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...
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 report 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. The efficiency of this method may be low. Miao, Duan, and Zhang (2006) proposed a two-class KNN classifier to make the final decision. Rest approach as the baseline SVM classifier.

Following the recommendation of Rifkin et al., we use the One-Vs-Rest SVMs is the class with the highest output value. The RKNN algorithm is also high.

2.1. Text classification technique

2.1.1. 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 $x_i \in \mathbb{R}^d$, $i = 1, \ldots, t$ 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):

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

subject to

$$
\begin{align*}
\langle \omega, x_i \rangle + b & \geq 1 - \xi_i, \quad \text{if } y_i = 1 \\
\langle \omega, x_i \rangle + b & \leq -1 + \xi_i, \quad \text{if } y_i = -1 \\
\xi_i & \geq 0, \quad j = 1, \ldots, t
\end{align*}
$$

Classification is then achieved according to the following function:

$$Q(x) = \text{sign}(\langle \omega, x \rangle + b)$$

where the data was mapped to a higher dimensional space by the function $\varphi$ 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 samples 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.

2.1.2. 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 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 $x_i \in \mathbb{R}^d$, $i = 1, \ldots, t$ 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):

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

subject to

$$
\begin{align*}
\langle \omega, x_i \rangle + b & \geq 1 - \xi_i, \quad \text{if } y_i = 1 \\
\langle \omega, x_i \rangle + b & \leq -1 + \xi_i, \quad \text{if } y_i = -1 \\
\xi_i & \geq 0, \quad j = 1, \ldots, t
\end{align*}
$$

Classification is then achieved according to the following function:

$$Q(x) = \text{sign}(\langle \omega, x \rangle + b)$$

where the data was mapped to a higher dimensional space by the function $\varphi$ 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 $c_j(i = 1, \ldots, r, j = 1, \ldots, n)$, the decision rule in KNN can be written as:

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

where $\text{sim}(x, x_i)$ 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:

$$X_1 \cup X_2 \cup \cdots \cup X_n = U$$

$$X_i \cap X_j = \emptyset \quad (i \neq j)$$  \hspace{1cm} (7, 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

$$X^L_R = \{x|[x]_R \subseteq X\}$$

$$X^U_R = \{x|[x]_R \cap X \neq \emptyset\}$$

$$BN_R(X) = X^U_R - X^L_R$$

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

$$c(X, Y) = 1 - \frac{\text{card}(X \cap Y)}{\text{card}(X)} \quad \text{if } \text{card}(X) > 0$$

$$c(X, Y) = 0 \quad \text{if } \text{card}(X) = 0$$  \hspace{1cm} (12)

where card denotes set cardinality.

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

**Definition 6.** $\beta$-lower approximation, $\beta$-upper approximation and $\beta$-boundary region.

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

$$X^L_\beta = \{x|c([x]_R, X) \leq \beta\}$$

$$X^U_\beta = \{x|c([x]_R, X) < 1 - \beta\}$$

$$BN_\beta(X) = X^U_\beta - X^L_\beta$$  \hspace{1cm} (13, 14, 15)

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

How to choose optimal $\beta$ value is a very difficult task. It often depends on our subjective nature 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 $\beta$ 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 $Q(x) = (\omega)^T \varphi(x) + b$ is a collection of discrimination function based on SVM. The pair $(U, Q)$ is called a modified approximation space, denoted by $K = (U, Q)$.

$$Q(x)$$ is a family of equivalence relations which satisfied the three properties namely reflexive, symmetric and transitive.

