Performance Comparison of Data Mining Algorithms for the Predictive Accuracy of Credit Card Defaulters

Dr. Maruf Pasha[†], Meherwar Fatima[††], Abdul Manan Dogar[†††] and Furrakh Shahzad[††††]
[†]Department of Information Technology, Bahauddin Zakariya University, Multan 60000, Pakistan
[††]Institute of CS & IT, The Women University Multan, Multan 60000, Pakistan
[†††]Department of Computer Science, CIIT, Sahiwal, Pakistan
[††††]Department of Computer Science, Pakistan Institute of Engineering and Technology, Multan 60000, Pakistan

Summary
The use of credit card for a secure balance transfer is a need of time. Fraudulent activities are also arising due to the fast growth of transactions. The motive of this research is to compare the predictive accuracy of customer's default payments using different data mining techniques. Accuracy can be predicted in more compact form than just describing binary result classification of "Credible" or "Not Credible" in respect of risk management. Normally, "defaulters" actual chance of default is mysterious. Six data mining techniques (FLDA, Naïve Bayes, J48, Logistic Regression, MLP, and IBK) are applied to the data-set. The results of this research indicate that the neural network performs best to predict the default of credit card clients and shows the highest accuracy.

Key words:
Data mining algorithms, Credit card defaulters, Performance of data mining, Predictive accuracy of credit card defaulters

1. Introduction

Data mining is burgeoning new research area for the detection of credit card defaulters. In banking Industry, Credit Card development is a remarkable occurrence. However, in 1730, the first credit card was issued. The usage of the plastic card has been massively increased to purchase goods. By using this plastic card, individual customers and company users are kept away from risks (e.g. fraud and robbery) [1]. Extensive use of credit cards is the cause of competition in the credit industry. Therefore, there is a need to expand and apply machine learning techniques to manage the data. It will save time and reduce errors [2]. Fraudulent Activities are also increasing due to the fast growth in credit card transactions. Fraud is widespread term [3]. Data mining contains a variety of techniques that are involved in investigating the accessible data and summarize it into valuable information. In computer science, data mining is used to detect patterns and relations between large amounts of data in the giant relational databases. For many ML algorithms, these patterns are used as input. Data mining algorithms are widely used for feature selection, classification, clustering and rule framing. The use of data mining in the banking sector is constantly increasing. This may also be used for decision support systems. Machine learning includes many classification algorithms that are used to divide data into several recommended number of categories [4]. Managing risk is essential for business. Different companies implement different applications to avoid risks. With the emergence of machine learning techniques, such models are being tried to build that can do risk analysis by examining the customer profile. For the experimental purposes, we used the data-set of Default of Credit Card clients from the 'UCI Machine Learning Repository.'

2. Related Work

Relative learning on the classifier's performance is done by Ajey et al. to predict credit cards defaulter. In this paper, the performance of data mining algorithms, name as Bayes Net, Meta-Stacking, Random Forest [21], Naïve Bayes, SMO [17] and Zero R is being discussed. For the performance analysis of algorithms, the data-set is taken from the UCI. "Feature selection" process has also been acknowledged. Both FS methods, "Correlation and information gain feature selection" provides most valuable or useful predictive features. Random Forest Ensemble method present the uppermost accuracy for the prediction of the default of credit card clients. It has been done with the experimental results [4].

Alejandro Correa Bahnsen et al., revealed the significance of the usage of the genuine monetary expenses of credit card fraud according to the algorithms of credit card fraud detection. Moreover, it is significant to have actual FN cost against transaction because it is not sufficient to have a fixed differentiation between false positive and false negative. Furthermore, their evaluations verified that comprising the real price by producing a cost sensitive scheme using the Bayes Minimum Risk classifier, meets to a great extent of fraud detection consequences in the condition of advance investments [5].

