Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(6):210-217

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

An improved incremental learning algorithm for text categorization using support vector machine

Cao Jianfang[1,2] and Wang Hongbin[1]

[1]*Department of Computer Science & Technology, Xinzhou Teachers' University, Xinzhou City, China*
[2]*College of Computer Science & Technology, Taiyuan University of Technology, Taiyuan City, China*

ABSTRACT

Text categorization which assigns natural language texts to one or more predefined categories based on their content is an important component in many information organization and management tasks. Different automatic learning algorithms for text categorization have different classification accuracy. SVM classification model is common powerful for text categorization task. It is based on probability and is of religious theoretic basis. In this paper the SVM categorization model is analyzed and an algorithm to perform text categorization using incremental model is presented. Compared with the Bayes learning method and the K-nearest neighbor method experimental results verify the effectiveness of the proposed algorithm. Experiments show that the incremental model dramatically reduces the training time and is a better classification algorithm.

Keywords: text categorization support vector machine feature extraction incremental learning algorithm

---

INTRODUCTION

Under the current situation of the rapid development of computer network how to deal with the massive information is one of the problems that must be solved currently. Text categorization—assignment of natural language texts to one or more predefined categories based on their content—is an important component in many information organization and management tasks [1]. So far it has been widely used in support of text acquisition positioning and filtering text categorization. Also it plays an important role in much more flexible dynamic and personalized information management tasks. However with the increase of obtained information especially online data of enterprises which have very strong real time new data will be continued to add to the database which requires we can retain historical analysis of the existing data when analyzing these new data.

Traditional text classification methods [2-4] put the update data and the previous data together whether it has been learned before to train and then get the new classification criteria. This approach is referred to batch classification method which is equivalent to forget all previous learning results and is very wasteful in space and time undoubtedly. Incremental learning method is the more meaningful ways to learn constantly update data namely it only learn new data on the basis of retaining previous learning results to form a continuous learning process. Using support vector machine (SVM) classification algorithm can well realize the incremental learning for multi-classification problems.

Some incremental learning methods based on SVM have been put forward. Literature [5] emphasized the role of the support vector--keep the support vector with concentrated old data join the new data to train obtain new support vectors and classification function. Literature [6] used the local incremental learning method. When the data cannot be classified correctly it puts the data around the new data into the data set and amends its original classification criteria. Vote method is used by Erdem et al. to determine the classification of test data. Firstly each batch of data is learned alone and classifiers are got respectively. Secondly the test data will be generated into each classifier to get

its classification marks and the mark signed by most classifiers is its classification. Although research on SVM is in full swing and some significant results have been achieved in recent years the research on incremental learning method in the field is still in its infancy and there are still many problems which need to be researched. In this paper an improved SVM incremental learning algorithm is proposed and applied to text classification. Compared with the Bayes learning method and the K-nearest neighbor method experimental results show the effectiveness of the proposed algorithm. It can greatly reduce the steps and time needed by classifier when the new class increases. Moreover it has advantage over the data size and expansion. SVM incremental learning method is a better classification algorithm.

II. THEORY OF SUPPORT VECTOR MACHINE AND CLASSIFIER

A. Support Vector Machine

The algorithm of support vector machine comes from statistical learning theory. The algorithm is based on structure risk minimization principle [7-11]. The original data set is compressed into the support vector set (typically is 3%—5% of the former) and the classification decision function is got by learning. The basic idea is to construct a hyperplane as the decision plane so as to obtain the maximum interval between the positive and negative mode.

The SVM method is put forward from the optimal classification plane in linear separable cases. As shown in Fig. 1 hollow circles and hollow squares represent the two types of training samples respectively. H is classification line which separates the two types correctly. H1 and H2 are lines which pass through the points that are the nearest to various types of samples and parallel to classification line. The distance between two lines is called classification interval. According to the principle of empirical risk minimization theory actual risk of SVM is decided by formula (1).

