International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 02 | Feb-2018 www.irjet.net p-ISSN: 2395-0072

DATA MINING TECHNIQUES AND APPLICATIONS

Prof. Neha Khatri-Valmik[1], Prof. Supriya Madhukar Salve[2], Syed Neha Samreen[3]

[1,2,3] *Assistant Professor, Dept of CSE, People's Education Society's College of Engineering, Aurangabad*

Abstract - *Data mining is examining huge pre-existing databases to generate new information. Data mining also known as knowledge discovery and it is a process of analyzing data from different perspectives also summarizing it into useful information. Definition of information can be used to increase revenue, cuts costs, or both. Data mining is analytical tool for analyzing data. The software allows users to analyze data from many different angles, categorize it, and summarize the relationships identified. Actually data mining is the process of searching patterns or correlations among various fields in large relational databases.*

Key Words: data, mining, tool, applications, knowledge

1. Data mining Overview

The development of Information generated large amount of databases and huge data. The research gave rise to an approach to store and manipulate this data for further decision making. The process of Data mining is extraction of useful information and patterns from huge amount data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis [1].

Data is any facts, text or numbers that can be processed by a computer system.

Information is pattern, association or relationship among all entire data that can provide information.

Knowledge is when **information** can be converted into knowledge about patterns and future trends.

2. Techniques:

Various algorithms and different techniques are Association Rules, Classification, Clustering, Decision Trees, Genetic Algorithm, Nearest Neighborhood method, Neural Networks, Regression, etc. are used for knowledge discovery from databases.

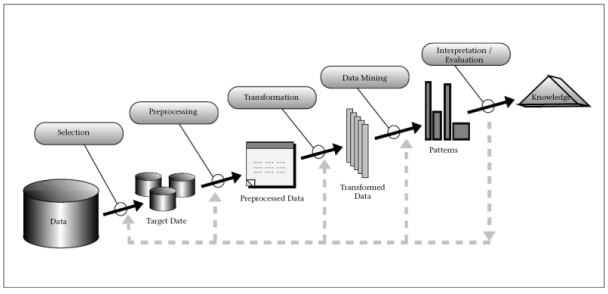

Fig -1: Knowledge Process [1]

2.1 Association rule

Association and correlation is many of the times used for find frequent item set findings amongst large data sets. This type helps businesses to make certain decisions, such as catalog design.

Association Rule must be able to generate rules with confidence values which should be less than one. The numbers of possible Association Rules for a given data set are larger values (if any) and a very high proportion of the rules have little values (if any).

Types of association rule

- Multi level association rule
- Multidimensional association rule
- Quantitative association rule

2.2 Classification

Classification method is assigning a class or label to a set of unclassified cases.

Two types of classification are:-

1. **Supervised Classification: It is** set of possible classes are known.
2. **Unsupervised Classification**: It is set of possible classes not known. After classification we can assign a name to that class. Unsupervised classification is also called as clustering.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification- The attributes are independent based on given class & this is called "Naïve" classifier because of these assumptions. Bayesian is a statistical classifier.

2.3 Clustering

Clustering is defined as identification of similar classes of objects. In clustering techniques we can identify dense and sparse regions in objects Also we can discover distribution pattern and correlations among data attributes. The classification technique is effective means of distinguishing groups/classes of objects. This method is costly so clustering is used as preprocessing approach [1].

Types of clustering methods

- Partitioning Method
- Hierarchical Agglomerative method
- Density based method

2.4 Decision Trees

In this Tree-shaped structure represents sets of decisions. Decisions generate rules for classification of dataset. Specific decision trees methods include Classification and Regression Trees and Chi Square Automatic Interaction Detection. Above are decision tree techniques used for classification of data set. They provide set of rules that can be applied to new unclassified data set [2].

Decisions trees are very simple to understand also provide proper results with small data.

Decision tree induction algorithms can also be used for classification in many areas like Educational, Medical, Manufacturing, Production, Financial, Fraud Detection as well as Astronomy

Advantages of Decision tree algorithm

This algorithm can handle continuous & discrete data.

a. Provides faster result in classifying the unknown records.
b. Also works well with large redundant attribute.
c. Provides results with small tree size.
d. A result does not affect the outliers.
e. Do not require preparation method, normalization.
f. Also works well with numeric data.

