International Journal of
Intelligent Systems and
Applications in Engineering
IJISAE
ISSN:2147-6799 www.atscience.org/IJISAE
Advanced Technology and Science
Original Research Paper

Estimation of Credit Card Customers Payment Status by Using kNN and MLP

Murat KOKLU*[1], Kadir SABANCI[2]

Accepted 3rd September 2016

Abstract:The Default of Credit Card Clients dataset in the UCI machine learning repository was used in this study. The credit card customers were classified if they would do payment or not (yes=1 no=0) for next month by using 23 information about them. Totally 30000 data in the dataset's 66% was used for training and rest of them as 33% was used for tests. The Weka (Waikato Environment for Knowledge Analysis) software was used for estimation. In estimation Multilayer Perceptron (MLP) and k Nearest Neighbors (kNN) machine learning algorithms was used and success rates and error rates were calculated. With kNN estimation success rates for various number of neighborhood value was calculated one by one. The highest success rate was achieved as 80.6569% when the number of neighbor is 10. With MLP neural network model the estimation success rates was calculated when there are different number of neurons in the hidden layer of MLP. The best estimation success rate was achieved as 81.049% when there was only one neuron in the hidden layer. MAE and RMSE values were obtained for this estimation success rate as 0.3237 and 0.388 respectively.

Keywords:*Data mining, Weka, MLP, kNN*

1. Introduction

Credit card is a physical card that is used for easily paying amount of a shopping. The cardholder could use it to give a paying promise as a requital to the cost of services and goods [1]. The issuer (possibly a bank) of the card assigns a credit for the cardholder to use it as cash advance or for payment to a dealer.For the banks the most important thing during credit card marketing is the payment capability of customers. In this study a payment status estimation have been proposed for credit card customers. For this purpose data mining algorithms have been used.

Data mining is a computational process that reveals patterns in data sets by using such methods like artificial intelligence, machine learning, statistics etc [2]. The methods used in data mining are investigated in two groups as predictive and descriptive. In predictive methods, a model is created by using a dataset whose results are known. For Example in a bank, the properties of customers who pay their credits back can be revealed and a model can be created by using previous data sets about funding of them. Afterward this model can be used on new customers for determining the possibility of pay their credits back. In descriptive methods, a relationship can be searched between two data sets. For example, the shopping habits of two different culture may be investigated for similarity [3].

Data mining methods can be divided into three groups due to their function.

1. Classification and Regression
2. Clustering
3. Association Rules

Data mining methods are used to classifying the data set. In order to learn a model which can divide an input data into given categories, training samples are used in classification process. The classification operations includes subsequent steps: a training dataset creation, determination of classes , describing attributes, determining more effective attributes, relevance analysis, model learning, usage of the model for classifying of an unknown data [4].

In this study, an estimation about whether the payment for next month is going to be done or not by the credit card clients in the default credit card clients data set with 23 attributes obtained from the UCI Machine Learning Repository, have been done. For estimation kNN and MLP algorithms have been used. The success rates and error values have been presented and compared with each other.

2. Material and Methods

2.1. Dataset

In this study the default of credit card clients data set obtained from UCI Machine Learning Repository have been used. This data set have been obtained from Credit cart customers' default payments in Taiwan. In this data set there are 23 attributes and a binary type class. These attributes and descriptions are as follow [5].

X1: The credit amount (NT dollar): it involves total credit that assign for the cardholder and his/her family.
X2: Sex (1:M ; 2:F).
X3: Edu. (1:Graduate; 2:Unv.; 3:High School; 4:Others).
X4: Marital (1:M; 2:S; 3:Oth).
X5: Age (in years).

[1] *Selcuk University, Faculty of Technology, Computer Engineering, Konya, Turkey*
[2] *KaramanogluMehmetbey University, Faculty of Engineering, Department of Electrical and Electronics Engineering, Karaman, Turkey*
* *Corresponding Author:Email: mkoklu@selcuk.edu.tr*
Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.

