Original Research

A New Multi-Stage Feature Selection and Classification Approach: Bank Customer Credit Risk Scoring

Farshid Abdi

Received: 2020-12-28 / Accepted: 2021-07-06 / Published online: 2021-08-21

Assistant Professor, Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran.
E-mail address: farshidabdi@azad.ac.ir

Abstract

Lots of customers information regularly are stored in the databases of banks. These databases can be used to assess the credit risk. Feature selection is a well-known concept to reduce the dimension of such databases. In this paper, a multi-stage feature selection approach is proposed to reduce the dimension of database of an Iranian bank including 50 features. The first stage is devoted to removal of correlated features. The second stage is allocated to select the important features with genetic algorithm. The third stage is proposed to weight the variables using different filtering methods. The fourth stage selects feature through clustering algorithm. Finally, selected features are entered into the K-nearest neighbor (K-NN) and Decision Tree (DT) classification algorithms. The aim of the paper is to predict the likelihood of risk for each customer based on effective and optimum subset of features available from the customers.

Keywords - Clustering; Credit risk prediction; filtering method; Genetic algorithm; Hybrid feature selection

INTRODUCTION

Due to recent credit crisis in financial institutions especially banks; credit risk prediction has become an increasingly important field in financial risk management. The credit crisis has significantly reduced the profit and causes bankruptcy of many banks. Credit risk is one of the most important risks in the banking system which can be defined as the possibility that a borrower does not fulfil its obligations and not repaid the loan to the bank (Danenas and Garsva 2015). Minimization of such risk while making correct credit granting decisions is critical for managing risk in financial institutions. Hence, existence of the automatic credit scoring systems become more important (Yu et al. 2011).

Credit scoring is usually accomplished through a set of decision making models and related techniques that help the loaners to make decisions about estimation of the credit of the customers (Thomas et al., 2002). In traditional credit scoring methods, the descriptive parameters of the customers are considered. Rating for each customer is calculated by subjective judgment of some bank experts. This type of is usually inexact, expensive and time-consuming. The automatic credit scoring methods improve accuracy, costs and time of prediction (Zhao et al. 2015). Vast amount of information and data that describes socio-demographic characteristics and economic conditions of the previous loan applicants are available in database of banks. The data in the databases can be used for credit risk assessments (Oreski and Oreski 2014).

Data mining techniques such as predictive models and classification can be utilized to construct the credit scoring models. Indeed, data mining techniques enables banks managers to analyse and explore useful information from their customer database (Yap et al. 2011). Therefore, historical data and demographic characteristics considered as an input of the data mining classifier and the output of it determines the credit conditions of the applicants (Marqués et al., 2012).

Bank databases usually have high dimensions. Pre-process of dataset to prepare it for classification and enhance the accuracy of prediction is an important task in data mining. Feature selection is a technique of data pre-processing that is usually implemented in the datasets with large number of variables and with the purpose of reducing irrelevant and redundant variables, facilitating understanding the data, improving the accuracy of prediction and enhancing the interpretability of the model (Oreski and Oreski, 2014; Oreski and Oreski, 2012).

The aim of this research is to predict the likelihood of risk for each customer based on effective and optimum subset of features available from the customers. The multi-stage feature selection approach combining genetic algorithm, filtering methods and clustering techniques is suggested for this purpose. The selected features are entered into the classification algorithms to predict customer credit risk. The proposed approach has been applied on a real case study.

The rest of the paper is organized as follows. In Section 2, a literature review on the subject of data mining techniques in credit risk scoring is presented. In Section 3, proposed methodology of the research is described. In Section 4, the results are represented and discussed. In Section 5 conclusions remarks and future research directions are presented.

LITERATURE REVIEW

Early detection of financial risks can help credit lenders and institutions to create appropriate policies for reduce losses and increase income. In recent years, several empirical studies have demonstrated that data mining techniques can be successfully used for credit risk management. It has been concluded that these techniques are performed better than traditional methods. Data mining methods do not assume subjective expertise and knowledge of the experts, but automatically extract information from past records of customers (Marqués et al., 2012). Artificial neural networks are the most common method for predicting credit risk (Zhao et al., 2015; Khashman, 2011; Khashman, 2010; Khashei et al., 2013). Zhao et al., (2015) proposed a multi-layer perceptron neural networks for credit scoring. Khashman (2010) trained three multi-layer supervised neural network based on the back propagation learning algorithm and under nine learning schemes. Khashman (2011) used neural network for credit risk evaluation under different learning schemes and suggested emotional neural network (EmNN) model. Khashei et al., (2013) presented a two-stage fuzzy hybrid classification method on the basis of traditional multilayer perceptron. Shen et al. (2019) proposed a novel ensemble model based on the SMOTE method and the PSO algorithm in order to address the problem of imbalanced data and used the combination of the AdaBoost algorithm with the optimised BP neural networks for credit risk evaluation.

