WUJNS
Wuhan University Journal of Natural Sciences

Vol.12 No.5 2007 912-916

Article ID: 1007-1202(2007)05-0912-05
DOI 10.1007/s11859-007-0040-x

A New Feature Selection Method for Text Clustering

□ **XU Junling[1], XU Baowen[1,2†], ZHANG Weifeng[3], CUI Zifeng[1], ZHANG Wei[1]**

1. School of Computer Science and Engineering, Southeast University, Nanjing 210096, Jiangsu, China;

2. State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, Hubei, China;

3. Department of Computer Science and Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, Jiangsu, China

Abstract: Feature selection methods have been successfully applied to text categorization but seldom applied to text clustering due to the unavailability of class label information. In this paper, a new feature selection method for text clustering based on expectation maximization and cluster validity is proposed. It uses supervised feature selection method on the intermediate clustering result which is generated during iterative clustering to do feature selection for text clustering; meanwhile, the Davies-Bouldin's index is used to evaluate the intermediate feature subsets indirectly. Then feature subsets are selected according to the curve of the Davies-Bouldin's index. Experiment is carried out on several popular datasets and the results show the advantages of the proposed method.

Key words: feature selection; text clustering; unsupervised learning; data preprocessing
CLC number: TP 393

Received date: 2007-02-18
Foundation item: Supported by the National Natural Science Foundation of China (60503020, 60373066), the Outstanding Young Scientist's Fund (60425206), the Natural Science Foundation of Jiangsu Province (BK2005060) and the Opening Foundation of Jiangsu Key Laboratory of Computer Information Processing Technology in Soochow University
Biography: XU Junling(1984-), male, Ph.D. candidate, research direction: statistical pattern recognition, machine learning and data mining. E-mail: junlingxu@gmail.com
† To whom correspondence should be addressed. E-mail: bwxu@seu.edu.cn

0 Introduction

With more and more high-dimensional data such as text data increasing, people need to partition the text collection into several parts according to some criteria such as the theme in order to obtain their expected data quickly. Clustering is a form of unsupervised learning which clusters similar objects together, so data belonging to one cluster are the most similar; and data belonging to different clusters are the most dissimilar. Majority work in text mining employs the "bag of words" approach and uses plain language words as features[1]. This representation of text data leads to an explosion in the number of features. Most clustering methods available can not handle text data well because of its high-dimension, so selecting a feature subset to represent the text and clustering on it is an effective method to solve the problem mentioned above. Motivated by the Iterative Feature selection (IF) method proposed in Ref. [2], in this paper, we propose a new feature selection method for text clustering based on expectation maximization and cluster validity. The major advantages of our method are:

① Feature subsets can be selected by user conveniently according to the performance curve;

② The complexity of the proposed feature selection method is very low and it has good performance in most cases;

③ The usability and effectivity of our method are better than IF method[2].

1 Related Works

Feature selection, an effective dimensionality re-

duction technique, is an essential preprocessing method to remove noisy features. Feature selection includes search or generation, and evaluation of subsets of features. Exhaustive methods guarantee optimality, but are impractical due to their exponential complexity in number of features. For supervised methods used in classification, the correlation of each feature with the class label is of importance. Dash *et al*[3] categorized the selection methods according to the generation procedures and evaluation functions. Koller *et al*[4] conducted further theoretical study on information theory based methods, and a framework for defining the theoretically optimal, but computationally intractable, method for feature subset selection was presented. Reviews for supervised selection can be found in Refs.[5,6].

As for unsupervised feature selection for clustering, Dash *et al* [7] proposed a filter method that is independent of any clustering algorithm, where feature importance is measured by its contribution to an entropy index based on data similarity. Martin *et al* [8] estimated how salient the individual features are. In addition, the number of clusters is estimated directly from the data. All these works proposed how to evaluate selection in clustering using the similarity between data objects. Thus, the problem of class label is avoided.

