IJSES International Journal of Social and Economic Sciences
Uluslararası Sosyal ve Ekonomik Bilimler Dergisi
E-ISSN: 2146-0078, 7 (2): 22-25, 2017

Credit Repayment Analysis Using Support Vector Machine And Principal Component Analysis

Emine BAHÇE ÇİZER[1*] Ayça AK[2] Vedat TOPUZ[3]
[1]Softtech, İstanbul, Türkiye
[2]Marmara University, Vocational School of Technical Sciences, Istanbul, Turkey
[3] Marmara University, Vocational School of Technical Sciences, Istanbul, Turkey

*Corresponding Author
E-mail: emine.bahce@softtech.com.tr

Received: July 15, 2017
Accepted: October 05, 2017

Abstract

Bank and lenders are required to conduct credit analysis to determine the creditworthiness of customers who applying for credit. These organizations apply a number of different methods in order to perform credit analysis with high accuracy, along with various statistical analysis tools. For this purpose, we will use the German Credit data set which is downloaded from UCI Machine Learning Repository open access based site. There are 1000 customer records in the data set and the credit status of these customers is encoded with the appropriate ones 1 and the credit status of these customers is encoded with the inappropriate ones 0. In the first step of this study, SVM analysis will be performed using 21 dependent variables and 1 independent variable in the data set. In the second step of this study, 21 dependent variables will be reduced by performing PCA analysis and SVM analysis will be performed with the dependent variables obtained after the PCA analysis. Will compare the performance of these two different analyzes in the outcome phase of the study.

Keywords: Support Vector Machine, Principal Component Analysis, Credit Analysis

INTRODUCTION

Banks need to manage their risks in the best way in order to manage their money resources effectively and efficiently. Generally, credit risk is the uncertainty of the customer's repayment status of the credit to provided by the bank. In this context, giving credit is one of the most important functions of the banks and as well as one of the most risky tasks of the banks. Banks uses various statistical techniques as data mining, fuzzy logic, regression analysis, classification analysis in order to minimize the risks that may arise in the repayment of the credit and they apply these techniques in their decision making systems. The use of all these techniques has one and the most important purpose is to classify the data set with the most appropriate algorithm.

It is seen that the accuracy of estimation of the results of analysis using SVM technique is higher than the other methods when studies on credit risk estimation are examined recently.In addition to credit prediction, the SVM analysis technique is successfully applied in many areas. This high prediction accuracy of the SVM technique is to determine how to drow the classifier boundary line between two groups in a plane. To draw this boundary line , the SVM draws close to each other and two parallel border lines in the data set and classify the data by approximating these two boundary lines and producing a common boundary line. The principal components analysis is a statistical analysis method which is used to separate the small number of unrelated variables called major components from the large data set. The purpose of the principal components analysis is to explain the minimum variance amount with the least number of main components. Principal component analysis is used for variable reduce on large data sets in many fields such as banking, finance, social research. When literature studies are examined, there are many studies on credit risk prediction using SVM.

Ghodselahi and Amirmadhi in their study called "Application of Artificial Intelligence Techniques for Credit Risk Evaluation" they designed a hybrid model for credit rating that applies collective learning in lending decision [Ghodselahi, Amirmadhi, 2011].

