
Financial credit risk evaluation model using machine learning-based approach

Pankaj Kumar Jadwal and Sonal Jain

Department of Computer Science and Engineering,
JK Lakshmipat University,
Jaipur, India
Email: pankajjadwal@gmail.com
Email: sonaljain@jklu.edu.in

Basant Agarwal*

Department of Computer Science and Engineering,
Indian Institute of Information Technology Kota,
MNIT Campus, Jaipur, India
Email: thebasant@gmail.com
*Corresponding author

Abstract: Credit risk associated with individual at the time of sanctioning credit is one of the most crucial tasks for the banks and financial institutions. Financial organisations use credit score of a customer to decide whether a loan should be granted or it should be declined. Most of the existing machine learning models are trained to compute the credit risk associated with a person. However, the performance of the existing models are limited due to the noisy nature of available training data. Therefore, the pre-processing of the data contributes a crucial role in the performance of the credit risk evaluation model. In this paper, we propose an approach to develop an efficient credit risk evaluation model based on integrating a clustering algorithm with a classification algorithm. More specifically, initially, we transform the data into clusters of similar nature, and further, we develop different classification models for each cluster. Experimental results show that the proposed model improves the performance of the model by reducing the effect of the noise.

Keywords: credit risk; clustering; classification; hybrid model; agglomerative clustering; k-mode clustering.

Reference to this paper should be made as follows: Jadwal, P.K., Jain, S. and Agarwal, B. (xxxx) 'Financial credit risk evaluation model using machine learning-based approach', *World Review of Entrepreneurship, Management and Sustainable Development*, Vol. x, No. x, pp.xxx-xxx.

Biographical notes: Pankaj Kumar Jadwal has over ten years of academic experience and is currently associated with Swami Keshvanand Institute of Technology Management and Gramothan, Jaipur. Before joining SKIT, he has worked with Poornima University. He has various publications to his credit and has presented research papers at National and International conferences.

He is a life member of Computer Society of India. His areas of interest for research are applications of machine learning, data mining and analysis of algorithms.

Sonal Jain has over 19 years of academic experience and is currently associated with JK Lakshmipat University, Jaipur. Prior to joining JKL, she has worked with Gujarat Law Society, Institute of Computer Technology, affiliated to GTU. She has designed and conducted various faculty development programmes. She has several publications to her credit and has presented research papers at National and International conferences organized by institutes like IITB, IITK and, NTU Singapore. She has co-authored several books including *Personal Computer Software* (Wiley India), *Simplifying C* (Dreamtech Publications) and six textbooks for Gujarat State Education Board. She is a life member Computer Society of India and has served as management committee member of Ahmedabad Chapter. She has also been visiting faculty at the University of J. Seyle, Slovenia, Hungary and also visited Amsterdam University of Applied Sciences under faculty exchange program at JKL. She is also the recipient of Young Scientist Award for Initiatives Discoveries and Developments in the discipline of Information Retrieval by Venus International Foundation, Chennai, December 2016 and received Vaigyanik Drashtikon Award at 4th Rajasthan Science Congress.

Basant Agarwal is working as an Assistant Professor in the Department of Computer Science and Engineering at Indian Institute of Information Technology Kota (IIIT-Kota), MNIT Campus, Jaipur, India. He has more than ten years of experience in research and teaching. He has worked as a PostDoc Research Fellow at NTNU, Norway under prestigious PostDoc Fellowship by European Research Consortium for Informatics and Mathematics (ERCIM) through 'Alain Bensoussan Fellowship Programme' in 2016. He has also worked as Research Scientist at Temasek Laboratories, National University of Singapore (NUS) Singapore. He has authored two books and published more than 50 reputed conferences and Journal. He is also serving as a senior member, technical program committee member, member of editorial board/reviewer board of various renowned international conferences and journals.

1 Introduction

Financial agencies including banks are associated with several forms of risks mainly credit risk, investment risk, business risk and operational risk. As the banking sectors are competing with each other, the traditional banking profit is reducing day by day due to biasing towards customer retention while ignoring the risk associated with it (Crook et al., 2007). Problem associated with lending money forced banks to move towards consumer banking to make banking institutes more profitable. However, the overall process of loan disbursement is relying on banks itself, and the biasing towards audit procedure generate more risk in lending funds to borrowers. As a result, the probable risk gradually rises.

