THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Data Mining Approach For Predicting Best Saving Scheme For Investment

Barjinder Singh*
Assistant Professor
Lovely Professional University, Phagwara, Punjab

Ravinder Singh
Assistant Professor
Lovely Professional University, Phagwara, Punjab

Ashish Dhall
Research Scholar
Lovely Professional University, Phagwara, Punjab

Amandeep Kaur
Research Scholar
Lovely Professional University, Phagwara, Punjab

Abstract- Investment is generally related to the allocation of funds in expectation of a positive rate of return in the future. Investment is very important because it helps to grow our money. But, the major in India is lack of awareness. So many investment options like bank deposits, mutual funds, equity, shares, gold, post office investments etc. are available in market yet people don't know the best investment scheme. This work will find traditional algorithm's efficiency and limitations and find the best approach to invest money.

***Index Terms*— Data Mining, Decision Tree, Investment, C4.5.**

I. INTRODUCTION TO DATA MINING

Data mining is technique which is helpful for several purposes in private and public sectors. Data mining was primarily used in the public sector for minor applications such as the discovery of scams and leftovers, but now a day's Data Mining is useful in many areas like reducing the cost, increasing revenue, improving performance of an organization and many more. As per the report, Data Mining has helped the US Government to recover million of dollars in fraudulent Medicare payments. Another Example of Data mining is Aviation safety the aviation department using the data mining techniques to regularize flight inspection standards and to enforce security to the planes. It is used to detect the reasons that are responsible for plane crash to recommend precautionary measures.[1]

Data mining is useful to discover new patterns from the present data by using different algorithms that have been

already purposed [2]. Data mining is helpful to analyze the representation of data which is used to discover new and better findings. Data mining always perform functionality keeping in view database management. A large dataset is analyzed using various data mining tools such as text mining tools, traditional data mining, and dashboards to find the relationship among data. The job of data mining is not just to collect and manage the records; it focuses mainly on new discoveries. Data mining is performed on different type of data for evaluation such as text data, images, forms, multimedia data etc.[3]

II. DATA MINING PROCESS

Data Mining is a repetitive process. The steps involved in data mining process are as under.

- Problem Definition
- Data Preparation
- Exploring Data
- Constructing models
- Exploring and Validating models

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

- Deploying Models

The brief explanation of these steps is as under:

Problem Definition: This step aims to study scope of problem. The predicted outcomes are evaluated. Basically in this step the actual problem is located and present relationships are build.

Data Prepration: Data prepration deals with two aspects

i) data identification and collection: it deals with identifying required data and collecting the scattered data from different places to one place.

ii) Cleaning the data: some time the data collected is not valid or have some ambiguity. That data is cleaned up whether by

removing the record or by modifying it.

Exploring Data: This step deals with studying the collected data in details and finding out some required calculations that are helpful in developing new studies or evaluations. Various statistical techniques such as standard deviation, variance, and regression analysis are applied to data to obtain useful results.

Constructing models: By using the information that has been analysed in the 3rd step new models have been developed. Models are basically an empty structure that just gives details of inputs, expected outputs and the process of applying algorithm to the data.

Exploring and Validating Models: In fifth step of data mining process the various purposed data mining models are tested. The outputs of all models are noticed. They are compared with the required output and the best model is selected.

Deploying Models: Actual implementation of the selected model takes place in this step. The selected model is available for various task such as

- To make some prediction that can be helpful in further decisions.
- Generate a query which satisfies the rules described in the model.
- Include data mining functionality into an application.
- Creating reports to make it easy for the users to directly query the existing model.
- Making updations automatically with the arrival of new data into the system.[4]

III. BACKGROUND

In this phase all the research study which is being done before start writing this research paper is mentioned in details.

Dr.Binod Kumar Singh [5] Mutual funds are those funds that are collected from many investors to invest in securities such as stocks, bonds and money markets. They are operated by money managers. One major benefit of mutual funds is that it is suitable for those investors that do not have awareness of market trends but have risk tolerance power. Mutual fund schemes can be classified as short term and long term funds. Selection between MFS and MFL depends upon various attributes such as investment purpose like (income, growth, tax saving) and loan facility requirement because loan facility is generally not available in many MFS schemes.