2.5 Genetic Algorithm

A genetic algorithm is search heuristic method that imitates the process of natural selection. Heuristic is also many times called as a Meta heuristic is used to generate useful solutions to search problems & optimization.

Genetic programming is vastly used in research in the last 10 years for solve data mining classification problems. The reason is that genetic programming is predicts rules that naturally represent in Genetic programming. Additionally, Genetic programming has been proven to produce good results with global search problems.

2.6 Nearest Neighborhood method

The Nearest Neighborhood Algorithm is used to determine solution to the travelling salesman problem. Nearest Neighbor are used for knowledge discovery for large data sets [3].

K-means clustering is used to partition n observations into k clusters, in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster [5].

The nearest neighbor technique is very easy to implement & executes quickly. Many times it misses shorter routes which are noticed with human perception because of its "greedy" nature.

K-nearest can perform better with missing data

a. Easy to implement & debug.
b. Provides accurate results.
c. Noise reduction techniques are used to improve accuracy of classifier.

Cluster is a group of similar things positioned/occurring closely together.

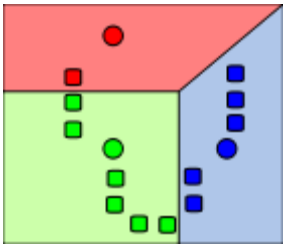

Fig -2: *k* clusters are created by associating every observation with the nearest mean [5]

Neural Networks

The area of neural networks belongs to the line between the artificial intelligence and approximation algorithm. Neural networks are modelled after the cognitive processes of brain. It is capable of predicting new observations from existing observations. A neural network consists of some interconnected elements namely units & nodes. Neurons are the collection of neurons like processing units, weighted connections. It is of many elements, one of this is called as nodes are connected in between. The connection between two nodes is weighted and training of the network is performed [4]. All the neurons within the network work together to produce a function as output. The computation is performed by collective neurons. Neural network can produce the output function even after single neurons are malfunctioned.

An advantage of Data Mining includes areas such as marketing, finance & government sector.

Disadvantages of data mining are privacy & security issues as well as misuse of information

Scope of Data Mining

Data mining itself says that it is similarities between searching for valuable business information in huge databases. Data mining techniques can produce benefits of automation on existing software & hardware platform. Also it can be implemented on systems like existing platforms are upgraded & new products are developed. Data mining can be used as faster processing; it means that users can automatically experiment with multiple models to understand huge complex data. Huge databases, in turn, generate improved predictions.

3. CONCLUSIONS

In this paper I have described how data mining techniques are useful & can be applied in educational field. It will help the educational institutions to decide the individual programs to enhance skills of the staffs as well as students to improve their performance in data mining. Data mining brings a lot of advantages to various fields like businesses, society, governments also containing the individuals. Privacy, security & misuse of information are the huge problems if they are not addressed and settled properly.

ACKNOWLEDGEMENT

First and foremost, I would like to thank my family for their guidance and support. I will forever remain grateful for the constant support and guidance extended them.

REFERENCES

[1] DATA MINING TECHNIQUES AND APPLICATIONS Mrs. Bharati M. Ramageri M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[3] Monika Goyal and Rajan Vohra, "Application of Data Mining in Higher Education", International Journal of Computer Science (IJCSI) Issues, Vol. 9, Issue-2, No.1, March 2012; pp-113-120.

[4] Gajendra Sharma, "Data mining and Data Warehousing and OLAP", Published by S.K. Kataria & Sons, New Delhi, India.

[5] https://en.wikipedia.org/wiki/K-means_clustering.

BIOGRAPHIES

Ms Neha Khatri-Valmik interests lie in various subjects like Digital Electronics, Computer Networks, Digital Image Processing, Open source systems, Software testing & Mining. I have a teaching experience of more than 8 years and currently working as Assistant Professor at PESCOE under the Department of Computer Science & Engineering, Aurangabad.

Ms. Supriya Madhukar Salve have a teaching experience of more than 10 years and currently working as Assistant Professor at PESCOE under the Department of Computer Science & Engineering, Aurangabad.

Ms. Neha Syed Samreen has completed her Masters in Computer Science & Engineering. She have teaching experience of more than 4 years and currently working as Assistant Professor at PESCOE under the Department of Computer Science & Engineering, Aurangabad.