International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Credit Risk Analysis Using Fuzzy Logic with Machine Learning Models

G. Arutjothi[1], Dr. C. Senthamarai[2]

[1]Ph.D. Research Scholar, Department of Computer Science, Govt. Arts College (Autonomous), Salem-7, Tamil Nadu, India

[2]Assistant Professor, Department of Computer Applications, Govt. Arts College (Autonomous), Salem-7, Tamil Nadu, India

Abstract:
Credit Risk is an important issue in the Banking Industry. Credit risk Prediction and assessment of credit is a difficult task for credit managers. Banking Industry has large amount of data related to the behavior of the customer and their credit history, but this raw data is not useful for making correct judgment in credit decisions. The banking industry is need of a correct credit decision making system, to distinguish between good customers and default customers. The data mining domain is suitable for assessing credit risk and making good decisions on credit. Feature selection is one of the main pre-processing step in the data mining. This paper use FuzzyRoughSetTheory (FRST) for finding feature subset. Four other feature selection methods are used to find the optimal feature subset. The four feature selection methods are Information Gain, Relief, Chi-Squared and Wrapper subset model. These different feature selection methods are compared in terms of accuracy and efficiency.

Keywords: Credit Risk, Feature Selection, FuzzyRoughSetTheory(FRST)

I. INTRODUCTION

In today's Banking Industry always used for credit evaluation Systems. Banks are having a tremendous amount of data related to the individual's and company's behavioral data and their credit history, but those are can't able to find good judgment of credit applicants. The process of lending can be divided into four main phases, including preapplication, application, performance and collection [15]. In this paper, we will find the credit risk problem which is critical issue in the application phase. Banks lose their capital amount based on this loan credit risk. Credit evaluation system is used to predict the credit score of the applicant's, those were belongs to either good credit or bad credit.

Credit risk analysis system is used statistical methods and data analysis techniques to evaluate the credit risk against the customers. These types of systems are used historical data for borrower behavior and credit related data. This system is mainly uses on predicting the creditworthiness of credit applicants. Some Banks uses to perform the credit risk of customers, but some are not use in still now. Most of the researchers concentrated to solve this credit risk problem in last decade but still now want to improve the credit risk evaluation gradually [17]. In this paper uses some benchmark credit dataset.

Data mining is a promising area for predicting and assessing the large amount of the credit dataset. Data Mining, also popularly referred to as Knowledge Discovery from Data (KDD), is the

automated extraction of patterns from large amount of data. Data methods are used for many applications like business, forecasting, predicting, banks and Governments etc,. To classify the customer as good and bad credit risk using this data mining classification techniques. Many classification methods are used to the credit risk analysis problem for last decades. Most used techniques are Decision Tree(DT), K-Nearest Neighbor(K-NN), Support Vector Machines(SVM) and Neural Network [10].

The main aim of the credit evaluation system is to correctly classify the customers. Banks dataset having the many irrelevant and redundant features, this way lead to low classification accuracy in credit evaluation systems. Credit evaluation system used for this irrelevant dataset is not ability to making a good decision. So that feature selection is important role in the data mining process. In that case feature selection needed for the credit risk dataset [16]. In order to select a subset of relevant features, feature selection is needed. The subset is sufficient to describe the problem with high precision. Feature selection is used to find the optimal feature subset which is provides the low cost and computational complexity.

In this study, we used fuzzyRoughSetTheory for feature selection based on various criteria at credit scoring tasks.

The remaining of the paper is structured as follows. In section II describes the basic concepts and literature survey on credit risk. In section III discuss on methodology and data to this work followed by result and discussions are made in section IV. In section V discuss the conclusion.

II. RELATED WORK

Feature selection can be defined as a process that chooses a minimum subset of features from the original set of features. Feature selection is used to reduce the feature sub space optimally. Feature selection is also known as the variable selection or attributes selection. Feature selection methods are classified as three categories, such as Filter methods, Wrapper methods and embedded methods

In filter type methods select variables regardless of the model. Filter methods suppress the least interesting variables. These methods are particularly effective in computation time and robust to overfitting[14]. Some techniques are Gini, information gain, the ratio of information gain, etc,.

In wrapper methods evaluate subsets of variables which allows, unlike filter approaches, to detect the possible interactions between variables. There are two main disadvantages of this method. The first one is increasing overfitting risk when the number of observations is insufficient and second one is significant computation time when the number of variables is large [7]. Common strategies are sequential wrapper Forward Selection (SFS) and reverse sequential Elimination (SBE). Embedded methods have been recently proposed that try to combine the advantages of both previous methods.

