Send Orders for Reprints to reprints@benthamscience.ae

474 *The Open Cybernetics & Systemics Journal*, 2014, *8*, 474-478

Open Access

Recognition of Chinese Entertainment News Words Using SVM-based Active Learning Strategy

Cao Jianfang[1,*], Chen Lichao[2] and Wang Huijun[1]

[1]*Department of Computer Science & Technology, Xinzhou Teachers University, Xinzhou, 034000, China;* [2]*College of Computer Science & Technology, Taiyuan University of Science and Technology, Taiyuan, 030024, China*

Abstract: The recognition of Chinese entertainment news words is an important task for Chinese information processing. In order to solve the problem of recognizing the Chinese entertainment news words, a SVM-based method has been introduced in this paper, which uses active leaning strategy. It selects new instances, incrementally, to be labeled and included in its training set to form an incremental course of learning. The results of the test show that the method is efficient and the precision and recall of entertainment news words recognition achieved are 78.92% and 86.42%, respectively. The method in this paper has gained good effect.

Keywords: Entertainment news words, Machine learning, SVM algorithm, SVM-Based active learning strategy, Text classification.

1. INTRODUCTION

The recognition of Chinese entertainment news words is an important task in Chinese information processing. With the development of the world, the importance of information has become increasingly prominent and analysis and identification of news words have attracted more and more attention. The entertainment news words should be seen as a branch of news research fields. Statistical data show that the ratio of entertainment news words is approximately 21.6%. These words cover a wide range and variety of shapes and patterns which change over time, and they have no obvious morphological markers, so they can be identified with difficulty.

Nowadays, the recognition methods for entertainment news words are divided into two categories: rule-based and statistics-based. The method of statistics-based is used more, including the hidden Markov model [1-4], the maximum entropy model [5, 6], etc.; this method has however obtained some achievements. The system performance which is obtained by statistical methods is closely related to the size and field of the labeled corpus. The composition of news entertainment discourse varies greatly in different areas and at times is very unstable and the tagging of the corpus consumes more energy and time. All these factors often not only become key issues for improving the system performance, but they also limit system migration between different application fields, which is difficult to meet the demands of practical applications.

The support vector machine (SVM) which was developed on the basis of statistical learning theory is a new general learning method. Compared with some statistical learning methods in the past, SVM is based on structural risk minimization principle and has shown superior performance. It can often get better results than other methods especially in the case of limited sample sets. Currently, SVM has been used in many areas of the natural language processing, such as text classification, shallow parsing, phrase recognition etc., and it has achieved good results. This paper attempts to apply SVM in the recognition of Chinese entertainment news phrases, use SVM to identify entertainment news phrases in correct segmentation corpus, improve learning control accuracy and reduce the workload of the manual sample tagging by combining with active learning strategies in the sample selection and training.

2. BASIC PRINCIPLES OF SVM

SVM algorithm has been derived from the statistical learning theory. The algorithm is based on the structural risk minimization principle, which can compress the collection of raw data to support the vector set (usually is the former's 3% -5%), through learning in order to get the classification decision function. The basic idea is to construct a hyper plane as the decision surface, so that the interval between the positive and negative mode is maximum.

SVM method is proposed when optimal classification surface is linearly separable, which is shown in Fig. (**1**).

The hollow circles and open squares represent two kinds of the training samples. *H* is the classification line that is separated correctly. *H1* and *H2* are lines which pass through the points which are nearest to various types of sample and parallel to the classification line. And the distance between the two lines is called the classification interval. In accordance with the principle of empirical risk minimization theory, SVM's actual risk is decided by the formula (1).

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

*Address correspondence to this author at the Heping West Street, Xinzhou, China. Postcard: 034000; Tel: 15635067998; E-mail: kcxdj122@126.com

1874-110X/14

Where, $R(\omega)$ represents the actual risk, $R_{emp}(\omega)$ represents the empirical risk, and Φ represents the confidence interval. Complete separation ensures that $R_{emp}(\omega)=0$ and maximum interval ensures that the minimum range of confidence interval is Φ, so that the real risk is minimum.

