International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Analysis of Improved ID3 Algorithm using Havrda & Charvat Entropy

Kirandeep[1], Prof. Neena Madan[2]
Guru Nanak Dev University Regional Campus, Jalandhar In Partial Fulfillment For Degree Of Master's In Computer Science & Engineering, Jalandhar, Punjab, India

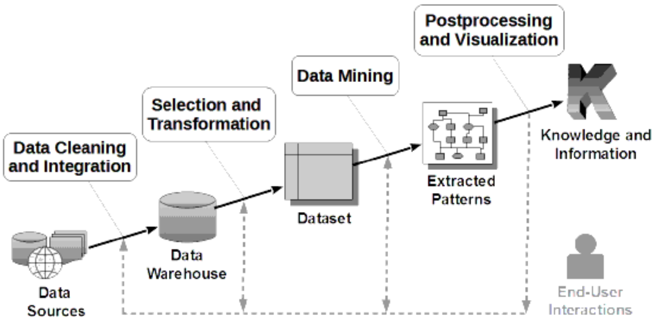
ABSTRACT

Data mining is often called as knowledge discovery procedure. We use data mining techniques in order to identify the patterns and relationships among them. Data mining consist of combinations of Machine learning, Visualization for identifying the patterns. Data mining consist of different techniques which can be used to complete a goal. From all the data we need to find the data which will give more information which will help to predict the performance. Data mining techniques allows us to discover hidden patterns, relationship form huge amount of data. Information extracted from large database is helpful in decision making. By referring the extracted information the processing methodologies are selected. The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data. Finding useful patterns in data is known by different names (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing)

Keywords : ID3,Improved ID3,Havrda & Charvat entropy

I. INTRODUCTION

- **Data cleaning**: It is also known as data cleansing; in this phase noise data and irrelevant data are removed from the collection.
- **Data integration**: In this stage, multiple data sources, often heterogeneous, are combined in a common source.
- **Data selection**: The data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation**: It is also known as data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining**: It is the crucial step in which clever techniques are applied to extract potentially useful patterns.
- **Pattern evaluation**: In this step, interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation**: It is the final phase in which the discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

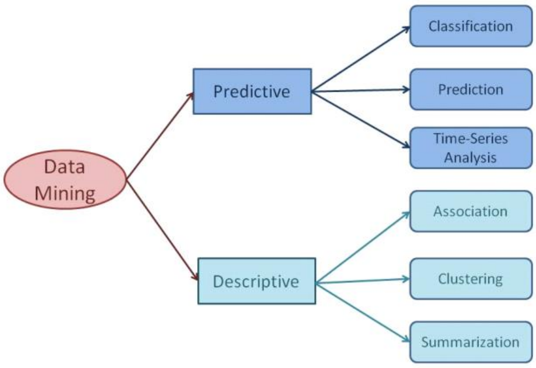

Postprocessing and Visualization
Data Mining
Selection and Transformation
Data Cleaning and Integration
Knowledge and Information
Extracted Patterns
Dataset
Data Warehouse
Data Sources
End-User Interactions

II. DATA MINING TECHNIQUES

Several core techniques used in data mining to describe the type of mining and data recovery operation. Data mining techniques are useful to transform the data into information. The various models to illustrate the data mining process are given below:

Predictive Model:

- The predictive model makes prediction about unknown data values by using the known values. Ex. Classification, Regression, Time series analysis, Prediction etc.
- Many of the data mining applications are aimed to predict the future state of the data.
- Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state.
- Classification is a technique of mapping the target data to the predefined groups or classes, this is a supervise learning because the classes are 52 predefined before the examination of the target data.
- In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.



Descriptive Tasks

- Find human-interpretable patterns that describe the data.
- The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Summarization, Association rule, Sequence discovery etc.
- Clustering is similar to classification except that the groups are not predefined, but are defined by the data alone. It is also referred to as unsupervised learning or segmentation. It is the partitioning or segmentation of the data in to groups or clusters.
- Summarization is the technique of presenting the summarize information from the data. The association rule finds the association between the different attributes.
- Association rule mining is a twostep process: Finding all frequent item sets, Generating strong association rules from the frequent item sets. Sequence discovery is a process of finding the sequence patterns in data. This sequence can be used to understand the trend.

Classification

Classification techniques in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. We can say that mining patterns that can classify future data into known classes, is classification. Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how “good” the algorithm is.

Classification Algorithms:- ID3, C4.5, Statistical Regression, Bayesian networks, Distance Simplest Near, K-Nearest Neighbors. Neural Network Propagation, NN supervised learning.

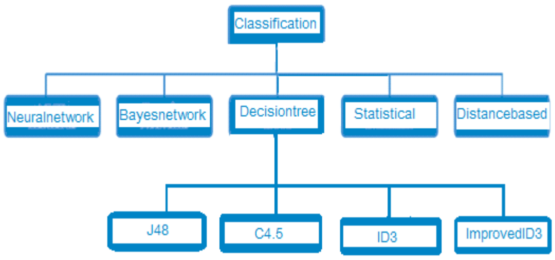


Figure 3. Classification algorithms

Improved ID3 Algorithm

The improved ID3 algorithm based on Gain Ratio decided whether the attributes should be modified or not by checking its weight. So, the inclination problem presented in traditional ID3 problem is overcome.