X6 to X11: Past payments.
X12-X17: Amount of bill statement (NT dollar).
X18-X23: Amount of previous payment (NT dollar).
X24: Class (0:Nopayment; 1:Payment done)

2.2. Software-WEKA

Developed by Waikato University in New Zealand, WEKA is an open-source data mining software with a functional graphical interface which incorporates machine learning algorithms [6]. WEKA includes various data pre-processing, classification, regression, clustering, association rules, and visualization tools. The algorithms can be applied on the data cluster either directly or by calling via Java code [7][8]. They are also suitable for developing new machine learning algorithms.

2.3. Machine learning algorithms

*K-Nearest Neighbours Algorithm:*A supervised learning algorithm, k-NN solves classification problems. Classification is the examination of the attributes of an image and the designation of this image to a predefined class. The critical point is the determination of the features of each category previously [9]. Conforming to the used classifcation algorithm k-NN based on the attributes drawn from the classification stage, the distance of the new individual that is wanted to be classified to all previous individuals is considered and the nearest k class is used. As an outcome of this procedure, the belonging of the test data is determined due to the k-nearest neighbour category which contains more exactly determined classes.In k-NN, the determination of the algorithm used for distance calculation and neighbour number are the critical optimization points.In the study, the optimum k number is appointed with experiments. In the calculation of distance, the Euclidean Distance is performed.
Euclidean calculation method [10]:

$$d(x_i,x_j) = \left(\sum_{s=1}^{p}(x_{is}+x_{js})^2)\right)^2$$

xi and xj are two points that is wanted to be learnt the distance between them.

Multilayer Perceptron: It is a feed forward type artificial neural network model which maps input sets onto appropriate output sets. A multilayer perceptron (MLP) is composed of multiple layers where each layer is connected to the other one. Each node is a processing element or a neuron that has a nonlinear activation function exceptthe input nodes. It uses a supervised learning technique named back propagation and it is used for training the network. The alteration of the standard linear perceptron, MLP is capable of distinguishing data which are not linearly separable [8].

3. Results and Discussion

In the study, WEKA software was used in order to estimate the payment for next month is going to be done or not by the credit card. Using the kNN algorithm, the estimation success rates of payment were obtained for different k-neighbour values. The Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) have been attained. The k-NN method's estimation success (in percentage), MSE and RMSE values have been presented in Table 1. A diagram showing the effect of number of neighbourliness on MAE and RMSE values have been presented in Figure 1.

Table1. The k-NN classifier’s estimation success and error values

Neighborliness Number (k)	*Estimation Success (%)*	*MAE*	*RMSE*
1	73.3039	0.2673	.5168
2	78.4608	0.2727	0.4525
3	77.1176	0.274	0.4285
4	79.3824	0.2731	0.4132
5	78.8627	0.2727	0.4049
6	80.049	0.2733	0.4002
7	79.8725	0.2735	0.3961
8	80.3137	0.2742	0.394
9	80.2843	0.2742	0.3913
10	80.6569	0.2743	0.3897

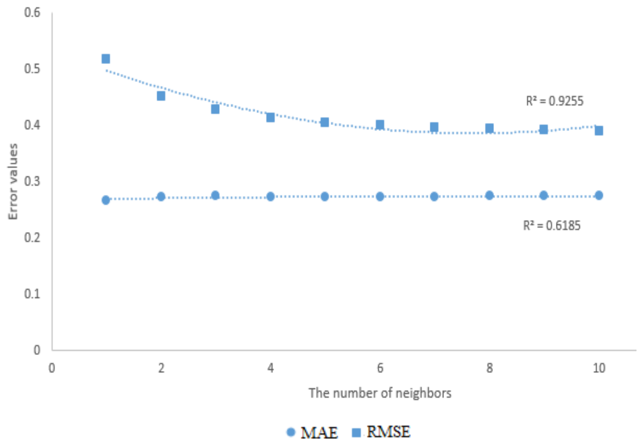

Figure 1. The change of error values versus the neighbourhood value

The data in the same dataset were processed using the MLP model, and the estimation success rates of credit card payments. The MAE, RMSE and estimation accomplishment have been attained while the number of neurons in the hidden layer is changing from 1 to 50. In the MLP model, the training was performed by taking the learning rate value as 0.3, momentum value as 0.2 and iteration number as 500. The estimation accomplishment, MAE and RMSE values have been presented in Table 2. A diagram showing the effect of number of neurons in the hidden layer on MAE and RMSE values have been presented in Figure 2.

