5th SASTech Iran |Mashhad May 12 - 17 | 2011
5th Symposium on Advances in Science & Technology

A hybrid mining model based on Artificial Neural Networks, Support Vector Machine and Bayesian for credit scoring

M. Siami[1], M.R.Gholamian[2], R.Nasiri[3]
[1]Iran University of Science and Technology, Tehran, Iran; m_siami@ind.iust.ac.ir
[2]Iran University of Science and Technology, Tehran, Iran; gholamian@iust.ac.ir
[3]Iran University of Science and Technology, Tehran, Iran; r_nasiri@ind.iust.ac.ir
Paper Reference Number:19
Name of the Presenter: Mohammad Siami

Abstract

In recent years, credit scoring is becoming one of the most important topics in the financial field. In consumer credit markets, lending decisions are usually represented as a set of classification problems. In this Paper, we have proposed a hybrid mining model for credit scoring, based on Artificial Neural Networks, Support Vector Machine and Naïve Bayesian to improve the accuracy of credit scoring classification task. To make these basic classifiers as an ensemble model, we have used majority voting technique to improve the prediction accuracy of existing credit scoring models. In order to approve the capability of our model in the field of credit scoring, Australian credit real dataset of UCI machine learning database repository has been applied. Finally we conduct a comparative assessment for the performance measuring of these methods, with three basic learners (Artificial Neural Networks, Support Vector Machine and Naïve Bayesian). Our findings lead us to believe that this hybrid method may provide better performance in the field of credit scoring.

Key words: Credit Scoring, Data Mining, Classifier ensemble, Support Vector Machine, Descion Tree, Naïve Bayesian

1. Introduction

Appearance of economical crisis in recent decade caused the banks and credit institutes to have more attention to credit risk (Wang et al 2011). So banks to reduce their credit risk used various kinds of credit scoring systems, therefore these systems developed and applied to support credit decisions (Hsieh and Hung 2010) Credit scoring is used to classify the applicants into two types: applicants with good and bad credit. Applicants with good credit have great possibility to repay financial obligation, and Applicants with bad credit have high possibility of defaulting(Wang et al 2011).

The importance of customer credit scoring emphasized in previous studies so that 1% false prediction about bad or good status of customers could make many problems (Zhang and Zhou et al 2010, Hsieh 2005)

As before mentioned, credit scoring is a binary classifier that is used for make distinction between bad or good applicants (Chen et al 2010) Many algorithms such as neural networks , (Hsieh 2005, Tsai and Wu 2008), decision trees (Zhang and Zhou, et al 2010, Huang et al 2007), genetic programming (On and Jeng et al 2005), support vector machine (Wang.G et al 2011, Chen et al 2010, Huang et al 2007, TunLi and Shiue et al 2006) and logistic regression analysis (Wang 2011, Huang et al 2007, Thomas 2000) have been proposed for better prediction of good applicant, but nevertheless this problem remains a hot topic in financial research.

In recent years, many studies have done to improve the accuracy of the models; all of these methods have tried to gain the maximum accuracy value (Zhang and Zhou et al 2010, Leea et al 2006, Tsai and Wu 2008, Hsieh and Hung 2010). But these methods may have some classifiers that may lead to have a good prediction, and other classifiers may not .in other words, in specific dataset some classifiers have high accuracy and other classifiers do not(Huang and Chen 2009). In this paper to overcome on this problem, we propose a hybrid mining model that combined three different classifiers include of support vector machine, artificial neural networks and naive Bayesian. We used this hybrid mining model to reduce the divination of our results on different data sets.

Besides the introduction, the organization of this paper is as follows. we review the credit scoring and related works in section 2, in section 3 introduce three single prediction models Used in this paper. In continues we proposed our hybrid mining model in Section 4. Explanation of experimental results on a benchmark dataset is organized in section 5. Finally we give our conclusion and further research work in Section 6.

2. Customer Credit Scoring

Credit scoring is a method that used for forecasting financial risk to customer lending. there are many definitions for credit scoring but most of them have expressed that credit scoring is a classification method which classifies customers into two main categories: customer with good and bad credit (Wang et al 2011,Zhang and Zhou et al 2010,Thomas 2000).

