# Providing Banking Loan to Customers Based on J48 Classifier Algorithm Combined with Neural Networks

Megha, Prof. Neena Madan
*Department of Computer Science & Engineering, Guru Nanak Dev University, Jalandhar, Punjab, India*

***Abstract*: Data mining is the transformation of large sizes of information into expressive patterns and rules. It is approximately describing the former and calculating the near future by means of information analysis. Data mining is really a multi-disciplinary area which combines, equipment understanding, data, database engineering and synthetic intelligence. The full total aim of the info exploration method is normally to extract information from the info set and convert it into a good understandable framework for more use. To predict bank failure events, apply the random forests method to the analysis of bank-level financial data in order to identify hidden patterns that can distinguish active and inactive banks. This paper mainly focuses on on to classify the customers as fake or fraud and non-fake or fraud. . By using J48 and Neural Networks to improve the accuracy rate further for detection of fraudulent customers.**

***Keywords* — Data mining; Random Forest Classifier; Neural networks; J48.**

## I. INTRODUCTION

Data Mining is probably the most pushing plus important region of research with the essence removing information and facts from considerable amount involving gathered information sets..The evolutionary path is witnessed inside the data source sector around the growth of the examples below benefits : information variety plus data source development, information supervision (including information storage space plus retrieval, plus data source contract processing), plus information research plus realizing (involving information warehousing plus information mining). Basically said, information exploration refers to removing or maybe \mining" find out corner from large amounts involving data. We're obtaining a several information, from uncomplicated mathematical dimensions plus textual content docs, for you to more complicated information and facts like spatial information, multi-media channels, plus hypertext documents. Information Exploration, likewise commonly known as Information Discovery around Sources (KDD), refers back to the nontrivial extraction involving implicit, previously unfamiliar plus possibly beneficial information and facts from information around databases. When information exploration and knowledge discovery around listings (or KDD) are likely to be treated while word, information exploration is actually portion of the awareness discovery practice [2].

The Knowledge Discovery in database method is definitely comprised of some steps leading via raw files collections to a certain amount of brand-new knowledge. The iterative process consists of the following steps:

a. Data cleaning: also referred to as facts washing, this can be a step wherein disturbance facts along with trivial facts are eliminated on the collection.
b. Data integration: at this time, numerous data resources, generally heterogeneous, might be combined within a frequent source.
c. Data variety: at this, the information highly relevant to the evaluation is selected and restored from the information collection.
d. Data transformation: also referred to as knowledge consolidation, it is a phase in your selected knowledge is changed into types befitting the mining procedure.
e. Data mining: it does not get necessary portion in that may brilliant techniques are applied to get designs perhaps useful.
f. Pattern evaluation: in this step, strictly helpful styles comprising understanding usually are identified based on given measures.
g. Knowledge representation: is actually the last period where the found understanding is actually successfully symbolized on the user. The following critical action employs visualization strategies to guide users recognize in addition to read the data mining results [2].

### *1.1 ALGORITHMS(TECHNIQUES) USED IN DATA MINING DURING BANKING LOAN:*

#### *1.1.1 J48*

J48 classifier is definitely a simple C4.5 conclusion pine intended for distinction, which produces a binary tree. It truly is most successful conclusion pine method for distinction problems. It constructs a pine to be able to design the particular distinction process. Following the pine is built, the particular formula is definitely put on to every single tuple from the data source to cause distinction for that tuple.

Algorithm J48 :

INPUT:
P//Training data

```
OUTPUT
DT //Decision tree
DTBUILD (*P)

{
DT=φ;
DT= Create root node and label with splitting attribute;
DT= Add arc to root node for each split predicate and label;
For each arc do
P= Database created by applying splitting
predicate to P;
If stopping point reached for this path, then
DT'= create leaf node and label with
appropriate class;
Else
DT'= DTBUILD(P);
DT= add DT' to arc;
}
```

While creating a decision tree, J48 omits this lost values i.e. the worthiness for the item can be estimated based upon just what exactly may be known with regards to the attribute values with regard to the other records. The key thought can be to separate the details in vary depending on the attribute values for the item that are determined throughout it test.

