THE USE OF GENETIC ALGORITHM, CLUSTERING AND FEATURE SELECTION TECHNIQUES IN CONSTRUCTION OF DECISION TREE MODELS FOR CREDIT SCORING

Mohammad Khanbabaei and Mahmood Alborzi

Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran

ABSTRACT

Decision tree modelling, as one of data mining techniques, is used for credit scoring of bank customers. The main problem is the construction of decision trees that could classify customers optimally. This study presents a new hybrid mining approach in the design of an effective and appropriate credit scoring model. It is based on genetic algorithm for credit scoring of bank customers in order to offer credit facilities to each class of customers. Genetic algorithm can help banks in credit scoring of customers by selecting appropriate features and building optimum decision trees. The new proposed hybrid classification model is established based on a combination of clustering, feature selection, decision trees, and genetic algorithm techniques. We used clustering and feature selection techniques to pre-process the input samples to construct the decision trees in the credit scoring model. The proposed hybrid model choices and combines the best decision trees based on the optimality criteria. It constructs the final decision tree for credit scoring of customers. Using one credit dataset, results confirm that the classification accuracy of the proposed hybrid classification model is more than almost the entire classification models that have been compared in this paper. Furthermore, the number of leaves and the size of the constructed decision tree (i.e. complexity) are less, compared with other decision tree models. In this work, one financial dataset was chosen for experiments, including Bank Mellat credit dataset.

KEYWORDS

Credit scoring, Genetic Algorithm, Feature Selection, Decision Tree, Clustering, Hybrid Approaches for Credit Scoring

1.INTRODUCTION

Gary and Fan (2008) [1] believed that, "Banks as economic institutions need to recognize customers' credit risk to offer credit facilities and manage their risk". Recently, non-parametric methods and data mining have been used in the customers' credit scoring techniques. Decision trees, as one of the classification techniques in data mining, can help to perform customer credit scoring with high ability of understanding and learning speed to build classification models. The main problem in this study is the construction of decision trees to classify bank customers optimally. There are several weaknesses in construction of recursive partitioning trees: 1. Greediness in the tree growing process and local optimization at each step in the node splitting

process, and there are other problems including instability and bias in splitting the rule selection [1]. 2. Tendency to construct large trees, and over-fit to training datasets [1], and generalization problem [2]. To cope with the interaction among attributes, genetic algorithms (GAs) are strong, flexible and better than most of rule induction algorithms in a global search. The reason is that GA applies a population of candidate solutions (individuals), and it evaluates them as a whole using fitness function. However, according to greedy rule induction algorithms, there is a single candidate solution every time and the evaluation is performed in a special candidate solution (based on local optimization). In addition, using probabilistic operators, GA prevents solutions to be locked in local optimization [2].

In this regard, genetic algorithms can help to select appropriate features and build optimum decision trees in credit scoring of bank customers. In feature selection (FS) in this paper, GA includes an optimization process, in which many combinations of features and their interactions are considered. Because GA searches for solutions efficiently in high dimensional and difficult response surfaces, it can be utilized for feature selection in a variety of problems and multivariate calibration in particular [3].

The purpose of this study is to propose an appropriate new hybrid model for credit scoring of bank customers. It is used to offer credit facilities for various classes of customers. A genuine recognition of features of bank customers is necessary to reach this objective. Moreover, it is required to build decision trees through a genetic algorithm with the following characteristics in classification: small size, simplicity, and high accuracy. In this paper, the development process in pattern recognition and CRISP-DM were used for credit scoring.

With respect to a study done by Tsai and Chen, the research area of hybridization approaches, which is performed to improve classification performance, is more active than single learning approaches [4]. The current study presents a hybrid mining approach in the design of an effective and appropriate credit scoring model based on genetic algorithm for credit scoring of bank customers in order to offer credit facilities to each class.

2. LITERATURE REVIEW

In this section, we shall review the literature of credit scoring and the commonly used techniques in modelling credit scoring problems. The statistical methods, non-parametric methods, and artificial intelligence approaches have been proposed to support the credit scoring process. Some of these techniques are as follows:

Artificial neural networks (ANNs) [5], naive Bayes, logistic regression(LR), recursive partitioning, ANN and sequential minimal optimization (SMO) [6], neural networks (Multilayer feed-forward networks) [7], ANN with standard feed-forward network [8], credit scoring model based on data envelopment analysis (DEA) [9], back propagation ANN [10], link analysis ranking with support vector machine (SVM) [11], SVM [12], integrating non-linear graph-based dimensionality reduction schemes via SVMs [13], Predictive modelling through clustering launched classification and SVMs [14], optimization of k-nearest neighbor (KNN) by GA [15], Evolutionary-based feature selection approaches [16], comparisons between data mining techniques (KNN, LR, discriminant analysis, naive Bayes, ANN and decision trees) [17], SVM [18], intelligent-agent-based fuzzy group decision making model [19], SVMs with direct search for parameters selection [20], SVM [21], decision support system (DSS) using fuzzy TOPSIS

[22], neighbourhood rough set and SVM based classifier [23], Bayesian latent variable model with classification regression tree [24], integrating SVM and sampling method in order to computational time reduction for credit scoring [25], use of preference theory functions in case based reasoning model for credit scoring [26], fuzzy probabilistic rough set model [27], using rough set and scatter search met heuristic in feature selection for credit scoring [28], neural networks for credit scoring models in microfinance industry [29].

Furthermore, there are some techniques related to ensemble credit scoring models. Neural network ensemble strategies [30], multilayer perceptron (MLP), neural network ensembles [31], ensemble of classifiers (Bagging, Random Subspace, Class Switching, Random Forest) [32], ensemble of classifiers (ANN, decision tree, naive Bayes, KNN and logistic discriminant analysis) [33], bagging ensemble classifier (ANN, SVM and Bayesian network) [34], Subagging ensemble classifier (kernel SVM, KNN, decision trees, adaboost and subagged classifiers) [35], SVM-based multi agent ensemble learning [36], Least squares support vector machines (LSSVMs) ensemble models [37].

Recently, there have been hybrid models in many credit scoring researches and there is a significant tendency to use hybrid intelligent systems for credit scoring problems. Nevertheless, there are few researches in the development of hybrid models for credit scoring [4]. In the following, some related works about hybrid-learning models are provided. Tsai and Chen expressed that the developed hybrid models are usually compared with those models that are based on a single machine leaning technique. Further, while hybrid models have better performance comparing single classification models, a question emerges that what kind of hybrid models can be the best choice in credit scoring problems [4].