Accurate and on-time prediction of credibility of borrowers is very crucial in the banks which involve a decision between to grant a loan and to reject the application form. It is a non-trivial and challenging task that requires in-depth analysis of different factors of the customer. Due to exponential and compound rising in the credit department, various statistical and machine learning models have been introduced and implemented by various researchers. Credit scoring is a statistical technique which is used to assess the credit risk of loan applicants. The prime objective of implementation of these credit scoring models is to analysis the risk associated with a loan application and classifies them into accepted or rejected categories (Huang et al., 2004). It is highly beneficial for the financial organisations to consider credit risk scoring as it reduces the expense of credit analysis, and limits the possible risks.

There are two types of approaches for predicting the credit risk associated with loan applicants:

- 1 statistical approaches
- 2 machine learning algorithms.

It is evident in the literature that the machine learning techniques provide more predictive and accurate results compared to statistical techniques for credit risk prediction (Huang et al., 2004; Jain and Kumar, 2007).

The main objective of this study to propose an optimal credit risk evaluation model for financial institutions using a hybrid approach of combining clustering and classification techniques. The proposed approach initially utilises the K-mode and agglomerative clustering techniques to construct optimal clusters of the loan applicants and further various classification algorithms are applied on these clusters that provides better prediction performance.

The main contributions of this paper are as follows:

- 1 a hybrid approach is proposed by combing clustering and classification techniques that provides better performance for credit risk evaluation
- 2 we propose a new hybrid feature selection technique to select optimal features to develop credit risk evaluation model.

This paper is organised as follows. Section 3 briefly reviews machine learning techniques (supervised and unsupervised) used in the paper. In Section 4, research methodology along with dataset, deployment of model and evaluation parameters are discussed. Finally, Sections 5 and 6 talk about experimental results and conclusion, respectively.

2 Related work

Most of the existing methods for the credit risk evaluation model are based on supervised learning along with soft computing-based techniques. The existing methods focused on efficient feature engineering, hypertuning the parameters of the algorithm, and filtering sample of the data for learning (Agarwal and Mittal, 2012a, 2012b). Existing statistics-based methods for credit risk evaluation uses logistic regression (Yeh and Lien, 2009; Zhao et al., 2009), naive Bayes (Yeh and Lien, 2009), Discriminant

analysis (Xu and Wang, 2009), and data envelope analysis (Min and Lee, 2008). In addition, many supervised learning algorithms have also been used for determining the credit score for loan applicants such as decision trees (Sung et al., 2002), support vector machines (Min and Lee, 2005), case-based reasoning (Lin et al., 2008) and k-nearest neighbours (Yeh and Lien, 2009).

Hybridisation of existing heuristic methods like genetic algorithm (GA), particle swarm optimisation (PSO), ant colony optimisation (ACO) and annealing simulation (AS) are state of the art techniques used by researchers in credit risk domain. Min et al. (2011) applied GA to retrieve the optimal set of inputs and parameters of the support vector machine model to improve the prediction capability of credit risk evaluation. Furthermore, Tsai et al. (2013) used GA to obtain most relevant features and instances in a K-nearest neighbour (KNN) and support vector machine (SVM) classifier and concluded that prominent feature and instance extraction produces significant results in classification. Swarm intelligence and PSO were applied to obtain most prominent features automatically (Lin et al., 2008; Marinakis et al., 2009). The second strategy to design a hybrid learning model via cascading down classifiers, where the classification results obtained from initial classifier are further filtered by the different classifier in the later stage. Chuang (2013) proposed a hybrid model that performed better as compared to individual CBR-based approach for bankruptcy prediction with respect to the precision score. Zhu et al. (2018) proposed a hybrid model containing convolutional neural network along with feature selection algorithm relief, applied on the dataset retrieved from Chinese consumer finance company. Xu et al. (2018) came up with a hybridised machine learning model based on RIPPER algorithm, applied on the Taiwanese credit card consumer dataset where Relief and smote algorithms are used as preprocessing techniques to remove redundant features and balancing the dataset respectively. In the third category, unsupervised methods (clustering) are utilised in form of preprocessing before classification to improve the prediction precision. Instances causes to degrade the performance of the classifier can be removed with clustering methods, after that the clustering outcomes are utilised as training data to recognise the patterns by supervised learning (Jain et al., 1999).

