*Send Orders for Reprints to reprints@benthamscience.ae*





# A Classification Algorithm for Chinese Verb Phrases Using Support Vector Machine

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

[1]*Department of Computer Science & Technology, Xinzhou Teachers University, No. 10, Heping west street, Xinzhou City, 034000, China*

[2]*College of Computer Science & Technology, Taiyuan University of Technology, No. 72, Yingze west street, Taiyuan City, 030024, China*

**Abstract:** Chinese verb phrases classification is to determine boundaries of verb phrases and divide them exactly, using brackets, by automatically analyzing and processing by computer after the sentences have been decollated and marked the characteristic or property of a certain word. SVM classification model is a common and powerful for classification tasks. In this paper, the SVM classification model is built by extracting static features and dynamic features of Chinese verb phrases, and an algorithm to perform Chinese verb phrases classification using support vector machine is proposed. Using 3500 sentences to train and test, experiment results show that the SVM model dramatically reduces the training time and steps. Compared with the method proposed in literature 15, classification precision rate is increased by approximately 8.0% using the algorithm in this paper, which fully illustrates that the performance of the proposed algorithm is superior classification algorithm.

**Keywords:** Feature extraction, machine learning, support vector machine, vector space model, verb phrase classification.

## 1. INTRODUCTION

With the rapid development of computer technology and the popularization of computers for nearly a dozen of years, the Internet has entered the homes and the on-line information is playing an increasingly important role in people's work, study and life. People increasingly need technologies' help such as search engines, machine translation, information retrieval and so on for work, educational and entertainment tasks. How to implement the computer auto-understanding of language and natural language processing has become an important research area in computer science fields. Natural language processing involves multi-level hierarchy such as speech tagging, phrase tagging, phrase word segmentation analysis and semantic understanding, etc [1, 2]. The task at each level requires the intervention of the formal language knowledge. The natural language processing has experienced word processing and vocabulary processing stages and it has achieved satisfactory results both in theory and in practice, and now has entered the stages of phrase processing and syntactic analysis. For Chinese, phrases have special important position. Its internal structure is stable, often interaction occurs with other ingredients in the sentences as a whole. And its constructed principles are consistent with the constructed principles of sentences. D. X. Zhu think, that if we have described structure and function of various phrases so clearly and in detail, structure of sentences can actually also be described clearly, because sentences are just the independent phrases [3]. In this sense, research on Chinese phrase recognition has the higher theoretical and practical value. Verb phrases are the most important and main form in Chinese phrases. Problems encountering in syntactic analysis such as syntactic ambiguity etc., also exist in verb phrase researches. Therefore, research on Chinese verb phrase automatic recognition is very valuable.

Chinese verb phrases classification is determined by boundaries of verb phrases and dividing them exactly using brackets by automatically analyzing and processing by computation after the sentences have been decollated and marked the characteristic or property of a certain word [2]. Because it involves a variety of complex issues, such as Chinese phrase structural analysis and phrases disambiguation, it is a very difficult piece of research work. However, it is the successful completion of automatic recognition of Chinese verb phrases that will be of great theoretical and practical significance for further syntactic analysis, information retrieval, machine translation, corpora studies.

SVM is a small sample learning method based on statistical learning theory in the mid 1990 by Vapnik et al. in Bell Lab, with a rigorous theoretical basis [4-7, 11]. Based on structural risk minimization criteria, it has stronger generalization ability and can better solve the practice problems such as the small sample, non-linear, highly practical problems and the local minimum dimension, thus it becomes one of

1874-4443/14 

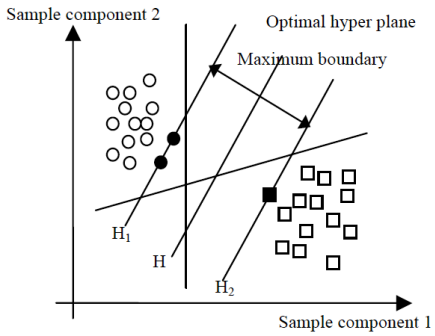


**Fig. (1).** Optimal classification face.

the fastest research directions of the recent developments in the field of machine learning. Although with recent international attention on SVM study, the research on Chinese phrases classification in this field is still in early stages. Therefore SVM classifier must be improved and expanded.