Support Vector Machines (SVM) are another type of learning mechanisms, which were utilized in credit risk prediction. Ping and Yongheng, (2011) proposed SVM models to evaluate the applicant's credit score. Yu et al., (2011) used weighted least squares support vector machine (LSSVM) and design of experiment (DOE) for credit risk evaluation. Hens and Tiwari (2012) proposed a hybrid approach on the basis of SVM and F score to reduce the computational time of sampling. Harris, (2015) suggested clustered support vector machine (CSVM) and compared the CSVM with other nonlinear SVM for credit scoring problem.

Pławiak et al. (2019) proposed a novel deep genetic cascade ensemble of SVM classifiers named DGCEC, for credit scoring. The proposed model combined the advantages of evolutionary computation, ensemble learning, and deep learning.

Several researches employed ensemble methods to enhance credit modelling performance (Marqués et al., 2012; Wang and Ma, 2012). Wang et al., (2012) proposed RS-Bagging decision tree (DT) and Bagging-RS DT in order to improve the accuracy of model by reducing the noisy data and redundant variables. Papouskova, M., Hajek (2019) proposed a two-stage credit risk model in order to predict expected loss. The authors applied class-imbalanced ensemble credit scoring with regression ensemble.

Xiao et al., (2012) focused on imbalanced datasets and combined ensemble learning with cost-sensitive learning and suggested a dynamic classifier ensemble method for imbalanced data. Hsieh and Lun-Ping Hung, (2010) introduced class-wise classification as a pre-processing step in order to improve the performance of ensemble classifier.

Another artificial intelligence method that has been used in the field of credit scoring is decision tree (Wang et al., 2012; Bijak and Thomas, 2012; Yap et al., 2011). Bayesian network classifier (Wu, 2011; Zhu, Beling and Overstreet, 2002), K-nearest neighbour (Henley and Hand, 1996; Laha 2007; Lessmann et al., 2015) have also been used in this filed.

One of the main issues in credit scoring is the database with high dimension and the selection of the most appropriate and important subset of the features (Hajek and Michalak, 2013; Oreski et al., 2012). Khalili-Damghani et al. (2018) proposed a two-stage hybrid approach based on the combination of filtering and TOPSIS method. Khalili-Damghani et al. (2018) applied Genetic algorithm and a combination of filtering methods to select the proper features. Wang et al., (2017) proposed a two-phase hybrid approach based on filter approach and multiple population genetic algorithm to reduce the dimension of the dataset. Maldonado et al., (2017)

presented two SVM-based strategies for simultaneous classification and embedded feature selection. Abdi et al., (2017) used Wrapper techniques (GA and forward method) and filter techniques (Gini Index, correlation, and information gain) separately and in hybrid form to find the most proper features.
Hajek and Michalak, (2013) compared several filter and wrapper approaches for feature selection. Oreski and Oreski, (2014) proposed a hybrid genetic algorithm with neural network (HGA-NN) in order to select the optimum subset of features and increase the accuracy of the prediction model. Arora and Kaur (2020) proposed a novel feature selection method named Bolasso (Bootstrap-Lasso) in order to select consistent and relevant features from pool of features.
Nalić et al. (2020) applied combination of various feature selection and ensemble learning classification algorithms to propose a new hybrid credit scoring model. Rtayli et al. (2020) proposed a Credit Card Risk Identification (CCRI) model and applied Random Forest Classifier as a feature selection method.
Present paper also considered the importance of feature selection and reducing irrelevant and redundant variables and proposed a multi-stage feature selection approach.

METHODOLOGY

I. Proposed Model

The proposed approach of this study, as shown in Figure1, is composed of two main phases including: Feature Subset Selection, and modelling procedure. The main purpose of this research is to predict the credit of bank customers based on effective and compact and optimum subset of features. Therefore, one of the main phases of the research focuses on the feature selection methods and a multi-stage feature selection method is represented. In high-dimensional datasets feature selection is an important phase in data pre-processing. Indeed, the purpose of this step is to remove redundant and irrelevant features and select an effective subset of the features with the acceptable prediction capability.
As can be seen in Figure 1, the first stage of the proposed feature selection approach is addresses removing high-correlated features. Second and third stages related to wrapper and filter methods. The genetic algorithm, Gini index, Information gain, Gain ratio, Correlation, Relief and Rule are used as filtering and wrapper methods in this research. Finally, in the last stage of the feature selection phase, the clustering algorithm namely X-means is employed in order to achieve compact subset of customers' features. Afterwards the selected features are entered into K-nearest neighbour (K-NN) and decision tree (DT) classification algorithms to estimate the credit score of the bank customers. In order to evaluate the methods, the classification accuracy metrics are calculated. The proposed approach has been applied in a real case study of credit customers of a bank in Iran. In this section the techniques used in the research are briefly explained.