Applying good applicants for credit scoring have many benefits .some of these benefits is: (Wang et al 2011)

- Evaluation of customer risk
- Decrease cost for credit assessment
- Decision making for loan applicant easily and quickly

Credit scoring is obtained from the customer experiences. Credit scoring analyst explores customer credit data in five perspectives that what mattered. These five perspectives is known 5Cs that include: the Character of the consumer, the Capital, the Collateral, the Capacity and the economic Conditions (Fig. 1) (Thomas 2000).

Customer credit scoring based on this framework, with the large number of customer credit data and increasing the number of requests for loan to conduct manually is impossible. So that methods and algorithms for credit scoring have been proposed to be able give us the best answers (Wang et al 2011).

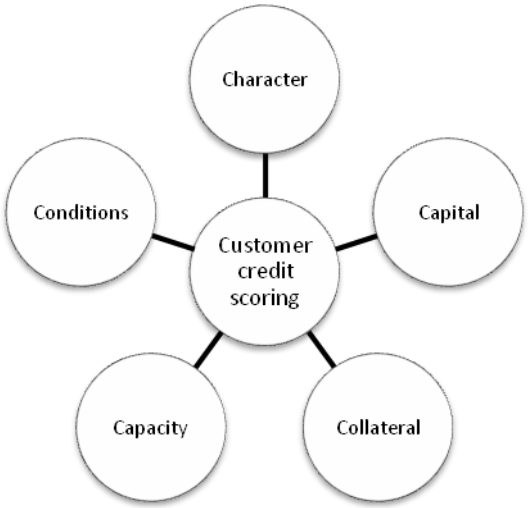

Fig. 1: Credit Scoring 5Cs Framework

Due to previous studies, Two categories of automatic credit scoring techniques, i.e., statistical techniques and Artificial Intelligence (AI) techniques, (Wang et al 2011, Nanni and Lumini 2009) Some statistical techniques have been widely applied to build the credit scoring models, these techniques require some assumption ,such as multivariate normality assumptions for independent variables, are frequently violated in the practice of credit scoring, which makes these techniques invalid for finite samples. first statistical methods that have been suggested for credit scoring are Discriminant Analysis (LDA) and Logistic Regression Analysis (LRA) (Thomas 2000). Methods of AI such that support vector machine (SVM) (Wang et al 2011,Huang et al 2007, TunLi and Shiue et al 2006), artificial neural networks (ANNs) (Hsieh 2005, Tsai and Wu 2008), decision trees (Zhang and Zhou et al 2010, Leea et al 2006, Chen an Li 2010) and genetic programming (On and Jeng et al 2005), is made to improve or solve defects of statistical methods but these methods also have some problems. We can not to be said certainly that one method of AI techniques is best. To being best method is related on the details of the problem, the data structure, the used characteristics, the extent to which it is possible to segregate the classes by using those characteristics, and the objective of the classification (Wang et al 2011).

According to above, we have three problems here:

- Low Accuracy of Statistical techniques (Wang et al 2011 and Thomas 2000)
- Don't exist best method for AI techniques (Wang et al 2011).
- When we use a single method for credit scoring, in specific dataset, we may have high accuracy and in other dataset lead to bad results (Wang et al 2011, Zhang and Zhou et al 2010)

In order to overcome on these problems, we presented a hybrid mining model that combined benefits of several methods and provide one unit model.

3. Research Methodology

In this section we introduce three single prediction models Used in this paper.

3.1 Support Vector Machine

SVM is a state-of-the-art AI technique that has proven their performance in many applications, such as credit scoring, financial time series prediction and so on. SVM is a binary classifier .minimize of classification error and simultaneously maximize the geometric margin are abilities of SVM. Support vector defined as input samples that are closest to the maximum margin hyper plane (Huang and Chen 2009).

If SVM divided data into two groups that caused hyper plane gain maximum margin, hyper plane is called optimal separating hyper plane (OSH) (fig 2). If linearly indivisible data in a lower dimensional feature space transform to linearly divisible data in a higher-dimensional feature space, OSH obtain. In this paper, we use sequential minimal optimization (SMO) algorithm for SVM due to its quick convergence. Stability of SVM is dependent on the maximum margin hyper plane is relatively stable (Huang and Chen 2009).