S. Saravana Kumar [6] Share Market termed as SM is another good option of investment that gives huge returns if the investment is done meticulously. This investment scheme is good for those investors that have good knowledge of share market and have high risk tolerance power. Because, this investment in shares gives no guarantee that you will actually realize any sort of positive return.

A.N. Agarwal , Rajesh Agrwal [7] A fixed deposit is a financial instrument provided mainly by banks. This is beneficial for savings because it gives saving and investment option for short and long period. Fixed deposits can be short term fixed deposits termed in this study as FDS and long term fixed deposits termed as FDL. Tenure of FDS varies from 7, 15 or 45 days to 1.5 year. FDL have tenure more than 1.5

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

year to 10 years. Some FDL's having tenure of 5 years and above also gives tax rebates to investors.

Public Provident Fund termed as PPF is another ideal investment scheme. PPF account is opened for the tenure of 15 years. It gives tax saving benefits to the investors. Another benefit of PPF is safe deposits and loan facility available against your deposits. employees are getting benefit of S P F(Statutory Provident fund) contributed by employer which is exempted from tax and even interest on SPF balance is also exempted. After this recognized institution are contributing for their employees in RPF(Recognized Provident Fund) contribution by employer up to 12% of basic salary is exempted and interest on RPf Balance is also excepted up to 9.5 % of such balance. It is also eligible for deduction section 80 C.

Yen-Liang Chen, Hsiao-Wei Hu, Kwei Tang [8] So far, we have studied that the classification is based upon Boolean class labels. The Boolean or categorical class labels are not suitable in many real life problems. In his paper the researcher has purposed a new way of tree classification using hierarchical

class labels. This newly purposed algorithm has been named as HCL(Hierarchical class Label classifier). The main focus of the researcher is accuracy in the results. The researcher has considered a training set of 16 hypothetical customer records. The purpose of the study is to find the interest of a customer towards the purchase of a particular brand of a computer. Some considered attributes in the study are: Gender of the customer, Customer's Career, Customer's Income, Preferred product. Among these attributes Preferred product is a dependent variable and all other are independent variables. The training data is further sub-divided based upon an attribute. The selection of that attribute is made on the basis of Gain ratio and Hentropy which calculates the maximum contributing factor. And further that attribute serves as a base for division of training data. The main drawback of this study is that, if there exists a gap between labels in the tree, then the accuracy level is not achieved.

Qiang Yang[9] Data mining is very useful in many areas. These technology can also be applied on customer relationship management, which is helpful to figure out those customers who are unfavorable towards your products and those who are well-wisher or favorable towards your product. After getting this knowledge manually some post processing techniques are enable which shows that how many customers who are favorable or unfavorable. Pre-processing technique show you result in virtualized way. But they don't suggest any thing which help us to increase profit .In his study Qiang Yang presents new algorithm that suggest some action which converts the opinion of customer from undesired to desired one. Which increase the profit .which is objective function of in his paper Quiang Yang decision tree as data mining technique.

Surjeet Kumar Yadav[10] Day by day the quantity of data stored in institutional database increasing rapidly. There are many classification technique can be applied on educational dataset to forecast the performance of the student in examination. These technique are Bayesian network, Decision tree, neural network etc. In his study 3 classification technique ID3, C4.5 and CART were used to predict the performance of student and also apply the 10 fold cross validation on the dataset to obtain the final result. Result comes in two forms in the form of accuracy and in the form of execution time to build the model. C4.5 is produced the maximum accurate result(67.777%) and execution time is 0.03 sec

Mohammed M Mazid, A B M Shawkat Ali, Kevin S Tickle [11] has focused to reduced the input space of data file, reducing computation time and increasing classification accuracy of C4.5 algorithm. Undoubtedly, C4.5 algorithm is a

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

powerful tool for classification. It is widely used because of its features as missing value handling and continuous number categorization. But C4.5 has some limitations also. The tree constructed by C4.5 some time has empty branches. These branches do not contribute to classification rules but make the decision tree more complex. Secondly, selection of discrete attributes creates equal number of branches in decision tree. This reduces the unsuitability of decision tree. Outfitting is the third drawback of c4.5. It can be removed by tree pruning technique, but this not a complete solution.

