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.

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

respect to category $c_j (i = 1,\ldots,t, j = 1,\ldots,n)$, the decision rule in KNN can be written as:

$$\text{assign } x \text{ in } c_j \text{ if } score(x,c_j) = \arg\max_{j=1}^{n} \sum_{i=1}^{k} sim(x,x_i) y(x_i,c_j) \quad (6)$$

where $sim(x,x_i)$ is the similarity between the test document x and the training document x_i.

3.2. Variable precision rough sets

Rough set theory, introduced by Pawlak (1982), is a formal mathematical tool to deal with uncertain, imprecise or incomplete information.

Definition 1 (*Approximation space).* Let U denote the universe (a finite non-empty set) and R be a family equivalence relation on U. The pair (U,R) is called an approximation space, denoted by $K = (U,R)$.

Definition 2 (*Partitions*). Let U be an universe, C be a family of subsets of U, $C = \{X_1,X_2,\ldots,X_n\}$. C is called a partition of U if the following properties are satisfied:

$$X_1 \cup X_2 \cup \cdots \cup X_n = U \quad (7)$$
$$X_i \cap X_j = \emptyset \quad (i \neq j) \quad (8)$$

Definition 3 (*Equivalence class*). Let U be an universe and R be an equivalence relation on U. We denote the equivalence class of object x in R by $[x]_R$. The set $\{[x]_R | x \in U\}$ is called a classification of U induced by R.

Definition 4. Lower approximation, upper approximation, and boundary region.

Let $K = (U,R)$ be an approximation space and X be a subset of U. The sets

$$\underline{X}_R = \{x | [x]_R \subseteq X\} \quad (9)$$
$$\overline{X}_R = \{x | [x]_R \cap X \neq \emptyset\} \quad (10)$$
$$BN_R(X) = \overline{X}_R - \underline{X}_R \quad (11)$$

are called lower approximation, upper approximation, and boundary region of X with respect to R in K, respectively.

Pawlak's rough set model is founded on classical set theory and information gathered from positive region will only be considered. So original rough set model cannot deal with datasets that have noisy data effectively and then some latent useful knowledge in boundary region may not be fully captured. In order to overcome the limitations, some extended rough set models have been put forward and variable precision rough set model as introduced by Ziarko (1993) is one of the most important extensions.

In VPRS model, standard inclusion relation is extended to majority inclusion relation, defined as Definition 5.

Definition 5 (*Majority inclusion relation*). Let X and Y be non-empty subsets of a finite universe U. The majority inclusion relation introduces the measure $c(X,Y)$ of the relative degree of misclassification of the set X with respect to set Y defined as

$$\begin{aligned} &c(X,Y) = 1 - card(X \cap Y)/card(X) && \text{if } card(X) > 0 \text{ or} \\ &c(X,Y) = 0 && \text{if } card(X) = 0 \end{aligned} \quad (12)$$

where *card* denotes set cardinality.

Based on this measure, three kinds of approximation regions can be defined as Definition 6.

Definition 6. β-lower approximation, β-upper approximation and β-boundary region.

Let $K = (U,R)$ be an approximation space and X be a subset of U, the sets

$$\underline{X}_\beta^R = \{x | c([x]_R, X) \leqslant \beta\} \quad (13)$$
$$\overline{X}_\beta^R = \{x | c([x]_R, X) < 1 - \beta\} \quad (14)$$
$$BN_\beta^R(X) = \overline{X}_\beta^R - \underline{X}_\beta^R \quad (15)$$

are called β-lower approximation, β-upper approximation and β-boundary region of X in K, respectively.

How to choose optimal β value is a very difficult task. It often depends on our subjectivity indeed or on some prior knowledge (Wang & Zhou, 2009). Some researches have focused on it (Beynon, 2004; Su & Hsu, 2006). Here, we provide an indirect method to determine optimal β value which will be discussed extensively in Section 5.1.

4. Proposed algorithm

In this paper, we consider single-label multi-class text categorization problem that assigning a single label to each example.

The task of text classification is hampered by the lack of large amounts of correctly labeled examples. The generation of training data is typically achieved by manual assignment of class labels to documents by experts. Manual annotations inherently exhibit a certain level of approximation or uncertainty (Ramakrishnan et al., 2005).

In this section, variable precision rough set based SVM, which can reduce the influence of noisily labeled examples, and modified KNN algorithm are proposed, respectively. Next section will discuss how to implement and combine these methods.

4.1. Variable precision rough support vector machine

As mentioned previously, in many scenarios, it is easy to generate a labeled dataset with some amount of noise in the labeling. A practical text learning algorithm needs to be resilient to such noisy labeling (Ramakrishnan et al., 2005). Take this consideration in mind, an extended SVM is proposed under the concepts of VPRS. The noise can be captured by using the modified approximation space and equivalence class.

Definition 7. Modified approximation space.

Let U denote the universal sample space (a finite non-empty set) and $\mathbb{Q}(x) = (\omega)^T \varphi(x) + b$ is a collection of discrimination function based on SVM. The pair $(U,\mathbb{Q})$ is called a modified approximation space, denoted by $K = (U,\mathbb{Q})$.

$\mathbb{Q}(x)$ be a family of equivalence relations which satisfied the three properties namely reflexive, symmetric and transitive.

Definition 8. Modified equivalence class based on discrimination function $\mathbb{Q}(x)$.

