is the main subject of the query, <*location*> represents the geographical scope of the query and <*spatial relationship*> determines the relationship between the subject and the geographical scope. For example, the triplet for the geographical query "*airplane crashes close to Russian cities*" would be <*airplane crashes*><*close to*><*Russian cities*>. Thus, a search for "*castles in Spain*" should return not only documents that contain the word "*castle*", but also those documents which have some geographical entity related to Spain.

Since a GIR system can be treated as a traditional search engine (the results for a query are displayed as a ranked list), it is important to pay attention to finding effective methods for query reformulation. These methods can take into account both lexical-syntactic features and geographical aspects. In this way, the search results will improve their quality and recall. The objective of this paper is to evaluate several geographic query reformulations for the GIR task, considering that a GIR system can perform as a IR system. To carry out this evaluation, we have used the most important evaluation framework in this context: GeoCLEF[1] [7,14].

The remainder of this paper is structured as follows: in Section 2, the most important works related to the geographic query reformulation in GIR are expounded; in Section 3, we describe the GIR system used for the experiments; Section 4 presents the main features of the query reformulations proposed; in Section 5, we describe briefly the evaluation framework; in Section 6 and Section 7, the experiments carried out and an analysis of the results are presented; finally, in Section 8, we draw some conclusions and future work is expounded.

2 Related Work

Jansen et al. [9] define the concept of query reformulation as the process of altering a given query in order to improve search or retrieval performance. Sometimes, query reformulation is applied automatically by search engines as with *relevance feedback* technique. It is a method that allows users to judge whether a document is relevant or not, so that automatic rewritings can be generated depending on it. At other times, query reformulation is carried out analysing

[1] http://ir.shef.ac.uk/geoclef/

the top retrieved documents without the user's intervention, taking into account term statistics. However, it has been found that users rarely utilize the relevance feedback options [19] and usually reformulate their needs manually [1].

The focus of this paper is geographic queries. According to Gravano [8], search engines are criticised because of their ignorance to the geographical constraints on users' queries and, therefore, retrieve less relevant results. This could be attributed to the way search engines handle queries in general as they adopt a keywords matching approach without spatially inferring the scope of the geographic terms. However, it shall be noted that a number of services to deal with this issue have recently been proposed in major search engines, but not in the general purpose tools.

Several authors have studied what users are looking for when submitting geographic queries [18,6,11]. One of the main conclusions of these studies is that the structure of geographic queries consists of thematic and geographical parts, with the geo-part occasionally containing spatial or directional terms. From a geographical point of view, Kohler [12] provides a research about georeformulation of queries. She concludes that the addition of more geo terms in the query is commonly used to differentiate between places that share the same name. This is also known as query expansion using geographic entities.

In the literature, we can find various works that have addressed the spatial query expansion. Cardoso et al. [4] present an approach for geographical query expansion based on the use of feature types, readjusting the expansion strategy according to the semantics of the query. Fu et al. [5] propose an ontology-based spatial query expansion method that supports retrieval of documents that are considered to be spatially relevant. They improve search results when a query involves a fuzzy spatial relationship, showing that proposed method works efficiently using realistic ontologies in a distributed spatial search environment. Buscaldi et al. [3] use WordNet[2] during the indexing phase by adding the synonyms and the holonyms of the encountered geographical entities to each documents index terms, proving that such method is effective. Finally, Stokes et al. [20] conclude that significant gains in GIR will only be made if all query concepts (not just geospatial ones) are expanded.

Our work could be positioned within those works that treat geographical part as textual terms, i.e., from a Natural Language Processing (NLP) point of view, exclusively. For this reason, the proposed query reformulations are based on expansions of the thematic and geographical parts detected in a geographical query, using synonyms and geospatial terms related with the keywords and geographical entities found in the query.

3 The SINAI-GIR Architecture

In this Section we describe an example of a GIR system. Specifically, we have used our own GIR system called SINAI-GIR [17]. GIR systems are usually composed of three main stages: preprocessing of the document collection and queries,

[2] http://wordnet.princeton.edu/

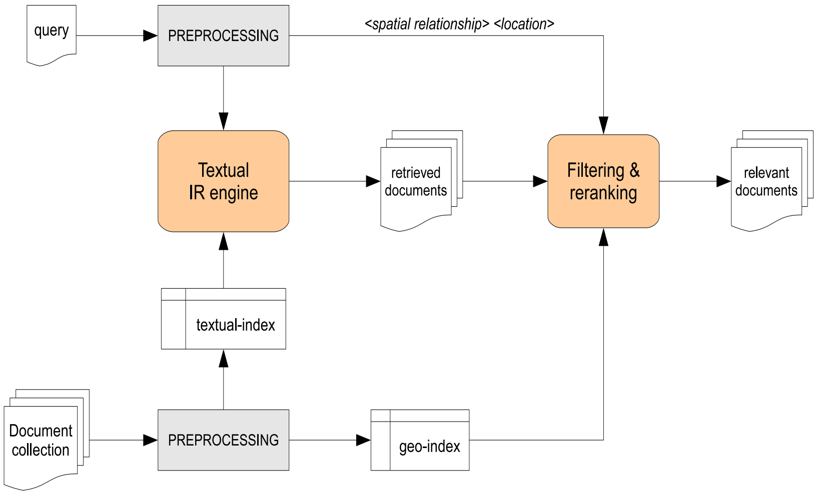


Fig. 1. Overview of the SINAI-GIR system

textual-geographical indexing and searching and, finally, reranking of the retrieved results using a particular relevance formula that combines textual and geographical similarity between the query and the document retrieved. This general architecture is shown in Figure 1.

With respect to the document collection processing, it is based on detecting all the geographical entities in each document and generating a geo-index with them. In this phase, the stop words are removed and the stem of each word is taken into account. We have used our own Named Entity Recognition (NER) tool to detect geographical entities. It is called GeoNER [16] and it is based on external knowledge resources such as GeoNames[3] and Wikipedia.

Regarding query processing, each query is preprocessed and analyzed, identifying the geographical scope and the spatial relationship that may contain. It also involves specifying the triplet explained in Section 1, which will be used later during the filtering and reranking process. To detect such triplet, we have used a Part Of Speech tagger (POS tagger) like TreeTagger[4], taking into account some lexical syntactic rules such as *preposition + proper noun*, for example. Moreover, the stop words are removed and the Snowball stemmer[5] is applied to each word of the query, except for the geographical entities. During the text retrieval

[3] GeoNames is a geographical database covers all countries and contains over eight million placenames that are available for download free of charge. http://www.geonames.org

[4] TreeTagger v.3.2 for Linux. Available in http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[5] Available in http://snowball.tartarus.org

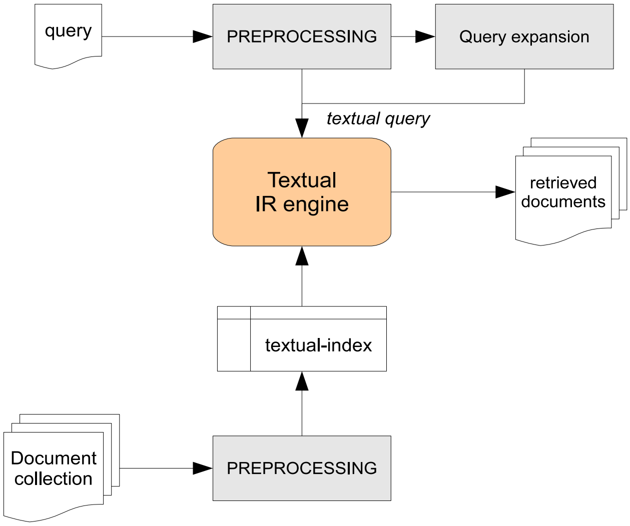


Fig. 2. Architecture of the GIR system employed to carry out the experiments

process, we obtain 1,000 documents for each query. We have used Terrier[6] as a search engine. According to a previous work [15], it was shown that Terrier is one of the most used IR tools in IR systems in general and GIR systems in particular, obtaining promising results. The weighting scheme used has been *inL2*, which is implemented by default in Terrier. This scheme is the Inverse Document Frequency (IDF) model with Laplace after-effect and normalization two. As a final step, each preprocessed query (including their geographical entities) is run against the search engine. The retrieved documents are filtered and reranked, setting in the last positions those documents that do not match with the geographical scope detected in the query. By contrast, those documents that fit the geographical scope detected, are set in the first positions.

Although GIR systems usually apply a geo-reranking process after the IR module (as can be seen in Figure 1), it is important to note that this process is not necessary for this work particularly, because we are interested in analyzing the behaviour of each proposed query reformulation from an IR point of view, i.e. evaluating their precision and recall scores without any reranking process that applies a geographic reasoning. Therefore, the architecture employed to carry out the experiments of this work follows the similar approach that is applied in traditional information retrieval systems but considering query reformulations, as shown in Figure 2.

[6] Version 2.2.1, available in `http://terrier.org`

4 Query Reformulations Proposed

Several query reformulations for the GIR task are analyzed in this work. They try to use both the thematic part and the geographical scope detected in the query. The objective of these query reformulations is to improve the retrieval process trying to find relevant documents that are not retrieved using the original query. Starting from the preprocessed original query, we have generated the following query reformulations:

– QR1: the geographical scope is removed, leaving only the thematic part of the original query.
– QR2: the thematic part is expanded, repeating its terms. In this way, we try to give more importance to the thematic part than the geo-part.
– QR3: the thematic part is expanded using only synonyms of the keywords detected in the thematic part of the query. We have considered as keywords the nouns recognized in such part. WordNet was used as external resource in order to extract the synonyms for each keyword.
– QR4: the geographical part is expanded using only synonyms of the geographical scope detected in the query. These synonyms were extracted from the GeoNames database.
– QR5: the geographical part is expanded using locations or places that match with the geographical scope and the spatial relationship detected in the query. Like the previous query reformulation, GeoNames was used as geographical knowledge base.
– QR6: the thematic and geographical parts are expanded, combining the QR3 and QR5 reformulations.

Table 1. Example of query reformulations generated for the query "*Visits of the American president to Germany*"

Reformulation	Text of the query
original	`visit American presid Germany`
QR1	`visit American presid`
QR2	`visit American presid visit American presid Germany`
QR3	`#and(#or(visit meet stay ) American presid Germany)`
QR4	`#and(visit American presid #or(Germany #3(Federal Republic of Germany) Deutschland FRG ) )`
QR5	`#and(visit American presid #or(Germany Berlin Hamburg Muenchen Koeln #2(Frankfurt am Main) Essen ) )`
QR6	`#and(#or(visit meet stay ) of the American presid ) #or(Germany Berlin Hamburg Muenchen Koeln #2(Frankfurt am Main) Essen )`

Table 1 shows an example of the different query reformulations generated for the query "*Visits of the American president to Germany*". As can be seen, QR2 and QR3 are query reformulations that expand only the thematic part of the queries and, on the other hand, QR4 and QR5 expand only the geographical part of them. Finally, QR6 can be considered a combination of expansions using both parts.

5 GeoCLEF: The Evaluation Framework

In order to evaluate the proposed query reformulations, we have used the GeoCLEF framework [7,14], an evaluation forum for GIR systems held between 2005 and 2008 under the CLEF[7] conferences. GeoCLEF provides a document collection that consists of 169,477 documents, composed of stories and newswires from the British newspaper *Glasgow Herald* (1995) and the American newspaper *Los Angeles Times* (1994), representing a wide variety of geographical regions and places. On the other hand, there are a total of 100 textual queries or topics provided by GeoCLEF organizers (25 per year). They are composed of three main fields: *title* (T), *description* (D) and *narrative* (N). For the experiments carried out in this work, we have only taken into account the *title* field. Some examples of GeoCLEF topics are: "*vegetable exporters of Europe*", "*forest fires in north of Portugal*", "*airplane crashes close to Russian cities*" or "*natural disasters in the Western USA*".

Regarding the evaluation measures used, results are evaluated using the relevance judgements provided by the GeoCLEF organizers and the TREC evaluation method. The evaluation has been accomplished by using the Mean Average Precision (MAP), Recall (R) and Precision at n (P@n). The MAP measure computes the average precision over all queries. The average precision is defined as the mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved. Recall is a measure of the extent to which relevant documents are found or retrieved. Recall is 1.0 when every relevant document is retrieved. Finally, Precision at n is the precision at the number of n relevant documents in the collection for the query. Precision is the fraction of the relevant documents divided by the total number of documents retrieved. Therefore, if P@n is 1.0, it means a perfect relevance ranking and a perfect recall at n documents retrieved.

6 Experiments and Results

The different results obtained using each query reformulation (QR) along with the result obtained using the original query are shown in Table 2. In such table, we show the average score of precision at the 5, 10 and 100 first documents retrieved, recall (R) and MAP for each query reformulation proposed. Although none of the proposed QRs improve the MAP score obtained using the original query, it is interesting to note that QR2 (the thematic part is expanded, repeating its terms) achieves the best P@10 score in three of the four topic sets.

[7] http://www.clef-initiative.eu/

Table 2. Evaluation results obtained for each query reformulation proposed

Topic Set	QR	P@5	P@10	P@100	R	MAP
2005	original	**0.5520**	0.4560	**0.1904**	**0.8364**	**0.3514**
	QR1	0.2640	0.2560	0.1260	0.6748	0.1638
	QR2	0.5200	**0.4920**	0.1840	0.8276	0.3353
	QR3	0.3680	0.3160	0.1400	0.7596	0.2035
	QR4	0.3120	0.2800	0.1212	0.6552	0.2242
	QR5	0.1440	0.1240	0.0772	0.5624	0.0952
	QR6	0.1600	0.1480	0.0780	0.5692	0.0942
2006	original	**0.2400**	0.1920	**0.0716**	**0.7288**	**0.2396**
	QR1	0.0560	0.0640	0.0252	0.4604	0.0615
	QR2	0.2320	**0.2040**	0.0664	0.6796	0.2314
	QR3	0.1440	0.1400	0.0604	0.7356	0.1419
	QR4	0.1920	0.1720	0.0636	0.6984	0.2064
	QR5	0.2240	0.1840	0.0612	0.6524	0.1811
	QR6	0.1840	0.1760	0.0580	0.6772	0.1486
2007	original	**0.3040**	**0.2560**	**0.1188**	**0.7156**	**0.2311**
	QR1	0.1600	0.1320	0.0796	0.4452	0.1255
	QR2	0.2640	0.2120	0.1072	0.6656	0.1871
	QR3	0.2000	0.1800	0.0884	0.6284	0.1774
	QR4	0.2160	0.2000	0.1020	0.6608	0.1687
	QR5	0.2240	0.2000	0.0928	0.6720	0.1874
	QR6	0.2240	0.2040	0.0836	0.6344	0.1763
2008	original	**0.3760**	**0.2680**	0.1104	**0.7368**	**0.2484**
	QR1	0.1760	0.1400	0.0928	0.5996	0.1301
	QR2	0.3440	**0.2680**	**0.1124**	0.7196	0.2381
	QR3	0.2960	0.2320	0.1024	0.6884	0.1972
	QR4	0.2640	0.1960	0.0924	0.6404	0.1619
	QR5	0.2720	0.2040	0.0964	0.6984	0.1906
	QR6	0.2720	0.2280	0.0948	0.7028	0.2028

At this point, we wonder if the QRs proposed were really retrieving relevant documents that the original query was not retrieving. Using the relevance judgements provided by the GeoCLEF organizers, we get the relevant documents retrieved by each QR that were not retrieved by the original query, as shown in Figure 3 and Table 3. The total number of documents retrieved was always 1,000. While in Figure 3 we can compare the behaviour of each query reformulation for the different topic sets regarding the number of relevant documents that were not retrieved by the original query, Table 3 presents these results numerically making a comparison with those obtained using the original query. It is also shown the total number of relevant documents for each topic set.

Analyzing these results in general, we can observe that all proposed query reformulations always retrieved relevant documents that were not retrieved using the original query. This does not mean that the proposed query reformulations achieve higher MAP scores than those obtained by the original query (see Table 2). The main reason for this behaviour is focused on the ranking process. In this

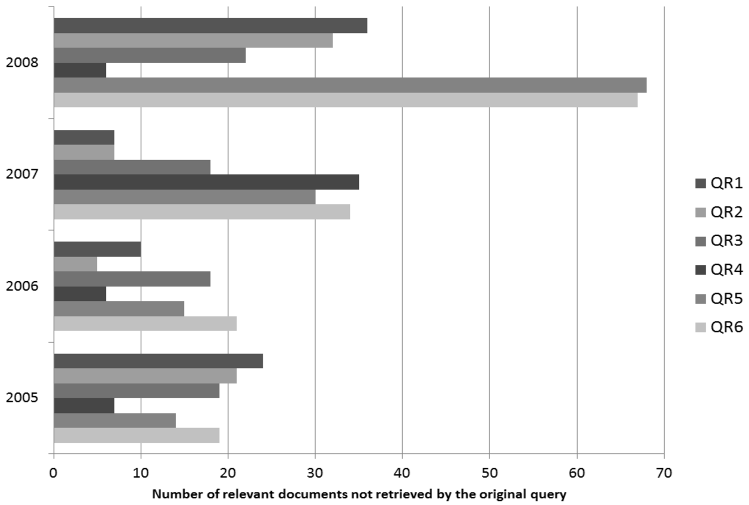


Fig. 3. Comparison of the number of relevant documents retrieved by each query reformulation that were not retrieved by the original query

Table 3. Number of relevant documents retrieved by each query reformulation compared with those obtained using the original query

Query set	Total num relevant docs	Num of relevant doc retrieved						
		original	QR1	QR2	QR3	QR4	QR5	QR6
2005	1028	**908**	735	895	813	706	579	583
2006	378	**284**	160	251	264	280	272	255
2007	650	**543**	391	521	489	483	493	464
2008	747	588	529	**597**	577	480	539	542

experiments we have not applied any spatial ranking process, only has been used the ranking provided by the search engine that does not employ any geographic reasoning. Another reason is that none of the query reformulations outperform the total number of relevant documents retrieved by the original query (except QR2 for the 2008 topic set), as can be seen in Table 3.

7 Analysis and Discussion

Following a general analysis, it is interesting to note the behaviour of the reformulations related to the geographical expansion (QR4 and QR5). Specifically, QR5 achieves a remarkable difference using the 2008 topic set with a total of 68 relevant documents not retrieved by the original query. In fact, this means that of 159 relevant documents not retrieved by the original query using the 2008

Table 4. MAP average results according to the query type

query type	MAP						
	original	QR1	QR2	QR3	QR4	QR5	QR6
Part-to relationship	**0.2582**	0.0819	0.2323	0.1686	0.1781	0.1411	0.1498
Adjacent-to relationship	**0.2130**	0.0861	0.2061	0.1273	0.1179	0.1126	0.0881
Non-geographic	**0.3771**	0.3510	0.3742	0.2974	0.3342	0.3342	0.2607

Table 5. P@10 average results according to the query type

query type	P@10						
	original	QR1	QR2	QR3	QR4	QR5	QR6
Part-to relationship	**0.2986**	0.1114	0.2971	0.2200	0.2114	0.1600	0.1829
Adjacent-to relationship	0.1813	0.1000	**0.1875**	0.1125	0.0938	0.1063	0.1375
Non-geographic	0.3929	0.3857	**0.4000**	0.3214	0.3500	0.3500	0.2786

topic set (747-588), 42.77% of them are retrieved using the QR5 reformulation. Another example occurs with QR4 that obtains the highest value for the 2007 topic set, retrieving 32.71% of the relevant documents not retrieved using the original query. On the other hand, the reformulations related to the thematic expansion (QR2 and QR3) also achieve good results in general, as can be seen for the 2005 and 2006 topic sets. All this makes that the reformulation that combines the QR3 and QR5 reformulations (QR6) also obtain good results, as shown for all topic sets. Finally, QR1 achieves the best score for the 2005 topic set, so the idea of removing the geographical part in the original query can sometimes be a good strategy. This may sound a little strange when we are working on GIR, but we have to take into account that sometimes a query can be considered as a geographic query because it contains a geographical term, but really it is not. For instance, the query "*Japanese rice imports*" might seem a geographic query because it contains the term "*Japanese*", but really it does not impose any geographical constraint.

In order to carry out a more in-depth analysis regarding the distinctive features each QR has, we will use the classification type given by Cardoso and Silva regarding spatial relationships [4]. They distinguish two main types (*part-of* and *adjacent-to*) in order to drive the query expansion strategies according to the proper relationships contained in the geographical ontology used for that purpose. *Part-of* relationships (for example *in*, *of*, *on the*, *at*, etc.) are the most common spatial relationships found on geographical queries [12], denoting that the user is interested on documents inside the boundaries of the given scope of interest. Adjacent-to relationships denote proximity (for example *around*, *next to*, *within X km of*, etc.) and their semantic may have distinct interpretations [5]. According to this classification and taking into account that a geographic query can be considered as a non-geographic query despite contain a geographic entity, we classified manually the 100 queries provided by GeoCLEF, resulting

Table 6. Number of relevant documents retrieved according to the query type

query type	Number of relevant documents retrieved						
	original	**QR1**	**QR2**	**QR3**	**QR4**	**QR5**	**QR6**
Part-to relationship	**1643**	1234	1623	1507	1336	1227	1231
Adjacent-to relationship	**284**	206	248	244	220	263	224
Non-geographic	**396**	375	393	392	393	393	389

14 as non-geographic, 16 as *adjacent-to* type and the remaining (70) as *part-of* type. Therefore, most of the GeoCLEF queries (70%) are considered as *part-of* type.

According to that classification and based on the results shown in Table 4, Table 5 and Table 6, we can observe some findings. As expected, the behaviour of the QR1 is good in general for those queries considered as non-geographic, although that reformulation type does not improve any of the results obtained for the original query on average. In view of the obtained results on average, we can not draw a clear conclusion about when is more desirable to apply one reformulation type according to the type of the spatial relationship detected in the query. However it is interesting to note the good performance of the QR2 for the P@10 measure in general, and for the *adjacent-to* and *non-geographic* query types in particular. This means that repeat the keywords of the thematic part detected in a query could be a good strategy in order to obtain more relevant documents in these systems, particularly when we submit *non-geographic* or *adjacent-to* query types. This behaviour can be explained because by repeating the keywords in the thematic part we are reinforcing the importance of the information need provided by the user in the query when the geographical constraint is not so important.

8 Conclusions and Further Work

In this paper we propose different NLP techniques of query reformulation related to the modification and/or expansion of both parts thematic and geospatial that are usually recognized in a geographical query. We have evaluated each of the reformulations proposed using GeoCLEF as an evaluation framework for GIR systems. This evaluation has been carried out from an IR point of view, that is, without taking into account any geo-reranking procedure after the retrieval process. The results obtained show that all proposed query reformulations retrieved relevant documents that were not retrieved using the original query although these did not improve the results obtained using the original query on average. We carried out a brief analysis according to the two main types of spatial relationships that can be recognized in a geographical query, but it did not provide us a clear conclusion. However, we noted that repeat the keywords of the thematic part detected in a query could be a good strategy in order to obtain more relevant documents in these systems, particularly when we submit *non-geographic* or *adjacent-to* query types.

For future work, we will study in depth when is more suitable to apply these techniques in a GIR system depending on the type of the query and providing a fusion method for collecting those relevant documents retrieved by each query reformulation that were not retrieved using the original query. Then we will work on the spatial reranking process after this fusion in order to sort the final list of documents according to the two criteria of relevance in these systems: thematic and geographical.

Acknowledgments. This work has been partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development through the FIRST project (FP7-287607). This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. It has been also partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER) through the TEXT-COOL 2.0 project (TIN2009-13391-C04-02) from the Spanish Government.

References

1. Anick, P.: Using terminological feedback for web search refinement: a log-based study. In: SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 88–95. ACM, New York (2003)
2. Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston (1999)
3. Buscaldi, D., Rosso, P., Arnal, E.S.: Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 939–946. Springer, Heidelberg (2006)
4. Cardoso, N.: Query expansion through geographical feature types. In: Purves, R., Jones, C. (eds.) GIR, pp. 55–60. ACM (2007)
5. Fu, G., Jones, C.B., Abdelmoty, A.I.: Ontology-Based Spatial Query Expansion in Information Retrieval. In: Meersman, R., Tari, Z. (eds.) OTM 2005. LNCS, vol. 3761, pp. 1466–1482. Springer, Heidelberg (2005)
6. Gan, Q., Attenberg, J., Markowetz, A., Suel, T.: Analysis of geographic queries in a search engine log. In: Proceedings of the First International Workshop on Location and the Web, pp. 49–56. ACM, Beijing (2008)
7. Gey, F.C., Larson, R.R., Sanderson, M., Joho, H., Clough, P., Petras, V.: GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 908–919. Springer, Heidelberg (2006)
8. Gravano, L., Hatzivassiloglou, V., Lichtenstein, R.: Categorizing web queries according to geographical locality. In: Proceedings of the 12th International Conference on Information and Knowledge Management, pp. 325–333 (2003)
9. Jansen, B.J., Booth, D.L., Spink, A.: Patterns of query reformulation during web searching. JASIST 60(7), 1358–1371 (2009)

10. Jones, C.B., Purves, R.S.: Geographical information retrieval. International Journal of Geographical Information Science 22(3), 219–228 (2008)
11. Jones, R., Zhang, W.V., Rey, B., Jhala, P., Stipp, E.: Geographic intention and modification in web search. International Journal of Geographical Information Science 22(3), 229–246 (2008)
12. Kohler, J.: Analysing search engine queries for the use of geographic terms. Master's thesis, University of Sheffield - United Kingdom (2003)
13. Larson, R.: Geographic information retrieval and spatial browsing. In: Smith, Gluck, M. (eds.) Geographic Information Systems and Libraries: Patronsand Mapsand and Spatial Information, pp. 81–124 (1996)
14. Mandl, T., Carvalho, P., Di Nunzio, G.M., Gey, F., Larson, R.R., Santos, D., Womser-Hacker, C.: GeoCLEF 2008: The CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 808–821. Springer, Heidelberg (2009)
15. Perea-Ortega, J.M., García-Cumbreras, M.Á., García-Vega, M., Ureña-López, L.A.: Comparing Several Textual Information Retrieval Systems for the Geographical Information Retrieval Task. In: Kapetanios, E., Sugumaran, V., Spiliopoulou, M. (eds.) NLDB 2008. LNCS, vol. 5039, pp. 142–147. Springer, Heidelberg (2008)
16. Perea-Ortega, J.M., Martínez-Santiago, F., Montejo-Ráez, A., Ureña-López, L.A.: Geo-NER: un reconocedor de entidades geográficas para inglés basado en GeoNames y Wikipedia. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) 43, 33–40 (2009)
17. Perea-Ortega, J.M., Ureña-López, L.A., García-Vega, M., García-Cumbreras, M.A.: Using Query Reformulation and Keywords in the Geographic Information Retrieval Task. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 855–862. Springer, Heidelberg (2009)
18. Sanderson, M., Kohler, J.: Analyzing geographic queries. In: Proceedings Workshop on Geographical Information Retrieval SIGIR (2004)
19. Spink, A., Jansen, B.J., Ozmultu, C.H.: Use of query reformulation and relevance feedback by excite users. Internet Research: Electronic Networking Applications and Policy 10(4), 317–328 (2000)
20. Stokes, N., Li, Y., Moffat, A., Rong, J.: An empirical study of the effects of nlp components on geographic ir performance. International Journal of Geographical Information Science 22(3), 247–264 (2008)

Adaptive Evidence Accumulation Clustering Using the Confidence of the Objects' Assignments

João M.M. Duarte[1,2], Ana L.N. Fred[2], and F. Jorge F. Duarte[1]

[1] GECAD - Knowledge Engineering and Decision Support Group,
Institute of Engineering, Polytechnic of Porto (ISEP/IPP),
Porto, Portugal
{jod,fjd}@isep.ipp.pt

[2] Instituto de Telecomunicações,
Instituto Superior Técnico,
Lisboa, Portugal
afred@lx.it.pt

Abstract. Ensemble methods are known to increase the performance of learning algorithms, both on supervised and unsupervised learning. Boosting algorithms are quite successful in supervised ensemble methods. These algorithms build incrementally an ensemble of classifiers by focusing on objects previously misclassified while training the current classifier. In this paper we propose an extension to the Evidence Accumulation Clustering method inspired by the Boosting algorithms. While on supervised learning the identification of misclassified objects is a trivial task because the labels for each object are known, on unsupervised learning these are unknown, making it difficult to identify the objects on which the clustering algorithm should focus. The proposed approach uses the information contained in the co-association matrix to identify degrees of confidence of the assignments of each object to its cluster. The degree of confidence is then used to select which objects should be emphasized in the learning process of the clustering algorithm. New consensus partition validity measures, based on the notion of degree of confidence, are also proposed. In order to evaluate the performance of our approaches, experiments on several artificial and real data sets were performed and shown the adaptive clustering ensemble method and the consensus partition validity measure help to improve the quality of data clustering.

1 Introduction

The general goal of data clustering is to find structure in data. Specifically, clustering consists of grouping a set of objects into clusters, such that similar objects are assigned to the same cluster and distinct objects are assigned to different clusters, according to some notion of similarity between data. A large number of clustering algorithms have been proposed over time. However, none of the clustering algorithms can alone discover all sorts of shapes and structures of clusters.

In the last decade, several clustering ensemble methods were proposed stimulated by the effectiveness of classifier ensemble methods. These methods combine multiple data

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 70–87, 2013.

partitions to improve data clustering robustness and quality [9], reuse single-run clustering algorithms solutions [18], cluster data distributively, speed-up clustering process and cluster data with heterogeneous features.

Boosting algorithms have been very successful in supervised learning. These algorithms combine *weak* classifiers iteratively, such that, objects misclassified in previous iterations have greater importance in the current learning iteration [10]. By focusing on regions containing objects more difficult to classify it is expected the combination of these *weak* learners lead to a *strong* classifier. On unsupervised learning the class of each training object is unknown making the identification of the misclassified objects very difficult. Topchy et al. [20] proposed a clustering ensemble construction method following the boosting principles by checking the consistency of the objects' assignments on the previous iterations. At each iteration a new data set is subsampled. An object consistency index is computed as the fraction of the maximal number of times an object was grouped is a certain cluster over the current number of data partitions. The probability of an object being selected is the weighted sum of the object consistency index plus the probability of the object in the previous iteration. Zhai et al. [23] proposed a fuzzy clustering ensemble method based on dual boosting. Fuzzy partitions produced from subsamples of the original data are iteratively mapped into a co-association matrix and the probability distribution of an object being selected is computed so that objects easy and hard to cluster have great importance in the clustering process. Saffari and Bischof [15] introduced an unified and generic boosting framework which builds the clustering ensemble using any model-based clustering algorithm.

We propose an adaptive clustering ensemble construction method for Evidence Accumulation Clustering [7]. The clustering ensemble is build iteratively using an object weight clustering algorithm which focuses the learning process on the objects with more weight. After building each partition the co-association matrix is updated and the object weights are computed given the degrees of confidence of assigning each object to its cluster. The degrees of confidence are estimated using the similarity space induced from the co-association matrix and are viewed as indicators of how good/bad the objects are clustered. Comparing with the boosting methods mentioned before, our approach does not rely on subsampling techniques or a model-based clustering algorithm, but rather on an object-weighted clustering algorithm. Also, in order to build the clustering ensemble, the number of clusters for each data partition is not required to be the *natural* number of clusters, which makes it possible to use the Evidence Accumulation Clustering's split-and-merge strategy [8]. In this paper, we study the effect of focusing on the objects hard to cluster, on the objects easy to cluster, and on a mix of the previous. We also use the notion of degree of confidence to assess the quality of the produced consensus data partitions.