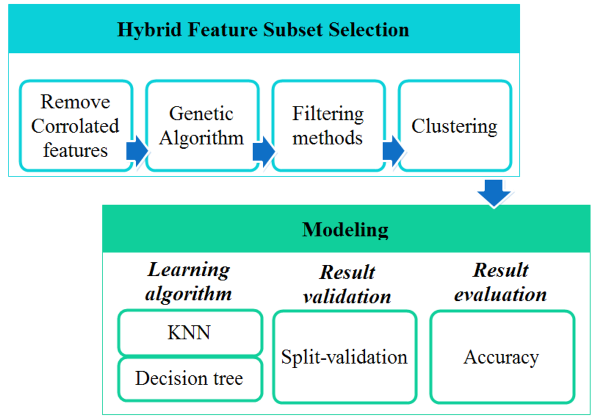

FIGURE 1
PROCESS OF THE RESEARCH

II. Feature Selection

Feature selection methods are divided into four broad groups: filter methods, wrapper methods, embedded methods, and hybrid methods (Moradkhani et al., 2015; Guyon and Elisseeff, 2003). The aim of these algorithms is to exclude irrelevant or redundant features in order to prepare data for classification and clustering (Tsai et al., 2013).
Using proper methods of feature selection improves the accuracy of classification algorithms, reduces the over-processing and complexity of calculations in algorithm, and makes the classification algorithm more generalized (Moradkhani et al., 2015; Oreski and Oreski, 2014).
In filtering methods, each feature is weighted on the basis of relationship between that feature and class variable and also the relationship between that feature and other features. In wrapper methods, a learning algorithm is used to evaluate the usefulness of subsets of features.

III. Clustering Algorithm

In this research X-means clustering algorithm is employed to conduct the last step of proposed hybrid feature selection method. In this method, there is no need to determine the value of clusters. The algorithm itself can estimate the proper number of clusters based on optimization of Bayesian Information Criterion (Pelleg and Moore, 2002).

IV. Classification Algorithm

In this research the K-Nearest-Neighbour (KNN) classification algorithm and Decision tree (DT) have been used to estimate the credit score of the bank customers. Learning process in K-NN classification algorithm is based on similarity. K-NN compares a specific testing record with a set of training records that are similar to it. K-NN is also called ‘‘instance based learner’’ and ‘‘Lazy learner’’. DT algorithm is a widely used algorithm for classification, with a top-down tree structure (Tsai et al., 2014; Larose, 2014).

EXPERIMENTAL RESULTS

I. Data Description

The dataset used in this study contains credit information of 221 customers of a bank in Iran. The dataset consists of 51 variables, with 50 predictor variables and 1 target variable. All of the 221 records of this database are divided into 176 records labelled "good applicants" and 45 records labelled "bad applicants". The initial predictor variables used in the study are represented in Table 1.

II. Hybrid Feature Selection Approach

Before performing the feature selection stage the data are pre-screened. Missing values are handled through imputation method. The feature selection method proposed in this research is a combination of genetic algorithm method, filtering feature selection methods and clustering.
In Figure 2, the proposed multi-stage feature selection approach of this study is presented. As can be seen in Figure 2, the first stage is devoted to removal of correlated features. The second stage is allocated to select the important features with genetic algorithm. The third stage is proposed to weight the variables using different filtering methods. Finally, the fourth stage selects features through clustering algorithm. In the following each stage of the proposed hybrid method is described.

- **First Stage:** *Remove Correlated Features*

As mentioned, there are 50 initial features for predicting the credit of the bank customers. Regarding Figure 2, in the first stage, the high-correlated variables are removed. High-correlated variables can add no meaningful information to our analysis. Table 2 shows the remaining features after removing the high-correlated features. As it can be seen in Table 2, 36 features have remained after removing the correlated features.

- **Second Stage**: *Feature Selection By Genetic Algorithm*

The second stage is related to select the features using genetic algorithm. Table 3 shows the selected variables using the genetic algorithm. As can be seen in this Table, 19 variables were selected from among 36 variables.

- **Third *Stage***: *Filtering feature selection*

In the third stage, the 19 variables which were selected by genetic algorithm are weighted using six feature selecting methods. Table 4 shows the weights assigned to the variables using the six filtering methods. The weighting methods are Gini Index, Information Gain, Information Gain Ratio, Correlation, Rule, and Relief. Figure 3 Shows the weights assigned to the 19 features using filtering methods.