Those papers in which data mining methods are applied in credit risk valuation were reconsidered by the researchers

Manuscript received March 5, 2017
Manuscript revised March 20, 2017

[6]. Ten data mining techniques in the credit risk assessment framework were extorted, and after that, they explored almost all papers from 2000 to 2010 which had focused on these data mining methods. In recent years, SVM [22] has been extensively useful. Given the fact that to enhance model's predictive performance, a technique for lessening the attribute subset is required. Several hybrid SVM based model has been planned. The majority of among these projected models, customers can only be classified into two classes "good" or "bad."From the aspect of management of risk, each applicant who applied for credit can be predicted by the possibility of a default, would be more consequential than organizing them into the binary classes. Models have been proposed for this purpose, and they are effortless to read and recognize.

Byanjankar A. et al., proposed a model for credit scoring by applying neural networks to classify "peer-to-peer" applications of the loan into the groups of default and non-default. They compare artificial neural network scheme with a logistic regression model. The result indicates that neural network performs more precisely in screening the loans of default. In selecting a loan application, smart decisions are made by this neural network model. However, earlier identification of loan applications of default permits the lenders to lessen their financial loss by eliminating the chance of investing in bad applicants. Result suggest that neural network performs best in showing bad applications. But this research only focuses on one Peer-to-peer lending case, and after this, analysis of resemblances and differences (in fallouts) from other Peer-to-peer lending cases will be noticed [7].

3. System Design

This system is designed for the Performance Comparison of Data Mining Algorithms for the Predictive Accuracy of Credit Card Defaulters. Diagrammatic presentation of the system is shown in Fig. 1.

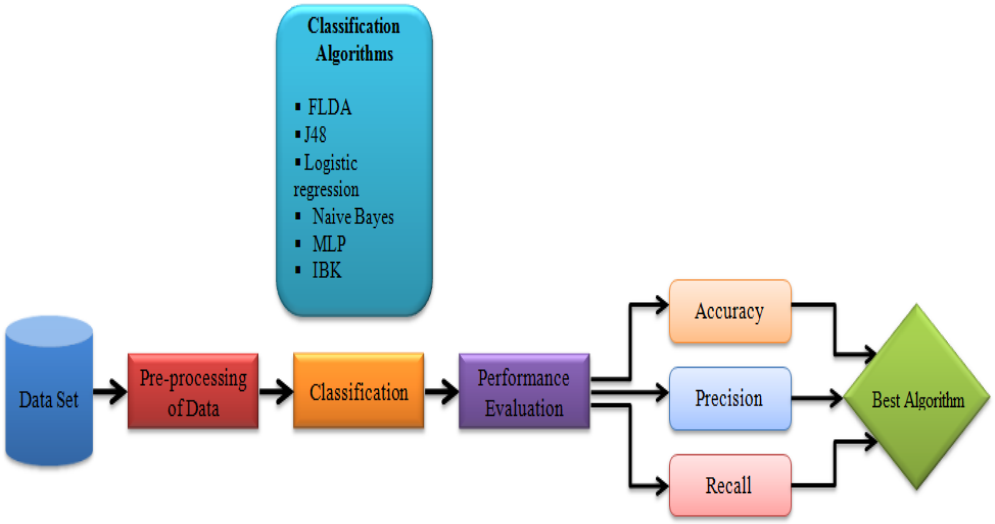

Fig. 1 System Design

This system contains a data-set that is obtained from UCI machine learning Repository [8]. Then data goes through the step of pre-processing. After that, Classification algorithms are applied to predict the performance. In data mining, classification phase recognizes the items in a group and places them under target categories. In this paper, Performance of algorithms is evaluated through Correct Classification Rate (Accuracy), In-Correct Classification rate, precision, and recall.

3.1 Data-set and Description

UCI machine learning Repository offered a data-set of customers default payments in Taiwan. Data-set comprises of 30000 instances and 24 attributes. There are no missing values in this data-set. This data set is in .xls format. Data-set detail is shown in Fig. 2.