This paper put forwards a Chinese verb phrases classification algorithm based on support vector machine, and the experiment proves the validity of this algorithm. It can greatly reduce the steps and time that classifier classification learning needs, has an advantage on the data scale and algorithm complexity, and is a better classification algorithm.

## 2. THEORY ON SUPPORT VECTOR MACHINE

Support vector machine algorithm dates from statistical learning theory. The algorithm is based on structural risk minimization principle, compressing the raw data collection to support vector set (usually is the former's 3%-5%), then learn to get the classification decision function. The basic idea is to construct a hyper plane as a decision-making graphic to make the interval between positive and negative modes maximum.

SVM method is proposed from the optimal classification in linearly separable cases. Being shown in Fig. (**1**), white circle and hollow boxes respectively represent two types of training samples, H is the classification line which does not separate the two types incorrectly. $H_1$ and $H_2$ are respectively the straight line which pass through the points that are the nearest to various types of samples and parallel to classification line. The distance between two lines is called classification interval. According to the experimental risk minimization principle, SVM's actual risk is determined by the formula (1).

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

In formula (1), $R(\omega)$ represents the actual risk, $R_{emp}(\omega)$ represents the empirical risk, $\Phi$ is the confidence interval. Completely separation makes $R_{emp}(\omega) = 0$ and maximum interval makes the minimum range of confidence interval $\Phi$, so that the real risk is minimized.

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

$$g(x) = \omega \cdot x + b = 0 \tag{2}$$

Take the discriminant function $g(x)$ normalized, and make all samples meet with $|g(x)| \ge 1$. Then the classification interval is $2/||\omega||$. The problem is changed into keeping the largest interval on condition that classification line may correctly classify all samples. It is symbolically defined as the formula (3).

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

In the condition of non-linear, some training samples don't accord to the condition of formula (3b). We must add a relaxation item $\varepsilon_i \ge 0$ in left of formula (3b). Minimize formula (3a) is to maximize $\Phi(\omega) = ||\omega||^2 / 2$. In accordance with the Lagrange function and Kuhn-Tucker condition, formula (3) finally can be changed into the formula (4).

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

We get the optimal classification function by solving formula (4) and it is shown as formula (5).

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

We can get $\varepsilon_i \ge 1 - y_i(\omega^T x_i + b)$ by the formula (4b), so the formula (4a) can be redefined as the formula (6).

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

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

The constant Q of formula (4a) is the equilibrium between the generalization ability and training accuracy. SVM has better generalization ability if Q is smaller; SVM has smaller training error if Q is larger. The formula (4b) introduces the slack variable. It allows some points to overstep boundary and increases SVM ability of noise immunity in case of non-separable. Because of slack variables, we define boundary relative to formula (5) for each sample as formula (7).

$$\gamma_i = y_i f(x) \tag{7}$$

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

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

Polynomial kernel function:

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

Radial basis function:

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

Two layer perceptron neural network sigmoid function:

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

Different kernel functions leads to different characteristics space, and so will have a different shape of the sampling distribution. In order to limit sample to transform to much larger feature space, this paper selects RBF kernel function as the classification kernel functions [4-6, 8].

In recent years, research on SVM focuses on the natural study of SVM and perfecting as well as increasing SVM applied research in both depth and breadth. To date, support vector machine has been applied to pattern classification, regression analysis, function estimation, and other fields. Isolated hand writing breaks recognition, page or text automatic classification, speech recognition, face recognition, gender classification, computer invasion detection, gene classification, remote sensing figure like analysis, target recognition, function return, function approximation, density estimation, time sequence forecast and the data compression, text filter, data mining, and nonlinear system control etc., all these problems have successful application of support vector machine.

## 3. CLASSIFICATION ALGORITHMS FOR CHINESE VERB PHRASES USING SVM

### 3.1. Analysis of Chinese Verb Phrases

For a Chinese sentence $H$ containing one or more than one verbs, we can use anthologies to establish the candidate set $can(H)$ of verb phrases. Let set $isvphrase=\{x \mid x \text{ is a verb phrase}\}$, set $\overline{isvphrase}=\{x \mid x \text{ is not a verb phrase}\}$. The problem of automatic classification of Chinese verb phrases in this paper is to judge every phrase $w$ in verb phrase candidate set, $can(H)$ is a verb phrase or not in the current sentence, namely, $w \in isvphrase$ or $w \in \overline{isphrase}$.