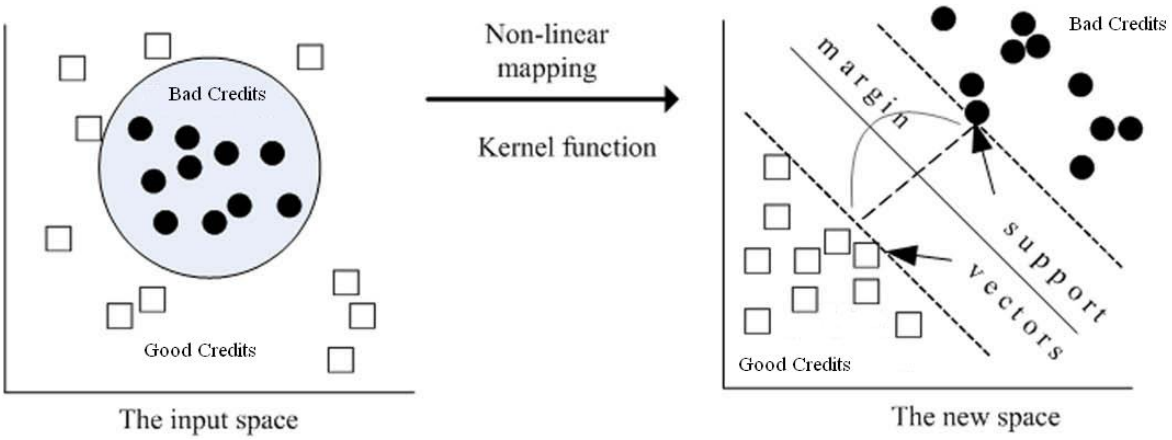

Fig. 2: Credit Scoring SVM example

3.2 Artificial Neural Networks

Structure of Artificial Neural Networks (ANNs) consists of several layers ,which one layer as input layer, one other for output layer, and a number of hidden layers existing in between. Each layer can have one or more nodes, and there are weights to connect the nodes in different layers. ANN(Wang et al 2011). We chose the most commonly used back-propagation networks for our experiments. In this paper, we use the multi-layer perceptron (MLP) is a universal function approximator that it has proven by Cybenco theorem (West 2000)

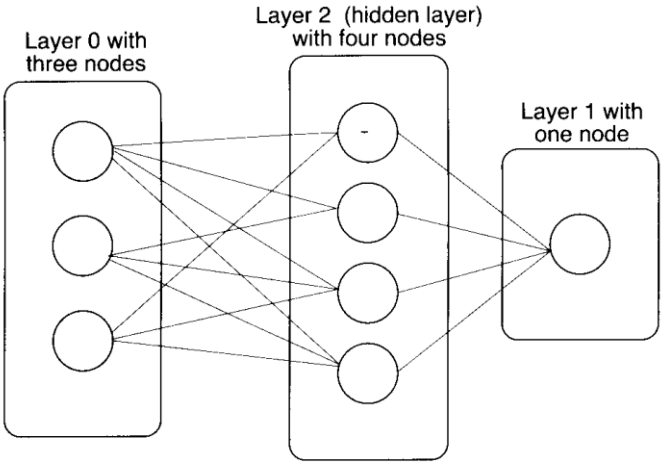

Fig. 3: Credit Scoring ANNs example

3.2 The naïve Bayesian classifier

The naïve Bayesian (NB) is a classifier for data mining that have better accurate in compared with another classifiers such that decision tree learning, rule learning, neural networks and instance based learning (Kononenko, 1991; Langley and Sage, 1994).

This classifier basically learns the class-conditional probabilities P (Xi = xi | C = cl) of each variable Xi given the Class label cl . A new test case (Xi = xi , . . . , Xn = xn) is then classified by using Bayes's rule to compute the posterior probability of each class cl given the vector of observed variable value (Ouali A. et al 2006)

$$P\,(C = c_l \,|X_1 = x_1, \ldots, X_n = x_n) = \frac{P\,(C = c_l)\; P\,(X_1 = x_1, \ldots, X_n = x_n \,|C = c_l)}{P\,(X_1 = x_1, \ldots, X_n = x_n)}.$$

The naïve Bayesian have one assumption that the variables are conditionally independent given the class label. Hence,[naïve]

$$P\,(X_1 = x_1, \ldots, X_n = x_n \,|C = c_l) = \prod_{i=1}^{n} P\,(X_i = x_i \,|C = c_i)\,.$$