Given k classes $c_1,c_2,\ldots,c_k$, let $K = (U,\mathbb{Q})$ be a modified approximation space. $Q_m(x) = (\omega)^T \varphi(x) + b_m$ is the discrimination function of class c_m, $1 \leqslant m \leqslant k$, and $Q_m(x) \in \mathbb{Q}(x)$. We denote the modified equivalence class of sample x based on variable precision rough SVM by $[x]_{RSVM}^m$. The set $\{[x]_{RSVM}^m | x \in U\}$ is called a classification of U induced by Q_m which satisfied:

$$[x]_{RSVM}^m = \{x_i | Q_m(x) Q_m(x_i) \geqslant 0\} \quad (16)$$

Definition 9. Modified β-lower approximation, β-upper approximation and β-boundary region.

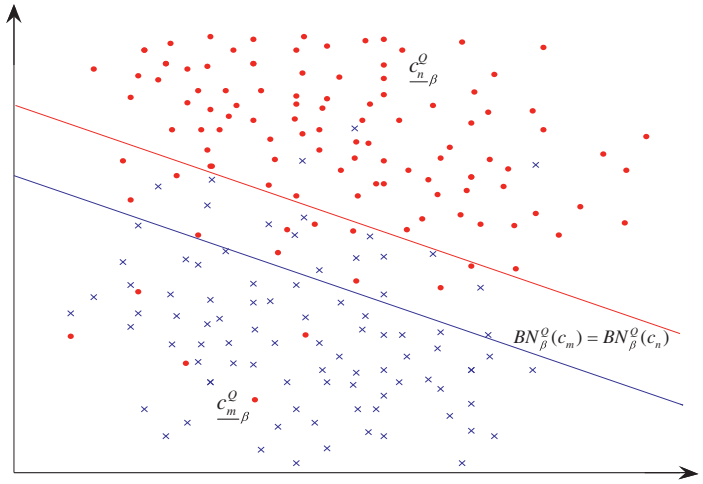

Fig. 1. VPRS based SVM for binary classification.

Let $K = (U, \mathbb{Q})$ be an approximation space, c_m be a subset of U includes documents which are manually labeled to category c_m. The sets:

$$\underline{c_m^Q}_\beta = \{x | c([x]_{RSVM}^m, c_m) \leqslant \beta\} \tag{17}$$

$$\overline{c_m^Q}_\beta = \{x | c([x]_{RSVM}^m, c_m) < 1 - \beta\} \tag{18}$$

$$BN_\beta^Q(c_m) = \overline{c_m^Q}_\beta - \underline{c_m^Q}_\beta \tag{19}$$

are called modified β-lower approximation, β-upper approximation and β-boundary region of c_m in k, respectively.

The variable precision rough set approach to SVM classification will allow us to create three kinds of approximation regions. For solving specific single-label text categorization problem, our emphasis is to determine the β-lower approximation set and β-boundary region. Fig. 1 shows β-lower approximation set for a classification problem with two classes c_m and $c_n (1 \leqslant m \leqslant n \leqslant k)$, where the objects have already been mapped into a higher dimensional space by mapping function φ.

Definition 10. The measure of classification quality in VPRSVM for category c_m with respect to discrimination function $\mathbb{Q}(x)$ is defined as:

$$K_\beta(c_m) = \frac{card\left(\underline{c_m^Q}_\beta\right)}{card(c_m)}, \quad 1 \leqslant m \leqslant k \tag{20}$$

and the overall classification quality is

$$K_\beta = \frac{\sum_{m=1}^k card\left(\underline{c_m^Q}_\beta\right)}{card(U)} \tag{21}$$

Classification must be correct absolutely in Lingras's rough set SVM model, and the boundary region will become large when deal with noisy dataset. Our simple extension of SVM classifier provides the basis for a more practical application with noise data.

VPRSVM is used to partition feature space both of training data and testing data, which makes it much easier for subsequent classification algorithm.

4.2. Restrictive k nearest neighbor algorithm

After adopting VPRSVM classifier, β-lower approximation and β-boundary region are obtained for each class. According to Definitions 8 and 9, texts in lower approximation set assign to corresponding category with high confidence. Otherwise, text in boundary region means the system does not have the ability to classify it, having no choice but make an unconfident decision. These classification results are unreliable. Nevertheless, efficient testing of memberships in boundary region is an interesting but tough research work.

KNN is suitable for little dataset and can achieve better performance. Especially, KNN is a type of instance-based learning where the function is only approximated locally and all computation is deferred until classification. Using the output of previous VPRSVM classifier, it is easy to find out the latent candidate category for each testing document. Thusly, we proposed the restrictive k nearest neighbor (RKNN) algorithm with which the scope of the k nearest neighbors only narrow into the limit training instances belong to the prospective categories mining through previous VPRSVM classifier. By this way, RKNN classifier has more discriminatory power in deciding whether the testing instance belongs to.

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 respect to category c_j ($i = 1, \ldots, t, j = 1, \ldots, n$), the decision rule in RKNN can be written as:

$$\text{assign } x \text{ in } c_j \text{ if } score(x, c_j) = \arg\max_{j=1}^{n} \sum_{i=1}^{k} sim(x, x_i) y(x_i, c_j) \tag{22}$$

$$c_j \in \text{candidate category} \tag{23}$$

An observation of boundary region can be drawn: boundary region texts contain many "ambiguous" Chinese words, such as 'movie', 'qualifications', 'treasury' and so on. These words always occur at different category (experimentation dataset and predefined category see Section 6.1) and even give error instruction. At the same time, there also exists some high discrimination word, e.g. 'resume', 'jobseeker', 'Andersen', 'constitutional', 'tumor', 'cancer'. With this situation in mind, we focus our efforts on efficient selecting high discrimination word with instruction function and a feature selection method based on Fisher criterion is given as follow.

$$J_{fisher}(t) = \frac{S_B(t)}{S_W(t)} \tag{24}$$

Given k classes $c_1, c_2, \ldots, c_k$, the following matrices are defined:
Between-class scatter matrix

$$S_B(t) = \sum_{i=1}^{k} \frac{n_i}{n} (m_i(t) - m(t))^2 \tag{25}$$

Within-class scatter matrix

$$S_W(t) = \frac{1}{n} \sum_{i=1}^{k} \sum_{x \in c_i} (x(t) - m_i(t))^2 \tag{26}$$

where t is the term of candidate features; $m_i(t)$ is mean value of t over class c_i; $m(t)$ is mean value of t over all classes; n_i is the number of documents in class c_i; and n is the total number of documents over all classes.