Fig. 2. Attribute Detail

Attributes are indicated as X1, X2..... X23. Y is the last attribute of data set. X1 attribute named as "LIMIT _BAL." X2, X3, X4 and X5 shows "SEX (Male=1, Female=2)", "EDUCATION(Graduate School=1, University=2,High School=3,Others=4.)", "MARRIAGE (Married=1,Single=2,Others=3.)" and "AGE(in year)" respectively. X6 to X11 attribute shows "PAY_0", "PAY_2", "PAY_3", "PAY_4", "PAY_5", " PAY_6", that describe the Repayment status in September 2005. X6 indicates that repayment status in 2005. X7, X8, X9, X10 and X11 shows the repayment status in August, July, June, May, and April 2005. X12 to X17 attribute describes the amount of bill statement. They are named as "BILL_AMT1", "BILL_AMT2", "BILL_AMT3", "BILL_AMT4", "BILL_AMT5" and "BILL_AMT6". X12 shows the Amount of Bill account in September 2005. X13, X14, X15, X16 and X17 describes Amount of Bill account in August, July, June, May, and April 2005. X18 to X23 attributes name are "PAY_AMT1", "PAY_AMT2", "PAY_AMT3", "PAY_AMT4", "PAY_AMT5" and "PAY_AMT6". It shows the amount of preceding payment. X18, X19, X20, X21, X22 and X23 attribute shows Amount paid in September, August, July, June, May, and April 2005 respectively. Y attribute present "DEFAULT PAYMENT NEXT MONTH."This is a response variable Values are Yes=1, No=0.

3.2 Data Pre-Processing

Data-set is in .xls format. It is first converted to .CSV format and then into Attribute-Relation File Format (ARFF) and passed into the WEKA: data mining tool [9]. Description of data-set states that data-set doesn't miss any data.

3.3 Classification Algorithms/Methodology

Different machine learning algorithms are used for different analysis. There are two types of learning that are commonly used: (i) supervised learning [24] and (ii) unsupervised learning [23].

Supervised learning: In supervised learning, correct targets are available, on the basis of these targets, algorithms respond appropriately to all possible inputs.

Un-Supervised learning: Correct responses are not available. On the basis of similarities, it classifies the data.

In this experiment, six techniques of supervised learning are used to find the default of credit card clients. Techniques are FLDA [18], Naïve Bayes [19], J48, Logistic Regression, MLP, and IBK [20].

3.3.1 Linear Discriminant Analysis

This is a famous classification technique. Its function optimally separates two groups that give utmost distance among their relevant means [10]. It is most commonly employed a technique for dimensionality reduction. It is used in preprocessing for pattern recognition in machine learning. The purpose of LDA is to reduce data dimension. It avoids over-fitting and under-fitting problems. In this way, the computational cost is also reduced [11]. R.A. Fisher presented LDA in 1936. The basic purpose of LDA was to use it as a classifier, to separate different classes in data. Original LDA classifier was proposed for the2-classproblem; later it was enhanced to multi-class classification problem.

The basic aim of Linear Discriminant analysis is to plan feature space onto a shorter subspace $k (k \leq n-1)$. In the meantime, the discriminatory information is maintained. Due to dimensionality reduction, we can reduce not only the computational cost but also the classification issues like over-fitting and under-fitting of data.

3.3.2 Naïve Bayes

The algorithm of Naïve Bayes is a simplest probabilistic classifier. The algorithm of Naïve Bayes calculates the set of probabilities by measuring the occurrence and counting the group of values in given specified data-set. Bayes's theorem and theory of probability are the basis for Naïve Bayesian classifier. It supposes that all attributes are independent given the value of a variable of the class. In real world applications, this conditional independence assumption hardly true. However, this algorithm performs very well and learn quickly in different supervised classification problem [12].