Entropy is a calculation of uncertainty of data. In his paper, researcher has used entropy to find out the central attribute of dataset it is calculated by the formula.

$$H(x)= \sum_{j=1}^{k} p_j \log_2 \frac{1}{p_j}$$

The attribute for which H(x) is minimum has been selected as central attribute which is used for further processing. Using this attribute correlation coefficient is calculated using Karl Pearson, Spearman and Kendall correlation methods. By applying these methods, relationship among attributes is calculated and the attribute that are less correlated are removed from the dataset. Mohammed M Mazid has used Matlab for calculation of entropy and correlation coefficients and then C4.5 algorithm is applied to this dataset using WEKA tool.

The results are shown by box plot using Matlab. In the result, it is concluded that median line for C4.5 is 76% and it is 81%, 79% and 80% for Karl Pearson, Spearman and Kendall correlation methods respectively. The processing time for improved C4.5 is also reduced in comparison of actual C4.5 algorithm.

In his study, the researcher had concluded that improved C4.5 algorithm does not perform better than actual C4.5 algorithm in every case. The experiment has been done on 8 datasets and it is concluded that Pearson's C4.5 performs better than Spearman C4.5 and Kendall C4.5. more experiments on different datasets can be done to make the comparison of improved and actual C4.5 algorithms

Clara Pizzuti and Domenico Talia [11] Each processor must perform their own task and reduces execution time. It can reduce processing time along with scalable parallel clustering for handle large data set. There are many strategies for parallelize data set of p auto class. Independent parallel search: - It can use classification technique for parallelism. Every processing element performs different classification and share data set. During the computation classification become performs in parallel to evaluate accuracy. Task parallel approach:- It can balance the computational cost of every processor.

METHODOLOGY

Decision tree algorithm only helps to classify data. According to existing decision tree algorithms, if a person X having some attributes invest in plan P_1 then another person Y having the same attributes must invest in plan P_1. But, in real sense, it might be possible that X's decision was wrong. X invested in wrong plan. In our research, the data is firstly converted to an optimal dataset by applying Expert rules and then feature reduction is performed on this dataset. Finally decision tree has been constructed from this optimal dataset. Following are the steps used in this research work. [12]

• First of all, Primary information was obtained through questionnaires and interviews from hundreds of individuals. Then Training data set was generated by randomly selecting 10 percent of the rows from this data. After this, the rules were created by the help of experts in this particular domain. Furthermore, the rules were created from this information in the Java Language Code. These rules have been applied to the data set.

• Moreover, this data set is minimized by measuring the gain ratio of each attribute and deducting three attributes with a minimum gain ratio resulting in an ideal data set. Finally,

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Construction of the optimal decision tree using the C4.5 algorithm. Last but not the least, there was a comparison between our existing approach and our approach.

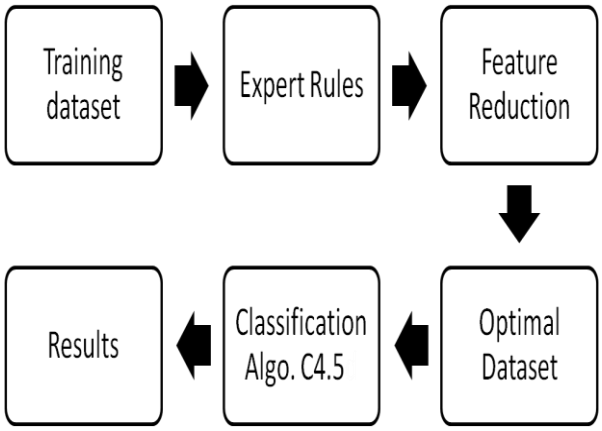


Figure 1: Showing approach used for prediction results

Applying Expert Rules on training dataset

Expert's knowledge has been gathered and from that knowledge some rules have been developed in java language and SQL/Server 2008. Netbeans IDE 6.0 platform is used to implement this java code. [13] Data collected for this study is in csv file format which is not directly supported by SQL/server. To apply expert rules on this dataset it is firstly converted into SQL/server required format by implementing java code. These rules are implemented to increase the accuracy of training dataset. [14]

The rules developed have been in the following format:

- If purpose = 'tax_saving' and awareness = 'No' and Risk_ tolerance = 'No' and Time_horizon = 'Long' and Income = 'High' and age = 'Young' then Investment_scheme = 'PPF'
- If purpose = 'Tax_saving' and awareness='No' and risk_tolerance = 'Yes' and time_horizon = 'Medium' and age = 'middle' then Investment_scheme = 'MFL'.