As to text processing, Yang *et al* [9] conducted a comparative study of supervised feature selection methods in statistical learning of text categorization, including document frequency (DF), information gain (IG), mutual information (MI), a χ^2-test (CHI), and term strength (TS). IG and CHI are found to be the most effective in their experiments. Although supervised feature selection methods have been successfully applied to text categorization, as text clustering is done on unsupervised data without class information, they could not work directly. Liu *et al*[2] demonstrated how to conduct supervised feature selection using the intermediate clustering result generated during iterative clustering to improve clustering performance. In their approach called IF, effective supervised feature selection is conducted in an iterative way during clustering. The essential of their method is an expectation maximization algorithm. We notice that in IF sometimes, an improper feature subset will be selected due to an impure cluster, and the performance will be degraded; and what is even worse is that IF discards some features in each iteration, so when the initial clustering is bad, the method would be ineffectual. Our method proposed in this paper is to overcome these disadvantages.

2 Feature Selection Method

In this section, we first explain the basic idea of our feature selection method. Then, our approach is compared with IF in Ref.[2].

2.1 Feature Selection Criteria

IG is widely employed for attributes merit criterion in machine learning. It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document[9]. Let $\{c_1, c_2, \cdots, c_m\}$ denote the set of categories in the target space, the information gain of term t is defined to be:

$$\begin{aligned} G(t) = & -\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i) \\ & + P_r(t)\sum_{i=1}^{m} P_r(c_i|t)\log P_r(c_i|t) \\ & + P_r(\overline{t})\sum_{i=1}^{m} P_r(c_i|\overline{t})\log P_r(c_i|\overline{t}) \end{aligned} \tag{1}$$

CHI measures the independence between feature and category[10]. Using the two way contingency table of a term t and a category c_i: where A is the number of times t and c_i co-occur, B is the number of time the t occurs without c_i, C is the number of times c_i occurs without t, D is the number of times neither c_i nor t occurs, and N is the total number of instances in the given dataset, the term goodness measure is defined to be:

$$\chi^2(t,c_i) = \frac{N\times(AD-BC)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{2}$$

$$\chi^2(t) = \max_{i=1,\cdots,m}\{\chi^2(t,c_i)\} \tag{3}$$

2.2 Index of Cluster Validity

DB index (Davies-Bouldin's index)[11] is a function of the ratio of the sum of within-cluster scatter to between-cluster separation, it uses both the clusters and their sample means. The value of this index will decrease as clusters become more compact and more distinctly separated. Therefore, when using this index to assess cluster validity, lower values are desirable.

The within ith cluster scatter and the between ith and jth cluster separation are defined as:

$$S_{i,q} = \left\{\frac{1}{|A_i|}\sum_{x\in A_i}\|x-v_i\|_2^q\right\}^{1/q} \tag{4}$$

$$d_{ij,t} = \left\{\sum_{s=1}^{p}|v_{is}-v_{js}|^t\right\}^{1/t} = \|v_i - v_j\|_t \tag{5}$$

where A_i is the set whose elements are the data points assigned to the ith cluster, v_i is the ith cluster center, $t \geqslant 1$, q is an integer and q,t can be selected independently of each other. $|A_i|$ is the number of elements in A_i. The DB index is defined as:

$$\mathrm{DB}(c) = \frac{1}{c}\sum_{i=1}^{c} R_{i,qt} \tag{6}$$

$$R_{i,qt} = \max_{j \in c, j \neq i}\left\{\frac{S_{i,q} + S_{j,q}}{d_{ij,t}}\right\} \tag{7}$$

2.3 Feature Selection Algorithm

Our method first does feature selection using DF criteria, and the feature subset selected by DF is used as the baseline of our experiment (start point). Then it repeats to cluster and do feature selection, and in each iteration it stores the selected feature subset and the DB index value of the new clustering which is clustered base on the selected feature subset, so the DB index value reflects the quality of the selected feature subset indirectly. Finally it draws the curve of the DB index value over iterating times. The detailed algorithm is shown as follows:

Algorithm Feature Selection:

1) Do feature selection using DF criteria;
2) For (i=0; i<N; i++) // N is given by user
{
3) Store the new selected feature subset $f_s[i]$;
4) Cluster in the new feature space;
5) Compute the DB index value db[i] for the new clustering;
6) Store db[i];
7) Do feature selection based on the new clustering using IG or CHI method;
}
8) Draw the curve of db over iterating times.