Supposing that if we have known $\forall w \in Train$ (Set $Train$ is the word pairs contained in the candidate set of verb phrases of train set), exist $n+1$ tuples $(v_1, v_2, \cdots, v_n, ci)$, in which $v_1, v_2, \cdots, v_n$ are the values of $w$ relative to features $s1, s2, \cdots, sn$, $ci$ is 0 or 1, (0 expresses $w$ is not a verb phrase, 1 expresses $w$ is a verb phrase). The problem to be solved is to determine whether $can'$ is a verb phrase or not according to the feature value $v_1', v_2', \cdots, v_n'$ of $can'$ for $\forall can' \in in$ ($in$ is the verb phrase candidate set of input sentence). We solve the above problem using the following three steps.

(1) Feature extraction: In all features of the candidate verb phrases, extract the meaningful feature for phrase classification.

(2) Establish classification mechanism: Establish a mechanism for automatic classification of verb phrases by learning a training set using SVM according to candidate verb phrases' characteristic values.

(3) Classify verb phrases automatically: For each candidate phrase in the input sentence, according to its characteristic values, judge the phrase is a verb phrase or not using the built verb phrase classification mechanism.

### 3.2. Feature Extraction of Verb Phrases

Whether a word pair in the candidate set can really pose a verb phrase is relevant to both their own property-related syntax and their context. Features of verb phrases are mainly reflected in these two aspects. This paper refers to its grammatical attributes as static features, and contexts with respect to the classification of verb phrases as dynamic features.

#### *3.2.1. Static Feature Extraction*

In the processing of the statistics and analysis of Corpus labor, we find that, copulas such as "is, am, are, become, look, feel, sound etc." often constitute a predicate-object structure verb phrase with noun or noun phrase later in the sentence. Directional verbs such as "come, go, up, down, etc." often constitute a complement structure verb phrases with the front verbs such as "can, would, demand, need, etc." We select such syntax properties as the static features of the verb phrases.

The extraction of static features on candidate word pairs of verb phrases in this study consults the grammatical knowledge dictionary of computational linguistics institute in Peking University. The dictionary has identified 46 properties on verb phrases [9, 10]. These properties can be classi-

fied into seven categories. Among the properties, the first class is on characteristics of verb itself, for example the verb is copula (non-copula), and auxiliary verbs (non-auxiliary verbs), and forms verb (un-forms verb) and trend verb (non-trend verb) and so on. The third class describes whether verbs can take really object (outside within), and take body speech object, and the speech object and associate the speech object (body associate), and take double object (double penn), and become the first verb in the pivotal sentences or not. The forth class describes whether verbs can carry existential object (existential), result complement, and trend complement or not. The fifth class describes whether verbs can be decorated by "not" and "haven't", and whether verbs itself can be decorated by the degree adverbs such as "very, extremely, too, etc.", and whether verbs can add auxiliaries tenses, and whether verb itself or verb with object can be decorated by adverbs "being" and so on. We extract the above 16 items as the static features of a candidate verb phrases. Therefore, the verb's static feature set is as follows.

{copula, auxiliary verb, form verb, trend verb, outside inside, quasi predicate object, double object, pivotal sentence, existential, verb result, verb trend, no, haven't, very, be, being}

#### *3.2.2. Dynamic Feature Extraction*

Static features gives candidate phrases the possibilities of constituting certain structural type of verb phrases, and whether it can really constitute verb phrase in the sentence depends on the context [11, 12]. Thus, dynamic characters can be divided in two types-environmental features on grammatically partial words and environmental features on candidates.

**(1) Environmental Feature Extraction on Grammatically Partial Words**

By the analysis of manual statistical results in the corpus, it can be concluded that, whether the candidate phrase can constitute verb phrase in the sentence is associated with the grammatically-partial words in the context. For example, when a verb is decorated by an adverb, a verb phrase is often formed. Therefore, the environmental feature on grammatically-partial word is determined as follows.

Isadverbfront: Whether there is an adverb in front of the verb to decorate the verb within the specified environment window for current candidate phrase.

Isprepfront: Whether there is a preposition or preposition phrase in front of the verb to decorate the verb within the specified environment window for current candidate phrase.

Isdirefront: Whether there is a directional word in front of the verb to decorate the verb within the specified environment window for current candidate phrase.

Isleback: Whether there is an auxiliary word "le" behind the verb within the specified environment window for current candidate phrase.