The rest of this paper is organized as follows. In Section 2, the clustering combination problem is introduced. An adaptive clustering ensemble approach and an object-weighted clustering algorithm are proposed in Section 3. The consensus clustering validity is addressed in Section 4. In Section 5, the experimental setup and results are discussed. Section 6 concludes this paper.

2 Clustering Combination

2.1 Problem Definition

Let $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a data set with n objects and $\mathbf{X} = \left[\mathbf{x}_1^T, \cdots, \mathbf{x}_n^T\right] \in \mathbb{R}^{n \times d}$ its matricial representation s.t. $\mathbf{x}_i = \left[x_{i1}, \cdots, x_{id}\right]^T \in \mathbb{R}^d$ is a vector containing the values for d attributes that describes $\mathbf{x}_i$. A clustering ensemble, $\mathcal{P}$, is defined as a set of N data partitions of $\mathcal{X}$:

$$\mathcal{P} = \{P^1, \cdots, P^N\}, \;\; P^c = \{C_1^c, \cdots, C_{K^c}^c\}, \tag{1}$$

where C_k^c is the k^{th} cluster in data partition P^c, which contains K^c clusters. Different partitions capture different views of the structure of the data. Clustering ensemble methods use a consensus function f which maps a clustering ensemble $\mathcal{P}$ into a consensus partition $P^* = f(\mathcal{P})$.

2.2 Related Work

Clustering ensemble approaches may be categorized according to the way data partitions belonging to clustering ensemble are produced – the *clustering generation step* – and to the combination scheme of them – the *consensus step*. Figure 1 shows an overview on multiple data clustering combination. The main approaches for the clustering generation and consensus steps are presented next.

Clustering Generation Step - The process of building the clustering ensemble defines how the data partitions which are going to be combined are generated. In this step, it is important to create diversity among the clustering ensemble in order to produce consensus partitions of superior quality [11]. In the clustering ensemble step the following options may be used separately or in combination.

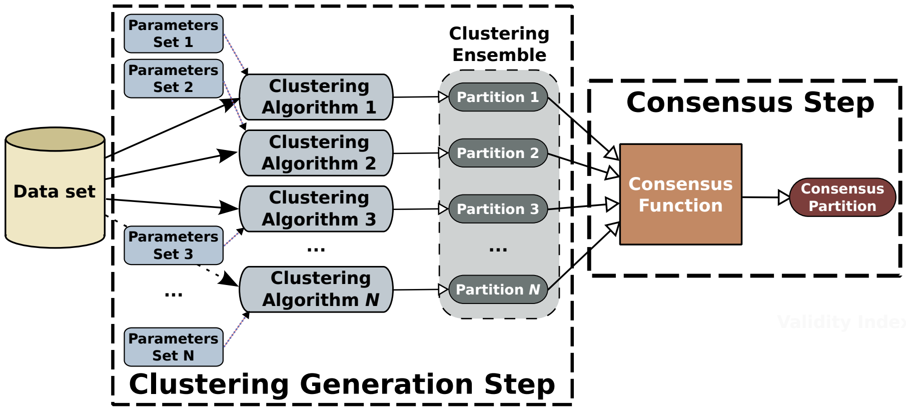


Fig. 1. Clustering ensemble steps

- Clustering algorithms - The data partitions may be produced using only one clustering algorithm or using several clustering algorithms [3]. In this case, diversity is created by optimizing distinct objective functions.
- Parameters and initializations - Even if only one clustering algorithm is used, diversity may be obtained by using different parameters and/or initializations. For instance, the k-means algorithm may be applied for each data partition using different number of clusters and initializations of centroids [7].
- Subsets of data objects - Each data partition may be produced using different sets of data objects. In real-life scenarios the data may be spread in different physical locations. Instead of concentrate all the data in one location, one may produce data clusterings at every locations, centralize only these clusterings, and then obtain the consensus clustering. Even if all the data is centralized, it may be advantageous to use different subsets of data. The use of resampling techniques may increase stability, robustness and quality in consensus clustering [14,20], and the use of subsampling techniques [3] can also speed-up the clustering generation step.
- Subsets of data features - The data partitions may be generated using all the features of the data set or by selecting distinct subsets of data features for each data partition [1]. Each subset of features can be considered as a partial view of the data, thereby, the clustering combination may be thought as an aggregation of distinct views of the data. Using subsets of data features also enables clustering data distributively, reduces memory usage, and enables the clustering of heterogeneous data.
- Projecting to subspaces - To prevent the use of noisy or irrelevant features, and to avoid the problem of the "curse of dimensionality" in high dimensional data, some clustering ensemble construction methods project the original data space into a lower dimensional data space before building the clustering ensemble. Fern and Brodley proposed the use of the random projection technique to build the clustering ensemble [5]. In this method, the original data features are linearly combined using random weights. Topchy et al. proposed to build ensembles of *weak* clusterings by projecting the feature space into only one dimension or by splitting the data by random hyperplanes [19].

Consensus Step - This step defines how the multiple data partitions are combined into consensus partitions. The most popular approaches are presented below.

- Majority voting - The majority voting approaches are the most commonly used in supervised classifier ensembles. Each classifier "votes" for the class of the object $\mathbf{x}_i$, and then $\mathbf{x}_i$ is given the class with more votes. The problem is more complex in unsupervised learning because the labels of the objects do not represent the objects classes, i.e., an object having the same label on different clusterings do not mean that the object was assign to the same class twice. Therefore, the cluster correspondence problem need to be solved to perform majority voting [22,4].
- Co-associations between pairs of objects - These methods store in a $n \times n$ matrix the frequency in which each pair of objects was grouped in the same cluster in all the partitions belonging to the cluster ensemble. This matrix may be viewed as a similarity matrix between objects, so a clustering algorithm can be used to produce the consensus partition [7].

- Searching for the median partition - Some approaches define the consensus clustering as finding the partition P^* that maximizes the average similarity between P^* and all the partitions belonging to the cluster ensemble. Topchy et al. proposed to maximize the Average Normalized Mutual Information by applying the k-means algorithm to a particular representation of the clustering ensemble [19]. Jouve and Nicoloyannis [12] proposed to represent the clustering ensemble as a categorical data set and search for the median partition using a categorical data clustering algorithm.
- Mapping the clustering ensemble into graph or hypergraph problems - Some approaches capture the relations between objects and transform them into graph problems. The CSPA [18] and IBGF [6] methods are some examples. Other approaches map the relations between the clusters in the clustering ensemble into graph problems (e.g. the CBGF [6] and WSPA [2] methods), or hypergraph problems (e.g. the HGPA and MCLA methods [18]). Another approaches, such as HBGF [6] and WBPA [2], represent both objects and clusters as vertices of a graph and map the object-to-cluster relations as edges.

For more informations on this topic, the interested reader may check the survey by Vega-Pons and Ruiz-Shulcloper [21].

2.3 Evidence Accumulation Clustering

The Evidence Accumulation Clustering method (EAC) [7] considers each data partition $P^c \in \mathcal{P}$ as an independent evidence of data organization. The underlying assumption of EAC is that two objects belonging to the same "natural" cluster will be frequently grouped together. A vote is given to a pair of objects every time they co-occur in the same cluster. Pairwise votes are stored in a $n \times n$ co-association matrix, $\mathbf{C}$, normalized by the total number of combined data partitions:

$$\mathbf{C}_{ij} = \frac{\sum_{l=1}^{N} vote_{ij}^{c}}{N}, \tag{2}$$

where $vote_{ij}^c = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ co-occur in a cluster of data partition P^c; otherwise $vote_{ij}^c = 0$. The consensus partition is obtained by applying some clustering algorithm over the co-association matrix, $\mathbf{C}$.

3 Proposed Combination Method

3.1 Adaptive Clustering Ensembles

In this section, an extension to the Evidence Accumulation Clustering method is proposed. It is inspired by the supervised learning Boosting algorithms, where a different weight is assigned to each object depending on its hardness to be well classified. We will refer to the proposed algorithm as Adaptive Evidence Accumulation Clustering (AdaEAC).

Our method relies on estimating the degree of confidence of assigning an object $\mathbf{x}_i$ to its cluster C_k, using the information contained in the co-association matrix $\mathbf{C}$. The idea is simple: if the average similarity of $\mathbf{x}_i$ with respect to the other objects belonging to the same cluster ($\{\mathbf{x}_j : \mathbf{x}_j \in C_k\}$) is higher than the average similarity to the objects belonging to the closest cluster (excluding C_k), then $\mathbf{x}_i$ probably was well assigned. Otherwise, the confidence of the assignment is low and $\mathbf{x}_i$ probably should have been assigned to the other cluster. The degree of confidence of assigning an object $\mathbf{x}_i$ to its cluster C_{P_i} is computed as

$$\mathrm{conf}(\mathbf{x}_i) = \left(\frac{1}{|C_{P_i}| - 1} \sum_{j:\mathbf{x}_j \in \{C_{P_i}\} \setminus \mathbf{x}_i} C_{ij} \right) - \left(\max_{1 \le k \le K, k \ne P_i} \frac{1}{|C_k|} \sum_{j:\mathbf{x}_j \in C_k} C_{ij} \right) \quad (3)$$

where $|\cdot|$ is the cardinality of a set.

While in the EAC approach all N data partitions belonging to the clustering ensemble are assumed to already exist, in AdaEAC the clusterings are produced in F folds. In each fold, an object-weighted clustering algorithm uses as input the object weights obtained in the previous fold to bias the production of L data partitions which are used to update the co-association matrix $\mathbf{C}$. After the co-association matrix $\mathbf{C}$ is updated, a consensus clustering algorithm is applied to $\mathbf{C}$ to obtain the current consensus partition P^* and the degree of confidence for each object is computed as described in Eq. 3. Finally, the weights of the objects for the next iteration can be update considering the degrees of confidence for the assignments of all objects. The proposed approach is summarized in Algorithm 1. In this paper, we assume the clustering ensemble $\mathcal{P}$ is built using the split-and-merge strategy. In this setting, the data partitions belonging to the clustering ensemble have an higher number of clusters than the *real* number of clusters, so, the clusters are smaller but more dense. The clustering ensemble is constructed by generating each data partition P^c with K^c clusters, where K^c is a random integer (different for each P^c) belonging to the set $\{K_{\text{min}}, K_{\text{min}}+1, \cdots, K_{\text{max}}-1, K_{\text{max}}\}$. K_{min} and K_{max} are parameters defined by the user.

We studied three distinct ways to compute the weights of the objects:

1. Emphasizing objects with **low degree of confidence:** The idea is to focus on the objects which have weak similarities with remaining objects of their group, according to the co-association matrix. As an example, if there are two touching clusters, the weak objects should be the ones that are positioned near the region the clusters touch. Concentrating the object-weighted clustering algorithm on this region should help the definition of the clusters borders. Equation 4 expresses this idea and this version of AdaEAC will be referred as *AdaEAC L*

$$w_i = \frac{\left[\max_{m=1,\cdots,n} \mathrm{conf}(\mathbf{x}_m)\right] - \mathrm{conf}(\mathbf{x}_i)}{\sum_{j=1}^{n} \left[\max_{m=1,\cdots,n} \mathrm{conf}(\mathbf{x}_m)\right] - \mathrm{conf}(\mathbf{x}_j)}. \quad (4)$$

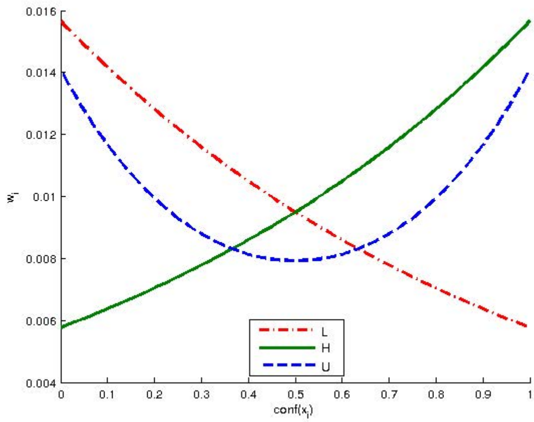


Fig. 2. Object weights w_i against degree of confidence $\mathrm{conf}(\mathbf{x}_i)$

2. Emphasizing objects with **high degree of confidence:** Focusing the objects with high degree of confidence, the ones more similar to the other objects of the same cluster, should reduce the problem of noisy points. This is expected because these noisy points will have low impact on the decisions taken by the object-weighted clustering algorithm. This idea if reflected in equation 5 and originates the version *AdaEAC H*

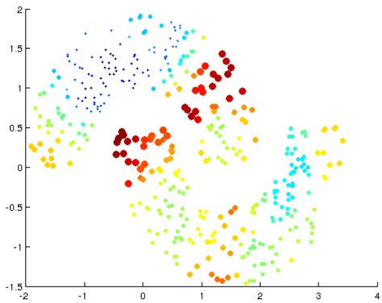
$$w_i = \frac{\mathrm{conf}(\mathbf{x}_i)}{\sum_{j=1}^{n} \mathrm{conf}(\mathbf{x}_j)}. \tag{5}$$

3. Emphasizing objects with **low and high degree of confidence:** With the combination of both previous ideas we expect the object-weighted clustering algorithm to focus both the clustering borders and well defined regions of the clusters. In order to compute the objects weights, the degree of confidences are first stretched to the $[0; 1]$ interval (Eq. 6) and then the *AdaEAC U* is derived from equation 7:

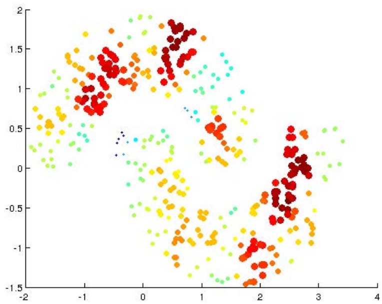
$$q_i = \frac{\mathrm{conf}(\mathbf{x}_i) - \min_{m=1,\cdots,n} \mathrm{conf}(\mathbf{x}_m)}{\max_{m=1,\cdots,n} \mathrm{conf}(\mathbf{x}_m) - \min_{m=1,\cdots,n} \mathrm{conf}(\mathbf{x}_m)}, \tag{6}$$

$$w_i = \frac{[1 - q_i(1 - q_i)]^2}{\sum_{j=1}^{n} [1 - q_j(1 - q_j)]^2}. \tag{7}$$

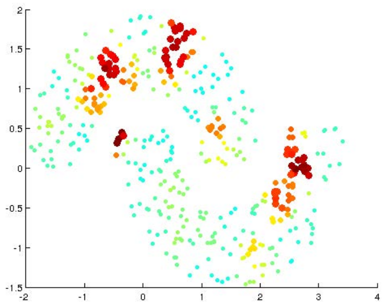
Figure 1 presents the behavior of object weights w_i against the confidence of the assignments $\mathrm{conf}(\mathbf{x}_i)$ and Figure 3 illustrates the corresponding weights in an artificial data set. Big (and red) points correspond to objects with high weights and small (and blue) points to objects with low weights.



(a) AdaEAC L



(b) AdaEAC H



(c) AdaEAC U

Fig. 3. Example: weights on each object for an artificial data set, according to Equations 4 to 7

Algorithm 1. Adaptive cluster ensembles using an object-weighted clustering algorithm

Input: Data set matrix $\mathbf{X}$; Number of folds F; Number of clusterings for each fold L; Minimum and maximum number of clusters K_{min}, K_{max}; and the *natural* number of clusters K^*.

$\mathbf{C} \leftarrow \mathbf{0}_{n,n}$ // Initialize co-association matrix
$\mathbf{W}^1 \leftarrow \left[w_1^1, \cdots, w_n^1\right]^T$,[2] $w_i^1 = \frac{1}{n}$ // Initialize object selection probabilities
$c \leftarrow 0$
for $f \leftarrow 1$ to F **do**
  for $l \leftarrow 1$ to L **do**
    $c \leftarrow c + 1$
    //Produce data partition using the distribution $\mathbf{W}^c$
    $K \leftarrow \text{RandomInteger}(K_{\text{min}}, K_{\text{max}})$;
    $P^c \leftarrow \text{ObjectWeightedClusterer}(\mathbf{X}, \mathbf{W}^c, K)$
    //Update co-association matrix
    for all $C_k^c \in P^c$ **do**
      for all $(\mathbf{x}_i, \mathbf{x}_j) \in C_k^c$ **do**
        $\mathbf{C}_{ij} \leftarrow \mathbf{C}_{ij} + 1$
      end for
    end for
    //Produce consensus partition
    $P^* \leftarrow \text{ConsensusClusterer}(\mathbf{C}, K^*)$
    //Update object confidence
    for all $(\mathbf{x}_i) \in \mathcal{X}$ **do**
      Compute $\text{conf}(\mathbf{x}_i)$ as in equation 3.
    end for
    //Update object weights
    for all $(\mathbf{x}_i) \in X$ **do**
      Update w_i^{c+1} using equations 4-7.
    end for
    $\mathbf{W}^{c+1} \leftarrow \left[w_1^{c+1}, \cdots, w_n^{c+1}\right]^T$
  end for
end for
return P^*

3.2 Object-Weighted k-Means

In this subsection, an object-weighted clustering algorithm is proposed. To incorporate distinct weights for different objects, a modification to the well-know k-means clustering algorithm [13] is presented. Given the desired number of clusters K, k-means algorithm proceeds by alternating between the assignment and update steps. During the assignment step, each object $\mathbf{x}_i$ is grouped in the cluster C_k with the closest center $\overline{\mathbf{x}}_k$. In the update step, the center of each cluster $\overline{\mathbf{x}}_l . \forall l \in \{1, \cdots, K\}$ is computed as the mean of the objects belonging C_l.

The proposed modification consists on modifying the update set in order to shift the center of the groups towards the objects with more weight. Thus, the centers of the clusters will be moved to more important regions, according to the object weights $\mathbf{W} = [w_1, \cdots, w_n]^T$. Algorithm 2 describes the proposed object-weighted clustering algorithm.

Algorithm 2. Object-weighted k-means

Input: Data set matrix $\mathbf{X}$; Object weights $\mathbf{W} = [w_1, \cdots, w_n]^T$; and the number of clusters K.

Randomly initialize clusters centroids $\overline{\mathbf{x}}_k, \forall k \in \{1, \cdots, K\}$.
repeat
  //Assign each object to the cluster of the closest centroid
  for $i \leftarrow 1$ to n **do**
    $C_{k^*} = C_{k^*} \bigcup \{\mathbf{x}_i\}$, s.t., $k^* = \arg\min_k ||\mathbf{x}_i - \overline{\mathbf{x}}_k||^2$
  end for
  //Compute new cluster centroids
  for $k \leftarrow 1$ to K **do**
    $\overline{\mathbf{x}}_k \leftarrow \frac{1}{\sum_{j:\mathbf{x}_j \in C_k} w_j} \sum_{\mathbf{x}_i \in C_k} w_i \mathbf{x}_i$
  end for
until Objects do not change cluster assignments
return $P = \{C_1, \cdots, C_K\}$

4 Consensus Partition Validation

After the consensus partition is generated, it may be useful to assess its quality, especially if one wants to choose the best partition among several consensus partitions. Given the definition of the degree of confidence of assigning an object to a cluster (subsection 3.1), a straightforward way to validate a consensus partition is the Average Confidence of assignment of the objects to its clusters:

$$AC(P^*) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{conf}(\mathbf{x}_\mathrm{i}) \tag{8}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\sum\limits_{j:\mathbf{x}_j \in \{C_{P_i}\}\setminus \mathbf{x}_i} C_{ij}}{|C_{P_i}| - 1} - \max_{1 \le k \le K, k \ne P_i} \frac{\sum\limits_{j:\mathbf{x}_j \in C_k} C_{ij}}{|C_k|} \right). \tag{9}$$

The value of $AC(P^*)$ is defined in the interval $[-1, 1]$. In the best-case scenario, where the co-associations of all objects with the objects belonging to the same cluster is 1 and the co-associations with objects belonging to the other clusters is 0, $AC(P^*)$ takes value 1. In the worst case scenario, where the co-associations between objects on the same cluster are 0 and belonging to different clusters are 1, $AC(P^*)$ takes value -1.

Figure 4 shows an example of a co-association matrix obtained using the split-and-merge strategy for the Iris data set. The objects are sorted by cluster: objects 1 to 50 belong to the first cluster, objects 51 to 100 to the second cluster, and the remaining objects to the third cluster. The similarities between objects (frequencies of co-associations) are represented in a gray scale. The co-association entries of highly similar objects ($C_{ij} = 1$) are shown in black, while the entries of very dissimilar objects ($C_{ij} = 0$) are shown in white. In a perfect-case scenario, the co-associations between objects in the

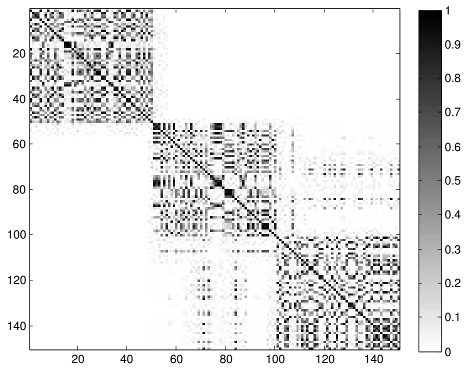


Fig. 4. Example of a co-association matrix for the Iris data set

same cluster should be 1 and the co-associations between objects belonging to different clusters should be 0, resulting in a figure with three 50×50 black squares. It can be seen that it did not occur for the given example (figure 4). One reason is due to some objects belonging to the second and third *natural* clusters being erroneously clustered together. Another reason is related to the use of the split-and-merge strategy: the number of clusters for each partition in the clustering ensemble is higher than the *natural* number of clusters, therefore some objects belonging to the same *natural* cluster have never been placed in the same cluster while building clustering ensemble. In these situations, where the intra-cluster co-associations are sparse, it may be helpful to assess the confidence of the assignments only on the neighborhood of each object. To do so, only the m^{th} nearest neighbors of each cluster should be considered while computing the average confidence. Let $V(\mathbf{x}_i, C_k, m)$ be the set of the m^{th} most similar objects of the cluster C_k to $\mathbf{x}_i$, according to the co-association matrix $\mathbf{C}$. The Average Neighborhood Confidence (ANC) of assigning the objects to its clusters is computed as

$$ANC(P^*, m) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\sum\limits_{j:\mathbf{x}_j \in V(\mathbf{x}_i, C_{P_i}, m)} C_{ij}}{|V(\mathbf{x}_i, C_{P_i}, m)|} - \max_{1 \leq k \leq K, k \neq P_i} \frac{\sum\limits_{j:\mathbf{x}_j \in V(\mathbf{x}_i, C_k, m)} C_{ij}}{|V(\mathbf{x}_i, C_{P_k}, m)|} \right). \quad (10)$$

Figure 5 shows the sum of the co-associations related to each individual object $\mathbf{x}_i$ for the matrix shown in figure 4, i.e. $\sum_j C_{ij}$. Each of these values is related to the average number of objects that were placed in the same cluster of each object. We observed the values are not constant for all the objects. In our example, the values vary from 6 to 17. This may be easily explained: the central objects of each cluster should be co-clustered with more objects than the peripheral objects. Another factor that may contribute for such variations are data sets with unbalanced size of clusters. Considering this fact, we propose an alternative version of ANC, where the neighborhood of each object $V_i(\mathbf{x}_i, C_{P_i}, m_i)$ has a dynamic size m_i. This alternative will be referred as

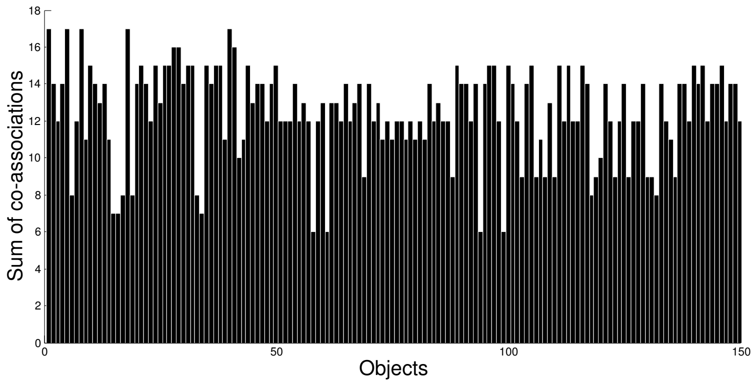


Fig. 5. Sum of the co-associations for each object of the Iris data set

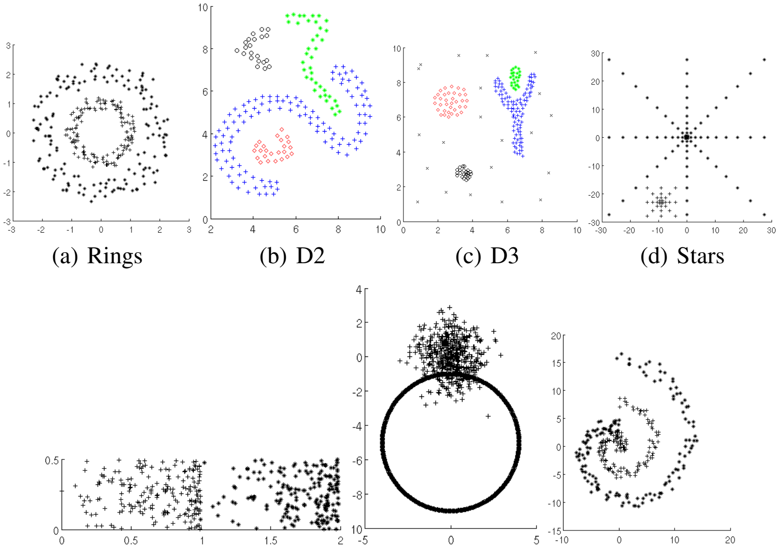
Average Dynamic Neighborhood Confidence (ADNC). The size of the neighborhood for each object m_i should be proportional to the sum of its co-associations. In this paper, we compute m_i as:

$$m_i = \left\lceil \alpha \sum_{j \in \{1,\cdots,n\} \setminus i} C_{ij} \right\rceil, \tag{11}$$

where $\alpha > 0$ is a parameter specified by the user.

5 Experimental Setup and Results

7 synthetic and 7 real data sets were used to assess the performance of the proposed approach on a wide variety of situations, such as data sets with different cardinality and dimensionality, arbitrary shaped clusters, well separated and touching clusters and distinct cluster densities. Table 1 presents the summary (number of objects n, number of dimensions d and the number of objects for each cluster) of all data sets used in our experiments and Figure 6 illustrates the 2-dimensional synthetic data sets used in our experiments. A brief description for each real data set is given next. The Iris data set consists of 50 objects from each of three species of Iris flowers (setosa, virginica and versicolor) characterized by four features. One of the clusters is well separated from the other two overlapping clusters. The Breast Cancer data set is composed of 683 objects characterized by nine features and divided into two clusters: benign and malignant. The Yeast Cell data set consists of 384 objects described by 17 attributes, split into five clusters concerning five phases of the cell cycle. There are two versions of this dataset, the first one is called Log Yeast and uses the logarithm of the expression level and the other is called Std Yeast and is a “standardized” version of the same data set, with mean 0 and variance 1. The Optdigits is a subset of Handwritten Digits data set containing only the first 100 objects of each digit, from a total of 3823 objects characterized by 64 attributes. The House Votes data set is composed of two clusters of votes for each of



(a) Rings (b) D2 (c) D3 (d) Stars

(e) Bars (f) Image2 (g) Spiral

Fig. 6. Synthetic data sets

the U.S. House of Representatives Congressmen on the 16 key votes identified by the The Wine data set consists of the results of a chemical analysis of wines grown in the same region in Italy divided into three clusters with 59, 71 and 48 objects described by 13 features.

To build the clustering ensembles we used the object-weighted k-means, proposed in Section 2, for both EAC and AdaEAC approaches. For EAC the object weights were set to $\frac{1}{n}$, making it equivalent to the standard k-means and the number of data partitions of the clustering ensemble was defined as $N = 200$. For AdaEAC, the number of folds was defined as $F = 10$ and the number of clusterings for each fold as $L = 20$ such that the number of partitions in both approaches were the same. The minimum and maximum number of clusters were defined as $K_{\text{min}} = \left\lfloor \min\left[\frac{2n}{20}, \max\left(\frac{2n}{50}, \sqrt{n}\right)\right]\right\rfloor$ and $K_{\text{max}} = \left\lceil \min\left[K_{\text{min}} + \max\left(\frac{2n}{50}, 2\sqrt{n}\right), \frac{n}{5}\right]\right\rceil$, respectively. Figure 7 shows the minimum and maximum number of clusters for $n = 1$ to 1000.

To extract the consensus partition from the co-association matrix the Average-link [17] and the Single-link [16] algorithms were applied and the number of clusters K^* was defined as the *real* number of clusters K^0 for each data set. Each clustering combination method was applied 30 times for each data set.

The Consistency index (Ci) [9] was used to assess the quality of the consensus partitions P^*. Ci measures the fraction of shared objects in matching clusters of the consensus partition (P^*) and the *natural* data partition (P^0) obtained from known labeling of data. The Consistency index is defined as $Ci(P^*, P^0) = \frac{1}{n}\sum_{k=1}^{\min(K^*,K^0)} |C_k^* \cap C_k^0|$, where it is assumed that consensus cluster C_k^* matches with the real cluster C_k^0.

Table 1. Data sets overview

Data sets	n	d	K	Cluster Distribution
Bars	400	2	2	$200 + 200$
D2	200	2	4	$116 + 39 + 21 + 24$
D3	200	2	5	$98 + 23 + 23 + 35 + 21$
Stars	114	2	2	$33 + 81$
Rings	300	3	2	2×150
Image2	1000	2	2	2×500
Spiral	300	2	2	2×150
Wine	178	13	3	$59 + 71 + 48$
Yeast	384	17	5	$67 + 135 + 75 + 52 + 55$
Optdigits	1000	64	10	10×100
Iris	150	4	3	3×50
House Votes	232	16	2	$124 + 108$
Breast Cancer	683	9	2	$444 + 239$

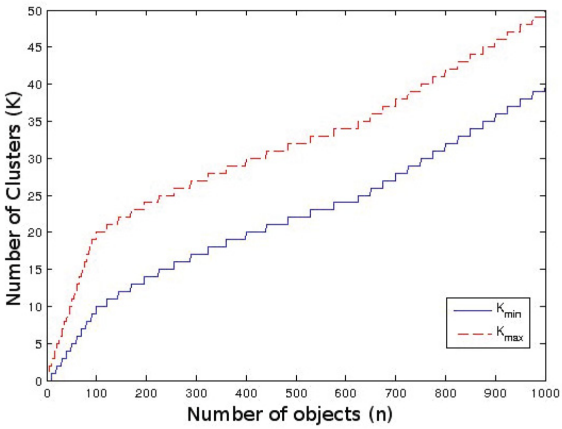


Fig. 7. Minimum and maximum number of clusters

Table 2 presents the average $Ci(P^*, P^0) \times 100$ values for the clustering ensemble methods using both Average-link (AL) and Single-link (SL) algorithms for extracting the consensus data partition (columns 2-9). Lines 3 to 9 show the results for the artificial data sets while lines 10 to 16 show the results for the real data sets. The clustering combination methods that achieved the best results in each data set are highlighted in bold. For the Bars data set, the best result was produced by AdaEAC L using both the average-link and single-link algorithms with 99.5%. For the D2 data set, the AdaEAC L and U using the single-link algorithm achieved the best results with 100%. The EAC outperform the adaptive approaches only in D3 data set using the single-link algorithm and in Std Yeast and Log Yeast data sets using the average-link algorithm. The best results for Stars and Wine data sets were obtained by AdaEAC U in combination with the

Table 2. Average $Ci(P^*, P^0) \times 100$ values for the clustering combination methods

Comb. Method	EAC		AdaEAC L		AdaEAC H		AdaEAC U	
Extraction Alg.	**AL**	**SL**	**AL**	**SL**	**AL**	**SL**	**AL**	**SL**
Bars	99.43	92.04	**99.5**	**99.5**	99.08	82.58	99.19	88.73
D2	73.55	98.3	57.28	**100**	73.28	99.15	74.03	**100**
D3	71.62	**90.55**	64.72	77.43	73.67	80.92	72.85	77.67
Stars	92.75	67.6	93.13	68.19	92.51	67.54	**93.57**	67.54
Rings	99.67	91.89	**99.8**	79.6	99.67	97.12	99.67	95.02
Image2	91.06	52.17	**91.4**	50.13	90.15	51.06	89.44	50.84
Spiral	80.2	**85**	77.26	83.82	80.79	**85**	81.7	**85**
Wine	72.21	72.19	72.23	71.05	72.17	62.4	**72.36**	64.72
Std Yeast	**68.35**	47.46	67.14	36.55	68.3	36.54	68.04	36.1
Optdigits	85.27	61.13	**88.03**	30.76	83.74	35.24	83.81	32.61
Log Yeast	**42.01**	36.52	38.99	36.4	41.23	36.73	41.4	36.76
Iris	89.93	74.67	**95.33**	85.07	90.16	74.76	90.18	75.73
House Votes	89.25	69.08	88.32	53.05	**89.71**	53.05	89.25	53.02
Breast Cancer	96.97	63.01	**97.06**	62.82	96.96	64.52	97.05	63.84

average-link algorithm with 93.57% and 72.36%, respectively. The AdaEAC L using the average-link algorithm, remarkably, achieved the best results for Rings, Image2, Optdigits, Iris, Breast Cancer and Bars data sets. We highlight the 95.33% result obtained in the Iris data set, which was superior to all the other methods by a margin higher than 5%. For the Spiral data sets the best result was 85% and was obtained by EAC, AdaEAC H and U using single-link algorithm. In summary, the AdaEAC L approach achieved the best result in 4 out of 7 synthetic data sets while the AdaEAC U approach obtained the best result in 3 out of 7 synthetic data sets, the EAC method in 2 out of 7, and the AdaEAC H in 1 out of 7 data sets. For the real data sets, the AdaEAC L approach obtained again the best result in 4 out of 7 data sets, the EAC in 2 and both AdaEAC H and U only in 1 data set. These results suggest that the AdaEAC L approach is a good option for combining multiple data partitions.

