Naivigyan Journal of Multidisciplinary Research **(NJMR),** 2025; 1(1): 32-38
Original Article | eISSN (O):
Published January 31, 2025

Implementation of Random Forest and Fuzzy-Based Data Mining Classification Model for the Banking Domain

Soni P M[1]* and Jayakrishnan S[2]

[1]Assistant Professor, Naipunnya Institute of Management and Information Technology, Pongam, Thrissur- 680308, Kerala, India

[2]Associate Professor, Naipunnya Institute of Management and Information Technology, Pongam, Thrissur- 680308, Kerala, India

*Corresponding Author's Email: sonipm@naipunnya.ac.in

Abstract

In many nations, loans particularly mortgage loans are essential to the banking sector. In many cases, personal loans are regarded as riskier than corporate loans. These two factors make the accuracy of loan repayment credibility forecast crucial for the well-being of households and the financial system, which in turn affects society as a whole. By analysing past data and forecasting client behaviour, data mining can resolve the issue. A variety of tools are used to make predictions. The most common modelling method for predicting a customer's ability to repay a loan is classification. Different algorithms can be used for classification, and the algorithms' accuracy varies as well. The main objective of this paper is to compare the results of classification using a random forest algorithm and fuzzy-based modelling.

Keywords: Feature selection, Classification, Accuracy, Performance

Introduction

Data mining technique involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set [Vivek Bhambri ,2011]. Today, customer relationship management in the banking industry is based purely on data mining techniques. The different areas in which Data mining Tools can be used in the banking industry are customer segmentation, Banking profitability, credit scoring and approval, Predicting payment from Customers, Marketing, Detecting fraud transactions and Cash management and Forecasting operations [Sudhamathy G , 2016]. Bankers should be vigilant of fraudsters because they will create more problems for the banking organisation, especially in finance. Banks hold huge volumes of customer behavior-related data from which they are unable to arrive at a judgment if an applicant can be a defaulter or not [Sudhamathy G , 2016]. Data mining techniques such as classification modelling and fuzzy-based modelling have an important role in solving such problems. One technique for data analysis that forecasts class labels is classification [Berson, et al , 1999]. Additional categorization techniques include rule-based, neural network-based, decision tree-based, statistical-based, and distance-based approaches [K. Chitra and B.Subashini, 2013]. Lot_Zadeh proposed fuzzy logic in 1965 as a way to deal with uncertainty.

Since Zadeh introduced fuzzy logic in 1965, it has been successfully used in numerous disciplines [Chen and Chiou, 1999]. Numerous fields, including operations research, artificial intelligence, medicine, and decision theory, have applications based on fuzzy logic. A considerable degree of human knowledge is used in fuzzy systems modelling to create fuzzy rules. Fuzzy theory represents information from zero (0) to one (1) and deals with imprecision and vagueness. By using a linguistic phrase, it can effectively describe vague knowledge of human subjective judgment. Additionally, fuzzy logic has been used to forecast a customer's ability to repay a debt. Given the unpredictability of human behaviour, it was thought to be crucial to examine the clients' behavioural patterns in relation to their capacity to repay the loan, assuming it was approved [Joseph Kobina and James Ben,2014].

This paper is organized as follows. The next section explains the dataset used for conducting the experiment. Section 3 discusses the methodology of the experiment. It includes classification using random forest and fuzzy logic Section 4 demonstrates the results and discussion. The conclusion is given in section 5 followed by references.

Methods and Materials

Information was gathered from a leading cooperative bank that offers loans to individuals, businesses, and other entities in order to satisfy the needs of all kinds of clients. The necessary information was obtained by interviewing the banking officer and observing the site. A thorough analysis of financial transactions and loan processing was also conducted for the same. The data set includes 15,000 mortgage loan customer details.

Sl	Attribute	Datatype
1	Loan No.	object
2	Loan Date	datetime64[ns]
3	Due date	datetime64[ns]
4	Loan amount	int64
5	Opening	int65
6	Payment	int66
7	Receipt	int67
8	int_rcvd	float64
9	fine_rcvd	float65
10	Mem No	object
11	action	object
12	secured	object
13	Loan Balance	int64
14	interest Rate	float64
15	Category	object
16	Purpose	object
17	gender	object
18	Occupation	object

Table 1- Credit Dataset

The data collected for the mining process may have missing numbers, noise, or inconsistencies. Consequently, conflicting information is produced by the mining process. Effective data mining results will come from a process that leverages high-quality data. To improve the effectiveness of the data mining process, the quality of the data, and eventually the mining outcomes, the acquired data must be pre-processed. Data preprocessing, which deals with preparing the initial data set and turning it into the final data set, is one of the most crucial phases in the data mining process. To transform the initial data set into the final data set, data preprocessing methods such as data cleaning, data reduction, data transformation, data integration are used. After data pre-processing, data mining techniques such as classification or fuzzy based modelling can be applied to the dataset for obtaining a pattern as knowledge. This knowledge can be used to assist bank officers in making proper decisions.

