RESEARCH ARTICLE OPEN ACCESS

A Proposal on Implementing Optimal Algorithm on Best Saving Services for Potential Investors by Using Data Mining Algorithms

Priya Babbar[1] , Barjinder Singh[2]
Department of Computer Science Engineering, Lovely Professional
University, Phagwara, (Punjab)

--************************----------------------------------

Abstract:

The main aspect of Data Mining is to divide large data sets into small datasets based upon some criterion. So far, many algorithms have been developed but they do not show uniformity in results. Some algorithms are reliable for some attributes and some for other attributes. The construction of classification trees separates these algorithms based upon their quality and scalability. This research will suggest many techniques that can be useful in improving the efficiency of the construction of classification trees (decision trees). The main problem in this approach is to identify the various resources that are required by an algorithm. To overcome this problem many mathematical models are used that studies the problem and finds the amount of resources that are required by an algorithm. These resources include time, memory storage, processor (it includes number and type of processors), buses, type of processing i.e. serial processing or parallel processing and many more. The objective of this study is to find the best investment scheme by using optimized algorithm. This algorithm tries to reduce the errors made by the predictive models.

Keywords — **Investment Schemes, Clustering, Classification, Prediction, C4.5 Algorithm.**

--************************----------------------------------

I. INTRODUCTION

With the help of Data Mining, potential investor can draw the idea which can help him to predict performance of various investment schemes. By searching these results and trends of various schemes Data Mining can help potential investors to find right investment scheme with good performance. Many Industries use Data Mining for extract the valuable information from the large database to minimize costs, enhance research, and increase sales i.e. banking, medicine, insurance, and retailing.Data Mining is extensively useful for full utilization of banking and other financial services as we know Banks and Post offices plays an extreme role in our society [1]. The services provided by these are highly secure and reliable so everybody wants to invest in their schemes. But it is not easy for the new investor to find the BEST SAVING SERVICES to increase the potential of wealth. Data Mining is useful tools to help the investor in finding best saving scheme. Data Mining is very effective in Business analysis also. It is used in market analysis, to identify the root cause of manufacturing problem to attract new customer, produce new and more effective product.

II. METHOD

The steps followed in this research will be as under:

- Primary data will be collected from 500 people through questionnaires and interviews.
- Training dataset will be created from this data by randomly selecting 38 rows.
- Knowledge elicitation from domain experts.
- From this knowledge, rules will be created in Java Language code.
- These rules will be applied to the data set.
- Feature reduction will be performed to this dataset by calculating gain ratio of every attribute and three attributes having minimum

gain ratio have been deducted which results in an optimal dataset.

- The optimal decision tree will be constructed by using c4.5 algorithm.

Finally, comparison has been made between existing approach and our approach.

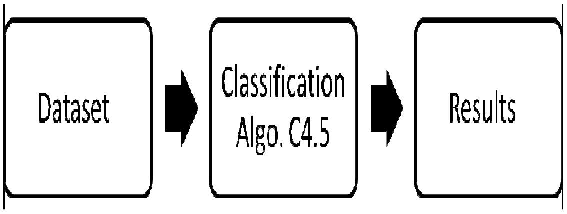

Figure 2.1: The diagram representing existing approach

Above figure represents the basic structure of existing approach. In this approach, training dataset is prepared and then classification algorithm C4.5 is directly applied to this training dataset which results in construction of decision trees and classification rules.

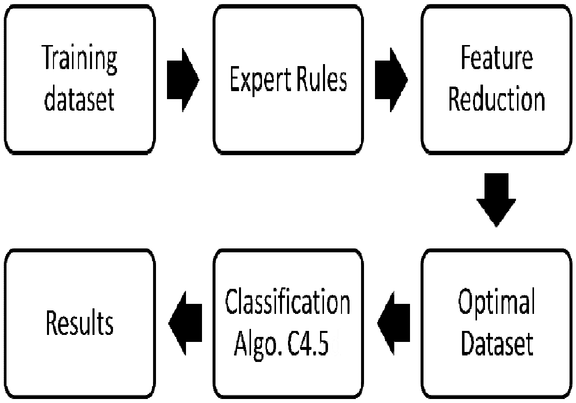

Figure 2.2: Representing our approach

But, in this study, expert rules are applied to the training dataset followed by feature reduction process that makes it an optimal dataset. This is done to improve accuracy and to decrease computation time of C4.5 algorithm.

The widely used technique in data mining is decision tree [3]. It is basically a representation of data in hierarchical shape. The top node is called root and last level nodes are called leaf nodes. The nodes between root node and leaf nodes are called internal nodes. The internal nodes in decision tree are represented by a rectangle and the oval is used to represent a leaf nodes. Decision tree is constructed on the principle of recursion in which root node (main attribute) is recursively divided into sub nodes (Sub attributes). The process is repeated until some class is not reached [4].

Let's suppose a record X having no class then simply insert the record at the root then using the classification rules the class is found. Construction of decision tree is basically splitting a record into sub-record based upon some attribute. This attribute selection is done using attribute selection measures such as information gain, gain ratio and gini index [5]. In information gain method, information gain of every attribute is calculated, the results are evaluated and the highest contributing independent factor is determined that effects the output of dependent variable. In the below considered example, admission of a child is dependent factor and age, father income, distance are the independent factors.