We studied various articles regarding subset selection of feature selection techniques on different tools; some of them are described here Credit risk assessment is one of the crucial issues in the financial organizations. Credit scoring is widely analyzed using classification techniques. Credit scoring is a typical data mining, classification problem. Feature selection algorithms identify the features that are relevant but not redundant to the solution. The major task is to rank the relevant features based on their fitness values. Most of the soft computing techniques are suitable for finding the best fitness values [1]. All type of feature selection methods are used for improving the classification accuracy [2]. Cintra [3]

applied fuzzy system based wrapper model to find the optimal feature subset and also get lowest error rate. Ephzibah[4] focused on soft computing based fuzzy feature selection. The combination of Genetic Algorithm (GA) and Fuzzy logic gives high performance than other techniques. Gönen [5] proposed a new feature selection with probit classifier based credit evaluation system. Fuzzy Rough Set based feature selection is used on high dimensional dataset and its provides high accuracy with optimal feature subset by Changyou [6]. Van-Sang [7] proposed a new hybrid feature selection methods which is combination of information gain and H2O gradient boosted model. AIS based fuzzy classifier system proposed by Kamalloo [8], he used Australia, German credit dataset from UCI. The combination of Multilayer Perceptron and Fuzzy based hybrid classifier was proposed by Mehdi [9] for credit risk evaluation. Louzada, Francisco, Anderson Ara [11], surveyed the classification techniques and find the best classification methods for credit scoring. Meenachi [12] applied Fuzzy Rough Set to cancer prediction data which is suited and get better accuracy than other techniques.

Finally, even if there is a hundreds of research, in feature selection, it is still hard to say which feature selection technique is the best. Each technique depends on a particular attributes set, so it is very important to optimize the feature subset. Effective variable selection methods are lacking.

III. RESEARCH METHODOLOGY

i. Proposed Model

In our proposed model, we have to identify the appropriate set of features by eliminating the irrelevant and redundant features to improve the performance of the classifier. The Fuzzy Rough Subset evaluation method is used in combine with a PSO Search algorithm for feature selection. PSO uses a number of agents (particles) that constitute a swarm, which is moving around the search space looking for the best solution [12]. The performance of the classifier is evaluated using evaluation classification accuracy.

ii. Proposed Architecture

Fig.1. shows the proposed architecture, the whole credit data are taken from Financial Institutions; it is analyzed to provide the useful information. This is really a complex or critical work on the financial institution. The proposed work makes decision whether the loan can be approved or rejected for the new potential customer.

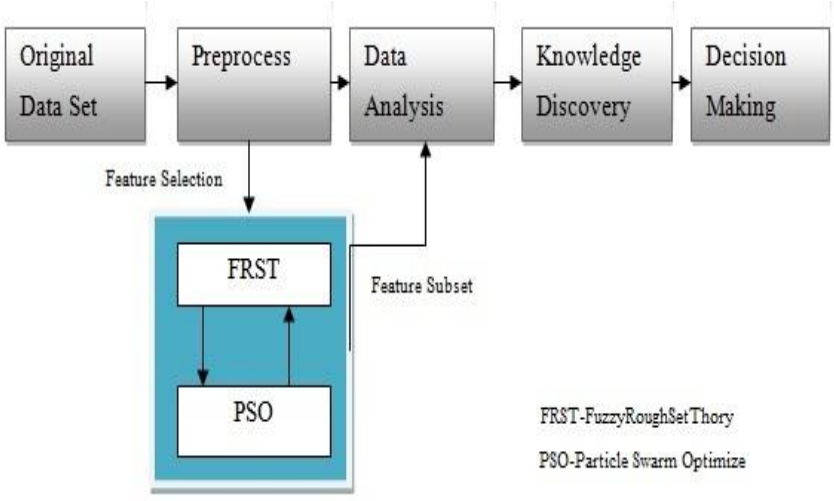

Fig*1.The proposed architecture for a credit evaluation system using FRST+PSO*

Fuzzy Rough Set based feature selection with combination of particle swarm optimization search method is used for this proposed work. This proposed model finds the optimized feature subset which is most relevant to the problem. This way we find the good and bad borrowers.

iii. Fuzzy Rough Set Theory

A fuzzy rough set is a derive from the approximation of a fuzzy set. The decision attribute values are fuzzy, the conditional values are crisp where corresponds to this case. The upper and lower approximations incorporate the extent to which objects belong to these sets, and are defined as:

$$\mu_{RX}([x]_R) = \inf\{\mu_X(x)|x \in [x]_R\}, \quad (1)$$

$$\mu_{RX}([x]_R) = \sup\{\mu_X(x)|x \in [x]_R\} \ldots\ldots \quad (2)$$

where $\mu_X(x)$ is the degree to which x belongs to fuzzy equivalence class $_X$, and each$[x]_R$ is crisp. The value <RX,RX> is called a rough-fuzzy set.