3.3.3 J48

The j48 algorithm is a simple Java implementation of a C4.5 decision tree for classification. The developer of the

C4.5 algorithm is Ross Quinlan. It is employed to produce decision tree. In classification problems, decision tree algorithm is very useful. The tree is constructed, by using the J48 algorithm, to model the classification process. When the tree is built, it is employed to each row of data-set. J48 algorithm avoids missing values (that item value can be forecasted by what is known about the attributes values for other records) during tree construction. Core intention is that data is partitioning into range. It relies on value for that entry that is in training example data. J48 algorithm permits classification through "DT" or rules produced from them [13],[12].

3.3.4 Multilayer Perceptron

Many neural networks have been built and examined that consist of Hopfield network, self-arranging neural networks, mean-field theory machine, RB (radial basis) function and multi-layer perceptron. For the problems of a larger domain, MLP is very important technique [14]. Feed forward neural networks are Multilayer perceptrons (MLPs) that are trained with standard backpropagation algorithm. Because this algorithm belongs to supervised learning, therefore they need correct targets to be trained. They broadly used in pattern classification, because of the capability to learn how to convert the input to required target. They can estimate nearly any input-output map with one or two hidden layers. Many neural networks engage Multi-layer perceptrons. Perhaps, it is the most widely used network architecture. Units are organized in a topology of layered feed-forward. The every unit performs a biased weighted sum of their inputs are performed by every unit. It then goes through the activation function, to generate their output. Multilayer perceptron has a model that contains input, output, weights and thresholds value [15].

3.3.5 Logistic Regression

Model of logistic regression is a statistical tool that normally used. It prognosticates the relationship of items between more than two groups. The constraint in this model is that nature of the target variables should be binary. Analogous to multiple regression, a powerful technique is offered by it. It also provides ANOVA for uninterrupted answers. A function of independent variables y1, y2, y3…yn with responses that are binary in nature, is a part of an exponential family with "log(∏1/(1-∏1),……log(∏n/(1-∏n)" as a canonical parameter. In scientific conditions the correlation among a canonical parameter and the vector of descriptive variables x is declared as:

$$\text{Log}(\prod i(1-\prod i)=x\beta i$$

Linear membership among Explanatory variables in vectors and logarithm of odds creates a membership of non-linear among the possibility of y that equals to 1 and explanatory variables in the vector.

$$\prod i\exp(x\beta i/(1+\exp(x\beta i))$$

For dealing with classification problems, logistic regression is an appropriate algorithm. However, calculated outputs can be showed as probabilities [10].

3.3.6 IBK

K-nearest neighbor classifier is IBK algorithm in Weka: data mining tool. This algorithm uses the similar distance metric. In object editor, the amount of nearest neighbor can be defined clearly. It can be automatically determined by using leave-one-out cross-validation emphasize to an upper bound given by the particular value. Dissimilar search algorithms are employed to find the nearest neighbors at high speed. Linear search is used as a default search, but there are further options containing "KD-trees," "ball trees" and "cover trees" [25]. As a parameter, distance function may be employed. The behind thing is similar as "IBL." That is a Euclidean distance; further choices are "Chebyshev," "Manhattan," and "Minkowski distances." Predictions/Guesses can be weighted, from one or more neighbor, following their distance from the test examples. Distance is converted into weights by applying two dissimilar formulas. Training examples that are held by the classifier can be limited by "window size" option. When novel examples are included, previous examples are removed to retain the number of training examples at this size [16].

3.4 Performance Evaluation

Performance evaluation is a major step in data mining. It highlights the progress of algorithms. In this experiment, Algorithm's performance is evaluated by Correct Classification (Accuracy), Incorrect Classification, Precision (Positive predictive value) and Recall (Sensitivity). Accuracy appears as a correctness of model when applied to data. When six Algorithms are applied to this data-set, the result of these algorithms in the form of Correct Classification (Accuracy), In-Correct Classification, Precision, and Recall is shown in Table 1.