This condition is necessary to reach the maximum prediction accuracy:

$$\text{the predicted class } L\,(X_1, \ldots, X_n) = \arg\max_c\,(P\,(C\,|X_1, \ldots, X_n)).$$

The naïve Bayesian is shown in fig.4 because the structure network of naïve Bayesian is static, only probability tables P (c) and P (Xi | c) need to be assessed

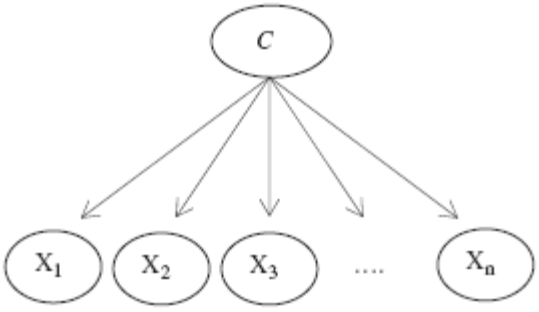

Fig. 4: naive bayes classifiers

4. Credit Scoring hybrid mining model

The learning process of our hybrid mining model for Credit Scoring was shown in Fig.1. We can partition two phases in the whole learning process. Firstly, by three different training classifiers such as Artificial Neural Networks(ANNs), Support Vector Machine(SVM) and Naïve bayesian classifiers will be learned (or trained) and generated; in this phase, we will have three different models that works independently. In the next Phase, for all observers in testing set, each Credit Scoring model will have a classification result Ci (‘‘good” or ‘‘bad”) and finally we vote to decide the final classification result of each observer with majority vote strategy.

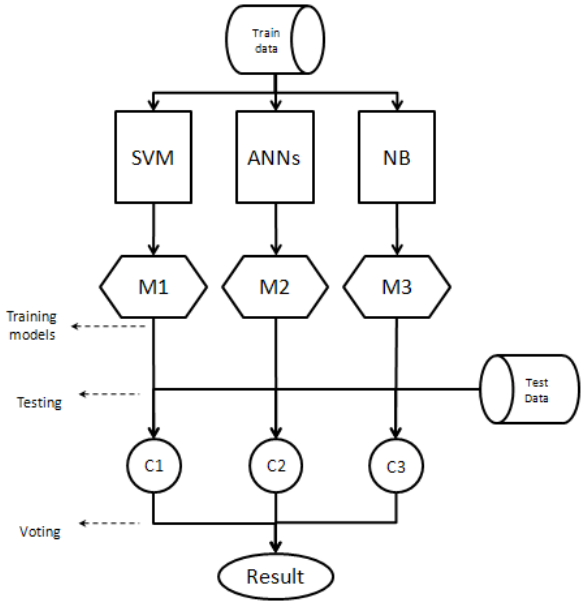

Fig. 5: Credit scoring hybrid mining model

4.1 Majority voting strategy

In this paper, we have applied majority voting technique to improve the prediction accuracy of three base credit scoring classifiers. The majority vote is a common voting strategy that combines multiple classifiers, that is to say, in the sequence of C1, C2, C3, if the number of ‘‘good” is greater than the number of ‘‘bad”, this sample will be identified as a good credit, otherwise it will be introduced as a bad credit (Zhang and Zhou et al 2010).

5. Results and Analysis

In the following, we describe the used dataset, our validation method and the results of our experiments.

5.1 Real World Data set

Australian credit real dataset from UCI Repository of Machine Learning Databases (http://www.niaad.liacc. up.pt/statlog/datasets.html) has been used to evaluate our classifier on the customer credit data. This dataset consists of 690 samples, with 307 good applicants and 383 bad ones. Each sample contains 15 features, including 6 nominal and 8 numeric features. Also, 15th feature is the class label which says that the customer have a good or bad credit. Table 1 described full information about Australian dataset.

name	Number of Classes	Instances	Nominal features	Numeric features	Total features
Australian	2	690	6	8	14

Table 1 : Real world dataset from UCI repository

5.2 Validation method

In order to test dataset, we applied ten-old cross validation method. Also, we have evaluated the performance of our hybrid mining model by comparing it with base classifiers. The performance of hybrid classifier has been compared with the applied classification methods.