**Definition 8.** Modified equivalence class based on discrimination function $Q(x)$.

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

$$[x]_{m, \text{RVM}} = \{x|Q_m(x)Q_m(x) \geq 0\}$$  \hspace{1cm} (16)

**Definition 9.** Modified $\beta$-lower approximation, $\beta$-upper approximation and $\beta$-boundary region.
Let $K = (U, Q)$ be an approximation space, $c_m$ be a subset of $U$ includes documents which are manually labeled to category $c_m$. The sets:

$$c^Q_m = \{ x \in Q(\mathcal{X}_{RPRSVM} - c_m) \leq \beta \}$$

$$c^Q_m = \{ x \in Q(\mathcal{X}_{RPRSVM} - c_m) < 1 - \beta \}$$

$$BN^Q_m(c_m) = c^Q_m - c^Q_m$$

are called modified $\beta$-lower approximation, $\beta$-upper approximation and $\beta$-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 $\beta$-lower approximation set and $\beta$-boundary region. Fig. 1 shows $\beta$-lower approximation set for a classification problem with two classes $c_m$ and $c_j (1 \leq m < n \leq k)$, where the objects have already been mapped into a higher dimensional space by mapping function $\phi$.

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

$$K_{\beta}(c_m) = \frac{\text{card}(c^Q_m)}{\text{card}(c_m)} \quad 1 \leq m \leq k$$

and the overall classification quality is

$$K_{\beta} = \frac{\sum_{m=1}^{k} \text{card}(c^Q_m)}{\text{card}(U)}$$

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, $\beta$-lower approximation and $\beta$-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. Thus, 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 $y(x_i, c_j) = \{ 1 \ x_i \in c_j \ 0 \ x_i \notin c_j \}$ 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:

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

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

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.

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

Given $k$ classes $c_1, c_2, \ldots, c_k$, the following matrices are defined: 

**Between-class scatter matrix**

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

**Within-class scatter matrix**

$$S_W(t) = \frac{1}{n} \sum_{i=1}^{k} \sum_{x_i \in c_i} (x(t) - m_i(t))^2$$
where \( t \) is the term of candidate features; \( m_j(t) \) is mean value of \( t \) over class \( c_j \); \( m(t) \) 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 analyze 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 \( \beta \). Given the value of \( \beta \), three regions are obtained adoption of VPRSVM algorithm. Despite their diverse applications in many domains, it is difficult to find out the optimal values and the values are often given subjectively. To the best of our knowledge, existing approaches adopted unified \( \beta \) 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 \( \beta \) 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 \( \beta \), 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 \( x_t \) is a testing sample, \( Q(x) = (\omega)^T \varphi(x) + b \) is a collection of discrimination function based on variable precision rough SVM. Assuming that \( Q_{n_1}^m(x_t) \geq Q_{n_2}^m(x_t) \geq \cdots \geq Q_{n_k}^m(x_t) \), \( l, m, n \in \{1, 2, \ldots, k\} \), 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 \( Q_{n_1}^r(x) \) 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 \( x_t \), \( Q^r = \{ Q_{n_1}^r(x_t), Q_{n_2}^r(x_t), \ldots, Q_{n_k}^r(x_t) \} \), \( r \geq 2 \), the standard deviation \( \sigma \) of \( Q^r \) is defined as:

\[
\sigma = \sqrt{\frac{1}{r-1} \sum_{i=1}^{r} (Q^r - \bar{Q}^r)^2}
\]

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 \( \sigma \) is lower, the probability of the text misclassified into the first predict class is higher, and it has potential to belong to the \( r \) predict class.

Statistics of the classification results validate the hypothesis by designing frequency histograms for the distribution of misclassified texts of \( \sigma \) 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 \( \sigma \), 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 \( \sigma \) for \( Q^r \) sequence of class \( c_l \) denoted by \( \sigma_{c_l} \) is determined by the confidence interval of quantile \( \rho \) for normal distribution. The optimum thresholds are determined by the following considerations: (1) \( \rho \) is large enough to contain most of the misclassified instances within the boundary region. (2) If \( \rho \) is too large, an amount of correct classification instances will flow into the boundary region and it will result in high computation cost. \( \rho \) is a tradeoff parameter between effectiveness and efficiency.

![Fig. 2. Frequency histogram for the distribution of misclassified texts of \( \sigma \).](image)
Given a definite value of $q$, different tailored value of $b$ is det-

deter-
ing for each category. Here, we note that our approach need not
calculate $b$ precisely, but obtained indirectly. Inspired by Beynon
et al. (2003), we also analyze the variation of approximation re-
gions. Each region is affected by $\rho$ value directly, and the quality
of classification will be influenced accordingly. The geometrical
analysis on 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 algo-

Fig. 3. Normality test for the distribution of misclassified texts of $\rho$.

