Zhi-Hua Zhou
Friedhelm Schwenker (Eds.)

LNAI 8183

Partially Supervised Learning

Second IAPR International Workshop, PSL 2013
Nanjing, China, May 2013
Revised Selected Papers

IAPR

Springer

Lecture Notes in Artificial Intelligence 8183

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

Randy Goebel
University of Alberta, Edmonton, Canada

Yuzuru Tanaka
Hokkaido University, Sapporo, Japan

Wolfgang Wahlster
DFKI and Saarland University, Saarbrücken, Germany

LNAI Founding Series Editor

Joerg Siekmann
DFKI and Saarland University, Saarbrücken, Germany

For further volumes:
http://www.springer.com/series/1244

Zhi-Hua Zhou · Friedhelm Schwenker (Eds.)

Partially Supervised Learning

Second IAPR International Workshop, PSL 2013
Nanjing, China, May 13-14, 2013
Revised Selected Papers

Springer

Editors
Zhi-Hua Zhou
National key Laboratory for Novel
Software Technology
Nanjing University
Nanjing
China

Friedhelm Schwenker
Abteilung Neuroinformatik
Universität Ulm
Ulm
Germany

ISSN 0302-9743 ISSN 1611-3349 (electronic)
ISBN 978-3-642-40704-8 ISBN 978-3-642-40705-5 (eBook)
DOI 10.1007/978-3-642-40705-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013950103

CR Subject Classification (1998): I.2.6, I.5, H.2.8, I.2, I.4, H.3

LNCS Sublibrary: SL 7 - Artificial Intelligence

© Springer-Verlag Berlin Heidelberg 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

Preface

Partially supervised learning (PSL) is a rapidly evolving area of machine learning, data mining and pattern recognition. In many applications unlabeled data may be relatively easy to collect, whereas labeling these data is difficult, expensive, and/or time-consuming as it requires the effort of human experts. PSL is a general framework for learning with partial supervision, or learning with partially labeled data. In this framework, the supervision information might be a crisp label, or a label plus a confidence value, or it might be an imprecise and/or uncertain soft label defined through a certain type of uncertainty model, or it might be that information about a class label is not available.

The PSL framework thus generalizes many kinds of learning paradigms including supervised and unsupervised learning, semi-supervised learning for classification and regression, transductive learning, semi-supervised clustering, multi-instance learning, weak label learning, policy learning in partially observable environments, etc. Therefore, PSL theories and algorithms are of great interest to various research communities. Research in the field of PSL is still in its early stages and has great potential for further growth, thus leaving plenty of room for further development.

PSL 2013 received 26 full submissions. The Program Committee consisting of 24 experts carefully reviewed the submissions, with the help of some external reviewers. Based on the reviews, 10 papers were selected for presentation at the workshop and inclusion in the post-workshop proceedings. Authors were requested to improve their manuscripts by incorporating reviewers' comments and feedbacks from the workshop audience, leading to the revised selected papers in this volume. The workshop program was significantly enhanced by the invited talk of Prof. Dale Schuurmans of the University of Alberta, Canada.

This workshop would not have been possible without the help of many individuals and organizations. First of all, we would like to thank the Program Committee members and reviewers for their great efforts in providing insightful comments on the submissions. We also wish to thank all the authors who have submitted their recent work to the workshop. The management of the papers, including the preparation of this proceedings volume, was done by the EasyChair conference management system. Special thanks go to the local arrangement and publicity chairs, Ming Li, Yang Yu, and Michael Glodek, for their outstanding contribution to the organization of PSL 2013.

This workshop was organized by the LAMDA Group of the National Key Laboratory for Novel Software Technology, Nanjing University, China, and the Institute of Neural Information Processing, University of Ulm, Germany. We thank the International Association for Pattern Recognition (IAPR), IEEE Computer Society Nanjing Chapter, and the National Science Foundation of China for their support.

July 2013

Zhi-Hua Zhou
Friedhelm Schwenker

Committee

Chairs

Zhi-Hua Zhou	Nanjing University, China
Friedhelm Schwenker	University of Ulm, Germany

Local Arrangement Chair

Ming Li	Nanjing University, China

Publicity Chair

Yang Yu	Nanjing University, China
Michael Glodek	University of Ulm, Germany

Program Committee

Songcan Chen	China
Xiaofei He	China
Tom Heskes	The Netherlands
Steven C.H. Hoi	Singapore
Rong Jin	USA
Wee Sun Lee	Singapore
Hang Li	China
Yu-Feng Li	China
Cheng-Lin Liu	China
Marco Loog	The Netherlands
Marco Maggini	Italy
Fabio Roli	Italy
Dale Schuurmans	Canada
Masashi Sugiyama	Japan
Edmondo Trentin	Italy
Grigorios Tsoumakas	Greece
Haifeng Wang	China
Wei Wang	China
Kai Yu	China
Shipeng Yu	Germany
De-Chuan Zhan	China

Min-Ling Zhang	China
Jerry Zhu	USA
Jun Zhu	China

Additional Reviewers

Giorgio Fumera
Nan Li
Wei Wu
Yan-Ming Zhang
Tingting Zhao

Supported by

National Science Foundation of China
IAPR (International Association for Pattern Recognition)
IEEE Computer Society Nanjing Chapter

Contents

Partially Supervised Anomaly Detection Using Convex Hulls on a 2D Parameter Space

Gabriel B. P. Costa[1], Moacir Ponti[1(✉)], and Alejandro C. Frery[2]

[1] Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, São Carlos, SP, 13566-590, Brazil
{moacir, gpbcosta}@icmc.usp.br

[2] Universidade Federal de Alagoas, Maceió, AL13566-590, Brazil
acfrery@lccv.ufal.br

Abstract. Anomaly detection is the problem of identifying objects appearing to be inconsistent with the remainder of that set of data. Detecting such samples is useful on various applications such as fault detection, fraud detection and diagnostic systems. Partially supervised methods for anomaly detection are interesting because they only need data labeled as one of the classes (normal or abnormal). In this paper, we propose a partially supervised framework for anomaly detection based on convex hulls in a parameter space, assuming a given probability distribution. It can be considered a framework since it supports any model for the "normal" samples. We investigate an algorithm based on this framework, assuming the Gaussian distribution for the not anomalous ("normal") data, and compared the results with the One-class SVM and Naïve Bayes classifiers, as well as two statistical anomaly detectors. The proposed method shows accuracy results that are comparable or better than the competing methods. Furthermore, this approach can handle any probability distribution or mixture of distributions, allowing the user to choose a parameter space that models adequately the problem of finding anomalies.

Keywords: Anomaly · Outlier · Semi-supervised learning

1 Introduction

Anomaly detection is the problem of finding patterns (samples, individuals) with an unexpected behaviour. Barnett and Lewis [1] defined outliers (anomalies) as observations which appear to be inconsistent with the remainder of that set of data.

Due to the nature of the problem, anomalies are often rare and dealing with them can help on applications such as fault detection, fraud detection, network intrusion, diagnostic systems, and medical condition monitoring. Anomaly detection methods take as input a sample or set of samples, and identify whether they are "normal" or "abnormal", according to what is expected to be found. Note that, in this text, we will refer normal as the data that is consistent with the

Z.-H. Zhou and F. Schwenker (Eds.): PSL 2013, LNAI 8183, pp. 1–8, 2013.
DOI: 10.1007/978-3-642-40705-5_1,
© Springer-Verlag Berlin Heidelberg 2013

expected pattern, with no relation with the probability distribution. On most applications, “normal” samples are widely available, while anomalies are scarce or frequently not available at all [2].

Detecting anomalies is critical in many systems, as they indicate an abnormal condition to be actuated upon [7]. Some examples are: an engine rotation defect, a flow problem on a pipeline, an intruder inside a system with malicious intentions, a fault on a factory production, and a disease or a dangerous medical condition.

Anomaly detection belongs to the wide class of classification problems and, therefore, there are three fundamental approaches: (1) unsupervised, (2) supervised, and (3) partially supervised. In the first, when there is no prior knowledge of the data, unsupervised learning algorithms such as clustering are used. The general idea is to identify outliers, observations that appear not to belong to any of the detected groups. The supervised approach models both normality and abnormality, requiring labeled samples from both classes. In this paper we focus on a semi-supervised (or partially supervised) method, which models only normality; situations which only model abnormality are rare [4].

According to Hodge and Austing [7], the advantages of using a partially supervised method to detect anomalies are: (a) it only needs data labeled as “normal”, (b) it is suitable for static or dynamic data, as it only learns one class, (c) most method are incremental, (d) it does not assume any distribution for the abnormal data.

In a review paper, Chandola et al. [2] pointed out that nearest neighbor and clustering-based methods may suffer with high-dimension datasets, since distance measures are not able to differentiate between normal and anomalous instances. Classification-based techniques can be a better choice in this scenario, but they require sufficient labeled data from both normal and abnormal classes. Statistical techniques are effective when both the dimensionality of data is low and the statistical assumptions are satisfied.

In this paper we propose a framework for anomaly detection based on convex hulls in a distribution parameter space (instead of a feature space). The convex hull is computed over pairs of estimates computed with normal data. We assume the univariate Gaussian distribution with two parameters, but the method can be employed with any distribution and high-dimensional data.

2 Anomaly Detection Based on Convex Hulls on a (μ, σ) Parameter Space

The training stage of our method consists in computing a set of estimates $\widehat{\theta}$ of parameters of a given distribution, which will be termed “points” henceforth. The points are computed using randomly selected normal instances of much smaller size than the complete sample, for instance, assuming a Gaussian distribution, $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma})$, a 2-dimensional parameter space is formed by the points.

In order to obtain representative points of the parameter space, every pair of feature vectors are concatenated (Fig. 1). After that, a convex hull H_N is

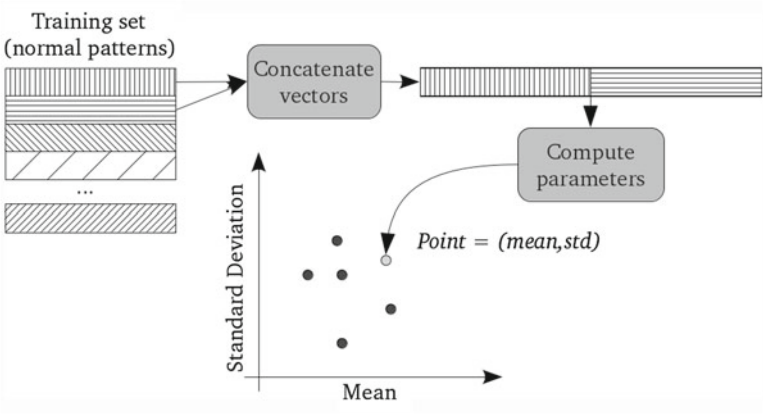


Fig. 1. Pairwise parameter estimation by concatenation of pairs of feature vectors

computed over the points, obtaining a geometric interpretation of the estimates produced by the normal data.

New instances are classified by computing a new set of estimate points; this time the points are obtained by pairing the new unknown instance with each point inside the normal convex hull, H_N. A new convex hull H_U is computed over this new set of points. Avoiding the points in the interior of H_N dramatically reduces the computational time of the second stage.

The detection algorithm uses the intersection of the convex hulls $H_N \cap H_U$. Assuming that the new pattern comes from the same law that produced the normal set of data, the intersection is expected to be high. If an anomaly is observed, as H_N was formed by the set of points obtained by pairs of normal observations, while H_U was obtained by merging the new pattern with only normal observations, the intersection will be reduced.

The algorithm is defined by the following steps:

1. **Normal class parameter estimation**, estimate $\widehat{\Theta}$, which is the set of parameters for normal data pairs (Algorithm 1 lines 1–6);
2. **New observation parameter estimation**, estimate $\widetilde{\Theta}$, which is the set of parameters for normal data pairs computed with the unknown data (Algorithm 1 lines 7–13);
3. **Convex hull intersection** is computed as a way to measure the perturbation inserted by the new pattern (Algorithm 1 line 14);
4. **Detection**: use the difference of the size of the intersection I in comparison to both convex hulls, H_N and H_U (Algorithm 1 lines 15–20).

Although the method uses an univariate model, it does not try to describe the data itself, but the relationship between samples, and capture perturbations when outliers are paired with expected-pattern data. For this reason we believe our algorithm is capable of deal with both univariate and mutivariate data.

Algorithm 1 Gaussian parameter space anomaly detection

```
estimate $\hat{\Theta} = \left\{\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \dots\right\}$
for each pair $i$ of observations $a, b$ with $a \neq b$ do
    $v_i \leftarrow$ concatenate $a$ and $b$
    estimate $\hat{\theta_i} = (\hat{\mu_i}, \hat{\sigma_i})$ over $v_i$
end for
compute $H_N$ using $\hat{\Theta}$
compute threshold $T$ using a validation set
for each new pattern $x$ do
    estimate $\widetilde{\Theta} = \left\{\widetilde{\theta}^{(1)}, \widetilde{\theta}^{(2)}, \dots\right\}$ with:
    for each point $y \in H_N$ do
        $c_i \leftarrow$ concatenate $y$ and $x$
        estimate $\widetilde{\theta_i} = (\widetilde{\mu_i}, \widetilde{\sigma_i})$ over $c_i$
    end for
    compute $H_U$ using $\widetilde{\Theta}$
    $I \leftarrow (H_N \cap H_U)$
    $d \leftarrow (H_N \setminus I) + (H_U \setminus I)$
    if $d < T$ then
        $x$ is considered normal
    else
        $x$ is considered an anomaly
    end if
end for
```

3 Experiments

In this paper we assume the Gaussian distribution for the "normal" data (no assumption is made by our algorithm for the anomalies). Therefore, our parameter space is $\mathbb{R} \times \mathbb{R}_+$.

A repeated random subsampling validation is used to compute the results, each experiment was repeated ten times, using 70 % for training, 15 % for validation and 15 % for testing. The mean and standard deviation are computed by these repetitions. The evaluation is based on an balanced accuracy value that takes into account the balance between the classes:

$$\text{Acc} = 1 - \frac{\sum_{i=1}^{c} E(i)}{2c},$$

where c is the number of classes, and $E(i) = e_{i,1} + e_{i,2}$ is the partial error of c, computed by:

$$e_{i,1} = \frac{FP(i)}{N - N(i)} \quad \text{and} \quad e_{i,2} = \frac{FN(i)}{N(i)}, i = 1, \dots, c,$$

where $FN(i)$ (false negatives) is the number of samples belonging to i incorrectly classified as belonging to other classes, and $FP(i)$ (false positives) is the number of samples $j \neq i$ that were assigned to i [11].

In order to find a threshold for d, c.f. step 4 of Algorithm 1, able to identify if a new observation is an anomaly, we used a validation set composed by 15 % of samples. Those normal validation samples were not used neither in the training nor the test stages. However, the value found for the threshold is used further to test the unseen instances.

The methods used to compare against our results are: the univariate and multivariate maximum likelihood classifiers (MLC) in their Anomaly Detection form [1], and the Naïve Bayes classifier [3], both under the Gaussian assumption. We also tested the One-class SVM which, instead of assuming a distribution for the observation process, fits a hyper-sphere to the data in order to describe it in a high dimensional space.

The experimental settings for each algorithm were:

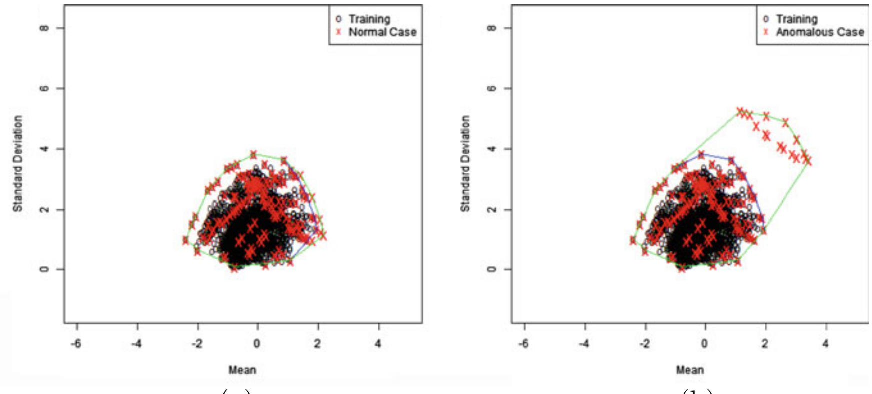
- **Univariate and multivariate MLC**: were trained with 70 % of the normal data, and the threshold for the probability obtained by using 15 % of both normal and anomalous data. The test step used 15 % of the dataset, including both normal and anomalous samples.
- **Naïve Bayes classifier**: was trained with 70 % of the normal data and 5 % of the anomalous data, as this is the proportion of available abnormal samples often found on datasets [2]. The test step used 15 % of the dataset, including both normal and anomalous samples.
- **One-class SVM**: same settings as the Naïve Bayes. A grid search with step 0.25 was performed to tune the parameters, using an evaluation dataset with 10 % of the samples.
- **Convex Hull on parameter space (Conv.Hull-PS)**: trained with 70 % of the normal data, and the threshold (difference between convex hulls) d obtained by using 15 % of normal samples. The test step used 15 % of the dataset, including both normal and anomalous samples.

Detailed information about the datasets used in the experiments are shown in Table 1, including synthetic and real data:

- **Normal-vs-2 distributions**: one hundred samples from the Gaussian law were simulated to describe the "normal" class. Another twelve samples were generated using a Lithuanian distribution (six samples) and a Banana-shaped distribution (six samples), to be used as anomalies.
- **Ionosphere**: Ionosphere data from UCI Machine Learning Depository [5]. Consists of a phased array of 16 high-frequency antennas that transmitted

Table 1. Dataset characteristics

Dataset	Type	#Samples	#Features	Anomaly rate (%)
Normal-vs-2	synthetic	112	2	10.7
Ionosphere	real	193	34	23.8
Parkinsons	real	351	23	36.0
BreastR	real	147	12	17.0
BreastW	real	699	9	34.5



(a) (b)

Fig. 2. Examples of convex hulls on parameter spaces using the dataset Normal-vs-2. The green line shows H_U and the blue line H_N. A normal sample detection is shown in (a) and an anomaly detection shown in (b).

around 6.4 kW. "Normal" radars are those that showed evidence of some type of structure in the ionosphere; "anomalous" radars are those whose signals pass through the ionosphere and, therefore, do not show any structure [12]. This dataset is composed by 351 instances of which 225 are considered normal and 126 abnormal. Every instace has 34 continuous attributes.

- **Parkinson**: a dataset of recorded speech signals from the UCI Machine Learning Depository [5]. The original study published the feature extraction methods for general voice disorders. It is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson disease [8].
- **BreastR**: breast cancer from the University of Roma "Tor Vergata". This dataset consists of 127 cases of benign tumors and 25 cases of malignant tumors. Each instance has 12 attributes obtained through the use of Gabor Wavelets [10].
- **BreastW**: Wisconsin Breast Cancer data [9] from the UCI Machine Learning Depository [5]. The dataset has 699 instances, each with with nine integer attributes; 458 of this instances are classified as benign, and 241 as malignant.

An example of the parameter space with pairwise estimated points, and convex hulls computed from a normal and an abnormal sample are shown in Fig. 2.

4 Results and Discussion

The sample mean and standard deviation of the balanced accuracy are shown in Table 2, with the best results highlighted in boldface. The comparison among results is made with respect to both the mean and the standard deviation of the balanced accuracy.

Table 2. Balanced accuracy: mean and standard deviation

Dataset	Normal-Univ (%)	Normal-Multiv (%)	Bayes Class. (%)	One-class SVM (%)	Conv.Hull-PS (%)
Normal-vs-2	$\mathbf{93.8 \pm 4.9}$	$\mathbf{95.2 \pm 1.4}$	50.0 ± 0.0	80.1 ± 0.1	$\mathbf{93.9 \pm 1.5}$
Ionosphere	71.2 ± 5.2	$\mathbf{78.2 \pm 2.4}$	$\mathbf{78.5 \pm 8.6}$	72.0 ± 0.1	$\mathbf{77.2 \pm 2.7}$
Parkinsons	62.9 ± 6.5	63.9 ± 5.2	54.9 ± 6.4	$\mathbf{67.3 \pm 0.2}$	$\mathbf{65.5 \pm 5.6}$
BreastR	67.2 ± 7.5	$\mathbf{70.1 \pm 6.6}$	50.0 ± 0.0	60.6 ± 0.7	52.1 ± 2.9
BreastW	$\mathbf{94.1 \pm 2.3}$	92.6 ± 2.0	85.1 ± 1.6	90.8 ± 0.5	$\mathbf{93.1 \pm 1.8}$

The proposed method achieved results comparable or better than the competing methods, except for the BreastR dataset. The average performance for this database is probably due to the high overlap rate presented by the normal and anomalous cases, which cannot be captured by the distribution model. This overlap rate makes it hard to identify the anomalies, since they are spatially mixed with the normal cases. The One-class SVM was better probably because it fits a hypersphere in a higher dimensional space. However, it is possible that by changing the parameter space used to classify the instances, the results presented by the proposed method can be significantly improved. Moreover, the results presented by the other datasets show that using a parameter space can help on an anomaly detection task when only normal samples are available.

The Naïve Bayes classifier suffered from the scarce anomaly data available, and it classified erroneously most anomalies. The statistical methods shows results comparable with our proposal, but they may not deal well with higher dimensional datasets, while our algorithm is more likely to be robust in such cases. The results supports this claim, specially by the results obtained with Parkinsons and Ionosphere datasets. Also, our proposal does not depend on model assumptions; as noted by Frery et al. [6], this is a major issue when using model-based classification techniques.

5 Conclusions

This paper reports results of a new anomaly detection framework based only on normal class samples. The interesting feature of this framework is that it can handle any probability distribution or mixture of distributions. It also includes all the advantages of partially supervised algorithms. Besides, it is possible to include information about anomalies in the validation step. The parameter space allows to specify parameters that better model the problem of finding anomalies. The use of a convex hull makes it possible to draw the boundary between normal and abnormal data behavior. Future works will explore variations by including the use of multiple parameters, exploring both the use of anomalous data in the training step.

Acknowledgment. This work was supported by FAPESP (grants n. 2011/16411-4 and n. 2012/12524-1).

References

1. Barnett, V., Lewis, T.: Outliers in Statistical Data. Wiley, New York (1994)
2. Chandola, V., Banerjee, A., Kumar, A.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 15 (2009)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, Chichester (2000)
4. Fawcett, T., Provost, F.J.: Activity monitoring: noticing interesting changes in behavior. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 53–62 (1999)
5. Frank, A., Asuncion, A.: UCI Machine Learning Repository. http://archive.ics.uci.edu/ml (2010)
6. Frery, A.C., Ferrero, S., Bustos, O.H.: The influence of training errors, context and number of bands in the accuracy of image classification. Int. J. Remote Sens. **30**(6), 1425–1440 (2009)
7. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artif. Intell. Rev. **22**(2), 85–126 (2004)
8. Little, M., McSharry, P., Roberts, S., Costello, D., Moroz, I.: Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Biomed. Eng. Online **6**, 23 (2007)
9. Mangasarian, O.L., Wolberg, W.H.: Cancer diagnosis via linear programming. SIAM News **23**(5), 1–18 (1990)
10. Mencattini, A., Salmeri, M., Casti, P., Pepe, M.L., Mangieri, F., Ancona, A.: Local active contour models and gabor wavelets for an optimal breast region segmentation. Comput. Assist. Radiol. Surg. (CARS 12) (2012)
11. Ponti, M.P.: Segmentation of low-cost remote sensing images combining vegetation indices and mean shift. Geosci. Remote Sens. Lett. IEEE **10**(1), 67–70 (2013)
12. Sigillito, V.G., Wing, S.P., Hutton, L.V., Baker, K.B.: Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Tech. Dig. **10**, 262–266 (1989)

Self-Practice Imitation Learning from Weak Policy

Qing Da, Yang Yu(✉), and Zhi-Hua Zhou

National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing 210046, China
{daq, yuy, zhouzh}@lamda.nju.edu.cn

Abstract. Imitation learning is an effective strategy to reinforcement learning, which avoids the delayed reward problem by learning from mentor-demonstrated trajectories. A limitation for imitation learning is that collecting sufficient qualified demonstrations is quite expensive. In this work, we study how an agent can automatically improve its performance from a weak policy, by automatically acquiring more demonstrations for learning. We propose the LEWE framework to sample tasks for the weak policy to execute, and then learn from the successful trajectories to achieve an improvement. As the sampling strategy is the key to the efficiency of LEWE, we further propose to incorporate active learning for the sampling strategy for LEWE. Experiments in a spatial positioning task show that LEWE with active learning can effectively and efficiently improve the weak policy and achieves a better performance than the comparing sampling approaches.

Keywords: Imitation learning · Active sampling

1 Introduction

In traditional reinforcement learning, an agent receives a reward after a sequential of actions, and tries to figure out the best actions according to the received rewards [1]. The high latency of the reward feedback forces the agent to search in a very large state space, which is of low efficiency and low effectiveness. Motivated by the teaching process in human society, imitation learning [2], alleviates this difficulty by introducing an expert mentor. The mentor provides demonstrations that successfully accomplish a specific task, and then the agent learns to follow the demonstrations that provide much more guiding information.

Imitation learning has drawn many attentions, and several approaches have been proposed (see [3] for a survey). These approaches roughly fall into two categories, inverse reinforcement learning and direct policy learning. In the first category, the aim is to learn a reward function, which associates actions with values, from the demonstration data instead of the delayed rewards, and the reward function is then used to derive a policy by reinforcement learning approaches [4,5]. In the second category, demonstration trajectories are used to directly

Z.-H. Zhou and F. Schwenker (Eds.): PSL 2013, LNAI 8183, pp. 9–20, 2013.
DOI: 10.1007/978-3-642-40705-5_2,
© Springer-Verlag Berlin Heidelberg 2013

learn a mapping function from agent's state observations to action [6,7]. In this work, we focus on the second category for deriving policies directly.

Direct policy learning requires sufficient demonstrations from an expert mentor, so it has several limitations in real-world situations. First, it is usually quite expensive to collect demonstrations from expert mentors; even if we are free from the budget problem, expert mentors may also produce suboptimal demonstrations; and in real-world reinforcement learning problems, the state space can be extremely large so that it is hard to have enough demonstrations for a good learning.

In this work, in order to tackle the problem of insufficient demonstration, we investigate approaches that an agent can automatically acquire more demonstrations to improve itself. We propose the LEWE (LEarning from WEak policy) framework that automatically improves a *weak policy*. The main idea of LEWE is, for a few tasks, the weak policy can lead to successful trajectories, which can serve as good examples for learning an improved policy. Therefore, LEWE samples the task space and runs the weak policy on the sampled tasks in order to acquire successful trajectories. The sampling strategy is a key to the efficiency of LEWE. We therefore further propose three sampling strategies utilizing different information.

We experiment the proposed approaches in a spatial positioning task. Experiment results verify the effectiveness of LEWE by showing that LEWE significantly improves the weak policy with various configurations. Moreover, the results also show that the proposed active sampling leads to a better performance than other sampling strategies.

The rest of this paper starts with an introduction of the background, and then presents the LEWE framework, which is followed by the experiment results, and ends with a section of conclusion.

2 Background

A Markov Decision Process (MDP) can be represented by a 4-tuple (S, A, T, R), where S is a set of states, A is a set of actions, $T(s_j|s_i, a) : S \times A \times S \rightarrow \mathbb{R}$ is the transition probability of reaching state s_j from state s_i after executing a, and $R(s, a) : S \times A \rightarrow \mathbb{R}$ defines the immediate reward by executing a from state s_i. Given an MDP, the goal is to derive a policy π that maps from S to A maximizing a long term reward. Imitation learning derives a policy by learning from a set of successful demonstration trajectories $D = \{d_i = (s_{it}, a_{it})_{t=0}^{T_i}\}_{i=1}^{n}$ without using explicit reward functions.

One branch of imitation learning research focuses on forming the reward function, also named *inverse reinforcement learning* (IRL) [8]. This method assumes that there exists a latent reward function R of a given task and the actual intention of the demonstrations is to maximize the expected discounted reward over time. Based on this assumption, IRL learns a reward function from the trajectories data and then use conventional reinforcement learning algorithms to derive a policy, such as methods in [4,5,9].

Another branch alternatively focuses on learning a mapping function directly from demonstrated trajectories, also named *direct policy learning*. These methods directly learn a policy π from D by taking the state-action pairs (s_{it}, a_{it}) collected from demonstrations as training examples, using supervised learning algorithms, such as kNN [10], local weighted learning [11], decision trees [12], etc.

Active learning, also referred as *query learning* or *optimal experimental design* in the statistics literature, is a subfield of machine learning, which only queries the labels of useful instances so that it achieves a good performance with a small amount of examples [13]. Some popular approaches for active learning are [13,14].

3 The Proposed Approach

Our aim is to train an agent to accomplish a class of tasks. We assume without loss of generality that a task is parameterized by a vector $\boldsymbol{p} = [p_1, p_2, ..., p_C]^T$. The i-th parameter is constrained by the range L_i respectively, i.e., $p_i \in L_i = [a_i, b_i]$. Then $\mathcal{L} = L_1 \times L_2 \times \ldots \times L_C \in R^C$ is the *task space*, and C is the dimensionality of $\mathcal{L}$. A weak policy then can be defined as a policy which is only able to successfully accomplish tasks in $U \subseteq \mathcal{L}$ with a small volume ratio $\frac{|U|}{|\mathcal{L}|}$, in the given time. It is relatively easy to obtain a weak policy in practice, either from an insufficient imitation learning, or from hand-written rules. Our goal is to make an agent improve itself starting from a weak policy, such that it will be able to accomplish unseen tasks.

3.1 The LEWE Framework

We propose the *LEarning from WEak policy* (LEWE) framework that outlines the self-improve procedure for an agent, as shown in Algorithm 1.

A weak policy π_0, the task space $\mathcal{L}$, the maximum number of sampled tasks N and the maximum number of iterations in a run T are provided as the inputs of the framework. The framework consists of two phases, sampling and learning. In the sampling phase, LEWE first invokes *TaskSampling* to generate a task in line 3, and then executes the weak policy for the task at most T steps, in lines 4 to 10. In each step of executing a task, the agent queries an action a from the weak policy π_0 on the current state s in line 6, then the agent executes a and updates its state as in line 7, where *execute* returns the state of the agent after taking the action. A task is considered to be accomplished successfully if the goal state is reached (determined by the function *TaskFinished*) within T steps. If the task is accomplished successfully, LEWE records the trajectory as a successful demonstration, and at the same time, records the task parameters as a successful task or a failed task, in lines 11 to 17. In the learning phase, an improved policy π_* is learned from the recorded data D through direct policy learning. Note that, although one can easily take the improved policy as a weak policy and use LEWE to refine it again, we rather choose to stay studying the effectiveness of this simple framework.

Algorithm 1 The LEWE Framework

```
Input:
    A weak policy $\pi_0$
    Task space $\mathcal{L}$
    The maximum number of sampled tasks $N$
    The maximum number of iterations $T$
Output:
    The learned policy $\pi_*$
$N_e \leftarrow 0$, $D \leftarrow \varnothing$, $P_{suc} \leftarrow \varnothing$, $P_{fail} \leftarrow \varnothing$
while $N_e < N$
    $\boldsymbol{p} = TaskSampling(\mathcal{L}, P_{suc}, P_{fail})$
    $s \leftarrow s_0$, $t \leftarrow 0$, $Demo \leftarrow \varnothing$
    while $!TaskFinished(\boldsymbol{p}, s)$ or $t < T$
        $a \leftarrow \pi_0(s)$
        $s \leftarrow execute(\boldsymbol{p}, s, a)$
        $Demo \leftarrow Demo \cup \{(s, a)\}$
        $t \leftarrow t + 1$
    end
    if $TaskFinished(\boldsymbol{p}, s)$
        $D \leftarrow D \cup \{Demo\}$
        $P_{suc} \leftarrow P_{suc} \cup \{\boldsymbol{p}\}$
    else
        $P_{fail} \leftarrow P_{fail} \cup \{\boldsymbol{p}\}$
    end
    $N_e \leftarrow N_e + 1$
end
$\pi_* \leftarrow directPolicyLearning(D)$
```

There are two important issues in order to achieve a successful application of LEWE framework. In LEWE framework, we expect that, by learning the successful demonstrations of a weak policy, an improved policy can be obtained via the generalization ability of the employed learning algorithm. Therefore, the generalization ability of the learning algorithm is important. The other issue is the sampling strategy that implements the *TaskSampling* function. We investigate three sampling strategies, *random sampling*, *evolutionary sampling* and *active sampling* in the follows.

3.2 Random Sampling

Random sampling simply selects a task $\boldsymbol{p}$ from the task space $\mathcal{L}$ uniformly at random. This strategy serves as the baseline approach, and is denoted as $TaskSampling_r$ function. A drawback of random sampling is that it is high likely to sample tasks that are too hard for a weak policy, which will waste a lot of time and resource.

3.3 Evolutionary Sampling

Evolutionary sampling is an improvement over random sampling on the basis of a simple observation: a weak policy will be more likely to succeed on tasks similar to the succeeded tasks in history rather than on totally strange tasks. Evolutionary sampling uses the mutation operator of evolutionary strategy algorithms [15] to generate new tasks by perturbating the succeeded tasks. Instead of sampling a task from $\mathcal{L}$ uniformly, with probability ρ, evolutionary sampling selects a succeeded task, and generates a new task by perturbating this task as following

$$p_i \leftarrow q_i + \mathcal{N}(0, \epsilon_i^2), i = 1, 2, ..., C \qquad (1)$$

where $\mathcal{N}(0, \epsilon_i^2)$ is a normal distribution with mean zero and standard deviation ϵ_i. With the remaining $1 - \rho$ probability, it does the random sampling. The probability ρ can be regarded as a trade-off between the possibility of success of the new task and the overall exploration in the task space.

3.4 Active Sampling

The evolutionary sampling, however, may not try best to exploit the whole task space, moreover, is only aware of the succeeded tasks but not the failed tasks. Ideally, we want to sample a task that

1. will result a successful trajectory with confidence.
2. is dissimilar to the past tasks as much as possible.
3. can produce the states dissimilar to the past collected states as much as possible.

The first property is with the same idea of evolutionary sampling. The second one is based on the assumption that different tasks tend to produce different states. The other one directly seeks such a task by estimating the distribution of possible states for a given task.

Based on this idea, we propose an active sampling strategy incorporating the active learning [14] idea. For an unseen task, we employ an SVM model to estimate the confidence that the weak policy will be successful, a Gaussian Mixture Model to estimate the distance to the explored area in task space, and the likelihood that unseen states will be visited. The implementation details of these three components are introduced as the follows.

To estimate the probability that a task can be successfully executed by the weak policy, we train a classifier to distinguish the succeeded tasks from the failed tasks. We combine the records P_{suc} and P_{fail} to get a training set $P_{train} = (\boldsymbol{p}_j, y_j)_{j=1}^n$, where n is the total number of recorded tasks, and y_j is the label of task $\boldsymbol{p}_j$ and is assigned to 1 for $\boldsymbol{p}_j \in P_{suc}$ and -1 otherwise. Then a standard SVM classifier $f(\boldsymbol{p}) = \mathrm{w}^T \Phi(\boldsymbol{p})$ is trained from P_{train}, where Φ is a function mapping $\boldsymbol{p}$ to a high-dimension space that appears implicitly in the kernel function $K(\boldsymbol{p}_1, \boldsymbol{p}_2) = \langle \Phi(\boldsymbol{p}_1), \Phi(\boldsymbol{p}_2) \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the inner product. For any unseen task $\boldsymbol{p}$, the higher $f(\boldsymbol{p})$ is, the higher probability that task $\boldsymbol{p}$ can be successfully executed by the weak policy.