Therefore, this paper compares the new proposed hybrid classification model with different types of credit rating models (entirely based on decision trees). The comparison is provided in terms classification accuracy, number of leaves, and size of the decision tree (complexity).

Some of hybrid models of learning that consider the application of hybrid techniques are as follows: Hybrid neural discriminant technique [38], hybrid model by probit and Classification and Regression Tree (CART) techniques [39], two-stage hybrid model using artificial neural networks and multivariate adaptive regression splines (MARS) [40], hybrid support vector machine technique [41], hybrid reassigning credit scoring model with MARS, ANN and case-based reasoning (CBR) [42], new two-stage hybrid approach by LR and back propagation network (BPN) [43], hybrid model via combining the classification and clustering techniques [4], neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model [44].

Many researchers have studied the application of genetic algorithms in feature selection and building decision trees. These studies mainly relate to other sciences and businesses that are listed as follows. We can use the results of them in research and development of credit scoring studies. In this regard, there are some studies about GA application in feature selection. Combining multiple classifier based on genetic algorithm [45], using GA in input variable selection [46], applying GA in variable selection with customer clustering [47] and use of GA to combine feature selection methods [48]. Furthermore, there are some studies about using GA to build decision trees that are provided accordingly. Classifier hierarchy learning by means of GAs [49], optimizing prediction models by GA (based on decision trees and neural networks) [50],

optimization of decision tree classifier through GA [51], fitness of binary decision tree by GA [52], classification tree analysis using TARGET [1], and utilization of the elitist multi-objective genetic algorithm for classification rule generation [53].

Therefore, the aim of this paper is to examine the performance of the new proposed hybrid classification model for credit scoring. It is enriched by studying the single classification and clustering techniques, as baseline models. The contribution of this paper is to figure out how the new proposed hybrid classification model can classify customers in credit scoring studies.

There are some advantages for the proposed hybrid classification model, different from the other models mentioned in the literature review section. 1. Application of data preparation and pre-processing methods in construction of the new proposed hybrid model. 2. Using clustering, as one of the data pre-processing methods, to increase the accuracy and decrease the complexity of customer classification. 3. Combining several feature selection algorithms based on the Filter, Wrapper and Embedded approaches to increase flexibility and classification accuracy in order to build decision trees (instead of using one single classifier). 4. Constructing and comparing variety of decision trees in the new proposed hybrid classification model. 5. Optimizing decision trees by GA in the new proposed credit scoring model in a hybrid context. Most of studies did not apply optimization techniques to improve the performance of their models in credit scoring. 6. Using complexity score along with the classification accuracy score in order to enhance the selection process of the best decision tree. 7. Using artificial intelligence, pattern recognition, and data mining approaches for credit scoring in complex conditions and non-linear relations in customer classification and feature selection. 8. Utilizing the development process in pattern recognition [54] and CRISP-DM process in construction of the final decision tree through the new proposed hybrid classification model for customers' credit scoring.

This paper is organized as follows. In section 3, it is briefly described about the methods used throughout the paper. In section 4, the research methodology is presented, including the development of the new proposed hybrid credit scoring model, the considered evaluation strategies, etc. The experimental results are elaborated in section 5. Finally, section 6 is devoted to discussions and conclusion of the article.

3. METHODS

3.1. The dataset

The description of the credit scoring features in the dataset collected from Bank Mellat of Iran is provided in Box 1. It includes 5173 cases of individual consumers' credit data for the first three months of 2003. The output (target) feature is 'type of record', which consists of three nominal class labels. Therefore, the credit ranking can be regarded as a three-class classification problem. The model is to classify new test cases into each of the three-class labels.

Box 1. The credit scoring features in the dataset collected from Bank

Date of contract (nominal), supervision code (nominal), branch number (nominal), request number (nominal), applicant category (nominal), local code (nominal), customer name (nominal), ID number (nominal), date of birth (nominal), code of the ID issue place (nominal), zone code (nominal), the corresponding code for category of the request (nominal), first period of payment (nominal), last period of payment (nominal), number of installments (numeric), bank quota (numeric), the amount of contract (numeric), the used amount of contract (numeric), allowable methods to use grants (nominal), the corresponding code for category of the collateral (nominal), the collateral price (numeric), last amount of debt (numeric), sector code (nominal), field of activity (nominal), category of contract (nominal), purpose of receiving grants (nominal), the place to spend grants (nominal), category of grants application (nominal), category of government-imposed grants (nominal), the record date (nominal), date (month and year) (nominal), name and last name (nominal), credit code (nominal), code of the record category (nominal), the account rubric code (nominal), last transaction date (nominal).

There are three categories of customers in the target feature. 1. Customer (1): They have paid back all of their credit facilities. 2. Customer (2): Three months have passed from the maturity date of their credit facilities. 3. Customer (3): They have non-performing credit facilities of more than six months [55].

In this paper, data preparation methods, those are considered for the Bank Mellat credit dataset in the new proposed hybrid classification model, are as follows: 1. removing attributes (features), those with unique values. 2. Elimination of some attributes, those indicate time trend for data input action. 3. Removing some of the attributes containing fixed or missing values. 4. Deleting instances (transactions) with missing values if we are unable to add value to them. 5. Omission of outliers and noises. 6. Assessment of the consistency in a unit of attribute measurement. 7. Converting textual values to numeric ones. 8. Normalization of attributes. 9. Discretisation of numeric and nominal values. 10. Merging values in nominal attributes. 11. Conversion of numeric to nominal values. 12. Converting dates to numeric values.

3.2. Credit scoring

Thomas defined that, "credit scoring is a technique that helps some organizations, such as commercial banks and credit card companies, determine whether or not to grant credit to consumers, on the basis of a set of predefined criteria" [19]. Some of the benefits of using credit scoring models are listed in [56] a study, which include the followings: cost reduction in credit analysis, quicker decision making regarding credit allocation, higher probability to collect credits, and lower amount of possible risks.

The history of credit scoring refers to the idea of statistical discrimination analysis, introduced by Fisher in 1936. In 1941 for the first time, David Durand used some techniques to classify good and bad loans. In 1960s, credit cards were appeared in banks and credit scoring became useful for banks. In 1980s, understanding the usefulness of credit scoring in credit cards, banks applied credit scoring for other products as well [57].

At first, credit scoring was performed based on judgmental view of credit analysts. After reviewing an application form, they said yes or no as a final decision on credit allocation. Their

decisions were based on the 3Cs, 4Cs, or 5Cs. These items are the character, capital, collateral, capacity, and conditions of the customer [57].

With respect to a study performed by [19], research in the field of credit scoring is increasing due to the significance of the credit risk evaluation. They included many statistical and optimization methods, such as linear discriminant analysis, logistic analysis, probit analysis, linear programming, integer programming, KNN, and classification tree.

