Expert Systems with Applications 38 (2011) 2030–2039

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Two-level hierarchical combination method for text classification

Wen Li [a,b,*], Duoqian Miao [a], Weili Wang [a,b]

[a] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[b] Information Engineering School, Nanchang University, Nanchang 330031, China

ARTICLE INFO

Keywords:
Text classification
Combination method
Variable precision rough sets
Support vector machine
k nearest neighbor

ABSTRACT

Text classification has been recognized as one of the key techniques in organizing digital data. The intuition that each algorithm has its bias data and build a high performance classifier via some combination of different algorithm is a long motivation. In this paper, we proposed a two-level hierarchical algorithm that systematically combines the strength of support vector machine (SVM) and k nearest neighbor (KNN) techniques based on variable precision rough sets (VPRS) to improve the precision of text classification. First, an extension of regular SVM named variable precision rough SVM (VPRSVM), which partitions the feature space into three kinds of approximation regions, is presented. Second, a modified KNN algorithm named restrictive k nearest neighbor (RKNN) is put forward to reclassify texts in boundary region effectively and efficiently. The proposed algorithm overcomes the drawbacks of sensitive to noises of SVM and low efficiency of KNN. Experimental results compared with traditional algorithms indicate that the proposed method can improve the overall performance significantly.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Text classification (TC), also known as text categorization, aims at automating the process that assigns documents to a set of previously fixed categories, has always been a hot topic. Many popular algorithms have been applied to text categorization. No Free Lunch (NFL) theorems (Wolpert & Macready, 1997) have shown that learning algorithms cannot be universally acceptable and any algorithm has its bias data. When the data fits the underlying classification strategy well, the system accuracy can be very high, and vice versa (Tan, Cheng, & Ghanem, 2005). Among the many well-known algorithms, support vector machine (SVM) (Joachims, 1998) and k nearest neighbor (kNN) (Cover & Hart, 1967) are widely used because their excellent learning performance both in theory and in practices. But despite their advantages, they also have weaknesses and limitations.

SVM is well founded in terms of computational learning theory and very open to theoretical understanding. The final classifier obtained by the SVM depends only on a small portion of the training samples, i.e. support vectors, which is good for implementation. However, this makes the SVM sensitive to noises or outliers and patterns that were wrongly classified lie near the separation hyper-plane (Zhang & Wang, 2008).

KNN is a well-known statically approach in pattern recognition. It is also known as one of the top-performing methods on the benchmark Reuters corpus (Yang & Liu, 1999). Because of using an instance-based learning algorithm, the KNN algorithm simply stores all of the training examples as classifier and delay learning until prediction phase. Under circumstance of huge amount of training data, considerable time would be paid during the classification process in KNN. Besides, the performance of KNN may be affected by noisy data (Srisawat, Phienthrakul, & Kijsirikul, 2006).

Researchers have long pursued the promise of harnessing multiple text classifiers to synthesize a more accurate classification procedure via some combination of the outputs of the contributing classifiers (Bennett, Dumais, & Horvitz, 2005). In this paper, we present a hybrid algorithm based on variable precision rough sets (VPRS) by combining the respective excellences of SVM and KNN in order to improve classification accuracy. The proposed method is based on a two-stage algorithm. First, by introducing the VPRS theory into the support vector machines, a variable precision rough SVM (VPRSVM) is presented. The transformed feature space is partitioned by using VPRSVM where lower and upper approximations of each category are defined. Second, on analysis of the characteristic of boundary region text, a modified KNN algorithm, namely restrictive k nearest neighbor (RKNN) classifier is put forward which built on the reduced candidate classes, and it only requires classifying testing document of boundary region effectively and efficiently.

Since uncertainties in the labeling are taken into account, our approach tries to provide a practical mechanism to deal with real-world noisy text data. Analysis of the different approximation

* Corresponding author at: Department of Computer Science and Technology, Tongji University, Shanghai 201804, China. Tel.: +86 15900799568.

E-mail addresses: jx_wenli@yahoo.com.cn (W. Li), miaoduoqian@163.com (D. Miao), ken.wlwang@gmail.com (W. Wang).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.07.139

space indicates that the VPRSVM algorithm partitions the feature space reasonably. Experimental results compared with traditional machine learning methods show that the proposed combination method improves the overall classification performance significantly.

The remainder of the paper is organized as follows: Section 2 gives an overview of related work. Section 3 introduces the basic background knowledge about VPRS and related text classification technique (i.e. SVM and KNN). Section 4 describes the proposed combination algorithm. Section 5 illustrates the implementation of the proposed algorithm in detail. Experimental results reports and discusses in Section 6. Finally, conclusions and future work are summarized in Section 7.

2. Related works

Much of the previous work about hybrid classification algorithms concentrated on combining various high performance classifiers in a hierarchical manner. As some examples, Silva and Ribeiro (2006) proposed a two-level hierarchical hybrid SVM–RVM model. The model first level uses an RVM (relevance vector machine) to determine the less confident classified examples and the second level makes use of an SVM to classify these texts. The drawback of this method is the difficulty on the definition of appropriate criteria for defining second level examples. Tang and Gao (2007) introduced a multi-model classifier that combines SVM with KNN to deal with the classification problem involves overlapping patterns. However, two round KNN algorithm is carried out to eliminate noisy pattern and extract boundary pattern. Then dual SVM classifier is trained to make the final decision. The efficiency of this method may be low. Miao, Duan, and Zhang (2009) combined the KNN and Rocchio techniques to enhance classification performance. The method need to calculate the similarity between any two training data to create equivalence classes. In addition, it not specially suited for dealing with noisy data.