5.3 Numerical Results

For implementation of base learners, such as ANN (MultiLayerPerceptron module), SVM (SMO module) and BN (naïvebayes module) we have used Weka (release 3.6.1) (Witten and Frank 2005) on the Australian credit data set, over the ten-fold cross validation. In continues, final result obtained from combination of each classifier result, with majority voting (Zhang and Zhou et al 2010).

We first compare our hybrid method and base classifiers (ANN, SVM, NB) with average accuracy that the value in each of them is shown in Fig6. Proposed our methodology have best accuracy against other classifiers.

Fig7 and Fig 8 show minimum and maximum accuracy obtained in 10-fold cross-validation. In these stages, my method with naive bayesian has the best answer among other methods. but as mentioned in Fig.6 and Fig.9 , our method have maximum average accuracy and minimum standard deviation therefore these method is better than others.

Fig. 8 illustrates standard deviation of different classifiers. Our method has first rank in this item. We have minimum value in this parameter; it means that almost in all of ten-fold cross validation, the accuracy of out hybrid method is more stable than other methods.

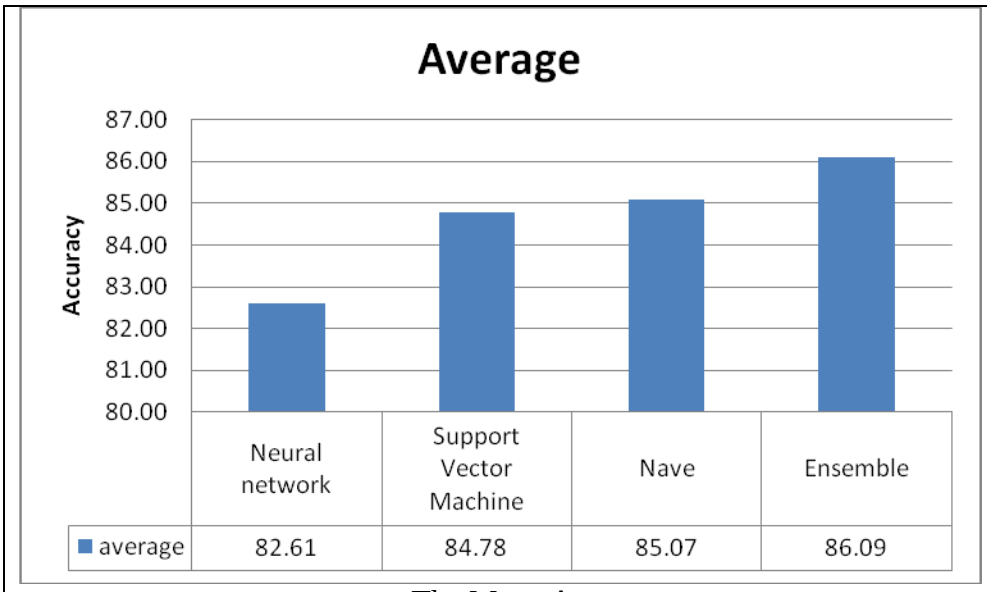

Fig. 6: The Mean Accuracy

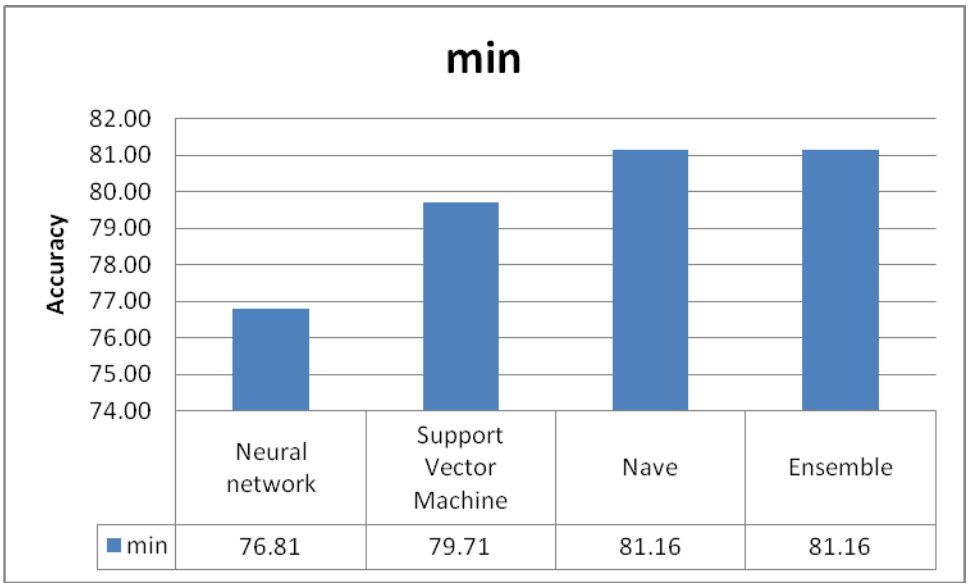

Fig. 7: The Minimum Accuracy

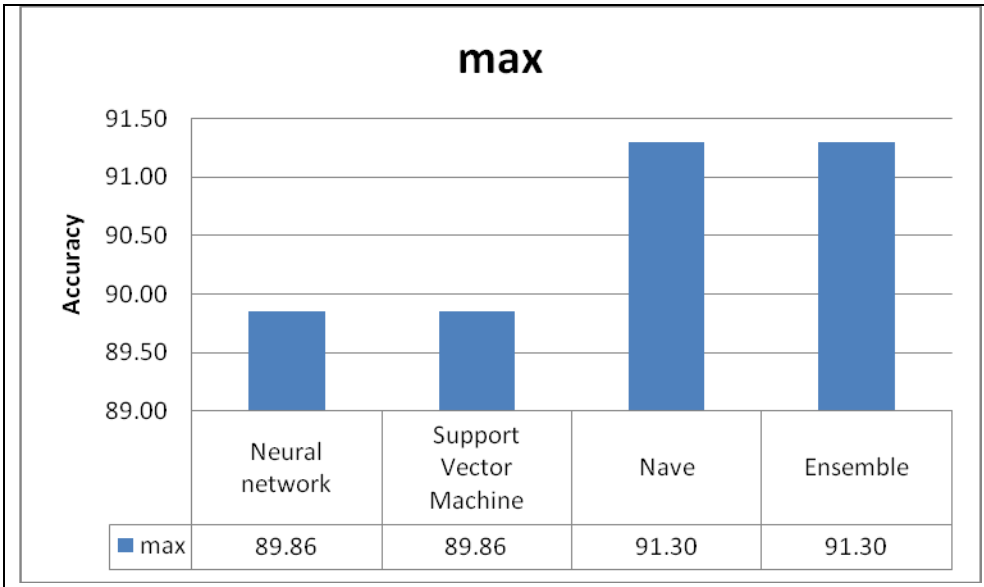

Fig. 8: The Maximum Accuracy

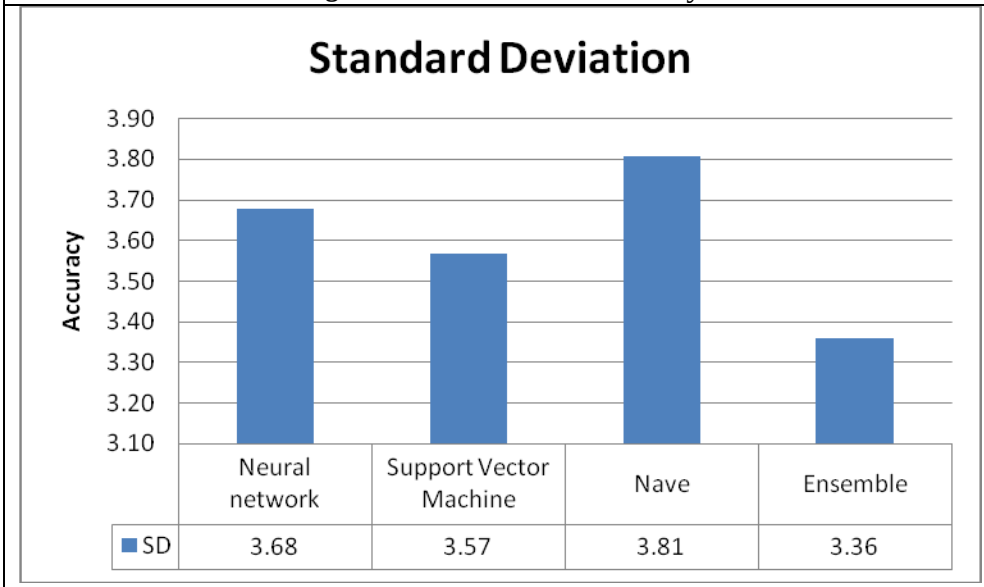

Fig. 9: The Standard Deviation

5. Conclusions

Recently, assessment of credit risk has been considered in financial discussions more than before.