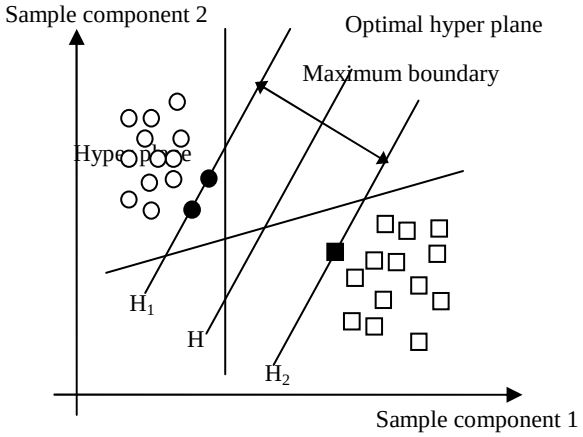

Figure 1. Optimal classification face.

$$R(\omega) \le R_{emp}(\omega) + \Phi \qquad (1)$$

Where $R(\omega)$ is actual risk $R_{emp}(\omega)$ is the empirical risk Φ is confidence interval. Completely separation makes $R_{emp}(\omega) = 0$ and maximum interval makes the minimum range of confidence interval Φ so that the real risk is minimized.

Supposing that linearly separable sample set is $(x_i, y_i), i = 1,...,n, x \in R^n, y \in \{+1,-1\}$. The common form of linear discriminant function in n-dimensional space is $g(x) = \omega \cdot x + b$. The formula of classified surface is:

$$g(x) = \omega \cdot x + b = 0 \qquad (2)$$

Take the discriminant function $g(x)$ normalized and make all kinds of samples meet with $|g(x)| \ge 1$. Then the classification interval is equal to $2/\|\omega\|$. The problem is changed into keeping the largest interval according to the condition that classification line may correctly classify all samples. It is symbolically defined as:

$$\begin{cases} \min\ f(x) = 2/\|\omega\| & (a) \\ s.t.\ y_i[(\omega \cdot x_i) + b] - 1 \geq 0, i = 1, 2, \cdots, n & (b) \end{cases} \quad (3)$$

In the case of non-linear some training samples don't meet the condition of formula (3b). We need add a relaxation item $\varepsilon_i \geq 0$ in the left of the condition formula (3b). Minimize formula (3a) is equal to maximize $\Phi(\omega) = \|\omega\|^2 / 2$. In accordance with the Lagrange function and Kuhn-Tucker conditions formula (3) finally can be converted to:

$$\begin{cases} \max \Phi(\omega, \varepsilon) = \frac{1}{2}(\omega^T \cdot \omega) + Q(\sum_{i=1}^{n} \varepsilon_i) & (a) \\ s.t.\ \varepsilon_i \geq 0, \forall i & (b) \\ y_i(\omega^T x_i + b) \geq 1 - \varepsilon_i, \forall i & (c) \end{cases} \quad (4)$$

The optimal classification function is got by solving formula (4) and shown as formula (5).

$$f(x) = \text{sgn}\{\omega^T \cdot x + b\} \quad (5)$$

We can get $\varepsilon_i \geq 1 - y_i(\omega^T x_i + b)$ by the formula (4b) so the formula (4a) can be redefined as:

$$\Phi(\omega, \varepsilon) = \frac{1}{2}\omega^T \omega + Q\sum_{i=1}^{n} \left|1 - y_i(\omega^T x_i + b)\right|_+ \quad (6)$$

$$|z|_+ = \begin{cases} 0, & if\ |z| \leq 0 \\ z, & other \end{cases}$$

The constant *Q* in formula (4a) is equilibrium between the generalization ability and training accuracy. If *Q* is smaller SVM has better generalization ability; If *Q* is larger SVM has smaller training error. Formula (4b) introduces the slack variable which allows some points overstep boundary and increases SVM ability of noise immunity in case of non-separable. Since introduction of slack variables we define boundary relative to formula (5) for each sample:

$$\gamma_i = y_i f(x) \quad (7)$$

Giving a training set $s = \{x_i, y_i\}_{i=1}^{N}$ formula (6) and (7) show that the aim of SVM learning algorithm is to find the function $f(x)$ so as to get the max boundary and $\max \sum_{i=1}^{N} \gamma_i$ of $f(x)$ relative to training set.

