Building Behavior Scoring Model Using Genetic Algorithm and Support Vector Machines*

Defu Zhang[1,2], Qingshan Chen[1], and Lijun Wei[1]

[1] Department of Computer Science, Xiamen University, Xiamen 361005, China
[2] Longtop Group Post-doctoral Research Center, Xiamen, 361005, China
dfzhang@xmu.edu.cn

Abstract. In the increasingly competitive credit industry, one of the most interesting and challenging problems is how to manage existing customers. Behavior scoring models have been widely used by financial institutions to forecast customer's future credit performance. In this paper, a hybrid GA+SVM model, which uses genetic algorithm (GA) to search the promising subsets of features and multi-class support vector machines (SVM) to make behavior scoring prediction, is presented. A real life credit data set in a major Chinese commercial bank is selected as the experimental data to compare the classification accuracy rate with other traditional behavior scoring models. The experimental results show that GA+SVM can obtain better performance than other models.

Keywords: Behavior Scoring; Feature Selection; Genetic Algorithm; Multi-Class Support Vector Machines; Data Mining.

1 Introduction

Credit risk evaluation decisions are crucial for financial institutions due to high risks associated with inappropriate credit decisions. It is an even more important task today as financial institutions have been experiencing serious competition during the past few years. The advantage of using behavior scoring models can be described as the benefit from allowing financial institutions to make better decisions in managing existing clients by forecasting their future performance. The decision to be made include what credit limit to assign, whether to market new products to these particular clients, and how to manage the recovery of the debt while the account turns bad. Therefore, new techniques should be developed to help predict credit more accurately.

Currently, researchers have developed a lot of methods for behavior scoring, the modern data mining techniques, which have made a significant contribution to the field of information science [1], [2], [3]. At the same time, with the size of databases growing rapidly, data dimensionality reduction becomes another important factor in building a prediction model that is fast, easy to interpret, cost effective, and

* This research has been supported by academician start-up fund (Grant No. X01109) and 985 information technology fund (Grant No. 0000-X07204) in Xiamen University.

Y. Shi et al. (Eds.): ICCS 2007, Part II, LNCS 4488, pp. 482–485, 2007.
© Springer-Verlag Berlin Heidelberg 2007

generalizes well to unseen cases. Data reduction is performed via feature selection in our approach. Feature selection is an important issue in building classification systems. There are basically two categories of feature selection algorithms: feature filters and feature wrappers. In this paper we adopt the wrapper model of feature selection which requires two components: a search algorithm that explores the combinatorial space of feature subsets, and one or more criterion functions that evaluate the quality of each subset based directly on the predictive model [4].

GA is used to search through the possible combinations of features. GA is an extremely flexible optimization tool for avoiding local optima as it can start from multiple points. The input features selected by GA are used to train a Multi-Class Support Vector Machines (SVM) that extracts predictive information. The trained SVM is tested on an evaluation set, and the individual is evaluated both on predictive accuracy rate and complexity (number of features).

This paper is organized as follows. In Section 2, we show the structure of the GA+SVM model, and describe how GA is combined with SVM. The experimental results are analyzed in Section 3. Conclusions are provided in Section 4.

2 GA+SVM Model for Behavior Scoring Problems

Firstly, we will give a short overview of the principles of genetic algorithm and support vector machines. Further details can be found in [5], [6].

In order to use SVM for real-world classification tasks, we should extend typical two-class SVM to solve multiple-class problems. Reference [7] gives a nice overview about ideas of multi-class reduction to binary problems.

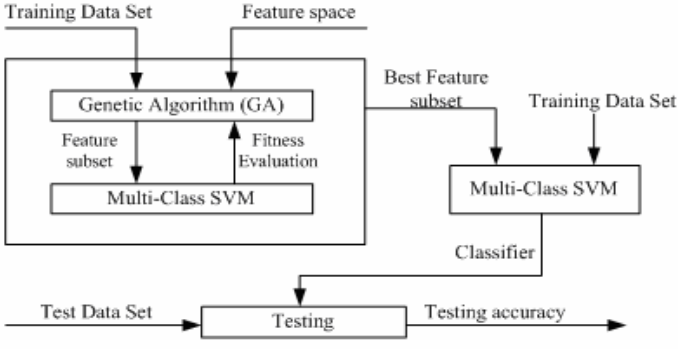

Fig. 1. A wrapper model of feature selection (GA+SVM)

Our behavior scoring model is a hybrid model of the GA and SVM procedures, as shown in Fig. 1. In practice, the performance of genetic algorithm depends on a number of factors. Our experiments used the following parameter settings: the population size is 50, the maximum number is 100, the crossover rate is 0.9, and the mutation rate is 0.01.