Actually, in order to evaluate the quality of the clustering, we also compute its error rate, F1-measure and entropy (in Section 3.2) of the clustering simultaneously with DB index in step 5).

Our feature selection method is also an iterative method based on expectation maximization like IF, but it overcomes the disadvantages in IF:

① It does not remove terms with lowest ranking scores outputted by IG or CHI at any iteration, so good features discarded before may be selected back in the subsequent iteration.

② It has a clear evaluation criterion (DB index value) for each feature subset, while IF can not decide which feature subset is good or bad if it does not remove terms.

3 Experiments and Analysis

3.1 Data Sets

As text clustering performance varies greatly on different datasets, we used three subsets of the two popular datasets (CT and 20 newsgroups dataset) in our experiment. CT is a subset of the four universities' datasets containing Web pages and hyperlink data. It was used for the cotraining experiments by Blum *et al*[12]. The 20 Newsgroups dataset is a collection of approximately 20 000 newsgroup articles, partitioned (nearly) evenly across 20 different newsgroups. Except few articles, all of them belong to one newsgroup precisely. The dataset has 20 sub-categories belonging to 4 categories: science, politics, sports and computer.

The following three subsets of the two datasets are used to measure the quality of generated clusters by comparing them with a set of categories created manually:

① Course2: A subset of the CT dataset, we removed hyperlink data from the original CT data set because of our assumption that only content is available;

② All4: A subset of 20 newsgroups dataset contains 4 000 articles which belong to one of the sub-categories of each category;

③ Sci4: A subset of 20 newsgroups dataset contains 4 000 articles which belong to the four sub-categories of category science.

3.2 Evaluation Criteria

Three kinds of measurements, error rate, entropy and F1-measure are used to evaluate the clustering performance.

Error rate is a commonly used method in statistics learning. F1-measure has been used to measure the quality of cluster when the cluster is compared with manually labeled class. Entropy measures the uniformity or purity of a cluster. The definition of these criteria can be found in Ref.[13], and they all need label information of the data.

3.3 Results and Discussion

K-means clustering algorithm was used to cluster datasets, and IG and CHI method were chosen in our feature selection experiment shown in Figs.1-6. Figures 1, 3, 5 display the error rate, F1-measure and entropy results on dataset Course2, All4 and Sci4 respectively.

Accordingly, Figures 2, 4, 6 display the DB index results on each dataset. We only show the results by using CHI as supervised feature selection method due to the limitations of page.

Our method selects features based on the DB index value, as can be seen from Fig.2, Fig.4 and Fig.6. Experiment shows that points (feature subsets at which iterating time) we should select have these characteristics:

① Its DB index value is a little smaller than the middle value, not too small.

② DB index value of points before it decreases slowly, and decreases sharply after its position (because of the instability of *K*-means algorithm).

③ DB index value of points before it decreases slowly, and increases slowly after its position.

Points 3 in Fig.1, 7 in Fig.3 and 13 in Fig.5 can be selected according to the criteria mentioned above and the curves of DB index in Fig.2, Fig.4 and Fig.6, and their quality is quite good as can be seen from Fig.1, Fig.3 and Fig.5 respectively. Our experiment also indicates that good feature subsets can be selected in about 20 times' iteration.

