International Journal of Scientific & Engineering Research, Volume 5, Issue 8, August-2014
ISSN 2229-5518

A Hybrid Approach for Automatic Credit Approval

E. Mortazavi, Dr. M. Ahmadzadeh

Abstract— Automatic credit approval is important for the operant processing of credit applications. It will prevent the credit card fraud. This paper proposes a hybrid approach, which combines the supervised tree classifiers with k-means clustering and feature selection to approve credit. Performance of this new approach is measured using the Credit Approval dataset and is shown to have high performance and accuracy.

Index Terms— Credit approval, Fraud, Supervised, Unsupervised, Accuracy.

1 Introduction

THE traditional information flow of data mining is used in implementation of data mining techniques for fraud detection, which begins with selection, preprocessing, transformation, data mining, interpretation and evaluation [1]. Today, it is common to use a credit card. Due to increasing credit card transactions, credit card fraud has become prevalent in recent years [2]. So nowadays, the need to detection systems is essential. By providing detection and prevention system from banks and financial institutions, card fraud is on the decline. Fraud prevention is the proactive mechanism with the aim of bringing down the incidence of fraud. Today credit approval has become an important issue in the banking sector. Automatic credit approval is the process of awarding credits to customers [3].

This paper combines the decision tree classifiers with k-means clustering and feature selection for automatic credit approval. First, feature selection is applied. Second, instances with missing values are eliminated. Third, the dataset is split into two clusters according to two classes by forming the clusters with k-means algorithm. Then the decision tree algorithms such as CART, random forest and J48 are applied on dataset and their performance are measured.

The rest of this paper is organized as follows: Section 2 describes the related work on credit card. Section 3 illustrates the methodology used which includes the supervised algorithms, unsupervised algorithms, Hybrid Approach and performance metrics. Section 4 gives the details of experiments and results. Section 5 concludes this paper.

2 Related Work

There are several types of studies in the domain of fraud detection such as credit card fraud detection.

Dheepa and Dhanapal [2] presented three methods to detect fraud. Firstly, they used clustering model to classify legal transaction. Secondly, they used Gaussian mixture model for modeling the probability density of credit card user’s past behavior. Finally, Bayesian networks are used in their study.

Chitra and Subashini [3] used the classification methods to build fraud detection models. In their work, the advantages of classification methods such as decision trees, Support Vector Machine (SVM) and Logistic Regression are shown to reduce the risk of banks.

Islam et al. [4] implemented naïve Bayes classifier and k-nearest neighbor classifier and applied them to the credit approval dataset. They showed that the performance of k-nearest neighbor classifier can be improved by varying the value of k.

Asha et al. [5] proposed a hybrid model for classification of the diabetic patient’s data. Hybrid model combines k-means clustering, k-nearest neighbor classification and correlation feature selection.

Da rocha and De souse [6] discussed on how decision trees are able to help in the prevention of bank fraud by the analysis of information regarding bank transactions. The information is obtained with the use of techniques and the CRISP-DM management model of data mining from internet bank transactions.

This paper investigates the usefulness of applying hybrid approach.

3 Methodology

3.1 Supervised Algorithms

E. Mortazavi, Student of school of Computer Eng. & IT, Shiraz University of Technology, Shiraz, Iran. Email: eleyeh.mortazavi@gmail.com
Dr. M. Ahmadzadeh, Head of School of Computer Eng. & IT, Shiraz University of Technology, Shiraz, Iran. Email: Ahmadzadeh@sutech.ac.ir

Supervised or classification is perhaps the best known data mining technique [3]. First, through the analysis of the training records of a dataset, a model is constructed. Then, the constructed model is used for classification. There are more classification methods that decision trees are one of them. In this paper, three decision tree algorithms are used.