Moreover, one of the problems in statistical and optimization methods is the ability to distinguish good customers from bad ones. Recently, studies in credit scoring are operated based on artificial intelligence (AI) techniques, such as artificial neural networks, evolutionary computation (EC), genetic algorithm, and support vector machine [19].

3.3. Hybrid models in credit scoring

One of the research issues to improve the classification performance is to apply hybrid-learning approaches instead of single ones. In other words, clustering methods composed with classification models can be used to pre-process the training dataset. In addition, classification models composed with clustering models can be applied as well [4].

Tsai and Chen found that, "to develop a hybrid-learning credit model, there are four different ways to combine two machine learning techniques. They are: 1. combining two classification techniques. 2. Combining two clustering techniques. 3. One clustering technique combined with one classification technique. 4. One classification technique combined with one clustering technique" [4]. As shown in Figure 2, two hybrid-learning approaches were applied in the model used in this paper. First, a clustering technique is combined with a classification technique. In this approach, SimpleKmeans clustering technique was firstly used to cluster the dataset. Then, several decision trees were used in each cluster to classify customers' credit scoring. Secondly, a combination of two classification techniques was utilized. This approach was employed in one part of the model. The second approach was the construction of decision tree with two branches using the GATree system [58] and then, construction of C4.5 in each branch (the classifier hierarchies approach, shown in Figure 2).

3.4. Classification and decision tree

Classification trees (decision trees) are one of the data mining techniques that predict the value (called class) of the dependent variable (target variable) using values of independent variables. Variables are also known as attributes (features). Values of target attribute are discrete. However, in independent attributes, they are either of discrete or continuous. Decision trees begin with the entire training dataset and they use a top-down induction method. Then, they apply the recursive partitioning approach to create branches in most informative attributes. It operates by splitting a particular subset based on values of the specified attribute. In the end, the final subset (known as leaves) is created using recursive partitioning method and they will receive their values (classes) consequently [46].

J48 (Java version of C4.5) was employed in Weka machine learning package for construction of decision trees. Furthermore, training and testing the decision tree models were performed based on 10-fold cross-validation. The evaluation of decision tree models was done by the correctly

classified instances (CCI) [46]. The inductive learning algorithm of C4.5 decision tree can be accessed in a work done by Larose [59].

3.5. Clustering

Clustering, as one of data mining techniques, classifies the records with similar objects. Records in one cluster are similar; however, they are dissimilar to records in other clusters. In clustering, there is no target variable and it is different from classification. Larose addressed that, "the clustering task does not try to classify, estimate, or predict the value of a target variable. Instead, clustering algorithms seek to segment the entire dataset into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized and the similarity to records outside the cluster is minimized" [59].

SimpleKmeans clustering algorithm was applied, which is part of Weka machine learning package. It has been discussed by Olson and Shi. It has been recommended by them that clustering, as a pre-processing stage in the dataset, can be used [60].

3.6 Feature (attribute) selection

Feature selection algorithms select appropriate features, usually as pre-processing stage of the model development. They are selected to increase the performance of the classification model, which is resulted from the training data. Some of the benefits of feature selection are mentioned as follows: 1. Noise reduction. 2. Achievement of an appropriate model through reduction in computational efforts. 3. Simplification of the final models obtained from the classification algorithms. 4. Uncomplicated application and updating of the model [61]. Then, using feature selection algorithms in construction of the new proposed hybrid classification model, we can reach better results in customer credit scoring.

In feature selection, there are three essential issues: the evaluation criterion, search method, and stopping rule. Often, there are five types of evaluation criteria: information, dependence, distance, consistency, and classification accuracy. The first four and the last one are related to the filter and wrapper approach in feature selection, respectively. FS in the filter approach is independent of classification algorithm, and selection of the features is operated based on the inherent quality of the data. In contrast, FS in the wrapper approach depends on classification algorithm to evaluate feature subsets. There are mostly three types of search methods in FS: complete search, heuristic search, and random search. The first two are used for a smaller search space that requires higher efficiency. The third one is applied for a larger search space [62]. Moreover, the stopping rules are provided: 'A predefined maximum iteration number has been attained', 'No better result can be obtained by adding or removing a feature', and 'The optimal feature subset has been found' [62]. One of the other feature selection methods is based on the embedded approaches. In these methods, FS is a part of classification method [61].

In this paper, we use the filter, wrapper, and embedded approaches for feature selection algorithms. The search method is random and based on genetic algorithm. The stopping rule is 'a predefined maximum iteration number has been attained'.

We use five feature selection methods as below. The first three are based on the filter approach. Fourth and fifth ones are according to the wrapper and embedded approaches, respectively. The first four are available in Weka version 3.5.8 machine learning tool

1. Feature selection based on correlation of Features Subsets with the Class and Intercorrelation between Features: As quoted in Weka version 3.5.8, "Evaluator function in this method evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class, while having low intercorrelation are preferred".

2. Feature selection based on consistency of the worth of a Subset of Attributes with the class values: As quoted in Weka version 3.5.8, "Evaluator function in this method evaluates the worth of a subset of attributes by the level of consistency in the class values when the training instances are projected onto the subset of attributes." Evaluator function in this method uses exhaustive search and in any subset, consistency must be higher than that of the full set of attributes.

3. Feature selection based on C4.5 classifier subset: In this method, attribute subsets are evaluated according to the training data or a separate holdout test set. A classifier (C4.5 in this case) estimates the 'advantage' of a set of attributes.

4. Feature selection based on the wrapper subset using C4.5 classifier: In this method, attribute sets are evaluated using a learning algorithm (C4.5 was used in this case).

5. Feature selection based on GATree system [58]: This method constructs a decision tree using GATree system. We can use its nodes as the final selected features for construction of the next decision trees.

3.7. Genetic algorithm

Genetic algorithms are general search algorithms based on Charles Darwin's principle of 'survival of the fittest'. They are utilized to respond to the complex optimization surfaces. These algorithms are applied in a population of chromosomes to generate candidate solutions in problem solving [46].

4. METHODOLOGY

4.1. Development of the new proposed hybrid model

The process of constructing the decision tree in the new proposed hybrid classification model for credit scoring of bank customers is represented in figure 2. Clustering can be used for data pre-processing [60]. According to figure 2, flowchart of total stages in construction of the new proposed hybrid classification model is shown in figure 1.

Firstly, 5668 transactions of real customers were collected. These customers had received government-imposed credit facilities in a contract format from Bank Mellat in the first three months of 2003. In the dataset, customers were divided into three classes (stated in section 3.1).