Iszheback: Whether there is an auxiliary word "zhe" behind the verb within the specified environment window for current candidate phrase.

Isandback: Whether there is a conjunction "and" behind the verb within the specified environment window for current candidate phrase.

We refer to the above six features as candidate phrases' environmental features on grammatically-partial words.

**(2) Environmental Feature Extraction on Candidates**

Whether a candidate phrase can constitute the verb phrase in the sentence is also relevant with the word property of the words in front of and behind verb. When used with certain types of words, it often constitutes a verb phrase, and when used with different categories of words, it doesn't constitute a verb phrase. Therefore, the environmental feature on candidates is determined as follows.

Isnounback: Whether there is a noun or noun phrase in back of the verb within the specified environment window for current candidate phrase.

Isverbback: Whether there is a verb in back of the verb within the specified environment window for current candidate phrase.

Istrendback: Whether there is a trend verb in back of the verb within the specified environment window for current candidate phrase.

We refer to the above three features as environmental features on candidates.

In this way, we have determined that the verb's dynamic feature set is as follows.

{isadverbfront, isprepfront, isdirefront, isleback, iszheback, isandback, isnounback, isverbback, istrendback}

In this paper's experiment, we use the extracted 25 features as the feature values of candidate phrases.

### 3.3. Classification Algorithm of Chinese Verb Phrase Using SVM

#### *3.3.1. Vector Space Model of Verb Phrases*

The basic idea of vector space model is using feature vectors $(w_1, w_2, \cdots, w_n)$ of text to represent text, in which $w_i$ is the weight of the $i^{\text{th}}$ feature item. The key step for classification based on vector space model is how to extract effective features for text [12]. We can select a character, word, or phrase as text feature, but it was generally felt that selecting words as features are better than characters and phrases. We have extracted the 25 feature properties for verb.

In this paper, if the total number of features of candidate phrases are $n+1$, in which $n$ is the number of feature properties, 1 is category tag, a vector space of $n+1$ dimensions is constituted. Each candidate phrase is expressed as a $n+1$-dimension vector $(w_1, w_2, \cdots, w_n, y)$. Vector's component in

each dimension is the weight of the feature in the candidate phrase. Weight is the corresponding features' critical element in the candidate phrases. In this paper, the total number of extracted features of candidate phrase verbs is 25, which are divided into 3 classes. According to the importance, weight should be treated differently. The first class is 16 static properties being extracted based on "the dictionary of modern Chinese grammar information" of Peking University, which are the most important for verb phrases' identification, so the 16 features' weights should be the largest. The extracted 3 candidate environmental features are description of verb phrases' context information, which are also important for recognition of verb phrases. But the extracted 6 environmental features on grammatically-partial words are the results of objective observation, which only serve local modifications for the verb and don't play a key role on verb phrase recognition, and the weights should be the smallest.

Let property flag sets are S1, S2, S3, where S1 is the static property set, S2 is environmental property set on candidates, S3 is environmental property set on grammatically-partial words. Then:

Weight set $W = \{w_x \mid x \text{ is } feature\ tag,\ x \in s1 \vee x \in s2 \vee x \in s3\}$, $w_x$ is the weight of features $x$. Through experiments, the calculation method of three kinds of feature weights is as follows.

(1) If the verb hasn't the 25 features, its weight is 0.

(2) If $x \in s1$, $w_x = 0.6$;

If $x \in s2$, $w_x = 0.6 \times \lambda$;

If $x \in s3$, $w_x = 0.6 \times \lambda \times \lambda$.

Where, $\lambda$ is an adjustment parameter, through experiments, $\lambda = 0.5$.

This method of calculating the weight reflects a different importance level of the various features to the verb phrase recognition and is feasible.

### *3.3.2 Construction of SVM Classifier*

Supposing that classification problems' training set is $E = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, Where $x_i \in R^n$ is feature property weight of the candidate verb phrase, $y_i \in \{-1, +1\}$ is classification value. If $y_i = -1$, It means that the $x_i$ cannot constitute a verb phrase with its back word or phrase. If $y_i = +1$, It means that the $x_i$ can constitute verb phrase with its back word or phrase. The problem to be solved first in this paper is to decide whether the verb can constitute verb phrases with the front or back word or phrase using minimum error probability for any candidate verb phrases.