In this paper has tried to provide one model for credit scoring that can be classified with more accurately. The model is presented in this paper was a hybrid model that combine 3 classifiers (ANN, SVM, NB) by majority vote. The accuracy of proposed method has assessed by using Australian data from UCI and 10 fold cross-validation .Finally ,we compare the answers from each classifiers and result showed that proposed method can be improved the accuracy.

First, the use of this method for different data set with larger size. Second, increase the number of original models that would be combined. Third, using of variety voting method for

combine original classifiers. Future researches could be present the extension of this model that include:

- Use of this model for different dataset with large value
- Increasing the number of base models
- Use of various voting method for combining base classifiers

Acknowledgements
We thank the Iran Telecommunication Research Center for financial support. We wish to thank the developers of Weka. We also express our gratitude to the donors of the different datasets and the maintainers of the UCI Repository.

References.

Chen.F and Li.F, (2010) "Combination of feature selection approaches with SVM in credit scoring" Expert Systems with Applications,vol. 37 pp. 4902-4909.

Chen.W, Ma.C, and Ma.L (2009)"Mining the customer credit using hybrid support vector machine technique," Expert Systems with Applications, vol.36, pp. 7611-7616.

Hsieh N.-C., Hung L.-P (2010) "A data driven ensemble classifier for credit scoring analysis Expert Systems with Applications, 37 (1), pp. 534-545.

Hsieh.N.C (2005) "Hybrid mining approach in the design of credit scoring models," Expert system with applications vol.28, pp. 655-665.

Huang.C.L, Chen.M.C, Wang.C.J (2007)"credit scoring with datamining approach based on support vector machine," Expert System with application vol.37, pp. 847-856.

Hung C., Chen J.-H. (2009) "A selective ensemble based on expected probabilities for bankruptcy prediction" Expert Systems with Applications, 36 (3 PART 1), pp. 5297-5303

Kononenko, I., (1991). "Semi-naïve Bayesian classifier". In: Proceedings of European Conference on Artificial Intelligence, pp. 206–219.

Langley, P., Sage, S., (1994). "Induction of selective Bayesian classifiers". In: Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence, Seattle,WA. Morgan Kaufman Publishers, Los Altos, CA, pp. 339–406.

Leea.T,Chiub.Ch.Ch,Y.-Ch.Chouc,Ch.J.Lud (2006) “Mining the customer credit using classification and regression tree and multivariate adaptive regression splines,” Expert system with applications, vol.50, , pp. 1113-1130.

Nanni.L, Lumini.A (2009)"An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring", Expert system with applications, vol.36, 3028-3033.

On.C.S, Jeng.J, Huang, G.HshiungTzeng (2005)“Building credit scoring model using genetic programming”, Expert System with application, vol.29, pp.41-47.

Ouali A., Ramdane Cherif A., Krebs M.-O(2006) “Data mining based Bayesian networks for best classification” Computational Statistics and Data Analysis, 51 (2), pp. 1278-1292.

Thomas.L.C (2000) “A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers” international journal of forcasting vol.16, pp.149-172.

Tsai.Ch.F, Wu.J.W (2008)”Using neural network ensembles for bankruptcy prediction and credit scoring,” Expert Systems with Applications, vol.34, pp. 2639-2649.

TunLi.S ,Shiue.W , Huang.M.H (2006)"The evaluation of consumer loans using support vector machines," Expert System with application, vol.30, pp.772-782.

Wang.G , Hao.J , Ma.J , Jiang.H (2011) “A comparative assessment of ensemble learning for credit scoring” Expert system with applications vol.38, pp.223-230.

West.D (2000) “Neural network credit scoring models” Computers and operation researches vol 27, pp 1131-1152.

Witten.H and Frank.E, Data Mining (2005) "Practical Machine Learning Tools and Techniques," 2nd ed, Morgan Kaufmann, <http://www.cs.waikato.ac.nz/ml/weka>.

Zhang, D., X. Zhou, et al (2010) “Vertical bagging decision trees model for credit scoring,” Expert system with applications, vol. 37, pp. 7838-7843.