Text Classifier Based on an Improved SVM Decision Tree

Zhenqiang Xu¹, Pengwei Li², Yunxia Wang¹

¹School of Information Science and Engineering Henan University of Technology, Zhengzhou 450001, China
²Teaching Department of Computer, Anyang Normal University, Huanghe Road, Anyang, Henan, 455000, China
xuzhenqiang@haut.edu.cn, lpw@aynu.edu.cn, wangyunxia@haut.edu.cn

Abstract

The classifier of SVM decision tree (SVM-DT) takes advantage of both the efficient computation of the tree architecture and the high classification accuracy of SVMs. The paper proposes a new effective approach to optimize the SVM-DT classifier while presents the research on text categorization using SVM-DT classifier. In this approach, a novel separability measure is defined based on Support vector domain description (SVDD), and an improved SVM-DT is proposed. Experimental results demonstrate the effectiveness and efficiency of the improved SVM decision tree.

© 2012 Published by Elsevier B.V. Selection and/or peer review under responsibility of ICMPBE International Committee.
Open access under CC BY-NC-ND license.
Keywords: SVM decision tree; Text Categorization; Support vector domain description

1. Introduction

With the development of the web, large numbers of documents are available on the Internet. Automatic text categorization becomes a key technology to deal with a large numbers of documents. It’s supervised learning task for automatically assigning documents to pre-defined classes of documents. More and more methods based on statistical theory and machine learning, such as Naive Bayes, k-nearest neighbor, support vector machines (SVM), decision tree have been applied to text categorization in recent years [1].

The recent results in pattern recognition have shown that support vector machine classifiers often have superior recognition performance in comparison to other classification methods. In most cases the classification of the text document organization and management depended on the multi-class text categorization [2]. However SVM is a learning approach for solving a two-class pattern recognition problem. For the conventional methods, an n-class problem is converted into n two-class problems or n(n-

* This work was supported in part by Natural Science Foundation of Education Department of Henan Province, China, under grant number 2011A120008.
1/2 two-class problems. A variety of strategies to the combination for the multi-class classification using Support Vector Machines have been proposed and widely used, for example 1vs1, 1vsRest, SVM-DT. Among them, the classifier architecture SVM-DT (Support Vector Machines utilizing Binary Decision Tree), Utilizing this architecture, N-1 SVMs needed to be trained for an N class problem. The performance of SVM-DT is superior to 1vs1 and 1vsRest [3]. In this paper, a new effective approach to optimize the decision-tree-based multiclass support vector machines has been proposed after presents the research on categorization using the decision-tree-based multiclass support vector machines, at the same time a novel separability measure on is defined based on support vector data description. Experimental results of text categorization showed that the proposed separability measure is very effective.

2. Document Representation

At present, documents are represented by the widely used vector-space model. In this model, each document is represented as a vector d. Each dimension in the vector \( V(d) \) stands for a distinct term in the term space of the document collection [4]. That is, take one document as a set of Term sequences, including term t and term weight w. Then the document will be made up of the pairs of \( <t_i, o_i(d)> \). We represent each document vector \( V(d) \) as

\[
V(d) = (t_1, o_1(d); t_2, o_2(d); \cdots; t_n, o_n(d))
\]

(1)

Where \( t_1, t_2, t_3, \cdots \) represent the features which express the document content and \( o_i(d) \) is the weight of \( t_i \) term of document d.

TFIDF is the most common weighting method used to describe documents in the Vector Space Model. The TFIDF measure is defined as follows:

\[
\omega_i(d) = \frac{tf_i(d) \times \log(N/n_i)}{\sqrt{\sum_j tf_j(d) \times \log(N/n_j)^2}}
\]

(2)

Where \( tf_i(d) \) is the raw frequency of \( t_i \) in document d, N is the total number of documents in the document corpus and \( n_i \) is the number of documents in the corpus where term \( i \) appears.

A major problem of text categorization is the high dimensionality of the feature space. Most of these dimensions are not relative to TC which results in reducing the performance of the classifier. An Feature Selection algorithm selects a subset of important features and removes irrelevant, redundant and noisy features for simpler and more accurate data representation. As a result, As a result, saving in the computational resources, storage and memory requirements could be achieved.

Some FS methods are usually used, such as document frequency (DF), mutual information (MI), a \( \chi^2 \) statistic (CHI), and term strength (TS).

Chi-Square statistic is the common statistical test that measures divergence from the distribution expected if one assumes the feature occurrence is actually independent of the class value. For the English text classification problem, IG and CHI better performance [5].

The CHI measure is defined as follows:

\[
\text{CHI}(t,c) = \frac{(N \times (AD-CB)^2)}{((A+C) \times (B+D) \times (A+B) \times (C+D))}
\]

(3)

Where A is the number of times t and c co-occur. B is the number of time the t occurs without c. C is the number of times c occurs without t. D is the number of times neither c nor t occurs. N is the total number of documents.