After data preparation, two clusters were generated from training and test dataset by SimpleKmeans clustering method. Feature selection based on genetic algorithm was accomplished by Filter, Wrapper and Embedded approaches for each cluster. Subsequently, C4.5 decision trees and the decision tree constructed using GATree system in each cluster were built with a set of selected features. The best decision tree in each cluster was selected by optimality criteria, such as number of leaves, size of the tree, and percentage of the correctly classified instances. Finally, two decision trees in each cluster were combined and the final decision tree was constructed for credit scoring of bank customers.

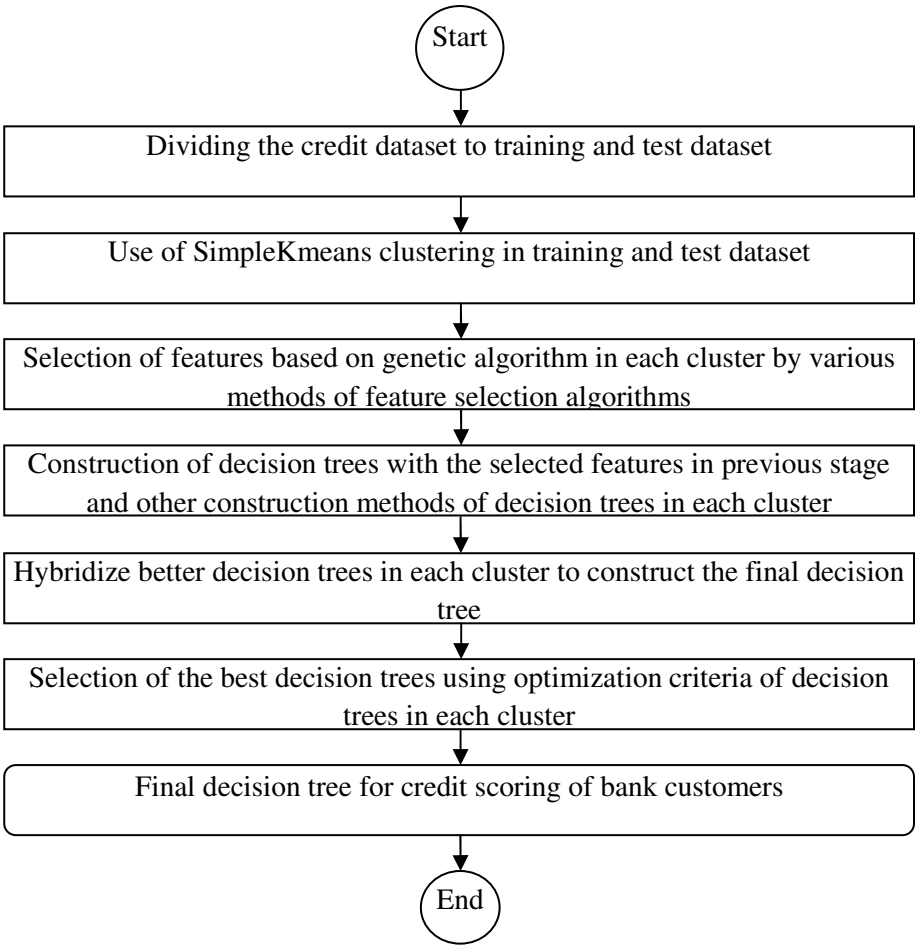

Figure 1. Flowchart of total stages to construct the new proposed hybrid classification model

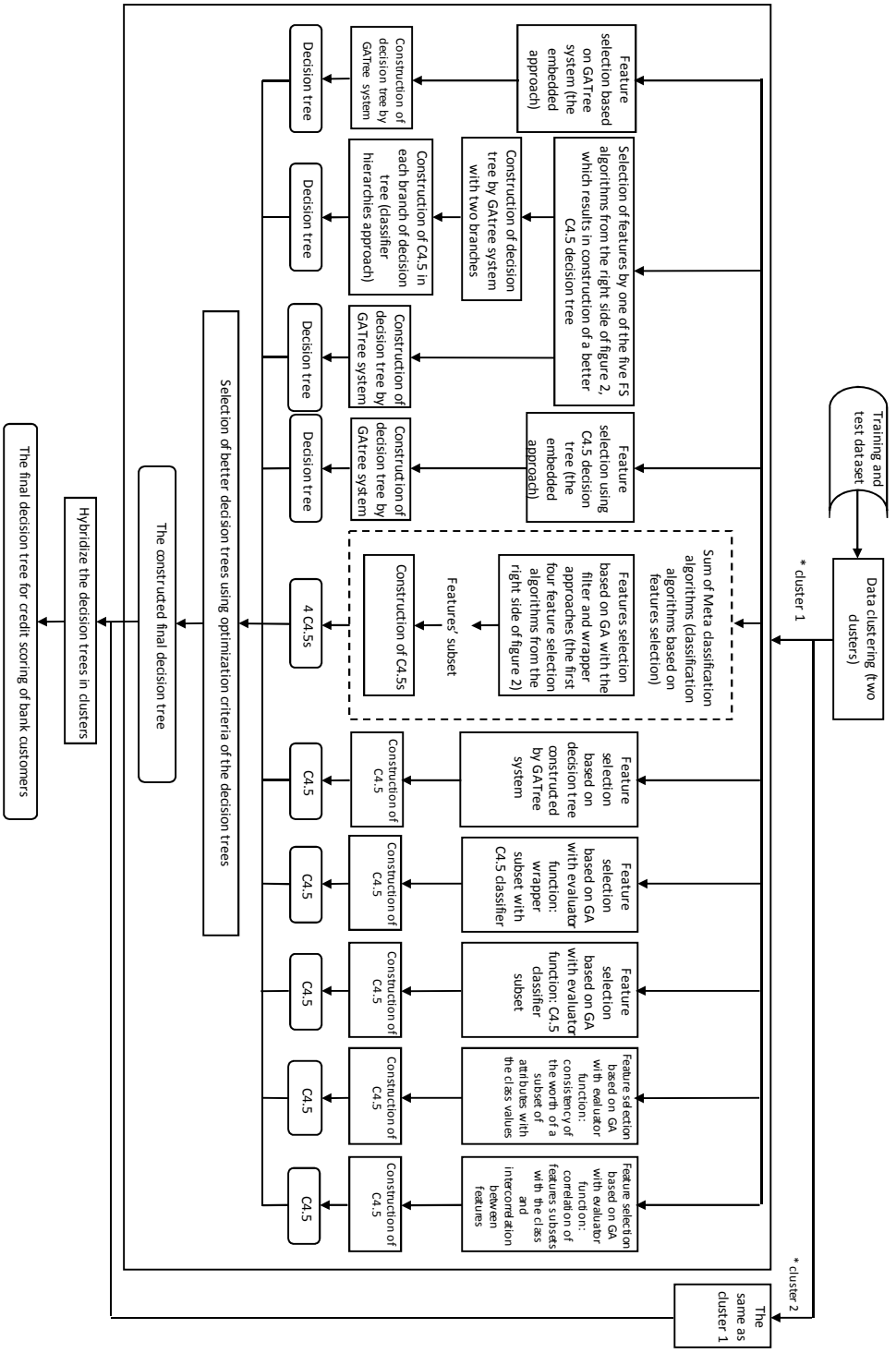

Figure 2. The construction process of decision tree, in the new proposed hybrid classification model for credit scoring of bank customers