Shahbudin, Hussain, Hussain, A.Samad, Tahir, in their study called "Analysis of PCA Based Feature Vectors for SVM Posture Classification" in the training process, two different solver accounts were used to analyze and classify human body position using SVM technique based on a combination of two different identification [Shahbudin, Hussain, Hussain, A.Samad, Tahir, 2010]. Martens, Baesens, Gestel, Vanthienen, in their study called, "Comprehensible Credit Scoring Models Using Rule Extraction From Support Vector Machines" they have recently presented a general overview of the proposed rule extraction techniques for SVMs [Martens, Baesens, Gestel, Vanthienen]. Huang, Chen, Wang, in their study called "Credit Scoring with a Data Mining Approach Based on Support Vector Machines" they have used three strategies to construct hybrid SVM-based credit scoring models to assess the credit rating obtained from the applicant's input characteristics [Huang, Chen, Wang, 2007]. Nguyen, in his study called "Tutorial on Support Vector Machine" he has written a tutorial work on a support vector machine with mathematical proofs and examples that help researchers to understand theoretically the fastest way to practice [Nguyen, 2015]. Sitt, Wu in their study called "Evaluation of Credit Risk" they have made a SVM-based credit grading classifier with 70% classification ability compared to standard credit ratings [Sitt, Wu, CS 229 – Machine Learning]. Ha, Nguyen in their study called "Credit scoring with a feature selection approach based deep learning" they have established a credit scoring model on deep learning and feature selection to assess the credit rating obtained from the applicant's input characteristics [Ha, Nguyen, 2016]. Hongjiu, Yanrong, Wuchong in their study called "An Application of Support Vector Machine for Evaluating Credit Risk of Bank" they have implemented SVM as a kind of advanced feedback network and they have applied this technique to how commercial credits in banks would evaluate the credit risk [Hongjiu, Yanrong, Wuchong,]. Madhavi, Radhamani in their study called "Improving the credit scoring model of microfinance institutions by support vector machine" They are investigating Microfinance institutions' credit scoring with a new non-parametric technique called Support Vector Machine [Madhavi, Radhamani]. Wu, Guo, Zhang, Xia in their study called "Study of Personal Credit Risk Assessment Based on Support Vector Machine Ensemble" they have introduced a SVM-based method based on fuzzy integral to distinguish the good creditor from the bad one [Wu, Guo, Zhang, Xia, 2010]. Dafincescu, in his study called "Learning Machines and Their Application in Credit Risk Prediction" focuses on early warning systems models, that are used to predict the default of a company based on patterns extracted from historical data [Dafincescu, 2013]. Min, Lee in their study called "Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernel Function Parameters" applies support vector machines (SVMs) to the bankruptcy prediction problem in an attempt to suggest a new model with better explanatory power and stability. To serve this purpose, they use a grid-search technique using 5-fold cross-validation to find out the optimal parameter values of kernel function of SVM [Min, Lee, 2005].

METHOD

Support Vector Machines

The main purpose of classification is to simplify the data and to provide users with more comprehensible information. Support vector machines are one of the effective and simple classification methods used for classification of data. The SVM was proposed by Boser, Guyon and Vapnik in 1992. The goal of a support vector machine is to find the optimal separating hyperplane which maximizes the margin of the training data.

SVM needs training data which means SVM is a both supervised learning algorithm and classification algorithm. Support vector machines can classify both linearly distinguishable and linearly indistinguishable data sets. With a proper conversion, the data can always be divided into two classes with a hyperplane. The hiperplane nearest learning data is called support vectors.

Notations and explanations mentioned below are quoted from source [Karagül, 2014]. For a classification problem of two classes, the primal model for SVM with flexible margins is expressed as:

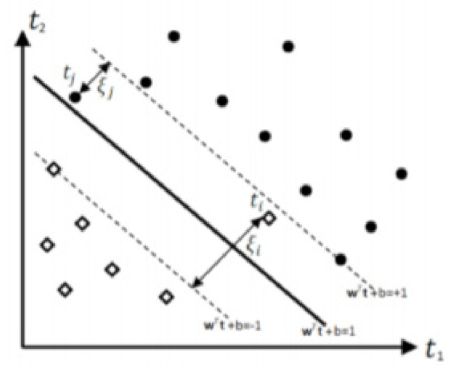


Figure 1 SVM explanation.

$$\min_{w,b} \quad \frac{1}{2} w^t w + C \sum_{i=1}^{N} \zeta_i$$

$$y_i (w^t t_i + b) \geq 1 - \zeta_i , i = 1, \dots , N$$

$$\zeta_i \geq 0, i = 1, \dots , N$$

The dual model for problem is obtained as follows:

$$\min_{\alpha} \quad \mathcal{L}(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \sum\nolimits_{j=1}^{N} \alpha_i \alpha_j y_i y_j \, t_i^T t_j - \sum\nolimits_{i=1}^{N} \alpha_i$$

$$\sum_{i=1}^{N} \alpha_i y_i \ , i = 1, .., N$$

$$0 \leq \alpha_i \leq C, i = 1, .., N$$

$t_i t_i$ variables are input vectors, $y_i y_i$ variables are output, α is Lagrange parameters. *C* is the penalty parameter. However, what makes the SVM approach more effective is the core functions that match the input space to the attribute space. Commonly used core functions are Gauss, Polynomial, Sigmoid, Linear and Radial Base Core Function (RBF).

The dual model due to the core function is expressed as:

$$\min_{\alpha} \quad \mathcal{L}(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \sum\nolimits_{j=1}^{N} \alpha_i \alpha_j y_i y_j \, K(t_i, t_j) - \sum\nolimits_{i=1}^{N} \alpha_i$$

$$\sum_{i=1}^{N} \alpha_i y_i \ , i = 1, .., N$$

$$0 \leq \alpha_i \leq C, i = 1, .., N$$

As a result of solving the quadratic programming problem, the classifier prediction model obtained in the dual space is obtained as follows:

$$\hat{y} = \sum_{i \epsilon S}^{\#SV} \alpha_i y_i \, K(t, t_i), i = 1, .., \#SV$$

K_{ij} denotes the kernel matrix, *S* denotes the set of support vectors, #SV denotes the number of support vectors, and $\hat{y}$ denotes the classifier estimate. With the obtained model, the classifier model can be used completely independent of the primal model.