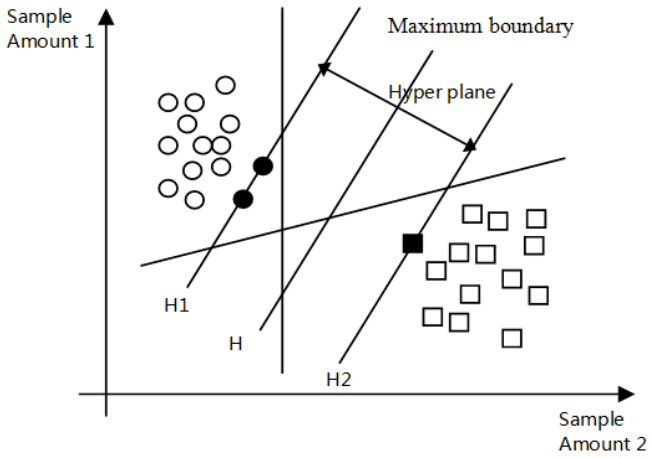

Fig. (1). The optimal classification surface.

Let linearly separable sample set be (x_i, y_i), $i = 1,...,n, x \in R^n, y \in \{+1,-1\}$. The general form of a linear discriminant function in n-dimensional space is $g(x)=\omega \cdot x+b$. The equation of classified surface is:

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

The discriminant function $g(x)$ is taken as normalized, and all types of samples accord with $|g(x)|\geq 1$. The classification interval is equal to $2/\|\omega\|$. The problem is transformed into keeping the largest interval in accordance with the condition that the classification line can correctly classify all samples. It is symbolically described as:

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

In the case of non- linear discriminant function, certain training samples do not meet the condition of expression (3b). So a relaxation item $\varepsilon_i \geq 0$ is added to the left of the condition formula (3b). Formula (3a) that is minimized is equivalent to maximize $\Phi(\omega)=\|\omega\|^2/2$. According to the Lagrange function and Kuhn-Tucker conditions, formula (3) eventually can be converted to:

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

The optimal classification function is obtained by solving equation (4) and is shown as equation (5).

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

$\varepsilon_i \geq 1-y_i(\omega^T x_i+b)$ is obtained by the formula (4b), so formula (4a) can be rewritten as:

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

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

The constant Q in formula (4a) is at equilibrium between the generalization ability and training accuracy. When Q is smaller, SVM has better generalization ability; When Q is larger, SVM has smaller training error. Formula (4b) introduces slack variable which allows some points to overstep the boundary and increases the SVM ability of noise immunity in the case of non-separable. Since the introduction of slack variables, a boundary is defined relative to formula (5) to each sample:

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

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

3. ACTIVE LEARNING STRATEGIES

3.1. Active Learning Strategies in Statistical Learning

Active learning is a machine learning technique, which chooses the most effective information samples from the ever labeled samples iteratively, and then can be marked manually. Each learner can choose some of the most effective training samples from a large number of unlabeled texts and reduce many labels which have little help to improve the accuracy of the learner, so the learner can obtain the same accuracy but requires fewer samples. Reducing the system's expectations error rate of statistical learner can help optimize the training data selection, which is the basis of active learning.

The sample selection based on uncertainty [7, 8] is a major active learning strategy. It makes the learner label the unlabeled examples, and gives the metric of confidence reliability of the sample; then it chooses the samples using the metric. This method is based on sample selection, and it can find the most uncertain sample which enhances the accuracy of the learner most effectively [9]. The focus of these active learning algorithms is to construct a reasonable measure of the output mechanism for assessing learning sequence annotation confidence reliability. Generally, the confidence is the entropy with different labels and probability distribution of samples for the learner which outputs probability values.

3.2. Active Learning Strategies of SVM

The active learning strategies are based on uncertain sample selections which often need to produce the output of probability value of the labeled sample. But SVM's output is not connected with probability. To solve the problem, a new confidence metric needs to be constructed. The output of the basic two categories SVM is a decision value, with its absolute value being the distance from the sample to the optimal separating hyperplane. So the distance can be marked as the learner label's confidence. Literature [10] based on the above idea proposed an approach in which SVM's output can be mapped to probability:

$$p(y=1 \mid f(x)=1/[1+\exp(Af(x)+B)] \tag{8}$$

Where, y is a marked sample, $f(x)$ is a determination value of SVM's output, A and B are the parameter values which need to be determined. It can be proved that the probability value has the same monotonicity with $f(x)$ when $A<0$. In fact, the probability value does not need to be known but its measure value. So the absolute value of SVM's output will be the measure of confidence in this article.

