2007－MPS－65 (6)
2007／6／25

Construction of a Paper Classification System Using SVM

Naomi Ashida* Masami Takata* Akira Sasaki‡ Hiroyasu Kamo† Naoyuki Nide† Kazuki Joe*
ashida@ics.nara-wu.ac.jp
* Graduate School of Human Culture, Nara Women's University
† Department of Computer and Information Science, Nara Women's University
‡ Japan Atomic Energy Research Institute

Abstract

A huge number of of journal papers on the web make it time-consuming for researchers to find the papers about particular topics they are interested in. Therefore, the searching process is highly expected to be automated. Conventional text classification methods are applicable to the automatic searching if the main texts of the papers are provided. However, we can not obtain the main texts but only the abstracts of the on-line papers freely. In this paper, we propose a paper classification technique using SVM for on-line physics papers . In this paper, we show the efficiency of our proposing technique.

SVMを用いた論文分類システムの構築

芦田 尚美* 高田 雅美* 佐々木 明‡ 鴨 浩靖† 新出 尚之† 城 和貴*
* 奈良女子大学大学院 人間文化研究科
† 奈良女子大学 理学部 情報科学科
‡ 日本原子力研究開発機構

概要

本論文は，SVMを用いた論文分類システムの構築を目的とする．Web上には膨大な数の論文が存在し，研究者は自分が必要とするごく少数の論文をその中から探し出す必要がある．また，通常Webから無償で入手することが出来るのはアブストラクト部分のみであるので，必要な論文であるか否かを本文を読まずに判断しなければならない．そこで，本論文では論文のアブストラクト部分のみを用いて論文を自動分類するシステムの構築を試みる．本論文では，分類器としてSVMを用いた論文分類システムを構築し，先行研究との比較を行い，提案手法の有効性を示す．

1 Introduction

Nowadays, it becomes common that journal papers are obtained from the web. The number of on-line papers is too huge to check all papers and classify them according to their topics by hand. Besides the huge number of on-line papers, there is a cost problem to download the main text of the papers while their abstracts are usually download free.

We have studied on-line paper classification by abstract from the web using LVQ [1]. In this paper, to improve the accuracy of our previous method, we propose an on-line paper classification using SVM[2]. We adopt PCA[3] and k-means clustering [4] as preprocessing of data, which is generated from abstracts.

In the rest of the paper, section 2 gives the motivation of this paper comparing with previous works. In section 2.1, we describe about the feature vectors extracted from abstracts of on-line papers. In section 3.3, experimental results are given to validate the proposed method.

2 Classification of On-line Papers

Conventional text classification methods require the main text of on-line papers if they are adopted in a straightforward way. Because of the use of just abstracts as described in [1], those methods will perform ineffective classification. We have proposed [1] to solve those problem. [1] use Learning Vector Quantization(LVQ) as a classifier.

In this paper, SVM is adopted as a classifier instead of LVQ. The target of SVM based on-line paper classification is the same as [1].

2.1 Feature Vectors

We use the same feature vectors which we have developed [1] for the classification of 16,070 on-line papers with atomic and molecular data. We define two categories for classification. These are called category-1 and category-0. Those definitions are given by [1]. Each sample is composed of a feature vector generated from an abstract, and consists of 3,557 dimensions.

2.2 Support Vector Machines

In our new method, we adapt Support Vector Machines (SVM) [2] as a classifier instead of LVQ in the previous method [1]. Since SVM has a simple mechanism and a good learning ability . In our SVM based method, we use LIBSVM [5] since it has detailed documents about implementation.

3 Classification using SVM

3.1 Reducing dimension

As described in the previous section, feature vectors consist of 3,557 dimensions generated.When an abstract of a paper contains an index word of the dictionary, the element of the feature vector corresponding to the word has positive value. Each abstract contains only a few index words at most. In other words, each feature vector of samples is a sparse form, therefore it has a lot of zero value.

In general, the smaller value of an element of feature vectors is presented, the less correlation between the element and the other elements. Therefore, some threshold value of elements should be used for the better classification. In this paper, we remove the elements where the number of positive element values is less than a threshold value to reduce the dimension size of feature vectors.

