Expert Systems with Applications 39 (2012) 6774–6781

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method

Akhil Bandhu Hens [a], Manoj Kumar Tiwari [b,*]

[a] Department of Mathematics, Indian Institute of Technology, Kharagpur, India
[b] Department of Industrial Engineering and Management, Indian Institute of Technology, Kharagpur, India

ARTICLE INFO

*Keywords:*
Support vector machine
Credit scoring
*F* score
Stratified sampling

ABSTRACT

With the rapid growth of credit industry, credit scoring model has a great significance to issue a credit card to the applicant with a minimum risk. So credit scoring is very important in financial firm like bans etc. With the previous data, a model is established. From that model is decision is taken whether he will be granted for issuing loans, credit cards or he will be rejected. There are several methodologies to construct credit scoring model i.e. neural network model, statistical classification techniques, genetic programming, support vector model etc. Computational time for running a model has a great importance in the 21st century. The algorithms or models with less computational time are more efficient and thus gives more profit to the banks or firms. In this study, we proposed a new strategy to reduce the computational time for credit scoring. In this approach we have used SVM incorporated with the concept of reduction of features using *F* score and taking a sample instead of taking the whole dataset to create the credit scoring model. We run our method two real dataset to see the performance of the new method. We have compared the result of the new method with the result obtained from other well known method. It is shown that new method for credit scoring model is very much competitive to other method in the view of its accuracy as well as new method has a less computational time than the other methods.

## 1. Introduction

Recently credit industry is growing rapidly with the rapid growth of financial sector, banking sector. The influence of the credit industry has been seen the consumer market of the developing and the developed countries. The competitiveness in credit industry is in an larger extent specially in emerging economies. Many financial firms have faced a tight competition in the market with respect to smooth service, efficient service, more benefit to the customers. Credit card is one of the services which every customer wants. Credit cards make the financial transaction a little bit easier. The number of applications for issuing credit cards is also increasing. Every financial firm wants to give better service to the customers and make large profit. But there is a constraint to issue more credit cards without any decision. There many some fraud applicant. They can misuse the credit cards. For banking institutions, loans are often the primary source of credit risk. Credit scoring models are used to evaluate the risk. Banks also want to issue more credit cards to increase their profits with minimizing the risk. It has been seen that some of the applications are fraud. Credit scoring models have been extensively used for the credit admission valuation. Many quantitative methods have been developed for the prediction of credits more accurately. Credit scoring models helps to categorize the applicants into two classes: in one case the applicants are accepted and in other cases applicants are rejected. During application procedure various information about the applicant like income, bank balance, profession, family background, educational background etc. are demanded. Not all of these characteristics are quantitative. Some characteristics are also qualitative. These qualitative characteristics are converted into quantitative characteristics with some standard procedure for the sake of computation. These characteristics are needed for credit scoring model specification. The objectives of credit scoring models are reduction of the cost for credit analysis, to make credit decision making comparatively faster, to make it efficient.

Recently, many researchers have done their research work on increasing the efficiency of credit scoring model, reducing the computation time. As a result various methods for the decision making of credit analysis have been proposed. These approaches help to detect fraud, assess creditworthiness etc. Baesens et al. (2003) presented the various classification techniques for credit scoring. An effective credit analysis also helps to issue loans with minimum risk. The previous historical data of the applicants for credit cards are necessary to create a credit scoring model. The credit scoring model has been created using the previous data consisting of the applicants' information and decision taken to their applications. Then this model is applied to a new applicant for credit cards. From

[*] Corresponding author. Tel.: +91 3222 283746.
*E-mail address:* mkt09@hotmail.com (M.K. Tiwari).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.12.057

this model creditor makes a decision on a new applicant whether he will be accepted or rejected. With the increase in the accuracy of credit scoring model, the creditors risk is decreased.

With the increasing importance of credit scoring model, this field has invoked interests to many researchers to work on it. Filter approach employs better results for credit scoring datasets as compare to wrapper approach (Somol, Baesena, Pudil, & Vanthienen, 2005). Many researchers have developed many methods for credit scoring. The computation of some model takes very long time. As a consequence, research works are being continued on reducing the computational method, increasing the overall efficiency of the credit scoring model. If it is needed much time to create a credit scoring model, then it is unworthy and it reduces the profit of the firm. Moreover the efficiency of the firm has been decreased. Computational time has a great significance for credit scoring model. Lower the computational time, it will be more efficient. In this study a new strategy has been proposed that reduces the computational time for credit scoring. Creditor takes many characteristics of the applicants for creating a credit scoring model. Some characteristics may not be so useful like other characteristics. If the creditor includes these characteristics, the computational time will be automatically increased. Sometimes the previous data are very large in volume. If the creditor takes the whole data set, the computational time will be very high. In this paper, we have tried to reduce the computational time by reducing the size of the data set to be considered for the computation and optimizing the characteristics (i.e. considering only the suitable characteristics from all characteristics) as well as our other focus is reduce the deviation of the result in our model from the actual result obtained from the whole dataset considering all characteristics.