4. IDENTIFICATION METHOD OF ENTERTAINMENT NEWS WORDS USING SVM-BASED ACTIVE LEARNING STRATEGY

4.1. Identification Strategies of Entertainment News Word using SVM

The task for entertainment news word recognition can be expressed by the following model: Suppose a sentence composed of n words is, $S=W_1,W_2,W_3,...,W_n$. W_i represents the ith words of the sentence; L_i is the mark of W_i. IOB labeling method is commonly used to label each word, in which 'B' represents the starting word of the entertainment news word; 'I' represents the non-starting word of the entertainment news word; 'O' represents the word beside the entertainment news word. For example: "" / O Dream of Red Mansions / B "/ O" / O Bao Yu / B "/ O candidates / I determine / O, / O20 years / O General / O boy / O starred / B. / O". Thus, the recognition process of entertainment news words becomes the process of labeling each word in the sentence. Furthermore, the process which is expressed as the classification problem for the sample unit can be accomplished by SVM. In fact, the entertainment news words I and B can be labeled into a category, so the problem becomes a basic two classification problem.

For characteristics selection of the samples, this paper chooses a word and its part of speech of context (both before and after the word), and the mark of entertainment news words is sample characteristic. In addition, taking into account the composition characteristics of entertainment news words, the paper also introduces feature of confidence for call word of entertainment news word. The call word of entertainment news word which often appears before the entertainment news word shows the person's name in entertainment world (such as Zhou Xun, RuoYing Liu, *etc*.). The call word of entertainment news word is the center word constituted by the entertainment news word, which plays an important role for recognition of entertainment news word. The calculating method of confidence of call word is:

$$p_h(w)=p_{h0}(w)/\sum p_{h0}(y) \tag{9}$$

$$p_{h0}(w)=\log_2(C(w)+2) \tag{10}$$

y represents the words in the entertainment news and $C(w)$ is the occurrences of those words. In order to calculate the confidence of call word of entertainment news word, there must be a call word vocabulary which contains frequency information of the word. Since there is no such vocabulary, this paper takes the following measures to establish the vocabulary: Combining training process for active learning, call words of entertainment news word are counted and the frequency information from the labeled corpus is used for training at the beginning and to build a call word vocabulary. Then, the vocabulary and frequency information are updated each time when a new selected sample joins the vocabulary. Thus, the final characteristics of sample are:

$$\begin{aligned} X=(&p_{i-2},l_{i-2},p_h(w_{i-2}),p_{i-1},l_{i-1},p_h(w_{i-1}),\\ &p_i,p_h(w_i),p_{i+1},p_h(w_{i+1}),p_{i+2},p_h(w_{i+2})) \end{aligned} \tag{11}$$

where, p_i represents the part of speech in current position; p_{i-1} represents the part of speech in previous position ; and so on ; l_i indicates the label of entertainment news of the current word; $p_h(w_i)$ represents the confidence when the current word is the call word of the entertainment news word. Considering the problem of data sparseness, the word itself is not regarded as the characteristic.

4.2. Learning Algorithm of SVM Using Active Learning Strategies

This paper introduced the active learning strategy for sample selection in order to use a large number of unlabeled corpuses better and reduce the workload of manual annotation. This strategy gives a better training effect for the same or less size sample training set. Selection strategy has already been described based on uncertainly sample applying SVM. Since the sentence is the basic unit in the selection of corpus sample, the uncertainty of defining sentence based on the uncertainty of word is:

$$S_i=1/\left|W\right|\sum_{w_i\in W} conf(w_i) \tag{12}$$

Among them, W is the set of words marked as word of the entertainment news. $conf(w_i)$ is the uncertainty of w_i. Thus, we can start to train the classifier from a set of marked samples and use the current classifier to classify the unmarked sample for calculating the uncertainty of each sentence. Then the m sentences with max uncertainty can be added to the set of samples. The whole training process is described as follows:

Input: a small amount of labeled corpus L and a large amount of unlabeled corpus U

Table 1. Recall, precision and F-value after each training (m = 50,90).