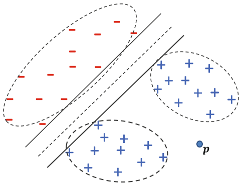


Fig. 1. An illustrative example of combining a SVM classifier and a GMM density estimator in task space

To estimate how far a task is from the past tasks, we use Gaussian Mixture Models (GMM) to estimate the visited area in task space. We approximate a distribution of task $\boldsymbol{p}$ belonging to the area we have explored with a GMM as

$$\Pr(\boldsymbol{p}|\tau_k, \mu_k, \Sigma_k) = \sum_{k=1}^{K} \tau_k \mathcal{N}(\boldsymbol{p}, \mu_k, \Sigma_k) \tag{2}$$

where K is number of components of this model, and $\{\tau_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ are model parameters to be estimated. We estimate these parameters through the Expectation Maximization (EM) algorithm from data $P_{suc} \cup P_{fail}$. For any unknown task $\boldsymbol{p}$, the lower $\Pr(\boldsymbol{p})$ is, the higher probability that the surrounding of task $\boldsymbol{p}$ has not been explored yet. We give an illustration of above discussion in Fig. 1. By the SVM, we separate the succeeded and failed tasks, so that we can estimate the task $\boldsymbol{p}$ has a high probability to be successfully executed by the weak policy, since it is far from the failed tasks. By the GMM, we can estimate that the task $\boldsymbol{p}$ is not in the explored area of the task space. Further, to estimate how far all possible states produced by a task from the past collected states, we firstly build the relationship between a task and all the states it produces. For task $\boldsymbol{p}$, all states $\{s_i\}_{\boldsymbol{p}}$ produced by this task are assumed to be generated from a linear model

$$s^T = \boldsymbol{p}^T B + \boldsymbol{\epsilon}^T \tag{3}$$

here $B \in R^{C \times d}$ is a coefficient matrix, C is the dimensionality of task space and d is that of state space. $\boldsymbol{\epsilon}$ is a noise sampled from $\mathcal{N}(0, \Sigma_{\boldsymbol{\epsilon}})$.

Denote M as the number of states collected so far, and $\hat{S} = [s_1, s_2, ..., s_M]^T \in R^{M \times d}$ are all the historical states, and $\hat{P} = [p_1, p_2, ..., p_M]^T \in R^{M \times C}$ are the corresponding tasks that produced the states. Then the likelihood function can be represented as

$$p(\hat{S}|\hat{P}, B, \Sigma_{\boldsymbol{\epsilon}}) \propto |\Sigma_{\boldsymbol{\epsilon}}|^{-\frac{M}{2}} exp(-\frac{1}{2} tr((\hat{S} - \hat{P}B)^T \Sigma_{\boldsymbol{\epsilon}}^{-1} (\hat{S}^T - \hat{P}B))) \tag{4}$$

Since we don't have any prior knowledge of the distribution of the parameter, we apply the classical frequentist least squares solution to estimate B using moore-penrose pseudoinverse

$$B = (\hat{P}^T \hat{P})^{-1} \hat{P}^T \hat{S} \tag{5}$$

So for any unknown task $\boldsymbol{p}$, the probability that a history state s_i can be produced by $\boldsymbol{p}$ is $p(s_i|\boldsymbol{p})$, we can now estimate the improvement of the diversity in state space that the task $\boldsymbol{p}$ brings by calculating the probability of historical states being produced by $\boldsymbol{p}$

$$h(\boldsymbol{p}) = \prod_{i=1}^{M} p(s_i|\boldsymbol{p}) \tag{6}$$

For any unknown task $\boldsymbol{p}$, the lower $h(\boldsymbol{p})$ is, the higher probability that task $\boldsymbol{p}$ can produce more unobserved states.

Considering all three components, active sampling strategy uses the following objective function

$$g(\boldsymbol{p}) = -\mathrm{w}^T \varPhi(\boldsymbol{p}) + c\sum_{k=1}^{K} \tau_k \mathcal{N}(\boldsymbol{p}, \mu_k, \varSigma_k) + \lambda \prod_{i=1}^{M} p(s_i|\boldsymbol{p}) \tag{7}$$

where c and λ are the trade-off coefficient. To find a task $\boldsymbol{p}^*$ that minimizes the objective function, we first employ the random sampling to generate a pool of tasks $P_{candidate}$, then we find $\boldsymbol{p}^*$ from the pool that

$$\boldsymbol{p}^* = \underset{\boldsymbol{p} \in P_{candidate}}{\arg\min}\ g(\boldsymbol{p}) \tag{8}$$

4 Experiment

We investigate three questions by experiments:

1. Can LEWE improve a weak policy?
2. Is active sampling better than the other two sampling strategies?
3. Is the generalization ability of the direct policy learning algorithm important to the performance of LEWE?

4.1 Spatial Positioning Task

We use a *positioning with heading* problem studied in [16] to validate our algorithms. This task consists of attaining a 2D planar target position with a target heading (x_g, y_g, θ_g) from the origin $(0, 0, 0)$, as shown in Fig. 2.

For this problem, the task parameters are the target position and the target heading, i.e., (x_g, y_g, θ_g). The corresponding task space is $\mathcal{L} = L_1 \times L_2 \times L_3 = [3m, 8m] \times [3m, 8m] \times [0, \frac{2\pi}{3} rad]$. For this class of tasks, some tasks are easy to perform (for example, task $(0, 3, 0)$ can be done by directly moving forward by 3 meters), while some other tasks are relatively hard such like task $(3, 3, \frac{2\pi}{3})$.

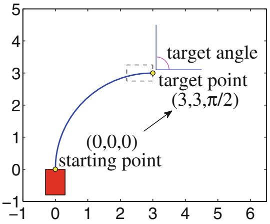


Fig. 2. The experimented task

4.2 The Weak Policy

We use a simple greedy-based method with a customized cost function R_0 as the weak policy π_0. This policy chooses the action that minimizes the expected cost of being executed at time t

$$\pi_0(s_t) = \underset{a_k}{argmin}\, \mathbf{E}[R_0(execute(s_t, a_k))] \tag{9}$$

The cost function R_0 is defined as a weighted summation of the axis distance between the agent and target position, absolute difference between the agent heading and the target heading and the distance to the *optimal line* (the line that passes through the target position (x_g, y_g) with the slope $tan(\theta_g)$), as

$$\begin{aligned} R_0(s_t) = &w_1|x_t - x_g| + w_2|y_t - y_g| + w_3|\theta_t - \theta_g| \\ &+ w_4 \frac{|Ax_t + By_t + C|}{\sqrt{A^2 + B^2 + C^2}} \end{aligned} \tag{10}$$

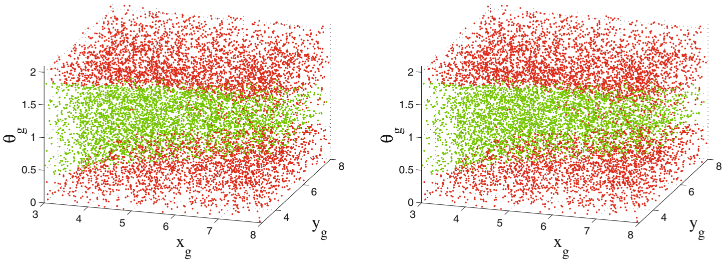
Here, A, B, C are the equations coefficients of *optimal line* and the weights used in the experiments are $w_1 = 0.1$, $w_2 = 0.1$, $w_3 = 0.15$ and $w_4 = 1$.

4.3 Experiment Setup

To measure the performance of our algorithms, policies are evaluated for success rate of task performing. A task is successfully executed if and only if the agent steps into $0.1\,m$ range of the target position and within $0.1\,rad$ error to the target angle, i.e., $\|x_t - x_g, y_t - y_g\| < 0.1\,m$ and $|\theta_t - \theta_g| < 0.1\,rad$, which can be considered as a more strict criterion comparing to the existing work (e.g. [16]).

In our experiments, each policy is tested with a test set containing 200 tasks uniformly sampled from $\mathcal{L}$. We employ 4 start-of-the-art supervise learning algorithms as the direct policy learning methods, i.e., kNN [17], decision tree [18] and random forest [19] implemented in WEKA [20] and SVM [21] in LIBSVM [22]. The parameters of the four classifiers are

- kNN: $k = 7$ (obtained by cross validation)
- Decision tree: Default settings of WEKA
- Random forest: Default settings of WEKA with ensemble size 100
- SVM: Radial Basis Function (RBF) kernel with $\gamma = 0.01$ and C is 200.



(a) real distribution (b) classification result by SVM

Fig. 3. The distribution of successful and failed tasks in tasks space

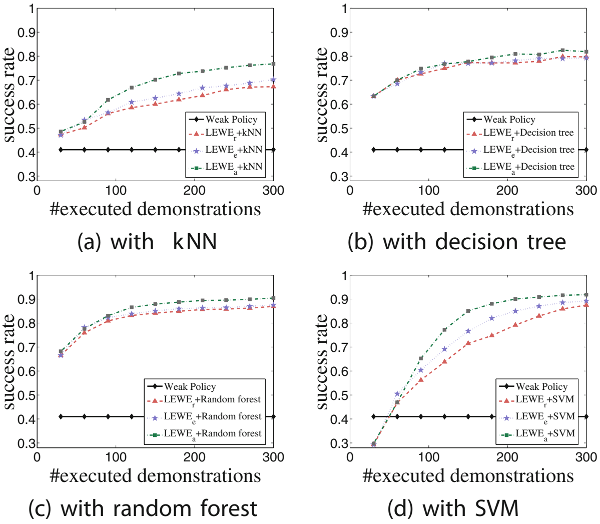
For task (x_g, y_g, θ_g), we extract 11 arbitrary geometry features from state (x, y, θ), including $x - x_g$, $y - y_g$, $\theta_g - \theta$, $\alpha - \theta$ (α is the angle of $(x_g - x, y_g - y)$), $\|x - x_g, y - y_g\|$ and some other geometry features.

We denote LEWE_r, LEWE_e and LEWE_a as LEWE with random sampling, evolutionary sampling and active sampling, respectively. For active sampling strategy, we also use LIBSVM as the SVM solver. Note that for both evolutionary sampling and active sampling strategies, there's no historical data in the very beginning, thus, random sampling strategy is used to sample first 30 successful demonstrations for both cases. The size of candidate set for active sampling is 100. The parameters ρ, c and λ are selected by cross validation. Finally, we run every configuration 200 times and report the mean success rate.

4.4 Experiment Results

We firstly investigate how the weak policy acts for this class of tasks. We randomly sample 10000 tasks from $\mathcal{L}$ and apply the weak policy to perform on these tasks. Figure 3a shows the task space, where the green points are the tasks accomplished (about 41 %) and the red ones are those not accomplished by the weak policy. It is easy to observe that the successful tasks are separated well from failed ones. Moreover, we apply a standard SVM classifier (with RBF kernel, C is 200 and γ is 0.1) to classify these tasks with a training set consists of only 75 randomly chosen tasks (30 successful v.s. 45 failed). The classification result is shown in Fig. 3b with an accuracy rate of about 88 %, which implies the learnability of easy and hard tasks.

We then investigate the first question, i.e., can LEWE improve the weak policy. Figure 4 plots the success rates against the number of executed demonstrations ranged from 30 to 300, for the weak policy as well as LEWE with the four learning algorithms and the three sampling strategies. We first observe that LEWE has a higher success rate than the weak policy, only except when using SVM with less than 60 demonstrations. It can also be observed that the success rate of LEWE increases as the number of the executed demonstrations increases. When 300 demonstrations are executed, LEWE achieves at least 0.67 success rate



(a) with kNN

(b) with decision tree

(c) with random forest

(d) with SVM

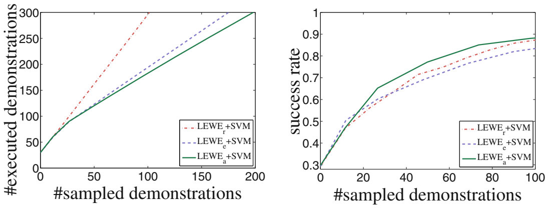
Fig. 4. The success rates of LEWE against the weak policy

Table 1. Success rates of LEWE after 300 executed demonstrations.

Algorithm	LEWE$_r$	LEWE$_e$	LEWE$_a$
*k*NN	0.67 ± 0.13	0.70 ± 0.11	**0.77 ± 0.06**
Decision tree	0.80 ± 0.11	0.79 ± 0.11	**0.82 ± 0.11**
Random forest	0.87 ± 0.05	0.88 ± 0.05	**0.90 ± 0.04**
SVM	0.87 ± 0.09	0.89 ± 0.07	**0.92 ± 0.03**

and as high as 0.92 success rate while that of weak policy is 0.41. Therefore, it is clear that LEWE effectively improves the weak policy.

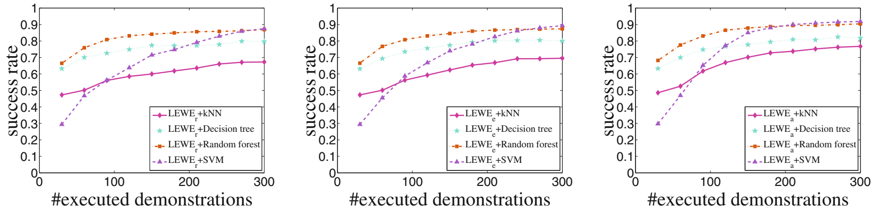
For the second question, i.e., is the proposed active sampling better than the other, we look into Fig. 4, where each plot also compares the three sampling strategy with a policy learning algorithm. The parameter selected for ρ, c and λ are 0.6, 2.5 and 1.0. For active sampling, the GMM with only one component in the task space shows to be the best in this case, which in fact degrades into a Gaussian distribution. It can be observed that the curves of active sampling are almost always above the comparing curves, particularly when the number of executed demonstrations is large. Table 1 compares the success rates after 300 executed demonstrations, where it can be found that active sampling is significantly better than the compared approaches by the t-tests with 95 % confidence. To further investigate how the sampling strategies work, we plot the number of executed demonstrations against the sampled ones in Fig. 5a. It can be observed that random sampling executes more demonstrations than evolutionary and active sampling for sampling the same number of successful demonstrations, which confirms our design that evolutionary and active sampling treat the samples much smarter so that the waste of failed executions is fewer. Meanwhile from



(a)

(b)

Fig. 5. Effectiveness of sampling strategies



(a) with random sampling strategy

(b) with evolutionary sampling strategy

(c) with active sampling strategy

Fig. 6. The success rates of LEWE with different direct policy learning algorithms

Fig. 5b, it is clear that active sampling makes a better utilization of the samples than evolutionary sampling, which confirms our design that active learning utilizes full information. Finally, we investigate the third question, i.e., is generalization ability of the policy learning algorithm important. Figure 6 compares LEWE with the four learning algorithms with each sampling strategy. It can be observed that random forest always takes the most advantage except that SVM is comparable with random forest when there are 300 demonstrations. It is consistent with the experience that random forest and SVM are the two state-of-the-art supervised learning approaches [23]. We recommend using random forest as it has a more stable performance and fewer parameters to be tuned.

5 Conclusion

In this paper, we propose the LEWE framework for an agent to improve its performance from a weak policy using imitation learning. We incorporate active sampling for LEWE to smartly choose demonstrations to practice. Experiments verified the effectiveness of the LEWE framework as well as the active sampling strategy. The work verifies the possibility that an agent can acquire demonstrations to improve its performance by itself. The future work will combine the learning policy with other reinforcement learning approaches towards a better performance.

Acknowledgements. This research was supported by the Jiangsu Science Foundation (BK2012303), the 2013 State Grid Research Project, and the Baidu Fund (181315P00651).

References

1. Sutton, R., Barto, A.: Reinforcement Learning. An Introduction. Cambridge University Press, Cambridge (1998)
2. Schaal, S.: Is imitation learning the route to humanoid robots. Trends Cogn. Sci. **3**(6), 233–242 (1999)
3. Argall, B., Chernova, S., Veloso, M., Browning, B.: A survey of robot learning from demonstration. Rob. Auton. Syst. **57**(5), 469–483 (2009)
4. Atkeson, C., Schaal, S.: Robot learning from demonstration. In: Proceedings of the ICML'97, San Francisco, USA, pp. 12–20, July 1997
5. Choi, J., Kim, K.: Inverse reinforcement learning in partially observable environments. In: Proceedings of IJCAI'09, Barcelona, Spain, pp. 1028–1033, July 2009
6. Jetchev, N., Toussaint, M.: Task space retrieval using inverse feedback control. In: Proceedings of ICML'11, Bellevue, WA, USA, pp. 449–456, June 2011
7. Zhang, D., Cai, Z., Nebel, B.: Playing tetris using learning by imitation. In: Proceedings of GAMEON'10, Leicester, UK, pp. 23–27, November 2010
8. Ng, A., Russell, S.: Algorithms for inverse reinforcement learning. In: Proceedings of ICML'00, Stanford, USA, pp. 663–670, June 2000
9. Ziebart, B., Maas, A., Bagnell, J., Dey, A.: Maximum entropy inverse reinforcement learning. In: Proceedings of AAAI'08, Chicago, USA, pp. 1433–1438, July 2008
10. Bentivegna, D.: Learning from Observation Using Primitives. Ph.D. thesis, College of Computing, Georgia Institute of Technology (2011)
11. Bentivegna, D., Atkeson, C.: Learning from observation using primitives. In: Proceedings of ICRA'11, Seoul, Korea, pp. 1988–1993, May 2001
12. Silver, D., Bagnell, J., Stentz, A.: Perceptual interpretation for autonomous navigation through dynamic imitation learning. Robot. Res. **70**, 433–449 (2011)
13. Settles, B.: Active learning literature survey. Computer Sciences Technical Report, University of Wisconsin-Madison (2009)
14. Huang, S., Jin, R., Zhou, Z.: Active learning by querying informative and representative examples. In: NIPS'11, pp. 892–900 (2011)
15. Beyer, H., Schwefel, H.: Evolution strategies-a comprehensive introduction. Nat. Comput. **1**(1), 3–52 (2002)
16. Argall, B., Browning, B., Veloso, M.: Learning robot motion control with demonstration and advice-operators. In: Proceedings of IROS'08, Nice, France, pp. 399–404, September 2008
17. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2001)
18. Quinlan, J.: C4.5: Programs for machine learning. Morgan kaufmann, San Franscisco (1993)
19. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
20. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Franscisco (2005)
21. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (2000)
22. Chang, C., Lin, C.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syss. Technol. **2**(3), 27 (2011)
23. Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms. In: Proceedings of ICML'06, Pittsburgh, PE, pp. 161–168 (2006)

Semi-Supervised Dictionary Learning of Sparse Representations for Emotion Recognition

Markus Kächele and Friedhelm Schwenker(✉)

Institute of Neural Information Processing,
Ulm University, 89069 Ulm, Germany
{markus.kaechele, friedhelm.schwenker}@uni-ulm.de

Abstract. This work presents a technique for the classification of emotions in human-computer interaction. Based on biophysiological data, a dictionary learning approach is used to generate sparse representations of blood volume pulse signals. Such features are then used for classification of the current emotion. Unlabeled data, i.e. data without information about class membership, is used to enrich the dictionary learning stage. Superior representation abilities of the underlying structure of the data are demonstrated by the learnt dictionaries. As a result, classification rates are improved. Experimental validation in the form of different classification experiments is presented. The results are presented with a discussion about the benefits of the approach and the existing limitations.

1 Introduction

In classification tasks, the performance of a classifier mostly depends on the quality of the features and the amount of available training data. The problem structure at hand often prohibits extensive training because not enough valuable training data is available. Often it is the case that pure recording of data is effortlessly possible – the difficulties lie in correctly labeling that data. In the context of emotion recognition based on biophysiological data, it is very challenging to manually label data points for a particular emotion. What typically happens instead is that a whole sequence of data is labeled as a particular emotion in the hope that the labeled emotion is appropriately reflected by the captured data. This asymmetry of labeled against unlabeled data has led to the field of partially supervised learning. This field deals with ways to incorporate unlabeled data into supervised classification schemes to improve the overall performance. The other important aspect of the performance of a classification system is the quality of the features. Discovering valuable features and optimal combinations of them is a highly demanding step especially when the signals are of high dimension. Dimensionality reduction techniques can be used to tackle high dimensionality, but often fail to capture the inherent structure of analysed data. Many real world signals however can be described with much less information than the raw sensor signals indicate. The structure can be described by using only a *few significant coefficients* of a suitable basis. In this basis, a signal is said to be *sparse*. Images,

Z.-H. Zhou and F. Schwenker (Eds.): PSL 2013, LNAI 8183, pp. 21–35, 2013.
DOI: 10.1007/978-3-642-40705-5_3,
© Springer-Verlag Berlin Heidelberg 2013

for example have sparse representations in the wavelet domain. For arbitrary signals, the so called *dictionary learning* approaches can be used to create a sparsity basis. We propose an approach that combines dictionary learning of blood volume pulse signals to achieve sparse representations with the paradigm of partially super vised learning to improve the quality of the dictionary.

The remainder of this work is structured as follows: In Sect. 2 an overview of related work is given. Our approach is presented in Sect. 3. Experimental validation is provided in Sect. 4, which is followed by a discussion and an outlook on future work in Sect. 5. Finally, Sect. 6 closes the paper with the conclusion.

2 Related Work

Emotion recognition and affective computing are very active fields of research that received increased attention in recent years [31,36]. The trends in these research areas seem to transcend, on one hand, from datasets with lab-like environments and acted emotions (e.g. the extended Cohn-Kanade dataset of facial expressions [19] or Emo-DB, a database of German emotional speech [5]) to datasets with real emotions in unconstrained environments (e.g. the Emo-Rec II dataset [40] or the AVEC 2011 dataset [11,33]). On the other hand, researchers tend to abandon datasets with single modalities (e.g. only audio or video recordings or biopotentials such as skin conductivity, respiration, electromyography or electroencephalography [29]) in favor of multi-modal datasets that incorporate many different modalities.

Emotion recognition in such datasets is a very challenging task because unlike in controlled environments, emotions might only show as nuances of what they are supposed to be or sensors might provide unreliable information (e.g. face detection might fail, because the recorded person turns away). Sophisticated machine learning methods are necessary in order to extract valuable information from the different channels. For emotion recognition from speech many different approaches can be employed from generic ones like computing a broad range of features and selecting the most informative ones [34,35] to highly specialized approaches like the one in [32]. In that work, model parameter values for voice quality are synthesized and learned by an artificial neural network (ANN) in order to create a generic model of glottal flow. The resulting outputs of the ANN are then used for classification. For recognition of facial expressions, different approaches exist like fitting parametric models [19], computing salient distance features [43] or using one or several of the well known features in the literature such as local binary patterns, motion history histograms or optical flow [4,21]. Emotion recognition from biophysiological data is particularly interesting due to the fact that the emotional state as manifested in the autonomic and parasympathetic nervous systems is directly reflected in different measurable modalities across the human body [26]. While not as thoroughly investigated as facial expressions yet, approaches like [15,23,28] show the significance of biophysiological measurements.

The inherent multi-modality of many datasets permits the analysis of more than one of the mentioned channels and asks for methods of combination for the information gained from the different signal sources. A so called multiple classifier system (MCS) [17,37] can be used to combine the different signals. Information is extracted from the different modalities and this information can be combined at several stages in the classification pipeline (compare *early*, *midlevel* and *late fusion* [9]). Fusion of noisy, unreliable and variably dense sampled data can be done using time integration and rejection sampling [28], by a hierarchical fusion cascade [16] or using a so called *Markov fusion network* [10].

Another way to deal with challenging data, that is particularly interesting in scenarios in which classifiers can be enriched using unlabeled data is partially supervised learning (PSL). Partially supervised learning is a paradigm in which labeled data is combined with unlabeled data to improve purely supervised methods. PSL is motivated by the fact that labeling data is difficult because it is time consuming and/or expert knowledge is needed. On the other hand, capturing large amounts of data is relatively effortless. Partially supervised learning deals with incorporating huge amounts of unlabeled data with few labeled samples. Some of the approaches that have already been proposed are *self training* [42] and *co-training* [3], which train classifiers on the labeled data, let the trained classifiers label the unlabeled data points, add them to the training samples and iterate until convergence is achieved. Recent applications of partially supervised learning for emotion recognition were done by Cohen et al. [8] for recognition of facial expressions and by Liu et al. [18] for emotion recognition in speech. In the context of emotion recognition the works [30] and [27] integrated a partially supervised preprocessing scheme based on clustering into a classification system, which is somewhat similar to our approach.

The recent work on *compressed sensing* [7] has induced a paradigm shift in the signal processing community. The theory states that optimal signal recovery is possible with far fewer measurements than specified by the *Nyquist limit* if the signal is *sparse* in any domain. This means that only a *few* significant coefficients account for the signal, while almost all other coefficients are zero or close to zero. Calderbank et al. [6] showed that this paradigm is potentially useful for the machine learning community. Dictionary learning utilises underlying sparsity to create a *dictionary* of representative base atoms which can then be combined to describe the data. Often, the amount of representative atoms in a dictionary exceeds the dimension of the inherent signal dimension. Such dictionaries are called *overcomplete*. Applications of dictionary learning include image denoising [1], sparse decompositions of speech signals [14] and in the medical field for ultrasonic imagery [38] and sparse decomposition of electrocardiography (ECG) signals for denoising [20]. There are many approaches to generate a dictionary from sample data. Olshausen et al. [22] use a maximum a posteriori approach to find a suitable dictionary while Aharon et al. [1] propose a generalization of the k-means clustering algorithm called *k-svd*, which will be used throughout this paper.

3 Sparse Emotion Recognition Using Dictionary Learning

3.1 The k-svd Algorithm

Given signals $\{\boldsymbol{x^{(i)}} \in \mathbb{R}^n\}_{i=0\ldots L-1}$, the desired goal is to compute sparse representations $\{\boldsymbol{\gamma^{(i)}} \in \mathbb{R}^m\}_{i=0\ldots L-1}$ for them so that atmost k elements are nonzero. Throughout this paper k is referred to as *sparsity value*. The nonzero elements are coefficients for base vectors of a *dictionary* $D \in \mathbb{R}^{n\times m}$. That means for every sparse representation the original signal value can be computed by a linear combination of dictionary atoms:

$$\boldsymbol{x^{(i)}} = D\boldsymbol{\gamma^{(i)}} \tag{1}$$

To find a suitable dictionary D that allows the transformation in Eq. 1, many algorithms exist. We will focus on the *k-svd*, which is a generalization of the k-means cluster algorithm in the sense that a signal vector is not only associated with a single prototype, but rather with several prototypes weighted according to their relevance. Just like k-means, it alternates between updating the prototypes and computing updated representations based on the improved dictionary and iterates until convergence. For the update step, a singular value decomposition is used. More formally, the k-svd algorithm solves the following optimization problem:

Given a matrix $X \in \mathbb{R}^{n\times L}$ whose columns $\{\boldsymbol{x^{(i)}}\}_{i=0\ldots L-1}$ are signals, the algorithm computes a dictionary $D \in \mathbb{R}^{n\times m}$ and a matrix $\varGamma \in \mathbb{R}^{m\times L}$ whose columns are the sparse representations of the signals $\boldsymbol{x^{(i)}}$ so that the error

$$\min_{D,\varGamma} \|X - D\varGamma\|_F \tag{2}$$

is minimized with the constraint that $\forall i \ \left\|\boldsymbol{\gamma^{(i)}}\right\|_0 \leq k$ with $\boldsymbol{\gamma^{(i)}}$ being the i-th column of $\varGamma$ and $\|\cdot\|_0$ denotes the l_0-norm [41], which counts the number of nonzero elements. For more details on the k-svd algorithm refer to [1].

To transform a data point into the space spanned by the dictionary, it is necessary to solve the following problem: A datapoint $\boldsymbol{x}$ can be (approximately) modelled as a linear combination of dictionary elements.

$$\boldsymbol{x} = \sum_{i=1}^{m} \boldsymbol{\gamma_i}(\boldsymbol{d^T})^{\boldsymbol{(i)}} + \boldsymbol{\epsilon} \tag{3}$$

$(\boldsymbol{d^T})^{\boldsymbol{(i)}}$ denotes the i-th row of D. Since the vector $\boldsymbol{\gamma}$ is by construction k-sparse, atmost k elements are nonzero. To find these elements, the so called *matching pursuit* algorithms can be used. In a greedy fashion, they calculate the dictionary element on which the projection of the vector $\boldsymbol{x}$ is largest, add the corresponding coefficient to the already computed coefficients and subtract the contribution of that dictionary element and repeat the process until either the k coefficients are found or an error threshold is reached. The $\boldsymbol{\epsilon}$ in Eq. 3 denotes that a residual error can remain.

3.2 The Dataset

In this work, the same data was used as collected and used in the paper [23]. It consists of 8 different emotions recorded over 20 days from a single test subject. Four different modalities including blood volume pulse, respiration, skin conductance and electromyography were recorded. The eight emotions comprise anger, hate, grief, reverence, platonic love, romantic love and joy and were induced by music, imagery or imagination. For baseline measurements, data was recorded without any emotion stimulus (labeled as 'no emotion'). Each daily measurement consists of about 2 minutes per emotion, recorded with a sample rate of 20 Hz. For every emotion, its location in the *valence-arousal space* according to [25] was determined. Arousal indicates the activity of the nervous system and valence is a measurement of how positive an emotion feels. Figure 1 shows the valence-arousal space with emotions and their locations.

3.3 Proposed Method

Robust emotion recognition is highly dependent on proper preprocessing of the signals but even more on reliably extracting features and selecting the valuable ones. In our approach, the feature extraction stage is indirectly included in the creation of the dictionary. This means that, besides basic preprocessing, no features have to be selected manually. For the conducted experiments, only the blood volume pulse channel was used because its structure and periodicity is very advantageous for a dictionary learning approach. Blood volume pulse is used to capture the heart rate and the heart rate variability, two measures that are defined over interpeak intervals of the so called QRS complexes [24]. In the

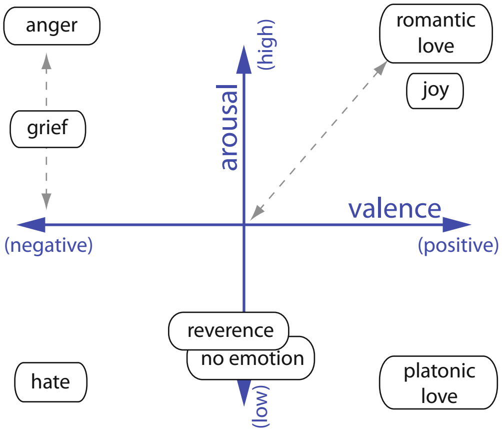


Fig. 1. The locations of the emotions categorized into valence and arousal values by the test subject of the study [12]. Because the quality of inducing some emotions differed during the recordings, there may be a drift of the emotions' true location. The dashed grey arrows indicate the direction of this drift.

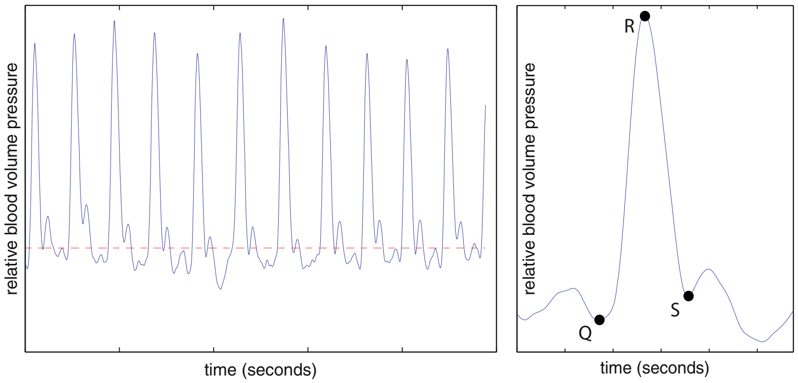


Fig. 2. Blood volume pulse signal after basic preprocessing. Major drifts in peak height have been removed and the mean has been adjusted. The characteristic QRS complexes are clearly visible. The small window shows a close up of a single QRS complex with its important points marked.

literature, heart rate is commonly used to differentiate emotions by their arousal level [15] but studies exist, which state that changes in the heart rate can also reflect changes in the mood [13]. Figure 2 shows a blood volume pulse signal together with a distinct QRS complex.

The different steps of the proposed approach in detail (for pseudocode refer to Algorithm 1):

Preprocessing and peak extraction: For the dictionary learning process, the signals should all approximately span the same domain. This is achieved by adjusting the mean to 0 for every sequence. Now, a modified version of the algorithm in [24] is used to locate the QRS complexes. To restrict changes in peak height of the QRS complexes, the whole sequence is scaled to set the median of the peak values to 1. The dataset was originally inspected to remove unusable recordings [12] and thus more involved preprocessing like linear detrending is not necessary.

Sequence cutting and alignment: To capture the whole spectrum of information in the heart rate, a window of appropriate length should be selected (for more details refer to [2] and [15]). In our experiments we choose a value of about 25 seconds. Using a small overlap between the sequences creates a more densely sampled data set and prevents important events to be lost or cut into different pieces. For emotion recognition, the relevant information is encoded in the interpeak intervals between neighbouring QRS complexes, so is has to be ensured that this information is not lost in the dictionary training process. This is accomplished by aligning the sequences to a certain peak. This alignment ensures that there is a common point of reference for every data segment. Because the alignment process can be inaccurate, the sequences are shifted based on the maximisation of a correlation value between the reference peak and the current peak in a small window.

Dictionary learning: The dictionary learning step is straightforward. After preparing the sequences, one has to define a sparsity value k, the desired number of dictionary elements m and which *pursuit algorithm* to use in the refinement process. We found that orthogonal matching pursuit (OMP) performed well in our scenario. Aharon et al. [1] found that other algorithms can result in slightly better dictionary quality but with higher computational costs. Building an overcomplete dictionary requires a considerable amount of data in order to accurately model the domain of the data points. Instead of merely relying on the signals that should be classified, additional data such as recorded sequences of other emotions can be used in a semisupervised manner to enrich the dictionary training process with additional data and thus ensure a more versatile dictionary. The effects of augmenting the training samples by additional unlabeled data is analysed in Sect. 4. A summary of the steps is given in Fig. 3.

Classification: After the dictionary has been trained, there is only one more step left. The training- and test-datasets have to be projected in the space

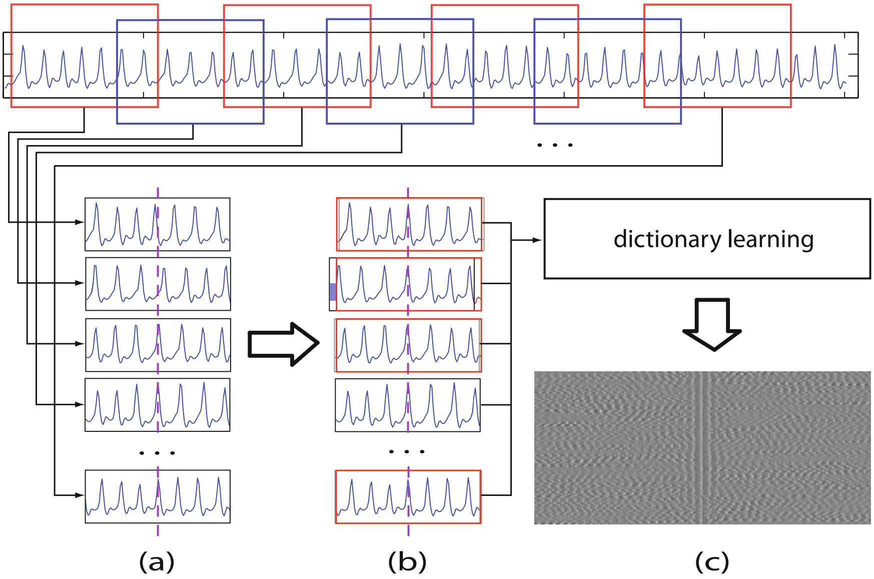


Fig. 3. In this figure, the workflow of the proposed algorithm is depicted. The upper part shows the input signal and the segmentation into smaller overlapping sequences. The segmentation takes the position of a peak as reference value and extracts a window of predetermined length around it. The results of this step can be seen in column (a). To correct for possible inaccuracies, a small window around the peak location is used to determine the optimal shift of the sequence to maximize a correlation value with a reference peak (column (b)). The effect can be seen by comparing the dashed purple reference lines. The aligned sequences are used as training samples for the dictionary learning process. The resulting dictionary can be seen in column (c) (bottom part) as a greyscaled image. A characteristic of the learnt dictionary is the column in the middle. It also shows the effects of the alignment process.

spanned by the dictionary. Once again, the chosen pursuit algorithm is applied. For a given data sample, it iteratively calculates the dictionary items for which the inner product is largest until all k coefficients are found. For more details refer to Sect. 3.1.