*1.1.2 Random Forest Classifier*

Random Forests are broadly considered the best possible "off-the-shelf" classifiers regarding high-dimensional data. Randomly jungles are a mixture of pine predictors in ways that every single pine will depend on within the values on the random vector sampled autonomously current exact same circulation for all those timber within the forest. A generalization blunder regarding jungles converges in order to a establish limit while the quantity of timber within the woods turns into large. A generalization blunder on the woods involving pine classifiers will depend on within the muscle of the people timber within the woods and also the association among them. A new part involving it data are selected, together with replacement unit, to learn every single tree. Outstanding teaching data are used to estimate blunder in addition to variable importance. Class work will be of the quantity of ballots via each of the timber for regression the standard involving the effects will be used. It is similar to bagged decision trees with hardly some key differences as given below:

1. For each split point, the search is not over all p variables but just over mtry variables (where e.g. mtry = [p/3])
2. No pruning necessary. Trees can be grown until each node contains just very few observations (1 or 5).

*1.1.3 Neural Networks*

Neural networks are an emerging artificial learning ability engineering this copies this mental faculties on the computer.These procedures are usually good parallel, handed out processing design. Your parallel struscture helps make nerve organs communities great at analyzing challenges with many variables. Any nerve organs multilevel style is made of a number of neurons that are structured in many tiers: an enter coating,a new concealed layer(s), as well as an output layer.The enter coating with neurons for this enter parameters on the network. Your concealed coating is really a connect between the enter coating and the output layer. Your neurons on this coating are usually mainly concealed from view, along with their quantity and understanding can easily typically end up being taken care of like a dark colored pack in order to those people who are undertaking this system.

The function of the hidden layer would be to process your suggestions variables. This is successfully done by means of summing upward just about all calculated advices, examining if thez amount of money matches your patience value plus utilizing the modification function. The actual loads regarding the issnput neuron plus secret neurons figure out as soon as every unit from the secret layer may possibly fireplace this is by adjusting these types of loads, your secret layer may possibly fireplace this is .In simple terms, your secret tiers discover the relationship between advices plus results in ways similar to those of your human brain by means of adapting your loads throughout the training process. The actual purpose of your productivity layer is a lot like those of your secret layer. Every suggestions due to this layer can be pressed like for example your secret layer. A specific nerve organs multilevel design relies on it has the topology, understanding paradigm plus understanding algorithm. Planning a nerve organs multilevel productively uses a very clear comprehending with the issue, and on deciding on after the majority of powerful suggestions variables. The operation involving creating a nerve organs multilevel design can be may process.

## II . LITERATURE SURVEY

Graham R. Wilkinson, Mick Schofield, Peter Kanowski,(march 2014) [3] described a genesis plus progress in the Tasmanian natrual enviroment techniques procedure, plus summarises all the different methods used to instill great degrees of complying, having a focus on training plus instruction, self-monitoring plus confirming by simply this is a, separate tracking by the Woods Routines Recognition, plus helpful behavior, supported by simply enforcement provisions. Consent tracking over 27 several years proves quick advancement from the decades next institution in the procedure, having regularly great degrees of success subsequently. Even so, larger corporate and business natrual enviroment executives regularly attain greater fees associated with complying than do small-scale natrual enviroment proprietors, plus redressing this specific imbalance has become a 2010 continual design throughout Tasmania's natrual enviroment techniques system.shyang Chen, Ching-Hsue Cheng, (feb 2013) [8], proposed standard bank insolvency or perhaps personal bankruptcy, specially of large banks, can easily greatly put in danger global financial stability. For that reason, providers and people immediately have to have a credit ranking warning to help