In the linearly non-separable cases the sample x is mapped to high dimensional feature space H which is mapped to the linearly separable case and uses a linear classifier in H. Thus only the inner product is calculated in the high-dimensional space and the inner product is realized using the function of the original space even if we don't know the form of transformation. According to the theory of functional--as long as a kernel function $K(x_i \cdot x)$ meets Mercer conditions it corresponds to an inner product in a certain space. Common forms of kernel functions are:

polynomial kernel function $K(x, y) = [(x \cdot y) + s]^d$;

the radial basis function $K(x, y) = \exp(-\sigma\|x - y\|^2)$;

two layer perceptron neural network Sigmoid function

$K(x, y) = \tanh(k(x \cdot y) - \mu)$.

Different kernel functions will result in different feature spaces; therefore there will be the different sample distributions. In order to limit the sample into the big feature space we choose a radial basis function as category kernel function.

B. Classifier

A classifier is a function which describes the mapping from the input characteristics $\overline{x} = (X_1, X_2, \ldots, X_n)$ to set $f(\overline{x}) = confidence(class)$ relying on the input. In the paper the propertie of the document is the words and its type is the text category. For N categories we can construct a collection of binary SVM classifier $C_N = \{S_2, S_3, \cdots, S_N\}$ which contains *N-1* classifiers but S_i is a binary classifier being used to distinguish collection $\{i\}$ and $\{1,2,\cdots,i-1\}$ .That is to say for the current classification system only a binary SVM classifier need to be constructed to distinguish its instance of and all the old instances when adding a new category.

III. SVM INCREMENTAL LEARNING ALGORITHM FOR TEXT CLASSIFICATION

A. Text Classification and Multi-class SVM

Automatic text classification technology is an important foundation of information retrieval and data mining. The main task is to learn the calibration sample feature in the given category tag set and to determine its category according to the semantic content of the text. In the process of text classification after feature selection for learning samples which have been pretreated (such as word segmentation punctuation etc.) each text d is represented as a vector x in the n-dimensional feature vector space and regard as input of machine learning algorithm. Currently there are many machine learning algorithms for text classification such as statistical learning method k- nearest neighbor method and SVM algorithm etc. Among them SVM algorithm reflects the performance level of the current text classification method. Moreover incremental update result in the SVM classifiers can not only add new samples or categories of knowledge to the classification model but also do not affect the performance of classifier and keep the number of categories.

B. Text Representation and Feature Selection

At present in the field of information processing text representation mainly uses the vector space model (VSM) [9]. The basic idea of vector space model is using vector $(w_1, w_2, \cdots, w_n)$ to represent text. Where w_i is the weight of the ith feature item. The feature items generally are characters words or phrase. According to the experiments generally words as the feature are better than characters and phrases. Therefore if we want to make the text be represented as a vector in the vector space we first make word segmentation for text and use segmented words as the dimension of vectors to represent text. Initially vector representation is completely 0 1 form. That is to say if the word appears in the text the dimension of text vector is 1 otherwise is 0. However this method does not reflect the degree of the role of words in the text so 0 1 are gradually replaced by more accurate word frequency. Word frequency is divided into absolute frequency and relative frequency. Absolute frequency uses frequency of words appeared in the text to represent text; Relative frequency is normalized frequency which is calculated mainly using TF-IDF formula:

$$W(t,\vec{d}) = \frac{tf(t,\vec{d}) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t \in \vec{d}} [tf(t,\vec{d}) \times \log(N/n_t + 0.01)]^2}}$$

Among them $W(t,\vec{d})$ is weight of word t in the text $\vec{d}$ and $tf(t,\vec{d})$ is word frequency of the word t in the text $\vec{d}$ N is the total number of training texts n_t is the number of text appeared t in training text sets the denominator of which is the normalization factor.

In order to improve the efficiency text feature selection should be removed some items according to the occurrence of characteristics and then select some features corresponding to classifiers. By comparing the feature selection methods we finally choose to use the mutual information entropy. The mutual information of feature X_i and classifier C is defined as:

$$MI(x_i,c) = \sum_{x_i \subseteq \{0,1\}} \sum_{c \subseteq \{0,1\}} p(x_i,c) \log \frac{p(x_i,c)}{p(x_i)p(c)}$$

Select K feature items with largest mutual information in the classifiers as input of the learning method. Let K =300 for the SVM learning algorithm according to the experimens which has not strict proof only because training effect is good when K =300.