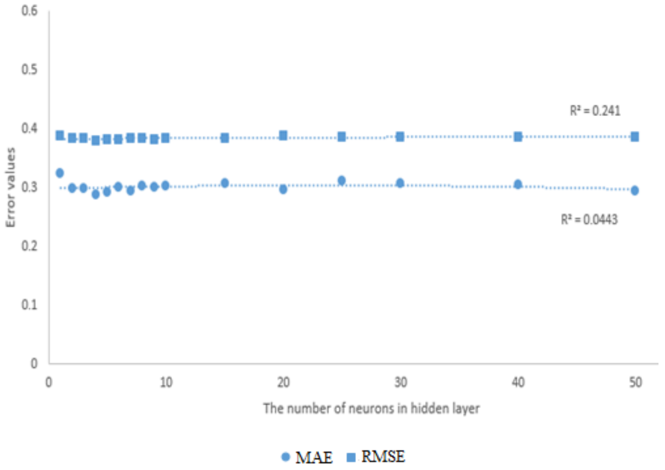

Figure 2. Change of error rate versus the number of neurons in the hidden layer

Table 2. The MLP classifier's success ratesand error values

The number of neurons in the hidden layer	*Estimation Success (%)*	*MAE*	*RMSE*
1	81.049	0.3237	0.388
2	79.1961	0.2982	0.3837
3	79.8137	0.2978	0.3838
4	80.3922	0.2878	0.38
5	80.3922	0.2926	0.3811
6	80.3137	0.3005	0.3807
7	80.1765	0.2934	0.3833
8	80.2451	0.3021	0.3838
9	80.5686	0.3002	0.3821
10	80.3137	0.3033	0.3838
15	80.2353	0.3066	0.3836
20	80.098	0.2973	0.3881
25	80.1569	0.3107	0.386
30	80.1176	0.3066	0.3853
40	80.4804	0.3038	0.3866
50	80.6961	0.2936	0.3855

4. Conclusion

In this study, credit card clients' behaviours about payment have been estimated. For this purpose machine learning algorithms like kNN and MLP have been used. The estimation success rates and error values of kNN and MLP were calculated. It was observed that the success rate was higher for the estimation performed by using the MLP algorithm. The highest estimation success rate was achieved when there was only one neuron in the hidden layer was 1 and the success rate was 81.049%. The MAE error value was 0.3237 and the RMSE value was 0.388 when there is only one neuron in the hidden layer. For the estimation success rates obtained using K-Nearest Neighbour Algorithm, the highest estimation success rate was achieved for 10 neighbourhood values, and it was 80.6569%. For this neighbourhood value, the MAE and RMSE value were obtained as 0.2743 and 0.3897 respectively.

References

[1] O'Sullivan, A., Sheffrin, S. M., (2003). Economics: Principles in action. Upper Saddle River, New Jersey 07458: Pearson Prentice Hall. p. 261. ISBN 0-13-063085-3.

[2] Chen, M. S., Han, J., & Yu, P. S., (1996). Data mining: an overview from a database perspective. Knowledge and data Engineering, IEEE Transactions on, 8(6), 866-883.

[3] Özekes, S., (2003). Data mining methods and application areas (Verimadenciliğimodelleriveuygulamaalanları). Istanbul TicaretUniversitesiDergisi, vol 3, 65-82 (Turkish).

[4] Sharma, T. C., & Jain, M. (2013). WEKA approach for comparative study of classification algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 2(4), 1925-1931.

[5] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.

[6] Witten I.H., Frank E., & Hall M.A. (2011). Data mining: practical machine learning tools and techniques. Elsevier, London.

[7] Patterson, D., Liu, F., Turner, D., Concepcion, A., & Lynch, R., (2008). Performance Comparison of the Data Reduction System. Proceedings of the SPIE Symposium on Defense and Security, Mart, Orlando, FL, pp. 27-34.

[8] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009). The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter, 11(1), 10–18.

[9] Wang, J., Neskovic, P., & Cooper, L. N., (2007). Improving nearest neighbour rule with a simple adaptive distance measure, Pattern Recognition Letters, 28(2):207-213.

[10] Zhou, Y., Li, Y. & Xia, S., (2009). An improved KNN text classification algorithm based on clustering, Journal of computers, 4(3):230-237.