Expected information needed to classify a record is calculated by the formula:

$$\text{Info(D)} = -\sum_{i=1}^{n}(pi)\log2(pi)$$

The contribution of each independent attribute is measured towards the dependent variable (admission in the considered example). This is done by the formula:

$$\text{Info}_A\text{(D)}= \sum_{j=1}^{v}\frac{|Dj|}{|D|}\times Info(Dj)$$

Finally the information Gain is evaluated as:

$$\text{Gain(A)= Info(D)-Info}_A\text{(D)}$$

In the next step, the required information for every field is evaluated and gain factor for each is calculated. After calculating gain factors for every attribute the comparison among them is made and the attribute having maximum gain factor is selected as a splitting attribute i.e. by considering this attribute the data can be split into further sub tables.

III. ALGORITHMS USED

1. ***Association:*** it refers to the linking of one incident to another for example if people buy bread then these people most likely buy butter also. If people buy software then they are most prone to buy a CD or an Antivirus.
2. ***Sequence:*** Second argument used in data mining is sequence it refers to the order of events in which they occur for example by the birth of a baby the parents are most likely to buy new clothes and other baby products.
3. ***Classification:*** in this technique data is divided into different groups based upon predefined classes. For instance a vehicle producing company classifies their product into classes like high demand, mild demand, and low demand. A model is derived based upon some features like price, mileage, gender of customer, brand etc...
4. ***Clustering***: [6] The process of dividing data into different groups called clusters is called clustering. In clustering the classes are not predefined. The objects of one cluster are same to other objects in that cluster but different from objects of another cluster. The benefit of clustering is that they are easy adaptable to changes. Example of clustering includes, a salesman wants to cluster the same type of population into one group and another type of population into another group to formulate the efficient marketing scheme for the product sale [2].

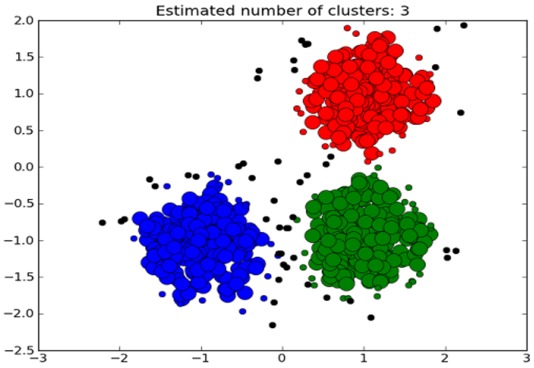

Fig. 1 Clustering technique

5. . ***C4.5 Algorithm:*** This algorithm is used to generate a decision tree. This decision tree is used for classification so it is referred to as statistical classifier.

A) *Attributes used:*

Income{low,medium,high},Age,Awareness{yes, no}, Loan Facility{yes, no}, Risk Tolerance, Gender{M,F}, Time_horizon, Investment_schemes has 6 distinct values {FDS, FDL, MFS, MFL, SM, PPF} short term mutual funds (MFS), long term mutual funds (MFL), short term fixed deposits (FDS), long term fixed deposits (FDL), share market (SM), and public provident fund (PPF) [8].

B) *Evaluation of performance:*

WEKA[7] software is used for Automatic Evaluation of Best Investment Options for Investors. Algorithm C4.5 is applied to an optimal dataset to construct an optimal decision tree. This tree will results in less error rate, consumes low memory and reduces computation time as compared to existing approach. From the investor's point of view, the resulted classification rules will be more realistic and gives more revenue to the investor. The comparison between existing approach and our approach is made by this algorithm on the basis of computation time, memory and accuracy.

IV. CONCLUSIONS

From the proposed methodology, unrevealed information could be extracted from large data set of personnel data that enhances the decision makers to have a better understanding and visualization of required knowledge. This knowledge can be useful to take right decision at right time.

ACKNOWLEDGMENT

All praise in the name of almighty God, who give us in the darkness and help in difficulties. The dissertation is the result of full semester of work whereby I have been accompanied and supported by many people. It is a pleasant aspect to that I have the opportunity to express my gratitude for all of them. I am also extremely indebted to my guide Mr. Barjinder Singh (Assistant professor, Department of Computer science, Lovely Professional University, Phagwara). I am very much thankful to Mr. Barjinder Singh for picking me as a student at the critical stage of my masters. I warmly thank him for his valuable advice, constructive criticism and his extensive discussions around my work.

REFERENCES

[1] Data Mining: Concepts and Techniques Second Edition Jiawei Han *University of Illinois at Urbana-Champaign* Micheline Kamber.

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

[3] Data Mining and Analysis Fundamental Concepts and Algorithms by Zaki & Meira (2014).

[4] C. Aggarwal, "Data Mining."

[5] **"**The Elements of Statistical Learning" by Freidman et al (2009).

[6] Ashish Dutt, Saeed Aghabozrgi, Maizatul Akmal Binti Ismail, and Hamidreza Mahroeian(2015) 'Clustering Algorithms Applied in Educational Data Mining' International Journal of Information and Electronics Engineering

[7] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/documentation.html. [Accessed: 26-Mar-2016].

[8] 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)