Algorithm 1: Pseudocode of the proposed algorithm

```
Input:
  – BVP classification sequences $S$
  – classification labels $l$ of $S$
  – BVP dictionary training sequences $U$
  – segment length $t$ in seconds
  – sparsity $k$
  – number of dictionary elements $m$

Result: Classifier $K$, Dictionary $D$
$C = \emptyset, \tilde{U} = \emptyset$ ;
foreach sequence $u$ in $U$ do
    $\tilde{u} = cutSequence(u, t)$;  //cut sequence according to $t$
    $\tilde{U} = \tilde{U} \cup \tilde{u}$;
$\tilde{U} = align(\tilde{U})$;             //create common reference frame by alignment.
$D := trainDictionary(\tilde{U}, k, m)$;  //$\tilde{U}$ contains cut and aligned sequences.
foreach sequence $s$ in $S$ do
    $\tilde{s} = cutSequence(s, t)$;
    $c = omp(\tilde{s}, D, k)$;  //project data point into dictionary space using
    orthogonal matching pursuit
    $C = C \cup c$;
Train classifier $K$ on $C$ using labels $l$;
Output: Classifier $K$, Dictionary $D$
```

4 Experiments and Results

Different experiments were conducted to evaluate the performance of the approach. The base of the experiments was the classification of the emotions *anger* and *platonic love* because both arousal and valence values differ. The feature vectors were computed by first creating a dictionary from a pool of training data samples and then projecting the classification training samples into the dictionary space using orthogonal matching pursuit [39]. The sparse feature vectors are then fed into a support vector machine (SVM) with a linear kernel for classification (Fig. 4).

4.1 Effects of Varying the Number of Dictionary Training Samples

The first experiment measures the effect of different training set sizes for the dictionary learning stage. The amount of informative training data used for the dictionary creation process is critical for the quality of the dictionary. In order to

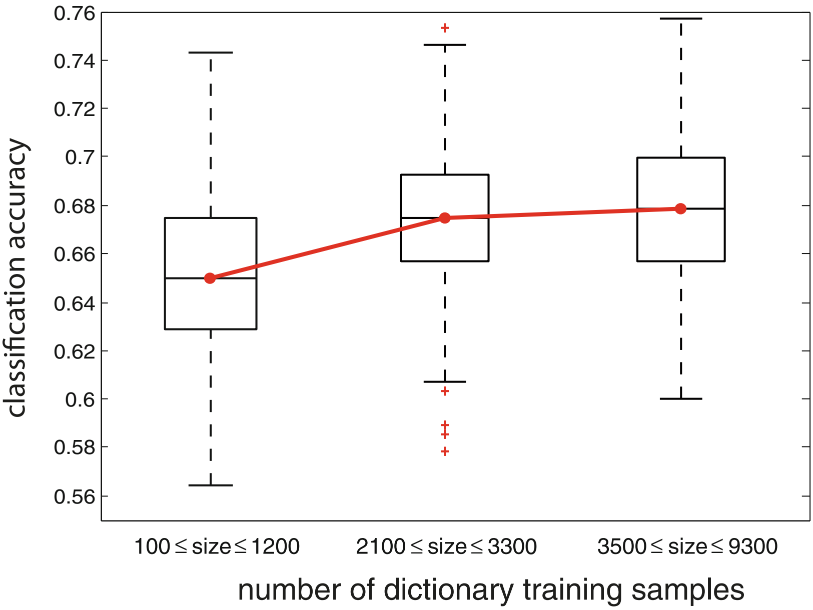


Fig. 4. The effects of different dictionary training set sizes in a cross validation context. The classification rate significantly improves using more data, including unlabeled samples. However, the improvement quickly reaches a level of satiation as can be seen by comparing the middle with the right boxplot. Each boxplots was created using at least 9 different training set sizes, spaced at least 100 samples apart. For every set size, the experiment was repeated 30 times.

get a versatile dictionary, many different training samples that cover the whole input space are needed. The training set size was gradually increased from 100 to 9300. The sparsity value was set to 3 and the dictionary dimension (number of atoms) was set to 60. In this experiment, the differentiation between the individual days was released thus allowing the training and test sets to contain data from every daily measurement. The validation of the classification result was done by cross validation with 20 % of the samples left out and the results are averaged over 30 trials.

4.2 Effects of Varying the Sparsity Value

The high degree of sparseness of the representations that can be achieved with dictionary learning renders it very useful for specific tasks. To examine the effect of different sparsity parameters on the classification result in the system at hand, the following experiment was carried out. The base of the experiment is the same as in Sect. 4.1, however the dictionary training size was fixed at 5000 samples while the sparsity parameter was increased from 1 to 50. The whole experiment was repeated 50 times and the results were averaged. In this context sparsity

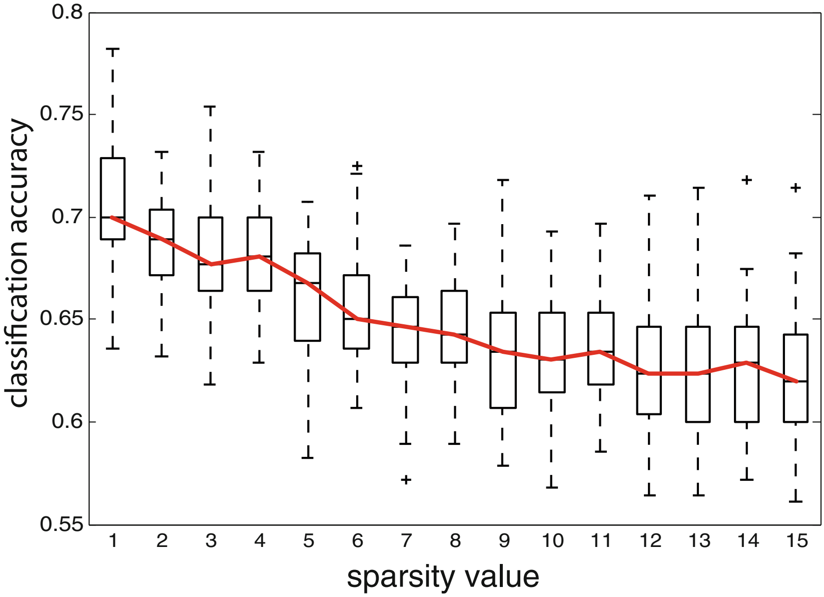


Fig. 5. The classification accuracy decreases with an increasing sparsity parameter. A dictionary with 5000 training samples is used for the comparison.

refers to the number of nonzero elements in the representation vectors. As can be seen in Fig. 5, the classification accuracy is decreasing with an increasing number of nonzero coefficients. It is interesting to see that the sparsity value of 1 performs best amongst all parameter choices. This could indicate that the choice of 60 dictionary atoms is already sufficiently high for classification. Thus by adding more coefficients a bigger overlap of the two classes is created, rendering separation of the two classes by a classifier more difficult.

4.3 Effects of Using Labeled and Unlabeled Data

This experiment should demonstrate the impact of unlabeled data on the quality of the dictionary. In comparison to the first experiment, this time, instead of cross validation, the differentiation between recordings from different days is kept and the classifier is validated with data of previously unseen days. First, only data from the emotions *platonic love* and *anger* is used to train the dictionary. The same emotions are classified in this scenario. In the second part, every recorded emotion is added to the dictionary training set. The amount of available training data is now much higher. The results were created by averaging over 50 trials. Figure 6 shows the classification rates with the respective number of training samples for the dictionary learning. It can be seen that the inclusion of additional data in terms of other recorded emotions improves the overall classification quality. Two more observations can be made:

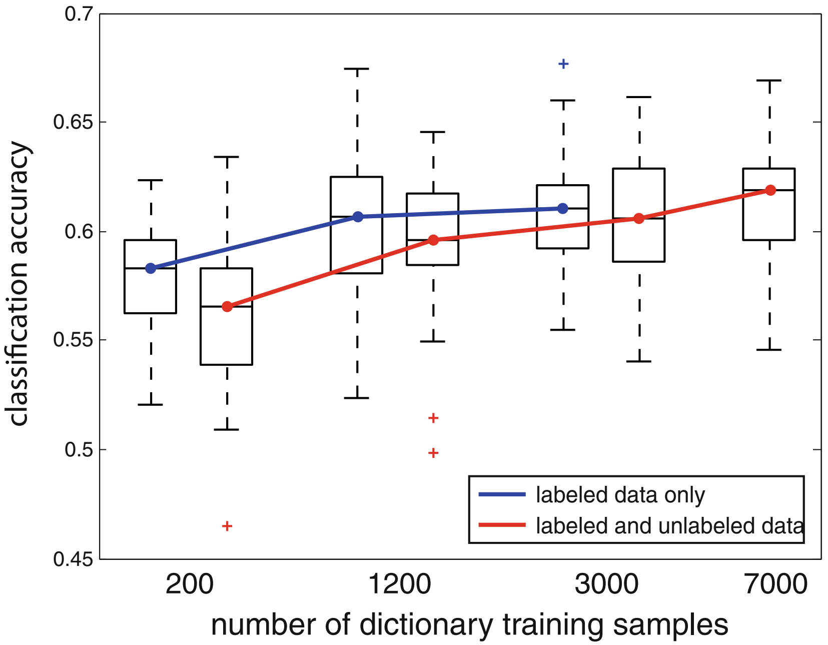


Fig. 6. Comparison of the classification quality of dictionaries trained with different sets of data. The boxplots connected by the blue line show the performance of dictionary training using only the emotion classes that are also classified. The boxplots connected by the red line show the performance of the learning process augmented by other emotions besides the ones classified. It is evident that using more, especially task unspecific data helps improving the classification quality.

- The slopes of the lines decrease rapidly. For the blue line, the step from 1200 to 3000 seems to result in a satiation. Although the red line also shows signs of satiation, the overall quality still increases up to a classification accuracy that is higher than using labeled information only.
- Dictionary training with data of the two classes that are classified seems to outperform training with the complete data set. This is certainly plausible because training with data of just the two classes that will be classified yield a dictionary that is much more tailored for exactly the task at hand. The augmentation, on the other hand, yields a more general dictionary that captures the underlying structure of the data more effectively.

5 Discussion

The provided results indicate that the usage of additional unlabeled data is promising for dictionary learning purposes. While the increase in classification accuracy between a minimum dictionary training set size of 100 and the maximum training size of about 9000 is about 5 %, the gain of using unlabeled data is slightly less at about 1.5–2 %. While in our approach extensive preprocessing and feature extraction were not necessary, the resulting features vectors were, although sparse, of very high dimension because of the number of trained dictionary atoms. In practice, a large number of dictionary atoms renders the approach impractical for large data sets. For small to medium scale problems, the approach seems promising. The analysed blood volume pulse signal is expedient for the purposes of this work because the structure is periodical and the irregularities in the periods are the important aspects that carry the information. Exactly those irregularities are captured by different atoms of the dictionary. For similar signals, the sparse representation will use similar dictionary atoms (e.g. with different scaling) while dissimilar signals are represented by combinations of different dictionary atoms. For a classifier, this is a favorable starting position.

For future work, the applicability of this approach to non-periodical signals could be examined with possible modifications to the alignment procedure. Another possible avenue to continue could be a combination of dictionaries for different modalities (e.g. respiration) or even a single dictionary from combined signals.

6 Conclusion

This paper presented an approach for emotion recognition based on blood volume pulse signals using sparse representations. Sparsity of the feature points is achieved using a dictionary learning approach directly on the raw signals. Excessive preprocessing, feature generating and feature selection are not necessary in our system. Furthermore, we showed that the quality of the dictionary can be improved by adding unlabeled data to the training samples and thus permitting a higher degree of sparsity. The effects of sparsity, number of dictionary training samples and enrichment by unlabeled training samples have been analysed.

Acknowledgements. The presented work was developed within the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG). The work of Markus Kächele is supported by a scholarship of the Landesgraduiertenförderung Baden-Württemberg at Ulm University.

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-svd: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11), 4311–4322 (2006)
2. Altimiras, J.: Understanding autonomic sympathovagal balance from short-term heart rate variations. Are we analyzing noise? Comp. Biochem. Physiol. A: Mol. Integr. Physiol. **124**, 447–460 (1999)
3. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT: Proceedings of the Workshop on Computational Learning Theory. Morgan Kaufmann Publishers (1998)

4. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Pajdla, T., Matas, J.G. (eds.) ECCV 2004, Part IV. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004)
5. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of german emotional speech. In: INTERSPEECH'05, pp. 1517–1520 (2005)
6. Calderbank, R., Jafarpour, S., Schapire, R.: Compressed learning: universal sparse dimensionality reduction and learning in the measurement domain. Technical report (2009)
7. Cands, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure. Appl. Math. **59**(8), 1207–1223 (2006)
8. Cohen, I., Sebe, N., Cozman, F.G., Huang, T.S.: Semi-supervised learning for facial expression recognition. In: Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR '03, pp. 17–22. ACM (2003)
9. Dietrich, C., Schwenker, F., Palm, G.: Decision templates for the classification of bioacoustic time series. In: Proceedings of the 2002 12th IEEE Workshop on Neural Networks for Signal Processing, pp. 159–168 (2002)
10. Glodek, M., Schels, M., Palm, G., Schwenker, F.: Multiple classifier combination using reject options and markov fusion networks. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI '12, pp. 465–472. ACM, New York (2012)
11. Glodek, M., et al.: Multiple classifier systems for the classification of audio-visual emotional states. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part II. LNCS, vol. 6975, pp. 359–368. Springer, Heidelberg (2011)
12. Healey, J.A.: Wearable and automotive systems for affect recognition from physiology. PhD thesis (2000)
13. Hughes, J.W., Stoney, C.M.: Depressed mood is related to high-frequency heart rate variability during stressors. Psychosom. Med. **62**(6), 796–803 (2000)
14. Jafari, M., Plumbley, M.: Fast dictionary learning for sparse representations of speech signals. IEEE J. Sel. Top. Sig. Process. **5**(5), 1025–1031 (2011)
15. Jonghwa, K., Ande, E.: Emotion recognition based on physiological changes in music listening. IEEE Trans. Pattern Anal. Mach. Intell. **30**(12), 2067–2083 (2008)
16. Kächele, M., Meudt, S., Arndt, I., Schwenker, F.: Cascaded fusion of dynamic, spatial, and textural feature sets for person-independent facial emotion recognition (2013) (Submitted to ICMI 2013)
17. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, New York (2004)
18. Liu, J., Chen, C., Bu, J., You, M., Tao, J.: Speech emotion recognition using an enhanced co-training algorithm. In: IEEE International Conference on Multimedia and Expo, pp. 999–1002 (2007)
19. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2010), pp. 94–101 (2010)
20. Mailhe, B., Gribonval, R., Bimbot, F., Lemay, M., Vandergheynst, P., Vesin, J.M.: Dictionary learning for the sparse modelling of atrial fibrillation in ecg signals. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '09, pp. 465–468. IEEE Computer Society (2009)
21. Ojala, T., Pietikäinen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)

22. Olshausen, B.A., Field, D.J.: Natural image statistics and efficient coding. Net. Comput. Neural Sys. **7**, 333–339 (1996)
23. Picard, R., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. Mach. Intell. **23**(10), 1175–1191 (2001)
24. Rudnicki, M., Strumillo, P.: A real-time adaptive wavelet transform-based qrs complex detector. IEEE Trans. Biomed. Eng. **46**(7), 281–289 (2007)
25. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**(6), 1161–1178 (1980)
26. Schachter, S.: The interaction of cognitive and physiological determinants of emotional state. In: Berkowitz, L. (ed.) Advances in Experimental Social Psychology, vol. 1, pp. 49–80+. Academic Press, New York (1964)
27. Schels, M., Kächele, M., Hrabal, D., Walter, S., Traue, H.C., Schwenker, F.: Classification of labeled and unlabeled bio-physiological data. In: Schwenker, F., Trentin, E. (eds.) PSL 2011. LNCS, vol. 7081, pp. 138–147. Springer, Heidelberg (2012)
28. Schels, M., Glodek, M., Meudt, S., Schmidt, M., Hrabal, D., Bck, R., Walter, S., Schwenker, F.: Multi-modal classifier-fusion for the classification of emotional states in WOZ scenarios. In: Proceedings of the 1st International Conference on Affective and Pleasurable Design (APD'12) [jointly with the 4th International Conference on Applied Human Factors and Ergonomics (AHFE'12)]. Advances in Human Factors and Ergonomics Series, pp. 5337–5346. CRC Press (2012)
29. Schels, M., Scherer, S., Glodek, M., Kestler, H.A., Palm, G., Schwenker, F.: On the discovery of events in eeg data utilizing information fusion. In: Computational Statistics: Special Issue: Proceedings of Statistical, Computing 2010 (2011) (online first)
30. Schels, M., Schillinger, P., Schwenker, F.: Training of multiple classifier systems utilizing partially labeled sequences. In: Proceedings of the 19th European Symposium on Artificial, Neural Networks (ESANN'11), pp. 71–76 (2011)
31. Scherer, S., Glodek, M., Schels, M., Schmidt, M., Layher, G., Schwenker, F., Neumann, H., Palm, G.: A generic framework for the inference of user states in human computer interaction: How patterns of low level communicational cues support complex affective states. In: Special Issue on Conceptual Frameworks for Multimodal Social Signal Processing, Journal on Multimodal User Interfaces (2012) (online first)
32. Scherer, S., Kane, J., Gobl, C., Schwenker, F.: Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. Comput. Speech Lang. **27**(1), 263–287 (2013)
33. Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., Pantic, M.: AVEC 2011-the first international audio/Visual emotion challenge. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part II. LNCS, vol. 6975, pp. 415–424. Springer, Heidelberg (2011)
34. Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A.:Acoustic emotion recognition: a benchmark comparison of performances. In: IEEE Workshop on Automatic Speech Recognition Understanding, ASRU 2009, pp. 552–557 (2009)
35. Schwenker, F., Scherer, S., Magdi, Y.M., Palm, G.: The GMM-SVM supervector approach for the recognition of the emotional status from speech. In: Alippi, C., Polycarpou, M., Panayiotou, C., Ellinas, G. (eds.) ICANN 2009, Part I. LNCS, vol. 5768, pp. 894–903. Springer, Heidelberg (2009)
36. Schwenker, F., Scherer, S., Morency, L.-P. (eds.): MPRSS 2012. LNCS (LNAI), vol. 7742. Springer, Heidelberg (2013)

37. Schwenker, F., Scherer, S., Schmidt, M., Schels, M., Glodek, M.: Multiple classifier systems for the recogonition of human emotions. In: El Gayar, N., Kittler, J., Roli, F. (eds.) MCS 2010. LNCS, vol. 5997, pp. 315–324. Springer, Heidelberg (2010)
38. Tosic, I., Jovanovic, I., Frossard, P., Vetterli, M., Duric, N.: Ultrasound tomography with learned dictionaries. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, pp. 5502–5505. IEEE (2010)
39. Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans. Inform. Theory **53**, 4655–4666 (2007)
40. Walter, S., et al.: Multimodal emotion classification in naturalistic user behavior. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part III, HCII 2011. LNCS, vol. 6763, pp. 603–611. Springer, Heidelberg (2011)
41. Wipf, D., Rao, B.: ℓ_0-norm minimization for basis selection. Adv. Neural Inf. Process. Sys. **17**, 1513–1520 (2005)
42. Yarowski, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In: Proceedings Thirty-Third Meeting of the ACL, pp. 189–196 (1995)
43. Zhang, L., Tjondronegoro, D.: Facial expression recognition using facial movement features. IEEE Trans. Affect. Comput. **2**(4), 219–229 (2011)

Adaptive Graph Constrained NMF for Semi-Supervised Learning

Qian Li, Liping Jing(✉), and Jian Yu

School of Computer and Information Technology,
Beijing Jiaotong University, Beijing 89069, China
{lpjing@bjtu.edu.cn}

Abstract. Recently, graph-based semi-supervised learning (GB-SSL) has received a lot of attentions in pattern recognition, computer vision and information retrieval. The key parts of GB-SSL are designing loss function and constructing graph. In this paper, we proposed a new semi-supervised learning method where the loss function is modeled via graph constrained non-negative matrix factorization (GCNMF). The model can effectively cooperate the precious label information and the local consistency among samples including labeled and unlabeled data. Meanwhile, an adaptive graph construction method is presented so that the selected neighbors of one sample are as similar as possible, which makes the local consistency be correctly preserved in the graph. The experimental results on real world data sets including object image, face and handwritten digit have shown the superiority of our proposed method.

Keywords: Semi-supervised learning · Graph constrained NMF · Adaptive graph construction

1 Introduction

With the exponential increasing of unlabeled data, the problem of learning from the expensive labeled data and a large amount of unlabeled data has attracted considerable attention in recent years. Such a learning problem is often called semi-supervised. After the value of unlabeled data was proved [12,18], many semi-supervised methods have been developed [21,22] for real applications in pattern recognition, computer vision, information retrieval and etc.

Among them, graph-based semi-supervised learning method (GB-SSL) is an important technique and has attracted a lot of interests. The early GB-SSL methods were designed by constructing a graph which takes data samples as nodes and the relation between connected samples as edges, and then seeking the minimum cut of the graph such that nodes in each connected component have the same label [3,4]. This kind of method only assigns the unlabeled sample into one category, called discrete prediction. Zhu et al. [23] extended the discrete prediction to continuous case by adapting the Gaussian random fields and label propagation strategy. They modeled the learning problem as a quadratic loss of

Z.-H. Zhou and F. Schwenker (Eds.): PSL 2013, LNAI 8183, pp. 36–48, 2013.
DOI: 10.1007/978-3-642-40705-5_4,
© Springer-Verlag Berlin Heidelberg 2013

the prediction function and used a normalized graph Laplacian as the regularizer to combine the relation between data samples. Subsequently, Belkin et al. [2] proposed a manifold regularization framework in Reproducing Kernel Hillbert Space (RKHS) based on the assumption that the data are distributed on a Riemannian manifold. This framework can provide similar output in a local region by regularizing the loss function with local smoothness.

In the above GB-SSL methods, the prediction value is obtained by a constant function about the unlabeled data information with spikes of few label information. Such solution may result in overfitting problem [13]. In many real applications, the original data matrix only contains non-negative values such as the term frequency in document sets, the gray value of the image pixel and etc. In this paper, thus, we introduce non-negative matrix factorization (NMF) technique to model the loss function between the data samples and the given limited label information. The basic idea of NMF is to find several non-negative bases from the original data and then linearly project data onto these bases [7]. Meanwhile, the nonnegative constraint leads NMF to a parts-based representation of the sample, which makes NMF popular in many fields.

Recently, researchers studied and extended the standard NMF to different semi-supervised models by considering the existing precious prior knowledge. These semi-supervised NMF methods can be grouped into two types, one is based on the pairwise constraints, the other is based on the explicit category labels. In order to balance the effect of pairwise constraints and data information, the existing semi-supervised NMF methods have to introduce extract parameters which require to be tuned carefully [16,24]. Lin and Wu [10] take the label information as additional hard constraints and propose a parameter free constrained NMF (CNMF). CNMF outputs a parts-based representation of data which has the consistent label with the original data. CNMF considers the label information, but ignores the local consistency between data points. However, such consistency is helpful to predict the category information of unlabeled data [2,5,17,20]. Among them, graph regularized NMF (GNMF) [5] is a typical unsupervised NMF method with the aid of local consistency.

In this paper, a new semi-supervised NMF model based on graph constraint is proposed to effectively cooperate the label information and intrinsic structure among data. Meanwhile, we design an adaptive graph construction to select the neighbors for each data point, which preserves the local consistency in graph and makes the algorithm parameter free.

The rest of this paper is organized as follows. Section 2 describes the related work. In Sect. 3, we introduce GCNMF model and the adaptive graph construction method. In Sect. 4, the experimental results on real-world data sets are listed and discussed. Section 5 provides our brief conclusion and future work.

2 Related Work

Given a set of n_l labeled samples $\{X_l, Y_l\} = \{(x_i, y_i)\}_{i=1}^{n_l}$ and a set of n_u unlabeled samples $X_u = \{x_j\}_{j=n_l+1}^{n_l+n_u}$, where $y_i \in \{1, 2, \cdots, c\}$ and c is the number of classes. The objective of GB-SSL is to infer the labels $\{y_{n_l+1}, \cdots, y_n\}$ for unlabeled data. In this case, the label information is denoted as $Y = [Y_l, Y_u]$, among them, $Y \in \mathbf{B}^{c\times n}$ is a binary matrix with $Y_{ji} = 1$ if x_i has label j, otherwise $Y_{ji} = 0$. Usually, the labeled data is very few, i.e, $n_l \ll n_u$ in real applications.

GB-SSL method works based on an undirected graph $G = \{X, E\}$, where the set of nodes or vertices is $X = \{x_i\}_{i=1}^n$ and the set of edges is $E = \{e_{ij}\}_{i,j=1}^n$. Typically, one uses some similarity function to calculate the weight of edge e_{ij}, which can be used to build a weight matrix $M = \{m_{ij}\}_{i,j=1}^n$. Meanwhile, the node degree diagonal matrix is defined as $D_{ii} = \sum_{j=1}^n m(i,j)$, the graph Laplacian is defined as $\Delta = D - M$ and the normalized graph Laplacian is $L = D^{-1/2}MD^{-1/2}$.

GB-SSL methods propagate label information from label nodes to unlabeled nodes by using edge-based affinity functions to estimate the weight of each edge [22]. Most methods define a continuous classification $F \in R^{c\times n} = [F_l, F_u]$ ($F_l \in R^{c\times n_l}$ and $F_u \in R^{c\times n_u}$) that is estimated on the graph to minimize a cost function. The cost function usually enforces a tradeoff between the smoothness of the function on the data graph and the accuracy of the function at fitting the label information. In particular, the local and global consistency method (LGC) [20] aims to optimize

$$\arg\min_F J_1(F) = tr(FLF^T) + \mu\|F - Y\|_F^2. \tag{1}$$

LGC is popular and has been empirically validated in the past. However, it only considers the intrinsic structure among the data set but ignores the information in ambient space. In order to solve this problem, Belkin et al. [2] combined the graph smoothness with the cost function in RKHS. Two popular methods LapRLS and LapSVM are proposed as follows.

$$\begin{aligned}&\arg\min_F J_2(F)\\ &= \tfrac{1}{l}\textstyle\sum_{i=1}^l V(x_i, y_i, F) + \lambda_1\|F\|_K^2 + \lambda_2 tr(FLF^T)\end{aligned} \tag{2}$$

where V is loss function, such as squared loss $(y_i - f(x_i))^2$ for RLS or the hinge loss function $\max[o, 1 - y_i f(x_i)]$ for SVM. The second term is an penalty term reflecting the complexity of the learner in the ambient space. The third term is to estimate the smoothness of graph. Parameters λ_1 and λ_2 require manual setting to control the effect of two later terms.

Recently, the intrinsic structure among data has been applied in NMF-based learning methods such as GNMF [5]. GNMF is implemented via solving the following optimization problem.

$$\arg\min_{W,F} J_3(W, F) = \|X - WF\|_F^2 + \lambda tr(FLF^T). \tag{3}$$

The aim of GNMF is to uncover the hidden semantics and simultaneously respect the intrinsic geometric structure. In (3), F can be taken as the compact representation of the original data X. The second term $\lambda tr(FLF^T)$ can guarantee that the new representation keeps the intrinsic geometric structure. Parameter λ should be carefully tuned. The initial purpose of GNMF is unsupervised learning, e.g., the clustering result can be obtained based on F. This method can incorporate the label information into the graph structure by modifying the weight matrix L to implement semi-supervised learning [10].

In addition, all the above GB-SSL methods have a strong assumption that the graph construction is reasonable and appropriate to reflect the intrinsic structure of data. Most existing graph construction algorithms, k nearest neighbor graph [15], local linear reconstruction graph [14], b-matching [6], and anchor-based [11] methods, have certain parameters which require manual setting and the parameters have a great impact on the structure of graph. Recently, Yan and Wang adopted sparse representation technique to build the graph and introduce it into semi-supervised learning model [17]. However, sparse representation does not consider the global constraints [25] when building the graph.

In next section, we will propose a semi-supervised NMF model with adaptive graph construction to solve the above problems.

3 Graph Constrained NMF

Given labeled examples $\{X_l, Y_l\}$ and unlabeled examples X_u, the semi-supervised learning can be implemented by optimizing the following objective function [2,13,22]

$$J^* = \arg\min_J \Omega(X_l, Y_l, J) + \Psi(X_u, J). \tag{4}$$

Both Ω and Ψ are loss functions between the known information and the prediction representation J. Among them, Ω evaluates the cost of labeled data, and Ψ measures the cost of unlabeled data.

3.1 Model Formulation

In this section, we propose a graph constrained NMF model (GCNMF) to estimate the cost of prediction. The traditional NMF model tries to decompose the given data matrix X into two parts, W for basis factor and H for encoding factor. Each data sample x_i can be represented by a linear combination of the bases W with coefficient $H_{i\cdot}$. Finally, $H_{i\cdot}$ can be taken as cluster indicator to predict the label information of unlabeled data x_i.

In order to make the coefficient factor H_l of X_l have the consistent label with Y_l, we define $H_l = Y_l S$. Here $Y_l \in R^{n_l \times c}$ is the known label information, and $S \in R^{c \times k}$ is an auxiliary matrix. Then, the loss function $\Omega(X_l, Y_l, J)$ for X_l can be defined as

$$\Omega(X_l, Y_l, J) = \|X_l - W(Y_l S)^T\|_F^2, \tag{5}$$

where $W \geq 0$ and $S \geq 0$. $Y_l(i,j) = 1$ if the ith point belongs to the jth category, otherwise $Y_l(i,j) = 0$. Meanwhile, the auxiliary matrix S is introduced to estimate the relationship between the known c categories $C_{1,\cdots,c}$ and the k bases $W_{\cdot t}$ $(t = 1, \cdots, k)$ being learned.

For unlabeled data X_u, the similarity between each pair of points $G \in R^{n \times n}$ is embedded to cooperate the intrinsic structure among data.

$$G = \begin{bmatrix} G_{ll} & G_{lu} \\ G_{ul} & G_{uu} \end{bmatrix}$$

Here G is a symmetric matrix. G_{ll} is the similarity between labeled data, $G_{ul} = G_{lu}$ is the similarity between unlabeled data and labeled data, and G_{uu} is the similarity between unlabeled data. Note that the matrix G can be calculated previously, and more detail on how to build G will be described in next section.

Since G_{ul} indicates the similarity between unlabeled data and labeled data, the label information Y_l is useful for predicting the label of unlabeled data. In other words, if a unlabeled point X_u^i and a labeled point X_l^j is very similar (G_{ul}^{ij} is larger), these two points should have the same label distribution. From this perspective, the coefficient factor H_u based on labeled information can be defined as $H_u^{(1)} = G_{ul} Y_l S$. This representation indicates that if one unlabeled data sample is more similar to a labeled sample, it should be assigned into the same class with the labeled sample.

At the same time, label prediction of each unlabeled sample is effected by its around unlabeled samples especially when there is a large amount of unlabeled data. If unlabeled points X_u^i and X_u^j are similar, i.e., G_{uu} is large, their corresponding coefficient vectors should be as similar as possible [5]. In order to keep this kind of local consistency, we formulate the coefficient factor H_u of X_u as $H_u^{(2)} = G_{uu} Z$, here $Z \in R^{n_u \times k}$ is an auxiliary matrix.

Finally, $H_u^{(1)}$ and $H_u^{(2)}$ are combined to represent H_u as $G_{ul} Y_l S + G_{uu} Z$. This setting guarantees that two points with higher similarity have similar part-based representations. Then, the loss function $\Psi(X_u, J)$ for X_u can be defined as

$$\Psi(X_u, J) = \|X_u - W(G_{ul} Y_l S + G_{uu} Z)^T\|_F^2 \tag{6}$$

where $W \geq 0$, $S \geq 0$ and $Z \geq 0$. Note that, when the unlabeled data are loosely correlated to the labeled data, namely when most of the elements within G_{ul} are small, this leads to $H_u = G_{uu} Z$. We refer to this case as weakly semi-supervised learning.

Based on (5) and (6), the semi-supervised learning objective function (4) can be written as

$$\begin{aligned} &\arg \min_{W \geq 0, S \geq 0, Z \geq 0} J_3(W, S, Z) \\ &= \|X_l - W(Y_l S)^T\|_F^2 + \|X_u - W(G_{ul} Y_l S + G_{uu} Z)^T\|_F^2. \end{aligned} \tag{7}$$

Here the loss function Ω and Ψ share the same term W which will output a solution considering both labeled and unlabeled data. We call this model as graph constrained NMF (GCNMF). In order to easily solve (7), let

$$A_l = X_l - W(Y_l S)^T,$$

$$A_u = X_u - W(G_{ul} Y_l S + G_{uu} Z)^T,$$

then the above optimization model can be written as

$$\arg\min_{W\geq 0, S\geq 0, Z\geq 0} J_3(W,S,Z) = \|A_l\|_F^2 + \|A_u\|_F^2 \tag{8}$$

Let matrix $A = [A_l, A_u]$, according to the property of matrix Frobenius norm, we have $\|A\|_F^2 = \|A_l\|_F^2 + \|A_u\|_F^2$. In this case, the optimization problem can be solved by finding the minimum of

$$\arg\min_{W\geq 0, S\geq 0, Z\geq 0} J_3(W,S,Z) = \|A\|_F^2. \tag{9}$$

Here,

$$\begin{aligned} A &= [A_l, A_u] \\ &= [X_l - W(Y_l S)^T, X_u - W(G_{ul}Y_l S + G_{uu}Z)^T] \\ &= [X_l, X_u] - [W(Y_l S)^T, W(G_{ul}Y_l S + G_{uu}Z)^T] \\ &= [X_l, X_u] - W[(Y_l S)^T, (G_{ul}Y_l S + G_{uu}Z)^T] \\ &= [X_l, X_u] - W\begin{bmatrix} Y_l S \\ G_{ul}YS + G_{uu}Z \end{bmatrix}^T \end{aligned}$$

and,

$$\begin{bmatrix} Y_l S \\ G_{ul}Y_l S + G_{uu}Z \end{bmatrix} = \begin{bmatrix} Y_l & 0 \\ G_{ul}Y_l & G_{uu} \end{bmatrix} \begin{bmatrix} S \\ Z \end{bmatrix}.$$

Let $U = \begin{bmatrix} Y_l & 0 \\ G_{ul}Y_l & G_{uu} \end{bmatrix}$, $V = \begin{bmatrix} S \\ Z \end{bmatrix}$, $X = [X_l, X_u]$. Then the proposed GCNMF method can be implemented by solving the following optimization problem.

$$\arg\min_{W\geq 0, V\geq 0} J_3(W,V) = \|X - W(UV)^T\|_F^2 \tag{10}$$

where U is the known information about label information Y_l and the similarity graph G on both labeled and unlabeled samples.

Following the standard theory of constrained optimization, the Lagrangian multiplier technique [8] can be used to solve (10) by introducing two multipliers α and β for constraints $W \geq 0$ and $V \geq 0$ respectively. Then, the variable W and V can be updated according to the KKT complementary condition as

$$W_{ij}^{(\tau+1)} = W_{ij}^{(\tau)} \frac{(XUV)_{ij}}{(W^{(\tau)}V^T U^T UV)_{ij}} \tag{11}$$

and

$$V_{ij}^{(\tau+1)} = V_{ij}^{(\tau)} \frac{(U^T X^T W)_{ij}}{(U^T UV^{(\tau)} W^T W)_{ij}} \tag{12}$$

Here τ is the iteration. The correctness and convergence of the updating process can be proved similar to [9,10]. Actually, CNMF [10] is a special case of our proposed model GCNMF when $G_{uu} = I$ and $G_{ul} = 0$, i.e, the relation between data points is ignored. In GNMF [5], the local consistency is only enforced on the prediction function F rather than the mapped representation (WF in (3)). However, the purpose of NMF-type model is to approximate X with WF, thus the local consistency in X should be preserved in WF not just F. In summary, our proposed model is different from the models in [5] and [10]. Once having V, we can get the data cluster-indicator $F = (UV)^T$ and the category of each point x_i, as $j^* = \arg\max_j \{F_{ji}\}$.

3.2 Adaptive Graph Construction

In the literatures, several methods have been proposed on graph construction for GB-SSL [1,6,11,14,15]. Their performance is significantly effected by the parameter (number of neighbors) which requires tuning carefully. In this paper, an adaptive graph construction is proposed to select the appropriate neighbors for each sample.