Further analysis of feature space found that the document vector in boundary region tend to be "drown" by the majority class. In order to analysis the influence of unbalanced dataset problem, where the training instances of the target class are significantly outnumbered by the other training instances (Wu & Chang, 2003), two feature selection scheme *Global* and *Local* conducted (How & Kiong, 2005):

Global: features are extracted from the documents under all categories according to J_{fisher} value.
Local: features are extracted from each category according to J_{fisher}value in different class.

5. Implementation of the proposed approaches

In this section, we will discuss the operational detail of implementing VPRSVM and RKNN algorithm in text classification problem.

5.1. Optimum determination of different regions

VPRS involved parameters β. Given the value of β, three regions are obtained adoption of VPRSVM algorithm. Despite their diverse applications in many domains, it is difficult to find out the optimal parameter and the values are often given subjectively. To the best of our knowledge, existing approaches adopted unified β value for different equivalence classes. However, as indicated in Japkowicz (2002), unbalanced datasets often appear in many practical applications. In an unbalanced dataset, the majority classes are represented by a large portion of all the examples, while the other, the minority classes have only a small percentage of all examples. Given a unified β value is not able to reflect the category diversity.

Very intuitively, for applying VPRSVM to single-label text classification problem, we focus our efforts on boundary region and the goal is to find out the misclassified texts. Keep this situation in mind, we do not find out the optimal parameter β, but get straight to the point: determine different regions. Based on the analysis of misclassified samples, a simple and approximate method for distinguish the boundary region are proposed to alleviate the impact of imbalance of text data.

Definition 11. Given k classes $c_1, c_2, \ldots, c_k$, and x_i is a testing sample, $\mathbb{Q}(x) = (\omega)^T \varphi(x) + b$ is a collection of discrimination function based on variable precision rough SVM. Assuming that $Q_l^1(x_i) \geqslant Q_m^2(x_i) \geqslant \cdots \geqslant Q_n^k(x_i)$, $l, m, n \in \{1, 2, \ldots, k\}$, we call, c_l is the first predict class, c_m is the second predict class and so on.

Error occurs when a text is not belonging to the first predict class.

One-Vs-Rest technique, however, may lead to an arbitrary decision. When largest value of the margin $Q_{(\cdot)}^1(x)$ cannot "overwhelm" the other decisions, it means that they do not have the ability to classify the text, having no choice but making an unconfident decision. These classification results are unreliable and should include in the boundary region.

Definition 12. Given a sequence of discrimination function values of testing text x_i $\mathbb{Q}^r = \left\{Q_{(\cdot)}^1(x_i), Q_{(\cdot)}^2(x_i), \ldots, Q_{(\cdot)}^r(x_i)\right\}$, $r \geqslant 2$, the standard deviation σ of $\mathbb{Q}^r$ is defined as:

$$\sigma = \sqrt{\frac{1}{r-1}\sum_{i=1}^{r}(Q^r - \overline{\mathbb{Q}^r})^2} \tag{27}$$

The standard deviation is one measure of statistical dispersion and it could describe the fluctuation of the discrimination function values.

Hypothesis. If the value of σ is lower, the probability of the text misclassified into the first predict class is higher, and it has potential to belong to the rth predict class.

Statistics of the classification results validate the hypothesis by designing frequency histograms for the distribution of misclassified texts of σ in different classes (negative range values obtained by mapping). Fig. 2 are frequency histograms of some classes. The histograms show that most of the error occur closely to the ordinate. With the increase of σ, the number of misclassified texts decrease.

According to Central Limit Theorem (CLT), the sum of a large number of independent random variables each with finite mean and variance will be approximately normally distributed. Further more, Fig. 3 are some normality tests on the misclassified samples and they show that most of the texts scattered along the straight line. It is proved that the distribution of misclassified texts also observes the normal distribution by means of statistical testing.

Thresholds to determine the boundary region are derived by the powerful statistical method. By randomly sampled 20% of training set for threshold tuning, the threshold of σ for $\mathbb{Q}^r$ sequence of class c_i denoted by $\theta_{c_i}^r$ is determined by the confidence interval of quantile ρ for normal distribution. The optimum thresholds are determined by the following considerations: (**1**) ρ is large enough to contain most of the misclassified instances within the boundary region. (**2**) If ρ is too large, an amount of correct classification instances will flow into the boundary region and it will result in high computation cost. ρ is a tradeoff parameter between effectiveness and efficiency.

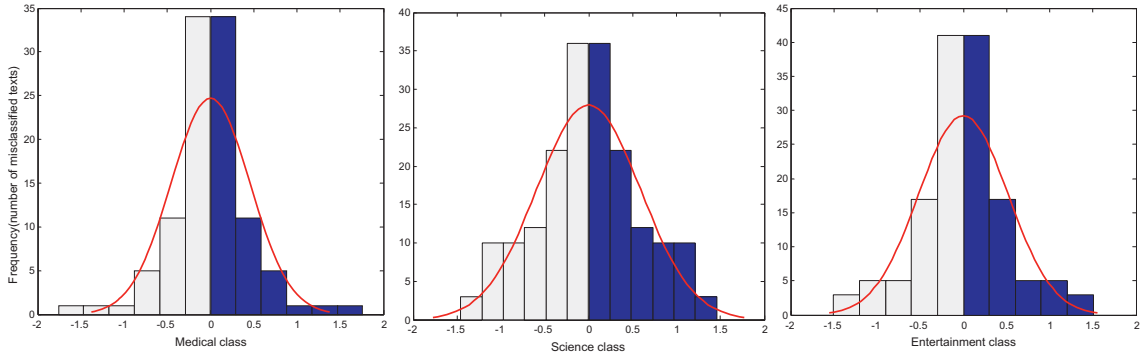

Fig. 2. Frequency histogram for the distribution of misclassified texts of σ.