Many modern data mining techniques are used for credit scoring models. For the last two decades various researchers has developed numerous data mining tools and statistical methods for credit scoring. Some of the methods are linear discriminant models (Reichert, Cho, & Wagner, 1983), logistic regression models (Henley, 1995), k-nearest neighborhood models (Henley & Hand, 1996), neural network models (Desai, Crook, & Overstreet, 1996; Malhotra & Malhotra, 2002; West, 2000) genetic programming models (Ong, Huang, & Tzeng, 2005; Koza, 1992). Chen and Huang (2003) presented a work to the credit industry that demostrates the advantages of Neural network and Genetic algorithm to credit analysis.). Each study has shown some comparative result with other methods. Neural network model is seen to be more effective in credit risk prediction comparative to other methods (Tam & Kiang, 1992). Logistic regressions, *K*-nearest neighborhood method for credit scoring also have a great significance in the view of accuracy and efficiency. Consequently, Neural network models are treated as the benchmark in the view of accuracy and solutions. But to get a good solution for a large dataset, more number of hidden layers is required in the neural network model. Consequently, the computational time is very high. So the efficiency is low.

Recent research works have focused on increasing the accuracy of the credit scoring model and developing some advanced methods. For example, Ho mann, Baesens, Martens, Put, and Vanthienen (2002) suggested a neuro fuzzy and a genetic fuzzy classifier. A integrated model based on clustering and neural networks for credit scoring was suggested by Hsieh (2005; Garson, 1991; Zhang, 2000). A hybrid system with artificial networks and multivariate adaptive regression splines was proposed by Lee and Chen (2005).There are more hybrid models. Lee, Chiu, Lu, and Chen (2002) suggested a hybrid credit scoring model with neural networks and discrimininant analysis. Recent research work involves integrating various artificial intelligence methods to data mining approach to increase the accuracy and flexibility.

There are many data mining and statistical approach for clustering. Support vector machine (SVM) is an important data mining tool which is being used for clustering, classification etc. Support vector machine was first proposed by Vapnik (1995). After that researches have been done to apply this SVM tool in a wide range of many applications. Now SVM is used in many applications like clustering of data, pattern reorganization, text categorization, biostatistics etc. A simple decomposition method for support vector machine is illustrated by Hsu, Chang, and Lin (2003). In credit scoring model classification is necessary. SVM model is also used in credit scoring model for making decision. In recent past few years there has been some comparative as well as hybrid approach studies between SVM and other computational approach. A comparative study between SVM and neural network model was done by Huang, Chen, Hsu, Chen, and Wu (2004). An integrated SVM model with genetic algorithm was suggested by Huang, Chen, and Wang (2007); Fröhlich and Chapelle (2003); Pontil and Verri (1998). They have shown that the result is comparable to benchmark methods.

All the previous studies about credit scoring are discussed about the accuracy of the model. Most probably no previous study discussed about the computational time. 21st century is for high performance computing. Everybody is conscious about the computational time of the method. Any method with low computational time is much more efficient and thus gives more profit to the company. There is not any concrete study to include to sampling methodology in SVM for credit scoring model with the integrating *F* score (Weston et al., 2001). In this study a new approach has been proposed about the reduction in computational time. In this paper, we have taken a stratified sample and we have done the credit scoring model with SVM and *F* score. Then we have done a comparative study with computation of the whole data and all characteristics.

This study is incorporated with the concept of reduction of unnecessary features with the calculation of *F* score. The reduction of features from the calculation of the sample data takes less time than the reduction of features from the calculation of whole dataset (Kohavi & John, 1997). Here the computational time is reduced for the two reasons: eliminating the unnecessary features and taking a sample instead of considering the whole sample It is shown in that paper that with reduction of features from the data set from the stratified sample gives similar accuracy with other benchmark methods. As a reduction of features and reduction of size of data, the computational time decreases significantly.

The paper organized as follows: Section 2 briefly describes the procedure of support vector model. The concept of sampling method and stratified sampling are described in Section 3.In Section 4, new strategy for reduction of computation time is discussed. In Section 5, empirical results are shown for real data set and comparative study is done here. In Section 6, the remarks and conclusion are drawn.

## 2. Concepts of support vector machine (SVM) classifier

SVM classifier was most probably first proposed by Vapnik (1995). Here we have illustrated the concept of SVM (support vector machine) and its application as a two class classifier. Basically SVM is a supervised learning method that analyzes data and recognizes patterns, used for statistical classification and regression analysis.

Suppose there are given some dataset of pairs $(x_i, d_i)$, $i = 1,2,3,\ldots,n$ where $x_i \in R^n$ and $d_i \in \{-1,+1\}$. The value of $d_i$ helps us to indicate the class to which the point $x_i$ belongs. Each $x_i$ is actually a $p$-dimensional real vector. With the help of we can find out the maximum margin hyper-plane that divides the points having $d_i = 1$ from those having $d_i = -1$. For this reason SVM is also known as maximum margin classifier. Any hyperplane can be written as the set of points $x$ satisfying

$$r \cdot x - b = 0 \tag{1}$$

where $\cdot$ denotes the dot product and the vector $\mathbf{r}$ is a normal vector which is perpendicular to the hyperplane. The offset of the