C. SVM Incremental Learning Algirithm

Supposing that the classifier of current $N-1$ category is C_{N-1}. When considering a new category N we only need train a binary sub-classifier S_N to distinguish between instances of N (positive cases) of all instances of the previous $N-1$ categories (regarding $N-1$ categories as a superclass SC_{N-1} negative cases) and get a new classifier $C_N = C_{N-1} \cup S_N$. In the processing of test C_N first gives instances to S_N and if S_N can accept it it is part of category N ; If S_N rejectes it the instances will be input the classifier C_{N-1} in order to obtain the decisions belonging to the previous

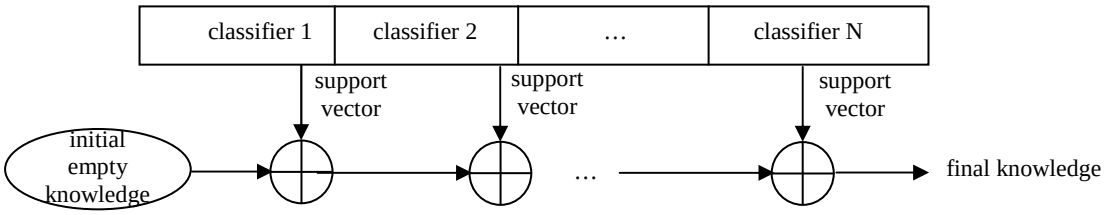

Figure 2. The process of incremental training.

When new categories are added we must replace the feature set which results in change of feature space. So we need adjust the method of feature selection. Supposing that the feature set of classifier C_{N-1} is F_{N-1} the feature space of classifier C_{N-1} is V_{N-1}. Before the training process begins we make local feature selection for the category N and superclass SC_{N-1} then get feature set the union of which and F_{N-1} is called F_N. Therefore $F_{N-1} \subset F_N$ and feature space V_{N-1} of classifier C_{N-1} is feature space V_N 's subspace of classifier C_N. So we train sub-classifiers S_N of two values in space V_N. The projection $x|V_N$ of the vector space V_N is input into the sub-classifier S_N for an instance x when testing. If it is refused and applied to classifier C_{N-1} $x|V_{N-1}$ will be taken as input. The learning procedure is as follows:

(1) Sample pretreatment: do original text segmentation remove punctuation and extract feature of text.
(2) Normalization process: according to the new attribute set and corresponding to the original data form a new training set and testing set and do normalization.
(3) Establish training model: use radial basis kernel function to get the training model for training samples.
(4) Classify: classify test samples using training model and output the results and classification accuracy.
The classification learning algorithm is as follows:

Input: feature set F_{N-1} the instances of *N* the instances of the superclass SC_{N-1} classifier C_{N-1}.

Output: the feature set F_N classifier C_N.

(1) $F_N = F_{N-1} \cup local\_feature\_selection(N, SC_{N-1})$.

(2) $S_N = SVM\_train(N, SC_{N-1}, F_N)$.

$N-1$ categories. When the classifier C_N of the previous N categories already exists the arrival of a new category $N+1$ can get the classifier C_{N+1}. Method of this paper is mainly based on ideas of incremental learning put forward by Drucker et al. which keeps support vectors of the old data set and join them into the new data set to train. The basic idea is shown in Fig. 2.

(3) $C_N = classifiers(C_{N-1}, S_N)$.

The constructed classifier C_N contains a binary classifier S_N and low-level classifier C_{N-1}. For each test instance x C_N processes its projection $x|V_N$ on V_N recursively. Only when $x|V_N$ does not belong to the current class will it be input to the lower classifier C_{N-1} and regard the results of C_{N-1} as its own decision. As a result if classifier error in a more high-level is the dominant error in the decision process the performance of the classifier will be largely influenced.