Rough-fuzzy sets can be generalized to fuzzy rough sets, where all equivalence classes may be fuzzy. When apply to data analysis, this means that both the decision values and the conditional values may be fuzzy or crisp. Fuzzy-rough set-based Feature Selection (FRST) is based on the notion of fuzzy lower approximation to reduce the datasets containing real valued features. The process becomes a crisp approach when dealing with nominal well-defined features. Positive value is defined as the union of lower approximations.

iv. Particle Swarm Optimization

Originally particle swarm optimization was devdelpoed by Kennedy and Eberhart (1995). The main idea of PSO is to mimic social behavior of birds. In PSO algorithm, each particle can move along the linear combination of its personal velocity, towards best global position and towards best local of its personal position in the problem space. In this way it is used to minimize the feature space. This paper, PSO is used for searching strategy for feature selection. The combination of PSO search based FRST system find the minimal feature subset. This features provides the high classification accuracy with classifier.

IV. RESULTS AND DISCUSSION

The FuzzyRoughSubsetEvaluator package in Python language has been used to demonstrate our proposed work. Feature selection is employed using the select attribute subset. Fuzzy rough subset evaluator is chosen for attribute evaluator and Particle Swarm Optimization (PSO) is chosen for select attribute. The German Credit dataset was collected from UCI public datasets repository which has 21 features [18]. The German credit dataset consists of 1000 loan applications, with 700 instances of good customer and 300 instances of default customer. For each applicant, 20 attributes describe the credit history, account balances, loan information and personal information; one is dependent or class attribute. After applying PSO search based FRST to evaluate the attributes and finally optimized the attributes. This optimal feature subset is best suitable for predicting credit risk and this is provide the high accuracy. This optimized feature subset has 12 attributes, which are irrelevant or removed from dataset.

Table 1: Comparison of accuracy on PSO-based FRST feature selection and the other feature selection methods for the German dataset.

Feature Selection	Reduct Features (original 21)	Iteration 1 Accuracy			Iteration2 Accuracy		
		Decision Tree	KNN	SVM	Decision	KNN	SVM
Information Gain	16	75.3	73.6	77.3	73.5	71	77.5
Relief	15	73.6	73	77.6	72	73	77
Chi-squared	16	76.3	70.3	78.6	73	71.5	78
Wrapper Model	10	77.3	74	76.6	75	73.5	76
FRST + PSO	12	73	**82.4**	75.3	78	81.6	75.5

Table 1 shows the Classification Results

This paper used five types of feature selection techniques: information gain, Relief, chi-squared, wrapper model and FRST, which are applied to the dataset and get a few feature subset. Then, Decision tree (J48), K-NN (1BK) and SVM (SMO) classify the dataset using these feature subsets. Results are presented in Table1. The best results are shown in bold.

As shown in Table 1 for comparing the accuracy of various methods, we saw that the accuracy of PSO search based FRST on the subset of newly selected features has been obviously improved and the number of features has been reduced. The average accuracy is low on the original data. After applying the feature selection, the average accuracy is 82.4%. This result emphasizes the efficiency of our method in terms of running time due to efficiently filtering the redundant features. Fig. 1 shows the results obtained with the different data mining techniques with different feature selection for classification on the German credit dataset.

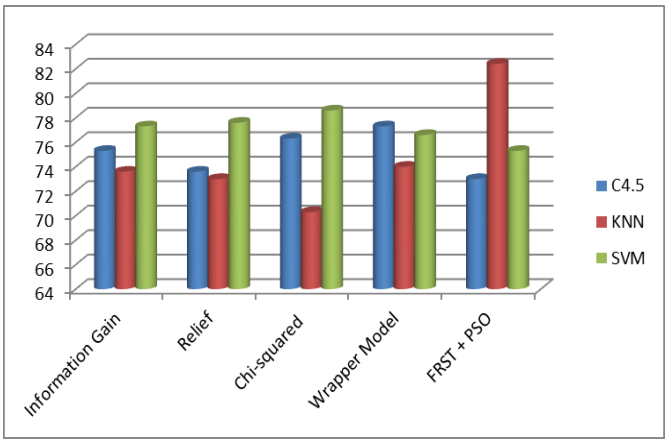

Figure 2 shows the averages of classification results.

