Domain Driven Classification of Customer Credit Data for Intelligent Credit Scoring using Fuzzy set and MC2

K.Shanmugapriya
Department of Information Technology
Sri Eshwar College of Engineering
Coimbatore, India
shanmugapriya.kumaresan@gmail.com

Abstract—**Credit scoring or credit risk assessment is an important research issue in the banking industry. The major challenge of credit scoring is to recruit the profitable customers by predicting the bankrupts. The credit scoring carried out by traditional data driven approaches resulted only in an imprecise solution. Also the domain-driven based multiple criteria and multiple constraint (MC2) level programming approach results only in a satisfying solution. In this paper, a fuzzy set based domain driven approach for classification of customer credit data has been provided. The multiple criteria and multiple constraint level programming are used for scoring the customers based on the classifier. The domain expertise knowledge is used for building the linear combinational sets of attributes for classification. This hybrid approach will identify the class of best, good, satisfactory, bad and worst customers. Experiments are based on publicly available datasets in the UCI Machine Learning Repository.**

Keywords- Credit scoring, classification, domain-driven approach, linear combination, fuzzy set, MC2.

I. INTRODUCTION

Due to prevalent adoption of internet banking, the use of credit cards has increased. The banks collect copious information about cardholder's transactions. Extracting knowledge from these transaction records and credit cardholder's personal record increases profit in banking sector. It is essential to classify credit card customers precisely into different classes to avoid losses due to bankrupts.

In general, credit scoring is used to analyze a sample of past customers behavior to differentiate the present customers into bankrupts and non-bankrupts (i.e., bad and good customers respectively). A mathematical model is devised by a set of attributes for carrying out the process of classification. By assigning weight to each attribute, score for each customer can be generated based on classification.

The credit scoring has become a challenging task due to lack of domain knowledge of banking industry. Previous works used only the linear and logistic regression models for scoring, which results in an imprecise solution. A new framework for credit scoring based on domain knowledge is proposed in this paper. The dataset used for experiment is collected from UCI machine learning repository.

II. BACKGROUND

A) Data Mining

Data mining refers to a set of methods and techniques used to extract some useful information and hidden patterns in the huge amount of data [1]. Data mining also refers to all aspects of an automated or semi-automated process for extracting useful information. This process consists of numerous steps such as integration of data from different sources, preprocessing of data and developing a model with some algorithms.

B) Classification

Classification is a supervised learning approach which involves finding rules that partition the data into disjoint groups. The classification is a data analysis task, where a model or classifier is constructed to predict categorical labels, such as "safe" or "risky" for the loan application data, "yes" or "no" for the marketing data. The categories can be represented by discrete values, where the ordering among values has no meaning.

Classification is the task of learning a target function f that maps each attribute set x to one predefined class labels y. The target function is also known informally as a "classification model". A classification model is used for both descriptive and predictive modelling [1].

C) Credit Scoring

Credit scoring is related to ranking the customers based on their past bank transactions. This helps to identify the bankrupt customers and to recruit the profitable customers.

Credit scoring is a process of analyzing the behavior of past customers to differentiate the bankrupt and non-bankrupt customers. This is a mathematical model which quantitatively measures the credit of the customer by applying some data mining techniques [9].

D) Domain-Driven Approach

The domain-driven data mining [2] generally targets actionable knowledge discovery in complex domain problems. It first aims to utilize and mine many aspects of intelligence such as,

a) Domain expertise
b) Environment
c) Real time human involvement

Therefore, domain-driven approach improves efficiency of the system compare to the data driven mining.

E) Fuzzy Set

Fuzzy set framework provides a natural way of dealing with problems in which the source of imprecision is the absence of sharply defined criteria of class membership rather than presence of random variables [3]. The set on the universe X that can accommodate "degree of membership" were termed as "fuzzy sets" by Zadeh.

F) Linear Combination

The attributes which have an impact on other attributes are grouped into a single linear combinational set. Based on the linear combinational set the classification of customers is done considering the impact of each attributes.

III. RELATED WORKS

In the past, some of the traditional credit analysis methods used are decision tree analysis [4], neural networks [5], support vector machine [6] and other. These are all data driven approach which results only in imprecise solution. Achieving accurate credit scoring is a challenge due to lack of domain knowledge of banking industry. They use linear and logistic regression methods for credit scoring. The regression methods cannot provide mathematically meaningful scores and may generate results beyond user's control.

The classification of credit cardholder behavior based on fuzzy linear programming [7] was developed which results in a fuzzy satisfying solution. Also, the classification approach based on multiple criteria and multiple constraint level programming methods [8] resulted in a solution which does not meet the domain requirement.