Jain et al. (1999) applied CLUTO clustering and two classification methods to improve the prediction accuracy of the clustering outcome. Esfandiary and Moghadam (2013) suggested a layered decision tree (LDT) strategy which make optimal segments of the dataset and arrange them in decreasing order, after that splits the dataset into groups to build one layer of decision tree. Vang et al. (2018) applied clustering process to connect various film industries and also discussed the valuable insights and gaps in the relevant domain. Pellegrini et al. (2018) explored the overall journey of financial sector in the china and also discussed the popularity of social lending in China. And Ribeiro and Chen (2012) suggested a two-step strategy to forecast the financial risk of French business companies, initially finding the bicluster patterns keeping instances and variables extremely correlated, followed by a SVM model of the forecasted data. As per latest work on unsupervised learning, Lourenco et al. (2015) concluded that creation of ensemble model from a set of base clustering algorithms is certainly a rising research direction and it will surely improve the prediction accuracy of classification (Esfandiary and Moghadam, 2013).

3 Machine learning techniques

3.1 Classification techniques

3.1.1 Decision trees

A decision tree is a supervised learning algorithm. The reason behind the popularity of decision trees is because it is not a requirement of domain-specific knowledge. The most popular CART and C4.5 have been applied for bankruptcy prediction problems (Li et al., 2010). In another scholarly study, CART was applied to choose predictors, after that a traditional logistic regression model was built structured on the dummy variables compliance with the tree nodes (Brezigar-Masten and Masten, 2012). (Delen et al., 2013) used C5.0, CHAID, CART and QUEST to analyse the effect of financial ratios on company overall performance. Among four decision tree algorithms C5.0 and CHAID provided the optimal prediction accuracy.

3.1.2 Logistic regression

Logistic regression is one of the most widely used statistical model uses logistic function in order to train response variable. Parameters of the logistic model are estimated by logistic regression. Logistic regression may be divided into three categories on the basis of type of response variables. Categories are binary, multinomial and ordinal logistics.

In the recent literature, a hybrid model having non-parametric models (SVM, DT) and parametric model (binary logistic regression model) are proposed as a effective pairing centred on multiple discriminant analysis (MDA) to improve over-all performance on default prediction.

3.1.3 K-nearest neighbour

KNN algorithm is usually an extremely basic, however well known and effective non-parametric technique broadly utilised in supervised algorithms. In this algorithm, the test cases are classified by selecting k closest training instances and predictive output is decided upon majority voting of k nearest neighbours (Malekipirbazari and Aksakalli, 2015). Case-based reasoning (CBR) is capable of solving complex problems and can provide suggestions by identifying identical experiences. But in terms of precision, other methods like neural networks are providing better results than CBR (Jo et al., 1997). Park and Han introduced an improved version of CBR model with associated weights with features (Park and Han, 2002).

3.1.4 Support vector machine

A SVM is a supervised learning-based classifier that separates the data with an optimal hyperplane. This hyperplane categorises the training data according to their class attributes where each class lay in one side of the hyperplane. SVM is a well-known nonlinear classifier. Among existing nonlinear classifiers, it produces significant pros like, generalisation capability, can deal with local minima and requirements of low number of parameters (Maldonado et al., 2017a, 2017b). Researchers have shown

interest in solving credit risk problems of banking domain with SVM (Maldonado et al., 2017a; Li et al., 2017).

3.1.5 *Random forest*

Random forests (RFs) are known as universal classifier and generated by making forest of different decision trees. Yeh et al. (2012) suggested a hybridised model, which integrates RST and RF to draw out beneficial and useful information, and used KMV to assess the credit rating of companies from the market-based details.

3.2 *Clustering techniques*

Clustering (unsupervised learning) algorithms can be viewed as the procedure of segmentation of dataset into optimal groups called clusters with high intraclass similarity and low interclass similarity. Clustering algorithms generally are divided into three broad categories: partitioning methods, hierarchical methods and density-based methods. In the literature, the K-mode clustering and agglomerative clustering algorithms are most popular for credit risk evaluation problem (Azizpour et al., 2018; Chou et al., 2017; Zhao et al., 2017).

K-mode clustering is an extended version of K-means clustering in which instead of distances and means, dissimilarities and modes respectively are used.

Agglomerative clustering approach (Zhang et al., 2012) is based on bottom-up clustering technique where clusters possess sub-clusters and sub-clusters have sub-clusters, etc. Agglomerative clustering approach begins with each and every object, considering each object as a cluster. After that, in each consecutive iteration, it merges (agglomerates) the closest couple of clusters by fulfilling some similarity conditions, until all the objects of the dataset merge into a single cluster.