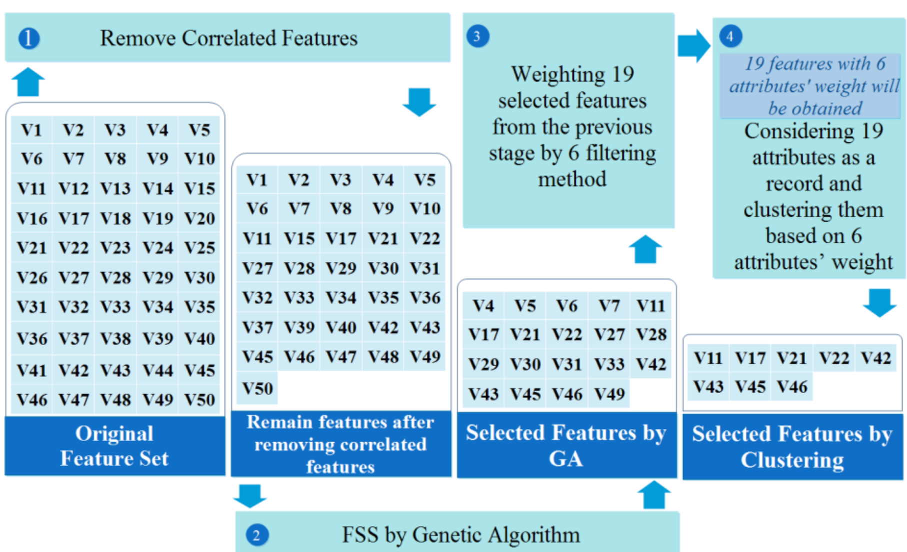

FIGURE 2
HYBRID FEATURE SUBSET SELECTION APPROACH

TABLE I
INITIAL PREDICTOR VARIABLES

Row	Variable	Row	Variable	Row	Variable	Row	Variable/ Type	Row	Variable
1	**V1:** industry and mine	**11**	**V11:** Stock	**21**	**V21:**Accumulated gains or losses	**31**	**V31:**Mangers history	**41**	**V41:**Current period assets
2	**V2:** agricultural	**12**	**V12:** Current assets	**22**	**V22:** shareholder Equity	**32**	**V32:** Type of company: Cooperative (=1, other =0)	**42**	**V42:** Prior period assets
3	**V3:** oil and chemical	**13**	**V13:** Non-current assets	**23**	**V23:** Sale	**33**	**V33:** Type of company: Stock Exchange(LLP) (=1, other =0)	**43**	**V43:** Two-Prior period assets
4	**V4:** infrastructure and service	**14**	**V14:**Total assets	**24**	**V24:**Gross profit	**34**	**V34:**Type of company: PJS (=1, other =0)	**44**	**V44:** Current period shareholder Equity
5	**V5**: Tax declaration	**15**	**V15:** Short-term financial liabilities	**25**	**V25:** Financial costs	**35**	**V35:** Type of company: Limited and others (=1, other =0)	**45**	**V45:** Prior period shareholder Equity
6	**V6:** Audit Organization	**16**	**V16:** Current liabilities	**26**	**V26:** Net profit	**36**	**V36:**Type of company: Stock Exchange (=1, other =0)	**46**	**V46:** Two-Prior period shareholder Equity
7	**V7:** Accredited auditor	**17**	**V17:** Long-term financial liabilities	**27**	**V27:** Active in internal market	**37**	**V37:** Experience with Bank(number of years in 5 categories)	**47**	**V47:** Current accounts creditor turn over
8	**V8:** Inventory cash	**18**	**V18:** Non-current liabilities	**28**	**V28:** number of countries that the company export to	**38**	**V38:** Current period sales	**48**	**V48:**Weighted Average Current Account
9	**V9:** Accounts receivable	**19**	**V19:** Total liabilities	**29**	**V29:** Target market risk (from 1 to 5)	**39**	**V39:** Prior period sales	**49**	**V49:** Average exports over the past three years
10	**V10:** Other Accounts receivable	**20**	**V20:** Capital	**30**	**V30:** Company history(number of years)	**40**	**V40:**Two-Prior period sales	**50**	**V50:** Last three years average imports

TABLE 2
REMAIN VARIABLES AFTER REMOVING HIGH-CORRELATED FEATURES

Row	Variable	Row	Variable	Row	Variable	Row	Variable
1	**V1**industry and mine	**10**	**V10**Other Accounts receivable	**19**	**V30:**Companyhistory(number of years)	**28**	**V40**Two-Priorperiod sales
2	**V2**agricultural	**11**	**V11**Stock	**20**	**V31:**Mangers history	**29**	**V42**Prior period assets
3	**V3**oil and chemical	**12**	**V15**Short-term financial liabilities	**21**	**V32**Type of company: (Cooperative =1, other =0)	**30**	**V43**Two-Prior period assets
4	**V4**infrastructure and service	**13**	**V17**Long-term financial liabilities	**22**	**V33**Type of company: Stock Exchange (LLP =1 ,other =0)	**31**	**V45** Prior period shareholder Equity
5	**V5**Tax declaration	**14**	**V21**Accumulated gains or losses	**23**	**V34**Type of company : (PJS=1, other =0)	**32**	**V46** Two-Prior period shareholder Equity
6	**V6**Audit Organization	**15**	**V22**shareholder Equity	**24**	**V35**Type of company : (Limited and others=1 ,other =0)	**33**	**V47**Current accounts creditor turn over
7	**V7**Accredited auditor	**16**	**V27:**Active in internal market	**25**	**V36**Type of company: (Stock Exchange =1, other =0)	**34**	**V48**Weighted Average Current Account
8	**V8**Inventory cash	**17**	**V28:**number of countries that the company export to	**26**	**V37**Experience with Bank (number of years in 5 categories)	**35**	**V49**Average exports over the past three years
9	**V9**Accounts receivable	**18**	**V29**Target market (risk (from 1 to 5	**27**	V39: Prior period sales	**36**	**V50**Last three years average imports