The fitness function has to combine two different criteria described to obtain better performance. In this paper we use F_{accuracy} and $F_{\text{compexity}}$ to denote the two criteria.

$F_{accuracy}$: The purpose of the function is to favor feature sets with a high predictive accuracy rate, SVM takes a selected set of features to learn the patterns and calculates the predict accuracy. The radial basis function (RBF) is used as the basic kernel function of SVM. With selected features, randomly split the training data set, the ratio of D_{train} and $D_{validation}$ is 2:1. In addition, since SVM is a stochastic tool, five iterations of the proposed method are used to avoid the affect of randomized algorithm. And the $F_{accuracy}$ is an average of five iterations.

$F_{compexity}$: This function is aimed at finding parsimonious solution by minimizing the number of selected feature as follows:

$$\text{Fcompexity} = 1 - (d-1)/(D-1) . \quad (1)$$

Where D is the dimensionality of the full data set, and d is the dimension of the selected feature set. We expect that lower complexity will lead to easier interpretability of solution as well as better generalization.

The fitness function of GA can be described as follows:

$$\text{Fitness}(x) = \text{Faccuracy}(x) + \text{Fcompexity}(x) . \quad (2)$$

3 Experimental Results

A credit card data set provided by a Chinese commercial bank is used to demonstrate the effectiveness of the proposed model. The data set is in recent eighteen months, and includes 599 instances. Each instance contains 17 independent variables. The decision variable is the customer credit: good, bad, and normal credit. The number of good, normal, and bad is 160, 225, and 214 respectively.

In this section, GA+SVM is compared with a pure SVM, back-propagation neural network (BPN), Genetic Programming (GP) and logistic regression (LR). The scaling ratio of the training and test data set is 7:3. In order to compare the proposed method with other models, five sub-samples are used to compare the predictive accuracy rate of those models. The predictive accuracy rates of the test data set are shown in Table 1. In the first sample, the feature subset selected by GA is shown in Table 2.

Table 1. Predictive accuracy rates of proposed models

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Overall
GA+SVM	0.8883	0.8994	0.9162	0.8771	0.8883	0.8940
SVM	0.8771	0.8715	0.8883	0.8492	0.8659	0.8704
BPN	0.8659	0.8676	0.8892	0.8431	0.8724	0.8676
GP	0.8827	0.8939	0.9106	0.8827	0.8883	0.8916
LR	0.8492	0.8659	0.8770	0.8436	0.8715	0.8614

Table 2. Features selected by GA+SVM in Sample 1

Feature Type	Selected Features
Custome's personal information	Age, Customer type, Education level
Custome's financial information	Total asset, Average of saving

On the basis of the simulated results, we can observe that the classificatory accuracy rate of the GA+SVM is higher than other models. In contrast with other models, we consider that GA+SVM is more suitable for behavior scoring problems for the following reasons. Unlike BPN which is only suited for large data sets, our model can perform well in small data sets [8]. In contrast with the pure SVM, GA+SVM can choose the optimal input feature subset for SVM. In addition, unlike the conventional statistical models which need the assumptions of the data set and attributes, GA+SVM can perform the classification task without this limitation.

4 Conclusions

In this paper, we presented a novel hybrid model of GA+SVM for behavior scoring. Building a behavior scoring model involves the problems of the features selection and model identification. We used GA to search for possible combinations of features and SVM to score customer's behavior. On the basis of the experimental results, we can conclude that GA+SVM obtain higher accuracy in the behavior scoring problems.

In future work, we may incorporate other evolutionary algorithms with SVM for feature subset selections. How to select the kernel function, parameters and feature subset simultaneously can be also our future work.

References

1. Chen, S., Liu, X: The contribution of data mining to information science. Journal of Information Science. 30(2004) 550-558
2. West, D.: Neural network credit scoring models. Computers & Operations Research. 27(2000) 1131-52
3. Li, J., Liu, J., Xu, W., Shi. Y.: Support Vector Machines Approach to Credit Assessment. International Conference on Computational Science. Lecture Notes in Computer Science, Vol. 3039. Springer-Verlag, Berlin Heidelberg New York (2004)
4. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence. 1(1997) 273-324
5. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
6. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag, Berlin Heidelberg New York (1995)
7. Allwein, E., Schapire, R., Singer, Y.: Reducing multiclass to Binary: A Unifying Approach for Margin Classifiers. The Journal of Machine Learning Research. 1 (2001) 113-141
8. Nath, R., Rajagopalan, B., Ryker, R.: Determining the saliency of input variables in neural network classifiers. Computers & Operations Research. 8(1997) 767–773