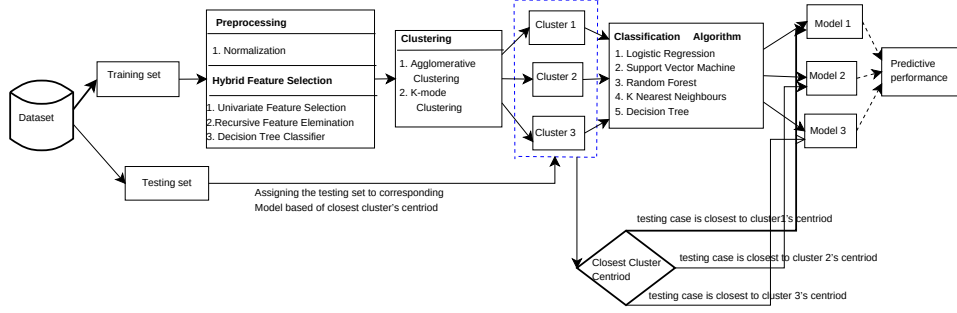
4 **Proposed approach**

In this paper, we focus on the problem of an efficient credit risk evaluation model that should be able to determine if a loan should be granted or declines to an applicant. The proposed approach combines a clustering algorithm and a classification algorithm to develop the final prediction model. We initially create different groups of data of applicants those are of similar nature, and further build prediction models each for these groups. At the testing time, we first find the closest group to the new sample, and use the prediction model corresponding to that group for prediction of the new sample.

More specifically, firstly, the dataset is pre-processed to impute missing values and feature engineering is applied to obtain prominent and optimal set of features. We propose a new hybrid feature selection algorithm to determine the best features by major voting of three algorithms namely univariate feature selection, recursive feature elimination, and decision tree classifiers feature selection algorithms. After extraction of prominent features, we use a clustering algorithm to divide the dataset into groups of similar kind of applicant. We use k-mode algorithm and agglomerative clustering algorithms for making the groups. We determine the number of clusters to be formed using elbow method. Finally, we build classification models for each of these groups

of data. The overview of the proposed approach for the credit risk evaluation model is shown in Figure 1. The detailed description of the proposed approach is given in subsequent subsections.

Figure 1 Overview of the proposed approach for credit risk evaluation model (see online version for colours)



4.1 Dataset and evaluation measures

To evaluate the proposed model, we use a popular and benchmark dataset of credit card applications, i.e., Australian credit dataset. The number of instances in the dataset is 690 and number of attributes are 15 including class attribute (six continuous and eight categorical). In the dataset, all inputs variables and values have been changed to symbols to guard the confidentiality of the dataset. Also, the dataset is comprised of nominal and continuous attributes. There are 37 missing values in the dataset which is 5% of the whole dataset.

4.2 Pre-processing of the dataset

In the pre-processing of the given data, the min-max normalisation technique is utilised for transforming all the attributes in a common range to avoid biasing among the attributes. The normalisation for an attribute x which has values in a range (Min, Max) can be transformed into a new range (New_Min, New_Max) using equation (1).

$$v' = \frac{v - Min(x)}{Max(x) - Min(x)} * (New_Max(x) - New_Min(x) + New_Min(x)) \quad (1)$$

4.3 Feature selection

The selection of the optimal features is very important to obtain good performance from the machine learning model. Irrelevant attributes degrades the performance of the model. Therefore, we propose a hybrid feature selection technique that combines three feature selection methods namely univariate, recursive feature elimination and decision tree techniques. The final optimal features are selected using majority voting. We apply three feature selection techniques to extract prominent features:

- 1 Univariate feature selection: statistical test can be used to choose attributes having the strong relation with the target class. In this paper, chi squared (χ^2) statistical test is used to select features that have strong correlation with the target class. The features with scores less than one are treated as irrelevant features, and features having score 1 or greater than 1 are considered as important features. Attributes along with relationship score with target class using univariate feature selection algorithm is shown in Table 1.