Classification is a data mining technique that assists in assigning a new data record to one of the target classes within a given data set [Berson et.al, 19993] stated that Classification aims to map a data item into one of several predefined categorical classes. For conducting this experiment, a random forest algorithm and a fuzzy-based classification algorithm were considered. One effective tree-learning method in machine learning is the Random Forest algorithm. During the training stage, it generates a number of Decision Trees. To measure a random subset of characteristics in each partition, a random subset of the data set is used to build each tree. By introducing variety across individual trees, this randomness lowers the possibility of overfitting and enhances prediction performance overall. The algorithm combines the output of every tree in prediction, either by average (for regression tasks) or voting (for classification tasks). An example of reliable and accurate findings is provided by this cooperative decision-making process, which is aided by the insights of several trees. For classification and regression tasks, random forests are frequently utilized because of their reputation for handling complex data, lowering. Table 2 briefly explains the process of the random forest classification algorithm

Step 1:	Select randomly **"k"** features from total **"m"** features.
Step 2:	Where **k << m**
Step 3:	Among the **"k"** features, calculate the node **"d"** using the best split point
Step 4:	Split the node into **daughter nodes** using the **best split**
Step 5:	Repeat **1 to 3** steps until the "l" number of nodes has been reached.
Step 6:	Build forest by repeating steps **1–4** for "n" number times to create **"n" number of trees**
Step 7:	End

Table 2 - Random Forest Algorithm

The act of transforming the input numerical values into membership functions is known as "fuzzification." Three language terms, such as low/medium/high, bad/average/good, or short/medium/long, can be used to describe the membership functions. Subjectively increasing the number of language terms—for example, by five or seven—will result in excellent accuracy. A membership function is a function that indicates the extent to which a certain input is a member of a set. Table 3 briefly explains the steps

Step 1 :	Read input values from the user and put into the database.
Step 2 :	Apply Fuzzification on the database to get the fuzzified input
Step 3 :	Using fuzzified input and inference rules, process the inference system for loan credibility prediction
Step 4 :	Apply Defuzzification to convert fuzzified output to user understandable output

Table 3 - Steps for Fuzzy based classification

The most important features and their corresponding feature importance that have been selected for the fuzzy classification algorithm are represented in Table 4.

Feature	Feature importance
int_rcvd	0.1418
opening	0.1363
fine_rcvd	0.1241
interest rate	0.1129

Table 4 - Most important Features

A membership function's output value is always between 0 and 1. Another name for the output value is a membership value or membership grade. Figure 2 portraits the membership function for the feature “fine_rcvd”. Figure 3 represents the code in python to generate fuzzy rules.

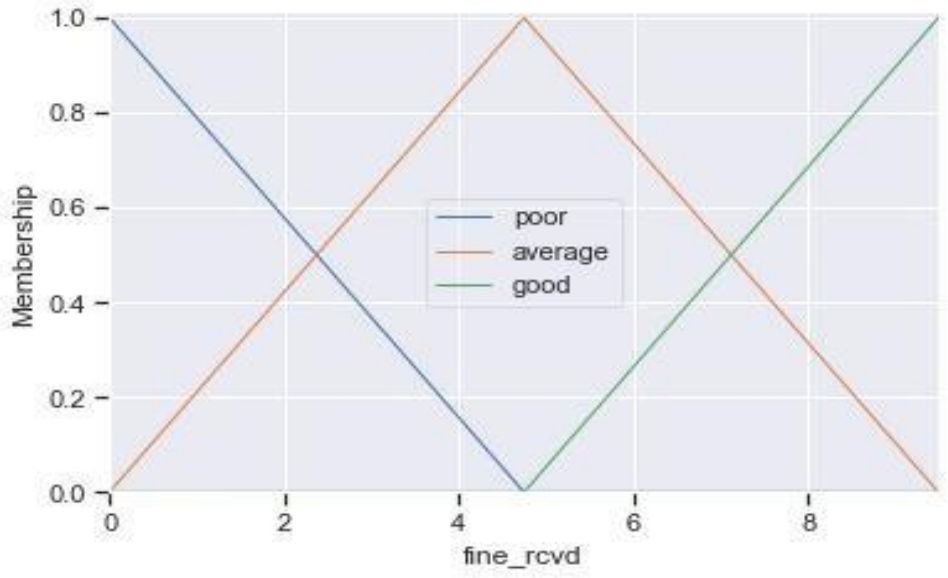

Figure 2 - Membership function for "fine _rcvd"

```
rule1 = ctrl.Rule (interest_rate ['good'] | fine_
rcvd ['good']| int_rcvd ['good']| days ['good'],
secured ['NO'])
rule2 = ctrl.Rule (interest_rate ['poor'] | fine_
rcvd ['poor']| int_rcvd ['poor']| days ['poor'],
secured ['YES'])
rule3 = ctrl.Rule (interest_rate ['average'] | fin
e_rcvd ['average']| int_rcvd ['average']| days ['a
verage'], secured ['YES'])
loan_ctrl = ctrl.ControlSystem ([rule1, rule2])
loan = ctrl.ControlSystemSimulation (loan_ctrl)
```

Figure 3 - Fuzzy Rules

Results and Discussions

The experiment is to analyse the accuracy obtained in two different classification models such as random forest algorithm and fuzzy-based modelling. The accuracy obtained by both classifiers is depicted in Table 5 and a graphical representation of the result is displayed in Figure 4.