The Fig:-1 describes the fuzzy linear programming model. This approach uses the attributes irrespective of their impact and identifies the classifier. The chosen classifier is evaluated using the multiple criteria models namely MSD-Minimizing the sum of the deviation and MMD- Maximizing the minimal distance.

A domain-driven based multiple criteria and multiple constraint level programming [9] is devised which classifies the customers into classes based on the user interaction. The approach resulted in a satisfactory solution and the method employed is computationally complex.

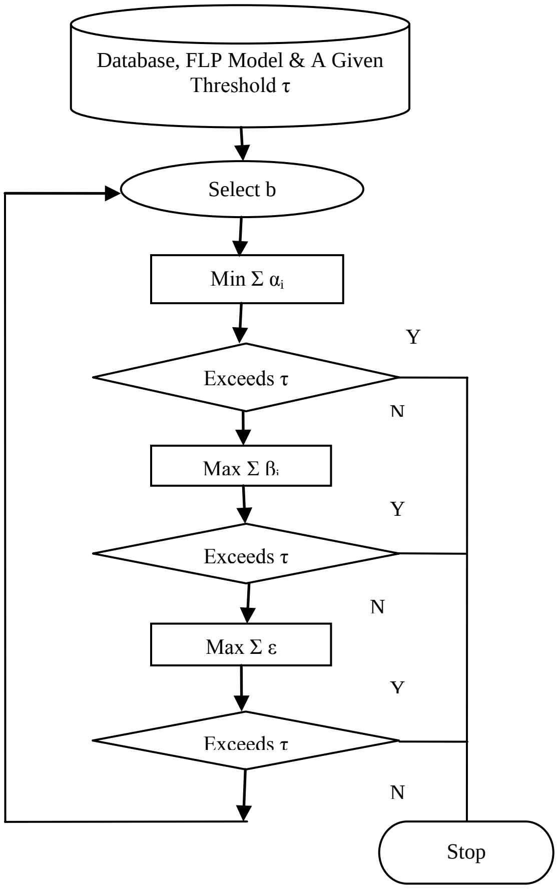

Fig:-1 Flowchart for fuzzy linear programming model

Fig: - 2 depict the classification of data based on MC2-domain driven approach. The attributes are selected based on domain knowledge and is verified against multiple conditions.

The best classifier is identified and the customer is positioned in particular class. Later the two classes (i.e.

restricted and unrestricted) are overlapped to obtain the satisfying solution.

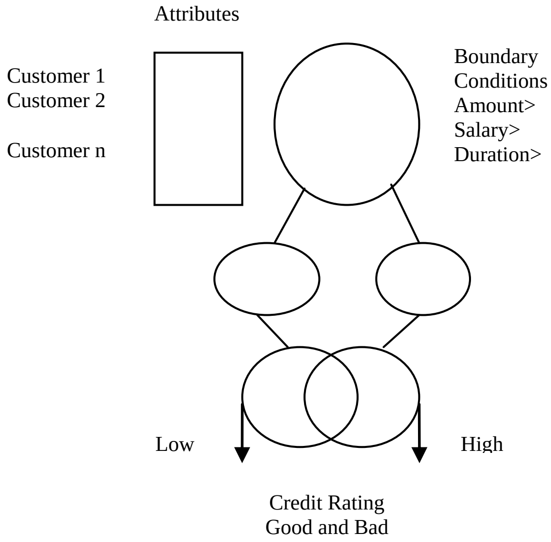

Fig:-2 Domain-Driven MC2 based approach

Fig: - 3 shows the fuzzy based MC2 approach which classifies the customers based on each attribute by identifying the membership. Finally the maximum membership among the individual memberships is determined. This helps in calculating score for customers. This approach results in best solution but which not concentrates on the domain knowledge.

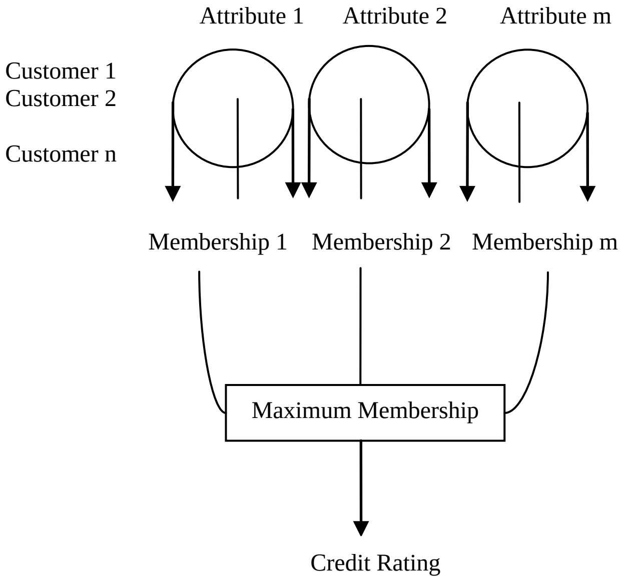

Fig:-3 Domain-Driven Fuzzy MC2 based approach

IV. PROBLEM DEFINITION

Given a set of r variables (attributes) about a cardholder a= (a_1, a_2…a_r), let A_i= (A_{i1}, Ai2 …A_{ir}) be the development sample of data for the variables, where i-1, 2 … n and n is the sample size. The best coefficient of the variable and the boundary value is determined for splitting and scoring customer data sets into best, good, ok, bad and worst classes.