Table 3 shows the average $Ci(P^*, P^0) \times 100$ values of all the consensus partitions produced for a given data set (column 2) and the average $Ci(P^*, P^0) \times 100$ of the partitions resulting of picking the best partition among the EAC and AdaEAC approaches (for a single run) according to the consensus clustering validity measures $AC(P^*)$ (column 3), $ANC(P^*, m)$ (columns 4 to 7), and $ADNC(P^*, m_i)$ (columns 8 to 10). Line 2 indicates the parameters used by the consensus measures. Four different values for the number of neighbors, $m = \{5, 10, 20, 40\}$, for the ANC measure were tested. For the ADNC consensus validity measure, we defined the value α as 0.5, 1 and 2. When the average quality of the partitions selected by the consensus validity measures outperform the average consensus results, the corresponding results are highlighted in bold. The highest average result for each data set is also underlined.

Table 3. Average $Ci(P^*, P^0) \times 100$ values of all consensus partitions and average $Ci(P^*, P^0) \times 100$ values for the consensus partitions selected by the consensus validity measures

Val. Measure	**Consensus**	**AC**	**ANC**				**ADNC**		
Parameters	**Average**		$\mathbf{m=5}$	$\mathbf{m=10}$	$\mathbf{m=20}$	$\mathbf{m=40}$	$\alpha=\mathbf{0.5}$	$\alpha=\mathbf{1}$	$\alpha=\mathbf{2}$
Bars	95.01	**99.5**	**99.20**	**99.10**	**99.15**	**98.1**	**99.38**	**99.18**	82.88
D2	84.45	73.78	73.90	73.82	73.07	72.53	74.18	74.92	74.12
D3	76.18	67.32	**90.13**	**87.47**	74.02	**77.98**	**89.78**	**85.52**	**80.98**
Stars	80.35	**93.42**	67.54	67.54	67.54	67.54	67.54	77.89	71.93
Rings	95.31	**98.24**	**99.67**	**99.67**	**99.67**	**99.8**	**99.67**	**99.67**	**99.66**
Image2	70.78	**91.4**	51.48	54.03	64.71	64.16	58.17	65.32	64.78
Spiral	82.35	**82.56**	77.88	77.62	79.78	82.07	78.48	78.44	79.69
Wine	69.92	**72.3**	**72.19**	63.58	61.27	61.31	**72.19**	67.04	62.45
Std Yeast	53.56	**67.9**	**66.97**	**68.47**	**68.49**	**68.19**	**68.2**	**68.25**	**68.22**
Optdigits	62.57	**87.57**	**83.24**	**83.49**	**83.36**	**83.64**	**83.35**	**83.54**	**83.12**
Log Yeast	38.76	**42.55**	**40.10**	**41.55**	**41.55**	**41.9**	**41.55**	**41.55**	**42.11**
Iris	84.48	**93.82**	76.36	77.38	74.49	74.67	78.07	80.11	74.67
House Votes	73.09	**88.32**	**74.02**	**87.47**	**88.32**	**88.32**	**86.08**	**88.41**	**88.51**
Breast Cancer	80.28	**95.89**	77.60	78.64	78.47	**97.06**	**87.35**	**90.48**	**91.53**

By picking the best consensus partition using the Average Confidence measure we obtained better results than the consensus average in 5 of the 7 artificial data sets and in all of the real data sets. We also notice that the average quality of the partitions selected by this index is close to the quality of the best clustering combination scheme in each data set. Overall, the AC measure selected better partitions that the other measures in 8 out of the 14 data sets. With respect to the Average Neighborhood Confidence measure, we verify that the measure is sensible to the neighborhood size m. It is not clear which value for m should be used for each data set. The ANC measure performed better using $m = 5$ neighbors in Bars, D2, D3, Wine and Iris data sets, while in the Rings, Spiral, Optdigits and Breast Cancer data sets ANC obtained better results using $m = 40$. In both cases, ANC achieved better results than the consensus average in 8 out of the 14 data sets. Regarding the Average Dynamic Neighborhood Confidence, it achieved better results than the consensus average in 9 out of the 14 data sets using $\alpha = 0.5$ and in 8 data sets using $\alpha = 1$. By comparing ADNC using $\alpha = 0.5$ and $\alpha = 1$ with the ANC results, we observe that the $Ci(P^*, P^0) \times 100$ values are usually similar to the best results achieved by ANC. These results point out that the Average Confidence index is a very good choice for performing consensus clustering selection. Although, in same situations, computing the confidence of the assignments in the neighborhood of each object is better. In this case, the dynamic definition of the neighborhood's size should be preferred over the static one.

6 Conclusions and Future Work

A clustering combination method, based in the Evidence Accumulation Clustering and the supervised learning boosting methods was proposed. Our approach is based on estimating the degree of confidence of the assignment of objects to clusters and then influence the process of constructing the clustering ensemble using an object-weighted clustering algorithm. We tested three distinct ways of computing the object weights: focusing on the objects hard to cluster, on the objects easy to cluster, and on a mix of the previous. Three consensus clustering validity measures based on the confidence of the objects' assignments were also proposed to selected the best consensus partition. Experimental results suggest that using the Adaptive Evidence Accumulation Clustering method, focusing the construction of the clustering ensemble on the objects that are harder to cluster, is a good choice to perform data clustering. It was also shown that the proposed consensus clustering measures can successfully be used to perform consensus clustering selection, in particular the Average Consistency index.

In the future, we pretend to study the influence of the size of the neighborhood for computing the Average Neighborhood Consistency and Average Dynamic Neighborhood Consistency measures.

Acknowledgements. This work is supported by FEDER Funds through the "Programa Operacional Factores de Competitividade - COMPETE" program and by National Funds through FCT under the projects FCOMP-01-0124-FEDER-PEst-OE/EEI/UI0760/2011 and PTDC/EIA - CCO/103230/2008, and grant SFRH/BD/43785/2008.

References

1. Al-Razgan, M., Domeniconi, C., Barbar, D.: Random Subspace Ensembles for Clustering Categorical Data. In: Okun, O., Valentini, G. (eds.) Supervised and Unsupervised Ensemble Methods and their Applications. SCI, vol. 126, pp. 31–48. Springer, Heidelberg (2008)
2. Domeniconi, C., Al-Razgan, M.: Weighted cluster ensembles: Methods and analysis. ACM Trans. Knowl. Discov. Data 2, 17:1–17:40 (2009)
3. Duarte, F.J., Fred, A.L.N., Rodrigues, M.F.C., Duarte, J.: Weighted evidence accumulation clustering using subsampling. In: Sixth International Workshop on Pattern Recognition in Information Systems (2006)
4. Dudoit, S., Fridlyand, J.: Bagging to Improve the Accuracy of a Clustering Procedure. Bioinformatics 19(9), 1090–1099 (2003)
5. Fern, X.Z., Brodley, C.E.: Random projection for high dimensional data clustering: A cluster ensemble approach, pp. 186–193 (2003)
6. Fern, X.Z., Brodley, C.E.: Solving cluster ensemble problems by bipartite graph partitioning. In: Proceedings of the Twenty-First International Conference on Machine Learning, ICML 2004, pp. 36–43. ACM, New York (2004)
7. Fred, A., Jain, A.: Combining multiple clustering using evidence accumulation. IEEE Trans. Pattern Analysis and Machine Intelligence 27(6), 835–850 (2005)
8. Fred, A., Jain, A.K.: Evidence Accumulation Clustering Based on the K-Means Algorithm. In: Caelli, T.M., Amin, A., Duin, R.P.W., Kamel, M.S., de Ridder, D. (eds.) SPR 2002 and SSPR 2002. LNCS, vol. 2396, pp. 442–451. Springer, Heidelberg (2002)
9. Fred, A.: Finding Consistent Clusters in Data Partitions. In: Kittler, J., Roli, F. (eds.) MCS 2001. LNCS, vol. 2096, pp. 309–318. Springer, Heidelberg (2001)
10. Freund, Y., Schapire, R.E.: A Decision-theoretic Generalization of Online Learning and An Application to Boosting. In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995)

11. Hadjitodorov, S.T., Kuncheva, L.I., Todorova, L.P.: Moderate diversity for better cluster ensembles. Inf. Fusion 7(3), 264–275 (2006)
12. Jouve, P., Nicoloyannis, N.: A new method for combining partitions, applications for distributed clustering. In: International Workshop on Paralell and Distributed Machine Learning and Data Mining (ECML/PKDD 2003), pp. 35–46 (2003)
13. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (eds.) Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press (1967)
14. Minaei-Bidgoli, B., Topchy, A., Punch, W.F.: Ensembles of partitions via data resampling. In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC 2004), vol. 2, pp. 188–192. IEEE Computer Society, Washington, DC (2004)
15. Saffari, A., Bischof, H.: Boosting for Model-Based Data Clustering. In: Rigoll, G. (ed.) DAGM 2008. LNCS, vol. 5096, pp. 51–60. Springer, Heidelberg (2008)
16. Sneath, P.H., Sokal, R.: Numerical Taxonomy: The Principles and Practice of Numerical Classification (1973)
17. Sokal, R.R., Michener, C.D.: A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin 28, 1409–1438 (1958)
18. Strehl, A., Ghosh, J.: Cluster ensembles — a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583–617 (2003)
19. Topchy, A., Jain, A.K., Punch, W.: Combining multiple weak clusterings, pp. 331–338 (2003)
20. Topchy, A., Minaei-Bidgoli, B., Jain, A.K., Punch, W.F.: Adaptive clustering ensembles. In: ICPR 2004: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR 2004), vol. 1, pp. 272–275. IEEE Computer Society, Washington, DC (2004)
21. Vega-Pons, S., Ruiz-Shulcloper, J.: A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence 25(03), 337–372 (2011)
22. Dimitriadou, E., Weingessel, A., Hornik, K.: Voting-Merging: An Ensemble Method for Clustering. In: Dorffner, G., Bischof, H., Hornik, K. (eds.) ICANN 2001. LNCS, vol. 2130, pp. 217–224. Springer, Heidelberg (2001)
23. Zhai, S.L., Luo, B., Guo, Y.T.: Fuzzy clustering ensemble based on dual boosting. In: Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2007, vol. 02, pp. 240–244. IEEE Computer Society, Washington, DC (2007)

An Explicit Description of the Extended Gaussian Kernel

Yong Liu and Shizhong Liao

School of Computer Science and Technology
Tianjin University,
Tianjin, China
szliao@tju.edu.cn

Abstract. Kernel methods play an important role in machine learning, pattern recognition and data mining. Although the kernel functions are the central part of the kernel methods, little is known about the structure of its reproducing kernel Hilbert spaces (RKHS) and the eigenvalues of the integral operator. In this paper, we first give the definition of the extended Gaussian kernel which includes the Gaussian kernel as its special case. Then, through a generalization form of the Weyl inner product, we present an explicit description of the RKHS of the extended Gaussian kernel. Furthermore, using the Funk-Hecke formula, we get the eigenvalues and eigenfunctions of the integral operator on the unit sphere.

Keywords: Integral operator, Reproducing kernel Hilbert space, Extended Gaussian kernel, Eigenvalues.

1 Introduction

The reproducing kernel Hilbert space (RKHS) and the eigenvalues of the integral operator recently have attracted more and more attentions in machine learning and data mining (comprehensive treatments are found in [15,18,9,16,12]). It is thus of crucial importance, for both practical and theoretical purposes, to have a deep understanding of the RKHS and the eigenvalues of the integral operator. Steinwart et al [13] first studied the structure of the RKHS induced by the popular Gaussian kernel, and they presented an orthonormal basis for this space. Minh [6] also discussed the RKHS of the Gaussian kernel and its orthonormal basis. Scovel et al [11] developed a general theory regarding mixtures of kernels, and analyzed the RKHS of the mixture in terms of the RKHSs of the mixture components. Sun and Zhou [14] explored the RKHS associated with the translation-invariant Mercer kernels, and derived some estimates for the covering numbers which form an essential part for the analysis of some algorithms in the learning theory. Kadri et al [5] explored the potential of adopting an operator-valued kernel feature space perspective for the analysis of functional data. Ferreira and Manegatto [3,4] analyzed the reproducing kernel Hilbert spaces of positive definite kernels on a topological space.

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 88–99, 2013.

In this paper, we generalize the results associated with the Gaussian kernel [13,6] to general kernel, namely as the extended Gaussian kernel. Compared to the Gaussian kernel, the extended Gaussian kernel can be used to solve the problems where the input data need to be scaled. In addition, we also present an explicit description for the eigenvalues and the eigenfunctions of the integral operator on the unit sphere, which can be used in the theoretical analysis of kernel principal component analysis [8] and other methods that need eigenvalue and eigenfunction.

The contribution of our paper mainly consists of two aspects:

- An explicit description of the RKHS with its orthonormal basis induced by the extended Gaussian kernel.
- An explicit description of the eigenvalues and the eigenfunctions of the integral operator associated with the extended Gaussian kernel on the unit sphere.

The rest of the paper is organized as following. In Section 2, we introduce the basic facts on an RKHS, In Section 3, we define the extended Gaussian kernel and present our main results, i.e., the explicit description of the RKHS and the eigenvalues of the extended Gaussian kernel. We conclude this paper in Section 4.

2 Preliminaries

Let $\mathcal{X}$ be a nonempty set. A function K is called a kernel on $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ such that for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ we have

$$K(\boldsymbol{x}, \boldsymbol{x}') = \langle \Phi(\boldsymbol{x}'), \Phi(\boldsymbol{x}) \rangle.$$

We call Φ a feature map and $\mathcal{H}$ a feature space of K. For any finite set of points $\{\boldsymbol{x}_i\}_{i=1}^N$ in $\mathcal{X}$ and $\{a_i \in \mathbb{R}\}_{i=1}^N$, if

$$\sum_{i,j=1}^{N} a_i a_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0,$$

then the function K is said to be positive definite kernel on $\mathcal{X}$.

For a given kernel, neither the feature map nor the feature space is uniquely determined. However, one can always construct a canonical feature space, namely, the reproducing kernel Hilbert space (RKHS). Let us now recall the basic theory of this space [1].

Definition 1. *Let $\mathcal{X}$ be a nonempty set and $\mathcal{H}$ be a Hilbert function space over $\mathcal{X}$, i.e., a Hilbert space that consists of functions mapping from $\mathcal{X}$ into $\mathbb{R}$.*

1. *The space $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) if for all $\boldsymbol{x} \in \mathcal{X}$ the Dirac functional $\delta_{\boldsymbol{x}} : \mathcal{H} \to \mathbb{R}$ defined by $\delta_{\boldsymbol{x}}(f) := f(\boldsymbol{x}), f \in \mathcal{H}$, is continuous.*
2. *A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if we have $K(\cdot, \boldsymbol{x}) \in \mathcal{H}$ for all $\boldsymbol{x} \in \mathcal{X}$ and the reproducing property*

$$f(\boldsymbol{x}) = \langle f, K(\cdot, \boldsymbol{x}) \rangle$$

holds for all $f \in \mathcal{H}$ and all $\boldsymbol{x} \in \mathcal{X}$.

A Hilbert function space $\mathcal{H}$ that has a reproducing kernel K is always an RKHS. Vice versa, i.e., every RKHS has a (unique) reproducing kernel (see [10]).

3 Main Results

In this section, we will first give the definition of the extended Gaussian kernel, and then we will present an explicit description of the RKHS and the eigenvalues of the integral operator associated with the extended Gaussian kernel.

3.1 Extended Gaussian Kernel

For a multi-index $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$, if $\boldsymbol{b} = \{b_1, \ldots, b_d\}^{\mathrm{T}} \in \mathbb{R}^d$, we write $\boldsymbol{x}^{(\boldsymbol{b})} = (x_1^{b_1}, \ldots, x_d^{b_d})^{\mathrm{T}}$, if $b \in \mathbb{R}$, we write $\boldsymbol{x}^{[b]} = (x_1^b, \ldots, x_d^b)^{\mathrm{T}}$.

Definition 2 (Extended Gaussian Kernel). *Let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty set. For $\boldsymbol{b} \in \mathbb{R}^d$, the extended Gaussian kernel $K_{\boldsymbol{b}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is written as*

$$K_{\boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{z}) := \exp\left(-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})} - \boldsymbol{z}^{(\boldsymbol{b})}\|^2}{\sigma^2}\right).$$

For $b \in \mathbb{R}$, the extended Gaussian kernel $K_b : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is written as

$$K_b(\boldsymbol{x}, \boldsymbol{z}) := \exp\left(-\frac{\|\boldsymbol{x}^{[b]} - \boldsymbol{z}^{[b]}\|^2}{\sigma^2}\right),$$

where $\sigma > 0$.

Remark 1. According to the definition of the extended Gaussian kernel, we know that the popular Gaussian kernel is a special case of the extended Gaussian kernel (when $\boldsymbol{b} = \{1, \ldots, 1\}^{\mathrm{T}}$ for $\boldsymbol{b} \in \mathbb{R}^d$ or $b = 1$ for $b \in \mathbb{R}$), thus the results associated with the extended Gaussian kernels can be easily applied to the Gaussian kernel. Moreover, in practice, the input data need to be scaled, so the extended Gaussian kernel with an advisable value of $\boldsymbol{b}$ may be more useful than the Gaussian kernel.

3.2 RKHS of Extended Gaussian Kernel

Let

$$\begin{aligned}
\boldsymbol{b} &= (b_1, \dots, b_d)^{\mathrm{T}} \in \mathbb{R}^d, d \in \mathbb{N};\\
\boldsymbol{\alpha} &= (\alpha_1, \dots, \alpha_d)^{\mathrm{T}} \in (\mathbb{N} \cup \{0\})^d;\\
|\boldsymbol{\alpha}| &= \sum_{i=1}^{d} \alpha_i;\\
\boldsymbol{x}^{\boldsymbol{\alpha}} &= \prod_{i=1}^{d} x_i^{\alpha_i};\\
\boldsymbol{x}^{\boldsymbol{b},\boldsymbol{\alpha}} &= \prod_{i=1}^{d} x_i^{\alpha_i b_i}.
\end{aligned}$$

We show the RKHS $\mathcal{H}_{\boldsymbol{b}}$ of the extended Gaussian kernel $K_{\boldsymbol{b}}$ in the following theorem.

Theorem 1. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty set, for every $\sigma > 0$, $\boldsymbol{b} \in \mathbb{R}^d$. Then the extended Gaussian kernel $K_{\boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})} - \boldsymbol{z}^{(\boldsymbol{b})}\|^2}{\sigma^2}\right)$ is the reproducing kernel of the space*

$$\mathcal{H}_{\boldsymbol{b}} = \left\{ f = e^{-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})}\|^2}{\sigma^2}} \sum_{|\boldsymbol{\alpha}|=0}^{\infty} w_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{b},\boldsymbol{\alpha}} : \|f\|_K^2 < \infty \right\}, \tag{1}$$

where the inner product $\langle \cdot, \cdot \rangle_K$ on $\mathcal{H}_{\boldsymbol{b}}$ is given by

$$\langle f, g \rangle_K = \sum_{k=0}^{\infty} \frac{k!}{(2/\sigma^2)^k} \sum_{|\boldsymbol{\alpha}|=k} \frac{w_{\boldsymbol{\alpha}} \nu_{\boldsymbol{\alpha}}}{C_{\boldsymbol{\alpha}}^k}$$

for

$$\begin{aligned}
f &= e^{-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})}\|^2}{\sigma^2}} \sum_{|\boldsymbol{\alpha}|=0}^{\infty} w_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{b},\boldsymbol{\alpha}},\\
g &= e^{-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})}\|^2}{\sigma^2}} \sum_{|\boldsymbol{\alpha}|=0}^{\infty} \nu_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{b},\boldsymbol{\alpha}},\\
& f, g \in \mathcal{H}_{\boldsymbol{b}} \wedge f, g : \mathbb{R}^d \to \mathbb{R}.
\end{aligned}$$

An orthonormal basis for $\mathcal{H}_{\boldsymbol{b}}$ is

$$\left\{ e_k(\boldsymbol{x}) = e^{-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})}\|^2}{\sigma^2}} \sum_{|\boldsymbol{\alpha}|=k} \sqrt{\frac{(2/\sigma^2)^k C_{\boldsymbol{\alpha}}^k}{k!}} \boldsymbol{x}^{\boldsymbol{b},\boldsymbol{\alpha}} \right\}_{k=0}^{\infty}. \tag{2}$$

Proof. See in Appendix.A.

Remark 2. Obviously, $\mathcal{H}_{\boldsymbol{b}}$ is a function space with Hilbert norm $\|\cdot\|_K$, and the inner product $\langle\cdot,\cdot\rangle_K$ in $\mathcal{H}_{\boldsymbol{b}}$ is a simple generalization of the Weyl inner product for the homogeneous polynomial space $\mathcal{H}_d(\mathbb{R}^d)$.

Remark 3. An orthonormal basis for the RKHS induced by the Gaussian kernel $K(\boldsymbol{x},\boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{\sigma^2}\right)$ has been known in the literature ([13] and references therein). We generalize this result to the extended Gaussian kernels $K(\boldsymbol{x},\boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})}-\boldsymbol{z}^{(\boldsymbol{b})}\|^2}{\sigma^2}\right)$. In addition, our approach using the Weyl inner product leads to a much shorter proof.

Remark 4. In [13], Steinwart et al discussed how to use the explicit description of RKHS to analyze support vector machines. Thus, we can use the above results to analyze support vector machines with the extended Gaussian kernels.

3.3 Eigenvalues and Eigenfunctions of Integral Operator

In the theoretical analysis of a broad variety of methods for machine learning and data analysis, such as kernel principal component analysis [8] and spectral clustering [17], the eigenvalues and the eigenfunctions of the integral operator play a crucial role. For this reason, we will study the eigenvalues and the eigenfunctions of L_{K_b} associated with the extended Gaussian kernel.

To state our results, we need the following connection between the theory of the reproducing kernels and the theory of the integral operators, which is manifested via Mercer's theorem. Let $\mathcal{X}$ be a complete, separable metric space, equipped with a finite Borel measure μ, that is $\mu(\mathcal{X}) < \infty$. Let $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a positive definite kernel on $\mathcal{X}$ satisfying

$$\kappa = \sup_{\boldsymbol{x}\in\mathcal{X}} \sqrt{K(\boldsymbol{x},\boldsymbol{x})} < \infty.$$

We consider the integral operator $L_K : L^2_\mu(\mathcal{X}) \to L^2_\mu(\mathcal{X})$,

$$(L_K f)(\boldsymbol{x}) = \int_{\mathcal{X}} K(\boldsymbol{x},\boldsymbol{t}) f(\boldsymbol{t}) d\mu(\boldsymbol{t}).$$

This is a self-adjoint, compact operator that has eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \ldots, \lambda_i, \ldots \geq 0,$$

with the corresponding L^2_μ-normalized eigenfunctions $\{\phi_k\}_{k=1}^\infty$ forming an orthonormal basis for $L^2_\mu(\mathcal{X})$. Mercer's theorem (we refer to [2] for more detail) states that

$$K(\boldsymbol{x},\boldsymbol{t}) = \sum_{k=1}^\infty \lambda_k \phi_k(\boldsymbol{x})\phi_k(\boldsymbol{t}),$$

where the series converges absolutely for each $(\boldsymbol{x}, \boldsymbol{t}) \in \mathcal{X} \times \mathcal{X}$ and uniformly on compact subsets of $\mathcal{X} \times \mathcal{X}$.

Let $S^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$ be the d-dimensional unit sphere, with surface area $|S^{d-1}| = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}$, where Γ is the gamma function defined by $\Gamma(k) = \int_0^\infty e^{-u}u^{k-1}du$. We review the concept of spherical harmonics which is defined in [7].

Definition 3 (Spherical Harmonics). *Let* $\Delta_d = -\left[\frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_d^2}\right]$ *denote the Laplacian operator on* $\mathbb{R}^d$*. A homogeneous polynomial of degree* k *in* $\mathbb{R}^d$ *is called a homogeneous harmonic of order* k *when its Laplacian vanishes. Let* $\mathcal{Y}_k(d)$ *denote the subspace of all homogeneous harmonics of order* k *on the unit sphere* S^{d-1} *in* $\mathbb{R}^d$*. The functions in* $\mathcal{Y}_k(d)$ *are called spherical harmonics of order* k*. We denote by* $\{Y_{k,j}(d;\boldsymbol{x})\}_{j=1}^{N(d,k)}$ *any fixed orthonormal basis for* $\mathcal{Y}_k(d)$ *where* $N(d,k) = \dim \mathcal{Y}_k(d) = \frac{(2k+d-2)(k+d-3)!}{k!(d-2)!}$, $k \geq 0$.

Theorem 2. *Let* $b \in \mathbb{R}, d \in \mathbb{N}, d \geq 2$*, be fixed. Let* $\mathcal{X} = S^{d-1}$ *and* μ *be the uniform probability distribution on* S^{d-1}*. If* $\langle \boldsymbol{x}^{[b]}, \boldsymbol{z}^{[b]} \rangle = \langle \boldsymbol{x}, \boldsymbol{z} \rangle^b$ *for all* $\boldsymbol{x}, \boldsymbol{z} \in \mathcal{X}$*, for the extended Gaussian kernel*

$$K_b(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x}^{[b]} - \boldsymbol{z}^{[b]}\|^2}{\sigma^2}\right), \sigma > 0,$$

the eigenvalues of $L_{K_b} : L^2_\mu(\mathcal{X}) \to L^2_\mu(\mathcal{X})$ *are*

$$\lambda_k = |S^{d-2}| \int_{-1}^{1} \exp\left(-\frac{2 - 2t^b}{\sigma^2}\right) P_k(d;t)(1-t^2)^{\frac{d-3}{2}} dt,$$

for all $k \in \mathbb{N} \cup \{0\}$*. Each* λ_k *occurs with multiplicity* $N(d,k)$*, and the corresponding eigenfunctions are the spherical harmonics of order* k *on* S^{d-1}*.*

Proof. See in Appendix.B.

Remark 5. Note that if $b = 1$ or $d = 1$, the assumption $\langle \boldsymbol{x}^{[b]}, \boldsymbol{z}^{[b]} \rangle = \langle \boldsymbol{x}, \boldsymbol{z} \rangle^b$ for all $\boldsymbol{x}, \boldsymbol{z} \in \mathcal{X}$ in the above theorem is satisfied. Thus, when we let $b = 1$, we can obtain the eigenvalues and the eigenfunctions of the integral operator induced by the Gaussian kernel.

Corollary 1. *Let* $d \in \mathbb{N}, d \geq 2$, $\mathcal{X} = S^{d-1}$*, and* μ *be the uniform probability distribution on* S^{d-1}*. For the Gaussian kernel*

$$K(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{z}\|^2}{\sigma^2}\right), \sigma > 0,$$

the eigenvalues of $L_K : L^2_\mu(\mathcal{X}) \to L^2_\mu(\mathcal{X})$ *are*

$$\lambda_k = |S^{d-2}| \int_{-1}^{1} \exp\left(-\frac{2 - 2t}{\sigma^2}\right) P_k(d;t)(1-t^2)^{\frac{d-3}{2}} dt,$$

for all $k \in \mathbb{N} \cup \{0\}$*. Each* λ_k *occurs with multiplicity* $N(d,k)$ *with the corresponding eigenfunctions being spherical harmonics of order* k *on* S^{d-1}*.*

Proof. Since the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{\sigma^2}\right)$ is a special case of extended Gaussian kernel when $b = 1$, we can prove the corollary by using the result of Theorem 2.

The radial kernel $K(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\sigma^2}\right), \sigma > 0$ is another popular kernel in machine learning and data mining. For both theoretical and practical purposes, we need to study the eigenvalues and the eigenfunctions of the integral operator associated with this radial kernel.

Theorem 3. *Let $d \in \mathbb{N}, d \geq 2$, $\mathcal{X} = S^{d-1}$, and μ be the uniform probability distribution on S^{d-1}. For the radial kernel*

$$K(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\sigma^2}\right), \sigma > 0,$$

the eigenvalues of $L_K : L^2_\mu(\mathcal{X}) \to L^2_\mu(\mathcal{X})$ are

$$\lambda_k = |S^{d-2}| \int_{-1}^{1} \exp\left(-\frac{\sqrt{2-2t}}{\sigma^2}\right) P_k(d;t)(1-t^2)^{\frac{n-3}{2}} dt,$$

for all $k \in \mathbb{N} \cup \{0\}$. Each λ_k occurs with multiplicity $N(d,k)$ with the corresponding eigenfunctions being spherical harmonics of order k on S^{d-1}.

Proof. See in Appendix.C.

4 Conclusion

In this paper, we have generalized the results of the explicit description of the reproducing kernel Hilbert space (RKHS) associated with the Gaussian kernel [13,6] to the extended Gaussian kernel. In addition, we have presented the explicit description for the eigenvalues and eigenfunctions of the integral operator. These results can be used in the theoretical analysis of the kernel principal component analysis and other methods which need analysis of the eigenvalues and the eigenfunctions.

We will apply the results of this paper to analyze the learning performance of SVM or other kernel-based methods, and to explore a new criterion for model selection of kernel methods.

Acknowledgments. The work is supported in part by the Natural Science Foundation of China under grant No. 61170019, and the Natural Science Foundation of Tianjin under grant No. 11JCYBJC00700.