IV. BAYES LEARNING METHOD AND K-NEAREST NEIGHBOR METHOD

A. Bayes Learning Method

Simple Bayes classifier [12-15] assumes that each word is conditional independence and all words node only have one parent node. Its principle is: Determine the class of each text in the training set and then calculate the probability estimation of word in training text. So the parameters of classifiers are composed of priori probability value and conditional probability value based on the class. Strictly speaking each class C_j has a document frequency $P(C_j)$ relative to all other classes. For each word W_t in the vocabulary V $P(W_t|C_j)$ indicates the frequency of occurrence that the expected word W_t of the classifier in the file of the class C_j. For standard tutor learning Bayes classifier classifier parameter is determined by the labeled training documents. We can draw a conclusion that Bayes learning method uses prior probability to judge assistantly so as to get the more accurate results. It is best in the sense of minimizing error probability and risk. But Bayes classifier needs know the conditional probability and its decision surface is often a hyper surface the shape of which is very complex and it is difficult to calculate and construct.

B. K-nearest Neighbor Method

The most simple and intuitive method in pattern classification field is classification method based on the distance function [12]. The core idea is to use the center of gravity of a class to represent the class. Then calculate the distance between the samples to be classified and the center of gravity. Finally put the samples to be classified into the class with the nearest distance. The Mahalanobis distance is often used in discriminant analysis which not only considers the mean of class but also contains the variance information of class. The using of information is full.The basic assumption of the Mahalanobis distance is all kinds are normal population. If all the sample point in a class can qualify as a representative of the class this is the nearest neighbor method. Nearest neighbor method not only compares the mean distance of various types but also calculates the distance between all sample points. As long as the distance is nearest it belongs to the class. The distance of the samples uses Euclidean distance. The Euclidean distance between two points $X=(x_1,x_2,\cdots,x_n)$ and $Y=(y_1,y_2,\cdots,y_n)$ is $d(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$.

V. EXPERIMENTS AND PERFORMANCE EVALUATION

Text classification is a mapping process in essence. The mark of evaluating text classification system is accuracy and speed of mapping in the condition of the continuous increase of data. Mapping speed depends on the complexity of the mapping rules. And the reference of evaluating mapping accuracy is the classification results of text by experts' judgment. The more similar with artificial classification results it is the higher accuracy of classification is. it implied the two indexes of evaluating the text classification system: accuracy and recall. This paper selects 2680 Chinese text corpus from the Internet using different classifiers and different amount of data to test and analyze efficiency and results of text classification SVM incremental learning algorithm. The corpus contain different types of documents and are classified by experts in the fields into 36 categories-- political military television etc. For experimental convenience this paper only divides corpus into 15 categories roughly. The method of selecting training sets and testing sets is as follows: randomly select 20% of data from corpus as open test set the remaining part as training set and closed test set. Repeat the experiment 15 times operate classification algorithm calculate the average value. This paper firstly tests the running time and error rate of SVM incremental learning algorithm in the condition of the continuous increase of data. The experimental results are shown in Table I.

As can be seen from the test results with the increasing of test data the classification computation time does not increase linearly and error rate is very low when using SVM incremental learning algorithm proposed in the paper.

Then compare different classification algorithms in order to verify the validity of SVM incremental learning algorithm in the closed and open test. The experimental results are shown in Table II.

TABLE I. DATA SET OF TESTING

Dataset	The amount of data	Computing time(s)	Error rates
Initial data set	800	94	4.0%
Data subset 1	130	5.0	3.2%
Data subset 2	200	5.8	2.8%
Data subset 3	280	6.3	3.0%
Data subset 4	350	6.8	2.9%
Data subset 5	410	7.0	3.1%

TABLE II. CLASSIFICATION ACCURACY OF DIFFERENT ALGORITHMS

Algorithm	Recall of closed test	Accuracy of closed test	Recall of open test	Accuracy of open test
SVM	89.6%	91.1%	82.4%	82.6%
Bayes	84.3%	85.8%	78.5%	76.7%
K-nearest neighbor	87.3%	89.3%	80.1%	80.4%

As can be seen from the results of the above closed and open test compared with Bayes and K-nearest neighbor method we can get higher recall and accuracy using SVM incremental learning algorithm which is a kind of effective method.