In GA, a chromosome is constructed by a string of genes. Each gene has a possible value [46]. GA selects the generated chromosomes randomly (based on the elitist strategy). It improves them by one-point crossover and one-point mutation operators with a given probability. In several generations, chromosomes are evaluated through fitness function. Then, new populations of chromosomes are generated using the Goldberg's selection method (roulette wheel) and genetic operators. Selection of chromosome is performed based on its fitness value in the population. The termination criterion is specified by the maximum number of generations.

The evaluator functions of attribute selection are based on GA that are demonstrated as follows: 1. Correlation of features subsets with the class and intercorrelation between features. 2. Consistency of the worth of a subset of attributes with the class values. 3. C4.5 classifier subset. 4. The wrapper subset with C4.5 classifier. The first three are based on the filter approach and the fourth one is based on the wrapper approach. In addition, the decision tree constructed using GATree system [58] is used for selection of features that it is based on the embedded approach in feature selection. This algorithm constructs decision tree, which its nodes can be used as the final selected features.

In feature selection based on the filter and wrapper approaches, a chromosome is a set of credit scoring features. A gene is a feature or an input variable. Encoding of a gene is binary. Where '1' (0) means there is (not) a given feature in the set of credit scoring features. The strategy, to find an optimal set of variables (features), is based on Goldberg genetic algorithm. Evaluation of input variable subset is performed based on subset evaluator function with n-fold cross-validation. Further, the subset of features is evaluated according to two directions: 1. individual predictive ability of each feature. 2. The degree of redundancy among features. Initial population, maximum number of generations, mutation, crossover probability, cross validation, and random seed number were 20, 20, 0.01, 0.9, 10, and 1, respectively.

In evaluator function (based on the wrapper subset with C4.5 classifier), number of folds, seed number, and threshold were 10, 1, and 0.01, respectively.

Parameters values of, the decision tree constructed using GATree system (the embedded approach) in GATree v2 software, and C4.5 decision tree in Weka version 3.5.8, are shown in tables 1 and 2, respectively.

Moreover, in order to construct decision trees using Meta classifier algorithm, we considered the cross-validation approach to train and test decision trees. Genetic algorithm was used for feature selection (in our case, four feature selection methods (only based on the filter and wrapper approaches) were considered).

Table 1. Parameters values of the decision tree constructed using GATree system (the embedded approach)

Parameter	**Parameter Value**
Using the cross-validation approach in training and testing the decision tree constructed using GATree system	10
Standard random crossover (default)	0.99
Standard random mutation (default)	0.01
Percent of genome replacement (default)	0.25
Error rate (increase the acceptable error rate to accelerate the evolution) (default)	0.95
Activate dynamic alteration of the decision trees preference (Dynamically alter the preference for smaller or more accurate trees)	Yes
Prefer the more accurate trees compared with the smaller ones in the beginning and end of the evolution	Yes
Number of generations	100
Initial population size	100
Random seed number (default)	123456789

Table 2. Parameters values of C4.5 decision tree

Parameter	**Parameter value**
Using miss-classification matrix to build C4.5 decision tree	No
Unit cost of miss-classification for all classes in the target variable	1
Prune the C4.5 decision tree	Yes
The confidence factor used for pruning (smaller values incur more pruning)	0.25
The minimum number of instances per leaf to prune the decision tree, and set size and complexity	2
Does it use binary splits for nominal attributes while building trees	No
Number of folds (Determines the amount of data used for the reduced-error pruning)	3
Reduced Error Pruning (Whether reduced-error pruning is used instead of C4.5 pruning)	No
Seed number	1
Sub tree raising (Whether to consider the sub tree raising operation during the pruning)	Yes
Number of cross validation (to train and test C4.5 decision tree)	10

In the new proposed hybrid classification model, parameters values of SimpleKmeans clustering were regarded as follows: Number of clusters and seed number were set as 2 and 1, respectively. The target variable (class) was ignored in the dataset. In addition, all of the data in the dataset were considered as training dataset.

There are several methods to construct decision trees in the new proposed hybrid classification model in this paper. 1. Construction of five C4.5 decision trees by five feature selection methods. 2. Building four C4.5 decision trees by Meta classifier algorithm (combining feature selection and

C4.5 decision tree algorithms) using four feature selection algorithms based on the filter and wrapper approaches. 3. Using C4.5 decision tree in selection of the features and building a decision tree constructed using GATree system with the selected features (based on the embedded approach). 4. Decision tree constructed using GATree system. 5. Using decision tree constructed using GATree system in selection of the features and building decision tree using GATree system with the selected features (based on the embedded approach). 5. Using classifier (in this case: decision tree) hierarchies. It is an alternative method among several methods to combine classifiers [49]. This hierarchy is used to arrange single classifiers in a tree. In this regard, we firstly used GATree system to construct decision tree with two branches. Then, we used C4.5 decision tree algorithm to build a decision tree in each branch.

4.2. Evaluation methods

There are three evaluation methods to evaluate the prediction performance (optimization of classification models) of the new proposed hybrid credit scoring model and all other decision trees used for comparison in this paper. 1. Percentage of the correctly classified instances. The complexity of the decision tree that is indicated through: 2. Number of leaves of the tree. 3. Size of the tree.

Decision trees, as one of the simple ways of knowledge representation, classify instances into classes. They include nodes, edges, and leaves. They are labeled by attribute names, possible values for attribute, and different classes, respectively [52]. They have internal nodes and leaves. Each internal node has child nodes [52].

Each node is a place for a decision. Final decisions are made into nodes, which can be either of discrete or continuous values. Decisions with discrete values are made in a developed classification tree [63].

With respect to Weka machine learning tool, size of the tree illustrates number of branches from node to leaves of the decision tree. It is equal to sum of leaves and nodes in the decision tree.