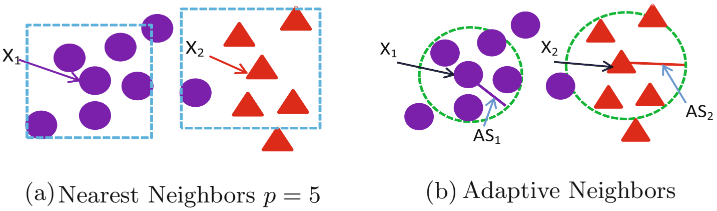
Given data set X, its similarity matrix $M \in R^{n\times n}$ can be calculated via some methods, e.g., Heat kernel [1] as follows.

$$M_{ij} = \begin{cases} \exp - \frac{\|x_i - x_j\|^2}{\sigma_i \sigma_j}, & x_j \in \mathcal{N}(x_i) \\ 0, & otherwise \end{cases} \tag{13}$$

Here σ_i is the variance of similarities between x_i and all samples. For each sample x_i, its average similarity AS_i is defined as $AS_i = \frac{\sum_{t=1}^{n} M_{it}}{n}$. If the similarity between x_i and x_j is greater than AS_i, then x_j becomes a neighbor of x_i. In this case, a hypersphere is constructed whose center and radius are x_i itself and the mean of the similarities (AS_i) respectively.

$$M_{ij}^{AG} = \begin{cases} M_{ij}, if \quad M_{ij} > AS_i \\ 0, \quad otherwise \end{cases} \tag{14}$$

Figure 1 indicates a toy example to show the process of adaptive neighbor selection. The toy data set contains 14 points belonging to two classes denoted as circle and triangle respectively. The neighbors of points x_1 and x_2 via fixing $p = 5$ are the points surrounded by the dash square in Fig. 1a. (In order to efficiently build the neighbor graph, p samples having higher similarity with the current



(a) Nearest Neighbors $p = 5$ (b) Adaptive Neighbors

Fig. 1. Neighbor selection based on similarity a) fixing number of neighbors ($p = 5$), b) adaptive selecting the neighbors based on AS.

point are considered. $AS_i = \frac{1}{p}\sum_{t=1}^{p} M_{it}$ is the average similarity between x_i and its top p similar samples.)

It can be seen that the neighbor distribution of each point is different. For example, x_1 has dense neighbors, while x_2 has relatively sparse neighbors. In this case, the neighbors of x_1 will play more important role to predict label for x_1 than the neighbors of x_2 predicting label for x_2. Meanwhile, the neighbors of x_2 come from two classes.

In Fig.1b, the sizes of neighbors of x_1 and x_2 (dash circle) become different because of introducing the threshold (average similarity AS_1 and AS_2). For x_2, only three same-class neighbors are determined. Even though this adaptive strategy can not guarantee that all selected neighbors of a sample are same-class, it makes sure that the neighbors are as more similar as possible. Thus, it is expected that the adaptive graph construction can effectively keep local consistency and then improve the performance of GB-SSL.

4 Experimental Results and Discussion

4.1 Dataset and Methodology

In this section, we would test GCNMF on four real data sets. Two of them are facial image data, one is object image data, and the forth one is handwritten digits data. Among them, Yale Face set [10] consists of 11 different images in each of 15 distinct persons. AT&T Face set [10] contains 400 images of 40 person and each has 10 facial images. COIL20 image library [5] contains 32×32 gray scale images of 20 objects viewed from varying angles and each object has 72 images. USPS handwritten digits set [2] has 10 digits $0, 1, \cdots, 9$ and 7291 samples where each sample is 16×16 digit image. In experiments, we randomly choose 250 samples from each class as our forth data set.

The proposed GCNMF is compared with the existing GB-SSL methods including LGC [20], LapSVM [2], CNMF [10] and GNMF [5]. According to [2,5] and [20], parameter μ in LGC is set to be 0.99, λ_1 and λ_2 in LapSVM are set to be 10^{-5} and $1/nz$ (here nz is number of nonzero entries in similarity matrix), and λ in GNMF is set to be 100. The label information is incorporated into the weight similarity M, i.e., $M_{ij} = 1$ if x_i and x_j have the same label, otherwise $M_{ij} = 0$.

Since the data sets are benchmark data which contain category information, all experimental results are evaluated via an external validation method, normalized mutual information (NMI) [19].

$$NMI = \frac{\sum_{j,l} n_{jl} log\left(\frac{n \cdot n_{jl}}{n_j n_l}\right)}{\sqrt{(\sum_j n_j log\frac{n_j}{n})(\sum_l n_l log\frac{n_l}{n})}}$$

where n_j, n_l are the number of samples in the jth class L_j and the lth cluster C_l respectively. n_{jl} is the number of samples occurring in both L_j and C_l, n is the total number of samples in the data set, and k is the number of classes.

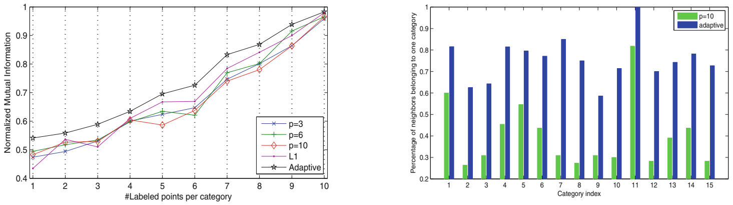
The larger the NMI, the better the learning performance. For all methods, the number of clusters is set to be the number of true classes.

4.2 Effect of Graph Construction

In this subsection, we carry out some experiments to test the effect of graph construction on the proposed method GCNMF. Traditionally, the local graph is built according to the adjacency matrix where the number of neighbors (p) for each point is predefined. However, it is not proper to fix p for all samples in real application as shown in Sect. 3, thus we introduced an adaptive method to select the neighbors for each sample according to its average similarity. The following experimental results will show the performance of GCNMF with additive graph construction.

Figure 2a shows the results of GCNMF with five different graphs on Yale Face data set. The first three experiments were conducted by fixing the number of neighbors p for all data points ($p = 3$, $p = 6$, $p = 10$ respectively), the fourth graph L_1 was built via sparse representation [17], and the fifth was conducted with the proposed adaptive graph construction method. It can be seen that the adaptive neighbor selection is helpful to improve the learning performance.

In order to investigate the properties of neighbors selected by the proposed method, we count the neighbors of points in each category as shown in Fig. 2b. Note that the category information is not used when adaptively selecting neighbors. For each category, the neighbors of all points are collected, and the percentage of neighbors belonging to the current category is calculated. From Fig. 2b, we can see that the adaptive method has ability to select more neighbors having the same category with the current point, which indicates that the selected neighbors and the current point have the similar structure. Therefore, the learning process can benefit from the adaptive graph construction method. For AT&T face data, COIL20 image data and USPS handwritten digits, the same results can be obtained.



(a) (b)

Fig. 2. Effect of graph construction on GCNMF for Yale Face data set: a) semi-supervised learning performance on unlabeled data under varying the labeled data sizes and b)comparison of the selected neighbor distribution by fixing p and adaptive method.

Table 1. Comparing NMI (%) on **Yale Face** under varying the labeled size.

♯ Labeled Data per Category	1	2	3	4	5	6	7	8	9	10
LGC	48.71	54.15	54.82	**66.48**	67.39	68.10	80.24	81.76	90.86	96.53
LapSVM	43.31	50.14	51.56	56.97	61.60	63.11	79.83	84.44	90.40	**98.28**
CNMF	48.88	54.39	53.37	59.72	59.33	58.41	72.55	75.77	82.20	96.53
GNMF	48.00	48.50	49.36	52.50	58.50	60.56	72.54	80.87	81.28	92.23
GCNMF	**54.11**	**55.84**	**58.90**	63.43	**69.57**	**72.58**	**83.30**	**86.88**	**93.86**	**98.28**

Table 2. Comparing NMI(%) on **AT&T Face** under varying the labeled size.

♯ Labeled Data per Category	1	2	3	4	5	6	7	8	9
LGC	66.25	66.39	67.86	68.75	70.83	73.00	75.00	77.50	80.00
LapSVM	73.33	83.75	85.00	86.67	87.86	88.75	89.83	90.00	91.00
CNMF	73.89	80.36	80.83	81.00	81.88	82.50	82.50	82.50	83.33
GNMF	**80.72**	82.64	85.89	87.81	88.38	91.27	92.53	93.66	96.96
GCNMF	80.56	**85.31**	**88.21**	**89.50**	**90.00**	**91.88**	**92.50**	**95.00**	**97.50**

Table 3. Comparing NMI (%) on **COIL20 Image** under varying the labeled size.

♯ Labeled Data per Category	4	8	12	16	20	24	28	32	36	40
LGC	60.39	63.52	66.85	68.73	74.50	77.09	84.82	86.93	87.95	87.69
LapSVM	68.02	68.36	73.22	76.07	80.00	82.90	84.57	86.14	85.60	85.73
CNMF	66.42	66.73	67.09	74.27	76.46	77.63	79.21	78.94	79.80	79.83
GNMF	68.83	**72.65**	**74.03**	75.35	76.04	79.53	81.92	86.71	87.50	88.11
GCNMF	**68.88**	69.64	72.89	**77.68**	**82.68**	**85.78**	**88.52**	**90.02**	**90.30**	**91.11**

Table 4. Comparing NMI (%) on **USPS Digit** under varying the labeled size.

♯ Labeled Data per Category	10	20	30	40	50	60	70	80	90	100
LGC	64.59	65.60	65.48	65.91	66.34	68.18	72.83	75.13	76.89	79.86
LapSVM	58.37	65.15	65.32	66.05	68.56	73.33	75.82	76.57	78.44	**83.63**
CNMF	54.86	65.12	65.43	66.66	66.79	67.49	68.31	70.53	72.63	72.34
GNMF	63.34	66.38	68.37	**71.71**	**73.70**	75.08	78.54	79.80	81.21	82.79
GCNMF	**64.86**	**68.47**	**69.87**	71.64	72.38	**75.13**	**79.34**	**80.96**	**82.49**	83.21

4.3 Effect of Labeled Data Size

In this subsection, we evaluate the influence of the labeled data size and compare GCNMF with LGC, LapSVM, CNMF and GNMF. For all methods, the adaptive method is used to build the graph. Tables 1, 2, 3 and 4 list the NMI values of learning results on four real world data sets. The best result in all techniques for the same data set is marked in dark.

For each data set, the number of labeled data per category is set according to the data size. For example, the size of labeled Yale Face per category is varying from 1 to 10. AT&T contains 40 categories and each category has 10 facial images, thus the number of labeled data is varying from 1 to 9. For COIL20 image and USPS digit, the number of labeled data is varying from 4 to 40, and from 10 to 100 respectively. For each labeled data size, we randomly selected labeled data ten times, and the average result is recorded. From Tables 1, 2, 3 and 4, it can be seen that the best results were obtained with the proposed GCNMF method in most cases. This indicates that the loss function based on graph constrained NMF is more adequate to these data sets than LGC and LapSVM.

Both GCNMF and GNMF perform better than CNMF, which demonstrates that the intrinsic structure among data is helpful to learn the prediction function in SSL. Comparing GCNMF with GNMF, they are different in terms of the graph introduction strategy (the graph contains both label information and the similarity between data samples). In GNMF, the graph only provides supervision on the updating process of F, while the graph supervises both W and H updating in GCNMF. Meanwhile, GCNMF is able to guarantee that data points sharing the same label can be mapped sufficiently close to each other, while GNMF can not, i.e., GNMF fails to make use of the label information.

5 Conclusions

In this paper, a new graph constrained non-negative matrix factorization model is designed for semi-supervised learning. The new model can make use of label information and the intrinsic structure among both labeled and unlabeled data. Another merit of the proposed model is parameter-free. Moreover, an adaptive graph construction method is provided to preserve the local consistency among data as correctly as possible. The experimental results on four real world data sets have shown the effectiveness of GCNMF. In this paper, we consider the Frobenius norm-based NMF cost function. In the future, the cost function based on KL-divergence and Earth Mover's distance will be further studied.

Acknowledgement. This work was supported by the Fundamental Research Funds for the Central Universities (2011JBM030, 2013JBZ005) and the National Natural Science Foundation of China (60905028, 61033013).

References

1. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Proceedings of NIPS (2001)
2. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. **735**, 2399–2434 (2006)
3. Blum, A., Chawla, S.: Learning from labeled and unlabeled data using graph mincuts. In: Proceedings of ICML, pp. 19–26 (2001)
4. Blum, A., Lafferty, J., Rwebangira, M., Reddy, R.: Semi-supervised learning using randomized mincuts. In: Proceedings of ICML, pp. 13–20 (2004)
5. Cai, D., He, X., Han, J., Huang, T.: Graph regularized nonnegative matrix factorization for data representation. IEEE Trans. PAMI **33**, 1548–1560 (2011)
6. Jebara, T., Wang, J., Chang, S.: Graph construction and *b*-matching for semi-supervised learning. In: Proceedings of ICML (2009)
7. Lee, D., Seung, H.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–791 (1999)
8. Lee, D., Seung, S.: Algorithms for non-negative matrix factorization. In: Proceedings of NIPS, pp. 556–562 (2001)
9. Lin, C.: On the convergence of multiplicative update algorithms for nonnegative matrix factorization. IEEE Trans. on Neural Netw. **18**, 1589–1596 (2007)
10. Liu, H., Wu, Z.: Non-negative matrix factorization with constraints. In: Proceedings of AAAI (2010)
11. Liu, W., Chang, J.S.: Large graph construction for scalable semi-supervised learning. In: Proceedigs of ICML (2010)
12. Miller, D., Uyar, H.: A mixture of experts classifier with learning based on both labeled and unlabelled data. In: Proceedings of NIPS, pp. 571–577 (1997)
13. Nadler, B., Srebro, N., Zhou, X.: Statistical analysis of semi-supervised learning: the limit of infinite unlabelled data. In: Proceedings of NIPS, pp. 1330–1338 (2009)
14. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**, 2323–2326 (2000)
15. Tenenbaum, J., Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000)
16. Wang, F., Zhang, T.C., Li, T.: Semi-supervised learning via matrix factorization. In: Proceedings of SIAM DM (2008)
17. Yan, S., Wang, H.: Semi-supervised learning by sparse representation. In: Proceedings of SIAM DM (2009)
18. Zhang, T., Oles, F.: A probability analysis on the value of unlabeled data for classification problems. In: Proceedings of ICML, pp. 1191–1198 (2000)
19. Zhong, S., Ghosh, J.: A comparative study of generative models for document clustering. In: Proceedings of SDW workshop on clustering high dimensional data and its applications, San Francisco, CA (2003)
20. Zhou, D., Bousquet, O., Lal, T., Scholkopf J.B.: Learning with local and global consistency. In: Proceedings of NIPS, Weston (2004)
21. Zhou, Z., Li, M.: Semi-supervised learning by disagreement. Knolwl. Inf. Syst. **24**, 415–439 (2010)
22. Zhu, X.: Semi-supervised learning literature survey. Technical Report 15304, University of Wisconsin, Madison (2006)
23. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of ICML (2003)

24. Zhu, Y., Jing, L., Yu, J.: Text clustering via constrained nonnegative matrix factorization. In: Proceedings of ICDM, pp. 1278–1283 (2010)
25. Zhuang, L., Gao, H., Huang. J., Yu, N.: Semi-supervised classification via low rank graph. In: Proceedings of ICIG, Huang (2011)

Kernel Parameter Optimization in Stretched Kernel-Based Fuzzy Clustering

Chunhong Lu(✉), Zhaomin Zhu, and Xiaofeng Gu

Key Laboratory of Advanced Process Control for Light Industry
(Ministry of Education), Department of Electronics Engineering,
Jiangnan University, Wuxi 214122, China
sharon0510@126.com, zhuzhaomin@gmail.com,
xgu@jiangnan.edu.cn

Abstract. Although the kernel-based fuzzy c-means (KFCM) algorithm utilizing a kernel-based distance measure between patterns and cluster prototypes outperforms the standard fuzzy c-means clustering for some complex distributed data, it is quite sensitive to selected kernel parameters. In this paper, we propose the stretched kernel-based fuzzy clustering method with optimized kernel parameter. The kernel parameters are updated in accordance with the gradient method to further optimize the objective function during each iteration process. To solve the local minima problem of the objective function, a function stretching technique is applied to detect the global minimum. Experiments on both synthetic and real-world datasets show that the stretched KFCM algorithm with optimized kernel parameters has better performance than other algorithms.

Keywords: Kernel fuzzy c-means · Kernel parameter · Optimization · Stretching technique

1 Introduction

Clustering algorithms are universally employed to partition patterns into a couple of smaller homogeneous groups. The fuzzy c-means (FCM) [1] algorithm, a typical one has been widely used in pattern recognition and image segmentation. The FCM algorithm, which applies Euclidean distance measure between objects and prototypes can obtain good clustering results for spherically-structured data, but cannot obtain effective clustering analysis for some complex distributed data such as the mixture structure with heterogeneous cluster prototypes and non-spherical geometry of data. The kernel-based fuzzy c-means (KFCM) [2] algorithm was then presented to overcome this drawback. A kernel function is defined to transform nonlinear distributed data to the higher dimensional feature space so that the naturally distributed data can be partitioned linearly. Obviously, the KFCM algorithm can improve the results of FCM algorithm by selecting appropriate kernel function and reasonable parameters. Due to its superiority to other kernel functions, the RBF kernel function has been employed in all four kernel clustering algorithms [3], which is thus also chosen in this work.

Z.-H. Zhou and F. Schwenker (Eds.): PSL 2013, LNAI 8183, pp. 49–57, 2013.
DOI: 10.1007/978-3-642-40705-5_5,
© Springer-Verlag Berlin Heidelberg 2013

However, the values of kernel parameters affect the performance of KFCM algorithm significantly. With respect to optimization of the kernel parameter, there are several methods in kernel-based techniques. Firstly, empirical, grid search and cross-validation methods are often applied to search the optimal kernel parameters [4]. On the one hand, users repeatedly execute their concerned algorithms for a couple of candidate kernel parameters according on their research experience until one corresponding to the best results is chosen as the final kernel parameter value. Unfortunately, because the number of these concerned candidate values is usually limited, the results of the kernel-based approaches are not distinctly preferable. On the other hand, the well-known cross-validation approach widely applied in the model selection can effectively obtain more favorable performance than selecting the parameter values empirically since the optimal value is chosen within a fairly broader range. But the kernel parameters need to be adjusted in real-time, which the cross-validation approach can not achieve on account of its time-consuming implementation. Secondly, the optimal kernel parameters can also be obtained by minimizing an objective function [5]. However, non-optimized parameters or sub-optimized values can be found by the above-mentioned methods. Although these kernel parameter learning methods can gain relative satisfactory evaluation results, searching the global optimized parameter is not promising. AL-Sultan and Fedjki [6] proposed a tabu search algorithm for globally solving fuzzy c-means clustering and obtained better performance in most of the test cases than the standard FCM algorithm.

In this paper, we propose an optimization method to select the kernel parameters for the KFCM algorithm, and employ a function stretching technique [7] to reformulate the objective function once it is trapped into a local minimum and to select the global optimized kernel parameter as the solution to the objective function. Incorporating the KFCM algorithm added by the stretching technique with optimization the kernel parameter is presented. Results of experiments on synthetic data sets and real data sets demonstrate that it is most important to select the desirable kernel parameters in kernel-based fuzzy clustering so that the KFCM algorithm with the best optimized parameters can outperform the other algorithms.

2 KFCM Algorithm

The classical FCM algorithm is remarkably effective only in partitioning spherical data based on the sum-of-squares error criterion. The KFCM algorithm has been adopted to overcome the problem in partitioning complex data with nonlinear boundaries by mapping nonlinear structure in the input space into a higher dimensional feature space. Given a data set $X = \{x_1,\ldots, x_N\}$, x_i $(i = 1\ldots N)$ from the input space R^p, a transformation function ϕ nonlinearly maps the data points to a higher dimensional feature space R^q. That is, $\phi : R^p \rightarrow R^q$, $p < q$. The inner product of two patterns obtained by the mapping function can be simply defined as a kernel function by using 'kernel trick' method:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j). \tag{1}$$

There are several examples of a kernel function. Using the RBF kernel as a kernel function in this paper, σ^2 as the variance parameter, we have

$$K(x_i, x_j) = e^{-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}}. \tag{2}$$

The kernel-induced Euclidean distance between patterns x_i and x_j in the input space can be calculated in the higher dimensional feature space through the kernel function $K(x_i, x_j)$as

$$d_{ij}^2(\phi(x_i), \phi(x_j)) = \left\|\phi(x_i) - \phi(x_j)\right\|^2 \tag{3}$$

$$\begin{aligned} &= \phi(x_i)\cdot\phi(x_i) - 2\phi(x_i)\cdot\phi(x_j) + \phi(x_j)\cdot\phi(x_j) \\ &= K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j). \end{aligned} \tag{4}$$

For a RBF kernel function, the distance measure in (3) using (2) can be easily described by

$$d^2(x_i, v_j) = 2(1 - 2K(x_i, v_j)). \tag{5}$$

The main goal of the KFCM algorithm is to minimize the following objective function as in the FCM.

$$J = \sum_{j=1}^{C}\sum_{i=1}^{N} u_{ij}^m d^2(x_i, v_j), \quad 2 \le C \le N, \tag{6}$$

where m represents the fuzzifier constant; C is the cluster number; N is the number of patterns; u_{ij} ∈[0,1], as elements of an order matrix $U_{C\times N}$, represents the membership degree of x_i in cluster j; V = (v_1,…,v_j,….v_c), v_j as centroids or prototypes. In addition, the elements of partition matrix U satisfy the following condition:

$$\sum_{j=1}^{C} u_{ij} = 1 \quad \text{and} \quad 0 < \sum_{i=1}^{N} u_{ij} < N, \forall j. \tag{7}$$

The memberships to minimize the objective function in (5) can be calculated by

$$u_{ij} = \frac{\left(1/d^2(x_i, v_j)\right)^{1/(m-1)}}{\sum\limits_{j=1}^{C}\left(1/d^2(x_i, v_j)\right)^{1/(m-1)}}, \tag{8}$$

and the formula of the prototypes v_j (j=1,2,…,C) proceeds as follows:

$$v_j = \frac{\sum\limits_{k=1}^{N} u_{jk}^m (x_k, v_j) x_k}{\sum\limits_{k=1}^{N} u_{jk}^m (x_k, v_j)}. \tag{9}$$

3 Learning the Kernel Parameters

The minimization problem of the objective function J given in (5) subjected to the constraint specified by Eq. (6) is solved by minimizing a constraint objective function defined as

$$\tilde{J}(\sigma) = \sum_{j=1}^{C}\sum_{i=1}^{N} u_{ij}^{m} d^2(x_i, v_j) + \sum_{j=1}^{C} \lambda_j (\sum_{i=1}^{N} u_{ij} - 1), \tag{10}$$

where λ_j ($j = 1,\ldots,C$) are Lagrangian multipliers. The optimal values of kernel parameter is obtained by minimizing (9),

$$\sigma^* = \arg\min_{\sigma} \tilde{J}(\sigma). \tag{11}$$

To calculate the kernel parameter σ, the general gradient method is applied to generate the optimal values σ^* by continually updating the following Eq. (11)

$$\frac{\partial \tilde{J}}{\partial \sigma} = -2\sum_{j=1}^{C}\sum_{i=1}^{N} u_{ij}^{m} K(x_i, v_j)\frac{\left\|x_i - v_j\right\|^2}{\sigma^3}, \tag{12}$$

$$\sigma^{(n+1)} = \sigma^{(n)} + \theta\left(\frac{\partial \tilde{J}}{\partial \sigma}\right). \tag{13}$$

Herein, θ is called the learning rate, commonly a positive constant and n is the iteration step. When $\left|\sigma^{(n+1)} - \sigma^{(n)}\right| < \varepsilon$, and ε is a very small positive number, the searching procedure reaches the convergence.

The objective function J is calculated according to Eq. (5) by serially updating the value of σ during executing kernel-based fuzzy clustering process until termination criteria satisfied. That is, the two consecutive Js almost remain unchanged. The iterative process can learn a minimizer of J. However, this acceptable minimizer may be local optimum in many cases. The stretching technique is applied for alleviating the local minima problem so that more optimized values of σ can be found. The idea of stretching technique is to perform a two-stage transformation of the objective function. In the first stage, local minima with higher functional values than the stretched local minimizer are eliminated, while other lower local minima remain unchanged; in second transformation stage, the neighbors of the local minimizer are stretched upwards and the local minimizer is turned into a local maximum. Thus the location of the global minimum is left unaltered in the final. This means a minimum of the objective function J can be obtained at the unchanged minimizer. Supposing an optimization objective function corresponds to a local minimize $\bar{\sigma}$ and the transformation function is defined as

$$G(\sigma) = J(\sigma) + \gamma_1\|\sigma - \bar{\sigma}\|(sign(J(\sigma) - J(\bar{\sigma})) + 1), \tag{14}$$

$$J_{new}(\sigma) = G(\sigma) + \gamma_2\frac{sign(J(\sigma) - J(\bar{\sigma})) + 1}{\tanh(\tau(G(\sigma) - G(\bar{\sigma})))}, \tag{15}$$

where γ_1, γ_2 and τ are arbitrary chosen positive constants, and $sign(\cdot)$ is the well-known triple valued sign function:

$$sign(x) = \begin{cases} 1, & if \quad x > 0, \\ 0, & if \quad x = 0, \\ -1, & if \quad x < 0. \end{cases}$$

Herein, $J(\sigma)$ is replaced by $J_{new}(\sigma)$ and a newly formulated $J(\sigma)$ is regarded as the objective function. The updating process repeating and stretching function transforming many times, the optimized solution to the original objective function can be obtained.

4 Stretched KFCM Algorithm with Optimal Kernel Parameter

Kernel-based fuzzy clustering algorithm is derived from the above section by the stretching technique and optimization of kernel parameter as follows:

Stretched kernel-based fuzzy c-means algorithm with optimal kernel parameter (SKFCM-opt *σ)*

Step 1. Set learning rate θ, the maximum iteration number n, stopping criterion ε, initial iteration k=0;γ_1>0, γ_2>0 and τ to a very small positive number.

Step 2. Initialize fuzzifier $m > 1$, usually set to 2, fuzzy partition U, prototypes v_j, kernel parameter σ_0 using Eq. (15).

Step 3. Update the memberships, cluster center based on formula (7) and (8); update kernel parameter σ according to Eq. (12) and $\sigma > 0$ must be satisfied during the iteration process. It is assumed that the kernel width exceeds zero at each iterating. If it is close to zero or a negative number, just giving 2ε to it in order to avoid the risk of degeneracy.

Step 4. Calculate the value difference of the objective function between consecutive iterations. Once a local minimizer is found, update the objective function J using newly-obtained J according to Eqs. (13) and (14).

Step 5. Repeat the total iteration process until termination criteria satisfied or maximum iterations reached.

Step 6. Select one minimizer that yields the best optimal solution of formula (5) and the minimizer is regarded as the best optimal value.

In this section, the computational complexity of the proposed algorithm is O ($C\ N\ q\ l$), where l is the iteration time in the algorithm implementing, C is clustering number, and N is the number of samples and q is feature attribute. The kernel matrix is calculated between the patterns and prototypes during each iteration. From the viewpoint of storage complexity, the space O (NC+Nq+$2Cq$) is used to storage samples, cluster prototypes and partition matrix. Additionally, the algorithm is proper for clustering some large datasets.

5 Experiments

The experimental results of SKFCM-opt are evaluated in this section to demonstrate effectiveness of the proposed method, compared with the other methods, KFCM with non-optimized σ (KFCM-σ_0), FCM with the tabu search technique (FCM-tabu) detailed in [6] and standard FCM without mapping respectively. An artificial data set Circles and two real datasets (Iris and Pendigits from UCI Machine Learning Repository [8]) are applied in these tests. We choose m=2 which is a common choice for fuzzy clustering. All obtained experimental results use the following parameters: average on 10 runs, iteration error ε=0.00001, max_iter=100, γ_1=3000, γ_2=0.5,$\tau = 10^{-5}$, θ=0.05.The initial value for the kernel parameter σ is set to

$$\sigma_0 = \frac{1}{C}\sqrt{\frac{1}{N}(\sum_{i=1}^{N} \|x_i - \bar{x}\|^2)}, \tag{16}$$

$\bar{x}$ is the centroid of the total patterns.

The Circles data set involves 200 data points in a 2-dimensional space and two classes which respectively contains 100 samples of rectangle (4 × 4) distribution and circular(radius is 4) distribution. The Circles data set is shown in Fig. 1.

The Iris data set contains 50 4-dimensional samples each of three species (Versicolor, Virginica, Setosa). One of the classes is well separated from the other two, while the remaining two are partly overlapped.

The Pendigits data set comprises of 16-dimensional 7494 samples each of ten classes. We randomly select 100 data for each of the four classes {1,3,5,7}, total 400 data used in the experiment.

The Defense Advanced Research Projects Agency (DARPA) 1998 Basic Security Module (BSM) datasets are increased for experimental results. Among these datasets,

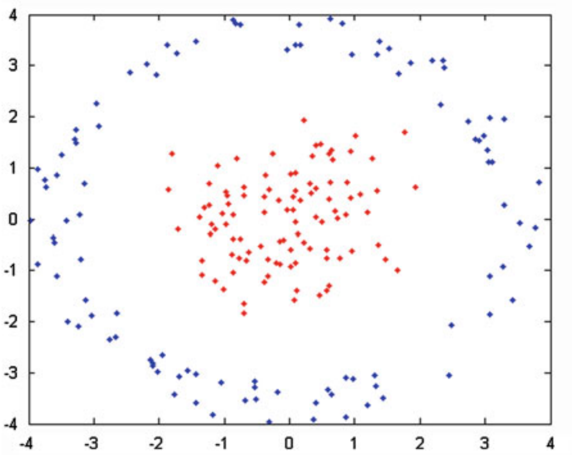


Fig. 1. The Circles data set

host-based BSM audit data from the seven-week training data and two weeks of testing data to evaluate the performance of the two algorithms, which involves two classes with 246 features, 7632 normal sessions and 456 attack sessions, respectively [9].

These data sets are used to analyze the quality of clustering. Table 1 shows the clustering accuracies of the presented method by comparing it with the other algorithms. The Circles data is turned into linear separation in the mapped feature space by the kernel function so that its clustering accuracy is evidently improved. In general, the performance of FCM-tabu method is better than the FCM algorithm, but is slightly worse than KFCM-σ_0 in the most cases. Additionally, SKFCM with the best kernel parameter (SKFCM-opt σ) obtains the best clustering accuracy among all these methods. It is apparently that stretching technique and optimization have an important influence on the clustering performance.

The clustering accuracies of SKFCM under a series of kernel parameter σ on Iris are shown in Fig. 2. It verifies the fact that the optimization of kernel parameters considerably affects the results of SKCM. Varying values of objective function under

Table 1. The clustering accuracies using FCM, FCM-tabu, KFCM and the proposed SKFCM with optimized σ for the aforementioned data sets.

Data sets	FCM (%)	FCM-tabu (%)	KFCM-σ_0 (%)	SKFCM-opt σ (%)
Circles data	51.16	78.24	93.24	100
Iris data	89.31	90.08	92.38	97.56
Pendigits data	42.82	48.34	59.74	67.23
DARPA data	40.52	54.05	53.61	65.29

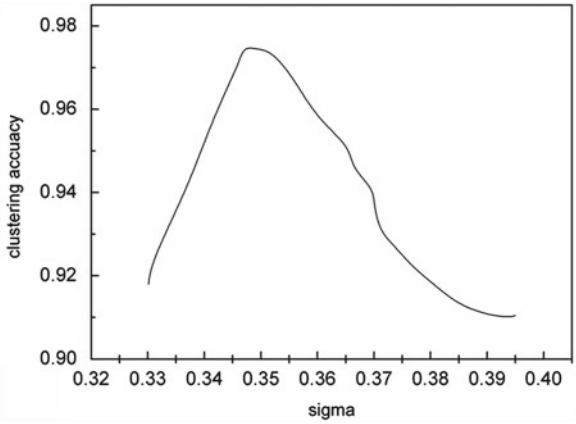


Fig. 2. Clustering accuracies under different kernel parameters σ on the Iris data

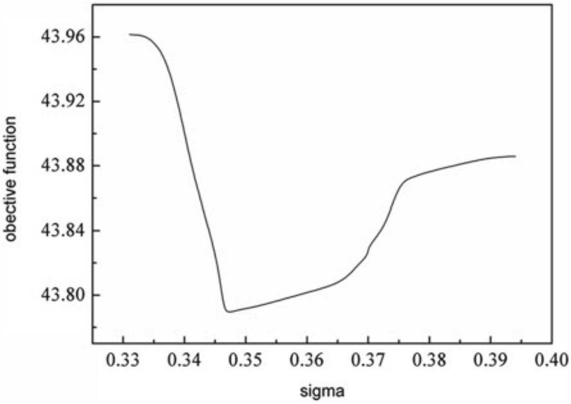


Fig. 3. Values of objective function under a series of σ values on the Iris data

a series of σ values on the Iris data is described in Fig. 3. From these two figures, we can observe when the proposed method achieves its best performance the objective function almost reaches its global minimum, given the same kernel parameter σ.

6 Conclusions

We have proposed a novel method for kernel-based fuzzy clustering. Based on the gradient method, the optimal parameter solution to the objective function of the KFCM algorithm is obtained, and a stretched technique reformulating the objective function can assure its optimal solution unaltered. Experimental results show that the proposed stretched kernel fuzzy clustering method with the optimal kernel parameter can be successfully applied compared with the KFCM algorithm with non-optimized σ, FCM with the tabu search algorithm and the standard FCM algorithm. However, there is no standard learning mechanism for evaluating the selection of the kernel parameters, a significant drawback of the kernel-based fuzzy clustering algorithms, which is worth of studying in the future work.

Acknowledgements. This work is supported by the PAPD of Jiangsu Higher Education Institutions, the Fundamental Research Funds for the Central Universities (JUDCF12027 JUSRP211A37) and the Fund of the State Key Laboratory of ASIC and System in Fudan University (11KF003).

References

1. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York (1981)
2. Wu, Z., Xie, W., Yu, J.: Kernel method-based fuzzy clustering algorithm. J. Syst. Eng. Electron. **16**, 160–166 (2005)

3. Girolami, M.: Mercer kernel-based clustering in feature space. IEEE Trans. Neural Netw. **13**, 780–784 (2002)
4. Wang, L., Chan, K.L.: Learning kernel parameters by using class separability measure. In: Proceedings of the Advances in Neural Information Processing Systems, NIPS (2002)
5. Zhang, D.Q., Chen, S.C., Zhou, Z.H.: Learning the kernel parameters in kernel minimum distance. Pattern Recognit. **39**, 133–135 (2006)
6. AL-Sultan, K.S., Fedjki, C.A.: A tabu search-based algorithm for the fuzzy clustering problem. Pattern Recognit. **30**, 2023–2030 (1997)
7. Parsopoulos, K.E., Vrahatis, M.N.: Recent approaches to global optimization problems through particle swarm optimization. Nat. Comput. **1**, 235–306 (2002)
8. Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J.: UCI machine learning repository. University of California, School of Information and Computer Sciences, Irvine. http://www.ics.uci.edu/~mlearn/MLRepository.html (1998)
9. Jeong, Y.S., Kang, I.H., Jeong, M.K.: A new feature selection method for one-class classification problems. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. **42**, 1500–1509 (2012)

Conscientiousness Measurement from Weibo's Public Information

Dong Nie, Lin Li, and Tingshao Zhu(✉)

Institute of Psychology, University of Chinese Academy of Sciences,
100190, CAS, Beijing, China
tszhu@psych.ac.cn, ginobilinie@gmail.com
http://ccpl.psych.ac.cn

Abstract. We apply a graph-based semi-supervised learning algorithm to identify the conscientiousness of Weibo users. Given a set of Weibo users' public information (e.g., number of followers) and a few labeled Weibo users, the task is to predict conscientiousness assessment for numeric unlabeled Weibo users. Singular value decomposition (SVD) technique is taken for feature reduction, and K nearest neighbor (KNN) method is used to recover a sparse graph. The local and global consistency algorithm is followed to deal with our data. Experiments demonstrate the advantage of semi-supervised learning over standard supervised learning when limited labeled data are available.

Keywords: Graph-based semi-supervised learning · Conscientiousness identification · KNN · SVD

1 Introduction

Personality can be defined as a set of characteristics which make a person unique, and the study of personality is of central importance in psychology. Among personality related researches, Big-Five theory is the mostly used one, it proposes five basic traits to form human personality: openness, conscientiousness, extraversion, agreeableness and neuroticism [10]. Conscientiousness, as one trait of Big-Five theory, is the state of being thorough, careful, or vigilant. Conscientious individuals are generally hard working and reliable. When taken to an extreme, they may also be "workaholics", perfectionists, and compulsive in their behavior. People who are low on conscientiousness are not necessarily lazy or immoral, but they tend to be more laid back, less goal-oriented, and less driven by success [16]. Conscientiousness measurement are mostly through self-report [27], which is time-consuming and sometimes, subjective. Some researchers made efforts to others-report [6,9,13], however, it is a big problem to achieve enough labeled data due to not only time-consuming and expensive, but also privacy problems.