Sample Number	m=50			m=90		
	Recall	Precision	F-Value	Recall	Precision	F-Value
300	73.24	65.74	68.94	73.23	65.74	68.92
350	76.53	69.05	72.25			
400	78.86	71.36	74.56	77.93	70.45	73.92
450	80.35	72.86	76.05			
500	81.97	74.47	77.67	80.87	73.76	76.65
550	82.47	74.97	78.17			
600	83.01	75.54	78.75	82.19	74.84	77.82
650	83.97	76.47	79.67			
700	84.77	77.26	80.46	83.74	76.27	79.59
750	85.38	77.87	81.07			
800	85.79	78.29	81.49	84.67	77.39	80.53
850	86.10	78.61	81.82			
900	86.42	78.92	82.31	85.45	78.01	81.97

Output: the SVM*M* by training set of samples *L*

(1) To start training from *L* and get SVM*M0*.

(2) Use *Mi* (*Mi* is the SVM with sub-sample selection and re-training) to classify and mark the corpus in *U*, then calculate the uncertainty of each sentence.

(3) Select m sentences which have the maximum value of Su from the *U* to submit them to the human-annotated category. Then transfer them from *U* to *L* while updating the vocabulary of the call word of entertainment news words.

(4) Repeat the cycle which is from step (1) to step (3) till it reaches the specified requirements or samples exhausted in *U*.

5. EXPERIMENTS

The above methods can be used to test the system by an open experiment, which uses the corpus from the network version of the entertainment news. First, a detailed analysis is conducted. Then manual segmentation label is used.

Experimental procedure is as follows: First, the corpus which contains some entertainment news of nearly 55,000 words, including 2056 entertainment news words is extracted and is randomly divided into three parts. Where: About 25 % of the corpus is the test set, about 10% of the corpus is the initial training set *L*, and the two parts of the entertainment news words are manually labeled. The remaining corpus is training data set of active learning in unlabeled set *U*.

Experiments are carried out in three groups: the previous two experiments set different values of m, being 90 and 50, respectively, in order to compare the sample selection effect of different coarseness. The third experiment selects samples randomly to compare the differences between the active learning and the non-active learning without using active learning strategies. The kernel function of SVM uses the radial basis kernel function, which is randomly selected in the experiment. Many experiments show that the performances of different kernel functions of SVM in classification are very similar to each other. Its error rate difference is less than 0.5% [11, 12].

Assuming a=the total of entertainment news words labeled correctly, b=the total number of entertainment news words marked, c= the total of entertainment news words in label results, the performance indicators of learning system used in entertainment news text recognition are defined as follows:

$$\text{Recall: } R = \frac{a}{b} \times 100\% \tag{13}$$

$$\text{Precision: } \mathrm{P} = \frac{a}{c} \times 100\% \tag{14}$$

$$F_\beta = [(\beta^2 + 1) \times P \times R] / (\beta^2 \times R + P) \tag{15}$$

Among them, β=1, the correct word mark is a type of entertainment news, and the borders have been correctly identified [12, 13]. The results are shown in Table **1**.

After a number of sample selection trainings, the original unlabeled samples concentrated to about 65% used for training, the final recall rate and precision rate stabilized at around 86% and 79%, and the F value was also stable at 82% (m=50) [14]. Observing the experimental results from different *M* values, better results can be obtained when more samples are selected, which may be related with the diversity of samples. In addition, the results using active learning strategies are also compared with the results without using active learning strategies in this paper when m=90, which is shown in Fig. (**2**).

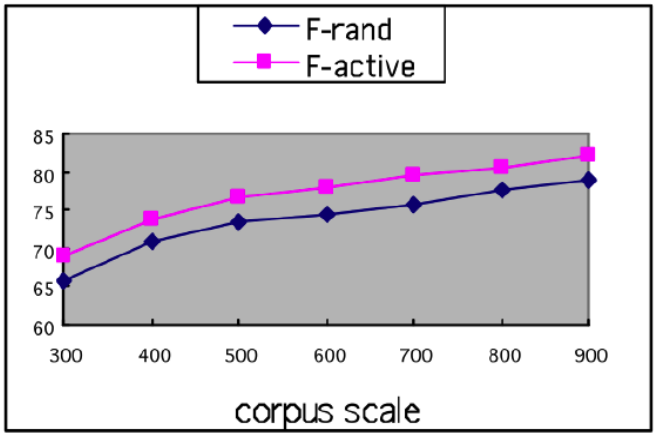

Fig. (2). Comparison of results between active and non- active learning.