3.1.1 CART

CART tree is a binary decision tree in which the node is split into two child nodes. Node splitting process is repeated until the tree is formed. To determine which node is the best choice for splitting, impurity measure is used. CART (Classification and Regression Tree) uses the Gini impurity measure. The node with the lowest value of Gini is selected. By summing the probability of each item being chosen times the probability of a mistake in categorizing that item, Gini impurity can be calculated [3]. The minimum value of Gini (zero) is when all the data in the node belong to a single target category.

3.1.2 Random Forest

Random forest is a class of ensemble methods which is designed for decision tree classifiers [7]. There is an original training set that bagging is used to produce the training dataset from an original training set. In bagging, N samples with replacement are randomly selected from an original training set. Random forest combines the predictions made by several decision trees [7].

Every decision tree uses a random vector. Random vector is produced from some fixed probability distribution. It randomly selects F input features (instead of all features at the training dataset) to split at each node of the decision tree. Then the tree grows without any pruning. After the formation of the trees, the predictions are combined using a majority voting scheme [7].

3.1.3 J48

J48 is one of the Decision tree techniques. J48 generates a decision tree from a set of labeled training data. It uses each attribute of the data to make a decision by splitting the data into smaller subsets. J48 uses the normalized information gain to determine which attribute will be decided for splitting. After calculating the normalized information gain, the attribute with the highest normalized information gain is selected for decision. Then, the algorithm repeats on the smaller subset. If all samples in a subset belong to the same class, the splitting process stops. J48 can handle training data with missing values [8].

3.2 Unsupervised Algorithms

Unsupervised or clustering is a data mining technique which divides the data into a number of clusters. Data within a cluster are most similar to each other and data in separate clusters are less similar to each other [7]. There are many methods for clustering such as partitioning methods, hierarchical methods and density based methods. In partition method, the most popular algorithm is k-means clustering which will be discussed in this paper.

3.2.1 K-means

K-means is one of the most popular clustering algorithms and is a centroid based technique. First, k initial centroids are chosen that k is the number of clusters. Each point is then assigned to the nearest centroid. Each group of assigned points to a centroid forms a cluster [7]. To assign points to the nearest centroids, proximity measure is needed. To measure proximity measure, Euclidean distance measure is used in this paper. When all the points are assigned to the centroids and created the clusters, the centroids recalculate based on the mean of points within the cluster. The process of assigning points to clusters is repeated until no change occurs in clusters. K-means algorithm is repeated until no change occurs in clusters and consists of three steps:

- Determine the initial centroids.
- Determine the distance of each point from the centroids.
- Assign each point to the centroid based on the minimum distance [9].

3.3 Hybrid Approach

In this paper, a hybrid approach is proposed to improve the performance of CART, random forest and J48 classifiers. The proposed hybrid approach as follows:

First, feature selection is applied. Feature selection is the process of identifying the attributes which are more important. It has the advantage of reducing the size of dataset and reducing the computation time. Thus, significant attributes are selected by using best Feature Selection method to a Credit Approval dataset.

Second, instances with missing values are eliminated from the dataset, in which case the size of dataset is reduced.

Third, by using the k-means algorithm the dataset is split into two clusters according to two classes. After clustering, the dataset is changed and cluster attribute is added to the dataset.

Finally, the decision tree algorithms such as CART, random forest and J48 are applied on new dataset and their performance are measured. Fig. 1 displays the diagram of this approach.