5. EXPERIMENTAL RESULTS

We used descriptive statistics, machine learning, and data mining tools to obtain experimental results. Weka 3.5.8 version and GATree v2 (an unregistered version) software and Microsoft Excel 2007 software have been employed to analyze the results in this paper. Table 3 elaborates characteristics of the decision tree constructed by the new proposed hybrid classification model in the Bank Mellat credit dataset.

In the next tables (4-7), we compare decision tree constructed by the new proposed hybrid classification model with the other C4.5 decision trees in the Bank Mellat credit dataset.

Also, table 8 compares classification accuracy (correctly classified instances) of some classification models in this paper with decision tree constructed by the new proposed hybrid classification model. These are: Naïve Bayes, KNN classifier (K=2), CHAID (Chi-squared Automatic Interaction Detection) tree, Random Forest classifier, multilayer perceptron,

sequential minimal optimization and logistic regression. Also, training and testing the compared classification models were performed based on 10-fold cross-validation.

Table 3. Characteristics of the decision tree constructed by the new proposed hybrid classification model, in the Bank Mellat credit dataset

Classification algorithm	**Total number of instances**	**number of the selected predictive attributes**	**Correctly Classified Instances**	**Percentage of the correctly classified instances**	**Number of Leaves**	**Size of the tree**	**Class accuracy of customers(1)**	**Class accuracy of customers(2)**	**Class accuracy of customers(3)**
The new proposed hybrid classification model	5173	17	4992	96.501%	213	290	0.9839	0.9594	0. 8794

Table 4. Characteristics of C4.5 constructed without feature selection and clustering, in the Bank Mellat credit dataset

Total number of instances	**Correctly Classified Instances**	**Percentage of the correctly classified instances**	**Number of Leaves**	**Size of the tree**	**Class accuracy of customers(1)**	**Class accuracy of customers(2)**	**Class accuracy of customers(3)**
5173	4964	95.96%	316	398	0.979	0.945	0.887

Table 5. Characteristics of C4.5 constructed with feature selection based on Genetic Algorithm and without clustering, in the Bank Mellat credit dataset

Evaluator Function of Attribute Selection, Based on Genetic Algorithm	**Total number of instances**	**Correctly Classified Instances**	**Percentage of the correctly classified instances**	**Number of Leaves**	**Size of the tree**	**Class accuracy of customers(1)**	**Class accuracy of customers(2)**	**Class accuracy of customers(3)**
Wrapper Subset with C4.5 Classifier	5173	4991	96.4817%	297	380	0.982	0.95	0.9
Correlation of the Features Subsets with the Class and Intercorrelation between Features	5173	4938	95.4572%	226	323	0.974	0.94	0.881
Consistency of the worth of a Subset of Attributes with the Class Values	5173	4926	95.2252%	238	326	0.976	0.923	0.885
C4.5 Classifier Subset	5173	4951	95.7085%	277	372	0.98	0.932	0.888

Table 6. Characteristics of C4.5 constructed with feature selection based on the Best First search Algorithm and without clustering, in the Bank Mellat credit dataset

Evaluator Function of Attribute Selection, Based on Genetic Algorithm	**Total number of instances**	**Correctly Classified Instances**	**Percentage of the correctly classified instances**	**Number of Leaves**	**Size of the tree**	**Class accuracy of customers(1)**	**Class accuracy of customers(2)**	**Class accuracy of customers(3)**
Wrapper Subset with C4.5 Classifier	5173	4998	96.6171%	318	389	0.984	0.965	0.879
Correlation of Features Subsets with the Class and Intercorrelation between Features	5173	4943	95.5538%	215	315	0.974	0.942	0.884
Consistency of the worth of a Subset of Attributes with the Class Values	5173	4980	96.2691%	276	352	0.979	0.958	0.886
C4.5 Classifier Subset	5173	4968	96.0371%	314	404	0.98	0.938	0.901

Table 7. Characteristics of C4.5 constructed with feature selection based on Genetic Algorithm and considering the feature of "type of cluster", in the Bank Mellat credit dataset

Evaluator Function of Attribute Selection, Based on Genetic Algorithm	**Total number of instances**	**Correctly Classified Instances**	**Percentage of the correctly classified instances**	**Number of Leaves**	**Size of the tree**	**Class accuracy of customers(1)**	**Class accuracy of customers(2)**	**Class accuracy of customers(3)**
Wrapper Subset with C4.5 Classifier	5173	4986	96.3851%	346	437	0.983	0.953	0.888
Correlation of Features Subsets with the Class and Intercorrelation between Features	5173	4943	95.5538%	215	315	0.974	0.942	0.884
Consistency of the worth of a Subset of Attributes with the Class Values	5173	4962	95.9211%	240	310	0.974	0.957	0.888
C4.5 Classifier Subset	5173	4970	96.0758%	301	400	0.979	0.949	0.889

Table 8. Classification accuracy of some classification models in this paper, compared with decision tree constructed by the new proposed hybrid classification model, in the Bank Mellat credit dataset

Classification Models	**Naive Bayes**	**KNN classifier (K=2)**	**CHAID tree**	**Random Forest classifier**	**multilayer perceptron (MLP)**	**sequential minimal optimization (SMO)**	**logistic regression**
Classification Accuracy (Correctly Classified Instances)	84.99%	95.5%	80.5%	95.8%	93.6%	82.25%	83.32%

6. DISCUSSIONS AND CONCLUSION

Banks require customer credit scoring to be able to appropriately offer credit facilities to their customers. Decision trees, as one of the classification techniques in data mining, can help to perform customer credit scoring. The main problem is the construction of decision trees to be able to classify bank customers optimally. This study has proposed a new hybrid classification model for designing a customer credit scoring model for banks (such as Bank Mellat). It is applied to offer credit facilities to each class. In the model, the development process in pattern recognition and CRISP-DM process are used in construction of the final decision tree for customers' credit scoring. The new proposed hybrid classification model is resulted from combination of clustering, feature selection, decision trees, and genetic algorithm techniques.The experimental results demonstrate the classification accuracy of the decision tree constructed by the new proposed hybrid classification model. It was higher than all of the compared decision trees throughout this paper. The only case that had better classification accuracy (approximately 0.1% higher) was C4.5 decision tree constructed by feature selection based on the best first search and wrapper evaluator function with C4.5 classifier. However, it was a large tree containing more leaves. Number of leaves and size of the tree in the decision tree (i.e. complexity) of the new proposed hybrid classification model in this paper were lower than all of the 13 compared decision trees. This shows that the decision tree of the new proposed hybrid classification model has even higher accuracy, and lower complexity. Moreover, with respect to tables 8 (related to Bank Mellat credit dataset), it is shown that classification accuracy of the decision tree constructed by the new proposed hybrid classification model is more than the entire other classification models compared in this paper. Therefore, the decision tree of the new proposed hybrid classification model was better than those decision trees, with lower complexity and higher classification accuracy and other classification models with higher classification accuracy. With respect to the issues addressed above, it is apparent that the new proposed hybrid classification model can be used for construction of the more suitable decision trees for credit scoring of bank customers.