These rules have been applied to the training dataset and those tuples have been identified whose class label attribute i.e. investment_schemes is not relevant to the input parameters and then this class label attribute is modified according to expert's rule.

This step gives us an optimal dataset that will increase accuracy of classification rules.

Feature Reduction

Feature reduction is a technique of identifying and removing those attributes that do not contribute towards classification of the dataset. [15] In this work, gain ratio technique is used to evaluate the worth of an attribute with respect to the class. Then rankers algorithm is used to arrange these attributes in descending order according to their gain ratio and last three attributes having lowest gain ratio are removed. Ranked attributes after applying gain ratio and Rankers algorithm[16]:

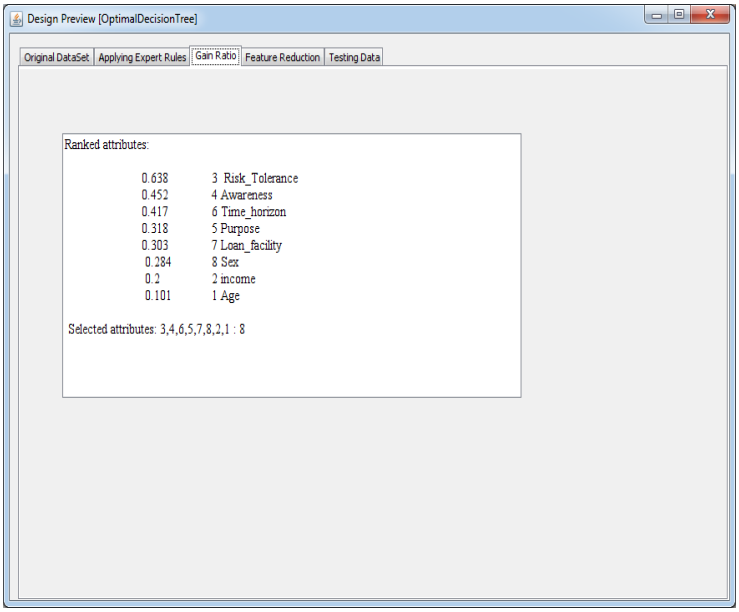


Figure 2: Representing Gain Ratio of all attributes

From the above figure, attributes Sex, Income and Age has minimum gain ratio. These three attributes will be removed from the dataset. It has been assumed that the factors with

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

minimum gain ratio do not contribute toward data classification i.e. results after implementing C4.5 algorithm will not get affected but, this will result in less computation time and less memory requirement.

Testing Data

In this phase supervised learning has been performed on optimal dataset. A Graphical User Interface has been designed for the user to receive input parameters. This received data is stored in arff format file named as testing_data.arff. Optimal dataset that is obtained from above steps, is converted into arff file format named as training_data.arff. A classsifier model is constructed on the basis of this file. Result of testing data is evaluated by this classifier model which suggests an appropriate target class for the investments.

RESULT AND DISCUSSION

In this study classifiers provided in WEKA software. The training dataset has been prepared by randomly selecting 38 instances from the actual dataset. Training dataset collected from investors is graphically represented in WEKA 3.4 using histograms. The view of training dataset in WEKA 3.4 is as shown below.