Using SVM to make classification, the first thing is to extract features from the original space, map samples of the original space into vectors of high-dimensional feature space to solve linear inseparable issues in the original space. This paper researches two-classification problem on verb phrase identification, for which the original space is linearly inseparable, and its input space is the point constituted by the verb. In order to use SVM to solve this problem, we need to find a mapping $\phi$, namely the kernel function mentioned earlier, to make set $\{z_i \mid z_i = \Phi(x_i) \in R^n, i = 1,2,...,l\}$ linearly separable. However, the kernel function is generally not sunk out. People generally use Mercer theory of function analysis and experiment experiences to put forward some common kernel functions described earlier. Selection of kernel functions is important, but currently there is not an accepted selection criteria, and people mainly use the experiment method to select the kernel function. For our experiment, we choose the RBF kernel function $K(x, x_i) = \exp(-\sigma \left\| x - x_i \right\|^2)$ because it has a significant statistical significance and very high classification accuracy. Where, $\sigma$ is the adjustable parameter of the kernel function. If a vector $x$ is SVM' support vector, it means that samples near $\sigma$ are within the larger range and are the same class with $x$ when $\sigma$ is smaller. Conversely, it means that samples near $\sigma$ are within the smaller range and are the same class with $x$. Therefore the selection of $\sigma$ have some effects on classification. The algorithm of constructing SVM Classifier is as follows.

Input: train file-trainfile.txt

Output: $(\omega, b)$

(1) Use the radial basis function $K(x, x_i) = \exp(-\sigma \left\| x - x_i \right\|^2)$ to get $z_i$. ($\sigma = 1.65$)

(2) Use the formula $\max W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j)$ to solve $\alpha$.

(3) Use the formula $\omega = \sum_{i=1}^{l} y_i \alpha_i z_i$ to get $\omega$.

(4) Use the formula $| f_{w,b}(z_i) | = 1$ to get $b$.

### *3.3.3. Classification Algorithm of Verb Phrases Based on SVM*

Input: test file-testfile.txt

Output: identification file--outcomefile.txt

(1) i=1.

(2) while ! (eof (testfile.txt )) do

(i) Select sample $x_i$, use $z_i = \varphi(x_i)$ to get $z_i$.

(ii) Compute the value of the objective function $f(z) = \mathrm{sgn}[\sum_{i=1}^{l} y_i \alpha_i (z \cdot z_i) + b]$.

(iii) If $f(z) = 1$ then

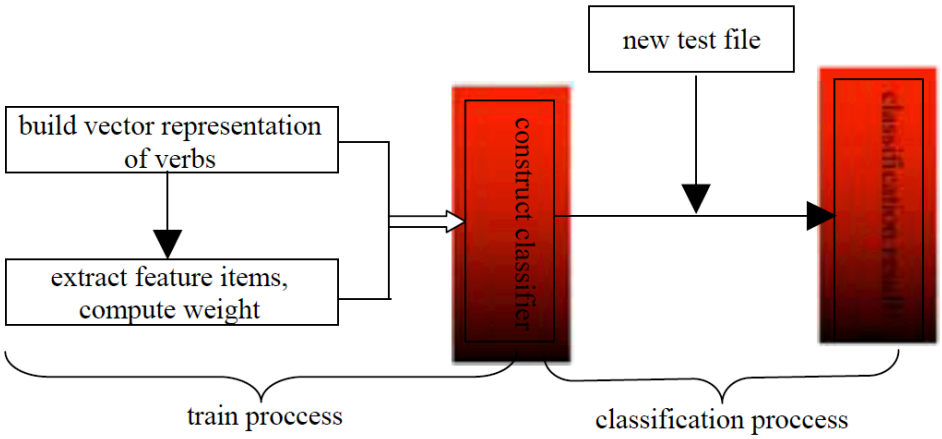


**Fig. (2).** The description of system.

$x_i$ can constitute verb phrases within the specified window.

else

$x_i$ can't constitute verb phrases within the specified window.

(iv) i=i+1.

(3) stop.

The description of system architecture is shown in Fig. (**2**).

## 4. EXPERIMENTS AND RESULTS

Institute corpus of this study is derived from Internet and includes 40 thousand words. Word segmentation and POS tagging has already been carried out. Normally, we have found it does not form verb phrases striding across punctuation. So we do pre-processing for corpus. Specific methods are as follows: use the punctuation marks as the unit to disconnect the sentences to constitute a smaller syntactic units, thus a sub-corpus is formed [13], which can greatly improve the accuracy of verb phrase recognition. These punctuation marks include: commas, semicolons, colons, periods, question marks, exclamation points and ellipses etc.