Sina Weibo is now one of the most popular service in mainland China, it has already attracted more than 300 million user to register the service [23], and many people spend much time on the service, thus, researchers say it has become

Z.-H. Zhou and F. Schwenker (Eds.): PSL 2013, LNAI 8183, pp. 58–67, 2013.
DOI: 10.1007/978-3-642-40705-5_6,

© Springer-Verlag Berlin Heidelberg 2013

an important part of user's life [7]. Many researches were done to found out the relationship between microblog usage and user's personality [11,13,20].

In recent years, there has been a substantial amount of work exploring how to incorporate unlabeled data into supervised learning, and several semi-supervised learning approaches have been proposed [3,4,28,30,31]. Successful applications have been made in many areas, such as computer vision [2,29], and information retrieval [24]. Semi-supervised learning has also been used in the context of microblog classification [17]. In many scenes, semi-supervised learning algorithms outperforms standard supervised learning algorithms.

In this paper, we propose a graph-based semi-supervised learning approach [28] to the problem of conscientiousness measurement. We randomly collected 562 Weibo users' public information using Sina Weibo API, and retrieved corresponding conscientiousness extent through self-report method. Then local and global consistency method was used to combine the labeled and unlabeled data.

The rest of this paper is organized as follows. In Sect. 2, we introduce some related work. Section 3 will describe the dataset in detail. Section 4 will talk about the graphs. In Sect. 5, we present the detailed algorithm. Experiment results will be discussed in Sect. 6. At last, we will give a conclusion in Sect. 7.

2 Related Work

Personality analysis based on social media has received considerable attention recently [1,11,13,20]. They mainly collected internet data and corresponding labeled data, and then applied supervised learning approaches, such as, classification and regression analysis, to build the mode. However, the problem is that training data is always scarce, meanwhile large scale of unlabeled data is often easy to retrieve.

Many methods have been proposed to deal with the problem. Among them, semi-supervised learning algorithms received great attention in the past few years, since they could perfectly make use of unlabeled data. Generative models are used for text classification [24] with both labeled and unlabeled data, but the assumption is that the data should obey a certain distribution. Co-training algorithm has been proposed in [5], much research has been done, but it requires features can be spilt into two conditionally independent sets. Transductive support vector machines (TSVMs) have also been investigated by [8,19]. These years, many computer scientists are devoted to graph-based semi-supervised learning methods and propose a series of effective algorithms [3,4,28,31], and these algorithms have been widely used in many fields.

In this paper, local and global consistency (LGC) method [28] are integrated, because people with similar Weibo data is supposed to have the same personality. We gathered a number of Sina Weibo users' public information, and asked part of them to finish a big-five inventory [21] online. We only focus on one trait:conscientiousness in the work. We used singular value decomposition (SVD) to select features, and added a graph sparsification step to optimize the graph.

The LGC semi-supervised learning algorithm was taken to give a ternary classification over the collected data, and thus users' conscientiousness extents were assessed.

3 Dataset

The dataset consists of 562 copies of Sina Weibo users' information together with corresponding conscientiousness scores and much more non-labeled data. The Weibo data is about the user items on Weibo service, and it can be spilt into several categories as follows:

1. information in user's personal profile, including nickname, address, gender, birthday, personalized domain name, description and so on.
2. information about friends and followers, for example, the number of friends.
3. statistical information for statuses, such as average count of statuses per day and proportion of originality statuses.
4. basic setting information, for example, whether to allow all to comment.
5. tags information
6. trends information
7. others

We didn't focus much on users' behavior information in the work, for example, user's specific Weibo statuses. The conscientiousness scores are measured in continuous value. Since only a few users are labeled, and more users' Weibo data are available, it is very suitable to use semi-supervised learning techniques for our problem.

3.1 Data Collection

Using Sina Weibo API[1], we first collected 999,9999 Weibo user IDs, then randomly chosen 10,000 user IDs. Using Sina Weibo API again, we crawled 10,000 users' Weibo data which has been described above from these 10,000 Weibo IDs. In this way, we successfully captured a 10,000-Weibo-user dataset. Thirdly, using Weibo service's @ function, we invited the users to be volunteers to finish big-five inventory online. At last, we collected 562 copies of qualified questionnaires. Hence, we had 562 copies of conscientiousness extent data. The whole process took over one month, and volunteers had got reimbursement in return.

The collected Weibo-user dataset (562 copies of labeled Weibo data) is the basis of our experiment.

3.2 Feature Extraction

As our collected Weibo dataset is relatively simple (not contain much behavior data), the preprocessing work is straight forward. For some features, we followed

[1] http://open.weibo.com

Table 1. Part of Extracted Features

Feature	Description
allow_all_comment	Whether to allow all users to freely comment
bi_all_followers_count	The number of user's followers
bi_all_friends_count	The number of user's friends
description	The length of user's description
statuses_weibo_count	The number of user's statuses
original_rate	The rate of original statuses
screen_name_length	The length of screen name
users_tags_count_100	The number of tags whose popularity is less than 100
users_tags_count_100_1000	The number of tags whose popularity is between 100 and 10,000
users_tags_count_100_1000	The number of tags whose popularity is over 10,000
first_weibo_time	The average time to give first status per day
last_weibo_time	The average time to give last status per day
verified	Whether the user's Weibo account has been verified
...	...

the original data directly, for example, statistical information about statuses. For others, we used simple statistical methods to deal with data to extract features. We had totally extracted 45 features for a user from the Weibo data. Some of the features are listed in Table 1.

After feature extraction, we normalize the Weibo data to make the data equally measured as follows.

$$x = (x - MinValue)/(MaxValue - MinValue)$$

where x is the value of a dimension for a user, while $MinValue$ and $MaxValue$ respectively represent the maximum and the minimum value of this feature dimension for all users.

3.3 Conscientiousness Scores Discretization

As we received 562 copies of effective big-five personality questionnaires, we paid attention to one trait: conscientiousness only in the work. We calculated the conscientiousness score for each user according to their questionnaire based on corresponding rule, however, it is measured in continuous value, and it should be discretized.

In various personality researches [11,20,26], level of grouping method is often used to discretize continuous value of personality trait. Specifically speaking, we first calculate the mean (μ) and standard deviation (σ) for the sample, and then subjects whose conscientiousness scores are greater than ($\mu + \sigma$) are grouped to be high level, while subjects whose conscientiousness scores are lower than ($\mu - \sigma$) are grouped to be low level, and subjects whose scores are between ($\mu - \sigma$) and ($\mu + \sigma$) are grouped to be normal level. According to the definition of conscientiousness, individuals with high level mean they are conscientious, even workaholics to some degree, on the contrary, individuals with low level mean they tend to be more laid back, less goal-oriented, and less driven by success, individuals with middle level are supposed to be normal. To have a simple description later, we abbreviate the three levels as "conscientious", "immoral" and "normal" respectively. Therefore, the conscientiousness label set C can be represented by these three levels:

$$C = \{\text{``}conscientious\text{''}, \text{``}immoral\text{''}, \text{``}normal\text{''}\}$$

4 The Graphs

The semi-supervised conscientiousness measurement problem is described as follows. There are 562 Weibo users $x_1, x_2 ... x_{562}$, each represented by a set of features discussed above. We randomly choose $l \leq 562$ Weibo users from the labeled dataset, and suppose the l Weibo users to be labeled with $y_1, y_2 ... y_l \in C$ respectively. The remaining data is set to be unlabeled. The goal is to predict the categories of the unlabeled points using method from [28].

4.1 Feature Reduction

Usually, multidimensional data may be represented approximately in fewer dimensions due to redundancies in data, which may improve the prediction accuracy [25]. Since the original feature space has 45 dimensions, we attempt to take singular value decomposition (SVD) method [12], which is a well known matrix factorization technique, to reduce dimensionality of feature space [15,22]. We simply describes the SVD methods as follows.

Suppose $\underset{n\times 45}{A}$ be our original data space, then use SVD technique to factor A into three matrices:

$$\underset{n\times 45}{A} = \underset{n\times r}{U} \underset{r\times r}{\sum} \underset{r\times 45}{V}$$

where, matrix $\sum$ is a diagonal matrix containing the singular values of the matrix A, here are exactly r singular values, where r is the rank of matrix A. The rank of a matrix is the number of linearly independent rows or columns in the matrix, and it means independent information in our data space here.

To accomplish we can simply keep the first k singular values in $\sum$, where $k \leq r$. This will give us the best rank-k approximation to original data space A, and thus has effectively reduced the dimensionality of our original space. In our experiment, the dimensionality of original feature space is reduced to 34.

4.2 Graph Construction

We first compute measure similarity between Weibo user x_1 and user x_2 by features, and a larger similarity implies that the two users have more chances to be the same conscientiousness extent. Details can be found in Sect. 5.

An undirected graph $G = (V, E)$ is formalized with n nodes V, and weighted edges E among the nodes. Each Weibo user is a node in the graph, including the unlabeled users. The node of labeled user is also labeled with conscientiousness extent in the graph. Each Weibo user is connected to any other users by similarity computed between the two users, no matter the user is labeled or not. Then a fully connected graph is constructed.

4.3 Graph Sparsification

As described above, the graph for our semi-supervised learning problem is a fully connected one. To ensure that the semi-supervised learning algorithm remain efficient and robust to noise, a sparse weighted subgraph from the fully connected graph is needed. There are few researches on graph construction [18], though graph-based semi-supervised learning has received much attention recently. K Nearest Neighbors (KNN) is the most common used algorithm to recover a sparse subgraph. Roughly speaking, each node merely connects to its k nearest neighbors to form a subgraph. Specifically, for each point in the fully connected graph, using similarities in the fully connected graph, searches for the k nearest points to it without considering itself. Thus we can recover a sparse subgraph with this method.

5 Algorithms

We use the simple Local and Global Consistency (LGC) algorithm [28] on the Weibo dataset, the following formula depicts the essence of this algorithm:

$$\min_f \left\{ \sum_{i=1}^{l} \left(f(x_i) - y_i\right)^2 + f^T \Delta f \right\}$$

where $f(x)$ is the decision function, y is the label for each node and Δ is the graph combinatorial Laplacian. The LGC method allows $f(x_l)$ to be different from y_l with penalty term, in other words, this method can accommodate noise.

We choose the following function to compute similarity between two users:

$$W_{ij} = \exp\left({}^{-d(x_i, x_j)}\!/_{2\sigma^2} \right)$$

where

$$d(x_i, x_j) = \sqrt{\sum\nolimits_{k=1}^{t} (x_{ik} - x_{jk})^2}$$

where t is the dimension of the feature space.

We now describe the LGC algorithm in detail:

1. Form the affinity matrix W using the function discussed above.
2. Recover a sparse affinity matrix using KNN method.
3. Construct the matrix $S = D^{-1/2} W D^{-1/2}$, where $D = diag(\sum_j w_{ij})$.
4. Iterate $F(t+1) = \alpha S F(t) + (1 - \alpha) Y$ until convergence, where $\alpha \in (0, 1)$.

5. Let F^* denote the limit of the sequence $\{F(t)\}$. Label each point x_i as a label $y_i = \arg\max_{j \leq c} {F_{ij}}^*$.

The above algorithm has been proved convergent, and we can computer F^* without iterations:

$$F^* = (I - \alpha S)^{-1} Y$$

where I is a identity matrix.

The LGC method described above solves a set of linear equations so that the predicted label of each example is the result of considering local and global consistency. The algorithm makes it that nearby points are likely to have the same label and points on the same structure are likely to have the same label, too. This algorithm successfully makes use of a amount of labeled and unlabeled points to build a classification, moreover it doesn't limit to binary classification problems, therefore, this method is suitable for our Weibo dataset.

6 Experiment Results

As it is very difficult to collect labels for the entire 10,000-Weibo-user dataset, we can only conduct experiments on the 562-Weibo-user dataset.

We evaluated LGC method on the Weibo user conscientiousness measurement tasks. For each task, we gradually increased the labeled set size systematically, performed 10 random trials for the labeled set size. In each trial we randomly sampled a labeled set with the specified 562-Weibo-user dataset, if a class was missing from the sampled labeled set, we redid the random sampling, and the remaining data were used as the unlabeled set. We select the parameter combination by a grid search with the three parameters (k, σ, α), and the results are as follows. The k in KNN sparsification is set to 4, the parameter in the weight function is selected as $\sigma = 2$, and the iteration coefficient (α) is set to 0.2.

We report the classification accuracies with LGC method on two graphs: fully connected graph and KNN sparse graph. And we also compare two feature space: original and compressed feature space. To compare the graph-based semi-supervised learning algorithm against a standard supervised learning algorithm, we choose one-vs-rest SVMs as baseline method, and we use SVM-Light (svmlight.joachims.org/) as the tool to implement our ternary classification problems. Specifically, we create three binary classifications, one for each class against all the other classes, and select the class with the largest margin. We choose RBF kernel for the SVM classifications, and the width of RBF kernel is set to 1.4. The results are presented in Fig. 1.

"SVD" means using Singular Value Decomposition method to deal with original feature space, "sparse" indicates using KNN methods to recover a sparse subgraph, and "original" denotes original feature space and fully connected graph. As shown in Fig. 1, we find that LGC method outperforms the RBF kernel SVM baseline a lot, especially when the labeled set size is small. LGC method with SVD processing and KNN sparsification performs best over all methods. When the labeled data set comes to 50, the accuracy can approximate 80 %, which is a good performance in conscientiousness measurement. The accuracy

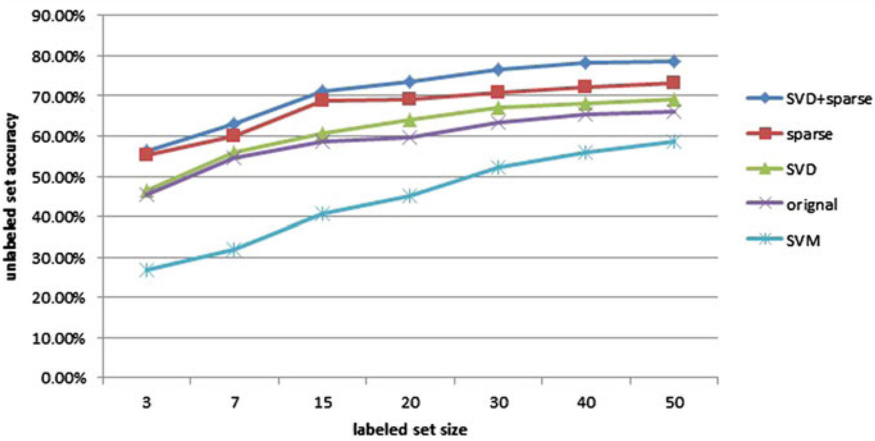


Fig. 1. The Accuracy of LGC and SVM.

can be improved to a certain extent if we use KNN method to construct a sparse subgraph. Using SVD method to compress feature space can also improve the classification accuracy in a small degree, it may be because manifold learning is not the best choice in semi-supervised learning for truly high dimensional data [14].

7 Conclusions

In this paper, we have investigated a local and global consistency based semi-supervised learning algorithm for conscientiousness measurement, which conforms two assumptions: similar examples should have similar labels and examples in similar structure should have similar labels. The algorithm makes full use of both labeled data and unlabeled data.

In our research, we obtained a set of Sina Weibo data, with 562 conscientiousness labels. We conducted our experiments over the dataset. Singular Value Decomposition (SVD) method was used to perform dimensionality reduction and K nearest neighbor (KNN) method was used to recover a sparse subgraph. Then the local and global consistency (LGC) algorithm was taken to give a ternary classification over the dataset. Our experiment shows that LGC algorithm achieves better performance when only very few labeled examples are available.

Apparently, there exists vast space we can do to promote the performance. Next, we will focus on model selection and parameter selection to better construct the graph. Meanwhile, we will try to incorporate more background knowledge into the learning process. In the future, we will use semi-supervised learning method to predict personality traits on large dataset.

Acknowledgments. The authors gratefully acknowledge the generous support from National High-tech R&D Program of China (2013AA01A606), NSFC (61070115), Institute of Psychology (113000C037), Strategic Priority Research Program (XDA06030800) and 100-Talent Project (Y2CX093006) from Chinese Academy of Sciences.

References

1. Hakura, J., Minamikawa, A., Fujita, H., Kurematsu, M.: Personality estimation application for social media. In: Fujita, H., Revetria, R. (eds.) Frontiers in Artificial Intelligence and Applications. IOS Press, The Netherlands (2012)
2. Balcan, M.-F., Blum, A., Choi, P.P., Lafferty, J., Pantano. B., Rwebangira, M.R., Zhu, X.: Person identification in webcam images: an application of semi-supervised learning. In ICML 2005 Workshop on Learning with Partially Classified Training Data (2005)
3. Niyogi, P., Sindhwani, V., Belkin, M.: On manifold regularization. In: Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (2005)
4. Blum, A., Chawla, S.: Learning from labeled and unlabeled data using graph mincuts. In: Proceedings of the 18th International Conference on Machine Learning (2001)
5. Mitchell, T., Blum, A.: Combining labeled and unlabeled data with co-training. In: Proceedings of the Workshop on Computational Learning Theory, COLT (1998)
6. Buchanan, T., Smith, J.L.: Using the internet for psychological research: personality testing on the world wide web. British J. Psychol. **90**(1), 125–144 (1999)
7. Cao, B.: Sina's weibo outlook buoys internet stock gains: China overnight. Technical report, Bloomberg (2012)
8. Chapelle, O., Sindhwani V., Keerthi, S.S.: Branch and bound for semisupervised support vector machines. In: Advances in Neural Information Processing Systems (NIPS) (2006)
9. Sumner, C., Byers, A., Shearing, M.: Determining personality traits and privacy concerns from facebook activity. In: Black Hat Briefings, 11 (2011)
10. Funder, D.C.: Personality. Annu. Rev. Psychol. **52**, 197–221 (2001)
11. Golbeck, J., Robles, C., Turner, K.: Prediciting personality with social media. In: Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems, pp. 253–262. ACM (2011)
12. Golub, G.H., Reinsch, C.: Singular value decomposition and least squares solutions. Numer. Math. **14**, 403–420 (1970)
13. Gosling, S., Augustine, A., Vazire, S., Holtzman, N., Gaddis, S.: Manifestations of personality in online social networks: Self-reported facebook related behaviors and observable prole information. Cyberpsychol. Behav. Soc. Netw. **14**, 483–488 (2011)
14. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: Weiss, Y., Saul, L.K., Bottou, L. (eds.) Advances in Neural Information Processing Systems 17. MIT Press, Cambridge (2005)
15. Kamp, Y., Bourlard, H.: Auto-association by multilayer perceptrons and singular value decomposition. Biol. Cybern. **59**, 291–294 (1988)
16. Ones, D.S., Hogan, J.: Conscientiousness and integrity at work. In: Hogan, R., Johnson, J., Briggs, S. (eds.) Handbook of Personality Psychology. Academic Press, San Diego (1997)
17. Zheng, H., Yoshinaga, N., Kaji, N., Toyoda, M.: A study on microblog classification based on information publicness. In: DEIM Forum (2012)

18. Jebara, T., Wang, J., Chang, S.-F.: Graph construction and b-matching for semi-supervised learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, pp. 441–448. ACM, New York (2009)
19. Joachims, T.: Transductive inference for text classification using support vector machines. In: Proceedings of the 16th International Conference on Machine Learning, pp. 200–209. Morgan Kaufmann, San Francisco (1999)
20. Qiu, J.R.L., Lin, H., Yang, F.: You are what you tweet: Personality expression and perception on twitter. J. Res. Pers. **46**, 710–718 (2012)
21. Deary, I., Whiteman, M., Matthews, G.: Personality Traits. Cambridge University Press, Cambridge (2006)
22. Rocha, L.M., Wall, M.E., Rechtsteiner, A.: Singular value decomposition and principal component analysis. In: Berrar, D.P., Dubitzky, W., Granzow, M. (eds.) A Practical Approach to Microarray Data Analysis, pp. 91–109. Kluwer, Norwell (2003)
23. Millward, S.: China's forgotten 3rd twitter clone hits 260 million users. Technical report, techinasia.com. 22 Oct 2012
24. McCallum, A.K., Thrun, S., Nigam, K., Mitchell, T.M.: Learning to classify text from labeled and unlabeled documents. In: AAAI-98, 15th Conference of the American Association for Artificial Intelligence, pp. 792–799 (1998)
25. Furnas, G.W., Landauer, T.K., Deerwester, S., Dumais, S.T., Harshman, R.: Indexing by latent semantic analysis. J. American Soc. Inf. Sci. **41**(6), 391–407 (1990)
26. Wiesner, W.H., Kichuk, S.L.: The big five personality factors and team performance: implications for selecting successful product design teams. J. Eng. Technol. Manag. **14**, 195–221 (1997)
27. Thompson, E.R.: Development and validation of an international english big-five mini-markers. Pers. Individ. Differ. **45**(6), 542–548 (2008)
28. Bousquet, O., Lal, T., Weston, J., Zhou, D., Schlkopf, B.: Learning with local and global consistency. Adv. Neural Inf. Process. Syst. **16**, 321–328 (2004)
29. Chen, K.-J., Zhou, Z.-H., Dai, H.-B.: Enhancing relevance feedback in image retrieval using unlabeled data. CM Trans. Inf. Syst. **24**, 219–244 (2006)
30. Zhou, Z.-H., Li, M.: Semi-supervised regression with co-training. In: International Joint Conference on Artificial Intelligence (IJCAI) (2005)
31. Ghahramani, Z., Zhu, X., Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions. In: The 20th International Conference on Machine Learning (ICML) (2003)

Meta-Learning of Exploration and Exploitation Parameters with Replacing Eligibility Traces

Michel Tokic[1,2](✉), Friedhelm Schwenker[1], and Günther Palm[1]

[1] Institute of Neural Information Processing, University of Ulm,
89069 Ulm, Germany
michel@tokic.com,
{guenther.palm, friedhelm.schwenker}@uni-ulm.de

[2] Institute of Applied Research, University of Applied Sciences,
Ravensburg-Weingarten, 88250 Weingarten, Germany

Abstract. When developing autonomous learning agents, the performance depends crucially on the selection of reasonable learning parameters, for example learning rates or exploration parameters. In this work we investigate meta-learning of exploration parameters by using the *"REINFORCE exploration control"* (REC) framework, and combine REC with replacing eligibility traces, which are a basic mechanism for tackling the problem of delayed rewards in reinforcement learning. We show empirically for a robot example and the mountain–car problem with two goals how the proposed combination can help to improve learning performance. Furthermore, we also observe that the setting of time constant λ is not straightforward, because it is intimately interrelated with the learning rate α.

1 Introduction

Controlling exploration and exploitation is one of the main challenges when developing autonomous learning agents. In general, exploratory actions lead to an increase in knowledge about the long-term utility of actions (long-term optimization), but often may cause the income of negative reward due to randomly selected bad actions. However, exploiting knowledge (short-term optimization) may also lead to sub-optimal action selections if the utility of an optimal action is underestimated. As a consequence, the *dilemma between exploration and exploitation* arises [1].

Several approaches exist to tackle this problem in reinforcement learning. Using action counters for determining confidence intervals is a popular approach in the domain of machine learning [2–4]. In contrast, neurobiologically inspired models utilize the immediate reward [5] or the temporal-difference error [6] to adapt the amount of exploration, e.g. by the meta-parameter τ of Softmax action selection. In a more recent approach, we proposed to use the *value difference* (the product between the temporal-difference error and the learning rate) as an indicator for the *uncertainty of knowledge about the environment* [7]. This indicator is used for controlling the action-selection policy between Greedy and

Z.-H. Zhou and F. Schwenker (Eds.): PSL 2013, LNAI 8183, pp. 68–79, 2013.
DOI: 10.1007/978-3-642-40705-5_7,
© Springer-Verlag Berlin Heidelberg 2013

Softmax, which aims at robustness with regard to stochastic rewards and even non-stationary environments [8].

In this paper we consider the problem of adapting the amount of exploration and exploitation in model-free reinforcement learning. We combine our recently proposed *"REINFORCE exploration control"* (REC) policies [9,10] with replacing eligibility traces [11], for tackling the problem of delayed rewards. We investigate the proposed algorithm on a reward model of a crawling robot that learns to walk forward through sensorimotor interactions, and also on the mountain–car problem with two goals.

2 Methods

We investigate the problem of maximizing an agent's cumulative reward over time, which in our experiments can be described as learning in a Markovian Decision Process (MDP) in discrete time $t \in \{0, 1, 2, \dots\}$ [1]. In general, an MDP consists of a finite set of states $\mathcal{S}$ and a finite set of possible actions in each state, $\mathcal{A}(s) \in \mathcal{A}, \forall s \in \mathcal{S}$. A transition function $P(s, a, s')$ describes the (stochastic) behavior of the environment, i.e. the probability of reaching successor state s' after selecting action $a \in \mathcal{A}(s)$ in state s. After the selection of an action, a reward $r \in \mathbb{R}$ is received from the environment and the agent finds itself in a successor state $s' \in \mathcal{S}$. The choice of action a is significant, and therefore the general goal is to find an optimal action-selection policy, $\pi^* : \mathcal{S} \rightarrow \mathcal{A}$, that maximizes the cumulative reward.

2.1 Action-Value Functions

Action-selection policies can either be learned using model-based techniques (the model of rewards and dynamics are approximated separately) or model-free techniques (only the value function is learned) [1]. Both require learning a value function for the prediction of future reward. In the following, we are particularly interested in optimizing model-free techniques, for the reason of being closely related to reinforcement learning in the brain [12,13].

Action-selection policies can be derived from value functions representing so far learned knowledge about the future reward [1]. An action-value function, $Q(s, a)$, approximates the cumulative discounted reward for following policy π, when starting in state s and taking action a,

$$Q(s,a) = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a \right\}, \tag{1}$$

where $0 < \gamma \leq 1$ is a discounting factor used for weighting future rewards in $Q(s, a)$. Since $Q(s, a)$ depends on rewards received in the future, the cumulative reward is considered to be an expected value $E_\pi\{\cdot\}$ which depends on the action-selection policy π. The parameter γ is allowed to take on the value of 1 only in episodic learning problems, i.e. an episode must terminate after a maximum of T steps, which prevents $Q(s, a)$ from growing to an infinite sum. In case a fixed horizon does not exist, γ must to be chosen < 1.

2.2 Q-Learning with Replacing Eligibility Traces

The action-value function is sampled from interactions with the environment. For this we use Watkin's Q-learning algorithm [14], which adapts reward estimates according to:

$$b^* \leftarrow \arg\max_{b \in \mathcal{A}(s')} Q(s', b) \tag{2}$$

$$Q(s,a) \leftarrow Q(s,a) + \alpha \underbrace{\left(r + \gamma Q(s', b^*) - Q(s,a)\right)}_{\text{Temporal-Difference Error } \Delta} \tag{3}$$

where $0 < \alpha \leq 1$ denotes a step-size parameter [15]. For an optimal action-value function, $Q^*(s,a)$, from which the optimal (greedy) policy π^* can be derived, the temporal-difference error is zero for all observation tuples (s, a, r, s').

As proposed by Singh and Sutton [1,11], we also combine Q-learning with *replacing eligibility traces*, which is known as $Q(\lambda)$-learning as shown in Algorithm 1. The advantage is that rewards are propagated faster to previously taken actions, which tackles the general problem of accounting delayed rewards in reinforcement learning. Visited state-actions pairs $e(s_e, a_e)$ are memorized in an *eligibility trace list*, where each entry in this list is associated with an additional memory variable, the *eligibility trace*, used for weighting the current temporal-difference error Δ also in the action value $Q(s_e, a_e)$ of previously taken actions. The trace for the last taken action is set to $e(s,a) \leftarrow 1$, which means that the temporal-difference error is fully credited.

All traces are decayed by $\gamma\lambda$ after taking an action. In this sense $0 \leq \lambda \leq 1$ denotes a time constant. For $\lambda = 0$, credit is only assigned to the last taken action, where Eq. (3) is only computed for the most recent observation tuple (s, a, r, s'). On the other hand, a time constant of $\lambda = 1$ refers to Monte-Carlo backups. Eligibility traces $e(s_e, a_e)$ are removed from the list as soon as their value becomes lower than a small threshold Θ, and thus the parameter λ does implicitly control over the amount of actions the current temporal-difference error is propagated into the past. The list of eligibility traces is also cleared in case an exploratory action is selected, because Q-learning learns the action-value function independently of the actual policy (called off-policy learning), assuming to follow a greedy policy in the limit for $t \rightarrow \infty$. Therefore, crediting the reward of exploratory actions to previously taken actions has no necessary relationship to a greedy policy [1].

3 Exploration and Exploitation

In the following we describe two basic strategies for deriving action-selection probabilities $\pi(s,a)$ from the action-value function, and afterwards we show how to combine them for meta-parameter learning.

The ε-Greedy policy selects a uniform randomly distributed action with probability $0 \leq \varepsilon < 1$ [1]. With probability $1 - \varepsilon$, a greedy action from the set $\mathcal{A}^*(s)$

Algorithm 1 Robot control: $Q(\lambda)$-learning with local REC adaption of ε-Greedy

```
Initialize $Q$ arbitrarily, e.g. $Q(s,a) = 0$ for all $s, a$
Initialize start state, e.g. $s \leftarrow \{g_x = g_y = 0\}$

repeat
  EXPLORATION / EXPLOITATION:
  $\xi \leftarrow$ random number from the interval $(0,1)$
  draw $\varepsilon(s)$ from a Gaussian distribution: $\varepsilon(s) \sim \mathcal{N}(\mu(s), \sigma(s))$

  if $\xi < \varepsilon(s)$ then
    $a \leftarrow$ random action from the set $\mathcal{A}(s)$
  else
    $a \leftarrow \arg\max_{b \in \mathcal{A}(s)} Q(s,b)$
  end if

  take action $a$
  observe reward $r$ and successor state $s'$

  COMPUTE TEMPORAL-DIFFERENCE ERROR:
  $\Delta \leftarrow r + \gamma \max_{b \in \mathcal{A}(s')} Q(s',b) - Q(s,a)$

  ELIGIBILITY-TRACE DECAY / CLEANUP:
  $e(s,a) \leftarrow 1$
  $V(s) \leftarrow \max_{b \in \mathcal{A}(s)} Q(s,b)$
  if $V(s) == Q(s,a)$ then
    for all $(s_e, a_e)$ in e-trace list: do
      $Q(s_e,a_e) \leftarrow Q(s_e,a_e) + \alpha \Delta e(s_e,a_e)$
      $e(s_e,a_e) \leftarrow \lambda\gamma e(s_e,a_e)$
      if $e(s_e,a_e) < \Theta$ then
        mark $e(s_e,a_e)$ for deletion
      end if
    end for
    delete marked eligibility traces
  else
    $Q(s,a) \leftarrow Q(s,a) + \alpha\Delta$
    clear e-trace list
  end if

  LOCAL REC ADAPTATION:
  $\rho \leftarrow r + \gamma \max_{b \in \mathcal{A}(s')} Q(s',b)$
  $\mu(s) \leftarrow bound\left[\mu(s) + \alpha(\rho - \bar{\rho}(s))\,(\varepsilon(s) - \mu(s))\right]$
  $\sigma(s) \leftarrow bound\left[\sigma(s) + \alpha(\rho - \bar{\rho}(s))\frac{(\varepsilon(s)-\mu(s))^2-\sigma(s)^2}{\sigma(s)}\right]$
  $\bar{\rho}(s) \leftarrow \bar{\rho}(s) + \alpha(\rho - \bar{\rho}(s))$

  $s \leftarrow s'$
until robot is switched off
```

of so far estimated optimal actions in state s is selected:

$$\mathcal{A}^*(s) = \arg\max_a Q(s,a)$$
$$\pi_{\mathrm{EG}}(s,a) = \begin{cases} \frac{1-\varepsilon}{|\mathcal{A}^*(s)|} + \frac{\varepsilon}{|\mathcal{A}(s)|} & \text{if } a \in \mathcal{A}^*(s) \\ \frac{\varepsilon}{|\mathcal{A}(s)|} & \text{otherwise} \ . \end{cases} \qquad (4)$$

Note that in any state s all selection probabilities $\pi_{\mathrm{EG}}(s,a)$ sum up to 1. A drawback of ε-Greedy is the choice of uniformly selected random actions, which might cause the income of negative reward due to bad actions. However, the ε-Greedy policy is reported for being hard to beat when a proper exploration parameter ε is configured [16].

The second policy we investigate is the Softmax policy, which selects an action according to its weighting in a Boltzmann distribution [1]:

$$\pi_{\mathrm{SM}}(s,a) = \frac{\exp\left(\frac{Q(s,a)}{\tau}\right)}{\sum_b \exp\left(\frac{Q(s,b)}{\tau}\right)} \ . \qquad (5)$$

This policy utilizes a positive parameter τ, called temperature, which controls between exploration and exploitation. High values of τ lead to equally distributed random actions, however low values to greedy actions. Again, in any state s all selection probabilities $\pi_{\mathrm{SM}}(s,a)$ sum up to 1.

In general, it is a problem to define global constants $\tau > 0$ and $\varepsilon \in [0,1]$ for achieving reasonable performance. Therefore, these parameters are typically initialized with a high value at the beginning of an experiment, being decreased over time. Especially for large state-action spaces, such fine-tuning is most often a time-consuming process.

3.1 Meta-Learning of Exploration/Exploitation Parameters

A drawback of ε-Greedy and Softmax is the *optimism in the face of uncertainty*. For this we recently proposed *"Value-Difference Based Exploration with Softmax"* (VDBE-Softmax) [7], which controls exploration and exploitation between Greedy and Softmax in a meta-learning fashion. A local exploration rate $\varepsilon(s)$, initialized by 1, is assigned to each state in the state space, which denotes the probability of selecting an exploratory action in state s. The selection probabilities $\pi(s,a)$ are adapted in dependence of fluctuations in the action-value function, which are considered to be a measure of the uncertainty in knowledge about the environment. Furthermore and in contrast to the ε-Greedy policy, exploratory actions are not selected equally distributed, but weighted according to the Softmax rule, which prevents selecting of very bad actions due to their weighting in the Boltzmann distribution. This idea of combining both policies was introduced by Wiering [17], called *Max Boltzmann Exploration* (MBE), but without learning of the policy parameters. Instead he configures ε and τ (globally) by hand.

The general idea of VDBE-Softmax is that high fluctuations of the action-value function should lead to a high degree of exploration, i.e. $\varepsilon(s) \to 1$, because the observation is insufficiently approximated by the prediction. On the other hand, when the prediction about future reward is well approximated, so far learned knowledge should be exploited, i.e. $\varepsilon(s) \to 0$. In this sense the corresponding exploration rate in state s is updated after each value-function backup for any action a in state s according to:

$$\varepsilon(s) \leftarrow \varepsilon(s) + \delta \left[1 - \exp\left(\frac{-|\alpha \mathit{\Delta}|}{\phi} \right) - \varepsilon(s) \right], \tag{6}$$

where ϕ is a positive parameter for the *sensibility* with regard to the absolute value difference $|\alpha \mathit{\Delta}|$ [7]. In case an exploratory action should be selected, all Q-values in state s are scaled into the interval $[-1, 1]$, and the action is selected according to Eq. (5) using $\tau = 1$. As proposed in [7,18] the learning rate δ can be determined online by the inverse of the number of actions, i.e. $\delta = \frac{1}{\mathcal{A}(s_t)}$ in the current state s_t.

Kobayashi and colleagues [6] proposed a similar approach (using the temporal-difference error $\mathit{\Delta}$), but adapting the parameter τ of Softmax in a global manner instead. In contrast, VDBE-Softmax is a state-based strategy, which has the advantage of selecting exploratory actions only in states where the observation is insufficiently approximated by the action-value function.

The extension of VDBE-Softmax for $Q(\lambda)$-learning is straightforward. We simply need to apply the learning rule right after each action-value update (line 19 of Algorithm 1), but additionally including the eligibility trace for state s_e and action a_e. Therefore Eq. (6) is slightly modified to:

$$\varepsilon(s_e) \leftarrow \varepsilon(s_e) + \delta \left[1 - \exp\left(\frac{-|\alpha \mathit{\Delta} e(s_e, a_e)|}{\phi} \right) - \varepsilon(s_e) \right] \quad . \tag{7}$$

3.2 REINFORCE Exploration Control

We recently proposed a further alternative for meta-learning of exploration parameters called *REINFORCE Exploration Control* (REC) [9,10]. The general idea is to control the exploration parameter of any above mentioned policies[1] by a gradient-following algorithm. The heart of REC is Williams *"REINFORCE with multiparameter distributions"* algorithm [19] for reinforcement learning in continuous action spaces.