identify the particular fiscal position and functional proficiency with banks.The credit ranking supplies fiscal organisations using an review with credit merit, financial commitment chance, and normal probability. Even though several models have also been consist of to fix credit ranking troubles, they have got the examples below drawbacks: (1) lack of explanatory strength; (2) addiction to the particular hard to stick to suppositions with precise methods; and (3) several factors, that lead to several size and complex data. To overpower these shortcomings, this work is applicable 2 multiple versions that will remedy the particular useful troubles around credit ranking class. Monica Fisher, Moushumi Chaudhury, Brent McCusker, (sept 2010) [10] introduced facts coming from outlying Malawi are used to assess the part involving jungles around outlying home version so that you can local climate variability, and also to look at the actual implications to get version so that you can long run local climate change. Even though jungles will not at this time be a factor around anticipatory version by outlying people, they generally do look of importance to reactive coping: providing meals while in shortages, plus a way to obtain funds to get coping with weather-related scalp failure. Look for people a lot of just a few jungles include lower revenue for each human being, can be found all around woods, and are also went by individuals who are more mature, far more chance averse, and fewer qualified compared to his or her cohorts. P. Ravi Kumar, V. Ravi, (july 2007) [12] presents an all-inclusive look at the project accomplished, over the 1968–2005, with the effective use of record and also brilliant techniques in order to resolve your bankruptcy prediction challenge encountered by means of finance institutions and also firms. The actual review is usually labeled through taking the type of technique given to fix this challenge since an important dimension. Appropriately, your reports are sorted with the examples below families of techniques: (i) record techniques, (ii) sensory cpa networks, (iii) case-based thinking, (iv) selection bushes, (iv) business study, (v) evolutionary methods, (vi) tough collection dependent techniques, (vii) other techniques subsuming unclear logic, support vector unit and also isotonic divorce and also (viii) comfortable computing subsuming easy hybridization of all above-mentioned techniques. Involving certain value is the fact with every newspaper, your review illustrates the origin of knowledge models, fiscal proportions employed, country connected with foundation, time collection of study and also the comparative operation connected with techniques with regard to prediction accuracy wherever out there.David L. Martell,(2001) [17] surveyed flame managing plans commonly consist of elimination methods to cut back the volume of people-caused that will fire which come about, prognosis systems to find that will fire when they're smaller, initial assault systems so that you can consist of that will fire just before that they burn up above huge areas, and large flame managing systems which will reduce the damage which comes from huge that will fire aren't governed simply by the 1st assault system. Additionally consist of gas modification methods so that you can mitigate a impact with that will fire which do occur plus using recommended flame to meet silviculture, animals an environment managing, as well as other territory managing objectives. Fire managing in addition involves cognizant decisions permitting several wildfires to burn readily or maybe to become subjected to restricted suppression action assuming the world wide web help of doing this can be thought to be positive. Charles E. Williams, Eric V. Mosbacher, William J. Moriarity,(feb 2000) [19], One of many key factors which affects this reliability associated with woodland ecosystems throughout portions of this eastern Mixed Says will be heavy exploring by way of overabundant populations associated with white-tailed deer (Odocoileus virginianus Zimmerman). Deer effects in order to upland reforested land on the Allegheny Plateau throughout northwestern PA, USA, have been in particular intense: higher degrees of deer exploring within the last few 60 decades have firmly affected woodland structure in addition to process. Riparian reforested land on the district tend to be full of herbaceous place types, although essentially practically nothing is well known on how deer exploring may influence the dwelling in addition to purpose of these kind of systems.

## III. RESULTS

### A. *PROPOSED METHODOLOGY*

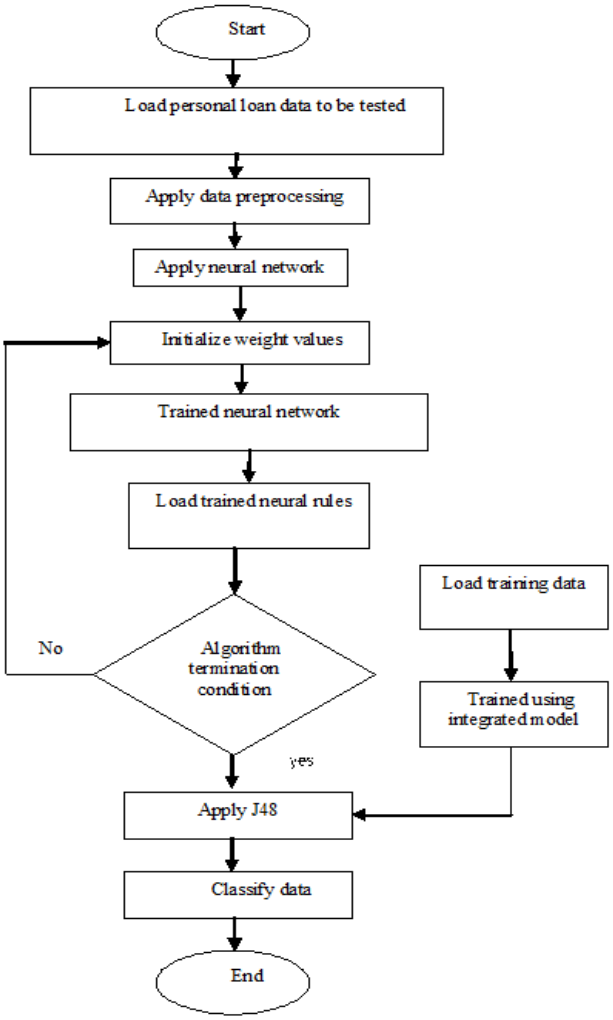


Fig 1: Flowchart of the proposed technique

*B. PROPOSED ALGORITHM*

1. Load credit data set
2. Divide data into train and testing forms
3. Apply unsuprwised filter
4. Train data using ann
5. Develop neural weithts
6. Update neural weights
7. Train data using neural network based j48 technique
8. Classify test instances
9. Stop

*C. PERFORMANCE ANALYSIS*

This paper has designed and implemented the proposed technique in MATLAB tool u2013a. The evaluation of proposed technique is done on the basis of following metrics i.e. overhead; energy consumption and time to live depends on the speed of vehicles. A comparison is drawn between our proposed technique and the existing work carried by Zhiwei Y. & Sherali Z. [4].