Note that we have 16,070 samples with 3,557 dimensions, so the above reduction of dimension size is not enough. Although taking a larger threshold value may be a way to reduce the size, we perform another approach for the dimension reduction in this paper. We apply PCA [3] to the dimension reduced samples .

3.2 Clustering using the K-means algorithm

As described in the previous section, we use a sample set consisting of category-0 with 15,944 samples and category-1 with 126 samples. This is a really lopsided data set for classification.

To balance the lopsided data set, we need to further categorize category-0 samples so that the resultant sub-categories of category-0 are balanced with category-1. The balanced categorization can be performed as follows. We explore various categorizations of all samples so that a resultant category contains the samples of the category-1 as many as possible and that the number of resultant categories is as large as possible. Because the two constraints are in a trade-off relation, we need to define the balanced categorization: Provided that at least 90 percent of category-1 samples are included in a resultant category, the number of resultant categories is the maximum. We call such a categorization a balanced one.

We adopt the k-means algorithm [4] for clustering all the samples. Using the k-means algorithm, we find the optimal number for k, which is the number of balanced categories for all the samples.

3.3 Application of SVM

According to the method described in subsection 3.1, the elements of feature vectors which have low appearance rates are removed. We count the number of positive values for each element of feature vectors of all samples. If the number is less than n, the corresponding element is removed such that n is enough small compared with the number of all samples. After the first reduction of the feature vector dimensions, we apply PCA to the feature vectors aiming at the further reduction of dimensions, and calculate the principal component of all the samples. Selecting the dimensions according to an accumulative contribution ratio over the particular rate, the data is converted into z-score [6].

The reduced data sets are then clustered using the k-means algorithm. The purpose of the clustering is to extract a data range containing as many category-1 samples as possible. Varying the value of k, we perform the clustering several times to obtain the best result (the optimal value of k), where a cluster contains most of category-1 samples and each cluster size is as balanced as possible.

Using the optimal value of k, the training samples are clustered using the k-meaning algorithm so that a single cluster contains as many category-1 samples

表 1: Experimental environment

CPU	Intel Pentium4 3.06 GHz
RAM	512 MB
OS	Vine Linux3.2
SVM library	LIB-SVM[5]

表 2: Clustering number and size

k	Number of clustering succeed	Size of cluster containing category-1
4	10 / 10	5135.4
5	6 / 10	5006.3
6	5 / 10	4755.4

of the training set as possible. We label the training samples in the single cluster, where most of category-1 samples are included, as "1", while the other training samples are labeled as "-1". Using the labels, an SVM learns the training samples.

We use C-SVM and the kernel function we use is the Gaussian radial basis function (RBF) [5]. Since the number of category-1 samples is extremely small compared with category-0 samples, we need to weight the category-1 samples so that each category-1 sample plays more important role in the SVM learning. The weighting learning is possible by changing the parameters C and γ as described in [5]. Varying the values of C and γ for the SVM learning, we investigate the relation between those parameter choices and the recognition, recall and precision rate for test samples.

4 Experiments

Experiments are performed with various parameters of C and γ. Table 1 summarize the experimental environment.

The data set is composed of 16,070 feature vectors of which dimension size is 3,557 generated by the method described in section 2.1. We select a half of all samples as the training data set, use the rest as the test data set. The method to select training data and test data is the same as in [1]. The number of samples in category-0 and category-1 is 15,944 and 126, respectively. Category-0 and category-1 in training data consist of 7,972 and 63 samples, respectively.

4.1 Evaluation measurement

In this experiment, we use three evaluation measures: the recognition, recall and precision rates. Those are defined in [1]. A higher recall rate means that the SVM recognizes category-1 samples without missing them. A higher precision rate means that the SVM does not inexactly recognize a category-0 sample as a category-1 sample. In this paper, the recall rate is emphasized since recognizing category-1 samples without any miss is our final goal.

4.2 Perform experiments

We assume that the elements of feature vectors with low appearance rates should not be used as classification indicators . Thus, the dimension of feature vectors is reduced to 1,667 from 3,557 by deleting the feature vector elements where the number of positive value in the element is less than five.