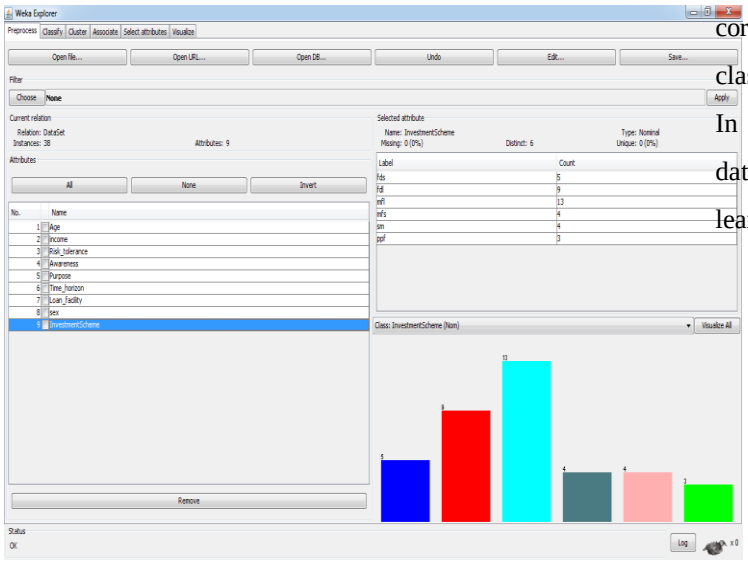


Figure 3: Representing training dataset in WEKA 3.4

It has been observed that most of the investors prefer to invest their money in mfl i.e. long term mutual funds and investment

in ppf accounts is the least choice of investors. But, this analysis may not be true because when investors do not have market awareness they prefer to follow others.

In the next step, this C4.5 algorithm has been applied to this training dataset. The accuracy of this training dataset is measured and the output is shown below.

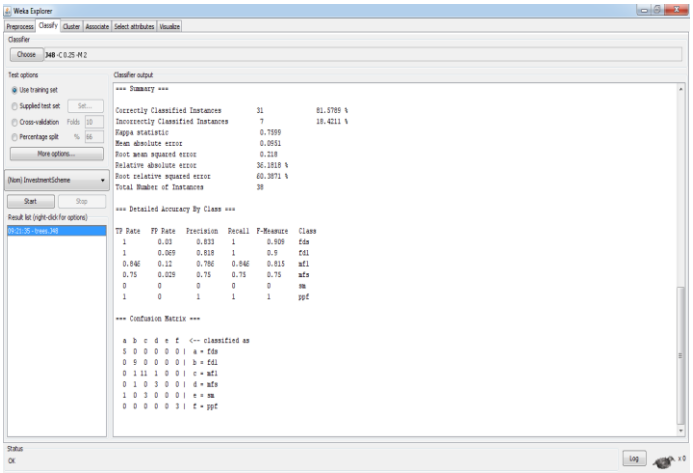


Figure 4: Representing accuracy of Existing Approach in WEKA tool

It has been observed that, this training dataset has 81.5789% correctly classified data and 18.4211% data is incorrectly classified.

In our work, training dataset is converted into an optimal dataset which is then fetched into WEKA 3.4 machine learning tool.

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

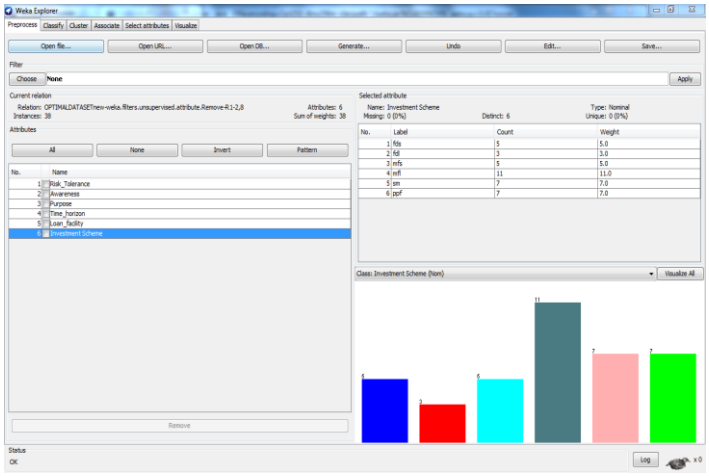


Figure 5: Representing view of an optimal dataset in WEKA

This figure shows the view of an optimal dataset. After applying expert rules, incorrect investment decisions of the investors are replaced by the correct decisions according to expert's opinions. Further, gain ratio of all the attributes is calculated and it has been observed that attributes age, income and sex have lowest gain ratio which are then removed from the dataset to make it an optimal dataset. This work is done on the assumption that, the attributes with lowest gain ratio does not contribute towards construction of decision trees.

In the above diagram it has been observed that, according to experts investment in long term mutual funds is the best option for many investors. PPF accounts come at second position. On the other hand, long term fixed deposits are least priorities of experts.