Appendix

This section gives the proofs for the theorems in the main text.

Appendix A

In order to prove Theorem 1, we first introduce the following lemma.

Lemma 1 (Aronszajn [1]). *Let $\mathcal{H}$ be a separable Hilbert space of functions over $\mathcal{X}$ with orthonormal basis $\{e_k\}_{k=0}^{\infty}$. $\mathcal{H}$ is a reproducing kernel Hilbert space iff*

$$\sum_{k=0}^{\infty} |e_k(\boldsymbol{x})|^2 < \infty$$

for all $\boldsymbol{x} \in \mathcal{X}$. The unique kernel K is defined by

$$K(\boldsymbol{x}, \boldsymbol{z}) = \sum_{k=0}^{\infty} e_k(\boldsymbol{x}) e_k(\boldsymbol{z}).$$

Proof (of Theorem 1). Note that for any vector $\boldsymbol{x}, \boldsymbol{z}$,

$$e^{\langle \boldsymbol{x}, \boldsymbol{z} \rangle} = \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{|\boldsymbol{\alpha}|=k} C_{\boldsymbol{\alpha}}^k \boldsymbol{x}^{\boldsymbol{\alpha}} \boldsymbol{z}^{\boldsymbol{\alpha}},$$

thus we can obtain that

$$\begin{aligned} K_{\boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{z}) &= \exp\left(-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})} - \boldsymbol{z}^{(\boldsymbol{b})}\|^2}{\sigma^2}\right) \\ &= \exp\left(-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})}\|^2}{\sigma^2}\right) \exp\left(-\frac{\|\boldsymbol{z}^{(\boldsymbol{b})}\|^2}{\sigma^2}\right) \exp\left(\frac{2\langle \boldsymbol{x}^{(\boldsymbol{b})}, \boldsymbol{z}^{(\boldsymbol{b})} \rangle}{\sigma^2}\right) \\ &= \exp\left(-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})}\|^2}{\sigma^2}\right) \exp\left(-\frac{\|\boldsymbol{z}^{(\boldsymbol{b})}\|^2}{\sigma^2}\right) \sum_{k=0}^{\infty} \frac{(2/\sigma^2)^k}{k!} \sum_{|\boldsymbol{\alpha}|=k} C_{\boldsymbol{\alpha}}^k \boldsymbol{x}^{\boldsymbol{b},\boldsymbol{\alpha}} \boldsymbol{z}^{\boldsymbol{b},\boldsymbol{\alpha}}. \end{aligned}$$

Let $\mathcal{H}_0 = \left\{ f = e^{-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})}\|^2}{\sigma^2}} \sum_{|\boldsymbol{\alpha}|=0}^{\infty} w_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{b},\boldsymbol{\alpha}} : \sum_{k=0}^{\infty} \frac{k!}{(2/\sigma^2)^k} \sum_{|\boldsymbol{\alpha}|=k} \frac{w_{\boldsymbol{\alpha}}^2}{C_{\boldsymbol{\alpha}}^k} < \infty \right\}$. For

$$f = e^{-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})}\|}{\sigma^2}} \sum_{|\boldsymbol{\alpha}|=0}^{\infty} w_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{b},\boldsymbol{\alpha}} \in \mathcal{H}_0,$$

$$g = e^{-\frac{\|\boldsymbol{x}^{(\boldsymbol{b})}\|}{\sigma^2}} \sum_{|\boldsymbol{\alpha}|=0}^{\infty} \nu_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{b},\boldsymbol{\alpha}} \in \mathcal{H}_0,$$

we define the inner product

$$\langle f, g \rangle_{K,0} = \sum_{k=0}^{\infty} \frac{k!}{(2/\sigma^2)^k} \sum_{|\boldsymbol{\alpha}|=k} \frac{w_{\boldsymbol{\alpha}} \nu_{\boldsymbol{\alpha}}}{C_{\boldsymbol{\alpha}}^k}.$$

We will show that $\mathcal{H}_0$ is itself a separable Hilbert space under $\langle\cdot,\cdot\rangle_{K,0}$. For simplicity, let $d=1$. Then

$$\mathcal{H}_0=\left\{f=e^{-\frac{x^{2b}}{\sigma^2}}\sum_{k=0}^{\infty}w_kx^{bk}:\ \sum_{k=0}^{\infty}\frac{k!}{(2/\sigma^2)^k}w_k^2<\infty\right\}.$$

It is clear that $\mathcal{H}_0$ is an inner product space under $\langle\cdot,\cdot\rangle_{K,0}$. Its completeness under the induced norm $\|\cdot\|_{K,0}$ is equivalent to the completeness of the weighted ℓ^2 sequence space

$$\ell^2_\sigma=\left\{(w_k)_{k=0}^{\infty}:\|(w_k)_{k=0}^{\infty}\|_{\ell^2_\sigma}=\left(\sum_{k=0}^{\infty}\frac{k!}{(2/\sigma^2)^k}w_k^2\right)^{1/2}\right\},$$

which is itself a separable Hilbert space. Thus $(H_0,\|\cdot\|_{K,0})$ is a separable Hilbert space.

If $\mathcal{X}\subset\mathbb{R}^d$ has non-empty interior, then the mononomials $\boldsymbol{x}^{\boldsymbol{b},\boldsymbol{\alpha}}$ are all distinct. From the definition of the inner product $\langle\cdot,\cdot\rangle_{K,0}$, it is easy to obtain that

$$\langle e_i,e_j\rangle_K=\begin{cases}0, & \text{if } i\neq j;\\ 1, & \text{otherwise;}\end{cases}$$

where e_k are given in (2). So $\{e_k\}_{k=0}^{\infty}$ are orthonormal under $\langle\cdot,\cdot\rangle_{K,0}$. Moreover, $\mathcal{H}_0=\operatorname{span}\{e_k,k=0,1,\ldots\}$, thus, $\{e_k\}_{k=0}^{\infty}$ forms an orthonormal basis for $(\mathcal{H}_0,\|\cdot\|_{K,0})$. By Lemma 1 and the following equation

$$\sum_{k=0}^{\infty}|e_k(\boldsymbol{x})|^2=K(\boldsymbol{x},\boldsymbol{x})=1<\infty,$$

we can obtain that $\mathcal{H}_{\boldsymbol{b}}$ is a reproducing kernel Hilbert space. Note that

$$\sum_{k=0}^{\infty}e_k(\boldsymbol{x})e_k(\boldsymbol{z})=K_b(\boldsymbol{x},\boldsymbol{z}),$$

and since the RKHS induced by a kernel on a set $\mathcal{X}$ is unique, thus $(\mathcal{H}_0,\|\cdot\|_{K,0})$ is the reproducing kernel Hilbert space of functions on $\mathcal{X}$ with the extended Guassian kernel $K_{\boldsymbol{b}}(\boldsymbol{x},\boldsymbol{z})$.

Appendix B

In order to obtain the eigenvalues and eigenfunctions of the integral operator associated with the extended Gaussian kernel, we first give the following lemma.

Lemma 2. *Let $d\in\mathbb{N},d\geq 2$ be fixed. Let $K:[-1,1]\to\mathbb{R}$ be a continuous function, giving rise to a continuous, positive definite kernel $K(\boldsymbol{x},\boldsymbol{t})=K(\langle\boldsymbol{x},\boldsymbol{t}\rangle)$ on $S^{d-1}\times S^{d-1}$. Let μ be the Lebesgue measure on S^{d-1}. The eigenvalues λ_k of*

$$L_K:L^2_\mu(S^{d-1})\to L^2_\mu(S^{d-1})$$

are given by

$$\lambda_k = |S^{d-2}| \int_{-1}^{1} K(t) P_k(d;t)(1-t^2)^{\frac{d-3}{2}} dt,$$

each with multiplicity $N(d,k)$*, for* $k \in \mathbb{Z}, k \geq 0$*, where* $P_k(d;t)$ *is Legendre polynomial of degree* k *in dimension* d*,*

$$P_k(d;t) = k!\Gamma\left(\frac{d-1}{2}\right) \sum_{l=0}^{\lceil \frac{k}{2} \rceil} \left(\frac{-1}{4}\right)^l \frac{(1-t^2)^l t^{k-2l}}{l!(k-2l)!\Gamma(l+\frac{d-1}{2})}.$$

The corresponding eigenfunctions for each λ_k *are the spherical harmonics* $\{Y_{k,j}(d;\boldsymbol{x})\}_{j=1}^{N(d,k)}$ *of the order* k*.*

Proof. Let $f : [-1,1] \to \mathbb{R}$ be a continuous function. Let $Y_k \in \mathcal{Y}_k(d)$ for $k \geq 0$. Then Funk-Hecke formula ([7], p 30) states that for any $\boldsymbol{x} \in S^{d-1}$:

$$\int_{S^{d-1}} f(\langle \boldsymbol{x}, \boldsymbol{t} \rangle) Y_k(t) dS^{d-1}(\boldsymbol{t}) = \lambda_k Y_k(\boldsymbol{x}), \tag{3}$$

where

$$\lambda_k = |S^{d-2}| \int_{-1}^{1} f(t) P_k(d;\boldsymbol{t})(1-t^2)^{\frac{d-3}{2}} dt \tag{4}$$

and $P_k(d;\boldsymbol{t})$ denotes the Legendre polynomial of degree k in dimension d. The spherical harmonics $\left\{\{Y_{k,j}(d;\boldsymbol{x})\}_{j=1}^{N(d,k)}\right\}_{k=0}^{\infty}$ form an orthonormal basis for $L^2(S^{d-1})$. So if the kernel K on $S^{d-1} \times S^{d-1}$ is defined by $K(\boldsymbol{x}, \boldsymbol{t}) = f(\langle \boldsymbol{x}, \boldsymbol{t} \rangle)$, via the Funk-Hecke formula, it is easy to verify that the eigenvalues of

$$L_K : L^2_\mu(S^{d-1}) \to L^2_\mu(S^{d-1})$$

are given precisely by (4), with the corresponding orthonormal eigenfunctions of $\{Y_{k,j}(d;\boldsymbol{x})\}_{j=1}^{N(d,k)}$. The multiplicity of λ_k is therefore $N(d,k) = \dim(\mathcal{Y}_k(d))$.

Proof (of Theorem 2). Note that

$$\exp\left(-\frac{\|\boldsymbol{x}^{[b]} - \boldsymbol{z}^{[b]}\|^2}{\sigma^2}\right) = \exp\left(-\frac{\|\boldsymbol{x}^{[b]}\|^2 + \|\boldsymbol{z}^{[b]}\|^2 - 2\langle \boldsymbol{x}^{[b]}, \boldsymbol{z}^{[b]} \rangle}{\sigma^2}\right),$$

since $\boldsymbol{x}, \boldsymbol{z} \in S^{d-1}$ and $\langle \boldsymbol{x}^{[b]}, \boldsymbol{z}^{[b]} \rangle = \langle \boldsymbol{x}, \boldsymbol{z} \rangle^b$, so it is easy to obtain that

$$\exp\left(-\frac{\|\boldsymbol{x}^{[b]} - \boldsymbol{z}^{[b]}\|^2}{\sigma^2}\right) = \exp\left(-\frac{2 - 2\langle \boldsymbol{x}, \boldsymbol{z} \rangle^b}{\sigma^2}\right).$$

Thus, using the Lemma 2, we know that the eigenvalues of

$$L_{K_b} : L^2_\mu(\mathcal{X}) \to L^2_\mu(\mathcal{X})$$

are

$$\lambda_k = |S^{d-2}| \int_{-1}^{1} \exp\left(-\frac{2-2t^b}{\sigma^2}\right) P_k(d;t)(1-t^2)^{\frac{d-3}{2}} dt,$$

and each λ_k occurs with multiplicity $N(d,k)$ with the corresponding eigenfunctions being spherical harmonics of order k on S^{d-1}.

Appendix C

Proof (of Theorem 3). On S^{d-1}, it is easy to verify that

$$\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\sigma^2}\right) = \exp\left(-\frac{\sqrt{(2-2\langle \boldsymbol{x}, \boldsymbol{z}\rangle)}}{\sigma^2}\right).$$

Thus, using the Lemma 2, we know that the eigenvalues of

$$L_K : L^2_\mu(\mathcal{X}) \to L^2_\mu(\mathcal{X})$$

are

$$\lambda_k = |S^{d-2}| \int_{-1}^{1} \exp\left(-\frac{\sqrt{2-2t}}{\sigma^2}\right) P_k(d;t)(1-t^2)^{\frac{d-3}{2}} dt,$$

and each λ_k occurs with multiplicity $N(d,k)$ with the corresponding eigenfunctions being spherical harmonics of order k on S^{d-1}.

References

1. Aronszajn, N.: Theory of reproducing kernels. Transactions of the American Mathematical Society 68, 337–404 (1950)
2. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bulletin of the American Mathematical Society 39(1), 1–49 (2001)
3. Ferreira, J.C., Manegatto, V.A.: Reproducing kernel hilbert spaces associated with kernels on topological spaces. Functional Analysis and Its Applications 46(2), 152–154 (2012)
4. Ferreira, J.C., Manegatto, V.A.: Reproducing properties of differentiable mercer-like kernels. Mathematische Nachrichten 285(8-9), 959–973 (2012)
5. Kadri, H., Rabaoui, A., Preux, P., Duflos, E., Rakotomamonjy, A.: Functional regularized least squares classification with operator-valued kernels. In: Proceeding of the 28th International Conference on Machine Learning (ICML 2011), pp. 993–1000 (2011)
6. Minh, H.Q., Niyogi, P., Yao, Y.: Mercer's Theorem, Feature Maps, and Smoothing. In: Lugosi, G., Simon, H.U. (eds.) COLT 2006. LNCS (LNAI), vol. 4005, pp. 154–168. Springer, Heidelberg (2006)
7. Müller, C.: Analysis of spherical symmetries in Euclidean spaces. Applied Mathematical Sciences, vol. 129. Springer, New York (1998)
8. Schölkopf, B., Smola, A.J., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10(5), 1299–1319 (1998)

9. Schölkopf, B., Smola, A.J.: Learning with kernels: Support vector machines, regularization, optimization, and beyond. The MIT Press (2002)
10. Schölkopf, B., Smola, A.J., Müller, K.-R.: Kernel Principal Component Analysis. In: Gerstner, W., Hasler, M., Germond, A., Nicoud, J.-D. (eds.) ICANN 1997. LNCS, vol. 1327, pp. 583–588. Springer, Heidelberg (1997)
11. Scovel, C., Hush, D., Steinwart, I., Theiler, J.: Radial kernels and their reproducing kernel Hilbert spaces. Journal of Complexity 26(6), 641–660 (2010)
12. Smale, S., Zhou, D.X.: Learning theory estimates via integral operators and their approximations. Constructive Approximation 26(2), 153–172 (2007)
13. Steinwart, I., Hush, D., Scovel, C.: An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. IEEE Transactions on Information Theory 52(10), 4635–4643 (2006)
14. Sun, H.W., Zhou, D.X.: Reproducing kernel Hilbert spaces associated with analytic translation-invariant Mercer kernels. Journal of Fourier Analysis and Applications 14(1), 89–101 (2008)
15. Vapnik, V.: The nature of statistical learning theory. Springer (2000)
16. Vito, E.D., Caponnetto, A., Rosasco, L.: Model selection for regularized least-squares algorithm in learning theory. Foundations of Computational Mathematics 5(1), 59–85 (2005)
17. Von Luxburg, U., Bousquet, O., Belkin, M.: On the Convergence of Spectral Clustering on Random Samples: The Normalized Case. In: Shawe-Taylor, J., Singer, Y. (eds.) COLT 2004. LNCS (LNAI), vol. 3120, pp. 457–471. Springer, Heidelberg (2004)
18. Wahba, G.: Spline models for observational data. SIAM (1990)

An Improved Genetic Clustering Algorithm for Categorical Data

Hongwu Qin, Xiuqin Ma, Tutut Herawan, and Jasni Mohamad Zain

Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang
Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia
{qhwump,xueener}@gmail.com, {tutut,jasni}@ump.edu.my

Abstract. Deng *et al*. [Deng, S., He, Z., Xu, X.: G-ANMI: A mutual information based genetic clustering algorithm for categorical data, Knowledge-Based Systems 23, 144--149(2010)] proposed a mutual information based genetic clustering algorithm named G-ANMI for categorical data. While G-ANMI is superior or comparable to existing algorithms for clustering categorical data in terms of clustering accuracy, it is very time-consuming due to the low efficiency of genetic algorithm (GA). In this paper, we propose a new initialization method for G-ANMI to improve its efficiency. Experimental results show that the new method greatly improves the efficiency of G-ANMI as well as produces higher clustering accuracy.

Keywords: Data mining, Clustering, Categorical data, Genetic algorithm.

1 Introduction

Clustering is an important data mining technique that groups together similar data objects. Most previous clustering algorithms focus on numerical data whose inherent geometric properties can be exploited naturally to define distance functions between objects. However, many fields, from statistics to psychology deal with categorical data. Unlike numerical data, categorical data cannot be naturally ordered. An example of categorical attribute is *color* whose values include *red*, *green*, *blue*, etc. Therefore, those clustering algorithms dealing with numerical data can not be used to cluster categorical data. Recently, the problem of clustering categorical data has received much attention [1-10].

Categorical data clustering has been defined as an optimization problem which aims to find an optimal partition of the objects according to an objective function [1-7]. Unfortunately, this optimization problem is NP-complete. Therefore most researchers resort to heuristic methods to solve it, such as ROCK [1], k-modes [2], COOLCAT [3], and k-ANMI [4]. However, these algorithms tend to find local optimal partition. Recently, some genetic clustering algorithms have been proposed to find globally optimal or near-optimal partition, such as ALG-RAND [6] and G-ANMI [7] algorithms. In the performance comparison conducted in [7], it has shown that

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 100–111, 2013.

G-ANMI is superior or comparable to ALG-RAND as well as other existing algorithms for clustering categorical data in terms of clustering accuracy. However, G-ANMI is very time-consuming. For instance, it takes G-ANMI 20759 seconds to mine 2 clusters from Mushroom dataset [11] with 8124 objects. Thus, it is necessary to improve its efficiency before it can be widely used in practice.

The low efficiency of G-ANMI is mainly caused by GA [8] which needs a lot of iterations to find the optimal solution. Given a population size, the efficiency of G-ANMI is dominated by the number of iterations. Hence, we have to reduce the number of iterations to improve the efficiency of G-ANMI. In a categorical data set, each attribute defines a partition of the objects. The aim of G-ANMI is to find a k-partition (k is the desired number of clusters) that shares the most information with the partitions defined by attributes (attributes partitions for short). In other words, G-ANMI tries to find a k-partition that is the closest to the attributes partitions. However, G-ANMI algorithm starts with a population of randomly generated k-partitions of objects. These randomly generated k-partitions are far from the attributes partitions when we process a larger data set. The farther these partitions are from the attributes partitions, the more iteration G-ANMI needs to reach the optimal k-partition. Hence, it is possible to reduce the number of iterations of G-ANMI by giving some better initial k-partitions which are closer to the attributes partitions in comparison with those randomly generated k-partitions.

In this paper, we propose a new initialization method for G-ANMI, in which some equivalence classes (the set of objects which has the same value on an attribute) in attributes partitions are directly integrated into the initial k-partitions. The initial k-partitions obtained by using the new method are closer to the attributes partitions in comparison with those randomly generated k-partitions, especially when we process a larger data set. As a result, less number of iterations is needed to reach the optimal k-partition. Experimental results show that the new method greatly improves the efficiency of G-ANMI, as well as produces higher clustering accuracy. The rest of the paper is organized as follows. Section 2 briefly introduces G-ANMI algorithm. Section 3 presents the new initialization method. Section 4 presents experimental results on UCI benchmark data sets. Finally, Section 5 presents conclusions and future work.

2 G-ANMI

G-ANMI employs basic GA to implement categorical data clustering, which works in the same way as the one used in ALG-RAND [6].

G-ANMI starts with a population of randomly generated partitions of objects, which are encoded as chromosomes. If the desired number of clusters is set to k, then each chromosome is encoded as a k-partition of objects. Suppose the integers between interval [0, k-1] are used as class identifier, a chromosome will be a string of integers which are between interval [0, k-1]. For example, suppose the number of objects is 20, and k is 4, a randomly generated chromosome is as follows

1 0 2 0 1 0 3 2 3 1 0 1 2 0 3 2 0 1 1 2

Then, G-ANMI uses the average normalized mutual information (ANMI) to evaluate the fitness of each chromosome in the current population. Given a set of r partitions defined by attributes: $\Lambda = \{\lambda^{(q)} \mid q \in \{1,2,...,r\}\}$ and a partition $\overline{\lambda}$, the average normalized mutual information (ANMI) between Λ and $\overline{\lambda}$ is defined as follows:

$$\phi^{(ANMI)}(\Lambda, \overline{\lambda}) = \frac{1}{r}\sum_{q=1}^{r} \phi^{(NMI)}(\overline{\lambda}, \lambda^{(q)}) \tag{1}$$

where $\phi^{(NMI)}(\overline{\lambda}, \lambda^{(q)})$ denotes the normalize mutual information between $\lambda^{(q)}$ and $\overline{\lambda}$. Without loss of generality, normalized mutual information between two partitions $\lambda^{(a)}$ and $\lambda^{(b)}$ is computed as follows:

$$\phi^{(NMI)}(\lambda^{(a)}, \lambda^{(b)}) = \frac{2}{n}\sum_{h=1}^{k^{(a)}}\sum_{g=1}^{k^{(b)}} n_g^{(h)} \log_{k^{(a)} * k^{(b)}}\left(\frac{n_g^{(h)} n}{n^{(h)} n_g}\right) \tag{2}$$

where $k^{(a)}$ and $k^{(b)}$ are the number of clusters in partition $\lambda^{(a)}$ and $\lambda^{(b)}$, respectively. $n^{(h)}$ denotes the size of cluster C_h in partition $\lambda^{(a)}$, n_g denotes the size of cluster C_g in partition $\lambda^{(b)}$, $n_g^{(h)}$ denotes the number of shared objects between C_h and C_g.

According to the fitness value, genetic evolution repeatedly changes the chromosomes in the current population to generate a new population. It is expected that chromosomes could be increasingly closer to the optimal partition with largest ANMI. Genetic procedure will halt when the best fitness in the current population is greater than the user-specified fitness threshold or there has been no relative improvement on best fitness after some consecutive iterations.

3 New Initialization Method

The basic idea of the new initialization method is that integrating some equivalence classes of the partitions defined by attributes into the generation of initial partitions. Two cases are considered:

i. If the population size *P* is greater than or equal to the number of attributes *M*, then the algorithm generates first *M* chromosomes from the *M* attributes partitions, and generates other *P-M* chromosomes randomly.

ii. If the population size *P* is less than the number of attributes *M*, then the algorithm generates *P* chromosomes from the first *P* attributes partitions.

Generating chromosomes from the attributes partitions is implemented by a one-one way, namely one chromosome is generated by one partition. Generating a chromosome from a partition means taking some equivalence classes of the partition as the part of the chromosome. How many equivalence classes should we take depends on the number of equivalence classes (*Nec*) in the partition and the specified number of clusters *k*. Different strategies are employed when the number of equivalence classes in the partition is greater than, less than and equals to the specified number of clusters, respectively. The details are described in Fig .1.

```
Begin

  For each partition Par
     if Nec in Par equals k
        Copy Par to the corresponding chromosome Chrom
     else
        if Nec is greater than k
           Copy first k equivalence classes of Par to the same locations in
             Chrom.
           Generate a random number between [0, k-1] for each of the remain-
             ing locations in Chrom.
        else
           Find a highest H which satisfies the following inequation
                          N - Sum >= k-H-1
          //where N is the length of a chromosome, Sum is the summation
          //of the size of first H+1 equivalence classes of Par.
          Copy first H+1 equivalence classes of Par to the same locations in
             Chrom.
          Generate a random number between [H+1, k-1] for each of the
             remaining locations in Chrom.
End.
```

Fig. 1. The procedure of generating a chromosome from a partition

Note that the purpose of inequation N - Sum >= k-H-1 is to ensure each number between interval [H+1, k-1] appears at least once in *Chrom* when generating a random number for each of the remaining locations in *Chrom*.

Next, we present an illustrative example of the new initialization method. For the comparison purpose, the G-ANMI algorithm with new initialization method is named improved G-ANMI (IG-ANMI).

Example 1. Suppose there is a data set with ten objects (O_1, O_2, …, O_{10}) and four attributes (A_0, …, A_3). Table 1 shows the partitions defined by the four attributes. The numbers 0, 1, 2, and 3 denote different equivalence classes (categories) in the partitions. We use the algorithms IG-ANMI and G-ANMI to cluster the objects, respectively. The parameter setting includes: the number of clusters k=3, the population size P=10, crossover rate=0.8, mutation rate=0.1, random seed=1, and the number of consecutive iterations without improvement=100.

Since the population size P is greater than the number of attributes, we generate first four chromosomes by using attributes partitions and generate remaining six chromosomes randomly. The attributes partitions are named *Par*[i], i=0, 1, 2, 3. The chromosomes are named C[j], j=0, 1, …, 9. The numbers of equivalence classes in *Par*[0] and *Par*[2] equal the specified number of clusters k, so we directly copy *Par*[0] and *Par*[2] to C[0] and C[2], respectively. The number of equivalence classes in *Par*[1] is less than k. According to the algorithm shown in Figure 1, we first seek an appropriate number H. In this example, there is only possible value for H, namely zero. Zero satisfies N - Sum >= k-H-1, thus H gets the value zero. Next, we copy the

Table 1. The partitions defined by four attributes

U	A_0	A_1	A_2	A_3
O_1	0	0	0	0
O_2	1	0	1	1
O_3	0	1	0	0
O_4	0	0	0	0
O_5	1	1	2	2
O_6	1	1	1	2
O_7	2	0	2	3
O_8	2	1	2	1
O_9	1	1	1	2
O_{10}	2	1	2	3

first equivalence class to the corresponding location in $C[1]$. Table 2 shows the status of $C[1]$ after copying the first equivalence class. There are still six locations need to be filled in $C[1]$. We generate a random number between interval [1, 2] for each of the six locations.

Table 2. The status of $C[1]$ after copying the first equivalence class

Location	0	1	2	3	4	5	6	7	8	9
$C[1]$	0	0		0			0			

The number of equivalence classes in $Par[3]$ is greater than k. According to the algorithm shown in Figure 1, we copy first three equivalence classes of $Par[3]$ to the corresponding location in $C[3]$. Table 3 shows the status of $C[3]$ after copying first three equivalence classes. There are still two locations need to be filled in $C[3]$. We generate a random number between interval [0, 2] for each of the two locations.

Table 3. The status of $C[3]$ after copying first three equivalence classes

Location	0	1	2	3	4	5	6	7	8	9
$C[3]$	0	1	0	0	2	2		1	2	

Following the method, the remaining six chromosomes $C[4]$, … $C[9]$ are randomly generated. At the end, ten chromosomes are obtained and summarized in Table 4. The numbers in bold style are randomly generated.

Note that the equivalence classes in each of the first four chromosomes are labeled by order 0, 1, 2. However, the equivalence classes in each of other six chromosomes are labeled unorderly. Actually, the numbers 0, 1, 2 in the partitions or chromosomes only denote different categories rather than order. That means the order of the labels doesn't affect the computation of fitness of a chromosome. Even if we change the order of the labels in some chromosomes, their fitness values keep invariable. For instance, we can change $C[0]$ from {0, 1, 0, 0, 1, 1, 2, 2, 1, 2} to {1, 2, 1, 1, 2, 2, 0, 0, 2, 0}, change $C[4]$ from {1, 1, 1, 2, 1, 1, 0, 1, 0, 0} to {0, 0, 0, 1, 0, 0, 2, 0, 2, 2}, and so on.

Table 4. Ten chromosomes generated by the new initialization method

Location	*C*[0]	*C*[1]	*C*[2]	*C*[3]	*C*[4]	*C*[5]	*C*[6]	*C*[7]	*C*[8]	*C*[9]
0	0	0	0	0	**1**	**2**	**1**	**2**	**1**	**1**
1	1	0	1	1	**1**	**0**	**1**	**1**	**1**	**2**
2	0	**1**	0	0	**1**	**2**	**0**	**0**	**0**	**0**
3	0	0	0	0	**2**	**1**	**1**	**1**	**1**	**0**
4	1	**1**	2	2	**1**	**2**	**1**	**1**	**0**	**2**
5	1	**1**	1	2	**1**	**2**	**1**	**1**	**0**	**0**
6	2	0	2	**0**	**0**	**2**	**2**	**0**	**1**	**0**
7	2	**1**	2	1	**1**	**2**	**2**	**0**	**0**	**1**
8	1	**1**	1	2	**0**	**0**	**1**	**2**	**2**	**2**
9	2	**2**	2	**1**	**0**	**0**	**2**	**1**	**0**	**0**

After the initialization, the next step of IG-ANMI is to calculate the fitness of each chromosome. According to the Eq. (1), we obtain the fitness of chromosomes as is shown in Table 5.

Table 5. The fitness of initial chromosomes of IG-ANMI

Chromosomes	fitness value	average
C[0]	0.654067	0.53927
C[1]	0.361562	
C[2]	0.615644	
C[3]	0.525807	
C[4]	0.252361	0.265807
C[5]	0.196573	
C[6]	0.403407	
C[7]	0.166014	
C[8]	0.323139	
C[9]	0.253349	
average	0.375192	

It can be seen from Table 5 that the average fitness value of first four chromosomes is higher than that of other six chromosomes, which indicates that the chromosomes generated from the attributes partitions are closer to the optimal partition than that generated randomly. With these fitness values, the algorithm IG-ANMI generates new population and goes into the next iteration. Since there has been no relative improvement on best fitness during 100 consecutive iterations, the algorithm IG-ANMI ends after the 100^{th} iteration. Finally, we get the optimal 3-partition {0, 1, 0, 0, 1, 1, 2, 2, 1, 2}.

We use G-ANMI algorithm to cluster the same data set below. Firstly, G-ANMI randomly generates *P* chromosomes as is shown in Table 6.

Table 7 shows the fitness values of the chromosomes in the initial population. Obviously, the average fitness as well as the best fitness of the chromosomes is less than that in the initial population generated by algorithm IG-ANMI. After 27 iterations, the best fitness reaches 0.654067, which equals to the best fitness of the initial population generated by algorithm IG-ANMI. Since there has been no relative improvement on best fitness during the subsequent 99 consecutive iterations, the algorithm G-ANMI ends after the 127th iterations. G-ANMI needs 27 more iterations than IG-ANMI due to the randomly generated initial population.

Table 6. Ten chromosomes generated by the initialization method of G-ANMI

Location	*C*[0]	*C*[1]	*C*[2]	*C*[3]	*C*[4]	*C*[5]	*C*[6]	*C*[7]	*C*[8]	*C*[9]
0	1	1	2	0	0	0	0	0	2	2
1	1	2	1	1	2	1	1	0	1	1
2	2	1	2	1	2	1	0	2	1	0
3	1	1	2	1	0	1	0	0	2	1
4	1	0	2	2	0	0	1	0	2	2
5	1	1	2	2	0	0	0	1	1	1
6	0	0	0	1	2	2	2	2	2	2
7	1	0	0	2	0	1	0	0	0	1
8	1	2	1	0	2	1	2	1	1	2
9	1	0	1	0	1	1	1	1	0	0

Table 7. The fitness values of the chromosomes in the initial population of G-ANMI

Chromosomes	fitness value
C[0]	0.211672
C[1]	0.469059
C[2]	0.338877
C[3]	0.227614
C[4]	0.139958
C[5]	0.181013
C[6]	0.216344
C[7]	0.263182
C[8]	0.377376
C[9]	0.166627
average	0.259172

4 Experimental Results

A series of experiments are conducted to evaluate the clustering efficiency and clustering performance of IG-ANMI. They are described below.