Given a definite value of $\rho$, different tailored value of $\beta$ is deter-

Fig. 4. Graphical interpretation of ranges of approximation regions for $\rho, \beta$ goes

rithm systematically. The main idea is classify all testing texts
using VPRSVM first, and feature space is partitioned into three re-
gions, i.e., lower approximation, upper approximation, and bound-
ary region, for each class. Optimal determination of boundary
region is obtained from validation on parameter tuning subset.

In order to determine hard-to-classify boundary region texts
will then adopting the RKNN classifier. With the recognition of pre-
vious VPRSVM, boundary region texts also narrow the scope of dis-

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
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, 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).

\[
\begin{align*}
TP (\text{True Positive}) &: \text{the number of documents correctly classified to that class.} \\
TN (\text{True Negative}) &: \text{the number of documents correctly rejected from that class.} \\
FP (\text{False Positive}) &: \text{the number of documents incorrectly rejected from that class.} \\
FN (\text{False Negative}) &: \text{the number of documents incorrectly classified to that class.}
\end{align*}
\]

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.

\[
\begin{align*}
pr(c_i) &= \frac{TP(c_i)}{TP(c_i) + FP(c_i)} \quad (28) \\
re(c_i) &= \frac{TP(c_i) + FN(c_i)}{TP(c_i)} \quad (29) \\
F_1(c_i) &= \frac{2 \cdot pr(c_i) \cdot re(c_i)}{pr(c_i) + re(c_i)} \quad (30)
\end{align*}
\]

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:

Micro-average-\(F_1 = \frac{\sum_{i=1}^{k} TP(c_i)}{\sum_{i=1}^{k} (TP(c_i) + FN(c_i))} \quad (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 \(ER_i, 2 \leq i \leq k\), is defined as follows:

\[
ER_i = \frac{\text{card(misclassified texts belong to the i\textsuperscript{th} predict class)}}{\text{card(misclassified texts)}} \times 100\% \quad (32)
\]

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

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

**Table 2** Contingency table for a set of binary decision.

<table>
<thead>
<tr>
<th>Class c(i)</th>
<th>System classification</th>
<th>(\text{Belong})</th>
<th>(\text{Not belong})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{Belong})</td>
<td>(TP(c_i))</td>
<td>(FP(c_i))</td>
<td>(FN(c_i))</td>
</tr>
<tr>
<td>(\text{Not belong})</td>
<td>(FN(c_i))</td>
<td>(TN(c_i))</td>
<td>(FP(c_i))</td>
</tr>
</tbody>
</table>