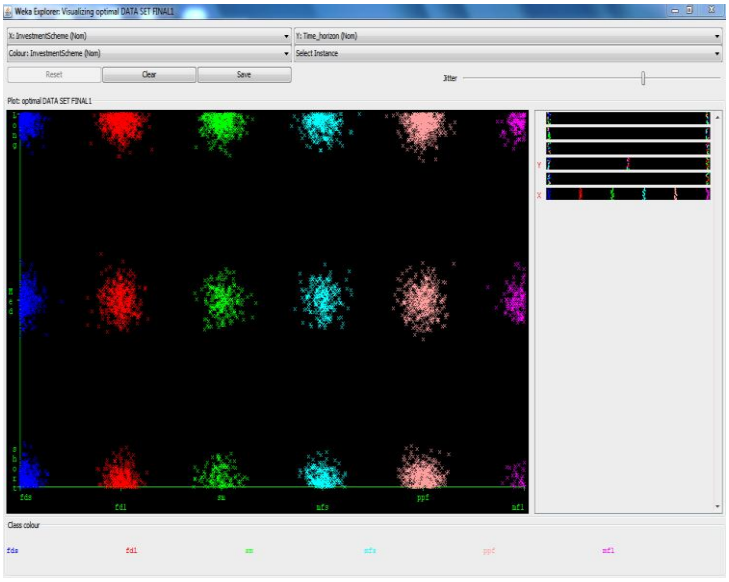


Figure 6: Shows the cluster size of Investment Schemes Fds, Fdl, SM, Mfs, PPF, Mfl on X axis and Time_horizon on Y axis

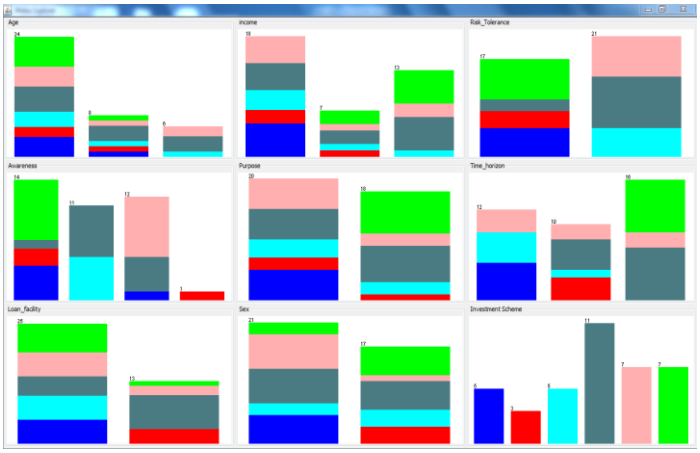


Figure 7: Representing all attributes showing data according to class attribute

This above discussed approach is implemented in Netbeans using java language.