4.1 Experiments Design

We aim to evaluate the influence of new initialization method on G-ANMI algorithm. Therefore, the experimental studies are devoted to the comparison between G-ANMI and IG-ANMI. Four real-life datasets obtained from the UCI Machine Learning Repository [11] are used in the experiments, including Zoo, Congressional Votes (Votes for short), Wisconsin Breast Cancer (Breast Cancer for short), and Mushroom. The reason for choosing these four datasets is that they are also used in G-ANMI for evaluation. The information about the data sets is tabulated in Table 8.

Table 8. The information about the four data sets

Data set	Number of objects	Number of Attributes	Number of classes
Zoo	101	16	7
Votes	435	16	2
Breast cancer	699	9	2
Mushroom	8124	22	2

The parameters required by G-ANMI and IG-ANMI are set to be the same as in [7]. In addition, population size has a great effect on the quality of clustering in G-ANMI and IG-ANMI. In our experiments we vary the population size to perform the comparison between G-ANMI and IG-ANMI. For the Zoo, Votes, and Breast Cancer data sets, the population size vary from 50 to 500, for the Mushroom data set, the population size varies from 50 to 200.

All the programs are written in C language and compiled on the Borland C++ version 5.02. All experiments are conducted on a machine with Intel Core2 Duo CPU T7250 @ 2.00GHz, 1.99 GB of RAM, running Microsoft Windows Vista.

4.2 Efficiency Analysis

In our experiments, the running time of algorithms is used as the criteria for efficiency evaluation. Figs. 2-5 plot the running time of G-ANMI and IG-ANMI in seconds on four data sets when population size is increased. It can be seen that IG-ANMI takes less running time than G-ANMI except for on the Zoo data set when population size is 500. It is worth noting that there is a very large difference between G-ANMI and IG-ANMI on the Mushroom data set, which indicates IG-ANMI can save much time when larger data sets are processed. Table 9 shows the concrete values of numbers of iterations and running time of G-ANMI and IG-ANMI on the Mushroom data set. When the population size is set to 200, G-ANMI takes 190998.485 seconds (53 hours) while IG-ANMI only take 1351.594 seconds.

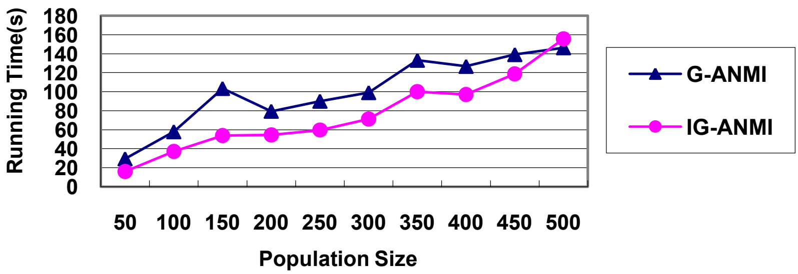


Fig. 2. Running [1]time vs. population size on the Zoo data set

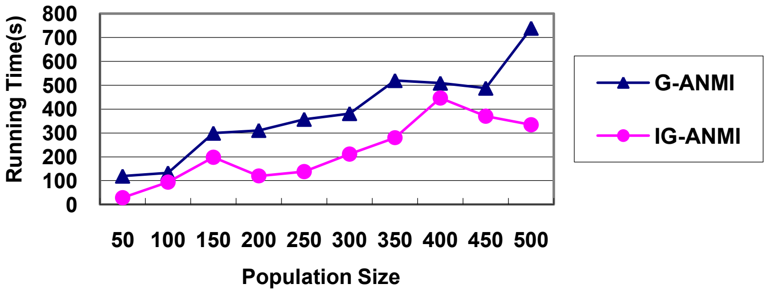


Fig. 3. Running time vs. population size on the Votes data set

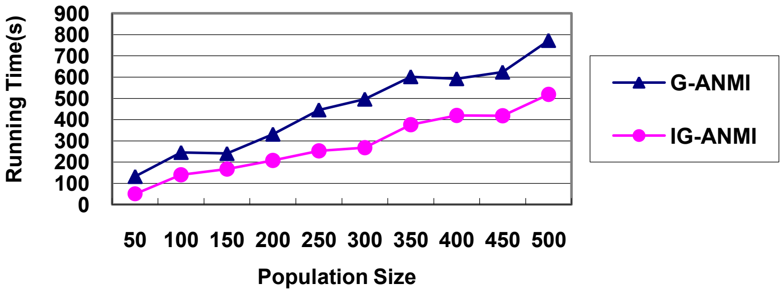


Fig. 4. Running time vs. population size on the Breast Cancer data set

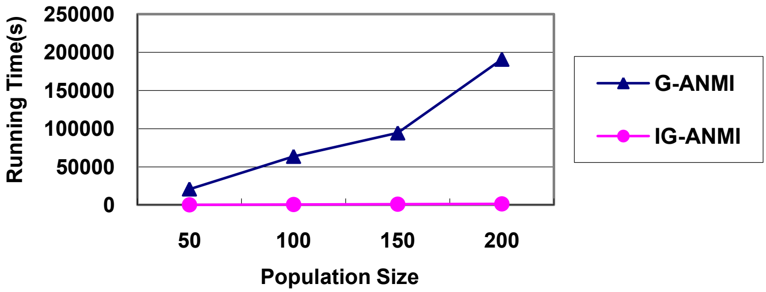


Fig. 5. Running time vs. population size on the Mushroom data set

Table 9. The numbers of iterations and running time of G-ANMI and IG-ANMI on the Mushroom data set

Population Size	Number of iterations		Running time (s)	
	G-ANMI	IG-ANMI	G-ANMI	IG-ANMI
50	10845	100	20759.969	201.25
100	14453	145	63574.047	606.312
150	13944	144	94324.032	880.875
200	17916	158	190998.485	1351.594

Table 10. The clustering accuracies of G-ANMI and IG-ANMI on four data sets

Data set	Algorithms	50	100	150	200	250	300	350	400	450	500	Average
Zoo	G-ANMI	0.832	0.851	0.911	0.881	0.861	0.891	0.931	0.812	0.851	0.921	0.874
	IG-ANMI	0.881	0.911	0.901	0.911	0.921	0.921	0.921	0.931	0.931	0.931	0.916
Votes	G-ANMI	0.874	0.862	0.876	0.864	0.880	0.874	0.869	0.864	0.876	0.874	0.871
	IG-ANMI	0.839	0.860	0.880	0.862	0.867	0.869	0.876	0.876	0.871	0.878	0.868
Breast Cancer	G-ANMI	0.924	0.962	0.960	0.963	0.972	0.975	0.975	0.975	0.977	0.978	0.966
	IG-ANMI	0.936	0.966	0.965	0.969	0.971	0.977	0.975	0.977	0.972	0.978	0.969
Mushroom	G-ANMI	0.547	0.568	0.546	0.538							0.55
	IG-ANMI	0.744	0.902	0.847	0.901							0.849

4.3 Performance Analysis

We use clustering accuracy to evaluate the performance of the IG-ANMI, which is one of the most widely used methods to evaluate the results of clustering algorithms. Given the true class labels and the required number of clusters, k, clustering accuracy is defined as $\frac{\sum_{i=1}^{k} a_i}{n}$, where n is number of objects in the dataset and a_i is the number of objects with the class label that dominates cluster i. A higher value of clustering accuracy indicates a better clustering result. The clustering accuracies of two algorithms on four data sets are summarized in Table 10. From the average accuracies, we can see that IG-ANMI has higher clustering accuracy on the Zoo, Breast Cancer, and Mushroom data sets. One exception is on the Votes data set, the clustering accuracy of G-ANMI is slightly higher than that of IG-ANMI. It is worth noting that IG-ANMI improves clustering accuracy greatly on the Mushroom data set.

5 Conclusions

In this paper, we propose a new initialization method for G-ANMI, namely integrating some equivalence classes of the attributes partitions into the generation of initial partitions. Experimental results on four real-life data sets show that the new method greatly improves the efficiency of G-ANMI, as well as produces higher clustering accuracy, especially on the larger data sets. The new initialization method could be more complicated so that producing better initial chromosomes. In the future work, we will develop other initialization methods to further improve the efficiency of G-ANMI.

Acknowledgments. This work was supported by PRGS under the Grant No. GRS100323, Universiti Malaysia Pahang, Malaysia.

References

1. Guha, S., Rastogi, R., Shim, K.: ROCK: a robust clustering algorithm for categorical attributes. Information Systems 25(5), 345–366 (2000)
2. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery 2(3), 283–304 (1998)
3. Barbara, D., Li, Y., Couto, J.: COOLCAT: an entropy-based algorithm for categorical clustering. In: Proc. of CIKM 2002, pp. 582–589 (2002)
4. He, Z., Xu, X., Deng, S.: k-ANMI: a mutual information based clustering algorithm for categorical data. Information Fusion 9(2), 223–233 (2008)
5. He, Z., Xu, X., Deng, S.: A cluster ensemble method for clustering categorical data. Information Fusion 6(2), 143–151 (2005)
6. Cristofor, D., Simovici, D.: Finding median partitions using information-theoretical-based genetic algorithms. Journal of Universal Computer Science 8(2), 153–172 (2002)

7. Deng, S., He, Z., Xu, X.: G-ANMI: A mutual information based genetic clustering algorithm for categorical data. Knowledge-Based Systems 23, 144–149 (2010)
8. Holland, J.H.: Adaptation in Natural and Artificial Systems. MIT Press (1992)
9. Bai, L., Liang, J.Y., Dang, C.Y.: An initialization method to simultaneously find initial cluster and the number of clusters for clustering categorical data. Knowledge-Based Systems 24, 785–795 (2011)
10. Herawan, T., Deris, M.M., Abawajy, J.H.: A rough set approach for selecting clustering attribute. Knowledge-Based Systems 23, 220–231 (2010)
11. UCI Machine Learning Repository (2011), `http://www.ics.uci.edu_/mlearn/MLRepository.html`

Instance-Ranking: A New Perspective to Consider the Instance Dependency for Classification

Xin Xia, Xiaohu Yang, Shanping Li, and Chao Wu

College of Computer Science and Technology, Zhejiang University
38 Zheda Road, Hangzhou, 310027, China
{xxkidd,yangxh,shan,wuchao}@zju.edu.cn

Abstract. Single-label classification refers to the task to predict an instance to be one unique label in a set of labels. Different from single-label classification, for multi-label classification, one instance is associated with one or more labels in a set of labels simultaneously. Various works have focused on the algorithms for those two types of classification. Since the ranking problem is always coexisting with the classification problem, and traditional researches mainly assume the uniform distribution for the instances, in this paper, we propose a new perspective for the ranking problem. With the assumption that the distribution for the instance is not uniform, different instances have different influences for the distribution, the Instance-Ranking algorithm is presented. With the Instance-Ranking algorithm, the famous K-nearest-neighbors (KNN) algorithm is modified to confirm the validity of our algorithm. Lastly, the Instance-Ranking algorithm is combined with the ML.KNN algorithm for multi-label classification. Experiment with different datasets show that our Instance-Ranking algorithm achieves better performance than the original state-of-art algorithm such as KNN and ML.KNN.

Keywords: Instance Ranking, KNN, ML.KNN, Multi-label Classification.

1 Introduction

Supervised learning problems are the one of the core topics in the machine learning related area. Generally speaking, in traditional supervised learning area, one instance is assigned to an unique single label λ from a set of disjoint labels $\mathcal{L}, |\mathcal{L}| > 1$ [1].However, single-label classification can't satisfy the requirements of many real-world applications. For example, a piece of DNA [2] from Yeast Gene can have the function of transcription and protein synthesis simultaneously; For text categorization [3], one news about the England Riots may have more than one categories, such as the economy, politics and ethics; For music categorization [4], a single song makes a person feel happy and exciting at the same time, while another song makes the person sad and nervous. The above examples show some common aspects, a single instance can belong to more than one label simultaneously. This phenomenon is called multi-label classification [5].

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 112–123, 2013.

More formally, let χ denote the input space and labels. Given the multi-label training dataset $D = \{(X_i, Y_i)\}_m^{i=1}$ where $X_i \in \chi \, and \, Y_i \subseteq \mathcal{L}$. Let $\mathcal{L} = \{1,2,3 \ldots |L|\}$ denote the $|L|$ finite associated labels, the goal of multi-label classification is to learn a hypothesis $h: \chi \rightarrow 2^Y$ which is used to predict the proper label set for a new instance.

Various methods have been proposed to solve the multi-label classification. Generally to say, those methods can be divided into two camps, *problem transformation method* and *algorithm adaptation method* [5]. The *problem transformation method* [6-8] transforms the multi-label classification task into a subset of multiple single-label classification tasks, usually it is based on the on binary relevance (BR) or label powerset (LP). *Algorithm adaptation method* extends specific learning algorithms in order to handle multi-label data directly. Those algorithms include lazy algorithm (ML-KNN [9] and Mr-KNN [10]), AdaBoosting.MH [3], Rank-SVM [11], BP-MLL[12] and decision trees [13].

Ranking problem is always coexisting with the classification problem. For single-label classification, the probability of the instance belong to each label in the associated label set is computed, and the instance is considered to belong to the label with the highest probability. For multi-label classification, after the probability is computed, the ranking threshold is computed; for each label of the instance, if the probability is above the threshold, then this label is added into the proper label set. Especially, for multi-label classification, one particular evaluation metrics called ranking loss is provided to evaluate average fraction of label pairs that are reversely ordered for the instance.

The traditional ranking analysis is built with the assumption the instance is uniform distributed, and each instance affects the labels distribution equally. In this paper, we propose another view to see the ranking and classification problem, which assumes that the instances are not uniformly distributed. Inspired by the PageRank which rank the web pages on the Internet, we rank the instances in the training datasets to generate the ranking scores. The ranking scores are used to describe the distribution for the instances in the datasets. The instance with higher scores will have deeper influence for the proper label sets. To validate the effective and feasibility of our Instance-Ranking algorithm, we modify the K-Nearest-Neighbors (KNN) [1] algorithm with the ranking scores (IR.KNN), and compare the IR.KNN with KNN. With the experiment showing that IR.KNN achieves a slightly better performance than KNN, we go further to apply the Instance-Ranking algorithm to multi-label classification. The ML.KNN algorithm is enhanced with our Instance-Ranking algorithm (IR.MLKNN). Experiments show that our IR.MLKNN achieves better than MLKNN.

The main contribution of this paper is summarized as follows:

- We propose a new perspective to view the instance dependency, and address a feasible instance ranking algorithm.
- We enhance the previous KNN and MLKNN algorithm with our instance ranking algorithm, and experiments show that our enhanced algorithm achieves better than the original ones.

The rest of this paper is organized as follows: the related works about PageRank, KNN and MLKNN are briefly reviewed in Section 2. In Section 3, the detail of

Instance-Ranking Algorithm is proposed. The Instance-Ranking Based KNN (IR.KNN) and the Instance-Ranking Based MLKNN (IR.MLKNN) are addressed in Section 4. In Section 5, the experiments and the results are presented. The final conclusion and future work are provided in Section 6.

2 Related Works

In this section, we briefly review the PageRank algorithm, the KNN and MLKNN algorithm. PageRank algorithm is most related to our Instance-Ranking algorithm, while the KNN and MLKNN algorithm will be enhanced later in the paper.

2.1 PageRank

The PageRank [14, 15] algorithm properly takes advantage of primitive matrix's convergence feature and the assumption of random surfer model perfectly. The basic idea of PageRank is that if page u has a link to page v, then the author of u is implicitly conferring some importance to page v. Therefore, for each page v that page u links to, the Rank of page v will be:

$$Rank_{i+1}(v) = \sum_{u \in B_v} Rank_i(u) / N_u \tag{1}$$

i is the iteration number. B_v is the set of pages that link to page v. N_u is the total number of pages u links to.

According this idea, a $n \times n$ transition matrix P is constructed (n is the number of pages). Each element P_{ij} represents the probability that random surfer move from $Page_i$ to $Page_j$.To Avoid the Rank Sink, the jump probabilities are added to dispatch rank scores of the dangling links and transfer transition matrix P to be a primitive one ($\bar{P}$). However, matrix $\bar{P}$ may not be an irreducible one as zero factors might exist in matrix $\bar{P}$. After dispatching some scores to pages not directly link to, transition matrix $\bar{\bar{P}}$ finally be transferred as an irreducible primitive matrix.

$$\mathrm{P} \rightarrow \bar{P} \rightarrow \bar{\bar{P}} \tag{2}$$

$$\bar{\bar{P}} = \alpha \bar{P} + (1 - \alpha) e e^T / N \tag{3}$$

2.2 KNN and MLKNN

K-nearest neighbor (KNN) [1] algorithm has a long history in the data mining area for single-label classification. Simply to say, KNN finds a group of nearest k instance in the training set, and bases the assignment of a label on the predominance of a particular label in its neighborhood.

ML-KNN [9] considers the labels for a new instance according to the labels of KNN , and then based on statistical information from the neighboring instance, maximum a posteriori (MAP) principle is utilized to determine the label set for the new instance.

3 Instance-Ranking Algorithm

In this section, we propose the details of the Instance-Ranking algorithm, which is an extension of our previous work [17].Unlike to the web pages which have different types of links, the instances in the datasets don't have the inherent link relationship. To rank the instances in the dataset, two jobs are required: The link model for the instances and the ranking algorithm for the instances.

3.1 Instance Ranking Graph

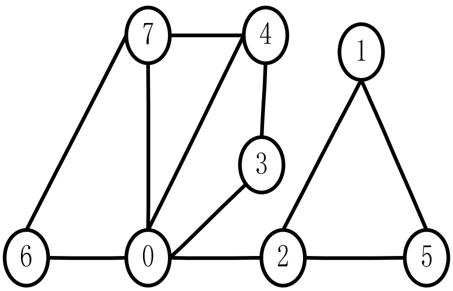
To build the instance ranking graph, we need to map the instances into the graph. For each instance X_i belongs to the dataset D $(X_i \in D)$, there is an unique vertex v_i representing the instance. The edges among the vertexes are defined as : for each instance X_i, find its K-nearest-neighbors in D, denote as $X_{i_1}, X_{i_2}, X_{i_3} \ldots X_{i_k}$, and there is an edge between v_i and v_m $(m \in \{i_1, i_2, i_3 \ldots i_k\})$.

Definition 1: the Instance Ranking Graph is defined as:

$$G = (V, E)$$
$$V = \{v_i | X_i \in D\} \tag{4}$$
$$E = \{(v_i, v_j) | v_i, v_j \in V, v_i \in KNN(v_j) \ or \ v_j \in KNN(v_i), i \neq j\}$$

ID	Input Space	2NN
0	2.5,3,3.5	{6,7}
1	6.5,7,8	{2,5}
2	3.5,4.5,5.5	{0,5}
3	6.5,3.5,2.5	{0,4}
4	4,1.5,2	{0,7}
5	5.5,6.5,7	{1,2}
6	1,3,2	{0,7}
7	2,2,2	{0,6}

(a)



(b)

Fig. 1. (a) an example dataset in the three-dimensional coordinate space; (b) the 2NN Instance Rank Graph with the dataset in (a)

The ***Instance Ranking Graph*** is a fully connected graph without any separate components as for each vertex v_i, as its K-Nearest-Neighbors can be always found, which avoids the existing of dangling vertex. Figure 1 (a) shows one example dataset with 8 instances in the three-dimensional coordinate space, Figure 1(b) shows the Instance Ranking Graph with 2NN. For example, the 2NN of the vertex 0 is vertex 6 and 7, so there are edges between vertex 0 and 6, vertex 0 and 7 in Figure 1(b); for the vertex 6, the 2NN is vertex 0 and 7, as there is already between vertex 0 and 6, only the edge between vertex 6 and 7 is added in Figure 1(b). With the ***Instance Ranking Graph*** built, the definition of the weights between two connected vertexes is required.

In this paper, we use the Euclid distance to describe the distance between two instances v_i and v_j(denoted as $distance(v_i, v_j)$).

Definition 2: the weight matrix W ***of the Instance Ranking Graph is:***

$$W_{ij} = \begin{cases} 0, i = j \; or \; (v_i, v_j) \notin E \\ \frac{1/distance(v_i,v_j)}{\sum_{v_p \in KNN(v_i)} 1/distance(v_i,v_p)}, (v_i, v_j) \in E \end{cases} \tag{5}$$

W is a Normalized matrix with $\sum_{j=1}^{m} W_{ij} = 1$. From the definition of weight matrix W, one can conclude that weight matrix is not a symmetrical matrix.

3.2 The Procedure of Instance Ranking

The random walk model [16] is used to rank the instances in the dataset. The general idea for instance ranking is: Visiting the instance rank graph from anyone of the vertexes v_0 . In anyone of the vertex v, it has the probability α to visit the neighbor vertex of v from the instance ranking graph, or it may also random teleport to any other vertex with the probability $(1 - \alpha)$, where α controls the priority jump to the vertex's neighbor as opposed to teleport to a random vertex in the graph. The whole process can be summarized in formula (6):

$$\boldsymbol{\pi}^{(k)T} = \boldsymbol{\pi}^{(k-1)T}(\alpha \cdot \mathbf{P} + (\mathbf{1} - \alpha) \cdot \mathbf{et}^{T}) \tag{6}$$

where $\boldsymbol{\pi}^{(k)}$ denotes the output ranking scores for the instances after k-th iteration, **P** denotes the adjacent probability matrix, α denotes the random teleport parameter (0<α<1), $\mathbf{t}$ denotes the probability of random walk to anyone of the vertexes. As the adjacent probability matrix **P** satisfies the property the sum of each row equals 1 ($\sum_{j=1}^{m} P_{ij} = 1$) , and the weight matrix W satifies $\sum_{j=1}^{m} W_{ij} = 1$. The formula (6) is rewritten as formula (7):

$$\boldsymbol{\pi}^{(k)T} = \boldsymbol{\pi}^{(k-1)T}(\alpha \cdot \mathbf{W} + (\mathbf{1} - \alpha) \cdot \mathbf{et}^{T}) \tag{7}$$

The whole algorithm of Instance Ranking is in Figure 2. The algorithm receives the parameters dataset D and the random teleport parameter **α**, after iterate formula (10) enough times, the algorithm will converge to a unique ranking scores. For different α, the algorithm will have different results. If α → 0, the iteration times are huge, and if α → 1, the Rank Matrix **RM** become much more volatile, and fluctuates noticeably for even small changes of Rank Matrix **RM**. The final ranking result for the example in Figure 1 (a) is Rank(0)=0.245, Rank(1)=0.067, Rank(2)=0.055, Rank(3)=0.019, Rank(4)=0.028, Rank(5)=0.083, Rank(6)=0.240, Rank(7)=0.264.

4 IR.KNN and IR.MLKNN

To verify the performance of our Instance-Ranking algorithm, we modify the two state-of-art algorithms for single-label and multi-label classification, KNN and

MLKNN. In this section, we present the modified KNN (IR.KNN) and MLKNN (IR.MLKNN) algorithms, and in the next section, the experiments for IR.KNN and IR.MLKNN are addressed.

Algorithm 1:The Instance Ranking Algorithm

Input: Dataset D, the random teleport parameter α
Output: Ranking Scores for each instance in D

Instance-Ranking (D, α)

1. Build the KNN based link graph G from D using (4);
2. Compute the weight matrix **W** using (5)
3. Random choose one of the rows in **W** as the initial ranking scores, $\boldsymbol{\pi}^{(0)} = \mathbf{W}_i^T$
4. Set $\mathbf{t} = (\frac{1}{m}, \frac{1}{m}, \frac{1}{m} \dots \frac{1}{m})^T$.
5. Compute the Rank Matrix $\mathbf{RM} = \boldsymbol{\alpha} \cdot \mathbf{W} + (\mathbf{1} - \boldsymbol{\alpha}) \cdot \mathbf{e}\mathbf{t}^T$
6. Iterate $\boldsymbol{\pi}^{(k)T} = \boldsymbol{\pi}^{(k-1)T} * \mathbf{RM}$ until $\boldsymbol{\pi}^{(k)}$ get convergent.
7. Return the final convergences $\boldsymbol{\pi}^{(\mathbf{final})}$

Fig. 2. The details of the Instance Ranking Algorithm

4.1 IR.KNN

The basic idea of IR.KNN is simple. Suppose current there are |L| different labels, for a new instance X_{new}, found the KNN of X_{new} (denote them as $X_{new_1}, X_{new_2}, \dots X_{new_k}$). For each label $Label_i$, compute the probabilities:

$$P_{label_i} = \sum_{X \in \{X_{new_1}, X_{new_2}, \dots X_{new_k}\} \&\& X \in Label_i} Rank(X) \tag{8}$$

X_{new} is considered to belong to the label with the highest probability. Figure 3 presents the detail of IR.KNN.

4.2 IR.MLKNN

As discussed in Section 2.2, the MLKNN algorithm can be transformed into the problem of computing the prior probability $P(H_b^j)$ and the posterior ty $P\left(E_{\vec{C}_t(j)}^j \middle| H_b^j\right)$. In the MLKNN paper, the prior probability is simply count the total number of instances which belong to the particular label divide by the total number of instances. More formally, the prior probability of MLKNN algorithm is:

$$P(H_1^j) = (s + \sum_{i=1}^{m} y_{x_i}(j))/(s * 2 + m), \tag{9}$$

$$P(H_0^j) = 1 - P(H_1^j), j \in \mathcal{L} \tag{10}$$

where s is a smoothing parameter controlling the strength of uniform prior.

In this paper, we reconsider the prior probability $P(H_b^j)$, and we embed our Instance-Ranking algorithm to enhance the prior probability. More formally, the prior probability of IR.MLKNN is:

$$P\left(H_1^j\right) = \sum_{i=1}^{m} Rank(X_i) * y_{X_i}(j), \tag{11}$$

$$P\left(H_0^j\right) = \sum_{i=1}^{m} Rank(X_i) * (1 - y_{X_i}(j)), j \in \mathcal{L} \tag{12}$$

Algorithm 2:The IR.KNN Algorithm

Input: Training Dataset D ($\mathbf{X_i, L_i}$), the random teleport parameter $\boldsymbol{\alpha}$, the new instance $\mathbf{X_{new}}$

Output: The label of the Instance $\mathbf{X_{new}}$

IR.KNN (D, $\boldsymbol{\alpha}, \mathbf{X_{new}}$)

1 Compute the Rank(X,L) ((X, L) ∈ D) using the Instance Ranking Algorithm

2 $\{(X_{new_1}, L_{new_1}), (X_{new_2}, L_{new_2}) \ldots (X_{new_k}, L_{new_k})\} = KNN(X_{new}, D)$

3 For each Label $label_i$ in the labelset

$$P_{label_i} = \sum_{(X,L)\in\{(X_{new_1},L_{new_1}),(X_{new_2},L_{new_2})\ldots(X_{new_k},L_{new_k})\}\&\&L\in Label_i} Rank(X, L)$$

4 Return the Label with the highest probability.

Fig. 3. The IR.KNN Algorithm

The detail of IR.MLKNN is in Figure 4. We mainly change the way of the Computing of Prior Probability. The array C[p] and $C'[p]$ is used to describe that with p neighbors of the instance belong to the label, whether the instance belongs to the label. If it is, C[p] will add one; else $C'[p]$ will add one.

5 Experiments and Results

In this section, we focus on the experiments on IR.KNN and IR.MLKNN, to verify our Instance-Ranking algorithm. The section includes two subsections: the experiments and results for IR.KNN comparing with KNN, and the experiments andresults for IR.MLKNN comparing with MLKNN. All the experiments are based on Weka [16] and Mulan [5].

5.1 Experiments and Results for IR.KNN

The dataset heart-statlog in the UCI[1] database is used to complete the experiment with IR.KNN. Heart-statlog dataset contains 270 instances, 13 attributes of the type

[1] `http://sourceforge.net/projects/weka/files/datasets/datasets-UCI/datasets-UCI.jar`

numeric, and two class "absent" and "present". We compare IR.KNN and KNN from K=5 to 50 using 10-fold cross-validation with $\alpha = 0.85$ and three evaluation metrics are provided: the correctly classified instances, mean absolute error, and relative absolute error.

Algorithm 3:The IR.MLKNN Algorithm

Input: Training Dataset D ($\mathbf{X_i}, \mathbf{L_i}$), the random teleport parameter $\boldsymbol{\alpha}$, the new instance $\mathbf{X_{new}}$
Output: The label set of the Instance $\mathbf{X_{new}}$

IR.MLKNN (D, $\boldsymbol{\alpha}, \mathbf{X_{new}}$)
--Compute the prior probability $\mathbf{P(H_b^j)}$
1 Compute the Rank(X,L) ($(X, L) \in$ D) using the Instance Ranking Algorithm
For each label $\ell \in \mathcal{L}$
2 $P\left(H_1^j\right) = \sum_{i=1}^{m} Rank(X_i) * y_{X_i}(j)$
3 $P\left(H_0^j\right) = \sum_{i=1}^{m} Rank(X_i) * (1 - y_{X_i}(j))$
-- Posterior Probability $\mathbf{P\left(E_{\vec{C}_t(j)}^j \middle| H_b^j\right)}$
For each label $\ell \in \mathcal{L}$
For each Instance $X_i \in$ D,
4 Identity KNN(X_i).
5 Compute the number of instances which belong to the label ℓ. $\delta = \vec{C}_{X_i}(j) = \sum_{b \in KNN(X_i)} y_b(\ell)$
6 If($y_{X_i}(\ell)$==1) $C[\delta]++$; else $C'[\delta]++$;
For each $j \in \{0,1,2,\dots,K\}$
7 $P\left(E_j^\ell \middle| H_1^\ell\right) = C[j]/\sum_{p=0}^{K} C[p]$; $P\left(E_j^\ell \middle| H_0^\ell\right) = C'[j]/\sum_{p=0}^{K} C'[p]$
--Prediction Period
For each label $\ell \in \mathcal{L}$
8 $\vec{C}_{X_{new}}(\ell) = \sum_{b \in KNN(X_i)} y_b(\ell)$
9 $y_{X_{new}}(\ell) = \arg\max_{b \in \{0,1\}} P(H_b^\ell)\, P\left(E_{\vec{C}_{X_{new}}(\ell)}^\ell \middle| H_b^\ell\right)$
Return $y_{X_{new}}(\ell)$ $\ell \in \mathcal{L}$

Fig. 4. The IR.MLKNN Algorithm

Figure 5 (a)-(c) presents the experiments results for heart-statlog dataset. The correctly classified instances metric has some disturbance. For K =15,40 and 50, the correctly classified instances of KNN is better than IR.KNN. But with the whole processes from K=5 to 50, IR.KNN achieves a slight better prediction than KNN, the average correctly classified instance percentage for IR.KNN is 83.03%, while for KNN is 82.37%. Although IR.KNN and KNN achieve almost the same correctly classified instance percentages, for other evaluation metrics, IR.KNN shows much better performance. The average mean absolute error for IR.KNN is 0.17669, while for KNN is 0.27161, IR.KNN achieves 53% better than KNN for the mean absolute error average; the average relative absolute error for IR.KNN is 35.77136%, while for KNN is 54.99154%, IR.KNN achieves 53.7% better than KNN for the relative absolute error average.