Table 1: Performance Evaluation of Data Mining Techniques

	Correct classification (Accuracy %)	Incorrect classification	Precision%	Recall%
FLDA	72.4	27.6	76.9	72.4
J48	80.3	19.7	78.2	80.3
Logistic regression	81	19	79.5	81
Naive Bayes	69.4	30.6	77	69.4
MLP	81.7	18.3	79.9	81.7
IBK	72.9	27.1	73	72.9

4. Result and Analysis

Classification algorithms are applied on the data-set through WEKA. Classification algorithm that performed best with the data-set of default of credit card clients is Multilayer Perceptron. Accuracy presented by MLP is 81.7%that is utmost as compared to other algorithm's predictive accuracy as revealed in Table 1. Precision and recall is 79.9% and 81.7% respectively. It's learning by example capability makes it superior to others. It has the highest coefficient of determination. The algorithm is fault tolerant through redundant information coding ability, whereas in other algorithms due to the low power of fault tolerance, partial destruction of network leads the algorithm to low efficient performance. Incorrect Classification rate of MLP is least as compared to other algorithms that are 18.3%. Fig. 3 shows the graphical representation of six data mining algorithm's performance.

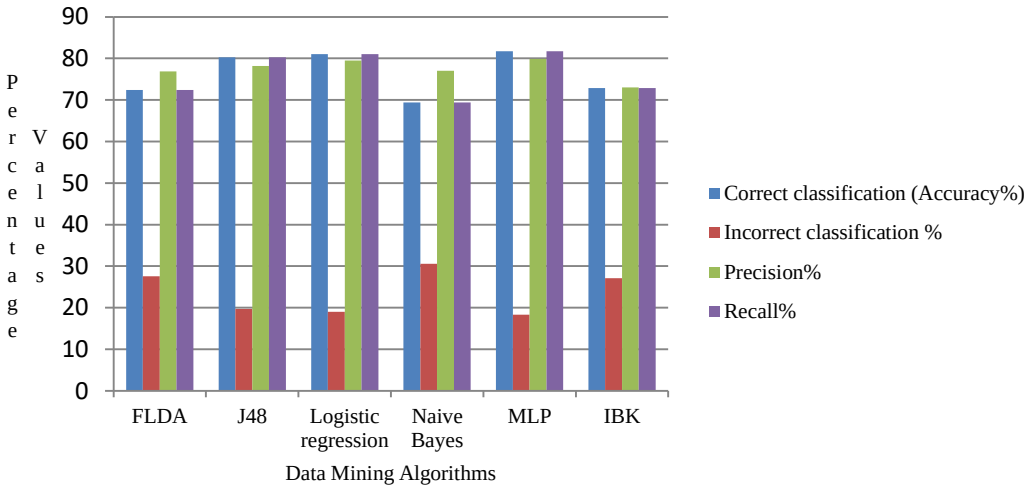

Fig. 3. Performance graph of Data Mining Techniques

Another strong point of multilayer perceptron is its reduced error rate, as it learns more and more with the passage of time and experience. Due to above mentioned rich features, Neural Network (multilayer perceptron) proved to be a good choice of other data mining algorithms.

5. Conclusion

Data mining algorithms play a vital role in the removal of manual errors and reduce dependency on human power. The research community is taking a keen interest in credit card fraud detection and defying the fraud. In this paper, Six data mining algorithms FLDA, J48, Logistic Regression, Naïve Bayes, MLP, IBK are applied to the data-set. Multilayer Perceptron performs best because of rich predictive features. It has prominent fault tolerance capability.J48, and Logistic Regression also shows a good predictive accuracy. These algorithm's applications are profitable and less error-prone. In banking area, application of classification algorithms is gaining strength, so it requires more analysis and expedition.

References

[1] Kiarie, Francis K., D. M. Nzuki, and A. W. Gichuhi. "Influence of Socio-Demographic Determinants on Credit Cards Default Risk in Commercial Banks in Kenya." (2015).