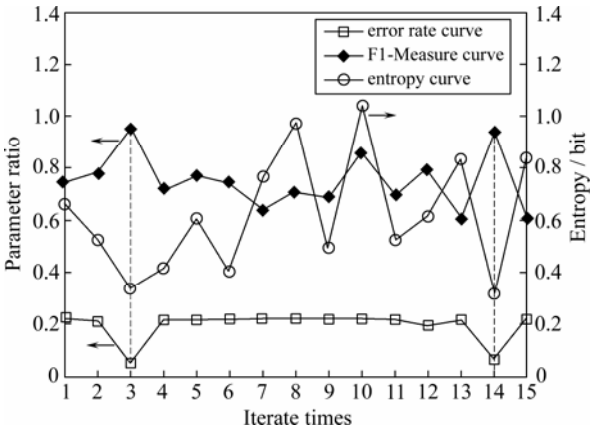

Fig.1 Error rate, F1-measure and entropy results on Course2

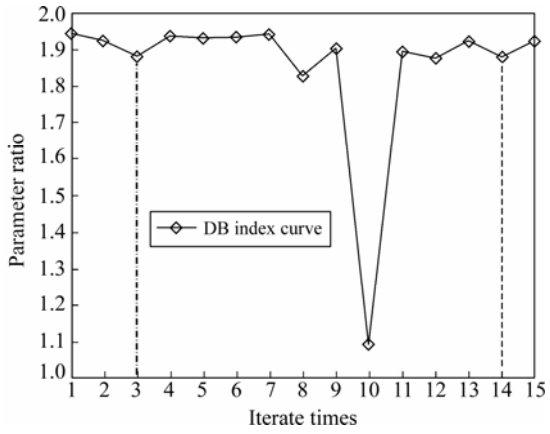

Fig.2 DB index result on Course2

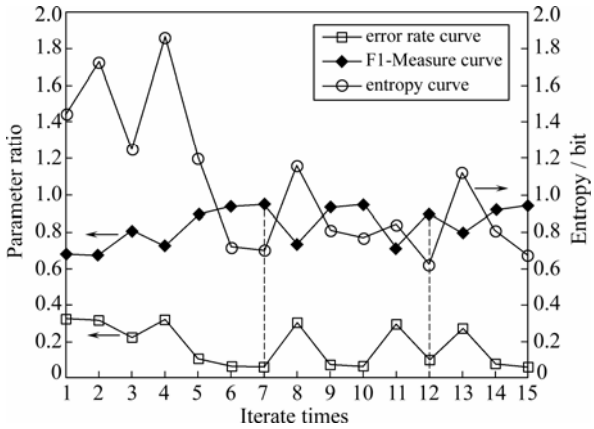

Fig.3 Error rate, F1-measure and entropy results on All4

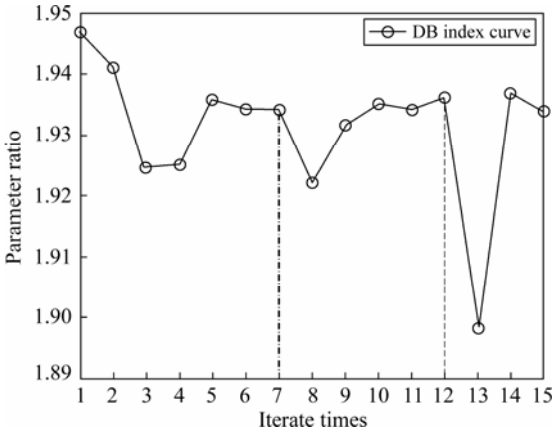

Fig.4 DB index result on All4

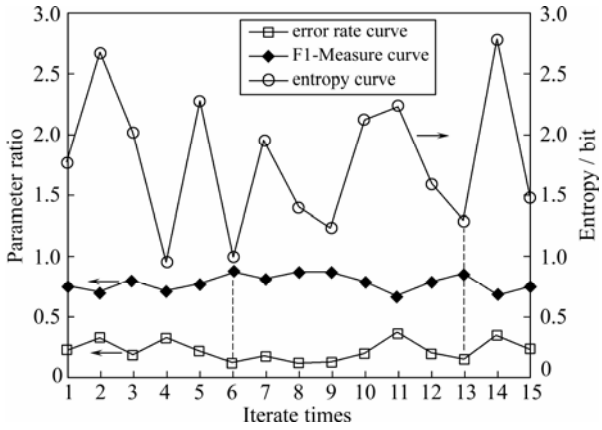

Fig.5 Error rate, F1-measure and entropy results on Sci4

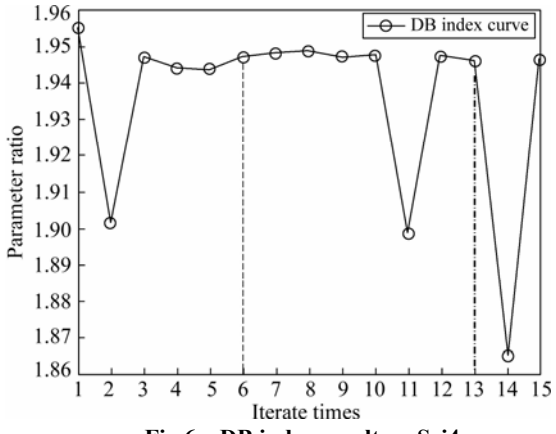

Fig.6 DB index result on Sci4

4 Conclusion

We have proposed a new feature selection method, which exploits cluster validity index as criteria to select feature subset, and good feature subsets can always be selected after a few iterations according to the criteria.

In the future, the proposed method will be tested on more data sets using some clustering algorithms other than hard *K*-means, e.g., soft clustering and density-based clustering.

References

[1] Sebastiani F. Machine Learning in Automated Text Categorization[J]. *ACM Computing Surveys*, 2002, **34**:41-47.

[2] Liu T, Liu S, Chen Z, *et al*. An Evaluation on Feature Selection for Text Clustering[C]//*Proceedings of the* 20*th International Conference on Machine Learning*. Washington D C: AAAI Press, 2003:488-495.

[3] Dash M, Liu H. Feature Selection for Classification[J]. *International Journal of Intelligent Data Analysis*, 1997, **1**(3): 131-156.

[4] Koller D, Sahami M. Toward Optimal Feature Selection [C]//*Proceedings of the* 13*th International Conference on Machine Learning*. Bari: Morgan Kaufmann, 1996:284-292.

