The Research of Support Vector Machine in Agricultural Data Classification

Lei Shi[1], Qiguo Duan[2], Xinming Ma[1], Mei Weng[1]
[1] College of Information and Management Science, HeNan Agricultural University, Zhengzhou 450002 China
[2] Zhengzhou Commodity Exchange, Zhengzhou 450008 China
sleicn@126.com, dqgcn@126.com, xinmingma@126.com, wengm@163.com

Abstract. The agricultural data classification is a hot topic in the field of precision agriculture. Support vector machine (SVM) is a kind of structural risk minimization based learning algorithms. As a popular machine learning algorithm, SVM has been widely used in many fields such as information retrieval and text classification in the last decade. In this paper, SVM is introduced to classify the agricultural data. An experimental evaluation of different methods is carried out on the public agricultural dataset. Experimental results show that the SVM algorithm outperforms two popular algorithms, i.e., naive bayes and artificial neural network in terms of the F_1 measure.

Keywords: Support Vector Machine, Agricultural Data, Classification

1 Introduction

As a very promising field with a huge growth potential, agricultural data classification is a hot topic in the agriculture and computer science communities. In recent years, many popular algorithms in the machine learning field have been applied in the agricultural data classification, such as decision tree [1], kNN [2], artificial neural network [3, 4], etc. Support vector machine (SVM) [5, 6], introduced by Vapnik and Chervonenkis in 1971, is a machine learning algorithm based on statistical learning theory. By using nonlinear kernel functions, SVM can map original input data into a high dimensional feature space to seek a separate hyperplane, and then it can perform classification by using the constructed *N*-dimensional hyperplane that optimally separates the data into two categories. For the past few years, SVM has been widely used in different fields and it can obtain high performance in many real world classification applications such as image retrieval [7], cancer recognition [8], text classification [9, 10] and credit scoring [11-13].

In this paper, SVM is introduced to classify the agricultural data. Experiments on real agricultural dataset have been conducted and the experimental results indicate that the SVM algorithm outperforms two popular algorithms, i.e., naive bayes and artificial neural network in terms of the F_1 measure. Thus, SVM is an effective method for agricultural data classification.

The remainder of the paper is organized as follows: Section 2 gives an introduction of SVM in detail. Section 3 reports and discusses the experimental results and finally Section 4 states the conclusions of our work.

2 SVM

As a popular machine learning algorithm, SVM is a new generation learning system based on recent advances in statistical learning theory. It realizes the theory of VC dimension and principle of structural risk minimum to constitute an objective function and then find a partition hyperplane that can satisfy the class requirement. The basic idea of SVM can be described as follows. Firstly, search an optimal hyperplane satisfies the request of classification. Secondly, use a certain algorithm to make the margin of the separation beside the optimal hyperplane maximum while ensuring the accuracy of correct classification. Then, the separable data can be classified into classes effectively [6]. As a kind of structural risk minimization based learning algorithms, SVM have better generalization abilities comparing to other traditional empirical risk minimization based learning algorithms. An illustration of the SVM is shown in Fig. 1.

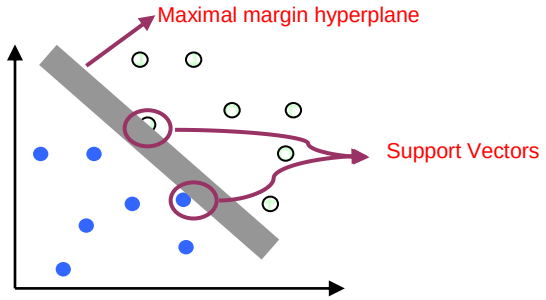

Fig. 1. An illustration of SVM

In a SVM classifier, let the training set be $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where x_i is an input vector and y_i its label. The partition hyperplane can be defined as [6]

$$\omega \cdot x + b = 0 . \tag{1}$$

where b is the offset of hyperplane; ω is the normal vector of the partition hyperplane. A partition hyperplane to make the bilateral blank area, i.e., $2/\|\omega\|$, maximum must be found to make the partition hyperplane as far from the point in training dataset as possible, which can be defined as follows.

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

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

$$y_i(\omega \cdot x_i + b) \geq 1 . \tag{3}$$

The lagrange function can be defined as:

$$L(\omega, b, \alpha) = \frac{1}{2}(\omega \cdot \omega) - \sum_{i=1}^{n} \alpha_i (y_i (\omega \cdot x_i + b) - 1). \quad (4)$$

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.

$$\max Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j). \quad (5)$$

The optimal class function can be defined as follows.

$$f(x) = \text{sgn}((\omega^* \cdot x) + b^*) = \text{sgn}(\sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^*). \quad (6)$$

An important advantage of SVM is that it can be analyzed theoretically using concepts from computational learning theory, and obtain state-of-the-art performance. Recently, it has also been applied to a number of real-world problems such as handwritten characters recognition, information retrieval and the classification of biomedical data. In this paper, SVM is introduced to classify the agricultural data for improving the classification performance of agricultural data.