- IID3's performance is also compared with J48 algorithm and Naïve Bayes classification algorithm. By doing performance analysis of all the three algorithms, a result is that IID3 algorithm is more efficient than other two algorithms.
- IID3 classified 93% of records accurately, whereas J48 and Naïve Bayes classified only 78.6% and 75% records respectively.
- If we want to get less number of node and leaf in a tree and to make it more effective and less complex, we can use this algorithm.
- In ID3 Algorithm, every attribute has the binary value domain (i.e. positive or negative). But in improved ID3 it is also possible that we have some specific attributes that have multiple valued domain (i.e. high, medium, low, etc.).
- ID3 was used only for the balanced datasets but this can be used for imbalanced datasets too (10% positive class and the rest 90% negative class).

Application of decision tree on performance of an employee:

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal.To Construct the Decision Tree, we use following method:

1. Select a variable of training samples as nodes, create a branch to every possible value of the variables. Accordingly, the training sample set is divided into several sub-set.

2. Do the same method to each branch, Training sample is the subset corresponding to the branches and one of the subsets which its parent node is divided into. When the node of all the training samples belongs to the same classification, or no remaining attributes can be used to further divide, Or the branch does not have samples, stop splitting the node branching and make it a leaf node.

III. LITERATURE SURVEY

Shubham Tupe, Chetan Mahajan, Dnyaneshwar Uplenchwar, Pratik Deo(2017)have discussed the decision tree method. Here we uses a decision tree for classifying data into variable types of classes which are useful for prediction purpose. All the task are performed by an admin. Task means adding training data set which required for calculating an entropy and information gain. Once these two modules are calculated the decision tree is created by using an ID3 algorithm.

- To improve the performance of an employee supervisor monitors the data very carefully.
- This kind of technique is very beneficial in order to predict the performance of an employee very accurately by considering the decision tree.
- Decision Tree method is used on an Employee database in order to analyse an employee data to make a prediction.

Ramanathan, Saksham Dhanda, Suresh Kumar(2016) used the Educational Data Mining (EDM) that is attracting a lot of researchers for developing methods from educational institutions' data that can be used in the improvement of quality of higher education. They had used the modified ID3 algorithm with gain ratio. The study was related to

- The prediction of students' performance in any higher institute has become one of the most important needs of that institute in order to improve the quality of the teaching process of that institution.
- Through this process we get to know the needs of the students and hence we can fulfill those needs to get better results.
- The WEKA tool was used for the analysis of J48 and Naive Bayes algorithms. The results are compared and presented.

Souvik Hazra,Satyaki Sanyal(2016) used the ID3 algorithm for recruitment prediction. The process is done in an efficient way with minimal cost and within minimal time. The main aim of the paper is to take into account relevant attributes based on quantitative and qualitative aspects of a candidate such as Years of Experience, Employment Status, Current Salary, Level of Education, Whether from a top-tier school or not and Internships and design a model accordingly to predict the hiring of a candidate.

- This model will help organizations to choose candidates efficiently and within a short period of time.
- This helps the recruiter to come into conclusion of hiring an applicant faster and choose the best for the job. Further implementation can be done for development and application for analysis of similar large datasets.

Hitarthi Bhatt, Shraddha Mehta, Lynette R. D'mello(2015) used to predict the placement of a student.The main aim of this paper is to identify relevant attributes based on quantitative and qualitative aspects of a student's profile such as CGPA, academic performance, technical and communication skills and design a model which can predict the placement of a student. For this purpose ID3 classification technique based on decision tree has been used.

- The result of this analysis will assist the academic planners to design a strategy to improve the performance of students that will help them in getting placed at the earliest.
- It can be concluded that classification algorithms can be used successfully in order to predict student placement.
- Further the implementation can be done in development and application of novel computational techniques for the analysis of large datasets.

Namita Puri, Deepali Khot, Pratiksha Shinde, Kishori Bhoite , Prof. Deepali Maste (2015) focuses on the classification of the student for the placement process in an engineering college. Campus placement of a student plays very important role in a college. So this paper suggests a system which makes the work of prediction of placement of student easy. For this we suggest to use ID3 algorithm- data mining classification algorithm.

- ID3 is a best algorithm which can be used for classification and prediction of student's placement in a engineering college.
- Result indicates that ID3 decision tree algorithm is best classifier with 95% accuracy. This study will also work to identify those students which needed special attention and guidance to increase their chances of placement.
- The generated knowledge will be quite useful for management of placement process to develop policies and strategies for better planning and implementation of educational program like improving soft skills and infrastructure to increase the placed student rate in University.