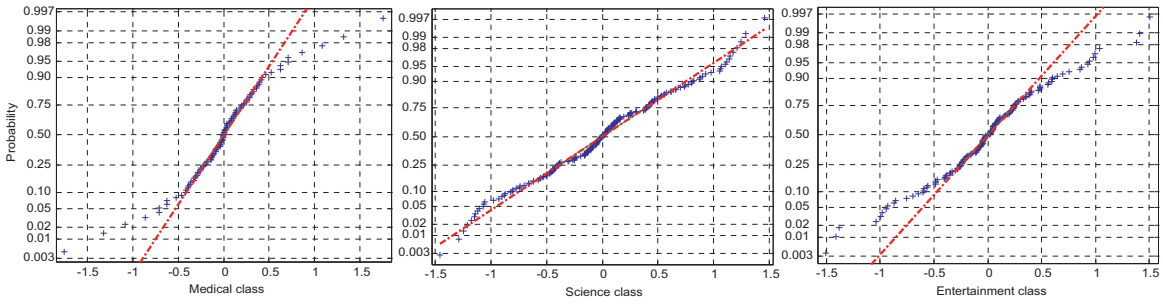

Fig. 3. Normality test for the distribution of misclassified texts of σ.

Given a definite value of ρ, different tailored value of β is determined for each category. Here, we note that our approach need not calculate β precisely, but obtained indirectly. Inspired by Beynon et al. (2003), we also analyze the variation of approximation regions. Each region is affected by ρ value directly, and the quality of classification will be influenced accordingly. The geometrical interpretation of the relationship among three region and parameter is illustrated in Fig. 4.

5.2. Hierarchical VPRSVM–RKNN algorithm

This section presents a two-level hierarchical hybrid VPRSVM–RKNN model whose objective is to combine the proposed two algorithm systematically. The main idea is classify all testing texts using VPRSVM first, and feature space is partitioned into three regions, i.e., lower approximation, upper approximation, and boundary region, for each class. Optimal determination of boundary region is obtained from validation on parameter tuning subset. Texts in lower approximation is the collection of all samples that make a decision with high confidence, boundary region is composed of all those texts which cannot be classified into any category with the certainty degree not lower than β by employing the current classifier model.

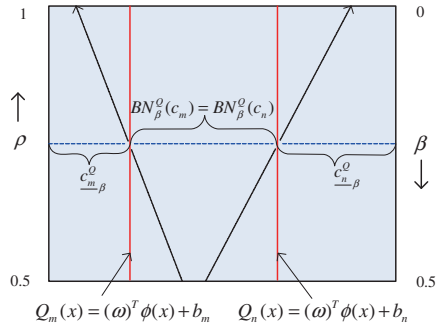

Fig. 4. Graphical interpretation of ranges of approximation regions for ρ, β goes from 0.5 to 1.

In order to determine hard-to-classify boundary region texts will then adopting the RKNN classifier. With the recognition of previous VPRSVM, boundary region texts also narrow the scope of discrimination by limiting it to several candidate category. Candidate categories were obtained by VPRSVM classifier, which means the promising categories for the testing example and determine from experimental analysis (see Section 6.4 Experiment 1). Therefore, the testing procedure of second step will only focus on the notable concern categories. The classification result of the VPRSVM–RKNN classifier system is the combination of the two-step discrimination result. The proposed architecture of VPRSVM–RKNN text classification is described in Fig. 5.

6. Experiment

6.1. Dataset

For a long time, there was no special benchmark for Chinese text categorization. Many reported results were achieved in some

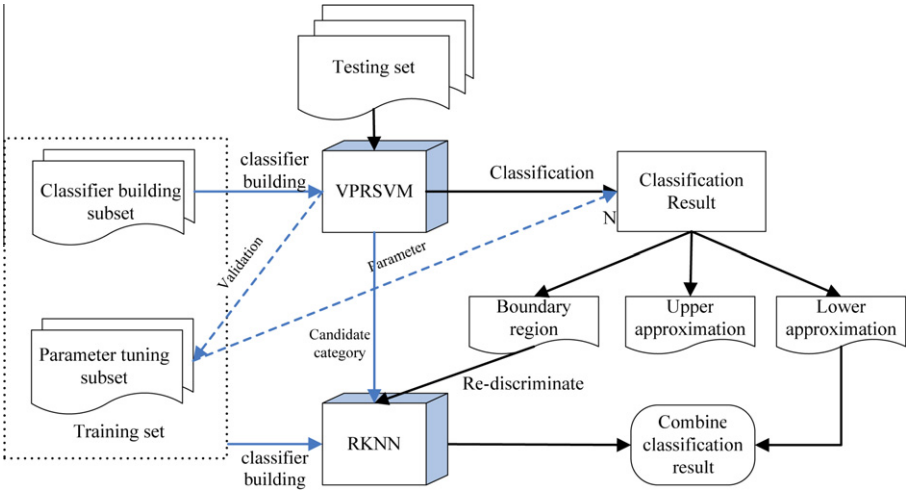

Fig. 5. Architecture of VPRSVM–RKNN classification.