Table 1 Relationship score of independent variables

Feature	A8	A9	A10	A7	A5	A14	A6	A3	A4	A2	A13	A12	A11	A1
Score	170.75	82.97	43.64	16.94	15.76	14.88	5.63	5.45	3.14	2.13	0.72	0.44	0.37	0.04

- 2 Recursive feature elimination: the recursive feature elimination (RFE) algorithm selects important features by recursively removing attributes that are not relevant and further building machine learning model using remaining attributes. On the basis of the model accuracy, it identifies the significant attributes and the combination of attributes that contribute towards predicting the target class. In our experiments, we use logistic regression to develop machine learning model to determine the significance of the attributes. The features with ranking one are treated as prominent features others are considered as irrelevant features. The attributes along with their ranking using RFE algorithm is shown in Table 2.

Table 2 Ranking of independent variables

Feature	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
Ranking	4	6	5	1	1	3	1	1	1	1	1	2	1	1

- 3 Decision tree classifier: this feature selection algorithm is inspired from a bagging of decision tree classifier. We determine the significance of a feature on the basis of where they occur in the decision trees. The features that occur at the leaf nodes are considered as irrelevant attributes. The features having scores less than or equal to two are considered as irrelevant features. Important features along with their significance score using decision tree classifier feature selection algorithm is shown in Table 3.

Table 3 Feature along with scores using decision tree classifier

Feature	A8	A9	A10	A13	A3	A5	A2	A7	A14	A6	A4	A12	A1	A11
Score	0.40	0.08	0.06	0.06	0.05	0.05	0.05	0.05	0.04	0.03	0.03	0.02	0.02	0.01

- 4 Hybrid feature selection algorithm: after obtaining the important features according the three feature selection algorithms, we combine the results of all these algorithms to finally determine the prominent features for developing the prediction model. We determine the most significant features using majority voting, in this, we consider the feature as important if two out of three algorithms have suggested it to be relevant otherwise it is considered as irrelevant.

Table 4 shows the importance of the features based on hybrid feature selection algorithm. It can be seen from the table that the features {A2, A3, A4, A5, A6, A7, A8, A9, A10, A13, and A14} are considered as important features.

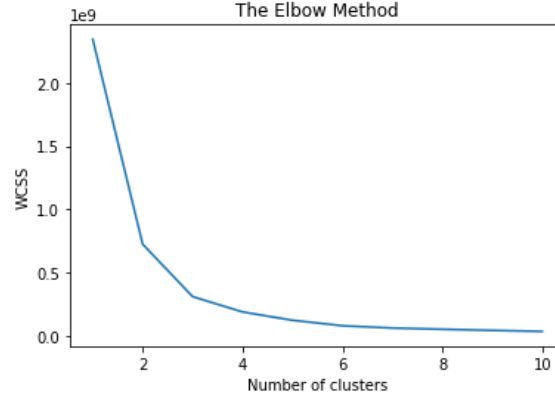
Table 4 Feature selection using hybrid feature selection method

<i>Feature</i>	<i>Univariate score</i>	<i>Recursive feature elimination ranking</i>	<i>Decision tree classifier score</i>	<i>Feature (Y/N)</i>
A1	0.04	4	0.02	N
A2	2.13	6	0.05	Y
A3	5.45	5	0.05	Y
A4	3.14	1	0.03	Y
A5	15.76	1	0.05	Y
A6	5.63	3	0.03	Y
A7	16.94	1	0.05	Y
A8	170.75	1	0.40	Y
A9	82.97	1	0.08	Y
A10	43.64	1	0.06	Y
A11	0.37	1	0.01	N
A12	0.44	2	0.02	N
A13	0.72	1	0.06	Y
A14	14.88	1	0.04	Y

4.4 Proposed hybrid prediction model

After obtaining the optimal features, we divide the dataset into groups of similar loan applicants using two clustering algorithms such as k-mode and agglomerative clustering algorithms. Further, we build different credit risk evaluation models for each of these groups. The proposed approach is a combination of supervised and unsupervised learning. Initially, clustering is used to make groups of similar credit card applicants. It is due to the reason that machine learning model works better on the dataset in which elements are identical to each other because of minimum variance between the elements of the dataset.

The number of clusters to be formed in the first phase of the proposed approach is decided by the elbow method. It draws a plot between number of clusters and obtained sum of squared error for the particular cluster size. As soon as number of clusters is increased, sum of squared error will be decreased. The elbow method try to find an elbow in the plot, after that, sum of squared error will be almost consistent for increasing cluster size (Ni et al., 2017). The elbow on the arm represents the optimal value of k and as per Figure 2 value of the elbow on the arm is 3. That is why, we have determined that three clusters should be formed. Plot between within cluster sum of squared errors and number of clusters is shown in the Figure 2.

