

## Role of Decision Tree Classification in Data Mining

A.Rajeshkanna<sup>1</sup> and K.Arunesh<sup>2</sup>

<sup>1</sup> Research Scholar , Department of Computer Science ,  
Sri S. Ramasamy Naidu Memorial College , Sattur , Tamilnadu  
[srnmcrjesh@gmail.com](mailto:srnmcrjesh@gmail.com)

<sup>2</sup> Associate Professor , Department of Computer Science ,  
Sri S. Ramasamy Naidu Memorial College , Sattur , Tamilnadu  
[arunesh\\_naga@yahoo.com](mailto:arunesh_naga@yahoo.com)

**Abstract.** In modern era data is very precious in each and every domain in the field of computer science and it has been growing drastically. Data has been categorized into two types as structured and unstructured data. Classification of data is a central task in Machine Learning as well as in Data Mining. There are several approaches for classification in literature, but Decision Tree (DT) remains as popular approach because of its simplicity and flexibility. For complex data, DT uses classification approach in top down manner. DT is considered as the best solutions to achieve optimal classification for complex problems. DT is not limited to the field of data mining but it also exists in different disciplines such as statistics, artificial neural networks and signal processing and so on. Entropy is the main parameter used in classification and for evaluating the data set. Using this parameter, it is convenient to estimate how much data is being classified and remained. Most of the experimental results with decision tree approaches significantly shows better classification. This paper, concentrates on the decision tree approach with respect to classification based on the proposed mobile users model. The learning process used in Decision Trees are also analysed.

**Keywords:** Decision Tree, Data Mining, Classification, Supervised Learning, Unsupervised Learning.

### 1 Introduction

In early stages, size of the data was very limited. Programming paradigms are used to extract knowledge data with limited size and handling it became very easier. In digital age, the size of the data is grown beyond Tera Bytes (TB) per second. Processing huge amount of data with traditional programming paradigm is

not suggestible. As cost of memory is very cheap, TB of data can be maintained easily. But one of the major problems is to extract knowledge data from the existing raw data. So many new paradigms are invented to extract the knowledge data, but the most popular approach among them is data mining. Data mining algorithms are the essential need for solving complex problems and the effective usage of these algorithms have been increased in recent years. Statistics, Artificial intelligence, Machine learning and databases are some of the relevant areas contributed to the development of data mining in its current form.[1].

Data mining is a process or a methodology to extract information or knowledge from huge data repositories which is being considered as critic but has some added potential values. It has been widely used in the recent areas such as government sectors, transportation, military appliances, Finance and banking sectors for prediction and analysis. Data classification is considered as a main task in Data mining and analyzing. Classification in data mining includes many methodologies and approaches. Decision tree based classification, Neural network approaches, Genetic Algorithms and other statistical models are some of the models used for classification [2]. Classification is a kind of decision support tool for organizations to consume and to resolve problems that occurred during the extraction. When it is handled with human, experts may miss some kind of hidden patterns predicting the necessary information. Most of the companies are having massive quantitative of data mining techniques which plays prominent role in analyzing massive data bases.

## 2 The Foundation of Data Mining

The evolution of data mining techniques started in 1990's. During the initial stage, data was available in computers, disks and tapes. IBM and CDC utilized the mining technique to extract the characteristics like retrospective and static data delivery. Later in 1980's Oracle, Microsoft and others used mining by focussing on dynamic data delivery. In 1990's the decision making approach is achieved by combining data warehousing with mining using On-Line Analytic Processing (OLAP) Servers and Multi-dimensional On-Line Analytic Processing (MOLAP) [1]. Data mining has become an emerging technique in digital age due to its efficiency in handling massive databases with multi-processing computers. The main characteristics needed to extract in digital era is prospective as well as proactive information delivery.