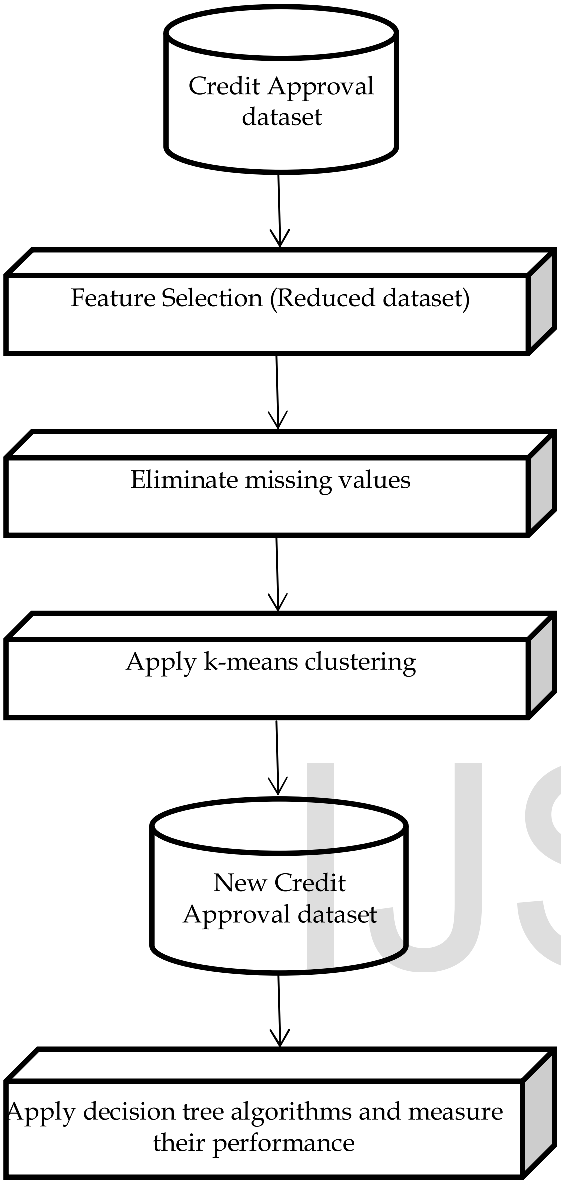
Credit Approval dataset
Feature Selection (Reduced dataset)
Eliminate missing values
Apply k-means clustering
New Credit Approval dataset
Apply decision tree algorithms and measure their performance

Fig. 1. Diagram of Hybrid Approach

3.4 Performance Metrics

To analyze the efficiency of the classifiers, a useful tool can be the confusion matrix. The confusion matrix is shown in Table 1.

- TP refers to the number of positive instances correctly classified by the classifier.
- FP refers to the number of negative instances incorrectly classified by the classifier.
- TN refers to the number of negative instances correctly classified by the classifier.
- FN refers to the number of positive instances incorrectly classified by the classifier [7].

TABLE 1
CONFUSION MATRIX

		Predicted Class	
Actual Class	Class	+	-
	+	True Positive (TP)	False Negative (FN)
	-	False Positive (FP)	True Negative (TN)

The performance metrics which uses in this paper are True Positive Rate (TPR), precision and accuracy. TPR and precision should be high to have high accuracy. TPR, precision and accuracy can be calculated using the following equations [7]:

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$$

TPR is the fraction of positive instances correctly predicted by the classifier. Precision defines the portion of cases really turns out to be positive in the group the classifier has announced as a positive class.

Besides the three evaluation measures, this paper further uses receiver operating characteristics (ROC) curves to compare the performances of the classifiers. ROC curve displays the tradeoff between true positive and false positive rate. It is a plot of TPR against FPR [7].

4 EXPERIMENTES AND RESULTS

In this paper, Credit Approval dataset [10] from UCI Repository of Machine Learning Databases and Domain Theories is used. This dataset concerns credit card applications and is provided by Quinlan in his studies. The dataset has 16 attributes. In Credit Approval dataset, all attribute names and values have been changed to meaningless symbols due to protection of the confidentiality of the data. The list of attributes is given in Table 2.

The dataset is interesting because there is a good mix of attributes: continuous, nominal with small numbers of values, and nominal with large numbers of values. There are 690 instances in this dataset, that 307 (44.5%) instances have positive class label (credit approved) and 383 (55.5%) instances have negative class label (credit denied).

International Journal of Scientific & Engineering Research, Volume 5, Issue 8, August-2014
ISSN 2229-5518

There are also a few missing values, 37 (5%) instances have some missing values.