Figure 2 Elbow method for obtaining optimal number of clusters (see online version for colours)

K-mode and agglomerative clustering is applied on the dataset and clusters of similar loan applicants are obtained. Now each cluster is assumed to be a independent dataset and different classification algorithms are applied on these clusters. Thus, we obtain different prediction models which are based on similar kind of loan applicants. At the testing of a new sample, we firstly determine the closest cluster centroid from the test sample, and further use the corresponding prediction model to identify the class label of the testing sample.

5 Experimental settings and results

We use 10 fold cross validation is used to evaluate the developed credit risk evaluation models. First of all, most widely used five supervised learning algorithms namely logistic regression, SVM, random forest, k-nearest neighbour, and decision tree algorithms are used to develop the prediction models. These five machine learning models are considered as baseline models to compare the performance of the proposed approach.

Table 5 shows the classification results of solitary baseline models. It can be seen from the results shown in Table 5 that the random forest classifier outperforms other classification algorithms with respect to the accuracy and F-score. Predictive accuracy, precision, recall and F-score is computed using equations (2), (3), (4) and (5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

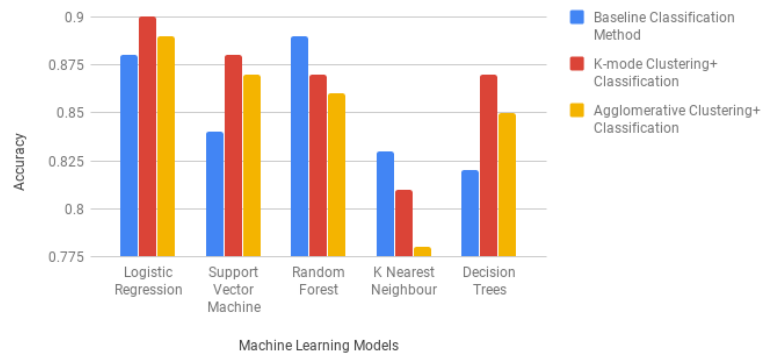
$$F - score = \frac{2 * precision * recall}{precision + recall} \quad (5)$$

Random forest gives an accuracy of 89% which is best as compared to other classifiers, whereas, the decision tree classifier provides lowest performance with accuracy of 82%.

Table 5 Performance results of various classification model

<i>Model</i>	<i>Prediction accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F-score</i>
<i>Baseline models</i>				
Logistic regression	0.880	0.848	0.778	0.812
SVM	0.840	0.909	0.751	0.822
Random forest	0.890	0.864	0.891	0.877
K-nearest neighbour	0.830	0.773	0.785	0.779
Decision tree	0.820	0.788	0.743	0.765
<i>Proposed hybrid models</i>				
Agglomerative+Logistic regression	0.890	0.716	0.787	0.749
Agglomerative+SVM	0.870	0.761	0.741	0.750
Agglomerative+Random forest	0.860	0.723	0.786	0.753
Agglomerative+K-nearest neighbour	0.780	0.716	0.753	0.734
Agglomerative+Decision tree	0.850	0.727	0.773	0.749
K-mode+SVM	0.880	0.841	0.794	0.816
K-mode+Random forest	0.890	0.811	0.763	0.786
K-mode+K-nearest neighbour	0.830	0.721	0.764	0.741
K-mode+Decision tree	0.820	0.761	0.718	0.738
K-mode+Logistic regression	0.900	0.871	0.779	0.822

Figure 3 Comparative performance of various classification models (see online version for colours)



We further evaluate the performance of the proposed models by including the clustering phase in the traditional classification models. We use K-mode clustering and agglomerative clustering to initially form the groups. Experimental results show that the k-mode clustering algorithm outperforms the performance of agglomerative clustering in all the cases. For example, logistic regression classifier with k-mode clustering gives the accuracy of 90%, whereas, LR with agglomerative clustering gives 89% accuracy as shown in Table 5. It can also be seen from the results in Table 5 that the

k-mode clustering algorithm along with classification algorithm performs better than the classification algorithm used without using any clustering algorithms in earlier phase.

The best performance is obtained by using k-mode clustering algorithm along with logistic regression classifier (as results shown in Table 5), it outperforms all the other machine learning models for the credit risk evaluation. We also show the comparative results of various prediction models in Figure 3.