Phrase classification is fundamentally a mapping process, so the mark of estimating the phrase classification system is the mapping accuracy and speed. Mapping speed depends on the complexity of the mapping rules, and the reference of estimating the mapping accuracy is the classification result through the thought of expert. The more the result is similar to artificial classification results, the higher the classification accuracy is [13]. This implies the two indicators of assessing text categorization system: precision and recall [14]. This paper uses the following indicators to measure the performance of the system: classification accuracy (precision), recall rate (recall), error rate (error) and leakage rate (leakage). The definition is as follows.

$a$ =total number of verb phrases identified correctly

$b$ =total number of non-recognized verb phrases.

$c$ =total number of verb phrases identified incorrectly.

$$\text{recall} = \frac{a}{a+b} \times 100\%$$

$$\text{precision} = \frac{a}{a+c} \times 100\%$$

$$\text{leakage} = \frac{b}{a+b} \times 100\%$$

$$\text{error} = \frac{c}{a+c} \times 100\%$$

recall+ leakage= 1

precision+ error= 1

For a test text, all verb phrases that it may contain are divided into two parts $a$ and $b$. The error recall number of verb phrases of classification system is $c$, and due to its greatest influence on system performance, the error rate must be made sure to be minimized. Experimental system is shown in Fig. (**3**).

The system selects 2500 sentences marked verb phrases as training set to construct SVM which contains 5136 verb phrases, and make closed test. And we use other 1000 sentences containing 2089 verb phrases to make open test. The test results are as follows in Table **1** and Table **2**.

In addition, in order to verify the performance of the proposed algorithm furthermore, we use the method proposed in literature [15] to classify the above 1000 sentences containing 2089 verb phrases. The classification precision rate is about 75.4%. However, the classification precision rate reaches 83.5% using the proposed method in this paper when $\sigma$ is 1.4. In contrast, classification precision rate is increased by approximately 8% using our method.

After the test and analysis, the following conclusions can be made: (1) The system's classification precision rate and recalled rate are higher, errors rate and leakage rate are lower on the condition of closed test and open test. Experiments have obtained ideal effect. (2) Closed test's precise rate and

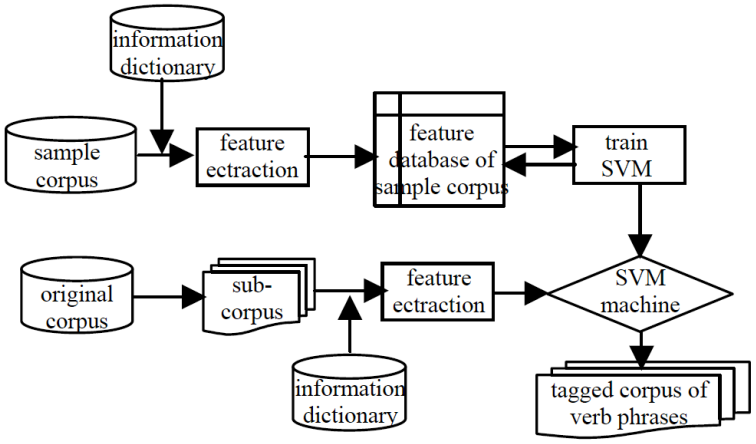


**Fig. (3).** Experimental system of verb phrases classification using SVM.

**Table 1. Results of closed test and open test.**

| Test Type | $\sigma$ | Precision | Recall | Error | Leakage |
|---|---|---|---|---|---|
| closed | 1.8 | 86.7% | 84.9% | 13.3% | 15.1% |
| open | 1.8 | 83.4% | 80.7% | 16.6% | 19.3% |

**Table 2. $\sigma$ value test.**

| $\sigma$ | Test Type | Precision | Recall |
|---|---|---|---|
| 2.2 | Closed | 76.2% | 75.1% |
| 1.4 | Closed | 83.5% | 81.7% |
| 1.1 | Closed | 75.5% | 72.4% |
| 0.9 | Closed | 74.8% | 71.6% |

recalled rate are higher than open test's if other conditions are the same. (3) In the case of all other things being same, if the value of $\sigma$ is not same, system identification accuracy rate and recall rate is different.