As the next step, PCA is applied to generate a smaller data set from the above feature vectors, and accumulative contribution is calculated. The new feature vectors with 754 dimensions are regenerated by using the top 754th elements of the principal components where the accumulative contribution percentage is over 90 percents.

Finally, the K-means clustering is applied to the restructured data set. Using $k = 4, 5$ and 6, the k-means algorithm is performed ten times. Table 2 shows the number of succeeding clustering out of ten and the average size of the cluster which contains category-1 samples. Table 2 shows $k = 4$ is the best. Consequently, $k = 4$ is applied to training data set as cluster centers.

The resultant cluster which contains most of category-1 samples is assumed as label "1" while the other clusters are assumed as label "-1" for SVM. In this way, the training data with labels are generated for SVM learning.

As described in 3.3, we need to find the best C and γ for better classification. By using a grid searching method, the best C and γ values are explored. SVM learning is performed with $2^{-6} \leq C \leq 2^{14}$ and $2^{-8} \leq \gamma \leq 2^{1}$. The grid step size is $2^{0.2}$ for both C and γ. Figure 1 and 2 show the recall and precision rate, respectively. When $C = 0.25$ and $\gamma = 0.00592$, the recognition rate, the recall rate and the precision rate are obtained 86.24%, 88.33% and 12.33%, respectively.

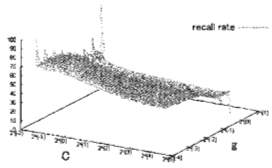

図 1: Recall rate

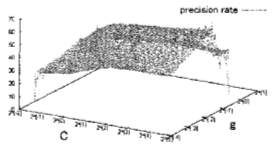

図 2: Precision rate

4.3 Discussions

As table 2 show, it becomes difficult to obtain a single cluster which contains most of category-1 samples as the value of k increases. However, the size of the cluster containing most of category-1 samples becomes larger as k gets smaller. Figure 3 shows the recall and the precision rates. As described in the figure, there is a trade-off between the recall and the precision rates. Namely, as one of the rates gets higher values, the other becomes lower. Configuring C and γ, we have obtained a more efficient result than that in the previous work [1]. Therefore, we confirm that LVQ is not only a method possible to classify on-line papers using their abstracts but also SVM.

5 Conclusions

In this paper, we adopted SVM for the classification of on-line papers by using their abstracts. We deleted some elements of feature vectors with low appearance rates, and applied PCA for the reduction of the dimension size. Then, we clustered them using the k-means algorithm for a more balanced data set.

Most of category-1 samples are put into a single cluster by the k-means algorithm. Although the resultant cluster also contains category-0 samples of which amount is about dozen of times of category-1 samples, they are enough to make SVM learn with some parameter choices.

For the evaluation measure of SVM, we adopted three grading scales called recognition, recall and precision rates. Recall rate is the most important out of three. In experimental results, we obtained better recall rates than our previous work [1]. Additionally we obtained the same precision rate as our previous work.

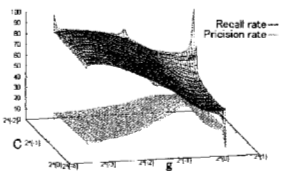

図 3: The relation between the recall and the precision rates

参考文献

[1] Kashiwagi, Watanabe, Sasaki and Joe. *Text Classification for Constructing an Atomic and Molecular Journal Database by LVQ.* The 2005 International Conference on Parallel and Distributed Processing Techniques and Applications, Vol.1, pp.481–487, 2002.

[2] Nello Cristianini, John Shawe-Taylor. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000.

[3] Lindsay I Smith. A Tutrial on Principal Components Analysis. http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf, Maintained by Cornell University, USA. 2002.

[4] J. B. MacQueen. Some Methods for classification and Analysis of Multivariate Observations *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297, 1967.

[5] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM : a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[6] Hervé Abdi. Z-scores. http://www.utdallas.edu/~herve/Abdi-Zscore2007-pretty.pdf