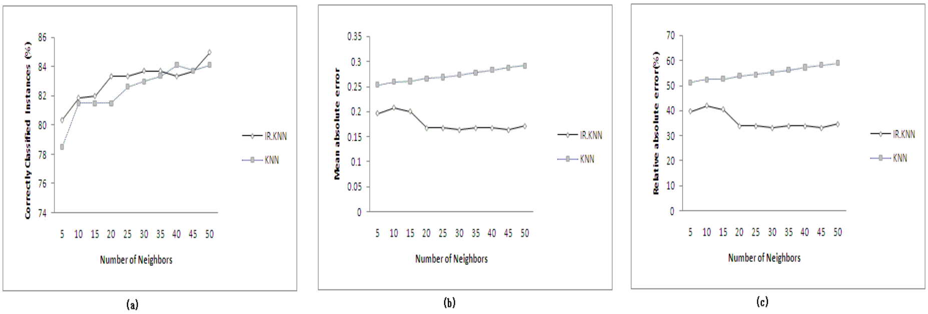


Fig. 5. The experiments results for IR.KNN and KNN for UCI heart-statlog dataset. (a) presents the correctly classified instances; (b) presents the mean absolute error; (c) presents the relative absolute error.

Table 1 presents our comparison of our IR.KNN with SVM, Naive Bayes, and decision tree [1]. We choose K=20 for IR.KNN, and the default parameters for the other algorithms as in Weka [16]. The experiment results show that our IR.KNN achieves better performance than Naive Bayes, and decision tree, and similar performance of SVM.

Table 1. Comparison of IR.KNN with SVM, Naive Bayes and Decision Tree

Evaluation Metrics	IR.KNN	SVM	Naive Bayes	Decision Tree
Accuracy	83.84%	84.07%	83.70%	76.67%
Mean Absolute Error	0.1635	0.1593	0.1835	0.2740
Relative Absolute Error	33.745%	32.25%	37.16%	55.48%

Table 2. The detail Description of the multi-label datasets emotions and CAL500

Name	Domain	Instances	Attributes	Labels	Cardinality	Density	Distinct
emotions	music	593	72	6	1.869	0.311	27
CAL500	music	502	68	174	26.044	0.150	502

5.2 Experiments and Results for IR.MLKNN

To verify the performance of IR.MLKNN, we used two multi-label datasets emotions and CAL500[2]. The description of those two datasets is in table 2. We compare

[2] http://mulan.sourceforge.net/datasets.html

IR.MLKNN with MLKNN from K=5 to 40 using 10-fold cross-validation with $\alpha = 0.85$. About four evaluation metrics [5] are provided: Hamming Loss, Average Precision, Ranking Loss, and One Error. As the evaluation metrics for multi-label classification is a bit different from single-label classification, we would firstly briefly introduce those evaluation metrics.

Results. Figure 6 and Figure 7 present the experiments results for IR.MLKNN with datasets emotions and CAL500 respectively. Table 3 summaries the average performance for those two datasets from the four evaluation metrics. Table 4 presents the results of the two datasets from the view of the maximum difference between IR.MLKNN and MLKNN. Generally speaking, IR.MLKNN achieves better performance than MLKNN consider the results of Figure 6, Figure 7, Table 3 and Table 4.

For Hamming Loss, we notice that for K=5, the IR.MLKNN is higher than the MLKNN for CAL500 in Figure 7(a), this issue happens since the neighbors for the Instance-Ranking is so small, the finally ranking cannot be used to represent the true ranking sequence for the instance of CAL500. Another interesting phenomenon is that the curves for our IR.MLKNN are closer to the curves for MLKNN in Figure 6 than those in Figure 7.

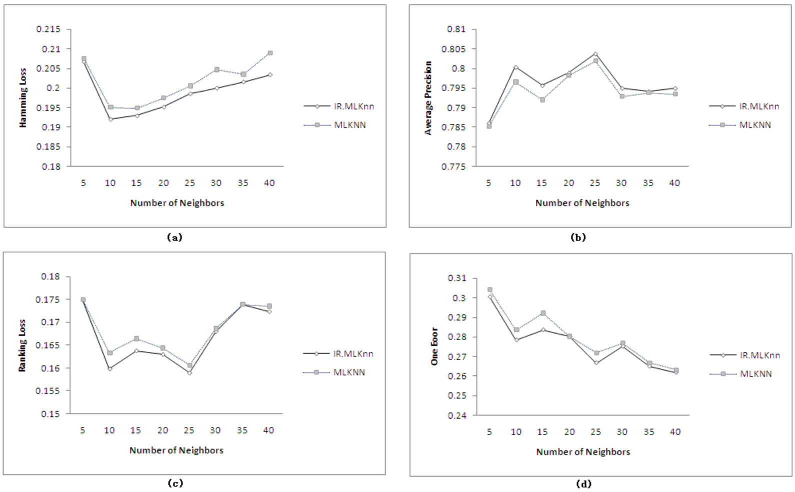


Fig. 6. The experiments results for IR.MLKNN and MLKNN for emotions dataset. (a) Hamming Loss; (b) Average Precision; (c) Ranking Loss; (d) One Error.

This is because the emotions dataset only has 6 labels, while CAL500 dataset has 174 labels. The prior probability enhancement for CAL500 is more significantly than emotions dataset. That is to say, for the datasets with more labels, the performance of our IR.MLKNN is much better.

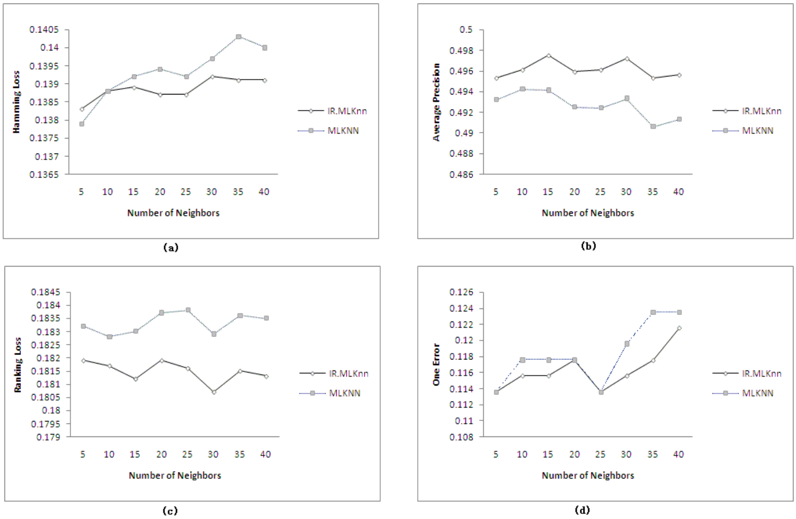


Fig. 7. The experiments results for IR.MLKNN and MLKNN for CAL500 dataset. (a) presents the Hamming Loss; (b) presents the Average Precision; (c) presents the Ranking Loss; (d) presents the One Error.

Table 3. The summarization information for IR.MLKNN and MLKNN for the datasets Emotions and CAL500 with four evaluation metrics from an average view

	Emotions		CAL500	
Evaluation Metrics	IR.MLKNN	MLKNN	IR.MLKNN	MLKNN
Hamming Loss	0.1987	0.2015	0.1388	0.1393
Average Precision	0.7960	0.7942	0.4961	0.4927
One Error	0.1668	0.1682	0.1814	0.1833
Ranking Loss	0.2618	0.2632	0.1215	0.1235

Table 4. The summarization information for IR.MLKNN and MLKNN for the datasets Emotions and CAL500 with four evaluation metrics from the view of maximum the margin

Evaluation Metrics	Emotions			CAL500		
	K	IR.MLKNN	MLKNN	K	IR.MLKNN	MLKNN
Hamming Loss	10	0.1920	0.1951	35	0.1391	0.1403
Average Precision	10	0.8003	0.7965	35	0.4953	0.4906
One Error	10	0.1599	0.1633	35	0.1814	0.1836
Ranking Loss	10	0.2784	0.2835	35	0.1175	0.1235

6 Conclusions

In this paper, we propose a new perspective to consider the instance dependency for classification. A novel algorithm called Instance-Ranking is presented to mining the instance dependency information. With Instance-Ranking algorithm, we modify the

KNN and MLKNN algorithm. The experiments show that the enhanced KNN and MLKNN (IR.KNN and IR.MLKNN) algorithm achieve better than the original algorithms. In the future, we will focus on the improvement of the Instance-Ranking algorithm, and apply it into more data mining applications, not only classification, but also cluster and semi-supervised problems.

Acknowledgment. This work is supported by the Ministry of Industry and Information Technology of China (No. 2010ZX01042-002-003-001).

References

1. Wu, X., Kumar, V.: The top ten algorithms in data mining. Chapman & Hall/CRC (2009)
2. Chen, X., Liu, M., Ward, R.: Protein function assignment through mining cross-species protein-protein interactions. PLoS One 3, e1562 (2008)
3. Schapire, R.E., Singer, Y.: BoosTexter: A boosting-based system for text categorization. Machine Learning 39, 135–168 (2000)
4. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multilabel classification of music into emotions. In: ISMIR (2008)
5. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3, 1–13 (2007)
6. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier Chains for Multi-label Classification. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part II. LNCS, vol. 5782, pp. 254–269. Springer, Heidelberg (2009)
7. Tsoumakas, G., Vlahavas, I.: Random *k*-Labelsets: An Ensemble Method for Multilabel Classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 406–417. Springer, Heidelberg (2007)
8. Zhang, M.L., Zhang, K.: Multi-label learning by exploiting label dependency. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 999–1008 (2010)
9. Zhang, M.L., Zhou, Z.H.: ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40, 2038–2048 (2007)
10. Lin, X., Chen, X.: Mr. KNN: soft relevance for multi-label classification. In: Proc. of the 19th ACM CIKM, pp. 349–358 (1999)
11. Weston, J.: A Kernel Method for Multi-Labelled Classification. Advances in Neural Information Processing Systems 14, 681–687 (2002)
12. Zhang, M.L., Zhou, Z.H.: Multilabel neural networks with applications to functional genomics and text categorization. IEEE Transactions on Knowledge and Data Engineering, 1338–1351 (2006)
13. Clare, A., King, R.D.: Knowledge Discovery in Multi-label Phenotype Data. In: Siebes, A., De Raedt, L. (eds.) PKDD 2001. LNCS (LNAI), vol. 2168, pp. 42–53. Springer, Heidelberg (2001)
14. Langville, A.N., Meyer, C.D.: Google page rank and beyond. Princeton Univ. Pr. (2006)
15. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web (1999)
16. Bouckaert, R.R., Frank, E., Hall, M.A., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: WEKA–experiences with a java opensource project. Journal of Machine Learning Research 11, 2533–2541 (2010)
17. Xia, X., Yang, X., Li, S., Wu, C., Zhou, L.: RW.KNN: A proposed random walk knn algorithm for multi-label classification. In: Proceedings of the 4th Workshop on Workshop for Ph. D. Students in Information & Knowledge Management, pp. 87–90. ACM (2011)

Triangular Kernel Nearest-Neighbor-Based Clustering Algorithm for Discovering True Clusters

Aina Musdholifah[1,2] and Siti Zaiton Mohd Hashim[1]

[1] Soft Computing Research Group, Universiti Teknologi Malaysia, Malaysia
[2] Department of Computer Science and Electronics, Universitas Gadjah Mada, Indonesia
aina_m@ugm.ac.id, sitizaiton@utm.my

Abstract. Clustering is a powerful exploratory technique for extracting the knowledge of given data. Several clustering techniques that have been proposed require predetermined number of clusters. However, the triangular kernel-nearest neighbor-based clustering (TKNN) has been proven able to determine the number and member of clusters automatically. TKNN provides good solutions for clustering non-spherical and high-dimensional data without prior knowledge of data labels. On the other hand, there is no definite measure to evaluate the accuracy of the clustering result. In order to evaluate the performance of the proposed TKNN clustering algorithm, we utilized various benchmark classification datasets. Thus, TKNN is proposed for discovering true clusters with arbitrary shape, size and density contained in the datasets. The experimental results on benched-mark datasets showed the effectiveness of our technique. Our proposed TKNN achieved more accurate clustering results and required less time processing compared with k-means, ILGC, DBSCAN and KFCM.

Keywords: cluster, classification, triangular, kernel nearest neighbor.

1 Introduction

Clustering is a powerful exploratory technique for the extracting knowledge of given data. It is a process of grouping unlabelled datasets based on the similarity and dissimilarity pair of data within the dataset. Recently, several clustering techniques have been introduced and proposed such as K-means [1-3], hierarchical clustering [4], nearest neighbor-based approach [5-7], DBSCAN-based methods [8], Support Vector Machine-based approach (SVM) [9], kernel fuzzy c-means (KFCM) [10-11] and DENCLUE [12-13]. Since they lack of valid statistical evaluation methods, the results of the hierarchical cluster analysis are subject to interpretation by the investigator. K-means technique is probably the most popular and a simple solution for clustering. However, the problem with this techniques is determining the proper number of clusters and potential to being trapped in local optimal [14].

Density based clustering algorithms such as SNN [5-7], DBSCAN [15] and DENCLUE [12-13] are used to determine the number of clusters automatically by the density of data points in a region. A cluster in these algorithms is a dense region of

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 124–140, 2013.

object that is surrounded by a region of low density. Density-based uses local cluster criterion in which the clusters are defined as region in data space whose objects are dense, and clusters are separated from one another with low-density region [15].

For instance, SNN technique [5], it scarify the similarity matrix by keeping only the k most similar neighbors, then constructs a correlative shared nearest neighbor graph and determine the core points based on the SNN density of each point. Except that, it also discards all noise points and assigns all non-noise, non-core points to the clusters. As for the algorithm complexity, because of the nested loops for outliers deletion stage, it would require $O(M^2)$ distance computations and $O(M^2/\text{blocksize})$ data access. Furthermore, there exists another important problem especially on data with arbitrary cluster: all or some of these core points may possibly belong to identical cluster and hence may present an empty cluster.

DBSCAN [15] can be used to discover clusters with arbitrary shapes. Each point of a cluster in the neighborhood of a given radius has to contain at least a minimum number of points, i.e. the density in the neighborhood has to exceed some threshold. The shape of a neighborhood is defined by the choice of a distance function for two points p and q, denoted by $dist(p,q)$. For example, in two dimensional data, the shape of the neighborhood is circle when Euclidean distance is used or rectangular when Manhattan distance is used. However, using density-based clustering (DBSCAN) for discovering clusters of arbitrary shapes have difficulties in discovering clusters with different densities. In addition, DBSCAN requires at least two parameters. DENCLUE (DENsity CLUstEring) improves DBSCAN through the use of kernel function as the influence function to express the contribution of each point to the overall density function. However, DENCLUE is computationally expensive than DBSCAN. In addition, DENCLUE is relatively resistant to noise [14].

In recent machine learning field, the trend was using the kernel method to construct a nonlinear [11]. Kernel fuzzy c-means algorithm (KFCM) [10] implements a new kernel-induced metric in the data space to change the original Euclidean norm metric in FCM and the clustered prototypes still lie in the data space so that the clustering results can be reformulated and interpreted in the original space. KFCM is robust to noise, outliers and incomplete data, and also tolerates unequal sized clusters [11]. Nonetheless, KFCM algorithm still has some weaknesses, such as the sensitivity to the initialization the number of clusters and cluster center.

Triangular Kernel Nearest Neighbor-based clustering (TKNN) proposed by [16] is a density based clustering algorithms which only requires one parameter; number of neighbors. TKNN provides a good solution for clustering non-spherical and high-dimensional data such as spatial-temporal data without the knowledge of data labels, as detail in [16]. However, there is no definite measure to evaluate the accuracy of the clustering result. To evaluate the performance of the proposed TKNN clustering algorithm, we utilized various benchmark datasets such as classification datasets UCI [17]. Thus, TKNN was applied to generate class labels for discovering true clusters of such datasets.

Two parameters were used in this study to analyze the performance of the proposed approach: accuracy of clustering result and time consumption. In addition, the proposed TKNN was compared against basic k-means, kernel fuzzy c-means (KFCM) and two other density-based clustering algorithms: ILGC [18] and DBSCAN [15].

The rest of the paper is organized as follow: Section 2 introduces the triangular kernel nearest neighbor based clustering, TKNN; Section 3 presents the experimental results; and finally Section 4 is the conclusion.

2 Triangular Kernel Nearest Neighbor-Based Clustering (TKNN)

TKNN is referred as a kernel nearest neighbor based clustering using triangular kernel function [16]. Kernel nearest neighbor based clustering is a density-based clustering algorithm [19]. The algorithm is able to tackle the difficulties of high dimensional data and cluster of very different densities. In order to determine the density of objects, it uses a nonparametric density estimation procedure. It combines k-nearest neighbor (KNN) and kernel density estimation, i.e. KNN density estimation is extended and combined with kernel function.

TKNN is an improvement from ILGC [18], another kernel nearest neighbor based clustering using Gaussian kernel function. Triangular kernel function is chosen to replace Gaussian kernel function and enhance performance of clustering [16] because it is less time consuming. Table 1 shows the triangular and Gaussian kernel functions for clustering, where Kt_{ω}, Kg_{ω} and $dist(x, x_i)$ refer to the triangular kernel density function of the cluster ω, the Gaussian kernel density function of the cluster ω and the distance between object x and x_i, respectively.

TKNN assigns each object x_i into only one cluster, where if the triangular kernel nearest neighbor densities function in cluster ω_i is maximized, and then the target data is considered belong to the cluster ω_i. The detail of TKNN algorithm can be found in [16]. However, the summary of TKNN algorithm is represented in Fig. 1.

Table 1. Kernel functions for clustering

Type	Function
Triangular	$Kt_{\omega}(x)=\frac{1}{n}\sum_{i=1}^{n}\left(1-dist(x,x_i)\right)$
Gaussian	$Kg_{\omega}(x)=\frac{1}{n}\sum_{i=1}^{n}e^{1/2dist(x,x_i)}$

TKNN Algorithm (Dataset, k neighbors) :

(i) For each data x_i calculate $dist(x_i,x_i)$ (i,j = 1,…,n ; and $i \neq j$)

(ii) Repeat

For each data x_i (i = 1,…,n) with k neighbors is assigned to cluster ω as:

$$Clust_{\omega} = Max\left(Kt_{\omega}(x_i)\right)$$

Re-index the clusters.

(iii) Until no changes in the data structure or iterations have converged.

Fig. 1. TKNN algorithm [16]

3 Experimental Results

3.1 Data Description

In this study, an experimental analysis was devised to assess the performance of proposed TKNN algorithm in discovering true clusters, both in terms of accuracy and efficiency. TKNN algorithm was applied to generate class labels for discovering true clusters of various benchmark datasets for analyzing the accuracy of proposed algorithm. In this work, we adjusted three schemes of experiment which are differentiated by the characteristics of the datasets as given in Table 2. It includes an artificial spherical dataset, shape sets provided by [20] and classification datasets from UCI Machine Learning Repository [17].

For the first experiment, the proposed TKNN algorithm was applied on an artificial spherical or globular dataset. The spherical dataset was used to evaluate the performance of TKNN algorithm for clustering on the dataset which all classes were convex object. An object is convex in Euclidean shape if every pair of points within the object; the considered object on the straight line segment that joins them, is also within the object. Therefore, the objective of first experiment was to evaluate the performance of the proposed algorithm on spherical dataset which was convex, centered, well separated and compact.

An artificial spherical dataset was generated through the use of a random number generator. This dataset had 200 instances with two attributes. All instances were grouped in 3 classes. The data distribution of each class was generated using various means $\mu_1 = [0,10]$, $\mu_2 = [20,10]$, and $\mu_3 = [10,20]$; variance $\sigma_1 = [1.7,1.7], \sigma_1 = [0.7,0.7]$, and $\sigma_1 = [1.0,1.0]$ and number of data $n_1 = 120$, $n_2 = 60$, $n_3 = 20$. The dataset distribution is shown in Fig. 2.

In the second experiment, eight synthetic sample datasets provided by [20] and depicted in Fig. 3 were utilized. Some datasets contained true clusters with soft boundary; the distance between two points of different true clusters was closer than such point of same true cluster. Table 3 summaries the characteristic of the true clusters within the datasets. Different colored symbols were used to demonstrate the different classes. From the Fig. 3, it is visualized that all datasets used had different shapes and number of classes (range from 2 to 31). The eight datasets with different shape were utilized in this experiment in order to evaluate the capability of the proposed TKNN algorithm on clustering dataset with arbitrary shape.

Table 2. Three distinct schemes of experiment

Experiment	Characteristics of dataset
1	Spherical and globular dataset
2	Dataset contained classes with arbitrary shapes
3	Real dataset restrained groups with arbitrary shapes

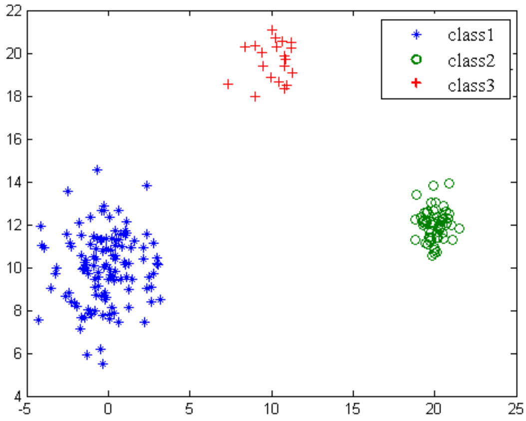


Fig. 2. Artificial spherical dataset

Table 3. Characteristics of eight synthetic sample datasets

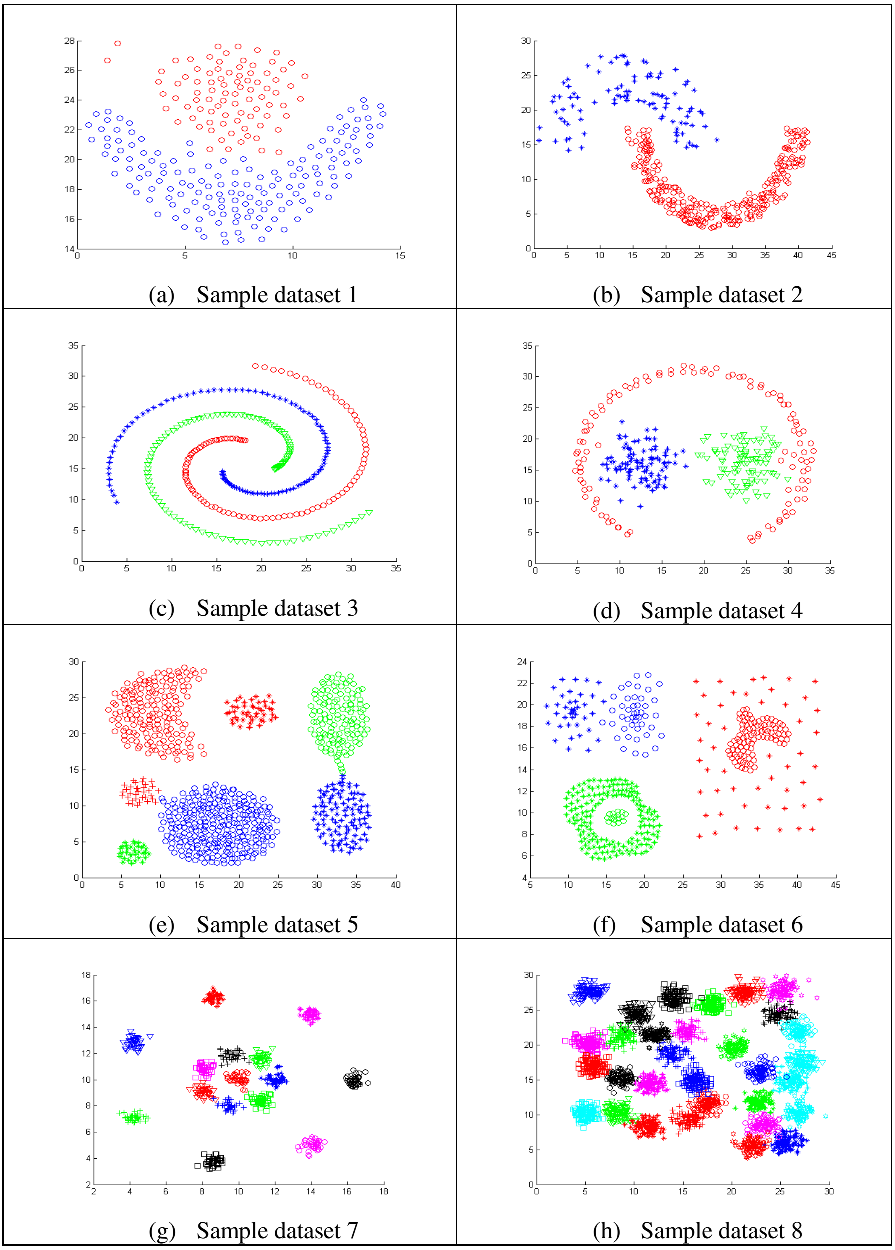
Sample dataset	Number of true clusters	Number of points	Boundary		Shape type	
			Soft	Hard	Convex	Non-convex
1	2	240	√	-	√	√
2	2	373	-	√	-	√
3	3	312	-	√	-	√
4	3	300	√	-	√	√
5	7	788	√	√	√	√
6	6	399	√	√	√	√
7	15	600	√	√	√	-
8	31	3100	√	√	√	-

In sample dataset 1 [21], there were two classes which one of them was non-convex shape and visualized by blue circle symbol on Fig. 3(a). The motivation of utilizing sample dataset 1 was to analyze the capability TKNN algorithm on the dataset that contained different shapes and soft boundary between two clusters.

Sample dataset 2 [22] presented in Fig. 3(b) formed two half-rings and contained 373 two-dimensional patterns (upper class had 97 points and lower class had 276 points). In addition, both of them were non-convex shape. Thus, it was interesting to evaluate the performance of TKNN on datasets, where all the classes within the dataset were non-convex.

Sample dataset 3 [23] composed three classes with spiral shape as shown in Fig. 3(c). A spiral can be defined as a curve which derives from a central point, which gets farther away as it resolves around the point. However, all points in this dataset were not classified based on the distance between the center points and the considered points.

Fig 3(d) visualizes sample dataset 4 [23]. It had two circle classes inside a circular class. Each class within the sample dataset 4 contained 100 data points. In addition, the points in both circles were generated by using Gaussian distribution. This sample could be observed by using two Gaussian classes with adding some Gaussian noise which tended to connect the classes together. However, it was interesting to analyze whether TKNN could discover three clusters as similar as natural classes or TKNN considers the circular class as noise data.



(a) Sample dataset 1 (b) Sample dataset 2

(c) Sample dataset 3 (d) Sample dataset 4

(e) Sample dataset 5 (f) Sample dataset 6

(g) Sample dataset 7 (h) Sample dataset 8

Fig. 3. Shape of eight synthetic sample datasets

The rest of sample datasets were more complex than previous sample datasets in term of number of classes and shapes of classes contained in the datasets. For instance, sample dataset 5 [24] was two-dimensional synthetic dataset that consist of seven perceptually distinct classes. Fig. 3(e) shows a link between two globular classes that is belongs to one of the two globular classes. Furthermore, this link makes the boundary between two clusters becomes soft.

In sample dataset 6 [25], there were five classes of different shape and size with additional noise as shown in Fig. 3(f). Within sample dataset 7, a ring [26] contained 3100 random patterns generated by 15 similar 2D Gaussian distributions. Whereas, sample dataset 8 [26] was largest dataset among dataset used in second experiment. It consists of 31 randomly placed 2D Gaussian classes of 100 patterns each. The objective of using sample dataset 8 is for assessing the performance of proposed algorithm on clustering data with many clusters.

For the third experiment, four classification datasets from [17] were selected with numerical real-value attributes, namely *Iris*, *Wine*, *Glass*, and *Ecoli*. Iris contained the measurements on different iris plants. Wine records results of chemical analysis of Italian wines were obtained from three different cultivars. The glass dataset contained the value of chemical components of glasses. Meanwhile, in the Ecoli dataset, each Ecoli instance was described by using the value of Escherichia Coli bacterium that was taken from different analysis techniques.

All the respective benchmark datasets contained labels or classes. The number of classes next called true clusters with ranges from 3 to 8, the number of instances ranges from 150 to 336, and the number of dimensions ranges from 4 to 8. In addition, most aforementioned datasets were not spherical and of high-dimensional data. Table 4 summarizes the main characteristics of the datasets and also information of the UCI classification datasets.

3.2 Result Analysis

TKNN was applied to generate the class labels for finding the true clusters of various classification datasets. All the aforementioned datasets had labels. We viewed the labels of the datasets as the objective knowledge on the structure of the datasets. Thus, the goal was to cluster the data without the knowledge of the labels and measure how well the clustering captures the true labels. We utilized the F-measure [27] which is one of the most commonly used external validity criteria to evaluate the performance of classification model [28]. The accuracy in F–measure is defined as the combination of both Information Retrieval notions' *Precision* and *Recall.*

Table 4. Characteristics of UCI classification dataset

Datasets	Instances	attributes	classes
Iris	150	4	3
Wine	178	13	3
Glass	214	9	6
Ecoli	336	7	8

Consider if, a dataset D has a collection of true class is $L = \{L_1, L_2, \ldots, L_c\}$ and the collection of resulted cluster is $\omega = \{\omega_1, \omega_2, \ldots, \omega_{nc}\}$. The precision of cluster ω_i with respect to class L_j is the fraction of cluster ω_i that consists of objects in class L_j that are defined as:

$$P_{ij} = \frac{|\omega_i \cap L_j|}{|\omega_i|} \tag{1}$$

Recall of cluster ω_i with respect to class L_j is the extent to which a cluster ω_i contains all objects in class L_j cluster and is defined as:

$$R_{ij} = \frac{|\omega_i \cap L_j|}{|L_j|} \tag{2}$$

Using the local values of precision and recall, the overall precision and recall are computed as:

$$P = \frac{1}{c} \sum_{j=1}^{c} \max_{i \in [1..nc]} P_{ij} \qquad\qquad R = \frac{1}{c} \sum_{j=1}^{c} \max_{i \in [1..nc]} R_{ij} \tag{3}$$

Both values; P and R are combined to measure the quality of clustering result w.r.t L by means of a single value. The overall F-measure is in the range [0,1] and computed as:

$$F = \frac{2PR}{P + R} \tag{4}$$

Furthermore, in order to analyze the groups founded by TKNN, we required visualizing the result. Regarding the number of attributes of the datasets, it was difficult to visualize the clustering result with high dimensional data. Thus, we need to reduce the dimensionality of the data without destroying the meaning. Principal component analysis (PCA) [29] provided solutions to reduce the complex data set to a lower dimension which could reveal the hidden, simplified structure that lay beneath it. The main objective of using PCA was to reduce dimensionality of the data with minimum of loss of information, by retaining as much variation in the original dataset [29] as possible.

In the first experiment, TKNN produces clusters which were equal to true clusters (classes) and 100% accurate. The clustering result is shown in Fig. 4. It was observed that each cluster was well separated while the distance between points in a cluster to data in another cluster was large. In addition, each cluster was mostly compact and well defined cluster, since the distance of points within same cluster was very close. According to the experimental results, it had been proven that TKNN has the ability to discover true clusters on spherical dataset.

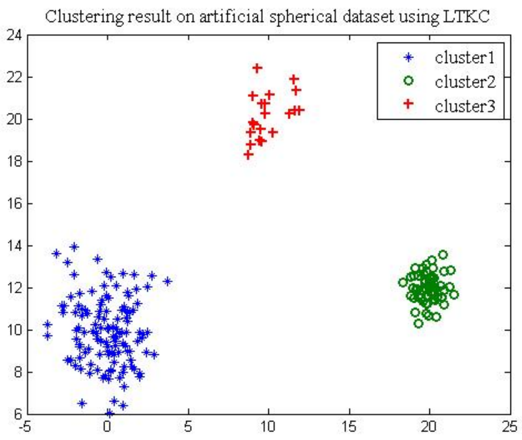


Fig. 4. Clustering on artificial spherical dataset using TKNN

In the second and third experiment, the proposed TKNN algorithm was compared to related the algorithms; ILGC [18], DBSCAN [15], KFCM [10] and basic k-means. A quantitative measure of classification accuracy i.e. F-measure was calculated to assess the performance of all algorithms. In addition, a visual assessment was also utilized for accuracy of inspection.