[2] Keramati, Abbas, and NiloofarYousefi. "A proposed classification of data mining techniques in credit scoring." *the Proceeding of 2011 International Conference of Industrial Engineering and Operations Management, Kuala Lumpur, Malaysia, Jurnal*. 2011.

[3] Raj, S. Benson Edwin, and A. Annie Portia. "Analysis oncredit card fraud detection methods." *Computer, Communication and Electrical Technology (ICCCET), 2011 International Conference on*. IEEE, 2011.

[4] Ajay, Ajay Venkatesh, ShomonaGracia Jacob. "Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers." International Journal of Computer Applications (0975 – 8887) Volume 145 – No.7, July 2016.

[5] Bahnsen, Alejandro Correa, et al. "Cost sensitive credit card fraud detection using Bayes minimum risk." *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*. Vol. 1. IEEE, 2013.

[6] Keramati, Abbas, and NiloofarYousefi. "A proposed classification of data mining techniques in credit scoring." *the Proceeding of 2011 International Conference of Industrial Engineering and Operations Management, Kuala Lumpur, Malaysia, Jurnal*. 2011.

[7] Byanjankar, Ajay, MarkkuHeikkilä, and JozsefMezei. "Predicting credit risk in peer-to-peer lending: A neural network approach." *Computational Intelligence, 2015 IEEE Symposium Series on*. IEEE, 2015.
[8] Default of credit cards client [online] http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
[9] Weka Data mining software [online] http://www.cs.waikato.ac.nz/ml/weka/
[10] Singh, Ravinder, and Rinkle Rani Aggarwal. "Comparative evaluation of predictive modeling techniques on credit card data." *International Journal of Computer Theory and Engineering* 3.5 (2011): 598.
[11] Linear Discriminant Analysis [online] http://sebastianraschka.com/Articles/2014_python_lda.html
[12] Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." *International Journal of Computer Science and Applications* 6.2 (2013): 256-261.
[13] Sharma, Aman Kumar, and SuruchiSahni. "A comparative study of classification algorithms for spam email data analysis." *International Journal on Computer Science and Engineering* 3.5 (2011): 1890-1895.
[14] Delashmit, Walter H., and Michael T. Manry. "Recent developments in multilayer perceptron neural networks." *Proceedings of the seventh Annual Memphis Area Engineering and Science Conference, MAESC*. 2005.
[15] Panchal, Gaurang, et al. "Behaviour analysis of multilayer perceptronswith multiple hidden neurons and hidden layers." *International Journal of Computer Theory and Engineering* 3.2 (2011): 332.
[16] Vijayarani, S., and M. Muthulakshmi. "Comparative analysis of bayes and lazy classification algorithms." *International Journal of Advanced Research in Computer and Communication Engineering* 2.8 (2013): 3118-3124.
[17] Sun, Zhaonan, et al. "Multiple kernel learning and the SMO algorithm." *Advances in neural information processing systems*. 2010.
[18] Gao, Quan-xue, Lei Zhang, and David Zhang. "Face recognition using FLDA with single training image per person." *Applied Mathematics and Computation* 205.2 (2008): 726-734.
[19] Karim, Masud, and Rashedur M. Rahman. "Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing." (2013).
[20] Salama, Gouda I., M. Abdelhalim, and MagdyAbd-elghanyZeid. "Breast cancer diagnosis on three different datasets using multi-classifiers." *Breast Cancer (WDBC)* 32.569 (2012): 2.
[21] Read, Jesse, et al. "Classifier chains for multi-label classification." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2009.
[22] Schuldt, Christian, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: A local SVM approach." *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. Vol. 3. IEEE, 2004.
[23] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. "Unsupervised learning." *The elements of statistical learning*. Springer New York, 2009. 485-585.
[24] Caruana, Rich, and AlexandruNiculescu-Mizil. "An empirical comparison of supervised learning algorithms." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.
[25] JacobGoldberger, SamRoweis, and RuslanSalakhutdinovGeoffHinton. "Neighbourhood components analysis." *NIPS'04* (2004).