Table 1
The distribution of TanCorp-12. (M = megabyte).

Class name	Number of texts	Size of class (M)	Class name	Number of texts	Size of class (M)
Art	546	1.42	Entertainment	1500	2.89
Car	590	0.89	Estate	935	1.80
Career	608	1.78	Medical	1406	2.64
Computer	2865	4.17	Region	150	0.49
Economy	819	2.60	Science	1040	1.97
Education	808	1.41	Sport	2805	4.20

benchmark of information retrieval that did not include the class information. In order to use this kind of benchmark, it is necessary to clustering the datasets before the task of categorization.

The TanCorp corpus,[1] a collection of 14,150 texts in Chinese language, has been collected and processed by Tan (2006). The corpus was divided into two hierarchical levels. The first level contains 12 big categories (art, car, career, computer, economy, education, entertainment, estate, medical, region, science and sport) and the second consists of 60 subclasses. This corpus can serve as three categorization datasets: one hierarchical dataset (TanCorpHier) and two flat datasets (TanCorp-12 and TanCorp-60). To evaluate the proposed approach, we have conducted experiments on TanCorp-12, and Table 1 shows the distribution of TanCorp-12.

6.2. Experimental setting

In our work, 70% documents randomly sampled for training set and the remaining 30% are used for testing set. The training set is further split into two subsets: the classifier building subset and the parameter tuning subset. The former is used to build the VPRSVM classifiers, while the latter is used to gain different parameter for forming the boundary region of each category that contains 20% of the training set documents.

During the preprocessing phase, we use a stop list to omit the most common words after word segmentation. It adopts the Vector Space Model (VSM) for text representation. In the VPRSVM classification phase, features are selected according to their weights, which are estimated by the IG (Information Gain) weighting technique. The technique was shown to be more promising than others (Yang & Pedersen, 1997). We used linear kernel for SVMs classifier since text classification problems are usually linearly separable. Learning parameters are set to penalty cost = 1.

For the experiment on each dataset, we used the 5-fold cross validation, and the average of all the results was used as the performance measure.

6.3. Evaluation metric

To analyze the performance of classification, we adopted the popular F1 measure. As shown in Table 2, four cases are considered as the result of classifier to the document (Yang & Liu, 1999).

TP (True Positive): the number of documents correctly classified to that class.
TN (True Negative): the number of documents correctly rejected from that class.
FP (False Positive): the number of documents incorrectly rejected from that class.
FN (False Negative): the number of documents incorrectly classified to that class.

Using these quantities, the performance of the classification is evaluated in terms of precision (*pr*), recall (*re*), and F_1 measure. Recall is defined to be the ratio of correct assignments by the system divided by the total number of correct assignments. Precision is the ratio of correct assignments by the system divided by the total number of the system's assignments. The F_1 measure is the combination of recall and precision with an equal weight.

[1] Available at http://www.searchforum.org.cn/tansongbo/corpus.htm.

Table 2
Contingency table for a set of binary decision.

Class c_i	System classification	
	Belong	Not belong
Expert classification		
Belong	$TP(c_i)$	$FP(c_i)$
Not belong	$FN(c_i)$	$TN(c_i)$

$$pr(c_i) = \frac{TP(c_i)}{TP(c_i) + FP(c_i)} \tag{28}$$

$$re(c_i) = \frac{TP(c_i)}{TP(c_i) + FN(c_i)} \tag{29}$$

$$F_1(c_i) = \frac{2 \cdot pr(c_i) \cdot re(c_i)}{pr(c_i) + re(c_i)} \tag{30}$$

For more than two classes, the F_1 scores are summarized over the different categories using the *Micro-average* scheme and *Macro-average* scheme. As previously discussed in Section 4.2, web pages in experimentation datasets are not equally distributed over categories. We use the *Micro-average-F_1* measure which has been widely used in information retrieval community to evaluate the methods (Lewis, 1991) which defined as:

$$\begin{aligned} \textit{Micro-average-}F_1 &= \frac{\sum_{i=1}^{k} TP(c_i)}{\sum_{i=1}^{k} (TP(c_i) + FN(c_i))} \\ &= \frac{\sum_{i=1}^{k} TP(c_i)}{\sum_{i=1}^{k} (TP(c_i) + FP(c_i))} \end{aligned} \tag{31}$$

6.4. Results and analysis

Experiment 1. Approximation space partition based on VPRSVM.

Definition 13. Error recall rate for the *i*th predict class denoted by $ER_i, 2 \leqslant i \leqslant k$, is defined as follows:

$$ER_i = \frac{card(\text{misclassified texts belong to the } i\text{th predict class})}{card(\text{misclassified texts})} \times 100\% \tag{32}$$

In our previous works, it was found that most of the misclassified text should belong to the second predict class. The distribution of errors is reported in Table 3.

According to Table 3, most of the error has the possibility to be revised limited in the first and second predict class. Furthermore, if *R* increases, it could cover most of the misclassified text at the cost of high computation. If so, more candidate categories will include in subsequent RKNN algorithm which influence the classification performance. The loss outweighs the gain. One of the most appropriate setting for parameter *R* is 2. This suggests that the candidate categories for RKNN algorithm are the first two predict class of each testing data employing VPRSVM classifier. Our following experiments are based on it.

Table 4 is detailed report about VPRSVM classification result. The first column in Table 4 shows the thresholds of σ to determine the lower approximation and boundary region for each category where confidence interval of quantile ρ is 0.95. Second and third column show the error coverage rate of category c_i (denoted by EC_{c_i}) and classification quality, where

Table 3
The distribution of error recall rate.