In REC the exploration parameter of a policy is considered to be an continuous action. In this sense, we proposed two variants: (1) the global (episodic) variant selecting an exploration parameter for the duration of an learning episode, and (2) the local (stepwise) variant selecting a state-based exploration parameter before actually selecting an action by one of the above policies. In any of both

[1] ε-Greedy: ε; Softmax: τ; MBE: ε; VDBE-Softmax: ϕ.

cases, the exploration parameter is drawn according to a Gaussian distribution with parameters μ (mean) and σ (standard deviation). For example, in the local variant the parameter of ε-Greedy is selected according to:

$$\varepsilon(s) \sim \mathit{bound}\left[\mathcal{N}\left(\mu(s), \sigma(s)\right)\right] \ , \tag{8}$$

where $\mathit{bound}[\cdot]$ ensures that $\varepsilon(s)$ stays within the interval $[0, 1]$. After taking action a according to the policy, and after observing the successor state s' and reward r, the local distribution parameters $\mu(s)$ and $\sigma(s)$ are adapted according to a reinforcement comparison scheme:

$$\mu(s) \leftarrow \mathit{bound}\left[\mu(s) + \alpha_R(\rho - \bar{\rho}(s))\frac{\varepsilon(s) - \mu(s)}{\sigma(s)^2}\right] \tag{9}$$

$$\sigma(s) \leftarrow \mathit{bound}\left[\sigma(s) + \alpha_R(\rho - \bar{\rho}(s))\frac{(\varepsilon(s) - \mu(s))^2 - \sigma(s)^2}{\sigma(s)^3}\right] \tag{10}$$

using performance measure

$$\rho = r + \max_b Q(s', b) \tag{11}$$

and its baseline

$$\bar{\rho}(s) \leftarrow \bar{\rho}(s) + \alpha(\rho - \bar{\rho}(s)) \ . \tag{12}$$

The learning rate α_R has to be chosen appropriately, e.g. as a small positive constant, $\alpha_R = \alpha\sigma^2$, as proposed by Williams [19]. Furthermore, all REC parameters must be bounded, e.g. $0 \leq \varepsilon(s), \mu(s) \leq 1$ and $0.001 \leq \sigma(s) \leq 5$. The bounds for Softmax, MBE and VDBE-Softmax can be taken according to [10]. In this paper, the performance measure ρ slightly differs from [10], which yielded to improved results in the experiments (especially for Softmax).

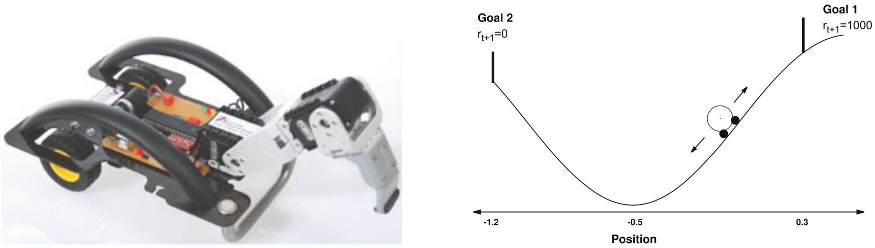
In contrast, the global variant of REC draws the exploration parameter at the beginning of an episode, for which reason the distribution parameters and baseline need only to be approximated for starting states [9,10]. When updating the distribution parameters at the end of episode i, the sum of rewards is taken as performance measure ρ_i, i.e.

$$\rho_i = \sum_{t=1}^{T} r_t \ . \tag{13}$$

In case a learning problem has only one starting state (such as in the investigated mountain–car problem with two goals), a stateless (global) approximation of μ, σ and $\bar{\rho}$ can be used.

4 Experiments

In this section, the two learning problems shown in Fig. 1 are investigated using $Q(\lambda)$-learning with meta-parameter learning of exploration parameters. First, a little crawling robot is investigated, which is a non-episodic learning problem on which the local variant of REC is applied. Second, the mountain–car problem with two goals is investigated, which is an episodic learning problem, and therefore the global variant of REC is applied.



(a) (b)

Fig. 1. Investigated learning problems: (a) the crawling robot [20], and (b) the mountain–car problem with two goals [9].

4.1 The Robot

We investigate a little crawling robot whose architecture was inspired from [21]. Figure 1(a) shows the corresponding hardware robot. The general aim is learning to crawl forward through sensorimotor interactions, which is achieved by a cyclic policy representing movements of the two joints g_x and g_y. The components of the corresponding MDP are defined as follows:

States: At each time step, the state $s \in \mathcal{S}$ consists of the arm's discrete joint positions at present, i. e. the state is fully described by $s = \{g_x, g_y\}$. Due to the small onboard memory of size 2 kB, each joint is discretized into 5 equidistant state positions resulting in total to 25 states as shown in [20].

Actions: The set of possible actions in each state $s \in \mathcal{S}$ consists of the four cardinal directions in the state-space model:

$$\mathcal{A}(s) \in \{\text{UP, DOWN, LEFT, RIGHT}\}\,.$$

Rewards: Performing an action a in the environment leads to a reward, $r \in \mathbb{R}$, which is measured as the number of accumulated wheel-encoder tics while repositioning the arm. The encoder tics trigger the external interrupt of the microcontroller. An interrupt service routine accumulates positive and negative encoder ticks for the reward signal r delivered to the learning algorithm.

For simplicity we do not model transition probabilities on the crawling robot, because actions (movements of the joints) always transition with probability 1 to the corresponding neighbor state.

We performed simulation experiments with a reward model sampled from the robot as shown in [20]. For simulating sensor noise, the reward from the model is perturbed with a Gaussian noise (mean 0 and variance 1). Figure 2 shows the results of our study, which are averages over 1000 runs, and using a discounting factor of $\gamma = 0.95$. In general it's observable that local REC adaptation behaves very robust independently of the policy and its parameter to be adapted. Using low learning rates of $\alpha = 0.1$, eligibility traces significantly improve the results the higher λ was chosen, which improves the standard Q-learning algorithm

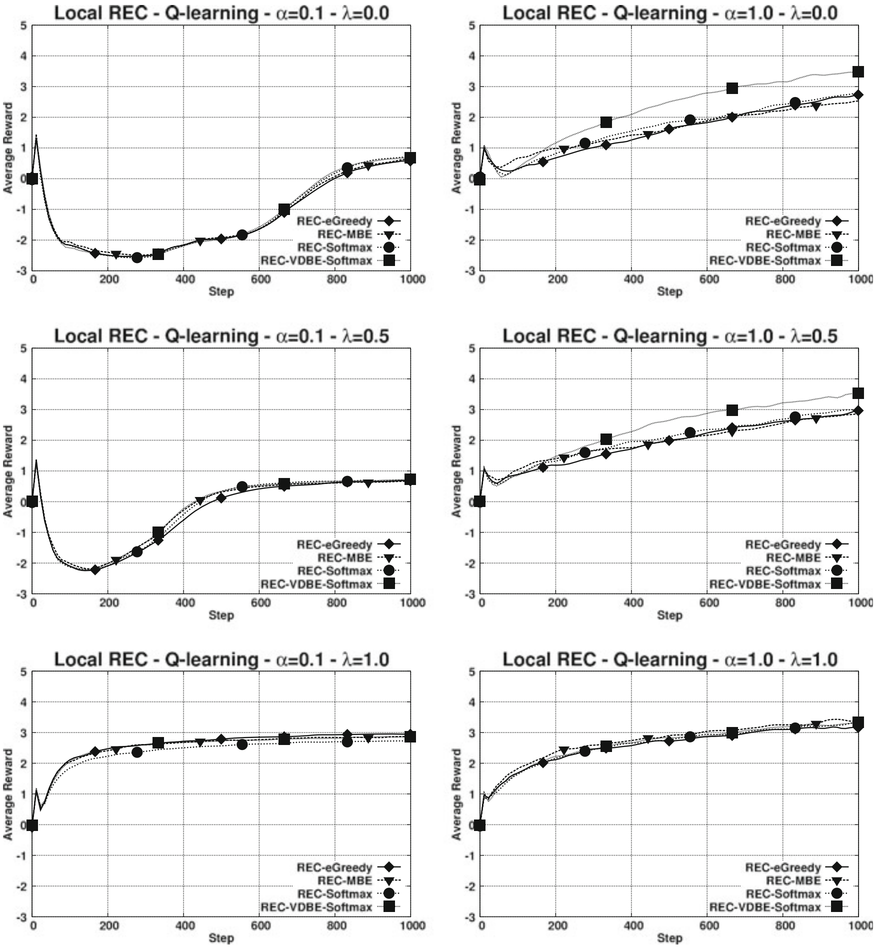


Fig. 2. The crawling robot: results for low learning rates of $\alpha = 0.1$ are on the left, high learning rates of $\alpha = 1.0$ are on the right. Results are smoothed and averaged over 1000 runs.

$(\lambda = 0)$ from Eq. (3). In contrast, high learning rates lead in general to better results, with little effect of using eligibility traces. From the results with high learning rates we further observe that the VDBE-Softmax policy is in general a bit better, but which is on cost of additionally memorizing $\varepsilon(s)$.

4.2 The Mountain–Car Problem with Two Goals

We investigate $Q(\lambda)$-learning on the *mountain–car problem with two goals* as proposed in [9,10] and depicted in Fig. 1b. This learning problem is an extension of the originally proposed mountain-car problem [11], but having two goal states and an enlarged action set.

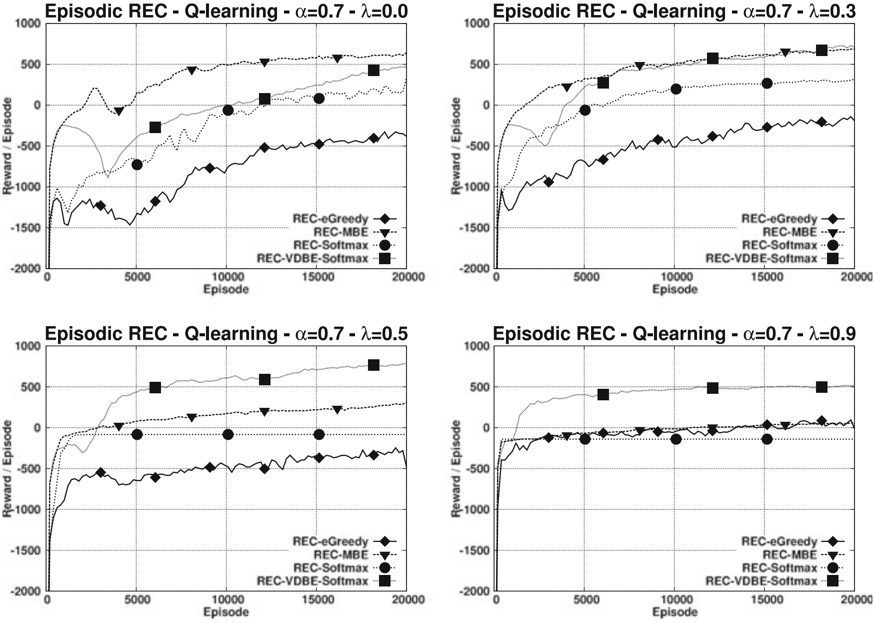


Fig. 3. The mountain–car problem with two goals: smoothed results for various λ; each averaged over 200 runs.

Our results shown in Fig. 3 reveal an interesting effect. The higher λ was chosen, the more policies tend to behave greedily, with the result of terminating an episode more likely at the left goal. For MBE eligibility traces seem to be contra productive. In contrast, the VDBE-Softmax results are improved in the range of $0.3 < \lambda < 0.5$, but also with degrading performance the more λ approaches 1.

5 Summary and Conclusion

We showed that replacing eligibility traces can improve the reward of an agent, which is consistent with the results shown by Singh and Sutton [11]. As shown on the robot example, learning improves the more λ approaches 1, but this is not the case for the mountain-car problem with two goals. Therefore it is definitely not straightforward to decide on the choice of time constant λ. The reason for this effect is that oversampling of actual performed actions can also lead to underestimating actions not selected so far (which might be better, than their current action-value predicts). As a result, it is more likely for $\lambda \rightarrow 1$ that a greedy behavior arises, as shown in the results of the mountain–car problem with two goals. Therefore, the optimal choice of λ is dependent on the learning rate α used for sampling the action-value function, which was also shown in the results of Singh and Sutton [11]. As a conclusion, meta-learning of λ and α in combination remains to be an interesting direction of further research.

Finally, it seems reasonable to initiate a discussion of the idea that reinforcement learning should also be considered as part of the relatively new research paradigm of *partially-supervised learning*, which by now had its emphasis on combinations of unsupervised and supervised learning. Our opinion is that the general framework of developing reinforcement learning agents fits well into this paradigm due to the following reasons: (1) reinforcement learning is utilized for sampling estimates about the future reward that are usually approximated by value functions [1], which (2) are most often learned using supervised learning algorithms, e.g. in a neural network fashion [22,23]. In previous research we successfully showed this relationship in the development of learning agents for board games [24,25]. Also in the context of neurobiology all three paradigms apparently interact with each other [26,27].

References

1. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)
2. Thrun, S.B.: Efficient exploration in reinforcement learning. Technical Report CMU-CS-92-102, Carnegie Mellon University, Pittsburgh (1992)
3. Auer, P.: Using confidence bounds for exploitation-exploration trade-offs. J. Mach. Learn. Res. **3**, 397–422 (2002)
4. Kocsis, L., Szepesvári, C.: Bandit based monte-carlo planning. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) ECML 2006. LNCS (LNAI), vol. 4212, pp. 282–293. Springer, Heidelberg (2006)
5. Schweighofer, N., Doya, K.: Meta-learning in reinforcement learning. Neural Netw. **16**(1), 5–9 (2003)
6. Kobayashi, K., Mizoue, H., Kuremoto, T., Obayashi, M.: A meta-learning method based on temporal difference error. In: Leung, C.S., Lee, M., Chan, J.H. (eds.) ICONIP 2009, Part I. LNCS, vol. 5863, pp. 530–537. Springer, Heidelberg (2009)
7. Tokic, M., Palm, G.: Value-difference based exploration: Adaptive control between epsilon-greedy and softmax. In: Bach, J., Edelkamp, S. (eds.) KI 2011. LNCS, vol. 7006, pp. 335–346. Springer, Heidelberg (2011)
8. Tokic, M., Ertle, P., Palm, G., Söffker, D., Voos, H.: Robust exploration/exploitation trade-offs in safety-critical applications. In: Proceedings of the 8th International Symposium on Fault Detection, Supervision and Safety of Technical Processes, Mexico City, Mexico, IFAC, pp. 660–665 (2012)
9. Tokic, M., Palm, G.: Adaptive exploration using stochastic neurons. In: Villa, A.E., Duch, W., Érdi, P., Palm, G. (eds.) ICANN 2012, Part II. LNCS, vol. 7553, pp. 42–49. Springer, Heidelberg (2012)
10. Tokic, M., Palm, G.: Gradient algorithms for exploration/exploitation trade-offs: Global and local variants. In: Mana, N., Schwenker, F., Trentin, E. (eds.) ANNPR 2012. LNCS, vol. 7477, pp. 60–71. Springer, Heidelberg (2012)
11. Singh, S., Sutton, R.S.: Reinforcement learning with replacing eligibility traces. Mach. Learn. **22**, 123–158 (1996)
12. Niv, Y., Daw, N.D., Dayan, P.: Choice values. Nat. Neurosci. **9**(8), 987–988 (2006)
13. Niv, Y.: Reinforcement learning in the brain. J. Math. Psychol. **53**(3), 139–154 (2009)
14. Watkins, C.: Learning from delayed rewards. Ph.D. thesis, University of Cambridge, England (1989)

15. George, A.P., Powell, W.B.: Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming. Mach. Learn. **65**(1), 167–198 (2006)
16. Vermorel, J., Mohri, M.: Multi-armed bandit algorithms and empirical evaluation. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) ECML 2005. LNCS (LNAI), vol. 3720, pp. 437–448. Springer, Heidelberg (2005)
17. Wiering, M.: Explorations in efficient reinforcement learning. PhD thesis, University of Amsterdam, Amsterdam (1999)
18. Tokic, M.: Adaptive ϵ-greedy exploration in reinforcement learning based on value differences. In: Dillmann, R., Beyerer, J., Hanebeck, U.D., Schultz, T. (eds.) KI 2010. LNCS, vol. 6359, pp. 203–210. Springer, Heidelberg (2010)
19. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. **8**, 229–256 (1992)
20. Tokic, M., Bou Ammar, H.: Teaching reinforcement learning using a physical robot. In: Proceedings of the Workshop on Teaching Machine Learning at the 29th International Conference on Machine Learning, Edinburgh, UK, pp. 1–4 (2012)
21. Kimura, H., Miyazaki, K., Kobayashi, S.: Reinforcement learning in POMDPs with function approximation. In: Proceedings of the 14th International Conference on Machine Learning, San Francisco, CA, USA, pp. 152–160. Morgan Kaufmann Publishers Inc. (1997)
22. Riedmiller, M.: Neural fitted Q iteration - first experiences with a data efficient neural reinforcement learning method. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) ECML 2005. LNCS (LNAI), vol. 3720, pp. 317–328. Springer, Heidelberg (2005)
23. Riedmiller, M., Montemerlo, M., Dahlkamp, H.: Learning to drive a real car in 20 minutes. In: Proceedings of the FBIT 2007 Conference, Jeju, Korea. Special Track on, autonomous robots (2007)
24. Faußer, S., Schwenker, F.: Learning a strategy with neural approximated temporal-difference methods in english draughts. In: Proceedings of the 20th International Conference on Pattern Recognition, pp. 2925–2928. IEEE Computer Society (2010)
25. Faußer, S., Schwenker, F.: Neural approximation of monte carlo policy evaluation deployed in connect four. In: Prevost, L., Marinai, S., Schwenker, F. (eds.) ANNPR 2008. LNCS (LNAI), vol. 5064, pp. 90–100. Springer, Heidelberg (2008)
26. Doya, K.: What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Netw. **12**(7–8), 961–974 (1999)
27. Bostan, A.C., Dum, R.P., Strick, P.L.: The basal ganglia communicate with the cerebellum. Proc. Nat. Acad. Sci. **107**(18), 8452–8456 (2010)

Neighborhood Co-regularized Multi-view Spectral Clustering of Microbiome Data

Evgeni Tsivtsivadze[1(✉)], Hanneke Borgdorff[2], Janneke van de Wijgert[3],
Frank Schuren[1], Rita Verhelst[4], and Tom Heskes[5]

[1] MSB Group, The Netherlands Organization for Applied Scientific Research,
Zeist, The Netherlands
{e.tsivtsivadze, f.schuren}@tno.nl

[2] Academic Medical Center of the University of Amsterdam and
Amsterdam Institute for Global Health and Development,
Amsterdam, The Netherlands
h.borgdorff@aighd.org

[3] Institute of Infection and Global Health, University of Liverpool,
Liverpool, UK
j.vandewijgert@liverpool.ac.uk

[4] International Centre of Reproductive Health,
Faculty of Medicine and Health Sciences, Ghent University,
Ghent, Belgium
Rita.Verhelst@UGent.be

[5] Institute for Computing and Information Sciences, Radboud University,
89069 Nijmegen, The Netherlands
t.heskes@science.ru.nl

Abstract. In many unsupervised learning problems data can be available in different representations, often referred to as views. By leveraging information from multiple views we can obtain clustering that is more robust and accurate compared to the one obtained via the individual views. We propose a novel algorithm that is based on neighborhood co-regularization of the clustering hypotheses and that searches for the solution which is consistent across different views. In our empirical evaluation on publicly available datasets, the proposed method outperforms several state-of-the-art clustering algorithms. Furthermore, application of our method to recently collected biomedical data leads to new insights, critical for future research on determinants of the cervicovaginal microbiome and the cervicovaginal microbiome as a risk factor for the transmission of HIV. These insights could have an influence on the interpretation of clinical presentation of women with bacterial vaginosis and treatment decisions.

1 Introduction

The multi-view paradigm [1–3] is particularly suitable for learning on datasets having more than a single data representation. A classic example is a web document classification task [1], where documents are represented via two different

E. Tsivtsivadze and H. Borgdorff contributed equally to this work.

Z.-H. Zhou and F. Schwenker (Eds.): PSL 2013, LNAI 8183, pp. 80–90, 2013.
DOI: 10.1007/978-3-642-40705-5_8,
© Springer-Verlag Berlin Heidelberg 2013

views - one that is based on the links and another one based on the text document. Complex, structured data with multiple representations are frequently encountered in the biomedical domain, making multi-view methods a natural application choice. Although in many circumstances the individual data representation can be sufficient for training a model, a combination of the multiple views can lead to more robust and accurate predictions compared to the ones obtained via the individual views.

The multi-view paradigm has been successfully applied to various learning problems such as semi-supervised classification e.g. [4], regression e.g. [5], preference learning e.g. [6] and clustering e.g. [3,7]. Our work concerns an unsupervised multi-view learning algorithm that builds upon the spectral clustering method [8,9]. The proposed algorithm is conceptually different from the existing methods as it uses novel neighborhood co-regularization technique to adapt cluster assignments for different views. Unlike in previous studies that consider aggregation of clusters based on the individual data representations e.g. [10–13], our method promotes consistent cluster assignments across multiple views and penalizes solutions that differ significantly. The closest in spirit to our method is the recent work described in [7], where for the first time a co-regularization framework was successfully applied to a clustering task. However, our co-regularization approach is fundamentally different and is geared towards solutions that capture local/neighborhood-based relations in the dataset. Furthermore, the optimization problem in our work differs from [7] and leads to a simpler closed form solution with fewer terms involved.

We apply the proposed method to a recently collected biomedical dataset from a study aimed at investigating cervicovaginal microbiome compositions. As the determination of microbial community compositions is becoming increasingly complex due to new molecular laboratory methods, unsupervised learning techniques have become an essential part of microbiome studies, e.g.[14,15]. We demonstrate that, unlike in previous studies, the proposed neighborhood co-regularized multi-view spectral clustering algorithm (NCMSC) identifies distinct clusters within the group of women with a “healthy” cervicovaginal microbiome and within the group of women with bacterial vaginosis (BV). Our observation will aid the analysis of the determinants of the cervicovaginal microbiome and the cervicovaginal microbiome as risk factor for other adverse outcomes, such as transmission of sexually transmitted infections (STIs) and HIV.

2 Neighborhood Co-regularized Multi-view Spectral Clustering

Consider we are given a dataset containing multiple representations. Let $X^{(v)} = \{x_i^{(v)}\}_{i=1}^n$. Note that here superscript v denotes the representation for a single view. Let $A^{(v)}$ denote an adjacency matrix of the graph constructed using the data representation in a view v. We can write the normalized Laplacian matrix as $L^{(v)} = D^{(v)-1/2}A^{(v)}D^{(v)-1/2}$, where $D^{(v)}$ is the corresponding degree matrix.

Following [8] the standard special clustering problem (or single view spectral clustering [7]) solves the optimization problem

$$\min_{Q^{(v)} \in \mathcal{R}^{n\times c}} tr\left(Q^{(v)T} L^{(v)} Q^{(v)}\right), \quad \text{s.t.} \quad Q^{(v)T} Q^{(v)} = I \tag{1}$$

where $Q^{(v)} \in \mathcal{R}^{n\times c}$ denotes the cluster assignment matrix and c is number of predefined clusters. In spectral clustering the final cluster membership is obtained by applying the k-means algorithm on the rows of the matrix $Q^{(v)}$. The algorithm we propose extends the standard spectral clustering framework by using neighborhood co-regularization techniques that naturally allow to leverage information from multiple views. Let us denote the cluster assignment matrix $Q^{(v)} = (\mathbf{q}_1^{(v)}, \ldots, \mathbf{q}_c^{(v)})^T$. Slightly overloading our notation, we denote the confidence that a data point $\mathbf{x}_j$ belongs to the cluster c as $[\mathbf{q}_c]_j$. For simplicity, in the derivations below we omit the cluster index.

We define the k neighbors of data point $\mathbf{x}_i^{(v)}$ as $N(\mathbf{x}_i^{(v)}) = \{\mathbf{x}_{i_1}^{(v)}, \ldots, \mathbf{x}_{i_k}^{(v)}\}$ or $X_i^{(v)} = (\mathbf{x}_{i_1}^{(v)}, \ldots, \mathbf{x}_{i_k}^{(v)})^T \in \mathcal{R}^{k\times d}$ where d is the dimensionality of the data point in a view v. Also, the corresponding cluster assignments can be written as $\mathbf{q}_i^{(v)} = (q_{i_1}^{(v)}, \ldots, q_{i_k}^{(v)})^T \in \mathcal{R}^k$. Below we generalize local linear regularization [16,17] to a multi-view setting. In our setting, for each data point $\mathbf{x}_i^{(v)}$, projections $W_i = (\mathbf{w}_i^{(1)}, \ldots, \mathbf{w}_i^{(M)})^T \in \mathcal{R}^{M\times d}$ are leaned via$\min\limits_{\mathbf{w}_i^{(v)}} J(W_i)$, where

$$\begin{aligned} J(W_i) = &\sum_{v=1}^{M} \sum_{\mathbf{x}_j^{(v)} \in N(\boldsymbol{x}_i^{(v)})} \|[\mathbf{q}^{(v)}]_j - \mathbf{x}_j^{(v)T} \mathbf{w}_i^{(v)}\|_F^2 + \lambda \sum_{v=1}^{M} \mathbf{w}_i^{(v)T} \mathbf{w}_i^{(v)} \\ &+ \nu \sum_{\substack{v,u=1 \\ v\neq u}}^{M} \sum_{\mathbf{x}_j^{(v)} \in N(\boldsymbol{x}_i^{(v)})} \|\mathbf{x}_j^{(v)T} \mathbf{w}_i^{(v)} - \mathbf{x}_j^{(u)T} \mathbf{w}_i^{(u)}\|_F^2. \end{aligned} \tag{2}$$

The first term in Eq. (2) stands for a multi-view version of the problem where we aim to find a weight vector that corresponds as closely as possible to the optimal clustering solution, the second term is the $L2$ regularization on the weight vector $\mathbf{w}_i^{(v)}$, and the third term is a co-regularization that promotes agreement among different views on the obtained clustering. Once the local predictors for all views have been constructed (see Appendix) we can compute the sum of the prediction errors for all clusters

$$\begin{aligned} J_c &= \sum_{v=1}^{M} \sum_{l=1}^{c} \|H^{(v)} \mathbf{q}_l^{(v)} - \mathbf{q}_l^{(v)}\|^2 \\ &= \sum_{v=1}^{M} \sum_{l=1}^{c} [\mathbf{q}_l^{(v)T} ((H^{(v)} - I)^T (H^{(v)} - I)) \mathbf{q}_l^{(v)}] \\ &= tr[\mathbf{Q}^T ((\mathbf{H} - \mathbf{I})^T (\mathbf{H} - \mathbf{I})) \mathbf{Q}], \end{aligned}$$

where $\mathbf{Q}$ is a $(Mn \times c)$ matrix containing the cluster assignments for all views and $\mathbf{H}$ is a $(Mn \times Mn)$ matrix containing predictions of the linear classifiers estimated via minimization of (2). Thus, the optimization problem we solve to determine cluster assignment matrices for all views is

$$\min_{\mathbf{Q} \in \mathcal{R}^{M} n \times c} tr[\mathbf{Q}^T((\mathbf{H}-\mathbf{I})^T(\mathbf{H}-\mathbf{I}))\mathbf{Q}] \quad \text{s.t.} \quad \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \tag{3}$$

The above problem is closely related to the standard spectral clustering and the solutions for all views are given by top-c eigenvectors of the matrix $\mathbf{L} = (\mathbf{H}-\mathbf{I})^T(\mathbf{H}-\mathbf{I})$. Similarly to [7] we can use any of the obtained cluster assignment matrices $Q^{(v)}$ in the final k-means step of the clustering algorithm. In our experiments, we observe no major dependence of the clustering performance on the choice of a particular $Q^{(v)}$.

2.1 Related Work

There have been a number of multi-view clustering algorithms proposed that build on the idea of leveraging information from different graphs/data representations. For example, in [10] the authors obtain a graph cut which on average is the most suitable for multiple graphs and provide a random walk formulation of the clustering problem. A clustering algorithm that constructs a graph based on nodes from two views and solves a standard spectral problem, is proposed in [11]. Other approaches for multi-view clustering fuse the information from multiple graphs based on matrix factorization [12] or consensus learning techniques [13].

The central idea behind all these methods is to construct clustering for the individual views and to reconcile them in the final solution. Our neighborhood co-regularized multi-view clustering algorithm is conceptually different from these methods as the cluster assignments for the individual views are adapted based on neighborhood co-regularization implemented via the third term in the equation (2). Informally, our method promotes consistent cluster assignments across multiple views and penalizes solutions that differ significantly.

Recently an unsupervised learning algorithm using a co-regularization approach has been proposed in [7]. The co-regularization we propose here is notably different and is geared towards clustering that captures local/ neighborhood-based relations in the dataset. This in turn leads to a simpler closed form solution with fewer terms involved. Furthermore, our algorithm can be straightforwardly formulated as a kernel-based method, which can make it suitable for learning non-linear dependencies when estimating cluster assignments.

We also note that the co-regularization framework has recently attracted considerable attention in the machine learning community and proved to work well on a wide range of learning problems e.g. [5,6,18]. Moreover, theoretical investigations demonstrate that the co-regularization approach reduces the Rademacher complexity by an amount that depends on the "distance" between the views [19,20]. We expect that a similar type of analysis can be applied to our algorithm and we aim to investigate this in the near future.

3 Experiments on Benchmark Datasets

To empirically validate the performance of the NCMSC algorithm, we compare the results of five algorithms on four benchmark datasets. Following [17] we select publicly available datasets, namely Newsgroup, UMIST, WebACE, and USPS.

To evaluate the performance of the algorithms, we compare the obtained clusters with the true class labels in each of the datasets. For this purpose we use two performance measures - clustering accuracy (ACC) and normalized mutual information (NMI) [13]. The clustering accuracy performance measure estimates the relationship between computed clusters and the class labels. Informally, it measures the extent to which data points contained in the clusters correspond to the class label and sums up matches between all class-cluster pairs. The normalized mutual information criterion reports mutual information between the obtained clustering and the true clustering, normalized by the cluster entropies. NMI ranges between 0 and 1 with a higher value indicating a closer match to the true clustering.

In our experiments we compare the performance of the proposed NCMSC algorithm with four other clustering methods, namely k-means (K-Means), hierarchical clustering (HC), spectral clustering (SC), and co-regularized multi-view spectral clustering (CMSC) [7]. For HC we use Euclidean distance. For NCMSC, CMSC and SC, the weights on data graph edges are computed by Gaussian functions. Similarly to [17] the variance is determined by local scaling. All regularization parameters in CMSC and our approach are determined by searching the grid {0.1,1,10}, and the neighborhood size is set by searching the grid {20, 40, 80}. In the experiments we consider two views for the NCMSC and CMSC algorithms, which are created by partitioning of the feature vector into two parts (random partitioning has been successfully used in previous studies e.g. [5,6]).

It can be observed from the Tables 1 and 2 that the NCMSC algorithm performs better compared to other clustering methods. We suggest that our algorithm outperforms CMSC due to the employed co-regularization procedure, which uses neighborhood-based cluster assignment models and, therefore, captures additional "local" relations in the data. Also, multi-view algorithms in

Table 1. Clustering accuracy results

Algorithm	HC	K-means	SC	CMSC	NCMSC
UMIST	0.4127	0.4372	0.6432	0.6824	**0.6914**
USPS	0.7354	0.7399	0.9312	0.9561	**0.9732**
Newsgroup	0.3223	0.3234	0.5211	0.5734	**0.5811**
WebAce	0.3123	0.3131	0.4553	0.5625	**0.5893**

Table 2. Normalized mutual information results

Algorithm	HC	K-means	SC	CMSC	NCMSC
UMIST	0.6441	0.6481	0.7623	0.8121	**0.8236**
USPS	0.8231	0.8521	0.9732	0.9832	**0.9917**
Newsgroup	0.2114	0.2212	0.4962	0.5213	**0.5283**
WebAce	0.1323	0.1431	0.3842	0.5125	**0.5298**

general tend to perform better than their single view counterparts and k-means or hierarchical clustering tend to perform poorer than SC-based approaches.

4 Experiments on a Cervicovaginal Microbiome Dataset

We also apply the proposed NCMSC algorithm to a new biomedical dataset containing results from experiments on a single channel phylogenetic microarray designed to characterize cervicovaginal microbiota [21]. The dataset is from an ongoing study aimed at identifying groups of women with similar cervicovaginal microbial compositions, the determinants of these compositions and their possible association with adverse reproductive health outcomes.

4.1 Dataset Preparation

Data from microarray experiments are available for 196 cervical samples from women who participated in an observational prospective cohort study aimed at estimating the HIV-1 incidence in Kigali, Rwanda [22]. BV was diagnosed using Gram stain and microscopy using Nugent scoring [23], which is considered the golden standard. BV is a disruption of the cervicovaginal microbiome characterized by a reduction of lactobacilli and an increase of mostly anaerobic bacteria that leads to an increased risk of preterm birth and sexually transmitted infections [24–26]. A Nugent score of 0-3 was considered as BV negative, 4-6 as intermediate microbiota and 7-10 as BV positive.

We have followed standard microarray preprocessing strategies as described, for example, in [27]. For each spot, signal over background ratios (S/B) are calculated. If the signal is not confidently above background [27], the S/B ratio is set at 1. Samples for which the positive controls (general bacterial probes) show a low S/B ratio, are discarded from the analysis. Lowess smoothing was performed for plate normalization.

The NCMSC algorithm is applied to the preprocessed dataset. The parameters and views are obtained as described in Sect. 3. We run NCMSC with predefined number of clusters ranging from 3 to 10. Next, we compute the co-occurrence matrix [13] (see Fig. 1) that reveals the true number of clusters in the data. For example, former research on the vaginal microbiome of asymptomatic American women described five clusters [15].

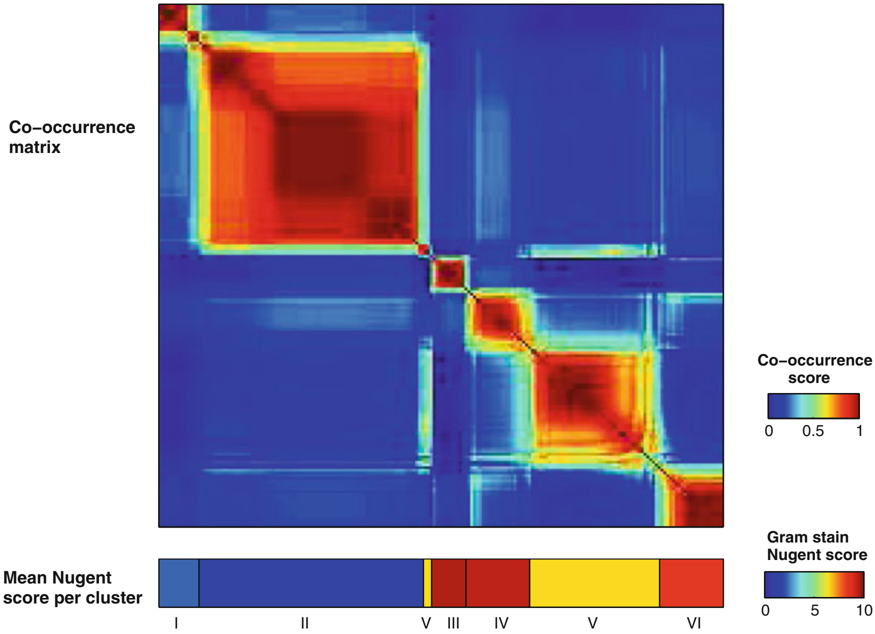


Fig. 1. Co-occurrence matrix based on the clusters obtained by the NCMSC algorithm. Patients that often co-occur in different solutions receive a high co-occurrence score.

I-VI: Number of the cervicovaginal microbiome community cluster. Average Nugent scoring per cluster based on gram stained microscopy denoted in colored bar below the co-occurrence matrix. A Nugent score of 0-3 is considered negative for bacterial vaginosis (BV), 4-6 as an intermediate state and 7-10 as BV-positive.