3. SVM decision tree classifier

The modal of Support vector machines utilizing a decision tree (SVM-DT) takes advantage of both the efficient computation of the tree architecture and the high classification accuracy of SVMs [6]. The basic idea of SVM-DT is that a class can be divided into the problem into a series of two-class problem, and the
two-class problem can be solved by SVM. Two architectures of decision tree are skewness tree and normal tree. In Fig.1, the figures are the examples of a four-class problem. In skewness tree, at first the hyperplane which separates Class 1 from classes 2, 3, 4 is calculated. Next, the hyperplane which separates Class 2 from Classes 3, 4 is calculated and finally the hyperplane which separates Class 3 from Class 4 is calculated.

In normal tree, the hyperplane which separates Class 1, 2 from Classes 3, 4 is calculated. Next the hyperplane which separates Class 1 from Class 2 and the hyperplane which separates Class 3 from Class 4 are calculated. The Capacity of skewness tree is Similar to normal tree. The object of study is the classification of skewness tree in this paper.

4. The measure of inter-class separability

4.1. Support Vector Data Description

The Support Vector Data Description is one of the methods which we will use to describe our data. It is inspired on the Support Vector Classifier of V. Vapnik. It is a method for the one-class problems. The idea of SVDD is to find a sphere with minimal volume which contains all data (or most of) the data objects [7].

Suppose we are given a data-set \( \{x_1, \ldots, x_N\} \in \chi \), where \( N \) is the number of samples, and \( \chi \) is the training set. The goal is find a hypersphere (in a high-dimensional Hilbert feature space where the samples have been mapped through a nonlinear transformation) of radius \( R \) and centre with a minimum volume containing most of these data objects. Therefore, minimize \( R^2 \) constrained to \( (x_i - \alpha)^T(x_i - \alpha) \leq R^2 \) is the goal. As usual in the SVM framework, the problem becomes

\[
\min R^2 + C \sum_i \xi_i \quad \text{s.t.} \quad (x_i - \alpha)^T(x_i - \alpha) \leq R^2 + \xi_i, \forall i, \xi_i \geq 0, i = 1, \ldots, l
\]

Where the parameter \( C \) controls the tradeoff between the volume of the hyperphere and the permitted errors (re-gularization parameter), a set of slack variables \( \xi_i (\xi_i \geq 0) \) are introduced for the distribution may contain outliers.

The primal function (4) is usually solved through its Lagrangian dual problem, which consists of solving

\[
\max \sum \alpha_i K(x_i, x_j) - \sum \alpha_i \alpha_j K(x_i, x_j) \quad \text{s.t.} \quad 0 \leq \alpha_i \leq C, \sum \alpha_i = 1, \quad j = 1, \ldots, l
\]

To determine whether a test point \( z \) is within the sphere, the distance to the centre of the sphere has to be calculated. A test object \( z \) is accepted when this distance is smaller than the radius, i.e., when
Expressing the center of the sphere in terms of the support vectors, we accept objects when \( K(z, z) - 2 \sum_{i, j} \alpha_i \alpha_j K(x_i, x_j) \leq R^2 \).

4.2 Inter-class Separability Base on SVDD

Give set of k types of samples \( \{X^i, X^2, \ldots, X^k\} \), where \( X^i = \{x_i, x_i', \ldots, x_i''\} \) is the ith class samples. We structure the hypersphere for each kinds of training sample respectively, and then the separability measures matrix is defined as formula 5 shows.

\[
D = \begin{bmatrix}
D_{11}(t) & \cdots & D_{1n}(t) \\
\vdots & \ddots & \vdots \\
D_{n1}(t) & \cdots & D_{nn}(t)
\end{bmatrix}
\]  

(6)

\[
D_{ij}(t) = \frac{N_i(t) + N_j(t)}{m_i + m_j} \in [0, 1]
\]  

(7)

Where \( N_i(t) \) was sample quantity which smaller than t (t was represent the distance from the ith class samples to the jth class hypersphere). Through analyzing, this matrix is symmetrical, if \( D_{ij}(t) \) was 1, that was show completely overlapped of the ith class and the jth class, if \( D_{ij}(t) \) was 0, that was show no crossing area of the ith class and the jth class. The smaller the value of \( D_{ij}(t) \) was, the better Separability of two kinds samples were.

4.3 Algorithm Description

On the basis of the optimized class separation solution in [3], the improved SVM-DT (ISVM-DT) algorithm of introducing separability measure base on Support vector domain description (SVDD) was designed.

Suppose \( S_1 \) is a set of positive example and \( S_2 \) is a set of negative example, \( C \) is the number of classification.

Step 1. With RBF kernel function, for a given train examples, the set of hyperphere were obtained, denoted by \( \{(r_1, c_1), (r_2, c_2), \ldots, (r_C, c_C)\} \), where \( r_i \) is the radius of the ith hyperphere and \( c_i \) is the center of the ith hyperphere.

Step 2. By formula (6), the separability measures matrix \( D \) were computed.