6 Conclusions

We focus on the developing a machine learning-based model for credit risk evaluation model that can determine if a loan should be granted or not for the loan applicants. In this paper, we propose a new technique that is based on combining clustering algorithm along with a classification algorithm. The proposed approach uses clustering algorithms to group the similar loan applicant's data, and further use various classification algorithms to develop the model. We use various classification algorithms such as logistic regression, SVM, random forest, and decision trees algorithm. Experimental results show that the proposed approach outperforms all the other existing models for the credit risk evaluation. The proposed approach is very generic in nature and can be applied to any other classification problem. In future, we would like to see the robustness of the proposed approach to other classification problem. Besides that, other clustering algorithms like graph clustering and spectral clustering can be applied to deal with the sparseness of the dataset.

References

- Agarwal, B. and Mittal, N. (2012a) 'Comparative study of feature reduction and machine learning techniques for spam detection', *2nd International Conference on Soft Computing for Problem Solving (SocPros-2012)*, pp.761–760, Springer.
- Agarwal, B. and Mittal, N. (2012b) 'Text classification using machine learning methods – a survey', *2nd International Conference on Soft Computing for Problem Solving (SocPros-2012)*, pp.701–710, Springer.
- Azizpour, S., Giesecke, K. and Schwenkler, G. (2018) 'Exploring the sources of default clustering', *Journal of Financial Economics*, Vol. 129, No. 1, pp.154–183.
- Brezigar-Masten, A. and Masten, I. (2012) 'Cart-based selection of bankruptcy predictors for the logit model', *Expert systems with applications*, Vol. 39, No. 11, pp.10153–10159.
- Chou, C-H., Hsieh, S-C. and Qiu, C-J. (2017) 'Hybrid genetic algorithm and fuzzy clustering for bankruptcy prediction', *Applied Soft Computing*, Vol. 56, No. 7, pp.298–316.
- Chuang, C-L. (2013) 'Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction', *Information Sciences*, Vol. 236, No. 9, pp.174–185.
- Crook, J.N., Edelman, D.B. and Thomas, L.C. (2007) 'Recent developments in consumer credit risk assessment', *European Journal of Operational Research*, Vol. 183, No. 3, pp.1447–1465.
- Delen, D., Kuzey, C. and Uyar, A. (2013) 'Measuring firm performance using financial ratios: a decision tree approach', *Expert Systems with Applications*, Vol. 40, No. 10, pp.3970–3983.
- Esfandiary, N. and Moghadam, A-M.E. (2013) 'LDT: Layered decision tree based on data clustering', *2013 13th Iranian Conference on Fuzzy Systems (IFSC)*, pp.1–4, IEEE.

- Huang, Z., Chen, H., Hsu, C-J., Chen, W-H. and Wu, S. (2004) 'Credit rating analysis with support vector machines and neural networks: a market comparative study', *Decision support systems*, Vol. 37, No. 4, pp.543–558.
- Jain, A. and Kumar, A.M. (2007) 'Hybrid neural network models for hydrologic time series forecasting', *Applied Soft Computing*, Vol. 7, No. 2, pp.585–592.
- Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) 'Data clustering: a review', *ACM Computing Surveys (CSUR)*, Vol. 31, No. 3, pp.264–323.
- Jo, H., Han, I. and Lee, H. (1997) 'Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis', *Expert Systems with Applications*, Vol. 13, No. 2, pp.97–108.
- Li, H., Sun, J. and Wu, J. (2010) 'Predicting business failure using classification and regression tree: an empirical comparison with popular classical statistical methods and top classification mining methods', *Expert Systems with Applications*, Vol. 37, No. 8, pp.5895–5904.
- Li, Z., Tian, Y., Li, K., Zhou, F. and Yang, W. (2017) 'Reject inference in credit scoring using semi-supervised support vector machines', *Expert Systems with Applications*, Vol. 74, No. 8, pp.105–114.
- Lin, S-W., Ying, K-C., Chen, S-C. and Lee, Z-J. (2008) 'Particle swarm optimization for parameter determination and feature selection of support vector machines', *Expert Systems with Applications*, Vol. 35, No. 4, pp.1817–1824.
- Maldonado, S., Bravo, C., Lopez, J. and Perez, J. (2017a) 'Integrated framework for profit-based feature selection and SVM classification in credit scoring', *Decision Support Systems*, Vol. 104, No. 12, pp.113–121.
- Maldonado, S., Pérez, J. and Bravo, C. (2017b) 'Cost-based feature selection for support vector machines: an application in credit scoring', *European Journal of Operational Research*, Vol. 261, No. 2, pp.656–665.
- Malekipirbazari, M. and Aksakalli, V. (2015) 'Risk assessment in social lending via random forests', *Expert Systems with Applications*, Vol. 42, No. 10, pp.4621–4631.
- Marinakis, Y., Marinaki, M., Doumpos, M. and Zopounidis, C. (2009) 'Ant colony and particle swarm optimization for financial classification problems', *Expert Systems with Applications*, Vol. 36, No. 7, pp.10604–10611.
- Min, J-H., Jeong, C-W. and Kim, M-S. (2011) 'Tuning the architecture of support vector machine: the case of bankruptcy prediction', *Management Science and Financial Engineering*, Vol. 17, No. 1, pp.19–43.
- Min, J.H. and Lee, Y-C. (2005) 'Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters', *Expert systems with applications*, Vol. 28, No. 4, pp.603–614.
- Min, J.H. and Lee, Y-C. (2008) 'A practical approach to credit scoring', *Expert Systems with Applications*, Vol. 35, No. 4, pp.1762–1770.
- Ni, Q., Pan, Q., Du, H., Cao, C. and Zhai, Y. (2017) 'A novel cluster head selection algorithm based on fuzzy clustering and particle swarm optimization', *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, Vol. 14, No. 1, pp.76–84.
- Park, C-S. and Han, I. (2002) 'A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction', *Expert Systems with Applications*, Vol. 23, No. 3, pp.255–264.
- Pellegrini, C.B., Pellegrini, L. and Sergi, B.S. (2018) 'Governance and banking system morphology in China: the evolution over the last two decades (1995-2015)', *World Review of Entrepreneurship, Management and Sustainable Development*, Vol. 14, No. 5, pp.549–580.