#Feature	ER_2 (%)	ER_3 (%)	ER_4 (%)	ER_5 (%)	ER_6 (%)	ER_7 (%)	ER_8 (%)	ER_9 (%)	ER_{10} (%)	ER_{11} (%)	ER_{12} (%)
2000	62.38	14.36	7.43	2.97	2.97	2.48	1.98	1.98	0.99	0.50	1.98
2500	61.19	13.93	7.46	4.48	2.49	2.99	2.49	1.00	1.49	0.00	2.49
3000	60.68	14.08	7.28	6.31	2.43	1.94	2.43	0.97	0.97	0.49	2.43

Table 4
Thresholds of σ, corresponding error coverage rate, classification quality and *Micro-average-F_1* for each category (for feature number = 1500).

Category	$\theta^2_{c_i}$	EC_{c_i} (%)	$K_\beta(c_m)$ (%)	*Micro-average-F_1*	
				Lower approximation (%)	Boundary region (%)
Career	1.4295	79.41	41.38	95.12	81.61
Sport	0.2771	76.19	95.29	99.27	20.83
Medical	0.7113	92.45	71.95	97.94	61.60
Region	0.1144	25.93	79.17	64.15	32.26
Entertainment	1.0965	95.71	44.07	94.00	77.61
Estate	0.7198	91.67	65.50	96.18	74.15
Education	1.0458	94.29	61.33	94.16	58.95
Car	1.0041	86.96	60.22	95.07	78.79
Computer	1.1373	90.91	78.66	97.60	71.96
Science	0.9498	86.02	42.90	89.97	60.45
Art	0.1578	29.31	78.42	71.65	26.09
Economy	0.1427	37.59	83.04	82.54	28.76
Overall	–	**75.8**	**68.6**	**95.22**	**64.99**

$$EC_{c_i} = \frac{card(\text{misclassified texts in boundary region})}{card(\text{texts in boundary region})} \times 100\% \tag{33}$$

The last two columns report the *Micro-average-F_1* of lower approximation and boundary region for each category, respectively.

Fig. 6 illustrates the dependence of the VPRSVM classification performance in terms of *Micro-average-F_1* with the feature number on two approximation regions. An observation can be drawn: both regions can obtain stable performance quickly. When feature number is 1500 which reduction degree is 98.02% (total feature number is 75,916 from training set), both regions achieve peak performance. Especially, overall *Micro-average-F_1* of lower approximation region is 95.22%, that is to say, less than 5% texts cannot be recognized.

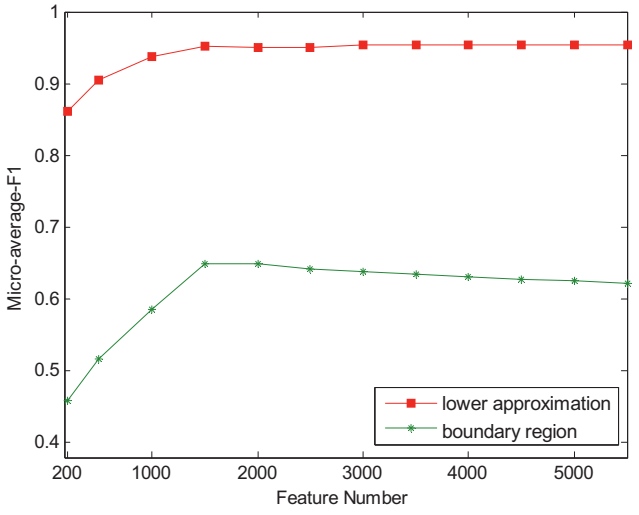

Fig. 6. Different region classification performance of VPRSVM with the feature number.

Manual analysis of misclassified texts belong to lower approximation region found that most of them have multi-category information and it is almost impossible to be revised correct. The accuracy of lower approximation forms the bottleneck for classification. On the contrary, we should focus on boundary region which is the decisive factor for classifier performance.

Experiment 2. Conduct on two feature selection scheme using RKNN.

In order to analyze the influence of skewed problem, two feature selection scheme: global and local are conducted on boundary region texts. The detailed classifier results of different pair of candidate category are as tabulated in Table 5.

For each cell contains three rows of data: First row is the accuracy of global scheme, second is accuracy of local scheme, and third row is correspond ratio (#majority category: #minority category). We note that the accuracy only focused on the text which have possible be classified into the correct category by RKNN classifier. If the manual labeled category not included in the candidate category, the text will never be classified correct and it is out of our consideration.

The observation of Table 5 indicates that: at most of the time, the global scheme is outperform the local scheme. But if the ratio is large enough, the results go by contraries (the data in bold). Inspired by the observation, a simple rule for make the best decisions on feature selection scheme be defined by rules (R1) and (R2).

(R1) If *ratio* $\geqslant$ 5.0, adopting global Fisher criterion scheme for RKNN feature selection.
(R2) If *ratio* < 5.0, adopting local Fisher criterion scheme for RKNN feature selection.

Experiment 3. Performance comparison.

To evaluate the efficiency of our proposed RKNN classifier in boundary region text classification, several extensively used algorithms, i.e., KNN and SVM, are implemented for comparison. Fig. 7 illustrates the *Micro-average-F_1* value with the feature number. Those curves were obtained by using 200, 500, 1000, 1500, 2000, 2500, ..., 10,000 features. For KNN and RKNN classifier, the value of k is 20. The observation in Fig. 7 indicated that: at most of the time, RKNN algorithm with Fisher criterion-based feature selection method outperforms the other methods.

Another interesting observation can be found that Fisher criterion based feature selection method have relatively poor performance with SVM and KNN algorithm, but excellent with RKNN algorithm. The reason to explain this observation is that the candidate category are limited in only two most hopeful classes, it can reduce the influence of unrelated features. Employed in RKNN algorithm fully exhibits the merit of Fisher criterion based feature selection method.