3 Experimental Results

To study the effectiveness of the SVM method for agricultural classification, we test it on agricultural data in this section. One agricultural dataset obtained from agricultural researchers in New Zealand, i.e., the white-clover dataset [14], is used in experiment. The objective of the white-clover dataset is to determine the mechanisms which influence the persistence of white clover populations in summer dry hill land.

We used the F_1 measure to evaluate the performance of algorithm. A confusion matrix contains information about actual and predicted classifications done by a classification system. The table 1 shows confusion matrix for two class classifier [15].

Table 1. Cases of the classification for one class

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

Several standard terms can be defined for the two class matrix. The recall is the proportion of positive patterns that were correctly identified, as calculated using the equation:

$$recall = \frac{TP}{TP + FN}. \tag{7}$$

Precision is the proportion of the predicted positive patterns that were correct, as calculated using the equation:

$$precision = \frac{TP}{TP + FP}. \tag{8}$$

Then, the performance of the classification can be evaluated in terms of F_1 measure.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

For SVM, we used the LIBSVM [16] for SVM implementation and set linear function as default kernel function of SVM. To evaluate the effectiveness of SVM in agricultural data classification, two popular algorithms, i.e., naive bayes [17] and artificial neural network [18], are implemented and used as benchmarks for comparison. Performance is evaluated by 10-fold cross validation.

Fig. 2 shows the classification results of SVM, naive bayes and artificial neural network in terms of F_1 measure on the dataset. The F_1 value of SVM is 67.3%, which is approximately 6.9% higher than that of naive bayes algorithm and 4.8% higher than that of artificial neural network algorithm.

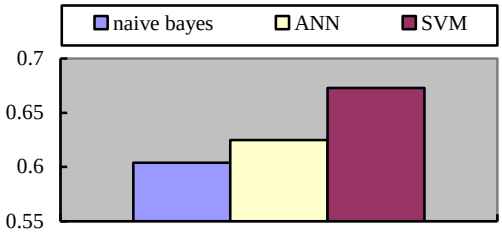

Fig. 2. Comparison of the F_1 of classification on dataset

4 Conclusion

The classification of agricultural data is an important application of information technology in agriculture. SVM is a powerful state-of-the-art classifier and has been applied in many fields. In this paper, SVM is introduced to classify the agricultural data for improving the classification performance. The experimental results show that the SVM is an effective method for classification of agricultural data.

References

1. Kirchner, K., Tolle, K.-H., Krieter, J.: The analysis of simulated sow herd datasets using decision tree technique. Comput. Electron. Agric. 42, 111--127 (2004)
2. Rajagopalan, B., Lall, U.: A k Nearest Neighbor Simulator for Daily Precipitation and Other Weather Variables. Water Resources Research 35 (10), 3089--3101 (1999)
3. Chedad, A., Moshou, D., Aerts, J.M., et al : Recognition System for Pig Cough based on Probabilistic Neural Networks. Journal of Agricultural Engineering Research 79(4), 449--457 (2001)
4. Schatzki, T.F., Haff, R.P., Young, R., et al: Defect Detection in Apples by Means of X-ray Imaging. Transactions of the American Society of Agricultural Engineers 40(5), 1407-1415 (1997)
5. Vapnik, W. N., Chervonenkis, A. Y.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications 16(2), 264--280 (1971)
6. Vapnik, V.: Statistical Learning Theory. John Wiley and Sons, New York (1998)
7. Tao, D.C., Tang, X. O., Li, X. L., Wu, X. D.: Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (7), 1088--1099 (2006)
8. Giorgio, V, Marco, M, Francesca, R.: Cancer recognition with bagged ensembles of support vector machines. Neurocomputing 56, 461--466 (2004)
9. Hyunsoo, K., Peg, H., Haesun, P.: Dimension Reduction in Text Classification with Support Vector Machines. Journal of Machine Learning Research 6, 37--53 (2005)
10. Simon, T., Daphne, K.: Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research, 45--66 (2001)
11. Bellotti, T., Crook, J.: Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications 36, 3302--3308 (2009)
12. Lee, Y. C.: Application of support vector machines to corporate credit rating prediction. Expert Systems with Applications 33(1), 67--74 (2007)
13. Huang, C. L., Chen, M. C., Wang, C. J.: Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications 33(4), 847--7856 (2007)
14. http://www.cs.waikato.ac.nz/~ml/weka/index_datasets.html
15. Yang, Y.: An evaluation of statistical approaches to text categorization. Journal of Information Retrieval 1, 67--88 (1999)
16. http://www.csie.ntu.edu.tw/~cjlin/libsvm/
17. Lewis, D. D.: Naive (Bayes) at forty: The independence assumption in information retrieval. In: 10th European conference on machine learning, pp. 4--15 (1998)
18. Bishop, C. M.: Neural networks for pattern recognition. Oxford University Press, Oxford, UK (1995)