Step 3. Under the assumption that the \( C_i \) class is one of \( S_1 \), by formula (8), \( M_{SIS2}^k \) and \( M_{S2}^k \) were obtained.

\[
\begin{aligned}
M_{SIS2}^k &= \frac{1}{C-1} \sum_{j \neq 2} D_{ij} \\
M_{S2}^k &= \frac{2}{(C-1)(C-2)} \sum_{j \neq \neq 2, i} D_{ij}
\end{aligned}
\]  

(8)

Step 4. The \( C_i \) class corresponding to the maximum \( M_{SIS2}^k + M_{S2}^k \) is positive; we can construct the optimal classification surface.

Step 5. Constructing the Decision Tree for the remaining class According to the above algorithm.

5. Experiments Results

In this section, we investigate the performance of our proposed improved SVM-DT algorithm for document classification. The system performance is compared with ovo, SVM-DT. In the following
experiment, the optimization parameters values in SVM are selected with leave one out cross validation. The LIBSVM software was used in our experiment to solve the SVM optimization problem.

For our experiments we used Reuters-21578 document collections. The documents in the Reuters-21578 collection appeared on the Reuters newswire in 1987. The documents were assembled and indexed with categories by personnel from Reuters Ltd. The Reuters-21578 data set contains 21578 news articles in 135 categories. We adopt the top eight classes of Sample quantity. 6574 documents in train set and 2315 documents in test set.

Documents were represented using a vector space model, where documents are represented by feature vector of terms. The preprocessing we carried out for the assigned data includes stop word elimination, stemming and sentence boundary determination. Each term is associated with a TF-IDF weight, where TF denotes the frequency of a term in a document, and IDF is calculated based on the distribution of the term in the training corpus. In all experiments the document vectors were normalized to unit length. For example stop word elimination is to filter out the words in a text which are generally regarded as ‘functional words’ and do not carry meaning.

To compare the performance of the classification methods, we use the well-known $F_1$ measure introduced, this measure is harmonic mean of precision and recall, which combines recall and precision in the following ways [8]:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$  \hspace{1cm} (9)

$$\text{Recall} = \frac{\text{Number of correct positive prediction}}{\text{Number of positive examples}}$$ \hspace{1cm} (10)

$$\text{Precision} = \frac{\text{Number of correct positive prediction}}{\text{Number of positive predictions}}$$ \hspace{1cm} (11)

We evaluated the systems with both macro and micro average $F_1$. The Micro- and Macro- $F_1$ emphasize the performance of the algorithm on common and rare categories, respectively. The experiment is conducted on the different methods including OvO, SVM-DT and improved SVM-DT (ISVM-DT).

By equation (6) to calculate the separability measure matrix base on SVDD, The algorithm described in paper is constructed to improve SVM decision tree classifier.

Table I shows classification performance and efficiency.

| Method   | Macro-F1 | Micro-F1 | Training time | Test time |
|----------|----------|----------|---------------|-----------|
| OvO      | 86.4%    | 85.92%   | 213s          | 46s       |
| SVM-DT   | 89.4%    | 90.34%   | 78s           | 30s       |
| ISVM-DT  | 91.3%    | 92.5%    | 67s           | 27s       |

The results in the table 1 show that the training time of Original SVM-DT algorithm is reduced obviously as compare with OvO but the performance is similar. The Improved SVM-DT based on svdd was significantly better compared to the other methods at the classification precision and training efficiency.

6. Conclusions

In this paper, a novel separability measure is defined base on Support vector domain description (SVDD), and an improved SVM decision tree is provided for solving multi-class problems of text categorization. The SVM-DT based on svdd was designed to provide superior multi-class classification performance. The experiments showed that this method is experimental result show that this algorithm has good performance in classification precision and efficiency.
Acknowledgment

The work is supported by Science Research Foundation of Henan University of Technology Grant 06XJC016 and 07XGG024.

References

[1] HAN Jia-xin, HE Hua-can. SVMDT Classifier and Its Application Research in Text Classification. Application Research of Computers, 2004, 21(1): 23-24.

[2] Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features[C]. Proc. of the 10th European Conf Machine Learning, 1999, 1372-142.

[3] ZHU Yuan-ping, DAI Ru-wei. Text Classifier Based on SVM Decision Tree. Pattern Recognition and Artificial Intelligence, 2005, 18(4): 412-414.

[4] Fabrizio Sebastiani. Machine Learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1): 1-47.

[5] Yiming Yang, Jan O. Pedersen. A Comparative Study on Feature Selection in Text Categorization [A]. Proceedings of ICML-97 [C]. 412-420.

[6] Takahashi F, Abe S. Decision-Tree-Based Multiclass Support Vector Machines. In: Proc of the 9th International Conference on Neural Information Processing. Singapore, Singapore, 2002, 1418-1422.

[7] Ban T, Abe S. Implementing Multi-class Classifiers by One-class Classification Methods[C]//2006 International Joint Conference on Neural Networks Vancouver. BC, Canada: Sheraton Vancouver Wall Centre Hotel, July 2006: 327-332.

[8] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys, Vol.34, No.1, 1-47, 2002.