V. PROPOSED SYSTEM

A hybrid approach of fuzzy set and multiple criteria and multiple constraints (MC2) with linear combination is used for credit scoring. With the use of linear combinational set of attributes the computational complexity will get reduced and the efficiency of the system increases. The MC2 is applied for identifying the best classifier for each predefined class involved in credit scoring.

a) System Architecture

The proposed credit scoring system follows the three-tier system architecture (Fig: - 4) as follows:

The lower tier (Data layer) represents data warehouse which is formed by combining different data sets from different sources.

The middle tier (Business layer) represents the credit scoring system process. It includes the process tracking (finding out the linear combinational set of attributes for classification), classifying (classifies the data set based on linear combinational set of attributes with best classifier using fuzzy set approach), scoring (scores the customers based on their satisfied linear combinational set of attributes) and analyzing (reviews the score and form the class of best, good, satisfactory, bad and worst class of customers).

The top tier (Presentation layer) represents the output format. The proposed credit scoring system outputs the result in two forms such as report and graphical representation of classes of customers.

b) Fuzzy Set (linear combinational set) based Classification

A set of linear combinational attributes are formed and the membership for each attribute (variable) is determined. The linear combinational sets are evaluated to determine the membership based on the fuzzy set approach [3].

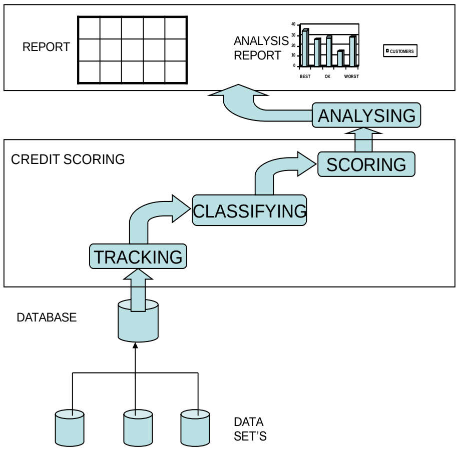

Fig:-4 System Architecture

Notion convention for fuzzy set when the universe of discourse X, is discrete and finite is,

$$U = \mu(x1)/x1 + \mu(x2)/x2 + \ldots + = \qquad (1)$$

Where, U is a fuzzy set

μ(x) is a degree of membership of element x in fuzzy set U. Also μ(x) ε [0, 1]

By applying some of the properties [3] for the membership value obtained from (1) associated with the fuzzy set, the membership for the best classifier of customers based on the linear combinational set is determined.

c) Scoring

The scoring is done based on the classifier obtained by each customer. To identify the best classifier based on the membership obtained some of the multiple criteria and multiple constraint level programming is used. Either by minimizing the sum of the deviation or by maximizing the minimal distance, the best classifier is obtained based on which the customer is scored.

d) Dataset

The dataset collected for evaluating the system is the “German credit data” collected from the UCI Machine Learning repository. The collected data is analyzed and pre-processed for error free and consistent data. The dataset contains 1000 instances i.e., records with 22 attributes, among which 13 categorical attributes, 7 numerical attributes, 1 class label and the attribute Cust_ID is the primary key.

VI. ALGORITHM FOR PROPOSED SYSTEM

Credit_scoring_alg

Input:

Processed dataset with different attributes- $A_{i(i =1 \text{ to } n)}$
Interval of cutoff b^l and b^u for predefined classes- best, good, satisfactory, bad and worst

Output:

Class separation with best classifier
Credit score for each customer

Step 1: Study different set of attributes in the dataset (i.e. A_i)
Step 2: Determine different groups of linear combinational set of attributes and assign score for each
LC = {G1, G2 … Gm}
Where, G1 = {linear combinational set of attributes, A_i}
Step 3: Identify the membership of each item set in the group (i.e. customer falling in the group range) using fuzzy set approach
Each attribute membership (MA_i)
$MA_i=$
Group membership (MG_i)
$MG_i=$
Linear combinational sets membership (M_i)
$M_i = 1/m (\sum_i^m MGi)$
Step 4: Repeat step 3 for all groups of linear combinational set of attributes
Step 5: Repeat step 3 and 4 for all customers (i.e. records) in the dataset
Step 6: Store the membership values identified for group and linear combinations of each customer
Step 7: Check the membership value against the predefined classes range and determine where the customer falls
Step 7.1: If satisfies the class model, place the customer
Step 7.2: If exceeds threshold, either by maximizing the internal distance or minimizing the sum of deviation, determine the best classifier
Step 8: Repeat step 7 for all customers (i.e. records) in the dataset
Step 9: Sum up the score for each customer based on their satisfying linear combinational sets of attributes
Score for customer i- S_i
Step 10: Display the best class of separation and the score of each customer in a score card

VII. EXPERIMENTAL RESULT

To explain the proposed model clearly, we use the data slice of the German credit data which is publicly available in the UCI machine learning repository. For the implementation of the proposed system, the categorical and numerical attributes are converted to ranges based on the domain knowledge of the system. The range falls between 0 and 1 as the classification is based on fuzzy set approach.

The groups of linear combinational attributes are formed as follows:
G1= {Status_Checkacc, Credit_History}
G2= {Purpose, Debators, Age}
and so on.

The membership for each is calculated using fuzzy set formula. Some of the linear combinational sets membership obtained are shown in Table:-I.

Table:-I Membership of Linear combinational set

LinearComb_Group
0.65802829595375
0.79737344726193
0.80418714472665
0.97630884161093
0.57610604247496
0.63359837358291
0.71782127431628
0.73269270464507
0.71550921076476
0.68541465509164

As the proposed system focuses on the impact of each attribute in a linear combinational format for classification the efficiency and the accuracy of the system is improved. Compared to the other traditional and domain driven approaches existing for credit scoring, the proposed system is effective because of the linear combinational set and domain knowledge. The comparison chart is depicted in Fig: - 5 efficiency and accuracy for German Credit Data.

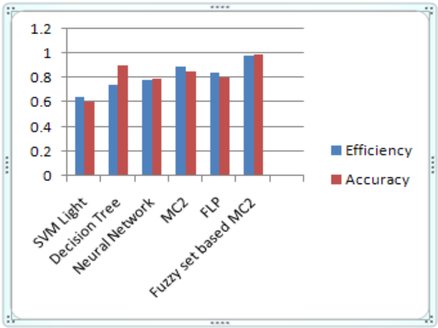

Fig: - 5 Efficiency and Accuracy for German Credit Data

VIII. CONCLUSION

In this paper, a new hybrid approach by using fuzzy set and MC2 is proposed for detecting the bankrupt customers and non-bankrupt customers based on their past transaction details. Also the score is calculated for each customer based on the best classifiers associated with the customers. The proposed method classifies customers into five distinct classes namely best, good, satisfactory, bad and worst. A major advantage of the proposed system over the existing systems is that, this is a domain-driven approach which built the hybrid model based on fuzzy set and MC2 functions. This ensures that one can achieve a best solution that meets the requirements of banking industry.

REFERENCES

[1] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, 2003.

[2] L. Cao and C. Zhang, “Domain-Driven Actionable Knowledge Discovery in the Real World”, IEEE Intelligent Systems, 2007.

[3] Timothy J. Ross, “Fuzzy Logic with Engineering Applications”, Second edition, Wiley-Indian edition.

[4] Ron Rhymon, “A SE-Tree Based Characterization of the Induction Problem”, Proceedings Machine Learning Conference, 1993.

[5] L. Yu, S. Whang and K. Lai, “Credit Risk Assessment with a Multistage Neural Network Ensemble learning approach”, Expert Systems with Applications, Vol 34, no 2, 2008.

[6] Y. Wang, S. Wang and K.K. Lai, “A New fuzzy support vector machine to Evaluate Credit Risk”, IEEE Transactions on Fuzzy Systems, Vol 13, no 6, 2005.

[7] J. He, X. Liu, Y. Shi, W. Xu and N. Yan, “Classification of Credit cardholder Behavior by using fuzzy linear programming”, International Journal of Information technology and decision making, Vol 3, no 4, 2004.

[8] J. Zhang, Y. Shi and P. Zhang, “Several Multi-Criteria programming methods for Classification”, Computers and Operations Research, Vol 36, no 3, 2009.

[9] J. He, Y. Zhang, Y. Shi and G. Huang, “Domain-Driven Classification based on Multiple Criteria and Multiple Constraibt-level programming for Intelligent Credit Scoring”, IEEE Transactions on Knowledge and Data Engineering, Vol 22, no 6, 2010.

[10] P.M. Murphy, D.W. Aha, “UCI Repository of Machine Learning Databases”, ww.ics.uci.edu /melearn/MLRepository.html.