An Intelligent SVM based Data classification in E-Commerce

Mrs.Suji Sukumaran
PG Scholar,Musaliar College of Engineering and Technology,Pathanamthitta,Kerala

***Abstract*—With the rapid development of e-commerce, financial data mining has been one of the most important research topics in the data mining community. Support vector machines (SVMs) and ensemble learning are two popular techniques in the machine learning field. In this paper, support vector machines and ensemble learning are used to classify financial data respectively. The experiments conducted on the public dataset show that compared with SVMs, ensemble learning achieves obvious improvement of performance.**

Keywords- Financial Data Mining; Support Vector Machines; Ensemble Learning

I. INTRODUCTION

The rapid growth of the e-commerce has stimulated the applications of financial data mining. Nowadays, financial data mining has been one of the most important research topics in the data mining community, and attracted much work for this task.

Support vector machines (SVMs) [1], introduced by Vapnik, are a kind of structural risk minimization based learning algorithms and have better generalization abilities comparing to other traditional empirical risk minimization based learning algorithms. By using nonlinear kernel functions, SVMs can map original input data into a high dimensional feature space to seek a separate hyperplane. As a popular machine learning algorithm, SVMs have been widely used in many fields such as data classification and pattern recognition in the last decade.

Ensemble learning is a kind of learning algorithms that construct a set of classifiers and classify new data by voting based on each prediction. The purpose of ensemble learning is to build a learning model to integrate a number of base learning models for obtaining better generalization performance [2]. Recently, ensemble learning is attracting much attention from pattern recognition and machine learning communities.

As two of the most popular techniques in the machine learning field, SVMs and ensemble learning are used to classify financial data respectively in this paper. The experiments are conducted on the public dataset. The experimental results indicate that compared with SVMs, ensemble learning achieves obvious improvement of performance.

II. SVMS

SVMs are based on the principle of minimizing structure risk and the aim of them is to constitute an objective function, then to find a partition hyperplane that can satisfy the class requirement. (x_i, y_i) is the linear separable dataset, $i = 1,..., n$, $x \in R^d$ and $y \in \{+1, -1\}$ is the class label. Then partition hyperplane can be defined as $\omega \cdot x + b = 0$, where ω is the normal vector of the partition hyperplane, and b is the offset of hyperplane. For making the partition hyperplane as far from the point in training dataset as possible, a partition hyperplane to make the bilateral blank area, i.e., $2/\| \omega \|$, maximum must be found, which can be defined as follows.

$$\text{Minimize} \phi(\omega) = \frac{1}{2} \| \omega \|^2$$

A constraint condition must be met, which is defined as follows.

$$y_i(\omega \cdot x_i + b) \geq 1$$

Then, lagrange function can be defined as:

$$L(\omega, b, \alpha) = \frac{1}{2}(\omega \cdot \omega) - \sum_{i=1}^{n} \alpha_i (y_i(\omega \cdot x_i + b) - 1)$$

Subject to the following two conditions, i.e., $\sum_{i=1}^{n} y_i \alpha_i = 0$ and $\alpha_i \geq 0$, then the following formula can be defined for seeking the minimum of lagrange function.

$$\max Q(\alpha) = \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j=1} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

The optimal class function can be defined as follows.

$$f(x) = \text{sgn}((\omega^* \cdot x) + b^*) = \text{sgn}(\sum_{i=1}^{n} \alpha^*_i y_i (x \cdot x) + b^*)$$

For nonlinear separable situation, nonlinear mapping $\kappa(x)$ can be used to map the instance x to higher dimensional feature space which is linear separable [1].

III. ENSEMBLE LEARNING

Boosting, one of the most popular ensemble learning, is introduced in this section.

Boosting is a powerful ensemble method for boosting the performance of any weak learning algorithm, which needs only to be a little bit better than random guessing [3]. As an improvement of the initial boosting algorithm, the AdaBoost algorithm was introduced. For making the learning algorithm to minimize the expected error over different input distributions, it changes the weights of the training instances after each trial based on the base classifier's misclassifications [4]. It explicitly alters the distribution of training data fed to every individual classifier, specifically weights of each training sample. The detailed AdaBoost algorithm is described as Alg.1.