4.2 Results

Based on the co-occurrence matrix six separate clusters are identified (see Fig. 1). As would be expected, lactobacilli are present in most samples. The first cluster is dominated by *Lactobacillus crispatus* (all 3 probes targeted at *L. crispatus/ L. kefiranofaciens* give a positive signal in 8/11 samples), the second and largest by *Lactobacillus iners* (75-78/83 samples have a positive signal for the three probes targeted at *L. iners*) and the third to sixth cluster by bacteria which are known to be associated with bacterial vaginosis, such as *Gardnerella vaginalis*, *Atopobium vaginae*, *Mobiluncus mulieris*, *Prevotella* spp., *Dialister* spp., *Sneathia* spp. and *Megasphaera* spp. Furthermore, comparing the obtained clustering to the diagnosis of BV by gram stain Nugent scoring shows that this diverse group is indeed associated with BV, while cluster I and II are associated with the absence of BV. Fig. 1 shows the average Nugent score per cluster. Interestingly, the BV-associated group also divides into several clusters.

Our analysis confirms the results of two recent studies, that is [14] in which 93% of women without BV had a vaginal microbial community dominated by either *Lactobacillus crispatus* or *Lactobacillus iners* and [15] where clusters with low Nugent scores were dominated by respectively *Lactobacillus crispatus*, *Lactobacillus iners*, *Lactobacillus gasseri* and *Lactobacillus jensenii*. Unlike in these studies, our approach clearly distinguishes between clusters within the lactobacilli-dominated group and within the BV-associated group.

Contrary to the benchmark datasets, there is no standardized way of comparing clustering methods on this biomedical dataset. Although most of the methods were able to separate the BV-negative from the BV-positive group, only NCMSC is able to clearly identify and differentiate between six cluster groups. Other methods do not separate the *Lactobacillus crispatus* from the *Lactobacillus iners* cluster, even though scientific evidence increasingly shows that these should be separate clusters [14,15].

5 Conclusion

As the determination of microbial community compositions is becoming increasingly complex due to new molecular laboratory techniques, unsupervised learning methods have become an essential part of microbiome studies. In this paper we propose a novel algorithm for the analysis of complex microbiome data. Our work extends the spectral clustering method [8,9] to a multi-view setting. We propose neighborhood co-regularization approach to promote consistent cluster assignments across multiple views and to penalize the solutions that differ significantly. Our approach is fundamentally different from existing multi-view methods and is geared towards solutions that capture local/neighborhood-based relations in the dataset.

We have evaluated the performance of the proposed algorithm on several publicly available datasets and applied our method to a recently collected microarray dataset. On all these datasets the NCMSC algorithm outperformed other clustering methods.

When applied to the microbiome data, besides confirming previous studies, the NCMSC algorithm identifies clusters within the lactobacilli-dominated group and within the BV-associated group. This observation will help to identify determinants of the cervicovaginal microbiome and cervicovaginal microbiome compositions associated with other adverse outcomes, such as transmission of STIs and HIV. BV is a difficult to treat condition and the separation of BV-positive women into clusters with different microbial compositions can potentially be important for clinical presentation and treatment decisions.

Acknowledgement. We acknowledge support from the Netherlands Organization for Scientific Research (grant number 639.023.604). Funding for the cervicovaginal microbiome study was received from the European and Developing Countries Clinical Trials Partnership (EDCTP), European Commission 7th Framework CHAARM project and the Aids Fonds Netherlands (grant number 201102). The views expressed in this paper are those of the authors and do not necessarily represent the official position of EDCTP, the EU, or the Aids Fonds Netherlands.

Appendix

Given the matrix formulation of our optimization problem, we can find the following closed form for the solution. Taking the partial derivative of $J(W_i)$ with respect to $\mathbf{w}_i^{(v)}$ we get

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}_i^{(v)}} J(W_i) = &- 2X_i^{(v)}(\mathbf{q}_i^{(v)} - X_i^{(v)T}\mathbf{w}_i^{(v)}) + 2\lambda \mathbf{w}_i^{(v)} \\ &- 4\nu \sum_{u,v=1,u\neq v}^{M} X_i^{(v)}(X_i^{(v)T}\mathbf{w}_i^{(v)} - X_i^{(u)T}\mathbf{w}_i^{(u)}).\end{aligned}$$

By defining $G^{\nu} = 2\nu(M-1)X_i^{(v)}X_i^{(v)T}$, $G^{\lambda} = \lambda X_i^{(v)T}$ and $G = X_i^{(v)}X_i^{(v)T}$, we can rewrite the above term as

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}_i^{(v)}} J(W_i) = &2(G + G^{\nu} + G^{\lambda})\mathbf{w}_i^{(v)} - 2X_i^{(v)T}\mathbf{q}_i^{(v)} \\ &- 4\nu \sum_{u,v=1,u\neq v}^{M} X_i^{(v)}X_i^{(u)T}\mathbf{q}_i^{(u)}.\end{aligned}$$

At the optimum we have $\frac{\partial}{\partial \mathbf{w}_i^{(v)}} J(W_i) = 0$ for all views, thus we get the exact solution by solving

$$\begin{pmatrix} G_1 & -2\nu X_i^{(1)}X_i^{(2)T} & \dots \\ -2\nu X_i^{(2)}X_i^{(1)T} & G_2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \mathbf{q}_i^{(1)} \\ \mathbf{q}_i^{(2)} \\ \vdots \end{pmatrix} = \begin{pmatrix} X_i^{(1)T}\mathbf{q}_i^{(1)} \\ X_i^{(2)T}\mathbf{q}_i^{(2)} \\ \vdots \end{pmatrix}$$

with respect to $\mathbf{w}_i^{(1)}, \dots, \mathbf{w}_i^{(M)}$. Note that the left-hand side matrix is positive definite and therefore invertible.

References

1. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pp. 92–100. ACM, New York (1998)
2. Sindhwani, V., Niyogi, P., Belkin, M.: A co-regularization approach to semi-supervised learning with multiple views. In: Proceedings of ICML Workshop on Learning with Multiple Views (2005)
3. Chaudhuri, K., Kakade, S.M., Livescu, K., Sridharan, K.: Multi-view clustering via canonical correlation analysis. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 129–136. ACM (2009)

4. Krishnapuram, B., Williams, D., Xue, Y., Hartemink, A.J., Carin, L., Figueiredo, M.A.T.: On semi-supervised classification. In: Advances Neural Information Processing Systems, vol. 17 (2004)
5. Brefeld, U., Gärtner, T., Scheffer, T., Wrobel, S.: Efficient co-regularised least squares regression. In: Proceedings of the International Conference on Machine learning, pp. 137–144. ACM, New York (2006)
6. Tsivtsivadze, E., Pahikkala, T., Boberg, J., Salakoski, T., Heskes, T.: Co-regularized least-squares for label ranking. In: Hüllermeier, E., Fürnkranz, J. (eds.) Preference, Learning, pp. 107–123 (2010)
7. Kumar, A., Rai, P., Daume III, H.: Co-regularized multi-view spectral clustering. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K. (eds.) Advances in Neural Information Processing Systems, vol. 24, pp. 1413–1421 (2011)
8. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: Advances in Neural Information Processing Systems, vol. 14, pp. 849–856 (2001)
9. Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**(4), 395–416 (2007)
10. Zhou, D., Burges, C.J.C.: Spectral clustering and transductive learning with multiple views. In: Proceedings of the 24th International Conference on Machine Learning, pp. 1159–1166 (2007)
11. de Sa, V.R.: Spectral clustering with two views. In: Workshop on Learning with Multiple Views, International Conference on Machine Learning (2005)
12. Tang, W., Lu, Z., Dhillon, I.S.: Clustering with multiple graphs. In: Proceedings of the 2009 Nineth IEEE International Conference on Data Mining, pp. 1016–1021 (2009)
13. Strehl, A., Ghosh, J.: Cluster ensembles – a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2003)
14. Srinivasan, S., Hoffman, N.G., Morgan, M.T., Matsen, F.A., Fiedler, T.L., Hall, R.W., Ross, F.J., McCoy, C.O., Bumgarner, R., Marrazzo, J.M., Fredricks, D.N.: Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. PLoS ONE **7**(6), e37818 (2012)
15. Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S., McCulle, S.L., Karlebach, S., Gorle, R., Russell, J., Tacket, C.O., Brotman, R.M., Davis, C.C., Ault, K., Peralta, L., Forney, L.J.: Vaginal microbiome of reproductive-age women. PNAS **108**(Suppl. 1), 4680–4687 (2011)
16. Wu, M., Schölkopf, B.: A local learning approach for clustering. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems, vol. 19, pp. 1529–1536. MIT Press, Cambridge (2007)
17. Wang, F., Zhang, C., Li, T.: Clustering with local and global regularization. In: Proceedings of the 22nd National Conference on Artificial Intelligence, pp. 657–662. AAAI Press (2007)
18. Sindhwani, V., Niyogi, P.: A co-regularized approach to semi-supervised learning with multiple views. In: Proceedings of the ICML Workshop on Learning with Multiple Views (2005)
19. Rosenberg, D., Bartlett, P.L.: The Rademacher complexity of co-regularized kernel classes. In: Meila, M., Shen, X., (eds.) Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, pp. 396–403 (2007)
20. Sindhwani, V., Rosenberg, D.: An RKHS for multi-view learning and manifold co-regularization. In: McCallum, A., Roweis, S. (eds.) Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008), Finland, pp. 976–983. Omnipress, Helsinki (2008)

21. Dols, J.A., Smit, P.W., Kort, R., Reid, G., Schuren, F.H., Tempelman, H., Bontekoe, T.R., Korporaal, H., Boon, M.E.: Microarray-based identification of clinically relevant vaginal bacteria in relation to bacterial vaginosis. Am. J. Obstet. Gynecol. **204**(4), 1–7 (2011)
22. Braunstein, S.L., Ingabire, C.M., Kestelyn, E., Uwizera, A.U., Mwamarangwe, L., Ntirushwa, J., Nash, D., Veldhuijzen, N.J., Nel, A., Vyankandondera, J., van de Wijgert, J.H.: High human immunodeficiency virus incidence in a cohort of Rwandan female sex workers. Sex. Transm. Dis. **38**(5), 385–394 (2011)
23. Nugent, R.P., Krohn, M.A., Hillier, S.L.: Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. J. Clin. Microbiol. **29**(2), 297–301 (1991)
24. Hauth, J.C., Macpherson, C., Carey, J.C., Klebanoff, M.A., Hillier, S.L., Ernest, J.M., Leveno, K.J., Wapner, R., Varner, M., Trout, W., Moawad, A., Sibai, B.: Early pregnancy threshold vaginal pH and Gram stain scores predictive of subsequent preterm birth in asymptomatic women. Am. J. Obstet. Gynecol. **188**(3), 831–835 (2003)
25. Cohen, C.R., Lingappa, J.R., Baeten, J.M., Ngayo, M.O., Spiegel, C.A., Hong, T., Donnell, D., Celum, C., Kapiga, S., Bukusi, E.A.: Bacterial vaginosis associated with increased risk of female-to-male HIV-1 transmission: a prospective cohort analysis among African couples. PLoS Med. **9**(6), e1001251 (2012)
26. Wiesenfeld, H.C., Hillier, S.L., Krohn, M.A., Landers, D.V., Sweet, R.L.: Bacterial vaginosis is a strong predictor of Neisseria gonorrhoeae and Chlamydia trachomatis infection. Clin. Infect. Dis. **36**(5), 663–668 (2003)
27. Quackenbush, J.: Microarray data normalization and transformation. Nat. Genet. **32**(Suppl.), 496–501 (2002)

A Robust Image Watermarking Scheme Based on BWT and ICA

Tao Wang[1,2(✉)], Jin Tang[1,2], Bin Luo[1,2], and Cheng Zhang[1]

[1]School of Computer Science and Technology, Anhui University,
Hefei 230601, China
{qingyu_53@126.com, ahhftang@gmail.com,
luobin@ahu.edu.cn, 109643008@qq.com}
[2]Key Lab of Industrial Image Processing and Analysis of Anhui Province,
Hefei 230039, China

Abstract. A robust image watermarking scheme combined with the human visual characteristics is proposed. Berkeley wavelet transform (BWT) which is used in watermarking embedding procedure simulates physiology characteristics of the mammalian primary visual cortex (V_1). Independent Component Analysis (ICA) which is blind separation technology will be adapted to the watermarking extracting procedure. By combining the advantages of BWT and ICA, a robust image watermarking scheme is proposed and a simulation of the scheme is designed. Experimental results demonstrate that the proposed watermarking technique combines the imperceptibility, robustness, real-time and high capacity of digital watermarking algorithms.

Keywords: Watermarking · Berkeley wavelet transform · ICA

1 Introduction

Currently, there are many achievements on digital watermarking technology [1, 2]. A practical robust digital watermarking algorithm should have the following positive characteristics: good visual transparency, robustness against some general signal operations, large capacities, low time and space complexity, etc.

Digital watermarking algorithms can be roughly divided into two categories: space domain watermarking algorithm [3] and transform domain watermarking algorithm [4]. The most representative watermarking algorithm in space domain is LSB algorithm [5]. This algorithm runs fast and can hide a great deal of information. However, it cannot accommodate to the geometric deformation and signal operation. So its robustness is not good and has less efficient performance in practice. The most representative watermarking algorithm in transform domain is DCT algorithm [6]. The DCT algorithm firstly carries out DCT transform on the original image in order to embed the watermarking image in DCT coefficients. Although the process is simple, the number of DCT transform frequency coefficient is very small. Thus the embedding capacity is limited.

Z.-H. Zhou and F. Schwenker (Eds.): PSL 2013, LNAI 8183, pp. 91–103, 2013.
DOI: 10.1007/978-3-642-40705-5_9,
© Springer-Verlag Berlin Heidelberg 2013

In recent years, there have been many digital watermarking algorithms based on wavelet transform [7, 8]. The algorithm decomposes the original image using wavelet, and embeds the watermarking in low-frequency. So in order to increase the watermarking capacity and optimize the watermarking embedding location in the transform domain, many algorithms based on mathematical analysis or machine learning methods appear. Lou [9] uses neural network and combines with human visual system model [10] to train the brightness, frequency, texture and average information entropy of different regions of images, in order to determine the embedding strength and capacity of different regions. Davis [11] uses the wavelet transform to embed the watermarking, and then adaptively generate the maximum embedding strength of the image content by using neural network. But neural network is easy to produce over-learning phenomena, the hidden nodes are determined by experience, the parameter setting is difficult, and the pre-processing is complex. Some other watermarking algorithms are based on feature point [12, 13]. For example, the research papers [12, 13] apply the multi-scale Harris to detect the stable feature point from the original image and generate the practical feature regions, and then embed the watermarking. But Harris corner are sensitive to noise points.

Overall, even though there are many technologies on digital watermarking, but few of them have efficiency in practice. However, Berkeley wavelet transform [14] shares many characteristics (spatial localization, oriented and frequency bandpass) with the receptive fields of neurons in mammalian primary visual cortex (V_1). And BWT is constructed by four complete orthogonal bases, the calculation of transformation and inverse transformation are fast. So, BWT has good primary visual simulation and well computability. Therefore, we propose one new digital watermarking algorithm based on BWT and ICA [15, 16], in order to meet the imperceptibility, robustness, real-time and capacity requirements of digital watermarking algorithms in the maximum extent.

2 The Watermarking Embedding Procedure Based on BWT

2.1 Berkeley Wavelet Transform

The mammalian perception which is from the outside world leads to a gradual transformation and extraction of image information between the visual paths. The mammalian primary visual cortex (V_1) has three properties: spatially localized, oriented and frequency bandpass. While Berkeley wavelet transform has same properties with the mammalian primary visual cortex (V_1) in choosing base. Thus BWT transform simulates the physiology characteristics of the receptive fields of neurons in mammalian primary visual cortex (V_1).

BWT transform is constituted by eight mother wavelets, in four pairs. Each pair has a different direction: 0, 45, 90 and 135. Within each pair, one wavelet is odd symmetric, the other is even symmetric. Using $\Theta = \{\theta\} = \{0, 45, 90, 135\}$ represents the degree, $\Phi = \{\varphi\} = \{o, e\}$ represents the symmetry, and $\beta_{\theta,\varphi}(x, y)$ represents mother wavelet. Then we can get four pairs of mother wavelet, $\beta_{\theta,\varphi}$, which are shown in Eqs. (1)–(8).

$$\beta_{0,o} = [-\mu(x,y/3) + \mu(x-2,y/3)]/\sqrt{6} \tag{1}$$

$$\beta_{0,e} = [-\mu(x,y/3) + 2\mu(x-1,y/3) - \mu(x-2,y/3)]/\sqrt{18} \tag{2}$$

$$\begin{aligned}\beta_{45,o} =& [-(\mu(x,y-2) + \mu(x-1,y) + \mu(x-2,y-1)) \\ &+ \mu(x,y-1) + \mu(x-1,y-2) + \mu(x-2,y)]/\sqrt{6}\end{aligned} \tag{3}$$

$$\begin{aligned}\beta_{45,e} =& [-(\mu(x,(y-1)/2) + \mu(x-1,y) + \mu(x-1,y-2) \\ &+ \mu(x-2,y/2)) + 2(\mu(x,y) + \mu(x-1,y-1) \\ &+ \mu(x-2,y-2))]/\sqrt{18}\end{aligned} \tag{4}$$

$$\beta_{90,o} = [\mu(x/3,y) - \mu(x/3,y-2)]/\sqrt{6} \tag{5}$$

$$\beta_{90,e} = [-\mu(x/3,y) + 2\mu(x/3,y-1) - \mu(x/3,y-2)]/\sqrt{18} \tag{6}$$

$$\begin{aligned}\beta_{135,o} =& [-\mu(x,y) + \mu(x-2,y-1) + \mu(x-1,y-2) \\ &+ \mu(x,y-1) + \mu(x-1,y) + \mu(x-2,y-2)]/\sqrt{6}\end{aligned} \tag{7}$$

$$\begin{aligned}\beta_{135,e} =& [-(\mu(x,y/2) + \mu(x-1,y) + \mu(x-1,y-2) \\ &+ \mu(x-2,(y-1)/2) + 2(\mu(x,y-2) + \mu(x-1,y-1) \\ &+ \mu(x-2,y)]/\sqrt{18}\end{aligned} \tag{8}$$

$\mu(x,y)$ is one binary function, which is defined as follows:

$$\mu(x,y) = \begin{cases} 1, & \mathit{if}\ 0 < x \le 1,\ 0 < y \le 1 \\ 0, & \mathit{otherwise} \end{cases} \tag{9}$$

Then, the mother wavelets, $\beta_{\theta,\varphi}$, are scaled and translated using the dilation Eq. (10) to produce daughter wavelets, $\beta_{\theta,\varphi}^{m,n,s}$, at multiple positions (m, n) in the x–y plane, and multiple scales, s:

$$\beta_{\theta,\varphi}^{m,n,s} = \beta_{\theta,\varphi}(3^s(x-m), 3^s(y-n))/s^2 \tag{10}$$

The BWT uses triadic scaling: the size of the daughter wavelets are scaled by powers of 3. Since the BWT is a complete, orthonormal set, it is self-inverting. An image can be reconstructed from its BWT coefficients using

$$I(x,y) = \sum_{\theta\in\Theta,\varphi\in\Phi} \sum_{m,n,s=0}^{\infty} \omega_{m,n,s}^{\theta,\varphi} \beta_{\theta,\varphi}(3^s(x-m), 3^s(y-n)) \tag{11}$$

Where $\omega_{m,n,s}^{\theta,\varphi}$ are the BWT coefficients representing of the image.

From the calculation equation above, Eq. (10) represents the Berkeley wavelet transform and Eq. (11) represents the inverse Berkeley wavelet transform. We can find the calculations of the orthogonal base at each position (x,y) have a relationship with the spatial position (x,y), which has four directions. So, the four orthogonal bases are spatially localized and oriented. Meanwhile, the biological experiments show that the human visual nerve cells have frequency selective response to the inputting of the spatial information, and have a certain frequency bandpass which ranges from 0.6 to 2.0 and averages by 1.3. The frequency bandpass of BWT transform ranges from 1.2 to 2.2, which averages by 1.675. So, the choice of Berkeley wavelet base is agreed with the mammalian primary visual cortex (V_1) by the property. The BWT transform simulates the physiology characteristics of the mammalian primary visual cortex (V_1).

2.2 The Digital Watermarking Embedding Procedure Based on BWT

As the Berkeley wavelet transform fits the human visual characteristics, we can embed the watermarking under BWT transform. Here, we propose one digital watermarking embedding algorithm based on BWT transform, the detail process is shown as follows:

Step 1: Pre-process the original image I and watermarking image W by scaling image I with power 3, then the new size are $3^{k}*3^{k}$ (k is an integer greater than 0). We scale the size of watermarking W as the same size before, and then use Key1 which belongs to the copyright owners to encrypt the watermarking. We call the pre-processed image Ip and Wp.

Step 2: Do BWT Transform on image Ip and Wp, then get BWT coefficient matrix I_{BWT} and W_{BWT}.

Step 3: Determine the watermarking embedding locations by using BWT transform coefficient matrix I_{BWT} and W_{BWT}. As matrix I_{BWT} and W_{BWT} have same size, we can superimpose the value of W_{BWT} onto the transform coefficient matrix I_{BWT} in corresponding position. The linear superimposing equation is defined as (12) and (13), and then we can get mixing matrix H_1 and H_2:

$$H_1 = I_{BWT} + \alpha_1 \times W_{BWT} \tag{12}$$

$$H_2 = I_{BWT} + \alpha_2 \times W_{BWT} \tag{13}$$

Where, α_1 is the watermarking embedding intensity factor, so in order to ensure high transparency of the image after watermarking embedding, this value can be small. α_2 is watermarking extraction factor, in order to ensure the high quality of extracting watermarking, this value can be higher.

Step 4: Get inverse transform matrix I_{BWT}' by doing inverse BWT transform on the mixing matrix H_1, and recover the size of I_{BWT}' with the same size of original image, then we can get image I' which embedded with watermarking. Using Key_2 which belongs to the copyright owners to encrypted H_2, then we can get key matrix Key_H, which should act as a validation watermarking key, and will be delivered to the next stage.

The details of watermarking embedding procedure flow chart are shown in Fig. 1.

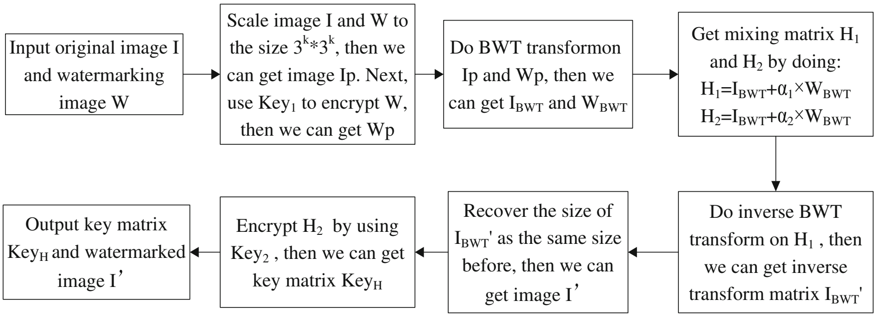

Input original image I and watermarking image W
Scale image I and W to the size 3^k*3^k, then we can get image Ip. Next, use Key1 to encrypt W, then we can get Wp
Do BWT transformon Ip and Wp, then we can get IBWT and WBWT
Get mixing matrix H1 and H2 by doing: H1=IBWT+α1×WBWT H2=IBWT+α2×WBWT
Do inverse BWT transform on H1 , then we can get inverse transform matrix IBWT'
Recover the size of IBWT' as the same size before, then we can get image I'
Encrypt H2 by using Key2 , then we can get key matrix KeyH
Output key matrix KeyH and watermarked image I'

Fig. 1. Flow chart of watermarking embedding procedure

3 The Watermarking Extracting Procedure Based on BWT and ICA

3.1 Blind Separation Technology ICA

ICA is to separate the observed signals which generated by independent sources and mixed by unknown ways. In this paper, we use ICA to detect the watermarking information. The advantage is that we do not need to consider the mixed ways and the strength of the watermarking embed into the host image. We can extract the watermarking if the random sequences which generated by the watermarking and the source image are satisfied with the independent conditions mutually.

In fact, we cannot determine signal sources uniquely in absence of some priori knowledge, because the blind separation problem has two inner uncertain solutions. Firstly, the output component's order is uncertain, so we cannot determine the recovered signal correspond to which component of the original signal. Secondly, the uncertainties of the output signal's amplitude, which make us cannot recover the true amplitude of the original signal sources. To the first question, as the key matrix is retained in the digital watermarking embedding procedure, we can obtain the mixing matrix by decrypting this key matrix, and then do inverse BWT transform to it. Thus, we can determine whether we separate the watermarking by the similarity of the separation matrix of ICA and the mixing matrix. The second question is the main reason which constraints ICA's extracted watermarking capacity. So if we use ICA to separate signal after extracting only part information of original watermarking image, mixing the signal separated with the rest of original watermarking signal will cause the distortion of the watermarking signal due to the change of amplitude. As we can embed watermarking information in the entire BWT transform coefficients, so embedding information are large. Here, we use the maximum embedded information ratio MEBR to measure the amount of the watermarking information which can be embedded in the original image.

$$\mathrm{MEBR} = \mathrm{WB}/\mathrm{IB} \tag{14}$$

Table 1. MEBR comparison results

Transform Style	MEBR of watermarking extracted by ICA
LSB	1
DCT	1/8
1 level DWT	1/2
2 level DWT	1/2+1/8
3 level DWT	1/2+1/8+1/32
BWT	1

Where, WB represents the maximum bit numbers of watermarking which can be embedded in the image, IB represents the bit numbers of the original image.

Table 1 shows that the maximum watermarking embedded ratio which are calculated with different watermarking extraction algorithms under different transform domains. When we embed the watermarking under DCT transform, firstly, we block the image and do DCT transform to each block, then select the frequency coefficients to embed watermarking. Often only 8 or less frequency coefficients can be embedded watermarking. Thus MEBR can reach up to 12.5 %. When we embed the watermarking under DWT transform, we often do one level wavelet transform to the entire image. Then we can get four regions, namely, LL_1,LH_1,HL_1,HH_1. And the watermarking embedding regions are LH_1 and HL_1, so the MEBR can reach up to 50 %. HH_1 region belongs to the high-frequency zone, the watermarking information can be filtered by using a low-pass filtering, and thus it is not suitable for embedding a watermarking. Although we can increase the watermarking capacity of LL_1 by using some statistical learning methods, the MEBR is unlikely to exceed 75 %. However, when the source image and the watermarking image are transformed by BWT, the whole watermarking transform matrix can be superimposed onto the source image transform matrix. As there are three reasons, firstly, BWT transform fits human visual characteristics; secondly, the size of the source image is equal to the watermarking image; thirdly, the watermarking embedding scheme is additive watermarking embedding scheme. Thus the MEBR can reach nearly up to 100 %.

3.2 The Watermarking Extracting Procedure Based on BWT and ICA

Here, we combine the advantage of ICA to process the blind separation with the BWT transform when we used in the embedding procedure, and then put forward one digital watermarking extraction algorithm based on BWT and ICA. The extraction process is shown as follows:

Step 1: We pre-process to the image which needs to be detected and use copyright key Key_2 to inverse decrypt validation watermarking key Key_H, and then we can get H_2 and its size. We scale the size of the detected image I' to the same size as before, then we can get image Ip'.

Step 2: Do BWT transform on Ip', then we can get BWT transform coefficient matrix I_{BWT}'.

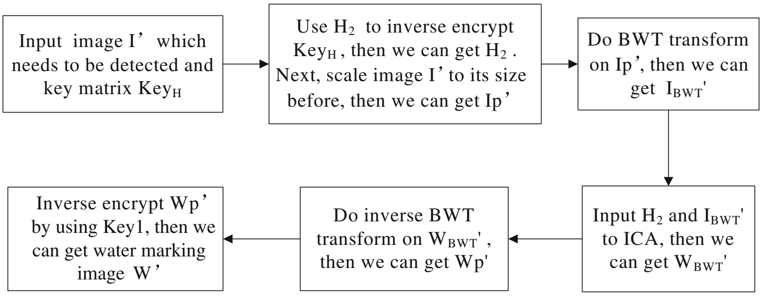

Input image I' which needs to be detected and key matrix Key$_H$
Use H$_2$ to inverse encrypt Key$_H$, then we can get H$_2$. Next, scale image I' to its size before, then we can get Ip'
Do BWT transform on Ip', then we can get I$_{BWT}$'
Input H$_2$ and I$_{BWT}$' to ICA, then we can get W$_{BWT}$'
Do inverse BWT transform on W$_{BWT}$', then we can get Wp'
Inverse encrypt Wp' by using Key1, then we can get water marking image W'

Fig. 2. Flow chart of watermarking extracting procedure

Step 3: We use I$_{BWT}$' and H$_2$ as our inputting, and use the blind separation technology ICA to separate mixing matrix, then we can get the watermarking coefficient matrix W$_{BWT}$'. And then do inverse BWT transform on W$_{BWT}$', then we can get Wp'.

Step 4: We decrypt Wp' by using the copyright owners key Key$_1$, then we can get the extracted watermarking image W'.

The details of the watermarking extracting procedure are shown in Fig. 2.

The image quality of the extracted watermarking can be measured by BER which defined as follows:

$$BER = 1 - \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} w(i,j) \cdot w'(i,j)}{\sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} w^2(i,j) \cdot \sum_{i=1}^{M} \sum_{j=1}^{N} w'^2(i,j)}} \tag{15}$$

4 Experimental Results and Analysis

4.1 Construction of Image Database

We extract part of typical images as our original testing image and watermarking image. Size are 243 × 243, as shown in Fig. 3 below.

In Fig. 3, image (3-a)–(3-i) are the original images, on behalf of rich color, rich texture, rich regional, detail plain, clear theme, color-scale and gray-scale images. Among them, Fig. (3-a)–(3-g) are color-scale images, Fig. (3-h), (3–i) are gray-scale images, Fig. (3-j) is a color-scale watermarking image, Fig. (3-k) is a gray-scale watermarking image. These images construct the database of our experiment.

4.2 Transparency Experiment

We choose the PSNR value to measure the quality of watermarking image. For the size of M × N gray-scale image, PSNR is defined as follows.

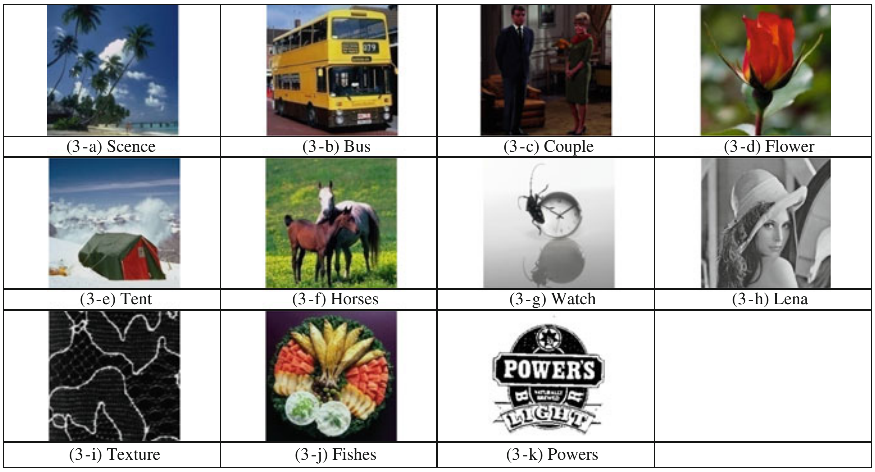

(3-a) Scence
(3-b) Bus
(3-c) Couple
(3-d) Flower
(3-e) Tent
(3-f) Horses
(3-g) Watch
(3-h) Lena
(3-i) Texture
(3-j) Fishes
(3-k) Powers

Fig. 3. Database images used in our experiments

$$PSNR = 10\lg\left(\frac{255^2 \times M \times N}{\sum_{i=1}^{M}\sum_{j=1}^{N}\left(I'(i,j) - I(i,j)\right)^2}\right) \tag{16}$$

For the size of M × N color-scale image, we use the average of three-channel RGB PSNR value to measure image quality.

$$\overline{PSNR} = (PSNR_R + PSNR_G + PSNR_B)/3 \tag{17}$$

In order to show the transparency of the watermarking algorithm, we embed watermarking image Fig. (3-j) to 24 bits Watch image Fig. (3-g). The results are shown in Fig. 4.

From Fig. 4, we can find that the watermarking transparency of this algorithm is good. In order to better illustrate the transparency of our algorithm, we embed the color-scale image (3-j) into the color-scale image (3-a)–(3-g) and embed the gray-scale watermarking (3-k) into the gray-scale image (3-h),(3-g). And then compute the PSNR value between the watermarking image and the original image. The generated results are shown in Table 2.

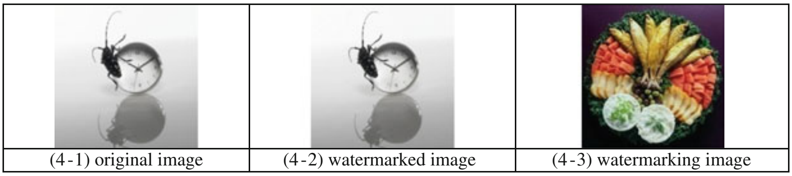

(4-1) original image
(4-2) watermarked image
(4-3) watermarking image

Fig. 4. Transparency experiment

Table 2. Experiment of watermarking transparency

	$\overline{PSNR}$	PSNR_R	PSNR_G	PSNR_B
3-a	45.7884	44.5741	45.7002	47.0909
3-b	45.6866	44.9539	45.6302	46.4757
3-c	45.1867	44.3810	45.0482	46.1310
3-d	45.6693	45.0324	45.6056	46.3700
3-e	46.0259	45.2031	46.0145	46.8602
3-f	45.8591	45.0048	45.9675	46.6050
3-g	46.6032	45.8702	46.2284	46.6032
3-h	44.810			
3-i	45.610			

From Table 2, we can find the PSNR value of watermarking image and original image are larger than 40. And from the experience, when PSNR is greater than 40, the quality of one image is ideal. Therefore, it is more robust to meet the standard of transparency. So the watermarking transparency of our algorithm is good. In order to further test whether this algorithm is better than others, we use three classic digital watermarking algorithms (LSB, DCT, DWT) to make comparison with our proposed algorithm. We have done the following PSNR comparison experiment, the experiment result is shown in Fig. 5.

From Fig. 5, we can find the PSNR value of our algorithm is slightly lower than that generated by the space domain LSB digital watermarking algorithm which is not robust to some general attacks, but better than transform domain digital watermarking algorithm DCT and DWT. It shows that our algorithm has similar transparency with LSB, and is better than the transparency of DCT and DWT. So the transparency of our algorithm is good.