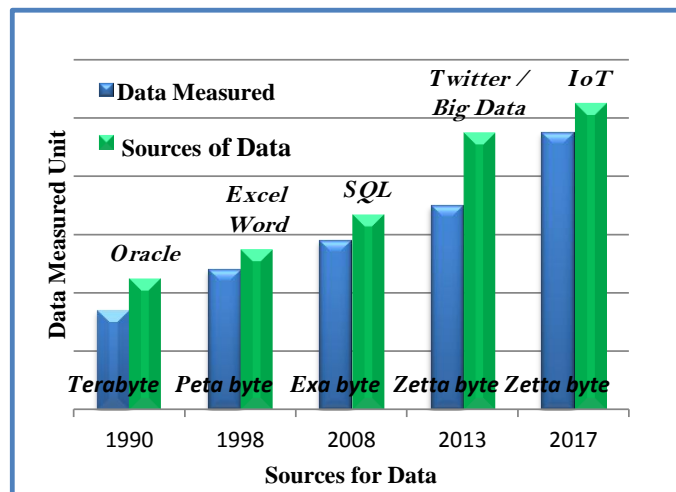


Fig.1.Representation of data magnitude from the year 1990-2017

Fig.1. represents the data magnitude from 1990’s to 2017. It is obvious from the graph that during the year 2017, the emergence of Internet of Things ( IoT ) has a tremendous increase in data.

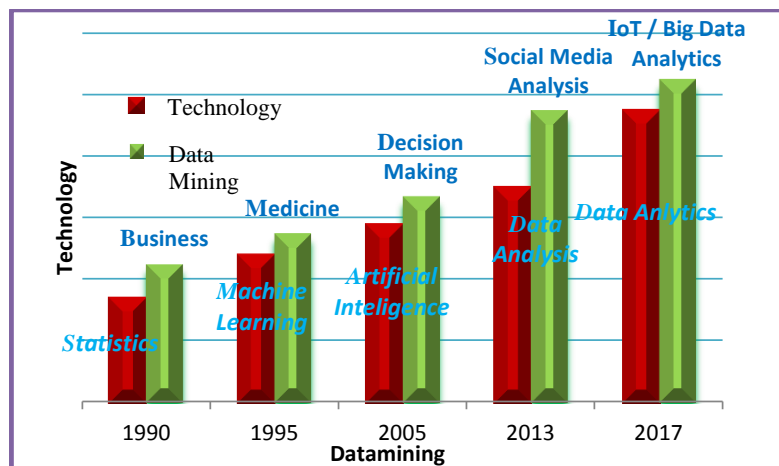


Fig. 2. Impact of Technology in Data Mining

Fig.2. represents the growth of Data Mining based on technology. Data Mining has its applications in various fields such as Business, Medicine, Decision Making Problems such as Stock Market, social Media analysis and in IoT. The impact of Machine Learning algorithms, Artificial Intelligence and Big data analytic tools reflected tremendous growth in data mining applications.

### 3 Architecture of Data Mining

Mining became flexible when it is integrated with data warehousing. Mining tools mainly operated outside of warehouse requiring extra steps to import and analyze the data. Integration of mining and warehousing simplifies analyzation of results. There are seven steps involved in mining as shown in the following Fig. 3. **Data integration:** It is the first step in mining in which data is collected from different sources and it is integrated. **Data Selection:** Extracting the useful information from the integrated data is the second step in the Architecture of data mining. Based on different parameters, only the appropriate attributes will be selected for further processing in Data selection step. **Data Cleaning:** Depending on the types of different sources there will be noisy data which causes inconsistency in data. Data cleaning is the efficient process which gets rid of such anomalies. **Data Transformation:** It is essential to apply data transformation techniques such as smoothing, normalization and some statistical measures to fit the data for further processing. **Data mining:** To discover the predicted information from refined available data, mining techniques such as clustering, association analysis and classification are applied. Above, other different techniques are also existed in mining to predict the required patterns and information. **Pattern Evaluation and Knowledge Presentation:** After mining, patterns are extracted from the databases. The role of this step is to select required patterns in the form of reports charts and to remove the redundancy for better visualization and transformation. **Decisions / Use of Discovered Knowledge:** This is the last step in mining process in which decisions are taken deeply on knowledge acquired from the existing patterns.