**TP Rate**

TPR refers to True Positive Rate. It is also called Sensitivity or Recall in some fields. TPR is defined as measurement of positive cases that are correctly identified. It is prediction of correctly identified instances.TPR can be expressed by using formula:

$$\text{TP Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

**FP Rate** FPR is called False Positive Rate. It is defined as ration those instances or objects that are incorrectly identified as positi It is also known as fall-out. FPR can be expressed by using formula:

$$\text{FP Rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

**Accuracy**

Accuracy refers to the ability of the model to correctly predict the class label of new or unseen data.It is calculated as-

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

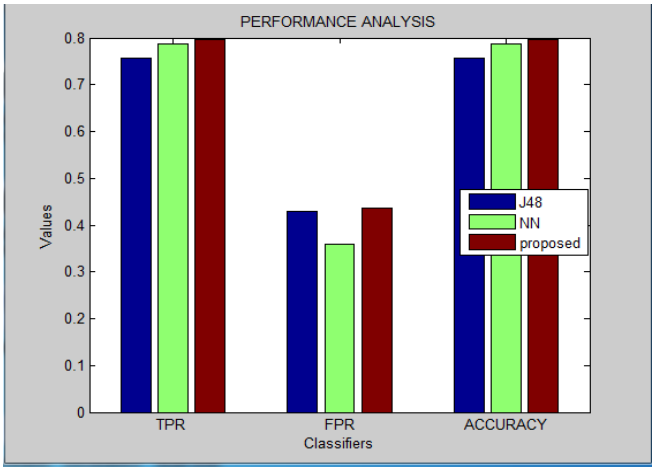


Fig.2 Output of different algorithms based on TPR,FPR,ACCURACY

**Precision**

Precision is defined as measurement of all positive cases that are identified when making calculations.. Precision is also known as positive predictive value. Higher Precision signifies that an algorithm significantly returned more relevant results when compared to irrelevant. Precision can be calculated by using the formula:

$$\text{Precision} = \frac{\text{TP}}{\text{FP} + \text{TP}}$$

**Recall**

Recall is the division of the written documents that are applicable to the question which have been winningly recovered. It is also called TP Rate or Sensitivity. It is defined as collection of positive cases.

Recall can be expressed as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

**F-Measure**

F-Measure is also called F1 score. It contains both precision and recall. It is generally use to check the accuracy and reliability. It computes the mean of precision and recall. Basically, it uses as best and 0 as worst when both precision and recall are used.

F-measure can be calculated with using the formula given as:

$$\text{F} - \text{Measure} = 2 * \frac{\text{P} * \text{R}}{\text{P} + \text{R}}$$

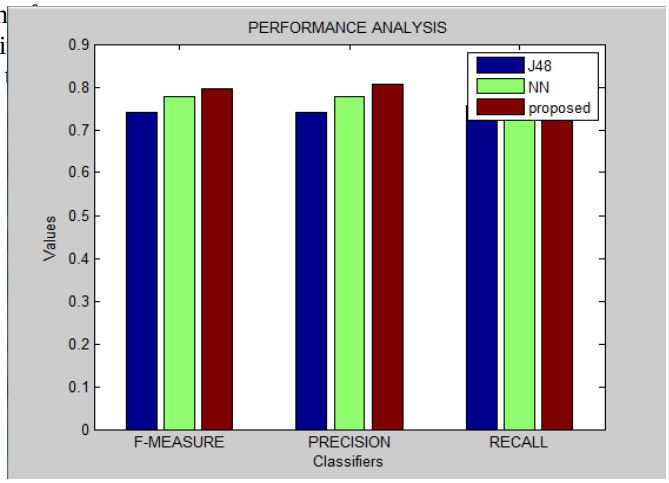


Fig.3 Output of different algorithms based on F-MEASURE,PRECISION,RECALL

## IV. CONCLUSION

In this paper, we have analyzed existing 'random forests-based early warning system for bank failures'. the proposed data mining based on j48 algorithm further by merging it with neural networks gives better results. this paper has shown comparison between exiting and proposed j48,neural network based algorithms are used for detection of fraudulent bankers on the basis of parameters like true positive rate,false positive rate,accuracy,f-measure ,precision and recall.This proposed technique detection of

fraudulent bankers shows better results as compared to the existing technique.