TABLE 2
THE CREDIT APPROVAL DATASET

Attribute	Type	Values
A1	Nominal	a, b
A2	Continuous	13.75 – 80.25
A3	Continuous	0 - 28
A4	Nominal	u, y, l, t
A5	Nominal	g, p, gg
A6	Nominal	c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff
A7	Nominal	v, h, bb, j, n, z, dd, ff, o
A8	Continuous	0 – 28.5
A9	Nominal	t, f
A10	Nominal	t, f
A11	Continuous	0 - 67
A12	Nominal	t, f
A13	Nominal	g, p, s
A14	Continuous	0 - 2000
A15	Continuous	0 – 100000
Class	Nominal	+, -

Experiments were performed in Weka (Waikato environment for knowledge analysis) which contains tools for data preprocessing, classification and clustering [11]. 10 folds cross-validation was used in the experiments.

4.1 Experimental Results of Decision Tree Classifiers

In this phase, decision tree algorithms such as CART, random forest and J48 are applied to classify the dataset. The performance of these models are measured based on TPR, precision, ROC and accuracy. Table 3 illustrates the performance of the decision tree algorithms.

TABLE 3
THEPERFORMANCE OF THE DECISION TREE ALGORITHMS

Classifiers	TPR		Precision		ROC	
	+	-	+	-	+	-
CART	0.91	0.81	0.79	0.92	0.84	0.84
Random Forest	0.84	0.87	0.83	0.87	0.91	0.91
J48	0.84	0.88	0.85	0.88	0.89	0.89

The accuracy of the decision tree algorithms is given in the Fig. 2.

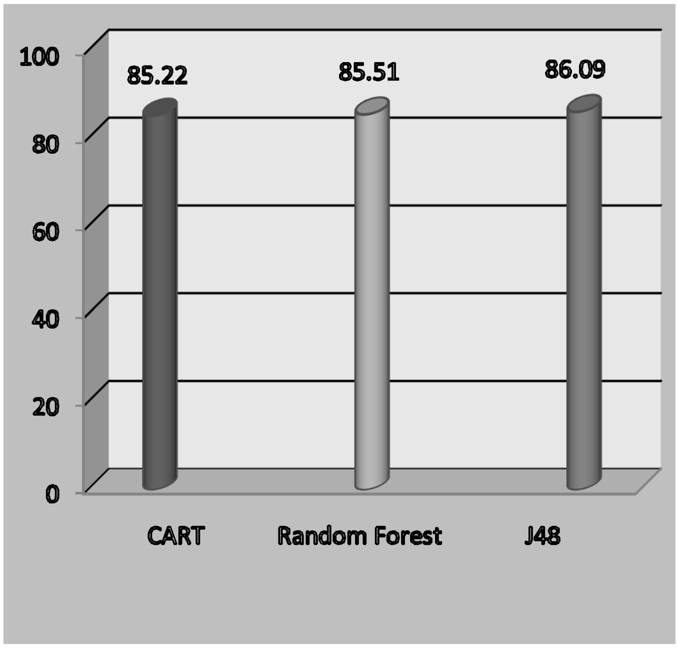

Fig. 2. Accuracy of Decision Tree Algorithms

Fig. 2 shows that the accuracy of J48 higher than the CART and random forest.

4.2 Experimental Results of Hybrid Approach

In this phase, the proposed hybrid approach is applied. First, the attributes which are most significant are selected by using best Feature Selection method. Credit Approval dataset has 16 attributes. In this step, the number of attributes is reduced to 8. Now, the dataset includes A4, A6, A8, A9, A11, A14, A15 and Class attributes.

Second step is eliminating the instances with missing values. The dataset has 690 instances. In this step, the number of instances is reduced to 674.