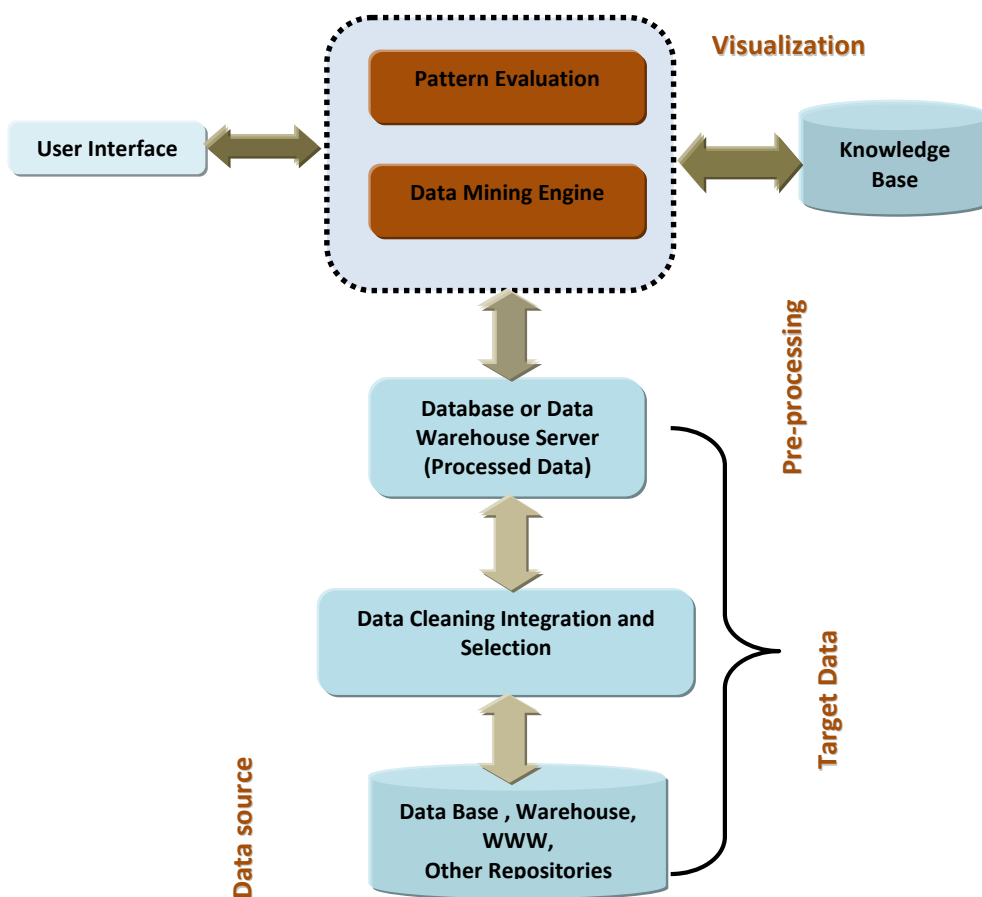


Fig.3. Architecture of Data Mining

#### 4 Classification Technique in Data Mining

Advanced techniques in Data mining are used to discover patterns from data. There are many tasks in data mining such as Association, classification, clustering and outlier detections. Classification is considered as the important and efficient task in data mining to solve complex real problems. Real world problems like

precise and prediction of risk can be solved by classification technique. Classification can be applied on both structured and unstructured data. It generates accurate and efficient clusters. The most popular classification algorithm includes NavieBayes, Support Vector Machine, Decision Trees and so on [3]. By considering all the algorithms, decision tree is the frequently used efficient algorithm with effective parameters such as greater speed, higher accuracy and easier understandability. Classification deals and performs only with structured data. In order to classify the unstructured data, it is essential to convert the data into a structured form.

## 5 Decision Tree

Decision Tree, one of the classification technique, is a constraint oriented construction in a graphical manner. It is very helpful in making the decisions. It requires systematic documentation to process and always identifies potential solution for a given data set.

### 5.1 Overview of Decision Tree

A decision tree is a predictive model in which the decision node plays a vital role for the possible outcome. Decision tree represents a classifier structure in which the root node and internal nodes are labelled. The labels are assumed as nominal or flat in traditional algorithms. In order to predict the future trends in recent complex classification problems, the variables may be of any type such as hierarchical variables, continuous variables or a combination of both type of variables. The hierarchical continuous labels can be constructed efficiently.[11]. Information gain approach is usually used to analyse the suitable property of each node of a generated decision tree. Using Information gain the attribute with the highest information gain can be selected as a test attribute. Information gain can be combined with attribute importance, and it is used as a new standard of attribute selection to construct decision trees [12].