## REFERENCES

[1]. Joel Janek Dabrowski, Conrad Beyers, Johan Pieter de Villiers, Systemic banking crisis early warning systems using dynamic Bayesian networks, Expert Systems with Applications, Volume 62, 15 November 2016, Pages 225-242, ISSN 0957-4174.

[2]. Katsuyuki Tanaka, Takuji Kinkyo, Shigeyuki Hamori, Random forests-based early warning system for bank failures, Economics Letters, Volume 148, November 2016, Pages 118-121, ISSN 0165-1765.

[3]. Graham R. Wilkinson, Mick Schofield, Peter Kanowski, Regulating forestry — Experience with compliance and enforcement over the 25 years of Tasmania's forest practices system, Forest Policy and Economics, Volume 40, March 2014, Pages 1-11, ISSN 1389-9341.

[4]. Rosalind L. Bennett, Haluk Unal, The effects of resolution methods and industry stress on the loss on assets from bank failures, Journal of Financial Stability, Volume 15, December 2014, Pages 18-31, ISSN 1572-3089.

[5]. Raymond A.K. Cox, Grace W.-Y. Wang, Predicting the US bank failure: A discriminant analysis, Economic Analysis and Policy, Volume 44, Issue 2, July 2014, Pages 202-211, ISSN 0313-5926.

[6]. Kimie Harada, Takatoshi Ito, Shuhei Takahashi, Is the Distance to Default a good measure in predicting bank failures? A case study of Japanese major banks, Japan and the World Economy, Volume 27, August 2013, Pages 70-82, ISSN 0922-142.

[7]. Robert DeYoung, Gökhan Torna, Nontraditional banking activities and bank failures during the financial crisis, Journal of Financial Intermediation, Volume 22, Issue 3, July 2013, Pages 397-421, ISSN 1042-9573.

[8]. Shyang Chen, Ching-Hsue Cheng, Hybrid models based on rough set classifiers for setting credit rating decision rules in the global banking industry, Knowledge-Based Systems, Volume 39, February 2013, Pages 224-239, ISSN 0950-7051.

[9]. Andreas Krause, Simone Giansante, Interbank lending and the spread of bank failures: A network model of systemic risk, Journal of Economic Behavior & Organization, Volume 83, Issue 3, August 2012, Pages 583-608, ISSN 0167-2681.

[10]. Monica Fisher, Moushumi Chaudhury, Brent McCusker, Do Forests Help Rural Households Adapt to Climate Variability? Evidence from Southern Malawi, World Development, Volume 38, Issue 9, September 2010, Pages 1241-1250, ISSN 0305-750X.

[11]. Melek Acar Boyacioglu, Yakup Kara, Ömer Kaan Baykan, Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey, Expert Systems with Applications, Volume 36, Issue 2, Part 2, March 2009, Pages 3355-3366, ISSN 0957-4174.

[12]. P. Ravi Kumar, V. Ravi, Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review, European Journal of Operational Research, Volume 180, Issue 1, 1 July 2007, Pages 1-28, ISSN 0377-2217.

[13]. Timothy J. Curry, Peter J. Elmer, Gary S. Fissel, Equity market data, bank failures and market efficiency, Journal of Economics and Business, Volume 59, Issue 6, November–December 2007, Pages 536-559, ISSN 0148-6195.

[14]. Marcia Millon Cornett, Jamie John McNutt, Hassan Tehranian, Long-term performance of rival banks around bank failures, Journal of Economics and Business, Volume 57, Issue 5, September–October 2005, Pages 411-432, ISSN 0148-6195.

[15]. Carlos A Molina, Predicting bank failures using a hazard model: the Venezuelan banking crisis, Emerging Markets Review, Volume 3, Issue 1, 1 March 2002, Pages 31-50, ISSN 1566-0141.

[16]. David L. Martell, Chapter 15 - Forest Fire Management, In Forest Fires, edited by Edward A. Johnson and Kiyoko Miyanishi, Academic Press, San Diego, 2001, Pages 527-583, ISBN 9780123866608.

[17]. Aigbe Akhigbe, Jeff Madura, Why do contagion effects vary among bank failures?, Journal of Banking & Finance, Volume 25, Issue 4, April 2001, Pages 657-680, ISSN 0378-4266.

[18]. Charles E. Williams, Eric V. Mosbacher, William J. Moriarity, Use of turtlehead (Chelone glabra L.) and other herbaceous plants to assess intensity of white-tailed deer browsing on Allegheny Plateau riparian forests, USA, Biological Conservation, Volume 92, Issue 2, February 2000, Pages 207-215, ISSN 0006-3207.