Principal Component Analysis

Principal component analysis is the method of expressing the data set consisting of the original p variables with new variables that are fewer in number and linear components of these variables. The analysis of the principal components is called the method of describing the number of variables with correlation between them and expressing them with k variables that have no correlation and are linear components of the original variables in number less than (k<p) the original number of variables. Eigenvalues and eigenvectors of the covariance matrix or the correlation matrix are found to find

the linear components of the variable p in the data matrix. Principal component analysis has three main objectives:

- Reduce the size of the data
- Making estimation
- View the dataset for some analysis

As a result of the PCA analysis, the actual size of the p-dimensional space is determined. This real dimension is called principal component analysis. The principal components have three characteristics.

- Without correlation
- The first principal component is the variable that most describes the total variability.
- The next principal component is the variable that most explains the remaining variability.

$$y_1 = a_1^t(x-\mu)$$)mponent,

$$y_2 = a_2^t(x-\mu)$$ l component,

$$y\,r_j = a_j t(x\,r - \mu)$$ es,

$$c_j = root(\lambda_j . a_j)$$ ors,

ANALYSIS MODEL DESIGN

Real life problems are composed of many different components, and these highly variable components can sometimes not be linearly classified. At this point, the classification feature of non-classifiable components of SVM is used. Today SVM technique is used to solve many different real life problems.

In this part of the study, the German Credit Data Set, which will be primarily used for analysis, has been downloaded from the UCI Machine Learning Repository site. The Creditability field was selected as a dependent variable within the data set, while the other variables were selected as independent variables. In the first step, 1 independent variable and 21 dependent variable were used in the data set to construct the working model. SVM analysis was performed using these variables with SPSS Modeler software.

In the SVM analysis, the data set divided by 70% training will be 30% test. SVM analysis performed here showed 97.29% correct classification of training data and 69.67% of test data correctly.

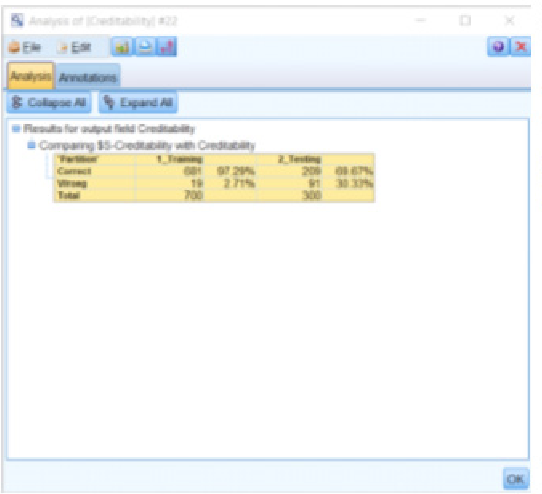


Figure 2 First SVM analysis result

Principal component analysis was performed using SPSS Modeler software in the second phase of the model to be created. The result of the Principal component analysis is that our german data set variables are ,

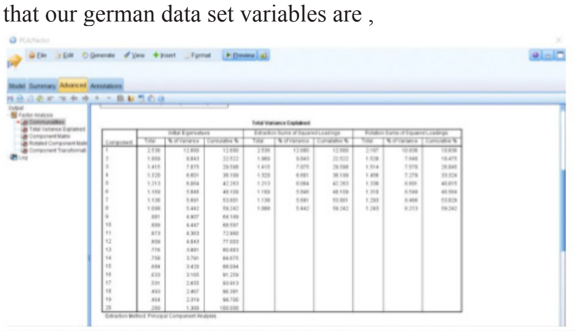
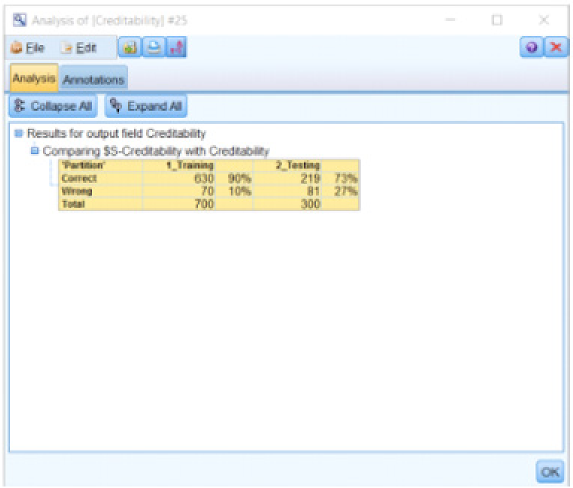