1 Available at http://www.searchforum.org.cn/tansongbo/corpus.htm.
Table 3
The distribution of error recall rate.

<table>
<thead>
<tr>
<th>#Feature</th>
<th>$E_R(%)$</th>
<th>$E_A(%)$</th>
<th>$E_M(%)$</th>
<th>$E_B(%)$</th>
<th>$E_L(%)$</th>
<th>$E_{LR}(%)$</th>
<th>$E_{BR}(%)$</th>
<th>$E_{LR}(%)$</th>
<th>$E_{BR}(%)$</th>
<th>$E_{LR}(%)$</th>
<th>$E_{BR}(%)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>62.38</td>
<td>14.36</td>
<td>7.43</td>
<td>2.97</td>
<td>2.97</td>
<td>2.48</td>
<td>1.98</td>
<td>1.98</td>
<td>0.99</td>
<td>0.50</td>
<td>1.98</td>
</tr>
<tr>
<td>2500</td>
<td>61.19</td>
<td>13.93</td>
<td>7.46</td>
<td>4.48</td>
<td>2.49</td>
<td>2.99</td>
<td>2.49</td>
<td>1.00</td>
<td>1.49</td>
<td>0.00</td>
<td>2.49</td>
</tr>
<tr>
<td>3000</td>
<td>60.68</td>
<td>14.08</td>
<td>7.28</td>
<td>6.31</td>
<td>2.43</td>
<td>1.94</td>
<td>2.43</td>
<td>0.97</td>
<td>0.97</td>
<td>0.49</td>
<td>2.43</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Category</th>
<th>$\rho_c^\alpha$</th>
<th>$EC_{ci} (%)$</th>
<th>$K_0(c_{ci}) (%)$</th>
<th>Micro-average-$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Career</td>
<td>1.4295</td>
<td>79.41</td>
<td>41.38</td>
<td>95.12</td>
</tr>
<tr>
<td>Sport</td>
<td>0.2771</td>
<td>76.19</td>
<td>95.29</td>
<td>99.27</td>
</tr>
<tr>
<td>Medical</td>
<td>0.7113</td>
<td>92.45</td>
<td>71.95</td>
<td>97.94</td>
</tr>
<tr>
<td>Region</td>
<td>0.1144</td>
<td>25.93</td>
<td>79.17</td>
<td>64.15</td>
</tr>
<tr>
<td>Entertainment</td>
<td>1.0965</td>
<td>95.71</td>
<td>44.07</td>
<td>94.00</td>
</tr>
<tr>
<td>Estate</td>
<td>0.7198</td>
<td>91.67</td>
<td>65.50</td>
<td>96.18</td>
</tr>
<tr>
<td>Education</td>
<td>1.0458</td>
<td>94.29</td>
<td>61.33</td>
<td>94.16</td>
</tr>
<tr>
<td>Car</td>
<td>1.0041</td>
<td>86.96</td>
<td>60.22</td>
<td>95.07</td>
</tr>
<tr>
<td>Computer</td>
<td>1.1373</td>
<td>90.91</td>
<td>78.66</td>
<td>97.60</td>
</tr>
<tr>
<td>Science</td>
<td>0.9498</td>
<td>86.02</td>
<td>42.90</td>
<td>89.57</td>
</tr>
<tr>
<td>Art</td>
<td>0.1578</td>
<td>29.31</td>
<td>78.42</td>
<td>71.65</td>
</tr>
<tr>
<td>Economy</td>
<td>0.1427</td>
<td>37.59</td>
<td>83.04</td>
<td>82.54</td>
</tr>
<tr>
<td>Overall</td>
<td>–</td>
<td><strong>75.8</strong></td>
<td><strong>68.6</strong></td>
<td><strong>95.22</strong></td>
</tr>
</tbody>
</table>

The formula to calculate $EC_{ci}$ is: $EC_{ci} = \frac{\text{card}(\text{misclassified texts in boundary region})}{\text{card}(\text{texts in boundary region})} \times 100\%$ (33)

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.

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 $\geq 5.0$, 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
Micro-average-F1 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.

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 reclassify 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—Tan-Corp— 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, 60870061) and the Ph.D. programs Foundation of Ministry of Education of China (No. 20060247039).

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

<table>
<thead>
<tr>
<th>0Care.</th>
<th>1Spor.</th>
<th>2Medi.</th>
<th>3Regi.</th>
<th>4Ente.</th>
<th>5Esta.</th>
<th>6Educ.</th>
<th>7Car</th>
<th>8Comp.</th>
<th>9Scie.</th>
<th>10Art</th>
<th>11Econ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>100%</td>
<td>100%</td>
<td>90.91%</td>
<td>100%</td>
<td>90.91%</td>
<td>95.83%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>77.78%</td>
</tr>
<tr>
<td>Ratio</td>
<td>(4.6:1)</td>
<td>(4.8:1)</td>
<td>(2.3:1)</td>
<td>(2.1:1)</td>
<td>(1.1:1)</td>
<td>(10:1)</td>
<td>(1.0:1)</td>
<td>(1.1:1)</td>
<td>(1.1:1)</td>
<td>(1.1:1)</td>
<td>(1.1:1)</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Rocchio</th>
<th>KNN</th>
<th>Winnow</th>
<th>NB</th>
<th>SVM</th>
<th>VPRSVM–RKNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro-average-F1</td>
<td>88.97%</td>
<td>90.29%</td>
<td>86.42%</td>
<td>91.32%</td>
<td>93.05%</td>
<td>94.11%</td>
</tr>
</tbody>
</table>

References