Vinayak Parmar, Rohit Kotalwar, Raj Chavan , Sunny Gandhi(2014) have discussed the data mining ID3 algorithm in order to predict the performance of an employee through various aspects. Decision Tree method is used on Employee's database to predict the Employee's performance on the basis of previous year database. Data mining is a powerful analytical tool that enables industrial institutions to better allocate resources and staff, and proactively manage employee's outcomes. They studied that

- The Industry superiors will have the ability to predict the employee's performance.
- This also helps the supervisor to find the employee's performance and those employees needed special attention for taking action at right time.
- The prediction of employee's performance with high accuracy is more beneficial. To improve their performance the supervisor will monitor the employee's performance carefully.

Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao(2013) discussed that an educational institution needs to have an approximate prior knowledge of enrolled students to predict their performance in future academics.

- This helps them to identify promising students and also provides them an opportunity to pay attention to and improve those who would probably get lower grades.
- As a solution, we have developed a system which can predict the performance of students from their previous performances using concepts of data mining techniques under Classification.
- We have analyzed the data set containing information about students and results in first year of the previous batch of students. By applying the ID3 (Iterative Dichotomiser 3) and C4.5 classification algorithms on this data, we have predicted the general and individual performance of freshly admitted students in future examinations.

Sonia Singh, Priyanka Gupta(2013) used decision tree learning algorithm. It has been successfully used in expert systems in capturing knowledge.

- The main task performed in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules.
- It is one of the most effective forms to represent and evaluate the performance of algorithms, due to its various eye catching features: simplicity, comprehensibility, no parameters, and being able to handle mixed-type data.
- There are many decision tree algorithm available named ID3, C4.5, CART and CTREE. We have explained three most commonly used decision tree algorithm in this paper to understand their use and scalability on different types of attributes and feature.

Nishant Mathur, Sumit Kumar, Santosh Kumar, and Rajni Jindal(2012) used the ID3 algorithm with Havrda & Charvat entropy based on decision tree .They studied that

- Instead of using Shannon Entropy, Havrda and Charvat Entropy has been used to find the information of different properties which is used as the node of decision tree.
- This modification has reduced the size of tree as well as decreased the rules, which will help to understand customer characteristics by which company growth and profit can be increased.
- In conclusion we can say that if we want to get less number of node and leaf in a tree and to make it more effective and less complex, we can use the Havrda and Charvat entropy instead of Shannon entropy and value of alpha (α) less than one will give decision tree with less number of nodes.

IV. REFERENCES

[1]. Shubham Tupe, Chetan Mahajan, Dnyaneshwar Uplenchwar, Pratik Deo," Employee Performance Evaluation System Using ID3 Algorithm. Vol. 5, Issue 2, February 2017

[2]. Souvik Hazra, Satyaki Sanyal," Recruitment Prediction Using ID3 Decision Tree",International Journal of Advance Engineering and Research Volume 3, Issue 10, October -2016

[3]. Hitarthi Bhatt, Shraddha Mehta, Lynette R. D'mello Use of ID3 Decision Tree Algorithm for Placement Prediction Volume 6, Issue 5, October -2016

[4]. Er. Kanwalpreet Singh Attwal ," Comparative Analysis of Decision Tree Algorithms for the Student's Placement Prediction", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, June 2015

[5]. Namita Puri, Deepali Khot , Pratiksha Shinde, Kishori Bhoite, Prof. Deepali Maste, " Student Performance Prediction Using ID3 Algorithm",International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 3 Issue III, March 2015

[6]. Implementation of Improved ID3 Algorithm to Obtain more Optimal Decision Tree. International Journal of Engineering Research and Development e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com Volume 11, Issue 02 (February 2015), PP.44-47

[7]. D. D. B. Rakesh Kumar Arora, "Placement Prediction through Data Mining," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 7, july 2014.

[8]. Soniya Singh & Priyanka Gupta , "Comparative Study ID3, CART and C4.5 Decision Tree Algorithm: A Survey", International Journal of Advanced Information Science and Technology (IJAIST) Vol.27,No.27,July 2014.

[9]. Sweta Rai, Priyanka Saini, Ajit Kumar Jain "Model for Prediction of Dropout Student Using ID3Decision Tree Algorithm", International Journal of Advanced Research in Computer Science & Technology, Vol. 2 Issue 1 Ver. 2 Jan-March 2014.

[10]. RohitKotalwar,RajChavan,VinaykParmwar, Sunny Gandhi," Evaluating Performance of Employee’s using Classification Algorithm Based on Decision Tree." Vol.4, No. 2, April 2014.

[11]. Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, " Predicting Students' Performance using ID3 and C4.5 classification algorithm", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013

[12]. Ramanathan L, Saksham Dhanda , Suresh Kumar D ,"Predicting Students’ Performance using Modified ID3 Algorithm", https://www.researchgate.net/publication/297699647_Predicting_Students'_Performance_using_Modified_ID3_Algorithm

[13]. Vibha Maduskar, Y.Kelkar ," A New Decision Tree Algorithm based on ID3 Volume 1-Issue 9, July 2013

[14]. Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, "Predicting Students' Performance using ID3 and C4.5 classification algorithm", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013.

[15]. Ming, H., Wenying, N. and Xu, L., (2009)"An improved decision tree classification algorithm based on ID3 and the application in score analysis", Chinese Control and Decision Conference (CCDC), pp1876