Figure PCA analysis resultant reduced variables

We performed the SVM analysis again with these variables, which we obtained as the result of principal component analysis, and structured our data set as 70% training and 30% test as we did in our first SVM analysis The results of SVM analysis with Principal Component analysis resultant variables showed that we correctly classified 90% of the SVM train data correctly and 73% of the test data in the SVM analysis with the variables we obtained.



CONCULUSION

With this study, SVM analysis was performed at the first stage with the information of the customers who have been previously applied for credit and have creditability information. As a result of this analysis, it is seen that 97.29% of the training data are correctly classified while 69.67% of the test data are correctly classified. In the second phase, we applied principal component analysis to our german credit data set which we used in our first analysis, and we performed SVM analysis again with the reduced variables we obtained. As a result of this analysis, we have seen that 90% of the training data are correctly classified and 73% of the test data are correctly classified. In the first SVM analysis, the correct classification rate of the training data is 97%, whereas in the second SVM analysis, the ratio drops to 90%. The first analysis showed that the test data were correctly classified as 69.67%, whereas in the second analysis this ratio increased 73%.

At this point we see that the correct classification rate of test data increases in the SVM analysis that we have done by reducing the variables with Principal Component Analysis. This shows that we can increase the accuracy percentages of the analyzes with minimum required variables without using too many variables to make a good classification.

REFERENCES

Karagül, K., (2014), The Classification Of The Firms Traded In Istanbul Stock Exchange By Using Support Vector Machines, Pamukkale University Journal of Engineering Sciences, Volume 20, Number 5, 174-178. İstatistik | Sınıf, (Web, A.T., 28.05.2017), https://goo.gl/hZ4KUR.

Atatürk Üniversitesi, (Web, A. T., 30.05.2017), https://goo.gl/WpNB0F

Heba E., Ashraf D., Aboul H., Ajith A, (2010), Principle Components Analysis and Support Vector Machine Based Intrusion Detection System, 10th International Conference on Intelligent Systems Design and Applications.

Chong W., Yingjian G., Xinying Z., Han X., (2010) Study Of Personal Credit Risk Assessment Based On Support Vector Machine Ensemble, International Journal of Innovative Computing, Information and Control, Volume 6, Number 5, 2353-2360.

Jae H. M., Young-Chan L., (2005), Bankruptcy Prediction Using Support Vector Machine With Optimal Choice Of Kernel Function Parameters, Expert Systems with Applications 28, 603-614.

Cheng-Lung., Mu-Chen C., Chieh-Jen W., (2007), Credit Scoring With a Data Mining Approach Based On Support Vector Machines, Expert Systems with Applications 33, 847–856.

Van-Sang H., Ha-Nam N., (2016), Credit Scoring With a Feature Selection Approach Based Deep Learning, MATEC Web of Conferences 54.

Asuri V. M., Radhamani .G, Improving The Credit Scoring Model Of Microfinance Institution By Support Vector Machine, International Journal of Research in Engineering and Technology, Volume 03.

Ruxandra D., (2013), Learning Machines And Their Application In Credit Risk Prediction, Bucharest Academy Of Economic Studies Financial And Stock Exchange Management Master.

Ahmad G., Ashkan A., (2011), Application Of Artificial Intelligence Techniques For Credit Risk Evaluation, International Journal of Modeling and Optimization, Volume 1, Number 3.

Marland S., Tony W., Evaluation Of Credit Risk, CS 229 – Machine Learning.

David M., Bart Baesens., Tony Van G., Jan V., Comprehensible Credit Scoring Models Using Rule Extraction From Support Vector Machines.

Shahrani S., Aini H., Hafizah H., Salina A.S., Nooritawati Md. T., (2010), Analysis Of PCA Based Feature Vectors For SVM Posture Classification, 6th International Colloquium on Signal Processing & Its Applications (CSPA).

Liu H., Hu Y., Wuchong, An Application of Support Vector Machine for Evaluating Credit Risk of Bank, Proceedings of the 7th International Conference on Innovation & Management.

Bc. Michal H., (2014), Support Vector Machines For Credit Scoring, University Of Economics in Prague Faculty of Finance, Master Thesis.