Meanwhile, the overall test accuracy of the whole dataset is also presented. To evaluate efficiency of the proposed algorithms, we used TanCorp dataset to allow a fair comparison with the other traditional learning machines. Different values of parameters have been tried on each algorithm to ensure that the experimental results faithfully reflect the performance of the algorithms. For KNN classifier, the value of k varies from 5 to 30 with step 5. The

Table 5
Accuracy comparison of global and local feature selection scheme on pairs of candidate category, and it's corresponding ratio.

	0	1	2	3	4	5	6	7	8	9	10
0Care.											
1Spor.	100%										
	100%										
	(4.6:1)										
2Medi.	90.91%	100%									
	81.82%	87.50%									
	(2.3:1)	(2.0:1)									
3Regi.	100%	–	**66.67%**								
	75%	–	**100%**								
	(4.1:1)	(18.7:1)	**(9.4:1)**								
4Ente.	100%	93.94%	100%	**62.50%**							
	95.83%	90.91%	94.12%	**75.00%**							
	(2.5:1)	(1.9:1)	(1.1:1)	**(10.0:1)**							
5Esta.	100%	100%	87.50%	85.71%	96.55%						
	100%	100%	83.33%	85.71%	93.10%						
	(1.5:1)	(3.0:1)	(1.5:1)	(6.2:1)	(1.6:1)						
6Educ.	95.45%	100%	85.14%	0	88.89%	–					
	93.18%	88.89%	85.14%	0	83.33%	–					
	(1.3:1)	(3.5:1)	(1.7:1)	(5.4:1)	(1.9:1)	(1.2:1)					
7Car	100%	100%	**75.00%**	100%	100%	100%	92.30%				
	100%	85.71%	**100%**	100%	50.00%	88.89%	92.30%				
	(1.0:1)	(4.8:1)	**(2.4:1)**	(3.9:1)	(2.5:1)	(1.6:1)	(1.4:1)				
8Comp.	**95.12%**	90.91%	94.74%	**66.67%**	88.10%	100%	76.67%	**95.83%**			
	97.57%	86.36%	89.47%	**100%**	83.33%	95.24%	63.33%	**100%**			
	(4.7:1)	(1.0:1)	(2.0:1)	**(19.1:1)**	(1.9:1)	(3.1:1)	(3.6:1)	**(4.9:1)**			
9Scie.	100%	100%	79.38%	**85.71%**	82.35%	66.67%	88.89%	88.89%	87.32%		
	100%	85.71%	76.29%	**100%**	76.47%	33.33%	77.78%	77.78%	87.32%		
	(1.7:1)	(2.7:1)	(1.4:1)	**(6.9:1)**	(1.4:1)	(1.1:1)	(1.3:1)	(1.8:1)	(2.8:1)		
10Art	100%	100%	100%	100%	81.72%	90.00%	70.00%	100%	71.43%	37.50%	
	100%	100%	66.67%	100%	80.65%	70.00%	50.00%	50%	64.29%	25.00%	
	(1.1:1)	(5.1:1)	(2.6:1)	(3.6:1)	(2.8:1)	(1.7:1)	(1.5:1)	(1.1:1)	(5.3:1)	(1.9:1)	
11Econ.	77.78%	50.00%	83.33%	**50%**	100%	100%	75.00%	84.62%	81.58%	45.00%	100%
	72.22%	50.00%	66.67%	**100%**	93.33%	0	62.5%	69.23%	80.26%	45.00%	100%
	(1.4:1)	(3.4:1)	(1.7:1)	**(5.5:1)**	(1.8:1)	(1.1:1)	(1.0:1)	(1.4:1)	(3.5:1)	(1.8:1)	(1.5:1)

Micro-average-F_1 value of each algorithm are presented in Table 6. The overall performance of proposed VPRSVM–RKNN is 94.11%, which is approximately 3.82% higher than KNN, and 1.06% higher than that of SVM algorithm. Consequently, the VPRSVM–RKNN is better than SVM and beats KNN algorithm.

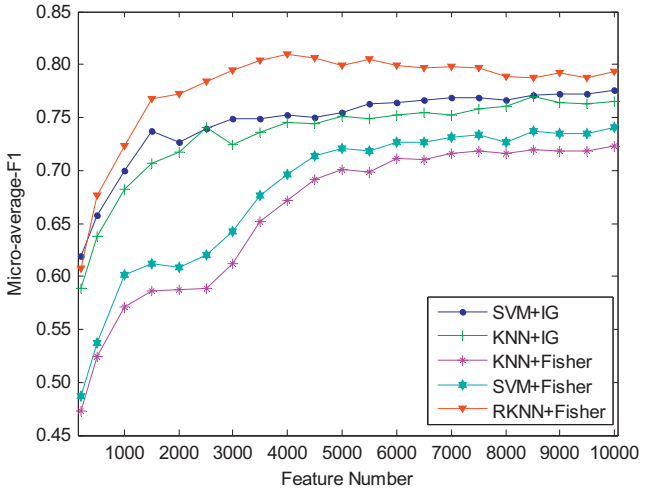

Fig. 7. Performance comparison of different classification algorithms and feature selection methods on boundary region text.

Table 6
Comparison of the efficiency of different algorithms.

Algorithm	Rocchio	KNN	Winnow	NB	SVM	VPRSVM–RKNN
Micro-average-F_1	88.97%	90.29%	86.42%	91.32%	93.05%	94.11%

7. Conclusion and future work

In this paper, a two-level hierarchical VPRSVM–RKNN algorithm by combining the strengths of SVM and KNN classifier based on variable precision rough sets is proposed to deal with text classification problem.