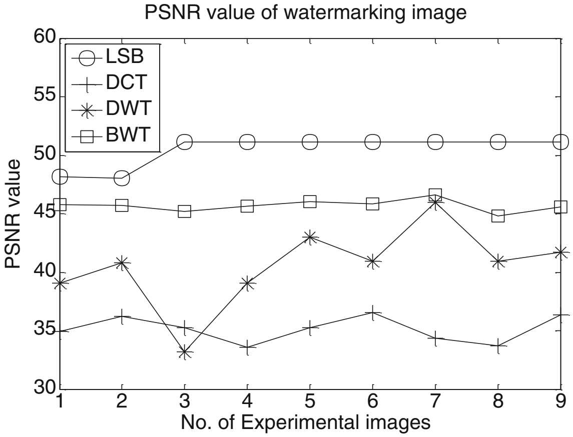
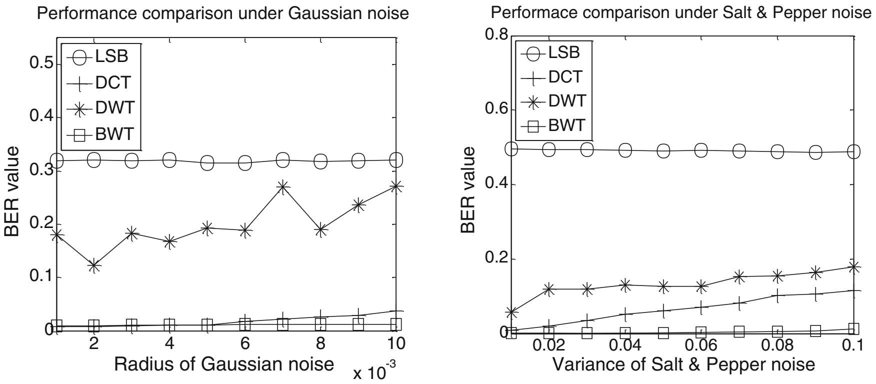


Fig. 5. Transparency comparison experiment



(6-a) (6-b)

Fig. 6. Robustness comparison under noises, (6-a) represents the Robustness comparison under Gaussian noise, (6-b) represents the Robustness comparison under Salt and Pepper noise

4.3 Robustness Experiment Under Additive Noise Attack

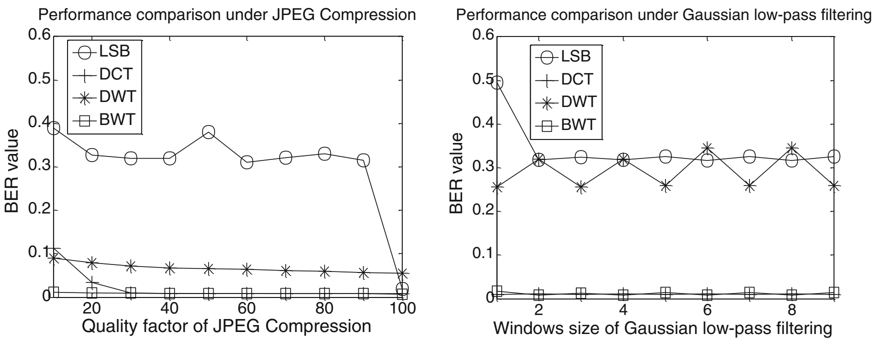
In order to test the robustness of our algorithm under additive noise, we have added two additive noises, they are the Gaussian noise and the salt & pepper noise. We calculate the BER value between the original image and the watermarking image which is extracted under the different Gaussian noise variance and the salt & pepper noise variance, the experimental results are shown in Fig. 6-a, 6-b.

As we can see from Fig. 6-a, our algorithm's robustness against the Gaussian noise is similar to digital watermarking algorithm based on DCT, and are superior to the robustness of LSB and DWT. Similarly, from Fig. 6-b, we can find that the proposed algorithm's robustness resistant to the salt & pepper noise are significantly better than other three watermarking algorithms. It can be concluded that our algorithm is robust to the additive noises.

4.4 Robustness Experiment Under JPEG Compression Attack

In order to further study the proposed algorithm, we have done the following experiment under JPEG compression. We perform JPEG compression processing on these images with ten different quality factors (10, 20,…, 100), and then extract the watermarking from these images. Meanwhile, we calculate the BER value between the original image and watermarking image. The experimental results are shown in Fig. 7-a.

From Fig. 7-a, we can find that the proposed algorithm is more robust to the JPEG compression than the traditional three watermarking algorithms. Even though the quality factors are very small, the BER values between original image and watermarking image are very small, the watermarking detected are still similar to original watermarking. So the algorithm proposed is more robust to the JPEG compression.



(7-a) (7-b)

Fig. 7. Robustness comparison under JPEG compression attack and Gaussian low-pass filtering attack, (7-a) represents the Robustness comparison under JPEG compression attack, (7-b) represents the Robustness comparison under Gaussian low-pass filtering attack

4.5 Robustness Experiment Under Gaussian Low-Pass Filtering Attack

In order to test the robustness of our algorithm, we add Gaussian low-pass filtering (which size ranges from 1*1 to 9*9) to these images, the experimental results are shown in Fig. 7-b. From Fig. 7-b, we can find that the proposed algorithm is more robust than DWT, LSB for Gaussian low-pass filtering attack and slightly better than DCT. Even though the size is larger, our algorithm is still robust. So from above, we can find that our algorithm is robust to Gaussian low-pass filtering processing.

4.6 Robustness Comparison Under Attack

By making comparison with the robustness experiments of the four algorithms before, we sum up the following form.

In Table 3, the symbol ">" means that robustness is better than other algorithms. Thus from Table 3, we can find that our proposed digital watermarking algorithm based on BWT has the similar transparency with LSB, and is better than other three algorithms against attack. Therefore, the algorithm presented in this paper combines with large capacity, good transparency and robustness. As BWT transform is composed by the four complete orthogonal matrices, so the transform speed and inverse

Table 3. Robustness comparison results

Robustness Index	Robustness comparison
PSNR Value	LSB>BWT>DWT>DCT
Resist Gaussian noise	BWT>=DCT>DWT>LSB
Resist salt & pepper noise	BWT>DCT>DWT>LSB
Resist JPEG compression	BWT>DCT>DWT>LSB
Resist Gaussian low-pass filtering	BWT>=DCT>DWT>=LSB

transform speed is fast. Therefore, this algorithm has lower complexity and combines the imperceptibility, robustness, real-time and high capacity of digital watermarking algorithms.

5 Conclusions

In this paper, we propose a image watermarking algorithm based on Berkeley wavelet transform which fits the human visual characteristics, and ensures the good transparency of the embedded watermarking image. It uses a blind separation technology ICA to detect the watermarking information. Experimental results show that the algorithm not only has good visual effects, but also robust to some common attacks, such as JPEG compression, additive noise, as well as Gaussian low-pass filtering attacks. The experimental results are much better than the commonly used DCT, DWT and LSB digital watermarking algorithms. Moreover, the algorithm can be embedded with a large amount of information, and the computation complexity is low. Thus, it has great value in practical application. Therefore, our future work will play its unique advantages by applying it to the engineering practice.

Acknowledgments. The author gratefully acknowledge the support from Natural Science Foundation of China (No.61073116 & No.61003038) and the support from key natural science fund of Anhui province (No. KJ2012A008 & No. KJ2010A006).

References

1. Van Schyndel, R.G., Tirkel, A.Z., Osbome, C.F.: A digital watermarking. In: Proceedings of the IEEE International Conference on Image Processing, Texas, USA, pp. 86–90 (1994)
2. Nikolaidis, N., Pitas, I.: Copyright protection of images using robust digital signatures. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Georagia, USA, pp. 2168–2171 (1996)
3. Chen, Y.Q., Hu, H.P., Li, X.T.: Digital watermarking restorable technique with synergetic association in space domain. Pattern Recognit. Artif. Intell. **5**, 535–540 (2005)
4. Cui, L.H.: Adaptive multiwavelet-based watermarking through JPW masking. IEEE Trans. Image Process. **4**, 1047–1060 (2011)
5. Nikolaidis, N., Pitas, I.: Multichannel L-filters based on reduced ordering. IEEE Trans. Circuits Syst. Video Technol. **5**, 470–484 (1996)
6. Cox, I.J., Kilian, J., Leighton, F.T.: Secure spread spectrum watermarking for multimedia. IEEE Trans. Image Process. **12**, 1673–1687 (1997)
7. Hsu, C.T., Wu, J.L.: Multiresolution watermarking for digital image. IEEE Trans. Circuits Syst.-II: Analog Digit. Signal Process. **45**, 1097–1101 (1998)
8. Kwitt, R.: Lightweight detection of additive watermarking in the DWT-Domain. IEEE Trans. Image Process. **2**, 474–484 (2011)
9. Lou, D., Liu, J.L., Hu, M.C.: Adaptive digital watermarking using network technique. In: Proceedings of the IEEE International Carnahan Conference on Security Technology, Taipei, Taiwan, pp. 325–332 (2003)

10. Siagian, C., Itti, L.: Rapid biologically-inspired scene classification using features shared with visual attention. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 300–312 (2007)
11. Davis, K.J., Najarian, K.: Maximizing strength of digital watermarkings using neural networks. In: Proceedings of the International Joint Conference on Neural Networks, Washington DC, USA, pp. 2893–2898 (2001)
12. Wang, X.Y., Wu, J., Niu, P.P.: A new digital image watermarking algorithm resilient to desynchronization attacks. IEEE Trans. Inf. Forensics Secur. **2**, 655–663 (2007)
13. Wang, X.Y., Hou, L.M., Wu, J.: Feature-based digital image watermarking scheme robust to geometric attacks. Acta Autom. Sinica **34**, 1–6 (2008)
14. Willmore, B., Prenger, R.J., Wu, M.C., et al.: The berkeley wavelet transform: a biologically-inspired orthogonal wavelet transform. Neural Comput. **6**, 1537–1564 (2008)
15. Rui, T., Shen, C.L., Tian, Q.: Comparison and analysis on ICA and PCA's ability in feature extraction. Pattern Recognit. Artif. Intell. **1**, 124–128 (2005)
16. Hyvrinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Netw. **13**, 411–430 (2000)

A New Weighted Sparse Representation Based on MSLBP and Its Application to Face Recognition

He-Feng Yin and Xiao-Jun Wu(✉)

School of IoT Engineering, Jiangnan University, Wuxi 214122, China
yinhefeng@126.com, wu_xiaojun@yahoo.com.cn

Abstract. Face recognition via sparse representation-based classification has received more and more attention in recent years. This approach has achieved state-of-the-art results, which outperforms traditional methods, especially when face image pixels are corrupted or occluded. In this paper, we propose a new weighted sparse representation method called WSRC-MSLBP which utilizes the multi-scale LBP (MSLBP) feature to measure similarity between face images, and to form the weight matrix. The proposed WSRC-MSLBP method not only represents the test sample as a sparse linear combination of all the training samples, but also makes use of locality of local binary pattern. Experimental results on publicly available databases show that the proposed WSRC-MSLBP method is more effective than sparse representation-based classification algorithm and the original weighted sparse representation method.

Keywords: Sparse representation-based classification (SRC) · Face recognition · Weighted sparse representation · Multi-scale LBP feature · Feature extraction

1 Introduction

Automatic face recognition [1] remains one of the most visible and challenging research topics in computer vision, machine learning and biometrics. It is widely applied to different fields including biometric authentication, security applications and human computer interaction. Compared with other biometrics, such as iris identification and palm identification, face recognition has the advantages of being convenient, immediate and well accepted.

In face recognition, face representation plays a vital part and its effectiveness intimately associates with the final recognition results. Face representation, in essence, is feature extraction for face images. It is a crucial issue in face recognition that which low-dimensional features are the most effective or informative for discrimination, and just from the outset, diverse pioneering approaches have been proposed to this end. Conventional facial features can be roughly divided into holistic features (PCA [2], LDA [3], LPP [4], etc.) and local features (LBP [5], SIFT [6], etc.). There are so many feature extraction methods that when dealing with a specific problem, practitioners tend to be in a dilemma of which features to use. However, recent progress in

Z.-H. Zhou and F. Schwenker (Eds.): PSL 2013, LNAI 8183, pp. 104–115, 2013.
DOI: 10.1007/978-3-642-40705-5_10,
© Springer-Verlag Berlin Heidelberg 2013

compressed sensing and sparse representation leads to novel algorithms for face recognition. Wright et al. [7] put forward a seemingly simple yet effective method called sparse representation-based classification (SRC), the training samples are used to form a structured dictionary, and the test sample is decomposed based on the dictionary to get its coding coefficient via l^1-norm minimization, then the test image is classified to the class which produces the minimum reconstruction error. In SRC, the precise choice of feature space is no longer critical, and it is robust to occlusion. However, SRC forms the dictionary by using all the training images, thus the generated dictionary may have a huge size, which makes adverse effects to the following sparse solver. To overcome this drawback, Yang et al. [8] proposed an unsupervised dictionary learning algorithm to obtain dictionary elements for each class. Yang et al. [9] presented a novel dictionary learning method which introduces Fisher criterion to the objective function in order to improve the pattern classification performance. Li et al. [10] came up with a local sparse representation based classification (LSRC) scheme, which performs sparse decomposition in a local manner. In LSRC, they exploited kNN rules to find k neighbors for the test samples, and the selected samples are utilized to construct the over-complete dictionary.

For general pattern classification problems such as dimensionality reduction, classification, clustering, etc., the locality structure of data has been observed to be critical [11, 12]. Lu et al. [13] took data locality into consideration, and imposed locality on the l^1 regularization. They utilized the distance between test samples and training samples to characterize their similarity, through this way, they formed the weight matrix. By solving a weighted l^1-minimization problem, they achieved impressive results on the Extended Yale B, AR databases and several datasets from the UCI repository. Nevertheless, Wang et al. [14] argued that similarities are not merely related to distance, and it is shown that traditional distance-based similarity measure may lead to high classification error rates even on several simple datasets. In addition, according to related researches about local binary pattern (LBP), features coded by LBP have highly discriminative power [15], this property makes it suitable for image classification tasks. Inspired by these findings, we intend to use the similarity of LBP features of images to form the weight matrix, this can better make use of data locality, thus boost the accuracy of face recognition.

The rest of this paper is organized as follows: LBP and sparse representation-based classification is reviewed in Sects. 2 and 3 respectively. Section 4 presents the proposed method. Extensive experiments were conducted on publicly available databases to verify the effectiveness of the proposed method in Sect. 5. Finally, conclusions are drawn in Sect. 6.

2 Local Binary Pattern

The LBP operator was first introduced by Ojala [16] and used as texture descriptor. Then Ahonen [5] applied it to face recognition and obtained outstanding results, which demonstrates that LBP is able to well describe face images.

The original LBP operator was defined as a window of size 3 × 3. This operator uses the value of the center pixel as a threshold, and the 8 surrounding pixels whose

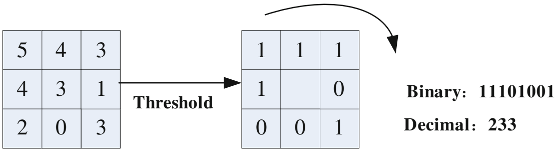


Fig. 1. The original LBP operator.

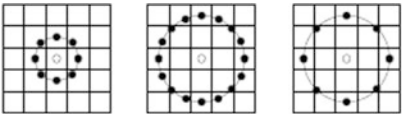


Fig. 2. The circular (8, 1), (16, 2) and (8, 2) neighborhoods.

value is higher than or equal to the value of the threshold is assigned a binary value 1, otherwise the value is 0. When this process is accomplished, 8 values can be read sequentially in the clockwise direction. The 8-bit binary number or its equivalent decimal number can be assigned to the center pixel and it can describe the texture information of an image. The basic LBP operator is illustrated in Fig. 1.

In order to facilitate the analysis of textures with different scales, the basic LBP operator is extended by combining neighborhoods with different radius. In this case, *P* points on the edge of a circle, whose radius is *R*, are sampled and compared with the value of the center pixel. For ease of presentation, the notation *(P,R)* is employed to formulate *P* sampling points on a circle of radius of *R*. See Fig. 2 for an example of circular neighborhoods.

However, not all the patterns coded by LBP are useful for describing the characteristics of textures, so it is necessary to choose which local patterns are account for a major part of all patterns. These patterns are referred to as uniform patterns. In their experiment, Ojala [15] found that uniform patterns provide about 90 percent of the 3 × 3 texture pattern in examined surface textures. A local binary pattern is called uniform if the binary pattern contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is considered circular [17]. For example, the patterns 11111111 (0 transitions), 00111000 (2 transitions) and 11100111 (2 transitions) are uniform whereas the pattern 00110110 (4 transitions) is not. Experimental results have demonstrated uniform patterns can describe most of the texture information, at the same time, they have strong ability to do classification tasks.

A histogram of the labeled image $f_l(x, y)$ can be defined as

$$H_k = \sum_{x,y} I\{f_l(x, y) = k\}, k = 0, \ldots, n - 1 \tag{1}$$

in which n is the number of different labels produced by the LBP operator and

$$I\{A\} = \begin{cases} 1, \ A \ is \ true \\ 0, \ A \ is \ false \end{cases}$$

This histogram consists of information about the distribution of the local micro-patterns, including spots, flat areas, edge ends, and curves.

Generally, when we extract features from face images, we can divide the face image into small blocks. And features are extracted from each block independently. The descriptors are then concatenated to form a global description of the face image. In this way we can obtain a description of the face image on local and holistic levels. Several possible similarity measures have been proposed for histograms, e.g. histogram intersection, log-likelihood statistic, χ^2 statistic etc. [5].

Experimental results have shown that the performance of χ^2 statistic is better than histogram intersection and log-likelihood statistic when using uniform patterns [5]. In this paper, since we use LBP uniform patterns of (8,1) neighborhood, so we choose χ^2 statistic to measure the similarity of histograms.

3 Sparse Representation-based Classification and Weighted Sparse Representation

3.1 Sparse Representation-based Classification (SRC)

Theoretical results show that well-aligned images of a convex, Lambertian object lie near a low-dimensional feature space of the high-dimensional image space [18]. This is the only prior knowledge about the training samples in SRC. The idea of SRC is presented as follows [7].

Suppose we have C distinct classes, given sufficient training samples of the i-th object class, the size of face images is $w \times h$, and the total number of samples of i-th class is n_i. We stack the n_i training images from the i-th class as columns of a matrix $A_i = \left[v_{i,1}, \ldots, v_{i,n_i}\right] \in R^{m \times n_i}$ $(m = w \times h)$. For a test sample $y \in R^m$ belongs to this class, according to linear subspace theory, y can be approximated by the linear combination of the samples within A_i, i.e.

$$y \approx \alpha_{i,1} v_{i,1} + \alpha_{i,2} v_{i,2} + \ldots + \alpha_{i,n_i} v_{i,n_i} \tag{2}$$

$$\alpha_{i,j} \in R, j = 1, 2 \ldots, n_i.$$

Since the initial identity of the test sample y is unknown, let A be the concatenation of the n training samples from all the C classes, where $\sum_{i=1}^{C} n_i = n$, then we can define a new matrix A:

$$\begin{aligned} A &= [A_1, A_2, \cdots, A_C] \\ &= \left[v_{1,1}, \ldots, v_{1,n_1}, \ldots, v_{i,1}, v_{i,2}, \ldots, v_{i,n_i}, \ldots, v_{C,1,\ldots,} v_{C,n_C}\right] \end{aligned} \tag{3}$$

If we use the new matrix A to represent the test image y, that is

$$y = Ax_0 \ \in \ R^m \tag{4}$$

where $x_0 = [0, \ldots, 0, \ldots, \alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,n_i}, \ldots, 0, \ldots, 0]^T \in R^n$ is a coefficient vector whose entries are zero except those associated with the i-th class.

In robust face recognition, the system $y = Ax$ is always under-determined, so it has an infinity of solutions, but we just need to find an optimal solution. Conventionally, this problem is settled by choosing the minimum l^2-norm solution. However, the solution is non-sparsity and it has no discriminative information. This motivates us to seek the sparsest solution to $y = Ax$, solving the following optimization problem:

$$(l^0)\ x_0 = \arg\ \min ||x||_0,\ subject\ to\ Ax = y \tag{5}$$

where $||.||_0$ denotes the l^0-norm, which counts the number of nonzero elements in a vector.

Unfortunately, the problem of finding the sparsest solution of an under-determined system of linear equations is NP-hard. Usually, one can use greedy pursuit algorithms to find a suboptimal yet sparse enough solution, e.g. matching pursuit [19], orthogonal matching pursuit [20], stage-wise orthogonal matching pursuit [21] etc.

Recent progress in the theory of sparse representation and compressed sensing reveals that if the solution x_0 is sparse enough, the solution to the l^0-minimization problem (5) is equal to the following l^1-minimization problem [22]:

$$(l^1)x_1 = \arg\ \min ||x||_1, subject\ to\ Ax = y \tag{6}$$

To solve the l^1-minimization problem, one can use gradient projection method [23], homotopy algorithm [24], iterative shrinkage-thresholding [25] etc.

When dealing with small dense error, we can modify the aforesaid l^1-minimization problem (6) to obtain a stable l^1-minimization problem:

$$(l_s^1)\ \ x_1 = \arg\min ||x||_1, subject\ to\ ||Ax - y||_2 \leq \varepsilon \tag{7}$$

where $\varepsilon > 0$ is error tolerance.

Once the coefficient vector $\hat{x}_i$ is obtained via (6) or (7), the test sample y is assigned to the class which minimizes the residual between y and $\hat{y}_i$:

$$\min_i r_i(y) = ||y - A\delta_i(\overset{\wedge}{x_1})||_2 \tag{8}$$

where $\delta_i(x)$ is an operator that selects the coefficients in x only associated with class i, $\hat{y}_i = A\delta_i(\hat{x}_i)$ is a sample that approximates the given test sample y.

With sufficient training samples, SRC does achieve excellent results. However, if the training samples is not enough, SRC may perform worse than conventional classifiers, thus makes SRC unstable. To overcome the drawback of SRC, Lu et al. [13] proposed a more robust weight sparse representation method which integrates both sparsity and data locality structure into a unified framework. The WSRC algorithm will be described in Sect. 3.2.

3.2 Weighted Sparse Representation-based Classification (WSRC)

SRC can ensure sparsity, but sometimes it may lose locality information. It has been shown that local sparse coding is effective for image classification. Weighted Sparse Representation-based Classification (WSRC) is a method for local sparse coding.

WSRC preserves the similarity between the test sample and its neighboring training data while seeking the sparse linear representation [13]. WSRC solves the following weighted l^1-minimization problem:

$$(Weighted\ l^1)\overset{\wedge}{x_1} = \arg\min ||Wx||_1\ subject\ to\ y = Ax \tag{9}$$

As mentioned above, A is the matrix that contains all of training samples, and each column of A is a sample, y is the input test sample, x is coding coefficient, W is a block-diagonal matrix, which is the locality adaptor that penalizes the distance between y and each training data. In [13], W is defined as

$$diag\,(W) = [dist\,(y, x_1^1), \ldots, dist(y, x_{n_c}^C)]^T$$

where $dist\,(y,\, x_i^c) = \left\|y - x_i^c\right\|^s$ denotes the distance between y and the i-th sample of c-th class, and s is the locality adaptor parameter.

When coping with occlusion, weighted l^1-minimization problem (9) can be extended to the following stable l^1-minimization problem:

$$(Weighted\ l_s^1)\hat{x}_1 = \arg\min ||Wx||_1\ subject\ to ||y - Ax|| \le \varepsilon \tag{10}$$

When obtaining the coding efficient x, the subsequent classification procedure is similar to that of SRC.

4 New Weighted Sparse Representation-based Classification

In WSRC, Lu [13] used the distance between test samples and training samples plus a locality adaptor parameter to form the weight matrix. In this way, they can utilize data locality while seeking the sparse linear representation. However, local feature is not extracted effectively.

When it comes to local feature extraction, features extracted by local binary pattern have highly discriminative power, which makes it suitable for image classification tasks. So in this paper, we make use of the similarity of LBP features between test samples and training samples to form the weight matrix. Considering wavelet transform has the nice features of space-frequency localization and multi-resolutions, we use wavelet transform to get the multi-scale LBP features of face images.

Wavelet transform has been introduced in our method to do the preprocess of the face images, it can reduce noise of images, and the low frequency component is a coarser approximation to the original image. Thus the wavelet image should be more suitable for recognition.

The LBP features that extracted from the low frequency component after 1-level wavelet transform have meaningful local and global features, and these features contribute a lot to face recognition. We divide the face image into $m \times n$ blocks. And the LBP features that extracted from the low frequency component after 2-level wavelet transform mainly have global features. When the aforesaid LBP features of two low frequency components are concatenated, we can gain the multi-scale LBP feature of a face image.

Given all that, the procedure of the proposed method WSRC-MSLBP is presented as follows:

1. Do wavelet transform to the training samples, and obtain the 1-level and 2-level low frequency components respectively.
2. Divide the 1-level and 2-level low frequency components into small blocks, then extract LBP features for each small block.
3. Concatenate the LBP features of all the small blocks to form the multi-scale LBP feature of the original face image.
4. The input test sample is also processed according to step 1–3 and obtain its multi-scale LBP feature.
5. Use χ^2 statistic, that is, $\chi^2(S,M) = \sum_k \frac{(S_k - M_k)^2}{S_k + M_k}$ (S denote a test sample and M denote the model) to measure the similarity of histograms between test samples and training samples, then form the weight matrix W in weighted l^1-minimization problem (10).
6. Solve the weighed l_s^1 problem (10) and gain the coding efficient $\hat{x}_i$ of the test sample y.
7. Calculate the reconstruction error $r_i(y) = \|y - A\delta_i(\hat{x})\|_2$ for each class i, then classify the test sample y based on which class yields the least reconstruction error.

5 Experiments and Analysis

In this section, we conduct experiments on publicly available databases for face recognition. The XM2VTS and AR databases are used to verify the performance of the proposed method and its competing methods, i.e. Nearest Neighbor Classifier (NNC), SRC and WSRC. As [13] does, we use three methods for dimensionality reduction, namely Eigenfaces [26], Fisherfaces [3] and Randomfaces [27]. We use the SPAMS package [28, 29] to solve the stable weighted l^1-minimization problem, and the basis function of wavelet transform is *coif4*. In SRC, the error tolerance ε is 0.005. In WSRC, the error tolerance ε and s are 10^{-4} and 0.5, 10^{-4} and 1.5 for XM2VTS and AR respectively. In Fisherfaces, LDA is preceded by PCA to avoid the problem of rank deficiency, and during our experiment, we choose components than account for 90 % energy.

One concern about Randomfaces is its stability, i.e., for an individual trial, the selected features could be bad. In order to reduce variation, when using Randomfaces, we generate 5 random projection matrices. And classification is based on the minimum average residual or distance.

5.1 Experiments on the XM2VTS Database

The XM2VTS database is a multi-modal database which consists of video sequences of talking faces recorded for 295 subjects at one month intervals. The data has been recorded in 4 sessions with 2 shots taken per session. From each session two facial images have been extracted to create an experimental face database of size 55 × 51.

Table 1. Recognition rate (%) of different methods on the XM2VTS database and the associated dimension of features

Dimension	10	30	50	70	90
(a) Eigenfaces					
NNC	64.25	80.25	82.75	84.50	86.25
SRC	63.50	86.75	89.50	91.50	92.50
WSRC	65.75	89.75	91.00	92.25	93.25
Proposed	96.25	96.25	96.25	96.25	96.25
(b) Randomfaces					
NNC	76.25	82.25	82.50	84.25	86.00
SRC	69.25	89.50	88.25	86.75	86.25
WSRC	73.75	88.75	91.50	91.75	92.75
Proposed	95.00	95.50	94.75	94.75	95.00
(c) Fisherfaces					
NNC	85.25	90.00	90.75	91.00	90.00
SRC	84.50	90.00	91.50	92.25	92.50
WSRC	83.75	90.50	91.75	93.00	92.25
Proposed	97.00	96.50	96.50	96.25	96.25

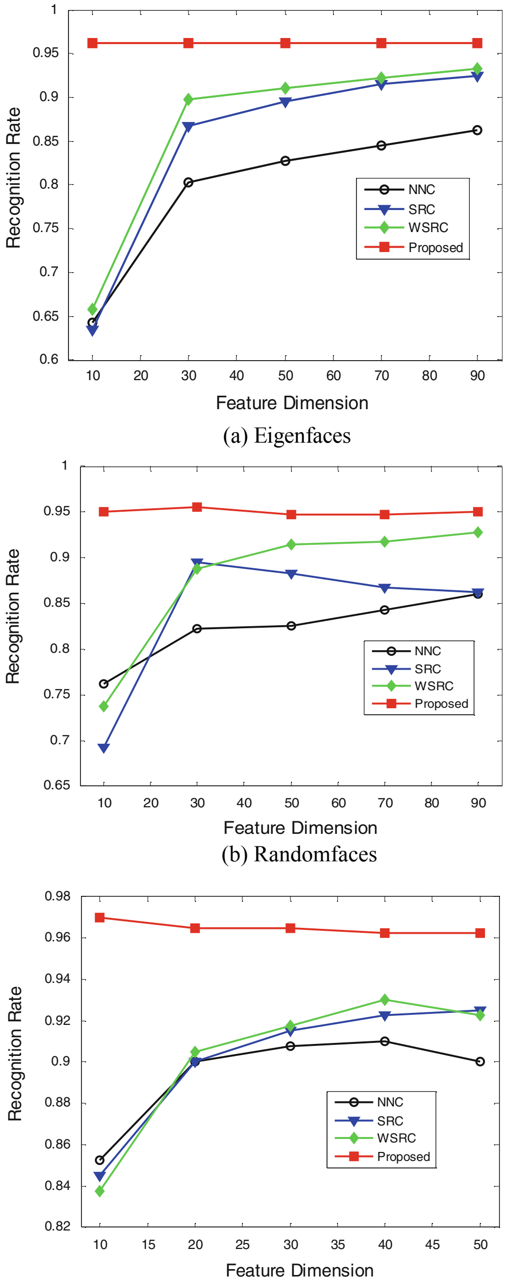
In our experiment, we chose a subset of the dataset consisting of 100 subjects. For each subject, four images are used as training samples, the rest for testing, and the face image is divided into 8 × 8 blocks when extracting the LBP features. Figure 3 shows the recognition performance for various features, in conjunction with four different classifiers: NNC, SRC, WSRC and the proposed method. Table 1 shows the detailed recognition accuracy of the methods considered.

5.2 Experiments on the AR Database

The AR face database consists of over 4,000 frontal images for 126 subjects. For each subject, 26 pictures were taken in two separate sessions. These images include more facial variations, including illumination change, expressions, and facial disguises. In the experiment, we chose a subset of the dataset (with only illumination and expression changes) consisting of 20 male subjects and 20 female subjects. For each subject, the seven images from Session 1 were used for training, and the other seven from Session 2 for testing, and the face image is divided into 2 × 5 blocks when extracting the multiscale LBP features. The comparison of competing methods is given in Table 2.

Based on the results obtained on the XM2VTS database, we have the following observations:

1. In lower dimensions (e.g. the first two rows in the first column in Table 1), SRC performs worse than NNC. The reason for this is that in lower dimensions, the linear measurements are insufficient, so the sparse recovery is not correct, this may have a direct impact on the performance of SRC.
2. WSRC outperforms SRC in most cases when using Eigenfaces and Randomfaces for dimensionality reduction, especially in lower dimensions. This indicates that



(a) Eigenfaces

(b) Randomfaces

(c) Fisherfaces

Fig. 3. Curves of recognition rate by different methods versus feature dimension on the XM2VTS database.

Table 2. Recognition rate (%) of different methods on the AR database and the associated dimension of features

Dimension	10	50	100	200	270
(a) Eigenfaces					
NNC	60.36	75.00	75.00	74.64	75.00
SRC	64.64	98.21	98.93	98.93	98.93
WSRC	70.36	98.57	99.29	98.93	98.57
Proposed	98.93	99.29	99.29	99.29	99.29
(b) Randomfaces					
NNC	65.00	72.50	72.50	74.64	73.93
SRC	29.64	74.29	79.64	83.93	84.29
WSRC	70.71	89.29	87.86	84.64	84.64
Proposed	89.64	91.79	93.57	93.93	92.50
(c) Fisherfaces					
NNC	85.00	97.50	98.57	98.57	98.57
SRC	72.86	90.71	97.14	98.57	98.93
WSRC	76.79	95.36	98.21	99.29	99.29
Proposed	99.29	99.29	99.29	99.29	99.29

in lower dimensions, the data locality contains more discriminative information than data linearity, and the imposed locality dose boost recognition performance.

3. It is interesting to find that when using Fisherfaces, in lower dimensions, the performance of NNC is better than that of SRC and WSRC, that is because the aim of LDA is to maximize the ratio of the between-class scatter matrix and the within-class scatter matrix, and the simple classifier NNC could separate data from different classes.
4. Whatever dimensionality reduction method it utilizes, recognition accuracy of the proposed method WSRC-MSLBP outperforms all the other three approaches significantly. This is because we also take data locality into consideration, in addition, we use the multi-scale LBP feature to measure similarity of test samples and training samples. Thus we can better preserve similarity between test samples and training samples, at the same time, we can make full use of the discriminative power of LBP features.

Similar results can also be seen on the AR database, the proposed method WSRC-MSLBP achieves state-of-the-art results.

6 Conclusions

In this paper, we propose a new weighted sparse representation method called WSRC-MSLBP, which uses the similarity of the multi-scale LBP features of face images to form the weight matrix, this can better make use of data locality, so as to boost the performance of face recognition. Experiments conducted on the XM2VTS and AR databases show the feasibility and effectiveness of the new method. However, in this paper, we do not consider the situation when face images are corrupted or occluded explicitly, and it is not uncommon in real-world situations, so in future, we will investigate a more robust method for face recognition.

Acknowledgments. This work was supported by the 111 Project of Chinese Ministry of Education, Grant No. B12018.

References

1. Chellappa, R., Wilson, C., Sirohey, S.: Human and machine recognition of faces: a survey. Proc. IEEE **83**(5), 705–741 (1995)
2. Turk, M., Pentland, A.: Face recognition using Eigenfaces. In: Proceedings of Computer Vision and Pattern Recognition, pp. 586–591 (1991)
3. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces versus Fisherfaces: recognition using class specific linear projection. IEEE PAMI **9**(7), 711–720 (1997)
4. He, X.-F., Yan, S., Hu, Y., Niyogi, P., Zhang, H.: Face recognition using Laplacianfaces. IEEE Trans. Pattern Anal. Mach. Intell. **27**(3), 328–340 (2005)
5. Ahonen, T., Hadid, A., Pietikäinen, M.: Face Recognition with Local Binary Patterns. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004, Part I. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
6. Bicego, M., Lagorio, A., Grosso, E., Tistarelli, M.: On the use of SIFT features for face authentication. In: Proceedings of IEEE Conference on Biometrics, in Association with CVPR Biometrics, pp. 35–35 (2006)
7. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE PAMI **31**(2), 210–227 (2009)
8. Yang, M., Zhang, L., Yang, J., Zhang, D.: Metaface learning for sparse representation based face recognition. In: Proceedings of the IEEE International Conference Image Processing (ICIP), pp. 1601–1604 (2010)
9. Yang, M., Zhang, L., Feng, X., Zhang, D.: Fisher discrimination dictionary learning for sparse representation. In: ICCV, pp. 543–550 (2011)
10. Li, C., Guo, J., Zhang, H.: Local sparse representation based classification. In: ICPR, pp. 649–652 (2010)
11. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
12. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Proceedings of CVPR, pp. 3360–3367 (2010)
13. Lu, C., Min, H., Gui, J., Zhu, L., Lei, Y.: Face recognition via weighted sparse representation. J. Vis. Commun. (2012) http://dx.doi.org/10.1016/j.jvcir.2012.05.003
14. Wang, M., Hua, X., Tang, J., Hong, R.: Beyond distance measurement: constructing neighborhood similarity for video annotation. IEEE Trans. Multimed. **11**(3), 465–476 (2009)
15. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
16. Ojala, T., Pietikinen, M.: A comparative study of texture measures with classification based on feature distribution. Pattern Recognit. **29**(1), 51–59 (1996)

17. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. PAMI **28**(12), 2037–2041 (2006)
18. Basri, R., Jacobs, D.: Lambertian reflectance and linear subspaces. IEEE Trans. Pattern Anal. Mach. Intell. **25**(2), 218–233 (2003)
19. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process **41**(12), 3397–3415 (1993)
20. Pati, Y., Rezaiifar, R., Krishnaprasad, P.: Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In: Signals, Systems and Computers Conference Record of The Twenty-Seventh Asilomar, vol. 1, pp. 40–44 (1993)
21. Donoho, D., Tsaig, Y., Drori, I., Starck, J.: Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. IEEE Trans. Inf. Theory **58**(2), 1094–1121 (2006)
22. Donoho, D.: For most large underdetermined systems of linear equations the minimal l^1-norm solution is also the sparsest solution. Commun. Pure Appl. Math. **59**(6), 797–829 (2006)
23. Figueiredo, M., Nowak, R., Wright, S.: Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE J. Sel. Topics Signal Process. **1**(4), 586–597 (2007)
24. Drori, I., Donoho, D.: Solution of l^1-minimization problems by LARS/Homotopy methods. ICASSP **3**, 636–639 (2006)
25. Hale, E., Yin, W., Zhang, Y.: A fixed-point continuation method for l^1-regularized minimization with applications to compressed sensing. Technical report CAAM Tech. Rep. TR07-07, Rice University (2007)
26. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cognit. Neurosci. **3**, 71–86 (1991)
27. Goel N., Bebis G., Nefian A.: Face recognition experiments with random projection. In: Proceedings of International Society for Optical Engineering Conference on Biometric Technology for Human Identification, pp. 426–437 (2005)
28. Mairal J., Bach F., Ponce J., Sapiro G.: Online dictionary learning for sparse coding. In: International Conference on Machine Learning, pp. 689–696 (2009)
29. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. **11**, 19–60 (2010)

Author Index