TABLE 3
SELECTED VARIABLES BY GENETIC ALGORITHM

Row	Variable	Row	Variable
1	V4: infrastructure and service	11	V29: Target market risk (from 1 to 5)
2	V5:Tax declaration	12	V30: Company history(number of years)
3	V6:Audit Organization	13	V31: Mangers history
4	V7:Accredited auditor	14	V33: Type of company: Stock Exchange(LLP) (=1, other =0)
5	V11:Stock	15	V42:Prior period assets
6	V17: Long-term financial liabilities	16	V43Two-Prior period assets
7	V21:Accumulated gains or losses	17	V45: Prior period shareholder Equity
8	V22: shareholder Equity	18	V46:Two-Prior period shareholder Equity
9	V27: Active in internal market	19	V49: Average exports over the past three years
10	V28: number of countries that the company export to		

TABLE 4
WEIGHTING FEATURES BY FILTERING METHODS

Variable	Gini Index	Information Gain	Information Gain Ratio	Correlation	Rule	Relief
V6: Audit Organization	0.000	0.000	0.000	0.153	0.000	0.504
V4: infrastructure and service	0.016	0.017	0.005	0.275	0.000	0.251
V33: Type of company: Stock Exchange(LLP) (=1, other =0)	0.118	0.140	0.028	0.639	0.000	0.143
V29: Target market risk (from 1 to 5)	0.152	0.171	0.011	0.023	0.000	0.155
V7: Accredited auditor	0.180	0.219	0.046	0.783	0.000	0.006
V30: Company history(number of years)	0.187	0.237	0.026	0.158	0.000	0.149
V31: Mangers history	0.201	0.193	0.026	0.486	0.000	0.230
V5: Tax declaration	0.249	0.236	0.093	0.916	0.000	1.000
V27: Active in internal market	0.279	0.352	0.074	0.969	0.000	0.072
V28: number of countries that the company export to	0.375	0.365	0.097	0.000	0.000	0.096
V21 : Accumulated gains or losses	0.454	0.496	1.000	0.416	0.933	0.006
V49: Average exports over the past three years	0.496	0.483	0.384	0.243	0.000	0.000
V46 : Two-Prior period shareholder Equity	0.648	0.659	1.000	0.971	0.422	0.002
V43 : Two-Prior period assets	0.654	0.649	1.000	0.751	0.533	0.002
V45 : Prior period shareholder Equity	0.746	0.791	1.000	0.737	1.000	0.008
V42 : Prior period assets	0.751	0.786	1.000	0.698	1.000	0.008
V11: Stock	0.804	0.794	1.000	0.680	0.956	0.014
V22 :shareholder Equity	0.904	0.942	1.000	0.767	1.000	0.008
V17 : Long-term financial liabilities	1.000	1.000	1.000	1.000	0.533	0.003

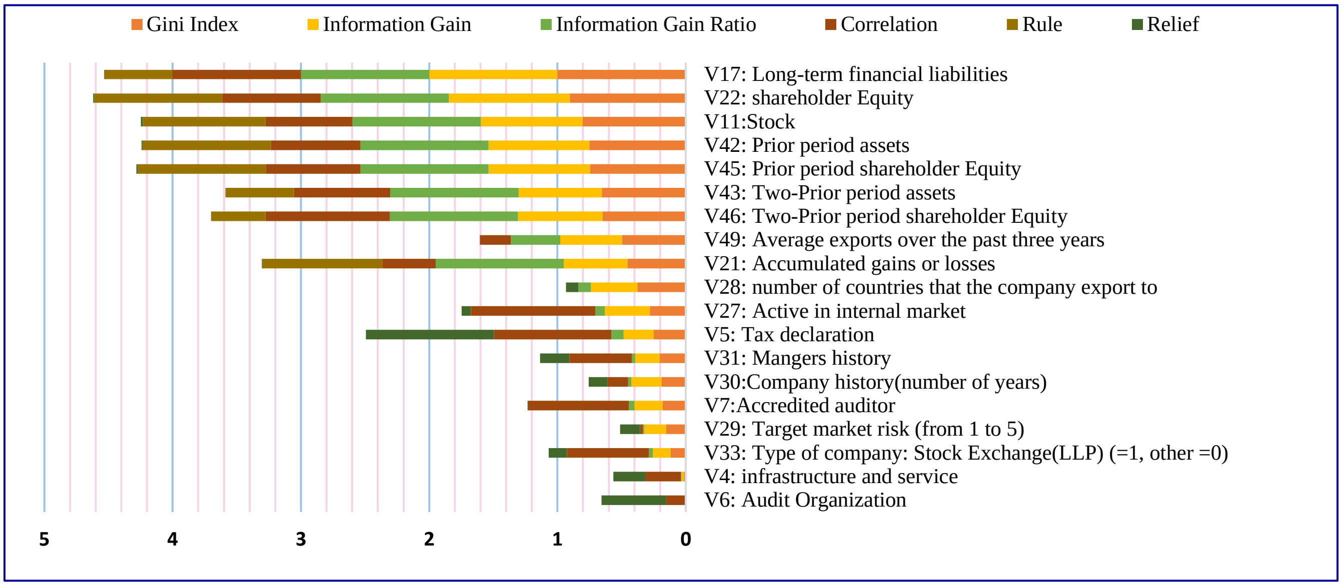

FIGURE 3
WEIGHTS OF FEATURES BASED ON FILTERING METHODS

As it can be seen in Table 4, the variable "shareholder Equity" has the largest sum of the weights (4.621). Then variable "Long-term financial liabilities" has the largest sum of the weights (4.536). Variable "Target Market Risk" has the smallest sum of the weights (0.521).