The proposed TKNN algorithm only required one parameter; *k* number of neighbors. In this experiment, TKNN was executed once for each *k* from 2 to *n*-1, where *n* is the number of objects. For each clustering results, the F-measure was calculated, and finally, the clustering with maximum F-measure value was chosen as a set of "true clusters. The setting approach of TKNN for choosing *k* parameter was also used to execute ILGC algorithm. Meanwhile, applying DBSCAN algorithm required two parameters: *MinPts* minimum number of neighbor points and *Pts* maximum distance value or radius. However, we eliminated the parameter *MinPts* to 4 for all datasets as suggested by [15]. In addition, to give K-means some advantages, the parameter number of clusters was set similar to the number of classes and maximum iteration equal to 100. Meanwhile, for KFCM algorithm, this experiment used Gaussian kernel function, 100 for maximum iteration, 10^{-5} for minimum convergence rate, and number of classes as number of clusters.

From the experimental result of sample dataset 1, TKNN achieved the highest accuracy, 99.17%. In addition, only two points were assigned to incorrect cluster due to their Euclidean distance values. Both points had relatively lower similarity values to other points within upper cluster than those points closer to the lower cluster. The result of sample dataset 1 demonstrated the ability of the TKNN approach to solve the "two clusters touching at a neck" problem with a medium-sized neighborhood (k = 10 as shown in Table 6). There was no distinct point pair distance inconsistency at the neck, so that the triangular kernel density function would be unable to break the data at the neck point.

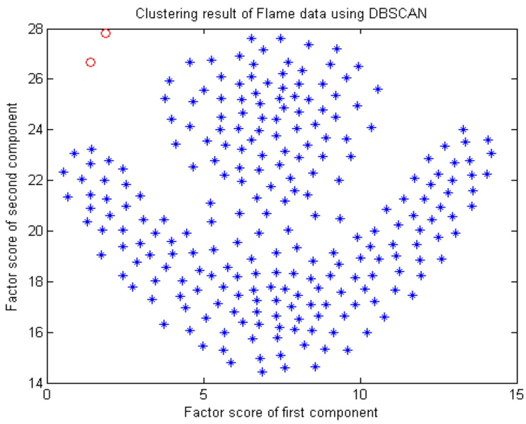


Fig. 5. Clustering on sample dataset 1 produced by DBSCAN

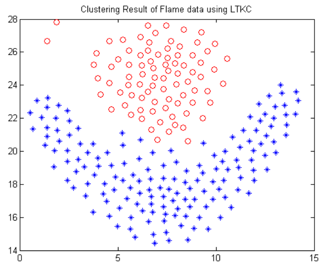
In addition, Fig. 5 shows the weakness of using DBSCAN for clustering the complex dataset, involving cluster with different densities and soft boundary. DBSCAN had difficulty to discover two groups instead of two natural clusters, but found one cluster and one group containing two noise points. However, DBSCAN identified the two points as noise instead of core or border points because their densities were very low.

Again, the performance of TKNN was evaluated for data with non-convex shape classes. In sample dataset 2, TKNN was proven capable to achieve two non-convex clusters as well as true class with 100% accuracy. However, the visual and quantitative analysis of experimental results on sample dataset 1 and 2 indicated that TKNN outperformed ILGC, DBSCAN, KFCM and k-means for clustering dataset that contained non-convex shape class.

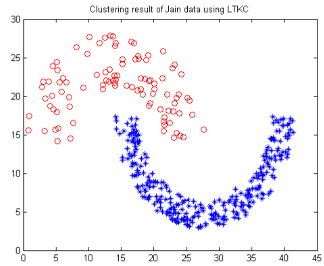
Next, sample dataset 3 with three spiral classes was considered. The privileged clustering result was achieved by using TKNN, ILGC and DBSCAN. Three density based clustering algorithms discovered correctly three spiral clusters, as shown in Fig. 6(c), by using neighboring concepts and inter-point Euclidean distance between the points within a cluster. The sample dataset 3 showed the effects of variation of neighborhood size quite clearly with respect to the continuity concept.

In this study, we also conducted experiments on the sample dataset 4 that contained three classes; a circular class with an opening near the bottom and two Gaussian distributed classes inside. All algorithms could not find the three clusters correctly. Again, the proposed TKNN algorithm gave a more satisfactory result, as shown in Table 3. However, some inter-cluster points were miss-clustered. Our method was successful in assigning closer to these points (and hence essentially detecting them as outliers). As a result, they had relatively lower similarity values to other points within a cluster than those points closer to the Gaussian centers or on the incomplete circle.

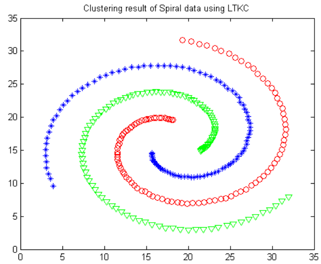
The clustering result for sample dataset 5 produced by TKNN algorithm was imperfect as shown in Fig. 6(e). In fact, this dataset contained seven classes but TKNN discovered five clusters only. Some points formed "narrow bridge" between two classes as these two classes were clustered into one cluster.



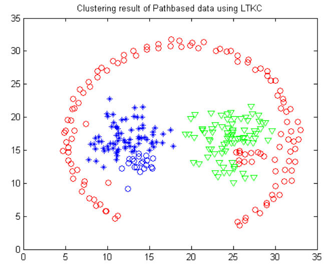
(a) Sample dataset 1



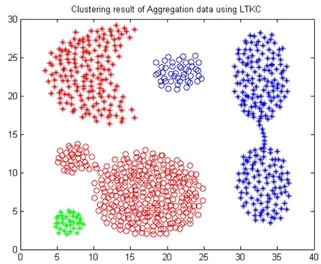
(b) Sample dataset 2



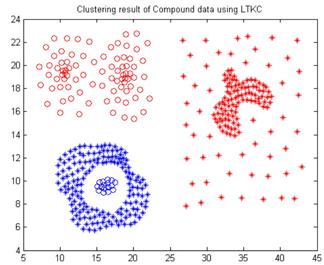
(c) Sample dataset 3



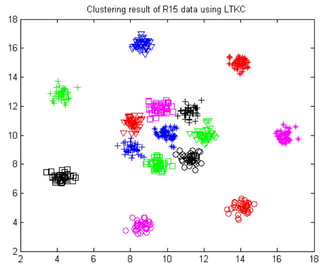
(d) Sample dataset 4



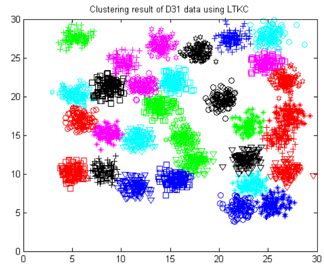
(e) Sample dataset 5



(f) Sample dataset 6



(g) Sample dataset 7



(h) Sample dataset 8

Fig. 6. Clustering results of second experiment produced by TKNN

Table 5. Accuracy of second experiment

Dataset	TKNN	ILGC	DBSCAN	k-means	KFCM
Sample1	99.17	64.58	64.58	84.17 $\pm$ 1.48	78.92% ± 9.82%
Sample2	100	100	73.99	77.48 $\pm$ 0.57	76.78% ± 0.96%
Sample3	100	100	100	34.29 $\pm$ 0.49	33.39% ± 0.62%
Sample4	87.00	73.33	80.67	63.33 $\pm$ 2.51	60.09% ± 16.65%
Sample5	78.43	78.43	82.23	32.87 $\pm$ 12.2	67.42% ± 15.25%
Sample6	87.22	87.22	96.99	65.67 $\pm$ 6.51	60.78% ± 4.48%
Sample7	99.67	99.67	53.33	53.33 $\pm$ 0.63	35.02% ± 1.23%
Sample8	97.55	97.52	62.54	86.97 $\pm$ 0.27	37.85% ± 1.57%

Table 6. Parameter of all algorithms used to produce true clusters on second experiment

Dataset	*TKNN*	*ILGC*	*DBSCAN*		*k-means*		*KFCM*	
	k	k	MinPts	Pts	max_iter	nc	max_iter	nc
Sample1	10	9	4	0.985	100	2	100	2
Sample2	22	17	7	2.446	100	2	100	2
Sample3	7	6	4	1.843	100	3	100	3
Sample4	20	17	4	1.839	100	3	100	3
Sample5	16	10	4	1.208	100	7	100	7
Sample6	15	12	4	1.393	100	6	100	6
Sample7	38	25	4	0.635	100	15	100	15
Sample8	71	46	4	0.535	100	31	100	31

In sample dataset 7, TKNN algorithm produced 15 circle clusters with 99.67% accuracy. It means only two points were not clustered correctly because the points were closer to the clusters than to the true class. Remarkably, DBSCAN, k-means and KFCM algorithms were not able to discover the original cluster structure. Moreover, the accuracy of clustering result produced by DBSCAN and k-means was less than 55%.

Again, TKNN algorithm outperformed ILGC, DBSCAN, k-means and KFCM in the largest sample dataset and number of clusters at most. 31 clusters were found by TKNN with 97.55 % accuracy. 76 points are miss-clustered since the Euclidean distance values between the points intra-cluster were lower than inter-cluster.

In addition, from the experimental results on eight sample datasets, the k-means algorithm was unable to identify the natural clusters. K-means had difficulties to discover clusters with arbitrary shape, such as non-convex shapes and spiral shapes. The k-means was suitable for clustering dataset that contained circle or globular class since it is partitioning clustering algorithm that represents cluster by gravity of center. Cluster center for k-means algorithm is a meaningful point to be associated with all members of clusters, and then using Euclidean distance for measuring distance between the cluster center and all points. In addition, Euclidean distance measure is suitable for classes with circle or globular shapes.

For another kernel-based clustering algorithm, KFCM was unable to discover the natural clusters. KFCM had difficulties to discover clusters with arbitrary shape and spiral shape. In addition, KFCM had least accuracy rate than other when finding clusters that contained points connecting as a line, such as sample dataset 7 and 8.

Otherwise, from the experimental results, it can be concluded that density based clustering algorithm such as TKNN, ILGC and DBSCAN are suitable for discovering true clusters as well as natural class with arbitrary shapes on datasets that contain points that are close within one class and or well separated. In addition, TKNN algorithms that utilized the Euclidean distance measure had accuracy close to 100% on the dataset which contained class with circle shapes, because Euclidean distance is suitable for objects with circle shapes.

In conclusion, from second experiment, the promising performance of TKNN algorithms demonstrated well in some difficult datasets with arbitrary shapes but there existed some other situations when these algorithms did not perform well. One such condition was when there were some points in the datasets which formed “narrow bridge” between clusters.

In third experiment, TKNN and other comparative algorithms are performed for discovering true clusters on four UCI data. Fig. 7 presents visual analysis of comparison between clustering results produced by TKNN and distribution of true clusters. Based on the four classification datasets from [17], we compared the accuracy and time consumption of the proposed TKNN algorithm with k-means, KFCM, DBSCAN and also ILGC algorithms. Table 7 presents the experimental results summary regarding the F-measure. On *glass* dataset, all algorithms achieved clusters with least accuracy rate than other data, because originally all objects in each class of this dataset were not compact and not well separated clusters as shown in Fig. 7 c). However, TKNN produced best clustering on all data compared to TKNN, ILGC, k-means and KFCM.

For evaluating the efficiency of TKNN and the comparing algorithms, time processing of each algorithm in clustering the benchmark datasets was measured. Fig. 8 shows comparison between the time processing (in millisecond) required by TKNN and k-means to cluster the benchmark datasets. In the figure, it can be observed that compared to k-means, the reduction execution time required by TKNN was from 0.02 to 0.14 millisecond. The slowness of k-means was essentially due to the computation time required to calculate the distance between each data and each cluster centre, and to update all cluster centers.

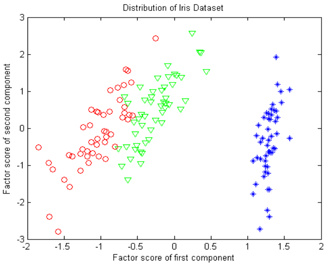
Based on the all experiments conducted, it has been proven that TKNN algorithm has the best clustering due to accuracy rate on discovering true clusters without the aids of labels. In almost all datasets, TKNN has shown improvements over k-means, KFCM, ILGC and DBSCAN.

Table 7. Clustering quality results (F-measure) on UCI data

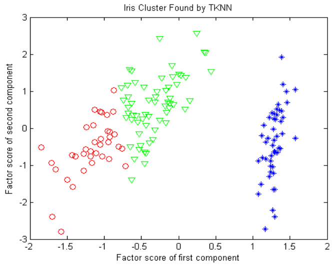
Data	TKNN	ILGC	DBSCAN	k-means*	KFCM
Iris	90.02%	86.03%	80%	90% ± 0.02%	78.78% ± 11.43%
Wine	95.76%	78.74%	69.48%	95.18% ± 1.44%	79.89% ± 6.05%
Glass	66.46%	32.45%	42.15%	58.14% ± 0.08%	52.67% ± 2.61%
Ecoli	92.97%	49.93%	46.91%	59.78% ± 1.02%	50.71% ± 2,92%

Visualization of data distribution using PCA	Visualization of clustering result by using TKNN

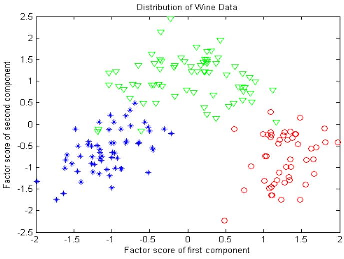
a1) True Clusters on *iris*



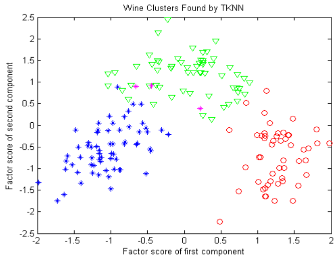
a2) Iris clusters found by using TKNN.



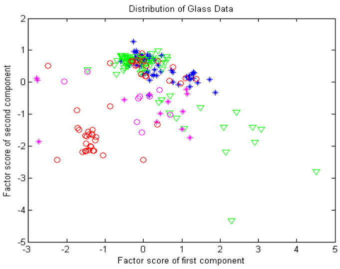
b1) True Clusters on *wine* dataset.



b2) Wine clusters found by using TKNN.

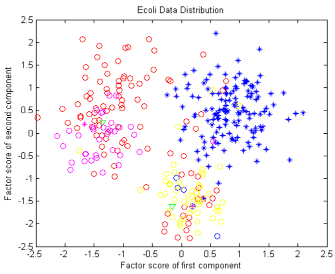


c1) True Clusters on *glass* dataset.



c2) Glass clusters found by using TKNN.

d1) True Clusters on *ecoli* dataset.



d2) Ecoli clusters found by using TKNN.

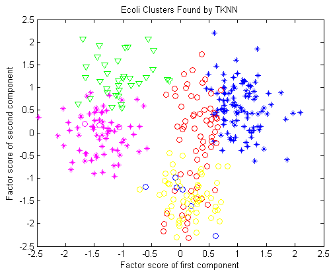


Fig. 7. Visualization of true clusters on a1) *iris* , b1) *wine*, c1) *glass* and c1) *ecoli* data compared againts clustering result produced by using TKNN on a2) *iris wine* b2) and *ecoli* c2) data

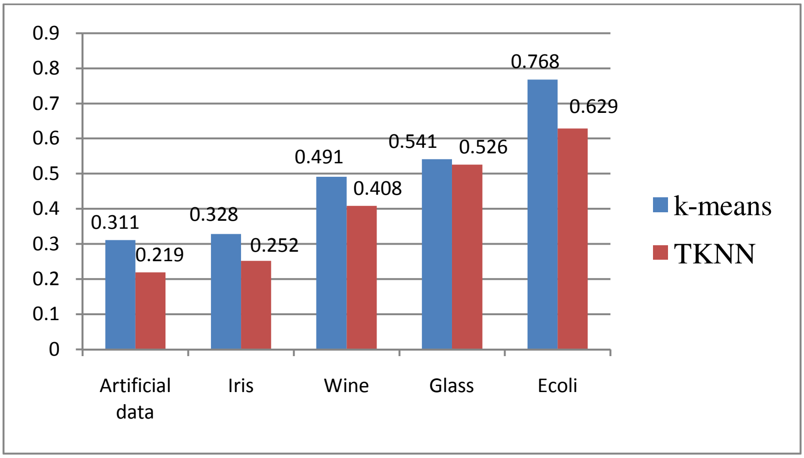


Fig. 8. Comparison of time processing required by k-means and TKNN

3.3 Computational Complexity Analysis

TKNN produced a distance matrix of size $(n(n-1))/2$ because redundant distance calculations were eliminated. The distance calculation process required $O(n^2)$ memory. However, for the huge dataset (over 10,000 objects), it was suggested to separate the distance calculation process and store the distance matrix in another file or variables. Thus, the computational complexity of TKNN depended mainly on the calculation of the k-nearest neighbor list through the kernel density estimation stage that requires the search for the data points failing in the neighborhood. In the results, their complexity was being relative to very expensive. The proposed TKNN has the simplest way to solve this problem. TKNN orders all the distances from the considered object to other objects, which leads to a complexity of $O(n \log(n))$.

Next, the computational complexity of three other clustering algorithms was described. K-means has computational complexity of $O(n * nc * I * d)$, where n is the number of data or objects, nc is the number of clusters, I is the number of iteration and d is number of attributes. It indicates that k-means has linear complexity and high cost time consumption. DBSCAN visits each point within the dataset, possibly multiple times. For practical considerations, however, the time complexity is mostly governed by the number of region Query invocations. DBSCAN executes exactly one such query for each point, and if an indexing structure is used to execute such a neighborhood in $O \log(n)$, an overall runtime complexity of $O(n\log(n))$ will be obtained. Since ILGC has a similar algorithm with TKNN, ILGC has the same computational complexity of $O(n \log(n))$.

4 Conclusion

This paper proposes the use of triangular kernel nearest neighbor (TKNN) for discovering true clusters with arbitrary shape, size and density in classification datasets.

In detail, TKNN combines the triangular kernel density and k-nearest neighbor (*k-NN*) density estimations to determine the number and member of cluster automatically. The experiments showed that the proposed approach could achieve more accurate clustering result and less time processing compared against k-means, ILGC, KFCM and DBSCAN.

Acknowledgments. This work is supported by a research grant from Universiti Teknologi Malaysia (UTM) VOT number QJ.130000.7128.01H12. The authors gratefully acknowledge many helpful comments by reviewers and members of Soft Computing Research Groups (SCRG) UTM Malaysia in improving the publication.

References

1. Anderson, T.K.: Kernel Density Estimation and K-means Clustering to Profile Road Accident Hotspots. Accident Analysis and Prevention 41(3), 359–364 (2009)
2. Golob, T.F., Recker, W.W.: A Method for Relating Type of Crash to Traffic Flow Characteristics on Urban Freeways. Transportation Research Part A: Policy and Practice 38(1), 53–80 (2004)
3. Shekhar, S., et al.: Data Mining and Visualization of Twin-cities Traffic Data, in Technical Report (TR 01-015), University of Minnesota (2001)
4. Skyving, M., Berg, H.Y., Laflamme, L.: A Pattern Analysis of Traffic Crashes Fatal to Older Drivers. Accident Analysis and Prevention 41(2), 253–258 (2009)
5. Steinbach, M., et al.: Discovery of climate indices using clustering. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC (2003)
6. Wang, M., Wang, A.P., Li, A.B.: Mining spatial-temporal clusters from geo-databases. In: 2nd International Conference on Advanced Data Mining and Applications, Xian, PEOPLES R CHINA (2006)
7. Lin, F., et al.: Discovery of teleconnections using data mining technologies in global climate datasets. Data Science Journal 6(suppl.), S749–S755 (2007)
8. Birant, D., Kut, A.: ST-DBSCAN: An algorithm for clustering spatial-temp oral data. Data & Knowledge Engineering 60(1), 208–221 (2007)
9. Chang, W., Zeng, D., Chen, H.C.: Prospective spatio-temporal data analysis for security informatics. In: 8th IEEE International Conference on Intelligent Transportation Systems (ITSC 2005). IEEE, Vienna (2005)
10. Zhang, D., Chen, S.: Kernel-based fuzzy and probabilistic c-means clustering. In: The International Conference on Artificial Neural Networks, Istanbul, Turkey (2003)
11. Zhang, D., Chen, S.: Clustering incomplete data using kernel-based fuzzy c-means algorithm. Neural Processing Letters 18, 155–162 (2003)
12. Hinneburg, A., Keim, D.A.: An effecient approach to clustering in large multimedia databases with noise. In: The Fourth International Conference on Knowledge Discovery and data Mining (KDD 1998). AAAI Press, Menlo Park (1998)
13. Hinneburg, A., Keim, D.A.: A general approach to clustering in large database with noise. Knowledge and Information Systems 5(4), 387–415 (2003)
14. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Addison Wesley (2006)

15. Ester, M., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceeding on the 2nd International Conference on Knowledge Discovery and Data Mining, Portland (1996)
16. Musdholifah, A., Hashim, S.Z.M.: Triangular kernel nearest neighbor-based clustering for pattern extraction in spatio-temporal database. In: The 10th International Conference on Intelligent System Design and Applications, Egypt (2010)
17. Classification data, UCI Repository of Machine Learning Database
18. Wasito, I., Hashim, S.Z.M., Sukmaningrum, S.: Iterative Local Gaussian Clustering for Expressed Genes Identification Linked to Malignancy of Human Colorectal Carcinoma. Bioinformation 2(5), 175–181 (2007)
19. Tran, T.N., Wehrens, R., Buydens, L.M.C.: KNN-kernel density-based clustering for high-dimensional multivariate data. Computational Statistics & Data Analysis 51, 513–525 (2006)
20. Clustering datasets, Speech and Image Processing Unit, School of Computing, University of Eastern Finland (2012)
21. Fu, L., Medico, E.: A novel fuzzy clustering method for the analysis of DNA microarray data. BMC bioinformatics 8(1), 3 (2007)
22. Jain, A.K., Law, M.H.C.: Data Clustering: A User's Dilemma. In: Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.) PReMI 2005. LNCS, vol. 3776, pp. 1–10. Springer, Heidelberg (2005)
23. Chang, H., Yeung, D.-Y.: Robust path-based spectral clustering. Pattern Recognition 41(1), 191–203 (2008)
24. Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation. ACM Transactions on Knowledge Discovery from Data 1(1), 1–30 (2007)
25. Zahn, C.T.: Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transaction on Computers 100(1), 68–86 (1971)
26. Veenman, C.J.: A maximum variance cluster algorithm. IEE Transaction on Pattern Analysis and Machine Intelligence 24(9), 1273–1280 (2002)
27. Van Rijsbergen, C.J.: Information retrieval. Butterworths, London (1979)
28. Gullo, F., Ponti, G., Tagarelli, A.: Clustering Uncertain Data Via K-Medoids. In: Greco, S., Lukasiewicz, T. (eds.) SUM 2008. LNCS (LNAI), vol. 5291, pp. 229–242. Springer, Heidelberg (2008)
29. Martinez, W.L., Martinez, A.R.: Exploratory data analysis with MATLAB. Chapman & Hall/CRC (2005)

DisClose: *Dis*covering Colossal *Close*d Itemsets via a Memory Efficient Compact Row-Tree

Nurul F. Zulkurnain[1, 3], David J. Haglin[2], and John A. Keane[3]

[1] Department of Electrical and Computer Engineering, Kuliyyah of Engineering, International Islamic University Malaysia, P.O. Box 10, 50728 Kuala Lumpur, Malaysia
nurulfariza@iium.edu.my
[2] High Performance Computing, Pacific Northwest National Laboratory, P.O. Box 999, MSIN J4-30, Richland, WA 99352, USA
david.haglin@pnl.gov
[3] School of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
{zulkurnn,jak}@cs.man.ac.uk

Abstract. A recent focus in itemset mining has been the discovery of frequent itemsets from high-dimensional datasets. With exponentially increasing running time as average row length increases, mining such datasets renders most conventional algorithms impractical. Unfortunately, large cardinality itemsets are likely to be more informative than small cardinality itemsets in this type of dataset. This paper proposes an approach, termed *DisClose*, to extract large cardinality (colossal) closed itemsets from high-dimensional datasets. The approach relies on a Compact Row-Tree data structure to represent itemsets during the search process. Large cardinality itemsets are enumerated first followed by smaller ones. In addition, we utilize a minimum cardinality threshold to further reduce the search space. Experimental results show that *DisClose* can achieve extraction of colossal closed itemsets in the discovered datasets, even for low support thresholds. The algorithm immediately discovers closed itemsets without needing to check if each new closed itemset has previously been found.

Keywords: Colossal closed itemset, high-dimensional dataset, minimum cardinality threshold.

1 Introduction

Rapid development in information technology has provided organizations with the ability to store, process and retrieve huge amounts of data. Nevertheless, there is a need to extract useful information and knowledge, efficiently and effectively, from these massive data stores. This serve to assist businesses, scientific and government related organizations to better plan, predict, and make decisions. This has led to the importance of data mining and the need to provide effective and efficient associated algorithm implementation.

Itemset mining has recently focused on the discovery of frequent itemsets from high-dimensional datasets with relatively few rows and a larger number of items [1],

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 141–156, 2013.

[2], [3], [8]. With exponentially increasing running time as average row length increases, mining such datasets renders most conventional algorithms impractical. Several papers have proposed the row-enumeration method to discover frequent itemsets based on the set of rows space instead of the itemset space [1], [2], [3], [8].

Nevertheless, due to the large number of frequent itemsets, discovering all frequent itemsets remains difficult. Strategies to provide more compact sets of frequent itemsets have been proposed such as finding only maximal frequent itemsets [4] or only closed frequent itemsets [5]. Closed itemsets provide a smaller set of results without information loss. Nonetheless, due to the density of high-dimensional data, it is difficult to enumerate all closed itemsets especially at the lower of the support spectrum [1], [2], [3], [8].

The most frequent itemsets tend to be both relatively smaller in size and larger in number. This quickly leads to insufficient memory when attempting to reach less frequent itemsets. Also, the most frequent itemsets can easily be extracted. In addition, applying the support constraint results in the pruning of many large cardinality itemsets that exist at this lower end of the support spectrum. Hence, discovery that starts from the largest cardinality itemsets in high-dimensional datasets may provide interesting insight into how these itemsets correlate.

Determining whether a mined pattern is closed is regarded as the main challenge in closed itemset mining [6]. Repeated checking to verify whether the itemsets are closed is costly in term of processing. Hence, discovering closed itemsets during the search process, and thus reducing the need for checking, should reduce computation.

This paper proposes discovery of colossal closed itemsets using a compact row-tree which reduces the memory required to store itemsets during the search process. The search for itemsets proceeds from the largest to the smallest by applying a search strategy that begins with the largest cardinality itemset and builds smaller itemsets. This strategy is combined with a bottom-up row enumeration search. We further utilize a minimum cardinality threshold to reduce the search space and focus on only colossal closed itemsets. We show that the algorithm immediately discovers colossal closed itemsets without the need to check each previously discovered closed itemsets.

The paper is structured as follows: Section 2 formulates the problem; search strategies are discussed in Section 3; in Section 4 the closedness-checking method is described; Section 5 presents the Compact Row-Tree; the algorithm and supporting theory are given in Section 6; experimental result are presented in Section 7; and Section 8 concludes.

2 Problem Formulation

Let T be a dataset table that consists of a collection of rows (transactions), $R = \{r_1, r_2, \ldots, r_m\}$ and a list of discrete items, $I = \{o_i, o_2, \ldots, o_n\}$. This set of transactions represents the number of rows (m) and the set of items signifies the number of columns (n) in T.

Table 1. Example of a discretized high-dimensional dataset

tid	Item													
	a	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***	***j***	***k***	***l***	***m***	***n***
1	1	1	1	2	2	1	2	1	2	2	2	2	2	1
2	2	1	2	2	2	2	2	2	1	2	2	2	1	2
3	1	1	2	1	2	2	2	2	2	1	2	2	2	2
4	2	1	2	1	2	2	1	2	2	2	2	2	2	2
5	1	1	2	1	1	2	2	2	2	1	2	2	2	2

A nonempty subset $\alpha \subseteq I$ is called an *itemset*. An itemset, α_k, which consists of k items, is described as k-itemset. Each row r_i is represented by a unique row identifier. Let $t\ (r_i)$ denote the itemset at row i of the table. Within a dataset, all of the row identifiers must be unique, but there may be duplicate row itemsets. That is, for $r_1 \neq r_2$, it may be that $t\ (r_1) = t\ (r_2)$. A set of rows is termed a *rowset*.

Example 1. (Table T) Table 1 illustrate as example of a discretized high-dimnesional dataset, T, that contains five rows and 14 items, so R = {1, 2, 3, 4, 5} and $I = \{a, b, c, d, e, f, g, h, i, j, k, l, m, n\}$.

Definition 1. (*Support Set*) Given an itemset α, the *support set* is represented as the set of rows in the dataset, T, that contain α. This is represented as:

$$r(\alpha) = \{r_i \mid \alpha \subseteq t(r_i)\} \tag{1}$$

Example 2. (*Support Set*) In Table 1, for an itemset $\alpha = \{a_1, b_1, d_1, e_2\}$, the support set $r_\alpha = \{3, 4\}$.

Definition 2. (*Support*) The support of an itemset α is the number of rows in which α occurs in T – denoted as $|\ r_\alpha\ |$.

Example 3. (*Support*) From Example 2, the support for itemset $\alpha = \{a_1, b_1, d_1, e_2\}$, $|\ r_\alpha\ | = |\ \{3,4\}\ | = 2$.

The relative support of α is $|r\ (\alpha)|/|r|$. It is well known that the support measure has an anti-monotonic property where $|r\ (\alpha_1)| \geq |r\ (\alpha_2)|$ for $\alpha_1 \subset \alpha_2$.

Definition 3. (*Frequent Itemset*) Given a dataset T and a minimum support threshold *minsup*, an itemset α is frequent if $|r\ (\alpha)| \geq minsup$.

Definition 4. (*Closed Itemset*) An itemset α is a closed itemset in dataset T if there is no proper superset α' exists $(\alpha \subset \alpha')$ such that the support of α is the same as the support of α'.

Definition 5. (*Closed Rowset*) A rowset β is a closed rowset in table T if not a proper superset β' exists $(\beta \subset \beta')$ such that the support of β is the same as the support of β'.

Definition 6. (*Closure*) Given a rowset β, we define $I(\beta) = \{o_j \in I \mid \forall_{r_k} \in \beta : o_j \in t(r_k)\}$. Following this, we can define $C(\alpha)$ as the closure of itemset α and $C(\beta)$ as the closure of rowset β as follows:

$$C(\alpha) = I(r(\alpha)) \tag{2}$$

$$C(\beta) = I(r(\beta)) \tag{3}$$

3 Search Strategies

3.1 Existing Itemset Search Strategies

The traditional search strategy explores the itemset space bottom-up: beginning from the smallest itemset that appears frequently and uses intermediate results to progressively build larger and larger itemsets. Conventional algorithms that use this strategy, such as *FP-Close* [5], are efficient for datasets containing relatively many rows and fewer columns (items) e.g. transactional data.