[5] Blum A, Langley P. Selection of Relevant Features and Examples in Machine Learning[J]. *Artificial Intelligence*, 1997, **1**(2):245-271.

[6] Jain A, Duin P, Chang M. Statistical Pattern Recognition: A Review[J]. *IEEE Trans Pattern Analysis and Machine Intelligence*, 2000, **22**(1):4-37.

[7] Dash M, Liu H. Feature Selection for Clustering[C]// *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Kyoto: Springer, 2000:110-121.

[8] Martin H, Mario A, Jain A. Feature Saliency in Unsupervised Learning[R]. Michigan: Michigan State University,2002.

[9] Yang Y, Pedersen J. A Comparative Study on Feature Selection in Text Categorization[C]//*Proceedings of the* 4*th International Conference on Machine Learning*. Nashville: Morgan Kaufmann Press, 1997:412-420.

[10] Galavotti L, Sebastiani F, Simi M. Feature Selection and Negative Evidence in Automated Text Categorization[C]// *Proceedings of the* 6*th International Conference on Knowledge Discovery and Data Mining*. Boston: ACM Press, 2000.

[11] Davies D, Bouldin D. A Cluster Separation Measure[J]. *IEEE Trans Pattern Analysis and Machine Intelligence*, 1979, **1**:224-227.

[12] Blum A, Mitchell T. Combining Labeled and Unlabeled Data with Co-Training[C]//*Proceedings of the* 11*th Annual Conference on Computational Learning Theory*. Madison: ACM Press, 1998:92-100.

[13] Huang S, Chen Z, Yu Y, *et al*. Multitype Features Coselection for Web Document Clustering[J]. *IEEE Trans Knowledge and Data Engineering*, 2006, **18**(4):448-459.

□