Decision tree classification scheme is also considered as the important classifier for Artificial intelligence as well as machine learning. A new data mining technique, cluster based fuzzy decision tree (GCFDT), is introduced for the mining task . [13]. The traditional ID3 algorithm has some challenging issues in solving the complex problems. The estimation of split measure function value for every node and the selection of the best attribute in the construction of the decision trees are the important factors to be considered. Many researchers are trying to justify the same mathematically[14]. DT also has some challenging issues such as to compute the Fourier spectrum and to construct a decision tree

from its Fourier spectrum. MobiMine, a mobile data stream mining system, is used to develop techniques for mining stock-market data from handheld devices. [15]. The construction of a Decision Tree for a mobile user's model is shown in Fig. 4.

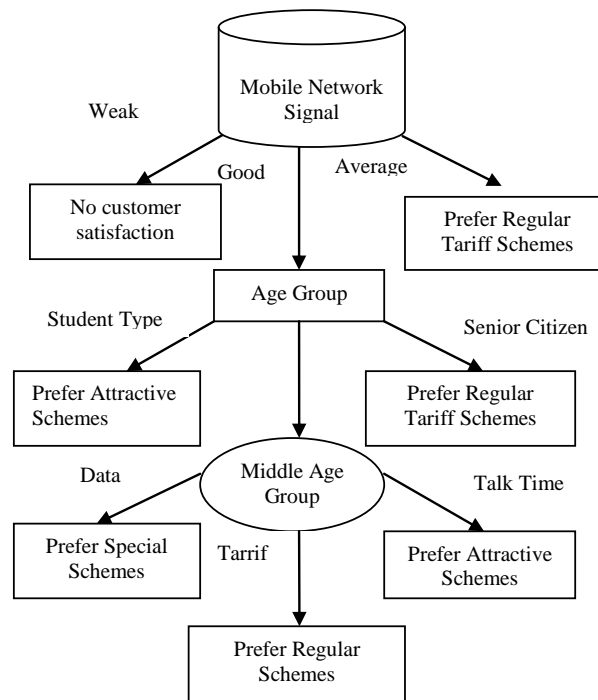


Fig.4. Decision Tree Model for Mobile Users

Decision Tree model for the selection of a recharge scheme for the particular mobile users is depicted in Fig.4. The selection of a particular network depends on the strength of the signal and the customer satisfaction is analysed. Based on the Age group, the customers are classified. According to the classification, the choice of the recharge schemes for Data, Talk time and Tariff are categorized. The result of the proposed model will yield the satisfied customers and the choice of the schemes for data and talktime and general tariff. Based on the output, the schemes can be determined. Thus Decision tree model will be very helpful for analyzing the type of customers and their choice, which in turn will be useful for satisfying the

buisness needs. Decision Tree can be applied to two kinds of learning algorithms such as supervised and unsupervised learning. Supervised learning is defined by the learning of a function which maps the input variables(x) and output variables or target variables(t). In general terms, it is defined as

$$t = f(X).$$

The supervised learning algorithm as the name indicates learns from the training data set and also can predict the desired output values depending on the input vector. The parameters chosen may be adjusted based on the optimal performance of the training set and the accuracy of the learned function can be evaluated. Supervised method is also used to discover the relationship between the input and output. In relationship discovery two main supervised models are categorised as classifier and regression [4]. There are many classifiers available to represent classification but the Decision tree is the most reliable approach. Originally it is studied in the field of statistics and in other areas. Regression maps input space into real world domain. There are several predefined classes available to provide a space for mapped input. Regression works on real world examples like stock market and advertisement. Unsupervised learning is categorized as having the input variables (X) and the unknown corresponding output variables [5]. Unsupervised learning has no prediction of target attributes and is based on learning without a teacher. Data has been clustered depending on statistical properties. Cluster will have significant labelling and the labelling is procedure carried out for small number of objects which are presented in the class. Unsupervised problems are grouped into two types as clustering and association [6]. Clustering is a problem which discovers and inherits groups depending on the behaviour of database. In clustering there are three popular types such as c-means, k-means and fuzzy clustering. Association discovers the rules that are used to divide the data into number of portions. The popular association rule algorithm is Apriori algorithm. For the DT classification, depending on the dataset both supervised and unsupervised learning are used to predict the outcome. In the proposed mobile user's model, both supervised and unsupervised learning can be applied depending on the previous dataset. Depending on the requirement of the outcome, the data may also be clustered or the association between the different parameters can be determined. Entropy is a measure of the homogeneity of the set of examples. In order to define information gain, we need to define a measure called Entropy. For a binary classification, Entropy is defined as,

$$Entropy(S) = -P_p \log_2 P_p - P_n \log_2 P_n$$

Where  $P_p$  represents the set of positive examples and  $P_n$  represents the set of negative examples in  $S$ .