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

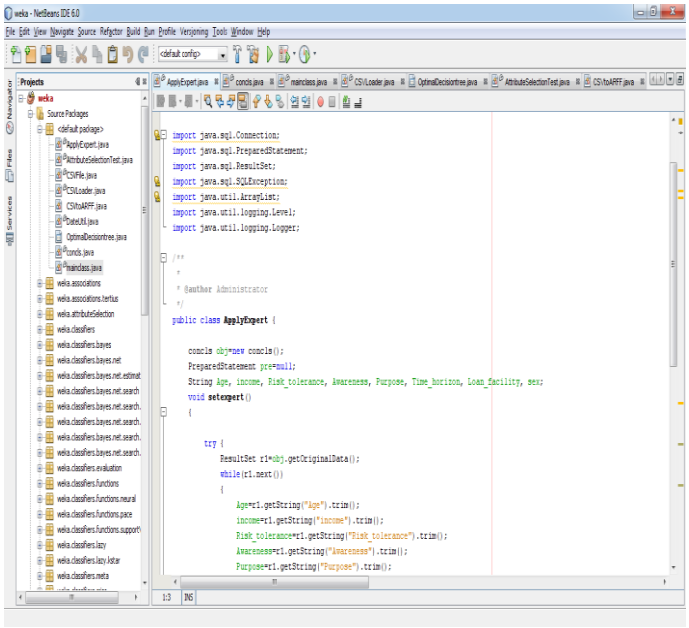


Figure 8: Representing view of coding implemented in java language

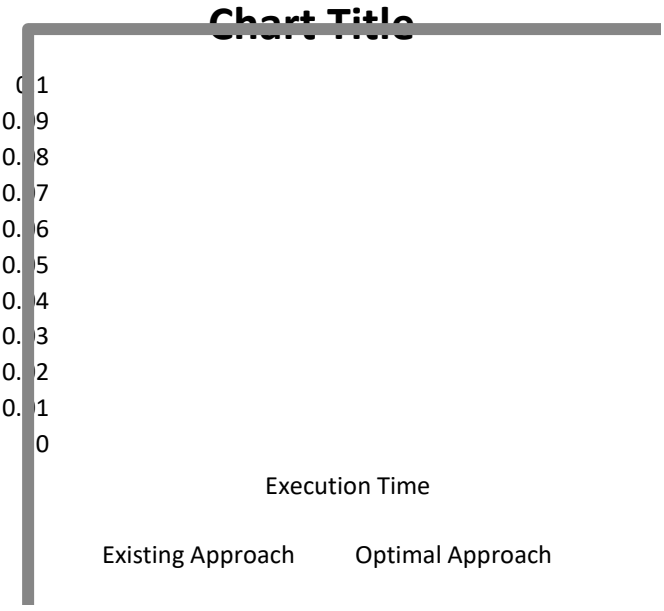


Figure 9: Representing time comparison of Existing and Optimal Approaches graphically

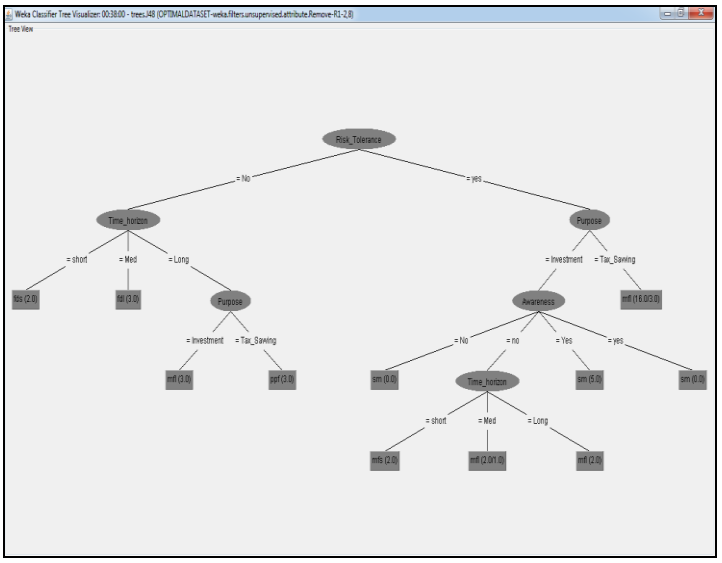


Figure 10 : Representing Decision tree constructed from optimal dataset

After converting dataset to an optimal dataset, C4.5 algorithm has been run on this dataset and fnally decision tree has been constructed. From this decision tree, classification rules are generated. These classification rules will help the investor to take right investment decision besed upon some input parameters.

.

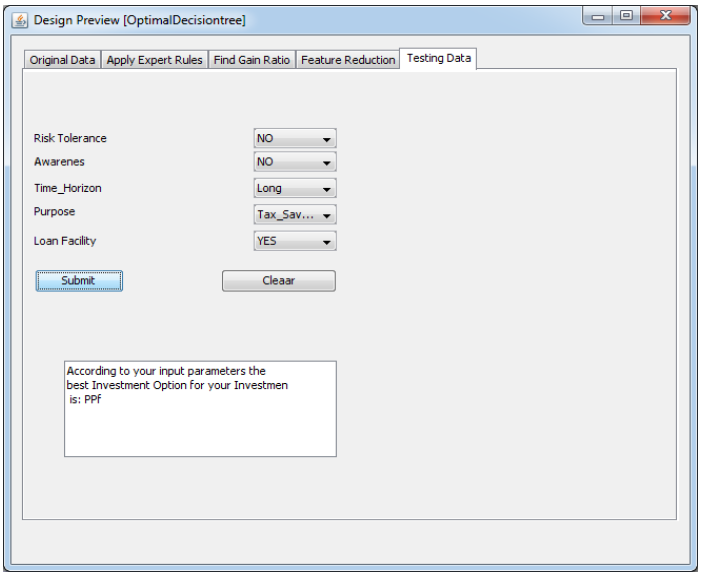


Figure 11: Representing graphical view of testing data.