## 5. CONCLUSION AND FUTURE WOKS

This paper presents classification algorithm of Chinese verb phrases using SVM based on the in-depth analysis and research of SVM classifier, and carries out automatic classification of Chinese verb phrases in the environment which is closer to the real environment. Experimental verification of these works is carried out and results are shown to be very effective.

At present, we have used the proposed algorithm to implement a text classification system, which can carry out verb phrase classification for plain text, and we have used a large amount of data to make tests, and achieved ideal results. In the future, we will continue to improve the performance of the system and attempt to combine with other valid phrase classification methods and further improve classification precision rate and recall rate of the system.

## CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

## ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China under Grant No. 61202163 and by the Natural Science Foundation of Shanxi Province under Grant No.2012011011-5 and No. 2013011017-2 and by the Technology Innovation Project of Shanxi Province under Grant No. 2013150 and by the Key disciplines supported by Xinzhou Teachers University under Grant No. XK201308. The authors are grateful for the constructive and valuable comments made by the many expert reviewers.

## REFERENCES

[1] D. X. Zhu, Ambiguity phenomenon in Chinese syntax. The Commercial Press: Beijing, 1980.

[2] M. A. K. Halliday, "An Introduction to Functional Grammar (3rd ed)". Foreign Language Teaching and Research Press: Beijing, 2008.

[3] F. Zamora-Martinez, M. J. Castro-Bleda, and S. Espana-Bquera, "A connectionist approachto Part-of-Speech tagging," In: *Proceedings of 1st International Joint Conferenceon Computational Intelligence (IJCCI 2009)*, Funchal, Portugal, pp. 421-426, 2009.

[4] N. Cristianini, and J. Shawe-Talor, "An introduction to support vector machines," *Beijing: Publishing House of Electronics Industry*, pp. 82-98, 2004.

[5] C. W. Hsu, and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415-425, 2002.

[6] O. Chapelle, V. N. Vapnik, and O. Bousquet, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 131-159, 2002.

[7] K. Takeya, and Y. Lepage, "Marker-based chunking for analogy-based translation of chunks," In: *Proceedings of the 2011 MT Summit 13*, Xiamen, China, pp. 338-345, 2011.

[8] Y. C. Wu, Y. S. Lee, and J. C. Yang, "Robust and Efficient Mul tic lass SVM Models for Phrase Pattern Recognition," *Pattern Recognition*, vol. 41, pp. 2874-2889, 2008.

[9] H. Y. Huang, and X. F. Zhang, "Part-of-Speech tagger based on maximum entropy model," In: *Proceedings of 2nd IEEE International Conference on Computer Science and Information*, Beijing, China, pp. 26-29, 2009.

[10] J. H. Xiao, X. L. Wang, and B. Q. Liu, "The study of a nonstationary maximum entropy markov modeland its application on the postagging task," *ACM Transactions on Asian LanguageInformation Processing*, vol. 6, no. 2, pp. 1-8, 2007.

[11] L. C. Yuan, "Improvement for the automatic part-of-speech tagging based on hidden Markov model," In: *Proceedings of 2nd International Conference on Signal ProcessingSystems (ICSPS 2010)*, Dalian, China, pp. 744-747, 2010.

[12] Y. S. Wang, "Research on part-of-speech tagging using decision trees in English-Chinese machine translation system," *Computer Engineering and Applications*, vol. 46, no. 20, pp. 99-102, 2010.

[13] J. Doucette, and M. I. Heywood, "GP classification under imbalanced data sets: active sub-sampling and AUC approximation," *Proceedings of the 11th European conference on Geneticprogramming*. Naples, Italy; Springer-Verlag, pp. 266-277, 2008.

[14] A. S. Garc, A. Fern Ndez, and J. Luengo, "A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability," *Soft Computing – A Fusion of Foundations, Methodologies and Applications*, vol. 13, no. 10, pp. 959-977, 2009.

[15] R. Li, J. H. Zheng, and M. Y. Guo, "Application Study of Hidden Markov Model Based on Genetic Algorithm in Noun Phrase Identification," *Computer Science*, vol. 36, no. 10, pp. 244-246, 261, 2009.

Received: September 22, 2014 Revised: November 05, 2014 Accepted: November 06, 2014