### Conclusions

This study addresses the important role of Decision Trees in the classification technique. It is also addressed that the parametric approach to handle structured and unstructured data using decision tree is an efficient methodology. From this study, it is concluded that the decision tree is the best approach to classify the data base in two criteria's such as splitting criteria and stopping criteria. Both the supervised and unsupervised learning approach will be helpful for predicting the outcome. Based on Information gain and entropy, the attribute selection can be effectively determined and the role of Decision tree will be very essential for solving complex problems.

### References

1. Ramzan, Majid, and Majid Ahmad. "Evolution of data mining: An overview.", in IT in Business, Industry and Government (CSIBIG) Conference on. IEEE, (2014).
2. Li, Linna, and Xuemin Zhang. "Study of data mining algorithm based on decision tree." Computer Design and Applications (ICCD), in International Conference on. IEEE, vol.1.(2010).
3. Karabadji, Nour El Islem, et al. "An evolutionary scheme for decision tree construction." Knowledge-Based Systems 119 ,pp.166-177(2017).
4. Rokach, Lior, and Oded Maimon. "Top-down induction of decision trees classifiers-a survey." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 35.4 ,pp.476-487(2005).
5. Balaji, S., and S. K. Srivatsa. "Unsupervised Learning in Large Datasets for Intelligent Decision Making." International Journal of Scientific and Research Publications, Volume 2, Issue 9, September 2012
6. Buttar, Harveen, and Rajneet Kaur. "Association Technique in Data Mining and Its Applications." International Journal of Computer Trends and Technology (IJCTT), vol.4, issue 4, (2013).
7. Kesavaraj, G., and S. Sukumaran. "A study on classification techniques in data mining." In Computing, Communications and Networking Technologies (ICCCNT), Fourth International Conference on. IEEE, (2013).
8. Gorade, Mr Sudhir M., Ankit Deo, and Preetesh Purohit. "A Study of Some Data Mining Classification Techniques." International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 04 Issue: 04 | Apr -2017.

9. Khalid, Balar, and NajiAbdelwahab. "A Comparative Study of Various Data Mining Techniques: Statistics, Decision Trees and Neural Networks." *International Journal of Computer Applications Technology and Research* 5,pp. 172-175,(2016).
10. Agarwal, Sonali, G. N. Pandey, and M. D. Tiwari. "Data mining in education: data classification and decision tree approach." *International Journal of e-Education, e-Business, e-Management and e-Learning* 2.2 ,140,(2012).
11. Hu, Hsiao-Wei, Yen-Liang Chen, and Kwei Tang. "A novel decision-tree method for structured continuous-label classification." *IEEE transactions on cybernetics* 43.6,pp 1734-1746(2013).
12. Jin, Chen, Luo De-Lin, and Mu Fen-Xiang. "An improved ID3 decision tree algorithm." *Computer Science & Education, ICCSE'09. 4th International Conference on. IEEE, (2009).*
13. Shukla, Sanjay Kumar, and Manoj Kumar Tiwari. "GA guided cluster based fuzzy decision tree for reactive ion etching modeling: a data mining approach." *IEEE Transactions on Semiconductor Manufacturing* 25.1 (2012): 45-56.
14. Rutkowski, Leszek, et al. "Decision trees for mining data streams based on the gaussian approximation." *IEEE Transactions on Knowledge and Data Engineering* 26.1 (2014): 108-119.
15. Kargupta, Hillol, and B-H. Park. "A fourier spectrum-based approach to represent decision trees for mining data streams in mobile environments." *IEEE Transactions on Knowledge and Data Engineering* 16.2 (2004): 216-229.
16. Ramzan, Majid, and Majid Ahmad. "Evolution of data mining: An overview." *IT in Business, Industry and Government (CSIBIG), 2014 Conference on. IEEE, 2014.*