Algorithm 1: The AdaBoost algorithm

1.Input: $S = \{(x_1, y_1),\ldots, (x_N, y_N)\}$,

Number of iterations T

2. **Initialize:** $d_n^{(1)} = 1/N$ for all $n = 1,\ldots, N$

3. **Do for** $t = 1,\ldots, T$

(a) Train classifier with respect to the weighted sample set $\{S, d^{(t)}\}$ and obtain hypothesis $h_t : x \to \{-1, +1\}$, i.e. $h_t = L(S, d^{(t)})$

(b) Calculate the weighted training error ε_t of h_t :

$$\varepsilon_t = \sum_{n=1} d_n^{(t)} I(y_n \neq h_t(x_n)),$$

(c) Set

$$\alpha_t = \frac{1}{2} \log \frac{1-\varepsilon_t}{\varepsilon_t}$$

(d) Update weights:

$$d_n^{(t+1)} = d_n^{(t)} \exp\{-\alpha_t y_n h_t(x_n)\} / Z_t,$$

Where Z_t is a normalization constant, such that

$$\sum_{n=1}^{N} d_n^{(t+1)} = 1 .$$

4. **Break if** $\varepsilon_t = 0$ or $\varepsilon_t \geq \frac{1}{2}$ and set $T = t - 1$.

5. **Output**: $f_T(x) = \sum_{t=1}^{T} \frac{\alpha_t}{\sum_{r=1}^{T} \alpha_r} h_t(x)$

IV. EXPERIMENTS

In this section, we present an experiment where the SVMs and ensemble learning are used for financial data mining. Two dataset from the UCI machine learning dataset repository, i.e., the German credit dataset and Credit Approval dataset [5], are used in experiment.

To analyze the performance of classification, we adopt the Accuracy. As shown in Table 1.

TABLE I. CASES OF THE CLASSIFICATION FOR ONE CLASS

Class *C*		Result of classifier	
		Belong	Not belong
Real classification	Belong	*TP*	*FN*
	Not belong	*FP*	*TN*

Four cases are considered as the result of classifier to the pattern [6].

TP (True Positive): the number of patterns correctly classified to that class.

TN (True Negative): the number of patterns correctly rejected from that class.

FP (False Positive): the number of patterns incorrectly rejected from that class.

FN (False Negative): the number of patterns incorrectly classified to that class.

Then, the performance of the classification can be evaluated in terms of Accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

We used the LIBSVM [7] for SVM implementation. We set radial basis function as default kernel function of SVM. For the ensemble learning algorithm, we use C4.5 [8] as base classifier for AdaBoost.M1. The number of base classifiers in boosting is set as 50. Performance is evaluated by 10-fold cross validation.

Table 2 shows the prediction results of various techniques in terms of Accuracy value on German credit dataset and Credit Approval dataset. On German credit dataset, the prediction Accuracy of boosting is 72%, which beats SVMs by about 2%. On Credit Approval dataset, The prediction Accuracy of boosting is 86%, which is approximately 37% higher than that of SVMs.

TABLE II. COMPARISON OF THE PERFORMANCES ON TWO DATASETS

Dataset	SVMs	Boosting
German credit	0.70	0.72
Credit Approval	0.49	0.86

V. CONCLUSIONS

This paper studied the finacial data mining based on SVMs and ensemble learning respectively. The experiments conducted on two public dataset, i.e., German credit and Credit Approval dataset, show that the ensemble learning technique outperforms SVMs obviously .

REFERENCES

[1] V. Vapnik, Statistical Learning Theory, John Wiley and Sons, New York, 1998.

[2] Dietterich, T., Ensemble methods in machine learning. In Kittler, J., & Roli,F. (Eds.), First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science, 2000, 1–15.

[3] Y. Freund, R.E. Shapire, A decision-theoretic generalization of online learning and an application to boosting. J. Comput. Syst. Sci. 55 (1) (1997)119–139.

[4] Bauer, E., Kohavi, R., An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning, 1999, Vol.36, 105-139.

[5] Blake, C.L., Merz, C.J., UCI repository of machine learning databases, 1998.

[6] Yang, Y. An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1999.

[7] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[8] Quinlan, J., C4.5:Programs for Machine Learning. Morgan Kaufmann, San Matteo, CA, 1993