Regarding the literature review and the new proposed hybrid classification model, there are two issues to be recommended in this paper. 1. Consideration of the miss-classification cost in decision tree algorithms and miss-selection of features in feature selection algorithms in the new proposed hybrid classification model. 2. Development of the new proposed hybrid classification model using other decision trees classification methods (like ID3, QUEST, CHAID and C&RT) or other classification models (like naïve bayes, KNN, neural networks, SVM and logistic regression). Moreover, there are some applicable recommendations to banks as follows: 1. applying the new proposed hybrid classification model in credit scoring of bank customers to be used for offering credit facilities beneficially. 2. Using the development process in pattern recognition in order to construct the classification models for better customer credit scoring in banks. 3. Designing and constructing a decision support system and applicable software, based on the new proposed hybrid classification model, for credit scoring of bank customers

REFERENCES

[1] Gray, J. B. and Fan, G. (2008) 'Classification tree analysis using TARGET', Computational Statistics & Data Analysis, Vol. 52, pp.1362-1364

[2] Carvalho, D. R. and Freitas, A. A. (2004) 'A hybrid decision tree/genetic algorithm method for data mining', Information Sciences, Vol. 163, pp.16 and 17

[3] Yoshida, H., Leardi, R., Funatsu, K., &Varmuza, K. (2001). Feature selection by genetic algorithms for mass spectral classifiers. Analytica Chimica Acta, 446, 486.
[4] Tsai, C. F. and Chen, M. L. (2010) 'Credit Rating by Hybrid Machine Learning Techniques', Applied Soft Computing, Vol. 10, pp.1,3
[5] Kim, Y. S. and Sohn, S. Y. (2004) 'Managing loan customers using misclassification patterns of credit scoring model', Expert Systems with Applications, Vol. 26, pp.567
[6] Hu, Y. C. and Ansell, J. (2007) 'Measuring retail company performance using credit scoring techniques', European Journal of Operational Research, Vol. 183, pp.1595
[7] Abdou, H. and Pointon, J. (2008) 'Neural nets versus conventional techniques in credit scoring in Egyptian banking', Expert Systems with Applications, Vol. 35, pp.1275
[8] Angelini, E., Tollo, G. d. and Roli, A. (2008) 'A neural network approach for credit risk evaluation', The Quarterly Review of Economics and Finance, Vol. 48, pp.733
[9] Min, J. H. and &Lee, Y. C. (2008) 'A practical approach to credit scoring', Expert Systems with Applications, Vol. 35, pp.1762
[10] Sustersic, M., Mramor, D. and Zupan, J. (2009) 'Consumer credit scoring models with limited data', Expert Systems with Applications, Vol. 36, pp.4736
[11] Xu, X., Zhou, C. and Wang, Z. (2009) 'Credit scoring algorithm based on link analysis ranking with support vector machine', Expert Systems with Applications, Vol. 36, pp.2625
[12] Bellotti, T. and Crook, J. (2009) 'Support vector machines for credit scoring and discovery of significant features', Expert Systems with Applications, Vol. 36, pp.3302
[13] Huang, S. C. (2009) 'Integrating nonlinear graph based dimensionality reduction schemes with SVMs for credit rating forecasting', Expert Systems with Applications, Vol. 36, pp.7515
[14] Luo, S. T., Cheng, B. W. and Hsieh, C. H. (2009) 'Prediction model building with clustering-launched classification and support vector machines in credit scoring', Expert Systems with Applications, Vol. 36, pp.7562
[15] Setiono, R., Baesens, B. and Mues, C. (2009) 'A Note on Knowledge Discovery Using Neural Networks and its Application to Credit Card Screening', European Journal of Operational Research, Vol. 192, pp.326
[16] Wang, C. M. and Huang, Y. F. (2009) 'Evolutionary-based feature selection approaches with new criteria for data mining:A case study of credit approval data', Expert Systems with Applications, Vol. 36, pp.5900
[17] 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, pp.2473
[18] Yoon, J. S. and Kwon, Y. (2009) 'A practical approach to bankruptcy prediction for small businesses: Substituting the unavailable financial data for credit card sales information', Expert Systems with Applications, Vol. 37, pp.1
[19] Yu, L., Wang, S. and Lai, K. K. (2009) 'An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support:The case of credit scoring', European Journal of Operational Research, Vol. 195, pp.942,943
[20] Zhou, L., Lai, K. K. and Yu, L. (2009) 'Credit scoring using support vector machines with direct search for parameters selection', Soft Computing, Vol. 13, pp.149
[21] Kim, H. S. and Sohn, S. Y. (2010) 'Support Vector Machines for Default Prediction of SMEs based on Technology Credit', European Journal of Operational Research, Vol. 201, pp.838
[22] Tansel, Y. and Yurdakul, M. (2010) 'Development of a quick credibility scoring decision support system using fuzzy TOPSIS', Expert Systems with Applications, Vol. 37, pp.567
[23] Ping, Y. and Yongheng , L. (2011) 'Neighborhood rough set and SVM based hybrid credit scoring classifier', Expert Systems with Applications, Vol. 38, pp. 11300-11304
[24] Kao, L. J., Chiu, C. C. and Chiu, F. Y. (2012) 'A bayesian latent variable model with classification and regression tree approach for behavior and credit scoring', Knowledge – Based Systems, Vol. 36, pp. 245-252.