- Ribeiro, B. and Chen, N. (2012) ‘Biclustering and subspace learning with regularization for financial risk analysis’, *International Conference on Neural Information Processing*, pp.228–235, Springer.
- Sung, T.K., Chang, N.S. and Lee, G.H. (2002) ‘Dynamics of modeling in data mining: Interpretative approach to bankruptcy prediction’, *Entrue Journal of Information Technology*, Vol. 1, No. 1, pp.63–78.
- Tsai, C-F., Eberle, W. and Chu, C-Y. (2013) ‘Genetic algorithms in feature and instance selection’, *Knowledge-Based Systems*, Vol. 39, No. 4, pp.240–247.
- Vang, J., Maher, S. and Brambini, A. (2018) ‘Emergence of film industries in small countries: on the needs for an integrated cluster and branding strategy’, *World Review of Entrepreneurship, Management and Sustainable Development*, Vol. 14, No. 5, pp.637–656.
- Xu, P., Ding, Z. and Pan, M. (2018) ‘A hybrid interpretable credit card users default prediction model based on ripper’, *Concurrency and Computation: Practice and Experience*, Vol. 30, No. 23, p.e4445.
- Xu, X. and Wang, Y. (2009) ‘Financial failure prediction using efficiency as a predictor’, *Expert Systems with Applications*, Vol. 36, No. 1, pp.366–373.
- Yeh, C-C., Lin, F. and Hsu, C-Y. (2012) ‘A hybrid KMV model, random forests and rough set theory approach for credit rating’, *Knowledge-Based Systems*, Vol. 33, No. 9, pp.166–172.
- Yeh, I-C. and 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*, Vol. 36, No. 2, pp.2473–2480.
- Zhang, W., Wang, X., Zhao, D. and Tang, X. (2012) ‘Graph degree linkage: Agglomerative clustering on a directed graph’, *European Conference on Computer Vision*, pp.428–441, Springer.
- Zhao, D., Huang, C., Wei, Y., Yu, F., Wang, M. and Chen, H. (2017) ‘An effective computational model for bankruptcy prediction using kernel extreme learning machine approach’, *Computational Economics*, Vol. 49, No. 2, pp.325–341.
- Zhao, H., Sinha, A.P. and Ge, W. (2009) ‘Effects of feature construction on classification performance: An empirical study in bank failure prediction’, *Expert Systems with Applications*, Vol. 36, No. 2, pp.2633–2644.
- Zhu, B., Yang, W., Wang, H. and Yuan, Y. (2018) ‘A hybrid deep learning model for consumer credit scoring’, *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp.205–208, IEEE.