- **Forth Stage**: *Selecting the Final Subset of Features by Clustering*

At the end of third stage, 19 features with 6 weights were created. These 19 features are clustered X-means algorithm based on these 6 weights. The 19 aforementioned features are grouped into two clusters.

Accordingly, the variables "Accumulated gains or losses", "Two-Prior period shareholder Equity"," Two-Prior period assets", "Prior period shareholder Equity", "Prior period asset", "Stock", "shareholder Equity" and "Long-term financial liabilities" are placed in cluster 1 and other variables are placed in the second cluster. Figure 4 shows the grouping of these 19 features according to 6 weights.

The clusters are analyzed based on the average of the filtering method weights assigned to each variable. This is conducted by calculating the average weights of filtering methods in each cluster for all variables in the particular cluster. The cluster with the largest weighted mean sum is selected and the associated variables are considered as the final selected variables. The results of weighted analysis of the clusters are shown in Table 5.

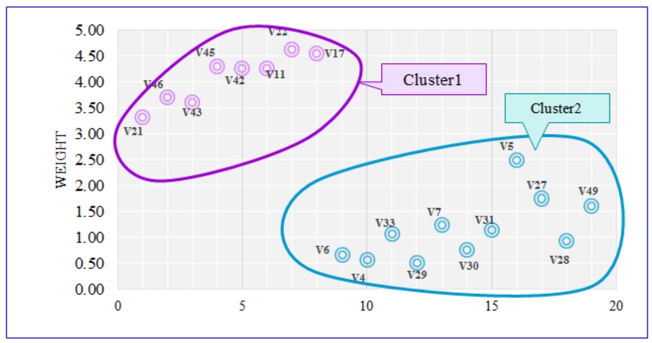

FIGURE 4
GROUPING 19 FEATURES BASED ON 6 WEIGHTS USING X-MEANS ALGORITHM

As can be seen in Table 5, the sum of the means of attributes' weights of the first cluster is equal to 32.53 and it is equal to 12.71 for the variables of the second cluster. Although the number of the variables in the first cluster (8 variables) is less than the number of the variables in the second cluster (11 variables), but the sum of weighted mean for the variables of the first cluster is larger than the second cluster. Therefore, the variables of the first cluster are more significant, and they are considered as the set of finally selected features.

TABLE 5
THE RESULTS OF CLUSTERING BY X-MEANS ALGORITHM

Cluster	Variables	Total Weight
cluster_1	**V21 :**Accumulated gains or losses	3.31
cluster_1	**V46 :**Two-Prior period shareholder Equity	3.70
cluster_1	**V43 :**Two-Prior period assets	3.59
cluster_1	**V45 :**Prior period shareholder Equity	4.28
cluster_1	**V42 :**Prior period assets	4.24
cluster_1	**V11:**Stock	4.25
cluster_1	**V22 :**shareholder Equity	4.62
cluster_1	**V17 :**Long-term financial liabilities	4.54
	Sum(Total)	**32.53**
cluster_2	**V6:** Audit Organization	0.66
cluster_2	**V4:** infrastructure and service	0.56
cluster_2	**V33:** Type of company: Stock Exchange(LLP) (=1, other =0)	1.07
cluster_2	**V29:** Target market risk (from 1 to 5)	0.51
cluster_2	**V7:**Accredited auditor	1.23
cluster_2	**V30:**Company history(number of years)	0.76
cluster_2	**V31:** Mangers history	1.14
cluster_2	**V5:** Tax declaration	2.49
cluster_2	**V27:** Active in internal market	1.75
cluster_2	**V28:** number of countries that the company export to	0.93
cluster_2	**V49:** Average exports over the past three years	1.61
	Sum(Total)	**12.71**

III. Classification

Eight variables of the first cluster are entered into classification algorithms in order to predict the credit risk of the customers. Credit prediction is conducted using K-Nearest Neighborhood (K-NN) and Decision Tree (DT) algorithms. Using KNN and DT and based on the selected features of a customer (as mentioned in Table 5), the estimation of credit the applicants of the loan can be determined.

In order to evaluate the validation of classification models, ten-fold validation method is used. Accuracy, precision and recall are selected to evaluate classification performance.

Evaluation metrics such as accuracy, precision, and recall are used to evaluate classification algorithms. These metrics can be explained with respect to a confusion matrix as shown in Table 6. (Kittidecha and Yamada 2018)

TABLE 6
CONFUSION MATRIX

		Prediction	
		Positive	Negative
Actual	Positive	True Negative (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Accuracy, precision, and recall are calculated respectively using Eq. (1), (2) and (3).

$$Accuracy = \frac{TN+TP}{TP+TN+FP+FN} \quad (1)$$

$$\Pr ecision = \frac{TP}{TP+FP} \quad (2)$$

$$\operatorname{Re} call = \frac{TP}{TP+FN} \quad (3)$$

The results of classification by KNN and DT are represented in Table 7.