Classifiers	Accuracy (%)
Random Forest	99.1
Fuzzy	84.6

Table 5 - Accuracy

Figure 4 depicts the input values given to the fuzzy classifier for evaluating the loan credibility behaviour of a customer.

interest_rate =16

interest_rcvd = 16

fine_rcvd = 12

days =1210

Figure 4 - Fuzzy Input

After giving the input values to the fuzzy classifier, it displayed the output as in Figure 5 which represents the probability of approving or rejecting a loan.

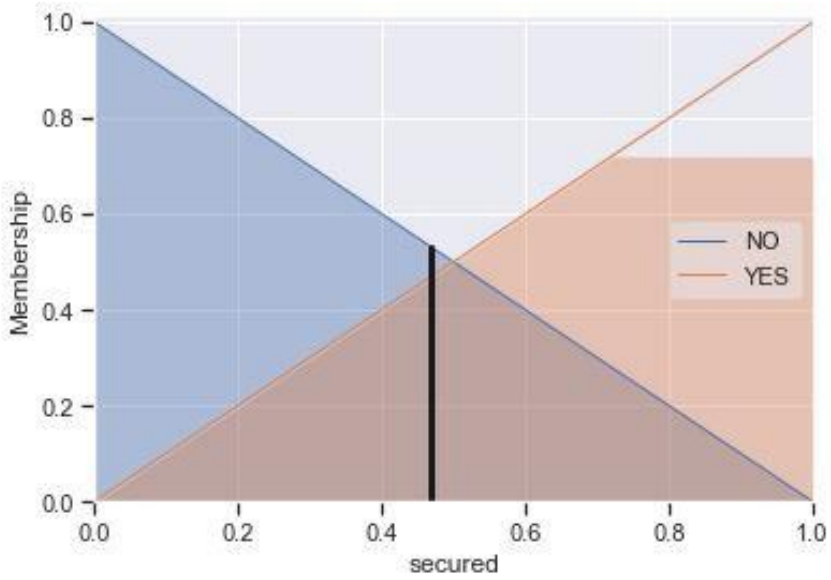

Figure 5- Probability of approving or rejecting the loan

Conclusion

In this paper, a comparative analysis between random forest classification and fuzzy-based approach is done to predict the loan repayment capability of a banking customer specially in the case of mortgage loans. It is very difficult for bank officers to determine whether to approve loan applicants or not. Now the experiment has proved that both techniques can help bank officers to make decisions more accurately.

References

Berson, A., Smith, S., and Thearling, K. (1999). Building Data Mining Applications for CRM. McGraw-Hill, New York

Bhambri, Vivek (2011). "Application of Data Mining in Banking Sector", *IJCSt* Vol. 2, Issue 2, June 2011, ISSN : 2229-4333(Print) | ISSN : 0976-8491(Online)

Chen, L-H., Chiou, T-W, (1999), "A fuzzy credit rating approach for commercial loans A Taiwan case". *Omega International Journal of Management Science* 27(1999), 407-419, 1999, https://doi.org/10.1016/S0305-0483(98)00051-6

Chitra, K. and B.Subashini (2013). "An Efficient Algorithm for Detecting Credit Card Frauds", *Proceedings of State Level Seminar on Emerging Trends in Banking Industry*, March 2013.

Joseph Kobina Panford , James Ben Hayfron-Acquah (2014). " Fuzzy Logic Approach to Credit Scoring for Micro Finance in Ghana: A Case Study of KWIQPLUS Money Lending" *International Journal of Computer Applications*, May 2014

Omaia Al-Omari1, Nazlia Omari (2019). "Enhanced Document Classification Using Noun Verb(NV) Terms Extraction Approach", *International Journal of Advanced Trends in Computer Science and Engineering* (IJATCSE), Vol.8, No.1, January –February 2019 https://doi.org/10.30534/ijatcse/2019/26822019

Sri Hari Nallamala, Dr. Pragnyaban Mishra, Dr. Suvarna VaniKoneru (2019). "Qualitative Metrics on Breast Cancer DiagnosiswithNeuro Fuzzy Inference Systems", *International Journal of Advanced Trends in Computer Science and Engineering* (IJATCSE), Vol. 8, No.2, March-April 2019.

Sudhamathy G. (2026). "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R" Vol 8 No 5 Oct- Nov 2016 https://doi.org/10.21817/ijet/2016/v8i5/160805414

Yi-Chung Hu a, Ruey-Shun Chen a, Gwo-Hshiung Tzeng (2003), "Finding fuzzy classification rules using data mining techniques", *Pattern Recognition Letters* 24 (2003), 509 –519 https://doi.org/10.1016/S0167-8655(02)00273-8