As seen from the comparison in Fig. (**2**), the method using active learning strategies is much better than that without using active learning strategies when the number of samples increase gradually [15]. The end result is that the value of F is better than the random sample selecting method from about 3-4%, which shows that the sample selection plays a key role in improving the accuracy of the learner. From the number of samples which can obtain the same values as F , the number of samples required is significantly less when active learning method is used, which shows that the selection of samples can reduce the redundant samples.

CONCLUSION

This paper studied the recognition method of entertainment news word which uses SVM and active learning strategies. Compared with other classifiers, SVM has better performance, whose generalization ability is very good especially in the case of small samples. As a machine learning method, active learning can effectively improve the learner's learning efficiency and accuracy by selecting the most valuable samples to submit to the learner. Experimental results show that this method has good effect. Further work will improve the strategy of active learning sample selection and avoid the diversity of samples to influence the classification accuracy rate.

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] A. Vlachos, "*Active learning with support vector machines*", MS: University of Edinburgh, CiteSeer^{x} 2004.

[2] S. D. Lu, and R. Y. Cui, "Semi-supervised clustering method based on active learning strategy", *Appl. Res. Comput.*, vol. 30, pp. 1718-1720, June 2013.

[3] X. J. Xiang, Y. Gao, L. Shang, and Y. Yang, "Parallel text categorization of massive text based on hadoop", *Comput. Sci.*, vol. 38, pp. 184-188, October 2011.

[4] R. Li, J. H. Zheng, and M. Y. Guo, "Application Study of Hidden Markov Model Based on Genetic Algorithm in Noun Phrase Identification", *Comput. Sci.,* vol. 36, pp. 244-246, 261, October 2009.

[5] G. Salton, and C. Buckley, "Term weighting approaches in automatic text retrieval", *Inform. Process Manag.,* vol. 24, pp. 513-523, May 1988.

[6] C. W. Hsu, and C. J. Lin, "A comparison of methods for multiclass support vector machines", *IEEE Trans. Neural Netw.*, vol. 13, pp. 415-425, February 2002.

[7] O. Chapelle, V. N. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines", *Mach. Learning*, vol. 46, pp. 131-159, January 2002.

[8] K. C. Di. "*Spatial data mining and knowledge discovery*", Wuhan University press: Wuhan 2000.

[9] H. L. Bai, J.M. Zhu, and C. P. Liu, "A fast license plate extraction method on complex background", In: *Proceedings of the Conference on ' IEEE Intelligent Transportation Systems*', USA: IEEE, pp. 985-987. 2003.

[10] C. H. Lu, G. J. Huang, and Z. B. Guo, "Identification of Chinese prepositional phrase", *Commun. Technol.*, vol. 43, pp. 181-183, 186, May 2010.

[11] U. Kiyotaka, and Q. Ma, "Named entity extraction based on amaximum entropy model and transformation rules", In: *Proceeding Conference 38th Annual Meeting on Association Computation Linguist., Hong Kong*: [s.n.], pp. 326-335, 2000.

[12] L. Zheng, and X. Q. Lv. "N+V+N", "V+N+N" structure phrase recognition in search engine query logs", *Comput. Eng. Appl.*, vol. 49, pp. 143-147, June 2013.

[13] K. W. Church, "A stochastic part s program and noun phrase parser for unrest ricted text", In: *Proceedings of the 2nd Conference on 'Applied Natural Language Processing*' Texas, USA, pp. 136-143. 1988.

[14] F. Erik, and S T Kim. "Memory-based shallow parsing", *Machine Learning Research*, vol. 2, pp. 559-594, March 2002.

[15] B. Mak, and E. Bocchieri, "Direct training of subspace distribution clustering hidden Markov model", *IEEE Trans. Speech Audio Proces.*, vol. 9, pp. 378-387, April 2001.

Received: September 22, 2014 Revised: November 30, 2014 Accepted: December 02, 2014