In practice, much works have been carried out on the combination of rough set theory and classification method (Lingras & Butz, 2007a, 2007b; Lingras, Chen, & Miao, 2009; Saha, Murthy, & Pal, 2007; Tan, Cheng, & Xu, 2007).

Saha et al. (2007) proposed Rough Set Meta (RSM) classifier to extract decision rules from trained classifier ensembles. The key idea of the algorithms is redundancy removal from the generated model and decision rule generation from reduced model. Experimental studied show the method improves accuracy uniformly. But ensemble methods need to generate models multiple times over different subset of the training examples. The time complexity and spatial complexity of rough set based classifier reduction algorithm is also high.

Lingras and Butz (2007a, 2007b) proposed a rough set interpretation of SVM and applied in classification that provide an instructive idea for expansion of SVM classifier. It is not difficult to find that the positive region must be absolutely correct in Lingras's definition, if adopting the method for classification problem with noisy data or outliers, the boundary region will become large and algorithm failure. Generally, the training data for text classification task is achieved by manual assignment of class labels to documents by experts. When faced with the challenge of selecting a class label from a set of similar or confusing class labels for a document, the expert often chooses a class label that seems the most plausible (Ramakrishnan, Chitrapura, & Krishnapuram, 2005). It is almost inevitable that there is some noise data in corpus we have collected. Based on this analysis, a refined rough SVM—VPRSVM is presented.

Further more, Lingras's techniques provide better semantic interpretations of the classification process, but how to deal with the boundary region has not yet been discussed. For automatic text classification problem, mining the correct class label of texts in the boundary region is a tough work. The RKNN algorithm is proposed to fulfill this task. That is to say, a systematical classification mechanisms is put forward in this paper.

3. Background knowledge

In this section, we review variable precision rough sets and the two text classification techniques applied in this paper, i.e. SVM and KNN algorithms.

3.1. Text classification technique

3.1.1. Support vector machine

SVM is a new machine learning method introduced by Vapnik (1995). It is based on Statistical Learning Theory (SLT) and Structural Risk Minimization (SRM) principle. SVMs become the hotspot of machine learning because of their excellent learning performance and generalization capability.

SVM is originally designed for binary classification. Given t training samples (x_1,y_1), $(x_2,y_2),\ldots,(x_t,y_t)$, where $x_i \in R^n$, $i = 1,\ldots,t$ and $y_i \in \{+1,-1\}$ is the class label of x_i, SVM seeks the optimal hyper-plane that best separates the two classes from each other with the largest margin which is equivalent to solving the following problem (Bottou, Cortes, & Denker, 1994):

$$\text{minimize} \quad J(\omega,b,\xi) = \frac{1}{2}\|\omega\|^2 + C\sum_{j=1}^{t}\xi_j(\omega)^T \tag{1}$$

$$\text{subject to} \quad (\omega)^T\varphi(x_j) + b \geqslant 1 - \xi_j, \quad \text{if } y_j = 1 \tag{2}$$

$$(\omega)^T\varphi(x_j) + b \leqslant -1 + \xi_j, \quad \text{if } y_j = -1 \tag{3}$$

$$\xi_j \geqslant 0, \quad j = 1,\ldots,t \tag{4}$$

Classification is then achieved according to the following function:

$$Q(x) = sign((\omega)^T\varphi(x) + b) \tag{5}$$

where the data was mapped to a higher dimensional space by the function φ and C is the penalty parameter that controls the tradeoff between training errors and the margin.

In order to extend them for multi-class classification, several schemes have been proposed and the three methods based on binary classifications: One-Vs-Rest, One-Vs-One, and directed acyclic graph SVM (DAGSVM) are widely used (Hsu & Lin, 2002).

The One-Vs-Rest method constructs k SVM models where k is the number of classes. Each classifier trained to distinguish the examples in a single class from all other examples. The final output of the k One-Vs-Rest SVMs is the class with the highest output value. While $k(k-1)/2$ classifiers where each one is trained on two classes data were constructed in both One-Vs-One and DAGSVM method. Rifkin and Klautau (2004)'s experiments show that simple One-Vs-Rest concept scheme is as accurate as any other approach. Following the recommendation of Rifkin et al., we use the One-Vs-Rest approach as the baseline SVM classifier.

3.1.2. KNN algorithm

KNN is a similarity-based learning algorithm. To classify an unknown document x, the KNN classifier finds the k nearest neighbors among the training documents and uses the categories of the k neighbors to weight the category candidates. Then majority voting among the categories of documents in the neighborhood is used to decide the class label of x.

Given n classes $c_1,c_2,\ldots,c_n$ and t training samples $x_1,x_2,\ldots,x_t$, and $y(x_i,c_j) = \begin{cases} 1 & x_i \in c_j \\ 0 & x_i \notin c_j \end{cases}$ is the classification for document x_i with

Download English Version:

https://daneshyari.com/en/article/385792

Download Persian Version:

https://daneshyari.com/article/385792

Daneshyari.com