Third step is applying k-means clustering. K-means splits the dataset into two clusters. In this step, the dataset is changed and cluster attribute is added to the dataset. Now, the new dataset includes A4, A6, A8, A9, A11, A14, A15, Cluster and Class attributes.

Final step is applying CART, random forest and J48 decision tree algorithms on new dataset. In this step, the performance of these algorithms are measured based on TPR, precision, ROC and accuracy. Table 4 illustrates the performance of the decision tree algorithms using hybrid approach.

TABLE 4
THEPERFORMANCE OFHYBRID APPROACH

Classifiers	TPR		Precision		ROC	
	+	-	+	-	+	-
CART	0.96	0.94	0.93	0.96	0.93	0.93
Random Forest	0.95	0.93	0.91	0.96	0.95	0.95
J48	0.96	0.94	0.93	0.97	0.93	0.93

The accuracy of the decision tree algorithms using hybrid approach is given in the Fig. 3.

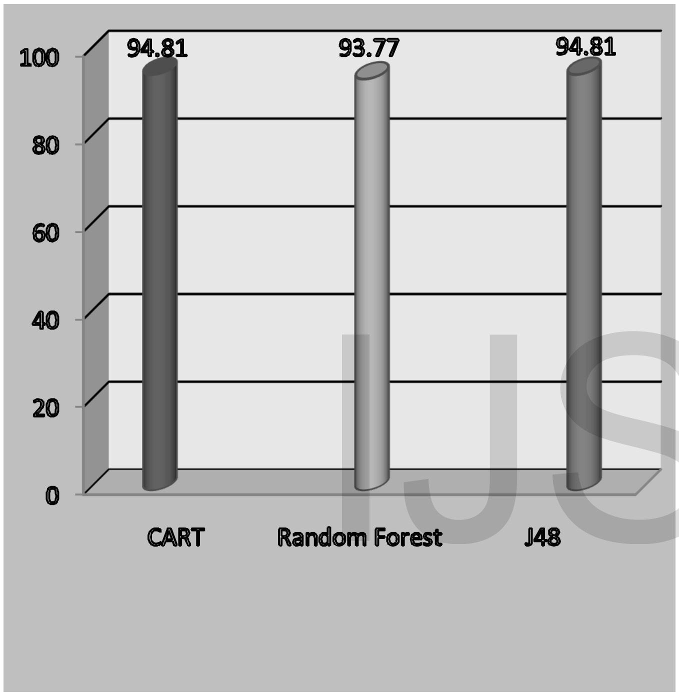

Fig. 3. Accuracy of Hybrid Approach

Fig. 3 reveals that the accuracy of CART and J48 is same and higher than the random forest.

4.3 Comparison of Accuracy of Hybrid Approach against Decision Tree Algorithms

In this phase, the accuracy of hybrid approach is compared against the decision tree algorithms. It is observed in section 3.1 that J48 decision tree has the highest accuracy among three decision tree algorithms mentioned above. Section 3.2 showed that the accuracy of CART and J48using hybrid approach higher than the random forest using hybrid approach. Fig. 4 illustrates the accuracy of hybrid approach and decision tree algorithms.