The highest accuracy in the credit risk dataset is provided based on proposed method. Four feature selection methods and proposed FuzzyRoughSet model result are analyzed. The proposed method shows high accuracy.

V. CONCLUSION

One of the biggest issue in the machine learning and statistics is to find the optimal feature subset. In classification accuracy is based on feature selection. The main objective of the feature selection is to reduce the size of dimensions, costs and increase the classification accuracy. This research paper uses a Particle Swarm Optimization (PSO) search based FuzzyRoughSetTheory (FRST) model for finding the optimal feature subset to evaluate the credit risk. Credit risk assessment is difficult task for the Banking Industry. This study is to compare the feature selection techniques with three classifiers, Decision Tree, SVM and K-Nearest Neighbour. Finally, we find the best credit risk evaluation technique. KNN with FRST feature selection is best for the credit risk prediction system. The result of this system is effective in credit risk evaluation process.

REFERENCES

1. Bernardo, Dario, Hani Hagras, and Edward Tsang. "A genetic type-2 fuzzy logic based system for financial applications modelling and prediction." *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*. IEEE, 2013.
2. Bouaguel, W. "*On Feature Selection Methods for Credit Scoring"*. Diss. Ph. D. thesis, Institut Superieur de Gestion de Tunis, 2015.
3. Cintra, Marcos Evandro, et al. "Feature subset selection using a fuzzy method." *Intelligent Human-Machine Systems and Cybernetics, 2009. IHMSC'09. International Conference on*. Vol. 2. IEEE, 2009.
4. Ephzibah, E. P. "Cost effective approach on feature selection using genetic algorithms and fuzzy logic for diabetes diagnosis." *arXiv preprint arXiv:1103.0087* (2011).
5. Gönen, Gülefşan Bozkurt, Mehmet Gönen, and Fikret Gürgen. "Probabilistic and discriminative group-wise feature selection methods for credit risk analysis." *Expert Systems with Applications* 39.14 (2012): 11709-11717.
6. Guo, Changyou, and Xuefeng Zheng. "Feature subset selection approach based on fuzzy rough set for high-dimensional data." *Granular Computing (GrC), 2014 IEEE International Conference on*. IEEE, 2014.
7. Ha, Van-Sang, Nam Nguyen Ha, and Hien Nguyen Thi Bao. "A hybrid feature selection method for credit scoring." *EAI Endorsed Trans. Context-aware Syst. & Appl.* 4.11 (2017): e2.
8. Kamalloo, Ehsan, and Mohammad Saniee Abadeh. "Credit risk prediction using fuzzy immune learning." *Advances in Fuzzy Systems* 2014 (2014): 7.
9. Khashei, Mehdi, and Akram Mirahmadi. "A Soft Intelligent Risk Evaluation Model for Credit Scoring Classification." *International Journal of Financial Studies* 3.3 (2015): 411-422.
10. Lahsasna, Adel, Raja Noor Ainon, and Ying Wah Teh. "Credit Scoring Models Using Soft Computing Methods: A Survey." *Int. Arab J. Inf. Technol.* 7.2 (2010): 115-123.
11. Louzada, Francisco, Anderson Ara, and Guilherme B. Fernandes. "Classification methods applied to credit scoring: Systematic review and overall comparison." *Surveys in Operations Research and Management Science* 21.2 (2016): 117-134.

12. Meenachi, L., et al. "Diagnosis of Cancer using Fuzzy Rough Set Theory." International Research Journal of Engineering And Technology , Volume: 03 Issue: 01 (2016).
13. Mei, Xueyan, and Yilin Jiang. "Association rule-based feature selection for credit risk assessment." *Online Analysis and Computing Science (ICOACS), IEEE International Conference of*. IEEE, 2016.
14. Ramya, R. S., and S. Kumaresan. "Analysis of feature selection techniques in credit risk assessment." *Advanced Computing and Communication Systems, 2015 International Conference on*. IEEE, 2015.
15. Sadatrasoul, Seyed, Mohammad Gholamian, and Kamran Shahanaghi. "Combination of Feature Selection and Optimized Fuzzy Apriori Rules: The Case of Credit Scoring." *International Arab Journal of Information Technology (IAJIT)*12.2 (2015).
16. Van Sang, Ha, Nguyen Ha Nam, and Nguyen Duc Nhan. "A novel credit scoring prediction model based on Feature Selection approach and parallel random forest." *Indian Journal of Science and Technology* 9.20 (2016).
17. Van-Sang, Ha, and Nguyen Ha-Nam. "Credit scoring with a feature selection approach based deep learning." *MATEC Web of Conferences*. Vol. 54. EDP Sciences, 2016.
18. UCImachinelearningreposistory