TABLE 7
THE RESULTS OF CLASSIFICATION BY KNN AND DT

Metrics/Learning algorithm	K-Nearest Neighbor	Decision Tree
Accuracy	86.41% +/- 6.86%	79.74% +/- 8.18%
Precision	78.72% +/- 16.04%	67.74%
Recall	74.50% +/- 17.24%	45.37% +/- 31.41%

As can be seen in Table 7, accuracy, precision, and recall values for K-NN algorithm are 86.41%, 78.72%, and 74.50%, respectively. K-NN algorithm perform better than DT method to predict credit risk of applicants.

CONCLUSION

In recent years, overdue loans made it necessary for banks to use credit scoring estimation systems. Credit scoring helps financial institutions to improve their profit and reduce possible risks. As, the databases of banks contain a large amount of customer information, they can be used to assess the credit risk. Using data mining techniques and analyzing customers' data could be an effective tool for credit scoring. Banks' databases usually contain several features of the customers. In these situations, it's important to identify an optimum feature subset and selected only important features. Some of the features may be redundant or irrelevant. In this regard, this research proposed a multi-stage feature selection approach. The proposed approach was applied to a customer's dataset of an Iranian bank. Optimum subset of the features was selected using combining genetic algorithm, filtering methods and well-known clustering methods. Afterwards, the selected features were entered into classification algorithms in order to predict the credit risk of the bank customers.

The proposed approach of this study can be customized and used in other service companies such as insurance, hospitals and supermarkets in order to achieve compact subset of customers’ features before conducting classification stage.

REFERENCES

[1] Abdi, F., Khalili-Damghani, K., Abolmakarem, S. (2017) "Solving customer insurance coverage sales plan problem using a multi-stage data mining approach", Kybernetes, 47(1) . https://doi.org/10.1108/K-07-2017-0244

[2] Apornak A., Raissi S., Keramati A., Khalili-Damghani K., (2020), optimizing human resource cost of an emergency hospital using multi-objective Bat algorithm, International Journal of Healthcare Management, 1-7

[3] Arora, N., Kaur, P.D. (2020) "A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment", Applied Soft Computing, 86, 105936, https://doi.org/10.1016/j.asoc.2019.105936

[4] Bijak K, Thomas L.C. (2012). "Does segmentation always improve model performance in credit scoring?" *Expert Systems with Applications* 39, 2433–2442

[5] Danenas P., Garsva G. (2015). "Selection of Support Vector Machines based classifiers for credit risk domain", *Expert Systems with Applications*, 42, 3194–3204

[6] Florez-Lopez R., Ramon-Jeronimo J.M., (2015) "Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal", *Expert Systems with Applications*, 42, 5737–5753

[7] Guyon S, Elisseeff A, (2003) "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research*, 3, 1157-1182

[8] Hajek P., Michalak K., (2013). "Feature selection in corporate credit rating prediction", *Knowledge-Based Systems*, 51, 72–84

[9] Harris T. (2015). "Credit scoring using the clustered support vector machine", *Expert Systems with Applications*, 42, 741–750

[10] Henley, W. E. (1995). "Statistical aspects of credit scoring. Dissertation", The Open University, Milton Keynes, UK.

[11] Hens A.B., Tiwari M.K. (2012) "Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method", *Expert Systems with Applications*, 39, 6774–6781

[12] Hsieh N-C, Hung L-P. (2010) "A data driven ensemble classifier for credit scoring analysis", *Expert Systems with Applications*, 37, 534–545

[13] Khalili-Damghani, K., Abdi, F., Abolmakarem, S. (2018) " Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries", *Applied Soft Computing*, 73, 816-828

[14] Khalili-Damghani, K., Abdi, F., Abolmakarem, S. (2018) "Solving customer insurance coverage recommendation problem using a two-stage clustering-classification model", *International Journal of Management Science and Engineering* Management, 14(1)9-19

[15] Khashei M,Rezvan M.T., A ZeinalHamadani, AND MBijari. (2013). "A Bi-Level Neural-Based Fuzzy Classification Approach for Credit Scoring Problems", *Complexity*, 18 (6), 46-57.

[16] Khashman A. (2010) "Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes", *Expert Systems with Applications*, 37, 6233–6239

[17] Khashman A. (2011) "Credit risk evaluation using neural networks: Emotional versus conventional models", *Applied Soft Computing* 11, 5477–5484

[18] Kittidecha C, Yamada K (2018) Application of Kansei engineering and data mining in the Thai ceramic manufacturing. Journal of Industrial Engineering International 14, 757–766

Int. https://doi.org/10.1007/s40092-018-0253-y

[19] Laha A. (2007). "Building contextual classifiers by integrating fuzzy rule based classification technique and K-NN method for credit scoring", *Advanced Engineering Informatics*, 21, 281–291

[20] Larose D. T., Larose C.D., (2014) "Discovering knowledge in data: an introduction to data mining", Second ed., John Wiley & Sons, Inc., Hoboken, New Jersey.