In contrast, high-dimensional datasets have a relatively large number of columns (items) and relatively few rows. If k is the maximum itemset size, there could be 2^k potential frequent itemsets. Exploring the dataset based on the number of items makes searching for closed frequent itemsets over the itemset space impractical.

CARPENTER [1] is an example of algorithms that search for closed frequent itemsets based on the rowset space. The algorithm conducts a bottom-up traversal of the row enumeration tree. Each node is checked to see if it is frequent and closed. As this criterion is based on the minimum support threshold, the nodes that do not satisfy the support constraint still need to be checked. As a result, the algorithm consumes both more memory and time in order to obtain the desired threshold.

By using the pruning power of the support threshold to reduce the search space, a top-down approach using a row enumeration tree has been proposed to discover closed frequent itemsets [2]. The search begins from the largest rowset and successively builds smaller and smaller rowsets. However, difficulties are still encountered in reaching the lower end of the support spectrum as much memory is consumed by the large numbers of closed frequent itemsets at the higher end.

A related problem is that of finding a formal concept (FC) [3]. Given a 0/1 matrix, a formal concept is a subset of k rows and l columns such that all of the matrix entries in one of the k rows and l columns contain a 1. Such a row and column subset is called a 1-rectangle. If the rows were rearranged so that all of the k subset rows appeared first (i.e. in rows 1 through k) and all columns were rearranged so that the l columns of the subset appeared in columns 1 through l, the upper-left k by l rectangle of the matrix would contain all 1 entries.

3.2 Proposed Search Strategy

A strategy for high-dimensional datasets is to search for closed itemsets based on the row number. Previous algorithms that use row-enumeration strategies to discover closed itemsets rely on the support constraint to reduce search space [1], [2], [3], [8]. As the frequency threshold reduces, the time and memory required for these algorithms to find closed itemsets dramatically increases. Yet the most valuable closed itemsets in high-dimensional data may have relative support values much closed to 1 than 100.

Therefore, we propose that rather than generating closed itemsets from the smallest set of items with higher supports, we search from the largest set of items that exists in a row possibly with very small support threshold. From this collection of closed itemsets, we can build increasingly smaller itemsets with increasingly higher support.

Bottom-Up Row-Enumeration Search

To extract the largest itemset from a dataset involves extracting the largest column that exists in the dataset. This implies that the search strategy can be based on a top-down column enumeration.

However, it can be observed that for a dataset with m number of columns (items), there will also be m number of levels for a top-down column enumeration tree. In addition, the maximum number of nodes (itemsets) that will exist in the top-down column enumeration will equal 2^m-1. For a high-dimensional dataset, the value of m is very large (i.e. hundreds of thousands); hence, enumerating the itemsets based on the number of columns is infeasible.

It makes sense to search for closed itemsets based on the number of rows because it is relatively smaller than compared to the number of columns in high-dimensional datasets [1], [2], [8]. The largest cardinality itemset initially exists in every single row of the high-dimensional dataset (unless duplicate rows occur). Therefore, most large closed itemsets begin from the infrequent end of the support spectrum. As a result, using the bottom-up row enumeration tree as the basis of the search strategy would appear to be more appropriate.

Transposed Table

Since its proposal, the transposition method [7] has been widely used by algorithms that discover closed itemsets from high-dimensional datasets [1], [2], [8]. Mining the closed itemsets directly from the original dataset can be complicated. Therefore, applying the method of transposition to the original dataset helps to simplify the extraction of closed itemsets in high-dimensional data. This is because when the original dataset is transposed, each column (item) value of the original dataset will become a row value in the transposed table, and will be represented by a set of rows (rowset) where that particular item occurs.

Transposed dataset provides a sparser representation of the original input dataset. As a result of this simplification, the method of transposition is utilized in the algorithm proposed here.

Minimum Cardinality Threshold, *mincard*

Definition 7. (*Cardinality*) The cardinality of an itemset α refers to the number of items in α. This is denoted as $|\alpha|$.

In contrast to the support threshold that helps to reduce the search space based on occurrence frequency, we stop the mining process for closed itemsets upon reaching the threshold parameter value for the minimum itemset cardinality, *mincard*.

Let $CI(T) = \{\alpha \mid \alpha \in T \text{ is a closed itemset}\}$. Large cardinality closed itemset mining involves enumerating all $\alpha \in CI(T)$ with $|\alpha| \geq$ *mincard*. We refer to these large cardinality closed itemsets as colossal closed itemsets (*CCI*).

Using the bottom-up row enumeration tree [1], branch exploration stops once the cardinality of the associated itemset falls below *mincard*. We can safely prune the search because of the anti-monotone property.

Property 1. (*anti-monotone*) If a rowset β has its associated $\alpha = I(\beta)$ such that $|\alpha| <$ *mincard*, then for any $\beta' \supseteq \beta$ it must be that $|I(\beta')| <$ *mincard*.

Combining both the anti-monotone property and the definition of closure gives the following property.

Property 2. (*at-threshold*) If a rowset β has its associated $\alpha = I(\beta)$ such that $|\alpha| ==$ *mincard*, then for any $\beta' \supseteq \beta$ it must be that $|I(\beta')| <$ *mincard*.

Note: Using a depth-first order in a serial implementation would result both in the most aggressive pruning of the search space and require the least memory.

4 Closedness-Checking Method

Mining for colossal closed itemsets has two restrictions: firstly, the need to check if an itemset is a colossal itemset and secondly, the need to check if it is closed. Using the minimum cardinality threshold in a bottom-up row enumeration search takes advantage of the first constraint. However, discovering only the colossal itemsets may lead to the production of several identical colossal itemsets,

Therefore, when a colossal itemset is found, the next step is to develop a method to efficiently identify whether it is a closed itemset. The method of identifying whether the itemsets discovered are closed is related closely to the search strategy proposed.

To take advantage of the second restriction in making the mining of colossal itemsets more efficient, a method which is based on a unique generator is developed. To define the unique generator, we begin by providing the definition for *itemset generator* and *tidset generator*.

Definition 8. (*Itemset Generator*) Given a dataset T, an itemset α is an itemset generator if no proper subset $\alpha' \subset \alpha$ exists such that the support of α is the same as the support of α'.

The equivalence class of itemsets with the same support set consists of exactly one closed itemset, potentially many itemset generators and potentially many itemsets that are neither closed nor generators.

Definition 9. (*Rowset Generator*) Given a dataset T, a rowset β is a rowset generator if no proper subset $\beta' \subset \beta$ exists such that the itemset of β is the same as the itemset of β'.

Similarly, the equivalence class of rowsets β_i with same itemset α such that $I(\beta_i) = \alpha$ consists of exactly one closed rowset, there are potentially many rowset generators and potentially many rowsets that are neither closed nor generators.

It can be observed that unlike the definition of frequent itemsets, the definitions of generators and closed sets do not depend upon any threshold parameter.

To construct smaller closed itemsets from larger ones, we use the following property:

Theorem 1. Suppose α_1 and α_2 are closed itemsets, with $\alpha_1 \neq \alpha_2$. Let $\alpha = \alpha_1 \cap \alpha_2$. If $\alpha \neq \emptyset$ then α is a closed itemset.

Proof: We have three cases to consider:

1. **Case 1:** $[\alpha_1 \subset \alpha_2]$. ***Observe that in this case*** $\alpha = \alpha_1$, ***so*** α ***is a closed itemset***.
2. **Case 2:** $[\alpha_2 \subset \alpha_1]$. ***Observe that in this case*** $\alpha = \alpha_2$, ***so*** α ***is a closed itemset.***

   For Case 1 and Case 2, in order for α_1 and α_2 to be closed itemsets with one a proper subset of the other, it must be the case (by definition of closed itemset) that they have different support. But we do know that such a situation exists.

   Consider any closed itemset α_1 with support larger than one, and pick any row r_i containing α_1 (i.e. $\alpha_1 \subset t(r_i)$). Now consider $\alpha_2 = t\ (r_i)$. Note that by definition all full-rowsets are closed. Clearly, this satisfies the conditions of Case 1. The rest of the case is just fundamental set theory, so the result holds.

3. **Case 3:** [α_1 and α_2 are incomparable]. ***Observe that*** $\alpha \subset \alpha_1$ ***and*** $\alpha \subset \alpha_2$.
   In this particular case, it is demonstrated that α is a closed itemset by contradiction. Assume that α is not a closed itemset, then there exists some item i such that $\alpha_i = \alpha \cup \{i\}$ has the same support as α. If $i \notin \alpha_1$, then all rows in $T_{\alpha_1} - T_\alpha$ are

not in T_{α_1}, but they are in T_α. Thus i must be in α_1. However, if $i \in \alpha_1$ (and not in α) then $i \notin \alpha_2$ and the same contradiction argument applies. Thus the assumption that α is not a closed itemset must be invalid.

Lemma 1. Every closed itemset that is not one of the entire transactions can be produced by intersecting some collection of closed itemsets.

Proof. Consider a closed itemset α and its corresponding rowset $\beta = T_\alpha$, as α is a closed itemset, $\alpha = \bigcap_{r_i \in \beta} \alpha_i$, where $(r_i, \alpha_i) \in T$. However, there may be many subsets of β for which $I(\beta) = \alpha$.

Using rowset enumeration as the control strategy for the search process, the same closed itemset would probably be found many times. The following observation allows a closed itemset to be found using only one of the rowsets. As stated above, for every closed itemset α, there is a unique rowset β that is a closed rowset.

Definition 10. (*Unique Generator*) Given the closed rowset $\beta = \{r_1, r_2, \ldots, r_k\}$, $r_i < r_j$ for all $i < j$, the smallest index for which $\beta_j = \{r_1, r_2, \ldots, r_j\}$ is a generator of β is a unique rowset generator for our itemset α.

It is simple to determine if a rowset β' is the unique generator. Let $\beta = T(I(\beta'))$. If β' $= \beta$, then the answer is that β' is the unique generator. If $\beta' \subset \beta$, β' is determined whether it is a prefix of β when the rowsets are written as lists in ascending order. If β' is not a prefix of β, then β' can be ignored and this branch of the search space is pruned.

The search for the unique generator will require relatively little computation when the number of rows is small; and this is the typical situation for high-dimensional datasets.

5 Compact Row-Tree

To assists the efficiency of the search, a compact tree data structure is built to store the itemsets from the transposed table, T^t. The *CR-Tree* is initially generated by building a set of nodes at the first level ($l = 0$) of the tree which represents each column value of the transposed table. These set of nodes are connected to each column of the transposed table through a set of pointers that link the node to the transposed table. The construction of the *CR-Tree* continues by adding the child nodes at each level of the tree. As the level of the tree increases, the number of child nodes decreases as the lowest node value from the previous level of the tree is discarded. A child pointer is then built to link between the nodes. In addition to the child pointer, an additional node link is made from the parent node to the child node that contains the same node value. The purpose of this node link is to assist in checking effectively for closed itemsets.

The structure of the *CR-Tree* is similar to the *FR-Tree* [2]. The *CR-Tree* is different in that instead of representing each branch of the tree to a rowset value, each node of the *CR-Tree* represents a group of rowset values. In this way, the *CR-Tree* becomes more compact as one node is shared by many rowset values. Each rowset value represents an itemset.

However, only one rowset value will be stored in each node of the *CR-Tree* during the search process. This is to ensure that a relatively small amount of memory is utilized during the process of mining the colossal closed itemset.

The characteristics of the *CR-Tree* are as follows:

1. The *CR-Tree* represents all values of the complete rowset.

   Let $N = \{n_i, n_{i+1}, \ldots, n_k\}$ be the set of nodes where $i = 1$ and k is the largest row value from the dataset. Let $M = \{m_j, m_{j+1}, \ldots, m_k\}$ be the set of child nodes where $j = i + 1$ and k is the largest row value of the dataset. Each $m_j = n_i \cup n_{i+l}$ where $l = \{1, 2, \ldots, k\text{-}1\}$. Therefore all β values are traversed until $i = k$.

Subsequent itemsets of the child nodes are obtained by intersecting the itemsets of the parent nodes. To reduce memory during the search for colossal closed itemsets, each node in the *CR-Tree* only stores one itemset value from the intersection of its parent nodes. Each rowset of the parent nodes is a subset of a rowset value of the child node.

2. Each node of the *CR-Tree* only stores one β value at a time for rowset β, with $|\beta|$ = node level.

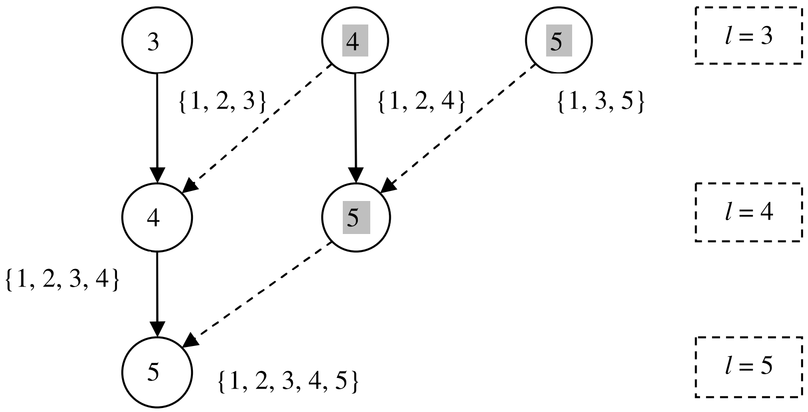


Fig. 1. Example of the second characteristics of the *CR-Tree*

Suppose at $l = 3$, $n_4 = \{1, 2, 4\}$ and $n_5 = \{1, 3, 5\}$. To obtain β for child node m_5, the union of the parent β values will produce, $n_4 \cup n_5 = \{1,2,4\} \cup \{1,3,5\} = \{1, 2, 3, 4, 5\}$. However, $\{1, 2, 3, 4, 5\}$ is not stored in m_5. This is because, based on the depth-first strategy, the itemset for $\beta = \{1, 2, 3, 4, 5\}$ will already have been discovered at $l = 5$.

The structure of the *CR-Tree* also assists in optimizing identification of the closed itemsets. For example, consider node 3 at the second level of the tree. Assume that the node contains an itemset with row values {1, 3}. Using the proposed closedness-checking method, the node will intersect with the nodes at the third level (Nodes 3, 4, 5) and then check with row values – {1, 2, 3}, {1, 3, 4} and {1, 3, 5} whether the itemset of {1, 3} exists in row 2, 4, or 5.

3. If discovered itemset, $\alpha_2 \subseteq \alpha_1$ where α_1 is the existing itemset in the node, the itemset α_2 will not replace α_1 although $\beta_1 \neq \beta_2$.

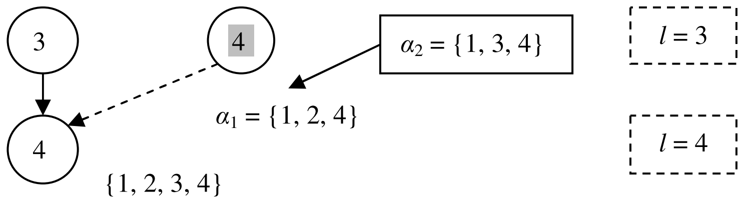


Fig. 2. Example of the third characteristics of the *CR-Tree*

In Fig. 2, suppose $\alpha_1 = \{\beta_1\} = \{1, 2, 4\}$ and $\alpha_2 = \{\beta_2\} = \{1, 3, 4\}$ at level $l = 3$, where $|\beta_1| = |\beta_2|$. If $\alpha_2 \subseteq \alpha_1$, this means that α_2 also exists in $\{\beta_1\}$. Therefore, $\beta_1 \cup \beta_2 = \beta$, where $|\beta| > |\beta_1|, |\beta_2|$. Thus α_2 will exists in $\beta = \{1, 2, 3, 4\}$ where β already exists at level, $l = 3$ of the *CR-Tree*.

6 Algorithm *DisClose*

To show the effectiveness of the search strategy, the closedness-checking method and the data structure proposed, a colossal closed itemset mining algorithm called *DisClose* has been designed to mine all colossal closed itemsets from the transposed table T^t of table T. *DisClose*, shown in Algorithm 1, will search the row enumeration space and, for each rowset, β, check whether it is the unique generator in the equivalence class of rowsets for $I(\beta)$. It is noted that using a depth-first order in a serial implementation would result in the most aggressive pruning of the search space and requires the least the amount of memory [1], [10]. For this reason, the general processing order for the rowsets is equivalent to the depth-first search of the row enumeration tree.

6.1 Major Steps of *DisClose*

Algorithm 1 shows the main steps of the algorithm *DisClose*. Assuming the *mincard* threshold value has been assigned, the algorithm begins by transforming table T into

Algorithm 1. *DisClose* algorithm

Input: Table T, and minimum cardinality threshold, *mincard*
Output: A complete set of colossal closed itemsets, *CCI*
Method:
1. Transform T into transposed table T^t
2. Build *CR-Tree*
3. Initialize $CCI = \emptyset$
4. Call Subroutine Colossal (T^t, *mincard*)

Subroutine Colossal (T^t, *mincard*)
Method:
5. **for** each node in the row enumeration space **do**
6. If | node [1][j] |.T^t | $\geq$ *mincard*
7. Store itemset at node [1][j]
8. Let β be the set of rows under consideration
9. node [l][j] $\rightarrow$ node [l+1][p] // pointing to child node
10. $\alpha = \alpha_1 \cap \alpha_2 = I(\beta)$, $\beta = \beta_1 \cup \beta_2$
11. ***Optimization S1***: If | α | < *mincard*, discard α
12. ***Optimization S2***: If | β | > current node level, discard β
13. ***Optimization S3***: If $\alpha \subseteq \alpha'$, discard α
14. Store α in node [l+1][p]
15. Call Subroutine **Closed** (*mincard*)

Subroutine Closed (*mincard*)
Method:
16. If node [l][j] == node [l+1][p] // checking for unique generator
17. Call Subroutine **Colossal** (*mincard*)
18. Store itemset in *CCI*
19. Call Subroutine **Colossal** (*mincard*)

transposed table T^t using the transposition operation. After the transposed table T^t is generated, the *CR-Tree* is built in Step 2 in order to access the colossal itemsets from T^t. The *CR-Tree* connects nodes at the first level to T^t through side-links. The side-link pointers enable direct access to the colossal itemsets from T^t. These pointers connect the node with the column T^t of equivalent value.

DisClose is compoased of two main subroutines: *Colossal* and *Closed*. After initialization of the set of colossal closed itemsets *CCI* to be empty at Step 4, the subroutine *Colossal* is called to deal with the transposed table T^t using the *CR-Tree* and find all colossal itemsets. Following the bottom-up row enumeration as the search order in step 5, the subroutine *Colossal* takes the transposed table, T^t and the *mincard* threshold as the parameter to ensure that itemsets with cardinality less than the specified is not stored as it is impossible for subsequent child nodes of the *CR-Tree* to contain itemsets of larger size.

There are seven sections in the subroutine *Colossal*, which will be explained one by one.

The first section is step 6 – step 7. The subroutine begins by accessing T^t through a side-link from the *CR-Tree*. The size of the itemset is checked for each column at Step 6. Only the itemsets that satisfy *mincard* are stored at the first level node of the *CR-Tree*; otherwise, it is not stored in the node as it will not contribute to obtaining larger itemsets. The advantage of this is that the algorithm does not require further access to the dataset, and hence, reduces the time required for repeated checking of the dataset. Note that this is the only role the transposed table T^t plays in the search process.

The second section is steps 8 – step 10. For each node in the *CR-Tree*, the intersection of the itemsets between the parent nodes continues using the depth-first search of the bottom-up row-enumeration tree in Step 10 is performed. By using depth-first search, *DisClose* produces the sequence $\beta \Rightarrow I(\beta)$. However, three optimization strategies are applied before the result of the intersection is stored in each child nodes.

At step 11, an optimization strategy *S1* is applied to stop further processing of the itemset if the size of the itemset does not satisfy the *mincard* constraint defined.

Step 12 performs the optimization strategy *S2* to prevent storage of itemsets with rowset values larger than the node level of the *CR-Tree*.

At step 13, optimization strategy *S3* is applied in order to ensure that the itemset obtained is not a subset of an already existing itemset in the child node.

Step 14 then stores the itemset that does not satisfy any of the three optimization strategies at the particular child node. The new itemset will replace the itemset that already exists in the node.

At step 15, the subroutine *Closed* is called when all the colossal itemsets of the child nodes have been discovered, in order to check whether the parent node is a closed itemset.

The subroutine *Closed* performs the closedness-checking method on the itemset. There are four main steps to this subroutine.

Step 16 sequentially compares the itemset α that exists in the parent node with the itemsets of its child nodes in order to identify the unique generator, based on a depth-first search of the rowset value in the row enumeration tree. Here, the node-link, which connects the parent and child node that contain the same node value, is used to perform the closedness-checking method. This is to ensure that it does not overlook existing child nodes with rowset β that contains a rid value that does not exist in rowset β' of the parent node.

7 Experimental Evaluation

Due to the space limitation, the experimental evaluation shows the comparison of *DisClose* with selected algorithms on one synthetic dataset.

The experiments were performed on a PC with a 2.66 GHz Intel Core2 Quad CPU Q9400 with 4.00 GB RAM and 150 GB hard disk. *DisClose* is implemented in C++.

The performance of *DisClose* was studied by comparing it with other state-of-the art algorithms. Each algorithm was selected to represent the different search strategies. These algorithms are: (i) *FP-Close* [5] - a representative of the column enumeration-based algorithms, (ii) *CARPENTER* [1] – a representative of bottom-up row enumeration-based algorithms, (iii) *D-Miner* [9] – a representative of constraint-based mining algorithms, and (iv) *TTD-Close* [2] – a representative of the top-down row enumeration search based set of algorithms.

All of the selected algorithms have been implemented in C++. Note: all of runtimes plotted in the figures include both computation time and I/O time.

Existing itemset mining algorithms – particularly those that find closed itemsets, - routinely present run-times for varying support thresholds. As *DisClose* uses a thresh old for cardinality, direct comparison is difficult. Given a support threahols greater than 1, existing algorithms would not find many large-cardinality closed itemsets; given a cardinality threshold greater than 1, *DisClose* would not find many frequent closed itemsets. The only fair way to compare the algorithms is to give both a threshold of 1, asking each to find all closed itemsets. The strength of *DisClose* is that it bypasses the huge number of small cardinality, high-frequency closed itemsets and focuses almost immediately on potentially valuable closed itemsets. However, this type of complete closed itemsets search does not really address the true purpose of either *DisClose* or the closed itemset mining algorithms.

Amongst the selected algorithms listed above, only *D-Miner* has been found to apply the minimum cardinality threshold, *mincard*. *D-Miner* is a constraint-based algorithm which uses the cardinality threshold in addition to the support constraint to discover concepts (closed itemsets). Hence, *D-Miner* is the closest comparison to *DisClose* with the exception that the support threshold in *D-Miner* is set to 1.

For other algorithms, an approach was to present the experimental results of *DisClose* with a secondary *x*-axis which represents the maximum support of the colossal closed itemsets discovered. Likewise, a secondary *x*-axis is also added to the results of *FP-Close*, *CARPENTER*, and *TTD-Close* which represents the maximum cardinality of the closed frequent itemsets discovered. Thus, by using this approach, it provides an observation on the ability and limitation of closed itemset mining algorithms that uses a support threshold in relation to *DisClose*, and vice-versa.

Another challenge in comparing performance of the algorithms is based on their implementation in identifying items in the datasets. For *FP-Close*, *CARPENTER*, and *D-Miner*, the algorithms were designed to identify each item based on the value present for each attribute of the dataset. However, for *TTD-Close*, each item in the dataset is read as a value that corresponds to the attribute of the data.

7.1 Synthetic Dataset

The synthetic dataset, generated using the IBM Quest Data Generator, consists of 100 rows, 4000 columns and an average itemset size of 2000. This dataset is represented as T100L2000N4000.

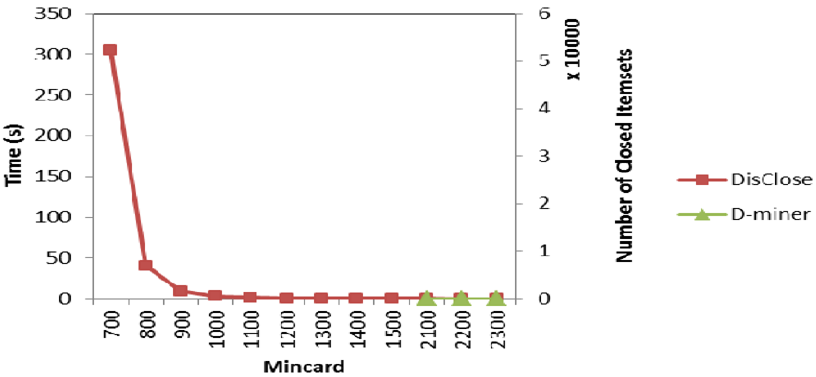
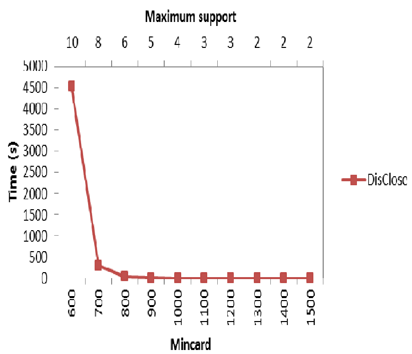


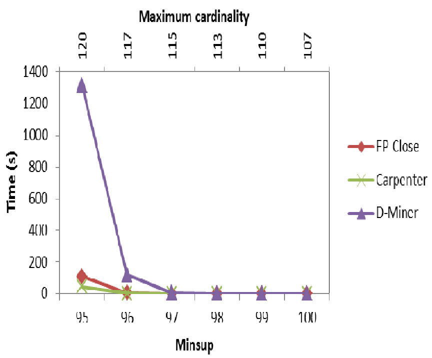
Fig. 3. Comparison with *D-Miner* using *mincard* threshold on T100L2000N4000

Fig. 3 shows the result of the performance between *DisClose* and *D-Miner*. It is observed that at a higher cardinality threshold, the difference in the time taken between the two algorithms is very small. However, as the *mincard* value decreases, *DisClose* largely outperforms *D-Miner*. Taking the maximum processing time of around 300 seconds, *DisClose* is able to discover colossal closed itemset with *mincard* = 700. For *D-Miner*, after *mincard* = 2100, the algorithm took more than 12 hours to discover the colossal closed itemsets.

As shown in Fig. 4(a), beginning with the largest closed itemsets, *DisClose* is able to discover the colossal closed itemsets with a maximum support of 10. The performance of *DisClose* sharply increases between *mincard* = 700 and *mincard* = 600. This is due to the large number of closed itemsets that exists between these thresholds. However, Fig. 4(b) shows that the algorithms are only able to reach *minsup* = 95 with the largest cardinality itemset of 120.



(a)



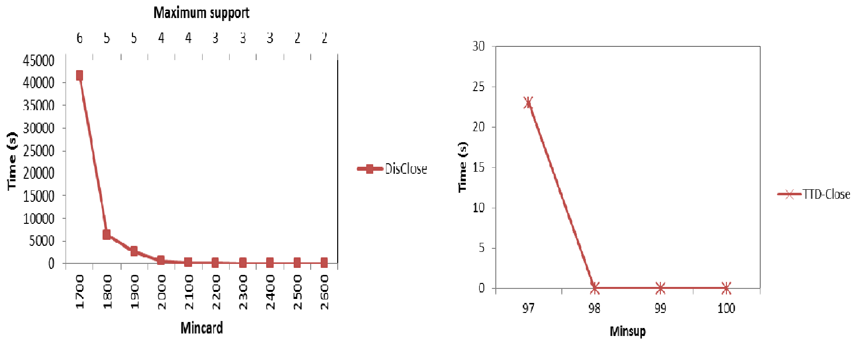
(b)

Fig. 4. Comparison with *FP-Close*, *CARPENTER* and *D-Miner* on T100L2000N4000

As the nature of *TTD-Close* is to read each item in the dataset as a value, the largest itemset that exists in the dataset is equivalent to its column value. Therefore, in this particular case, more colossal closed itemsets are discovered.

Fig. 5(a) shows that as the *mincard* value increases, the time required to discover the colossal closed itemsets also increases. *DisClose* is able to reach closed itemsets with *mincard* = 1700. There are a total of 78,717,638 closed itemsets that exists when *mincard* = 1700 having the maximum support of 6. This shows that for dense dataset, even at a high *minsup* threshold, the size of itemsets can become very large.

Fig. 5(b) shows that *TTD-Close* could only reach *minsup* = 97 with a total of closed itemsets of 58,505. *TTD-Close* runs out of memory probably due to the existence of larger cardinality itemsets at smaller *minsup* thresholds. This shows that for dense dataset, even at a high *minsup* threshold, the size of itemsets can become very large.



(a) (b)

Fig. 5. Comparison with *TTD-Close* on T100L2000N4000

8 Conclusions and Future Work

This paper has introduced an algorithm *DisClose* that searches for colossal closed itemsets from the largest cardinality itemsets that exist in the dataset. This search is integrated with the bottom-up row enumeration strategy. We propose *mincard* to further reduce the search space. The closedness-checking method used reduces the need to check whether a newly discovered closed itemset already exists. The new approach bypasses the huge number of small-cardinality, high-frequency closed itemsets and focuses on the potentially most valuable large closed itemsets.

The results show that the algorithm exhibit scalable performance with run-time being almost correlated with closed itemset count. *DisClose*'s performance appears to be linear with respect to the number of closed itemsets. This suggests there may be a relationship between the colossal closed itemsets identified and their corresponding support sets. In particular, do any of the closed itemsets actually identify known classes of examples in the datasets?

Further evaluation of the algorithm on real datasets is also required in order to calibrate behavior and efficiency.

Acknowledgments. We would like to thank to the authors of *D-miner* and *TTD-Close* for providing the executable. We would also like to thank to Christian Borgelt for providing the implementation codes for *FP-Close* and *CARPENTER* through his website.

References

1. Pan, F., Cong, G., Tung, A.K.H., Yang, J., Zaki, M.J.: CARPENTER: Finding closed patterns in long biological datasets. In: Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003), pp. 637–642. ACM (2003)
2. Liu, H., Wang, X., He, J., Han, J., Xin, D., Shao, Z.: Top-down mining of frequent closed patterns from very high dimensional data. Information Science 179(7), 899–924 (2009)
3. Zhu, F., Yan, X., Han, J., Yu, P.S., Cheng, H.: Mining colossal frequent closed patterns by core pattern fusion. In: Proc. International Conference on Data Engineering (ICDE 2007), pp. 706–715. IEEE (2007)
4. Bayardo, R.J.: Efficiently mining long patterns from databases. In: Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD 1998), pp. 85–93. ACM, New York (1998)
5. Grahne, G., Zhu, J.: Efficiently using prefix-trees in mining frequent itemsets. In: Proc. 1st IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI 2003), pp. 123–132 (2003)
6. Han, J., Cheng, H., Xin, D., Yan, X.: Frequent pattern mining: current status and future directions. Data Mining and Knowledge Discovery 15(1), 55–86 (2007)
7. Rioult, F., Boulicaout, J., Cremilleux, B., Besson, J.: Using transposition for pattern discovery from microarray data. In: Proc. 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2003), pp. 73–79. ACM (2003)
8. Cong, G., Tan, K.-L., Tung, A., Pan, F.: Mining Frequent Closed Patterns in Microarray Data. In: Proc. Fourth IEEE Int'l Conf. Data Mining (ICDM), vol. 4, pp. 363–366 (2004)
9. Besson, J., Robardet, C., Boulicaout, J.-F., Rome, S.: Constraint-based mining and its application to microarray data analysis. Intelligent Data Analysis Journal 9(1), 59–82 (2005)
10. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 1–12 (2000)

Author Index