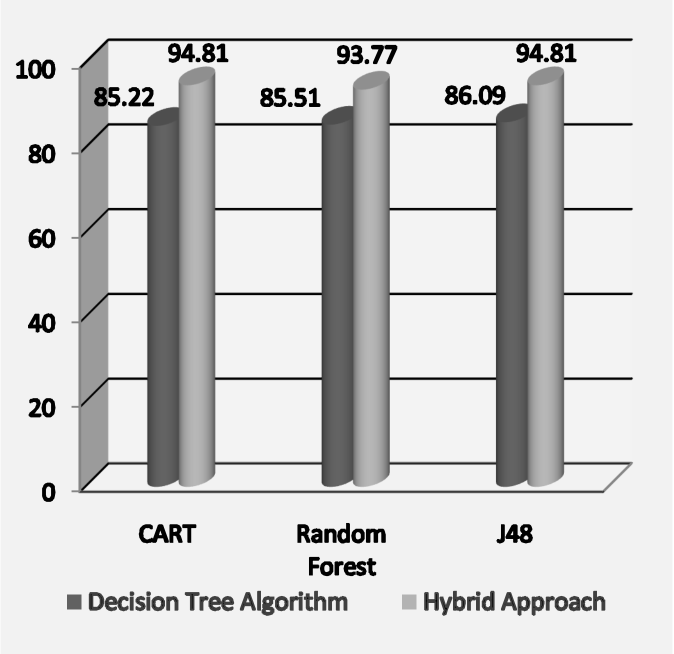

Fig. 4. Accuracy of Hybrid Approach and Decision Tree Algorithms

Fig. 4 shows that the CART and J48 using hybrid approach outperform all other algorithms to perform credit approval in Credit Approval dataset.

5 CONCLUSION

Credit approval is an important process for credit transactions. To detect and prevent credit card fraud, creation an efficient credit approval system is one of the key tasks for the banks and financial institutions. In this paper, a Hybrid Approach is proposed to improve the classification accuracy, which is based on combining feature selection, clustering and classification. The aim of this approach is to build fraud detection models. Experimental results of this new approach on Credit Approval dataset were compared with decision tree algorithms such as CART, random forest and J48. Among the algorithms CART and J48 using hybrid approach outperform all other algorithms to perform credit approval in Credit Approval dataset. The results showed that the CART using hybrid approach improved accuracy from 85.22% to 94.81%.

REFERENCES

[1] S. Sowjanya Chintalapati and G. Jyotsna, "Application of Data mining Techniques for Financial Accounting Fraud Detection Scheme," International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), vol. 3, pp. 717-724, issue 11, Nov. 2013,ISSN: 2277-128X.

[2] V. Dheepa and Dr.R. Dhanapal, “Analysis of Credit Card Fraud Detection Methods,” International Journal of Recent Trends in Engineering, vol. 2, no. 3, Nov. 2009.

[3] Dr.K. Chitra and B. Subashini, “Automatic Credit Approval Using Classification Method,” International Journal of Scientific & Engineering Research (IJSER), vol. 4, pp. 2026-2029,issue 7, Jul. 2013,ISSN: 2229-5518.

[4] M.J. Islam, Q.M.J. Wu, M. Ahmadi, and M.A. Sid-Ahmed, “Investigating the Performance of Naïve Bayes Classifiers and K-Nearest Neighbor Classifiers,” IEEE International Conference on Convergence Information Technology, pp. 1541-1546, Nov 21-23. 2007, Gyeongju, China, DOI: 10.1109/ICCIT.2007.148.

[5] Asha.T, S. Natarajan, and K.N.B. Murthy, “A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification,” Journal of Computing, vol. 3, issue 4, Apr. 2011, ISSN: 2151-9617.

[6] B.C. da Rocha and R.T. de Sousa, “Identifying Bank Frauds Using CRISP-DM and Decision trees,” International Journal of Computer Science & Information Technology (IJCSIT), vol. 2, no. 5, pp. 162-169, Oct. 2010. DOI: 10.5121/ijcsit.2010.2512.

[7] P.N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Pearson Education, 2006.

[8] J.R. Quinlan, C4.5: Programs for Machine Learning.San Mateo, CA: Morgan Kaufmann, 1993.

[9] D. Lavanya and Dr.K. Usha Rani, “A Hybrid Approach to Improve Classification with Cascading of Data Mining Tasks,” International Journal of Application or Innovation in Engineering & Management (IJAIEM), vol. 2, issue 1, Jan. 2013, ISSN: 2319-4847.

[10] C.L. Blake and C.J. Merz, “UCI Repository of Machine Learning Data bases,” University of California, Irvine, 1998, http://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/crx.data.

[11] Weka: Data Mining Software in Java, University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/ml/weka.