VPRSVM has two roles. One is to filter the noisy data, which can reduce the impact on subsequent RKNN classifier; the other role is to partition the feature space into different regions. We also present a practical approach to obtain the partition of feature space pragmatically. Then, RKNN algorithm is employed to reclassification texts in boundary region. The proposed algorithm overcomes the problem of sensitive to noises of SVM and low efficiency of KNN. A series of experiments on Chinese benchmark data—TanCorp— show that the VPRSVM estimates noisy data and partition feature space effective, and the proposed algorithm outperforms state-of-the-art machines learning methods.

The future work should be done on the issues of optimization combination of different methods. It is also necessary to conduct experiments on some other English benchmarks to verify its adaptability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 60775036, 60970061) and the Ph.D. programs Foundation of Ministry of Education of China (No. 20060247039).

References

Bennett, P. N., Dumais, S. T., & Horvitz, E. (2005). The combination of text classifiers using reliability indicators. *Information Retrieval, 8*(1), 67–100.

Beynon, M. J. (2004). The elucidation of an iterative procedure to β-reduct Selection in the variable precision rough sets model. In *Proceedings of the 4th international conference on rough sets and current trends in computing (RSCTC'04), LNAI 3066* (pp. 412–417).

Beynon, M. J. (2003). The introduction and utilization of (l,u)-graphs in the extended variable precision rough sets model. *International Journal of Intelligent Systems, 18*(10), 1035–1055.

Bottou, L., Cortes, C., Denker, J. S., et al. (1994). Comparison of classifier methods: A case study in handwriting digit recognition. In *Proceedings of the 12th international conference on pattern recognition (ICPR'94)* (pp. 77–87).

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27.

How, B. C., & Kiong, W. T. (2005). An examination of feature selection frameworks in text categorization. In *Proceedings of the second Asia information retrieval symposium (AIRS'05). LNCS 3689* (pp. 558–564).

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Netowrks, 13*(2), 415–425.

Japkowicz, N. (2002). Learning from imbalanced data sets: A comparison of various strategies. In *Proceedings of the learning from imbalanced data sets, AAAI work shop, technical report*.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European conference on machine learning (ECML)* (pp. 137–142).

Lewis, D. (1991). Evaluating text categorization. In *Proceedings of the speech and natural language workshop* (pp. 312–318).

Lingras, P., & Butz, C. (2007a). Precision and recall in rough support vector machines. In *Proceedings of the international conference on granular computing* (pp. 654–658).

Lingras, P., & Butz, C. (2007b). Rough set based 1-v-1 and 1-v-r approaches to support vector machine multi-classification. *Information Science, 177*, 3782–3798.

Lingras, P., Chen, M., & Miao, D. (2009). Rough cluster quality index based on decision theory. *IEEE Transactions on Knowledge and Data Engineering, 21*(7), 1014–1026.

Miao, D., Duan, Q., Zhang, H., et al. (2009). Rough set based hybrid algorithm for text classification. *Expert Systems with Applications, 36*, 9168–9174.

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences, 11*(5), 341–356.

Ramakrishnan, G., Chitrapura, K. P., Krishnapuram, R., et al. (2005). A model for handling approximate, noisy or incomplete labeling in text classification. In *Proceedings of the 22nd international conference on machine learning (ICML'05)* (pp. 681–688).

Rifkin, R., & Klautau, A. (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research, 5*, 101–141.

Saha, S., Murthy, C. A., & Pal, S. K. (2007). Rough set based ensemble classifier for web page classification. *Fundamenta Informaticae, 76*, 171–187.

Silva, C., & Ribeiro, B. (2006). Two-level hierarchical hybrid SVM–RVM classification model. In *Proceedings of the 5th international conference on machine learning and applications (ICMLA'06)* (pp. 89–94).

Srisawat, A., Phienthrakul, & Kijsirikul, B. (2006). SV-KNNC: An algorithm for improving the efficiency of K-nearest neighbor. In *Proceedings of the 9th Pacific Rim international conference on artificial intelligence (PRICAI'06)* (pp. 975–979).

Su, C. T., & Hsu, J. H. (2006). Precision parameter in the variable precision rough sets model: An application. *The International Journal of Management Science, 34*(2), 149–157.

Tan, S. (2006). An effective refinement strategy for KNN text classifier. *Expert Systems with Applications, 30*, 290–298.

Tan, S., Cheng, X., & Xu, H. (2007). An efficient global optimization approach for rough set based dimensionality reduction. *International Journal of Innovative, Computing Information and Control, 3*(3), 725–736.

Tan, S., Cheng, X., Ghanem, M. M., et al. (2005). A novel refinement approach for text categorization. In *Proceedings of the 14th ACM international conference on information and knowledge management (CIKM'05)* (pp. 469–476).

Tang, Y., & Gao, J. (2007). Improved classification for problem involving overlapping pattern. *IEICE Transaction on Information and Systems, E90-D*(11), 1787–1795.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Wang, J., & Zhou, J. (2009). Research of reduct features in the variable precision rough set model. *Neurocomputing, 72*(10–12), 2643–2648.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67–82.

Wu, G., & Chang, E. (2003). Class-boundary alignment for unbalanced dataset learning. In ICML 2003 workshop on learning from unbalanced data sets, Washington, DC.

Yang, Y., & Liu, X.(1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM conference on research and development in the information retrieval (SIGIR'99)* (pp. 42–49).

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th international conference on machine learning (ICML'97)* (pp. 412–420).

Zhang, J., & Wang, Y. (2008). A rough margin based support vector machine. *Information Sciences, 178*, 2204–2214.

Ziarko, W. (1993). Variable precision rough set model. *Journal of Computer and System Sciences, 46*(1), 39–59.