Main objective of this research is to help the investor in finding best investment option for his savings. Testing data

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

feature give best decision to the investor based upon the inputs
received from the investor. To make handling of this option easier for the user, graphical view has been presented to the investor.

CONCLUSION:

Based on C4.5 methods, this optimal algorithm is improved. This introduces new functionality on a large dataset, such as applying expert guidelines and increasing features. There have been several attempts to design and improve the generic data mining method, but no truly functional framework has been found. Therefore, the assistant of the domain expert is mandatory for each domain. The domain experts will be directed by the program to use their expertise effectively to use data mining systems to produce the required knowledge.

Domain experts are required to determine the variety of information to be collected in the specific problem domain, to select specific data for data mining, to clean and transform data, to extract patterns for knowledge generation, and finally to interpret patterns and generate knowledge. We will find out the important features that will affect investment decisions by applying this optimized approach to Investor data.

Moreover, supervised learning has been implemented. A Graphical user interface has been provided to the user to input various parameters. Based upon these parameters an output showing best investment scheme for the investor is displayed on the screen. This approach is better to an existing approach. The accuracy of decision has been improved. Computation time and memory requirement has also reduced.

5.2 Future Scope:

This approach is capable to handle multivariate data which makes it suitable for many other applications. This research is not bounded to particular area. In this work, this optimal technique is applied to improve accuracy of investment decisions. But, can be used in other applications also likewise Analysis of education patterns, Human talent management, risk evaluation etc

REFERENCES

[1] Ji Dan, Qiu Jianlin, 2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010).

[2] Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, Surjeet Kumar Yadav (WCSIT) ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012

[3] Clara Pizzuti and Domenico Talia, P-AutoClass: Scalable Parallel Clustering for Mining Large Data Sets, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 15, NO. 3, MAY/JUNE 2003 Vol. 02, No. 08, 2010, 2526-2534

[4] Souptik Datta, Chris R. Giannella, and Hillol Kargupta, Senior Member, IEEE IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 10, OCTOBER 2009 Approximate Distributed K-Means Clustering over a Peer-to-Peer Network.

[5] Dr.Binod Kumar Singh "A study on investors' attitude towards mutual funds as an investment option", International Journal of Research in Management ISSN 2249-5908 Issue2, Vol. 2 (March-2012)

[6] S. Saravana Kumar in his article "An Analysis of Investor Preference Towards Equity and Derivatives" published in The Indian journal of commerce, July-September 2010

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

[7] Income Tax and Investment Journal – (AY-2008-09)–by A.N. Agarwal (Income tax expert), Rajesh Agrwal(CA), Sanjay Kulkari (CA), and Dr. Gajanan Patil.- ABC Publication- Nagpur.

[8] Yen-Liang Chen, Hsiao-Wei Hu, Kwei Tang, Constructing a decision tree from data with hierarchical class labels, Expert Systems with Applications 36 (2009) 4838–4847.

[9] Qiang Yang, Senior Member, IEEE, Jie Yin, Charles Ling, and Rong Pan, "Extracting Actionable Knowledge from Decision Trees" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 1, JANUARY 2007.

[10] Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, Surjeet Kumar Yadav *(WCSIT)* ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012

[11] Clara and Talia Domenico (2003) "P-AUTO CLASS: SCALABLE PARALLEL CLUSTERING FOR MINING LARGE DATA SET". Page no 629-641. ISSN 1041-4347.

[12] Souptik Dutta (2009)"Distributed K-Means Clustering over a Peer-to-Peer Network", IEEE transactions on information and data engineering,vol. 21, no.10

[13] Mohammad Ali Farajian and Shahriar Mohammadi(2010) "Mining the bank customer behavior using clustering and association rule method". International journal of industrial engineering and production research. Volume 21, no.4, PP 139_245.

[14] C. Science and M. Studies, "Literature Survey on Clustering Algorithms," vol. 7782, no. 2010, pp. 447–453, 2015.

[15] Mohammed M Mazid, A B M Shawkat Ali, Kevin S Tickle, 'Improved C4.5 Algorithm for Rule Based Classification', Recent Advances in Artificial Intelligence, Knowledge Engineering and Data Bases.

[16] Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database Volume 3, Issue 3, March 2013 ISSN: 2277 128X

Copyright © 2019Authors