[21] Lessmann S, Baesens B, Seow H-V, and Thomas L.C., (2015), "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research", *European Journal of Operational Research*, 247(1),124-136

[22] Maldonado S, Perez J, Bravo C (2017) "Cost-based feature selection for Support Vector Machines –An application in credit scoring", *European Journal of Operational Research*, 261 (2) 656–665

[23] Marqués A.I, Garcia. V., Sanches J.S. (2012) "Two-level classifier ensembles for credit risk assessment", *Expert Systems with Applications*, 39, 10916–10922

[24] Moradkhani M, Amiri A, Javaherian M, Safari H, (2015) "A hybrid algorithm for feature subset selection in high-dimensional datasets using FICA and IWSSr algorithm", *Applied Soft Computing*, 35, 123-135

[25] Nalić, J., Martinović, G, Žagar, D. (2020) "New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers", *Advanced Engineering Informatics*, 45, 101130

[26] Nourian R. , Meysam Mousavi S., Raissi S., (2019) A fuzzy expert system for mitigation of risks and effective control of gas pressure reduction stations with a real application, Journal of Loss Prevention in the Process Industries,59, 77-90.

[27] Oreski S, Oreski D, Oreski G, (2012) "Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment", *Expert Systems with Applications*, 39, 12605–12617

[28] Oreski S, Oreski G. (2014). "Genetic algorithm-based heuristic for feature selection in credit risk assessment", *Expert Systems with Applications*, 41 (4) 2052-2064

[29] Papouskova, M., Hajek, P. (2019) "Two-stage consumer credit risk modelling using heterogeneous ensemble learning", *Decision Support Systems*, Vol.118, pp.33-45

[30] Pelleg D., Moore A. (2002). "X-means: Extending K-means with Efficient Estimation of the Number of Clusters", *Proceedings of the Seventeenth International Conference on Machine Learning*, PP. 727-734.

[31] Pławiak, P., Abdar, M., Acharya, UR. (2019), "Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring", *Applied Soft Computing*, Vol. 84, 105740, https://doi.org/10.1016/j.asoc.2019.105740

[32] Ping Y., Yongheng L. (2011). "Neighborhood rough set and SVM based hybrid credit scoring classifier", *Expert Systems with Applications*, 38, 11300–11304

[33] Rtayli, N. Enneya, N. (2020) "Selection Features and Support Vector Machine for Credit Card Risk Identification", Procedia Manufacturing, 45, 941-948.

[34] Shen, F., Zhao, X., Li. Z., Li. K., Meng. Z. (2019) "A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation", Physica A: Statistical Mechanics and its Applications, Vol. 256, 121073, https://doi.org/10.1016/j.physa.2019.121073

[35] Thomas L. C., Edelman D. B., Crook J. N. (2002). "Credit scoring and its applications". Philadelphia, PA: SIAM.

[36] Tsai C-F, Eberle W, Chu C-Y, (2013). "Genetic algorithms in feature and instance selection", *Knowledge-Based Systems*, 39, 240–247

[37] Tsai C-F., Hsu Y-F., Yen D.C., (2014) "A comparative study of classifier ensembles for bankruptcy prediction", *Applied Soft Computing*. 24, 977–98.

[38] Wang G, Ma J, Huang L, Xu K, (2012) "Two credit scoring models based on dual strategy ensemble trees", *Knowledge-Based Systems*, 26, 61–68

[39] Wang G, Ma J, (2012). "A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine", *Expert Systems with Applications*, 39, 5325–5331

[40] Wang D, Zhang Z, Bai R, Mao Y. (2018). "A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring", *Journal of Computational and Applied Mathematics*, 329, 307-321

[41] Wu, W.-W. (2011). "Improving classification accuracy and causal knowledge for better credit decisions". *International Journal of Neural Systems*, 21(04), 297–309

[42] Xiao J, Xie L, He C, Jiang X, (2012). "Dynamic classifier ensemble model for customer classification with imbalanced class distribution", *Expert Systems with Applications*, 39, 3668–3675

[43] Yap B. W., Ong S.H., Mohamed Husain N.H. (2011). "Using data mining to improve assessment of credit worthiness via credit scoring models", *Expert Systems with Applications*, 38, 13274–13283

[44] Yu L., Yao X., WangSh., LaiK.K.(2011). "Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection", *Expert Systems with Applications*, 38, 15392–15399

[45] Zhao Z, XuSh, KangB, KabirM.M.J., LiuY, and Wasinger R. (2015) "Investigation and improvement of multi-layer perceptron neural networks for credit scoring", *Expert Systems with Applications*, 42 (7) 3508–3516

[46] Zhu, H., Beling, P. A., and Overstreet, G. A. (2002). "A Bayesian framework for the combination of classifier outputs". *The Journal of the Operational Research Society*, 53(7), 719–727.