[25] Hens, A. B. and 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, Vol. 39, pp. 6774-6781
[26] Vukovic, S., Delibasic, B., Uzelac, A. and Suknovic, M. (2012) 'A case-based reasoning model that uses preference theory functions for credit scoring', Expert Systems with Applications, Vol. 39, pp. 8389-8395
[27] Capotorti, A. and Barbanera, E. (2012) 'Credit scoring analysis using a fuzzy probabilistic rough set model', Computational Statistics and Data Analysis, Vol. 56, pp. 981-994
[28] Wang, J., Hedar, A. R., Wang, S. and Ma, J. (2012) 'Rough set and scatter search metaheuristic based feature selection for credit scoring', Expert Systems with Applications, Vol. 39, pp. 6123-6128
[29] Blanco, A., Mejias, R. P., Lara, J. and Rayo, S. (2013) 'Credit scoring models for the microfinance industry using neural networks: Evidence from Peru', Expert Systems with Applications, Vol. 40, pp. 356-364
[30] West, D., Dellana, S. and Qian, J. (2005) 'Neural network ensemble strategies for financial decision applications', Computers & Operations Research, Vol. 32, pp.2543
[31] Tsai, C. F. and Wu, J. W. (2008) 'Using neural network ensembles for bankruptcy prediction and credit scoring', Expert Systems with Applications, Vol. 34, pp.2639
[32] Nanni, L. and Lumini, A. (2009) 'An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring', Expert Systems with Applications, Vol. 36, pp.3028
[33] Twala, B. (2010) 'Multiple classifier application to credit risk assessment', Expert Systems with Applications, Vol. 37, pp.1
[34] Hsieh, N. C. and Hung, L. P. (2010) 'A data driven ensemble classifier for credit scoring analysis', Expert Systems with Applications, Vol. 37, pp.534
[35] Paleologo, G., Elisseeff, A. and & Antonini, G. (2010) 'Subagging for credit scoring models', European Journal of Operational Research, Vol. 201, pp.490
[36] Yu, L., Yue, W., Wang, S. and Lai, K. (2010) 'Support vector machine based multiagent ensemble learning for credit risk evaluation', Expert Systems with Applications, Vol. 37, pp.1351
[37] Zhou, L., Lai, K. K. and Yu, L. (2010) 'Least squares support vector machines ensemble models for credit scoring', Expert Systems with Applications, Vol. 37, pp.127
[38] Lee, T. S., Chiu, C. C., Lu, C. J. and Chen, I. F. (2002) 'Credit scoring using the hybrid neural discriminant technique', Expert Systems with Applications, Vol. 23, pp.245
[39] Jacobson, T. and Roszbach, K. (2003) 'Bank lending policy, credit scoring and value-at-risk', Journal of Banking & Finance, Vol. 27, pp.615
[40] Lee, T. S. and Chen, I. F. (2005) 'A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines', Expert Systems with Applications, Vol. 28, pp.743
[41] Chen, W., Ma, C. and Ma, L. (2009) 'Mining the customer credit using hybrid support vector machine technique', Expert Systems with Applications, Vol. 36, pp.7611
[42] Chuang, C. L. and Lin, R. H. (2009)'Constructing a reassigning credit scoring model', Expert Systems with Applications, Vol. 36, pp.1685
[43] Lin, S. L. (2009) 'A new two-stage hybrid approach of credit risk in banking industry', Expert Systems with Applications, Vol. 36, pp.8333
[44] Akkoç, S. (2012) 'An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data', European Journal of Operational Research, Vol. 222, pp. 168-178
[45] Kim, E., Kim, W. and Lee, Y. (2002) 'Combination of multiple classifiers for the customer's purchase behavior prediction', Decision Support Systems, Vol. 34, pp.167
[46] D'heygere, T., Goethals, P. L. and Pauw, N. D. (2003) 'Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates', Ecological Modelling, Vol.160, pp.291-293
[47] Liu, H. H. and Ong, C. S. (2008) 'Variable selection in clustering for marketing segmentation using genetic algorithms', Expert Systems with Applications, Vol. 34, pp.502

[48] Tan, F., Fu, X., Zhang, Y. and Bourgeois, A. G. (2008) 'A genetic algorithm-based method for feature subset selection', Soft Computing, Vol. 12, pp.111

[49] Martinez-Otzeta, J. M., Sierra, B., Lazkano, E. and Astigarraga, A. (2006) 'Classifier hierarchy learning by means of genetic algorithms',Pattern Recognition Letters, Vol. 27, pp.1998

[50] D'heygere, T., Goethals, P. L. and Pauw, N. D. (2006) 'Genetic algorithms for optimization of predictive ecosystems models based on decision trees and neural networks', Ecological Modelling, Vol. 195, pp.20

[51] Huang, M., Gong, J., Shi, Z., Liu, C. and Zhang, L. (2007) 'Genetic algorithm-based decision tree classifier for remote sensing mapping with SPOT-5 data in the HongShiMao watershed of the loess plateau, China', Neural Computing & Applications, Vol. 16, pp.513

[52] Sorensen, K. and Janssens, G. K. (2003) 'Data mining with genetic algorithms on binary trees', European Journal of Operational Research, Vol. 151, pp.253-255

[53] Dehuri, S., Patnaik, S., Ghosh, A.and Mall, R. (2008) 'Application of elitist multi-objective genetic algorithm for classification rule generation', Applied Soft Computing, Vol. 8, pp.477

[54] Kennedy, R. L., Lee, Y., Roy, B. V., Reed, C. D. and Lippmann, R. P. (1998) Solving Data Mining Problems through Pattern Recognition, Copyright, Unica Technologies.Inc 1995-1997. Prentice Hall PRT New Jersey.

[55] Sharifi, K. (2009) 'Credit scoring in Bank Mellat', (Khanbabaei, M. interviewer)

[56] Ong, C. S., Huang, J. J. and Tzeng, G. H. (2005) 'Building credit scoring models using genetic programming', Expert Systems with Applications, Vol. 29, pp.41

[57] Thomas, L. C. (2000) 'A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers', International Journal of Forecasting, Vol. 16, pp.151-152

[58] Papagelis, A. and Kalles, D. (2001) 'Breeding Decision Trees Using Evolutionary Techniques'. Paper presented at the International Conference on Machine Learning, Williamstown, Massachusetts, pp.1-7

[59] Larose, D. T. (2005) Discovering Knowledge in Data, an Introduction to Data Mining, New Jersey: WILEY.

[60] Olson, D. and Shi, Y. (2007) Introduction to Business Data Mining, Singapore: McGraw Hill Education.

[61] Salappa, A., Doumpos, M. and Zopounidis, C. (2007) 'Feature selection algorithms in classification problems:an experimental evaluation', Optimization Methods and Software, Vol. 22, pp.199,200,202

[62] Wang, Y. Y. and LI, J. (2008) 'Feature-selection ability of the decision-tree algorithm and the impact of feature-selection/extraction on decision-tree results based on hyperspectral data', International Journal of Remote Sensing, Vol. 22, pp.2994

[63] Hsu, P. L., Lai, R., Chiu, C. C. and Hsu, C. I. (2003) 'The hybrid of association rule algorithms and genetic algorithms for tree induction: an example of predicting the student course performance', Expert Systems with Applications, Vol. 25, pp.51