Through experimental tests we can draw a conclusion that the advantages of the proposed algorithm is mainly reflected in the following aspects: (1) Adapt to large-scale data: With linear growth of classification data classification time and error rate is not linear growed; (2) Easy to expand: When adding new categories the original classification system established is not destroyed and we only need make some relative calculations for the new classification which makes classification easy to operate no duplication of work have some expansion capability .

CONCLUSION

This paper proposes SVM incremental learning algorithm on the basis of analyzing SVM classifier in-depth. And we make comparasion with Bayes learning method and K-nearest neighbor method realize text classification in an environment close to the real world. The above works are all verified by experiments and the results show that the proposed algirithm is very effective.

At present we have used the algorithm to complete a text classification system which can be used to classify text and Webpage text. The system has been tested using a large amount of data and obtained the quite ideal effect. In the future we will continue to improve the system tries to combine with other effective text classification methods and further improve the classification accuracy of the system.

Acknowledgement
This work was supported by National Natural Science Foundation of China under Grant No. 61202163 and by the Natural Science Foundation of Shanxi Province under Grant No. 2013011017-2 and by the Technology Innovation Project of Shanxi Province under Grant No. 2013150 and by the Key disciplines supported by Xinzhou Teachers University under Grant No. XK201308. The authors are grateful for the constructive and valuable comments made by the many expert reviewers.

REFERENCES

[1] Jyothi B. S. and Dharanipragada J. *Peer-to-Peer Networking and Applications* vol. 4 **2011** pp. 289-308.
[2] Guan R. C.Marchese X. H.Yang M.et al. *IEEE Trans on Knowledge and Data Engineering* vol. 23 2**0**11 pp. 627-637.
[3] Shamir O. and Tishby N. *Machine Learning* vol. 80 **2010pp**. 213-243.
[4] Jin R. M. Goswami A. and Agrawal G. *Knowledge and Information Systems* vol. 10 **2006pp**. 17-40.
[5] Tsang I. W.Kwok J. T.and Li S. "Learning The Kernel in Mahalanobis One-class Support Vector Machines" In Proceedings of *the* **2006** *International Joint Conferenceon Neural Networks (IJCNN* **2006***) IEEE* 2006pp. 2148-2154.
[6] Maggi F. Matteucci M.and Zanero S.Detecting Intrusions through System Cal- Sequence and Argument Analysis IEEE Transactions on Dependable and Secure Computing 7 (4)**2010** pp. 381-395.
[7] Francois J. Abdelnur H. State R. et al. *IEEE Transactions on Network and Service Management* vol. 7 **2010** pp. 244-257.
[8] Joseph J. F. C.Lee B.-S. Das A.et al. *IEEE Transactions on Dependable and Secure Computing* vol. 8 **2011** pp. 233-245.
[9] Yi Y. Wu J. and Xu W*IEEE Transactions on Network and Service Management* vol. 38 **2011** pp. 7698-7707.

[10] Hofmann A. and Sick B. *IEEE Transactions on Dependable and Secure Computing* vol. 8 **2011** pp. 282-294.
[11] Psorakis I. Damoulas T. and Girolami M. A. *IEEE Transactions on Neural Networks* vol. 21 **2010** pp. 1588-1598.
[12] Hsieh T. W. and Taur J. S. *Journal of Signal Processing Systems* vol. 60 **2009** pp. 105-114.
[13] Zhang T. Ramakrishnan R. and Livny M. "BIRCH: An Efficient Data Clustering Method for Very Large Databases" *In Jagadish H. V. and Mumick I. S.editors Proceedings of ACM SIGMOD International Conference on Management of Data* ACMPress **1996** pp. 103-114.
[14] Guha S. Rastogi R. and Shim K. "CURE: an efficient clustering algorithm forlarge databases" *In Haas L. M. and Tiwary A. editors Proceedings of the 1998 ACMSIGMOD international conference on Management of data* ACM Press New YorkNY USA **1998** pp. 73-84.
[15] Chiang J. H. and Hao P. Y. *IEEE Transactions on Fuzzy Systems* vol. 11 **2003** pp. 518-527.