International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 3 | ISSN : 2456-3307
DOI : 10.32628/CSEIT1953167

Study of Data Mining Techniques and Its Types

Alok Kumar Awasthi
Department of Computer Science and Engineering, Saraswati Higher Education and Technical College of Engineering, Varanasi, Uttar Pradesh, India

ABSTRACT

Data mining, classification, clustering, decision tree, Associative rule embedded processor technology moving towards faster and smaller processors and systems on a chip, it becomes increasingly difficult to accurately evaluate real time performance. This research describes an evaluation method using an embedded architecture software emulator that models the Motorola M-CORE processor architecture. This emulator is used to evaluate and compare the real-time performance of a public-domain experimental Real-Time Operating System (RTOS) against a bare-bones multi-rate task scheduler. The results of the experiment, as shown in arrival time JITTER, response-time DELAY, and CPU BREAKDOWN figures, show the trade-offs between job load, job frequency, and kernel overhead. This research suggests full-system software emulation to be a valid method of evaluating embedded systems' behavior and real-time performance.

Keywords : Data Mining, Classification, Clustering, Decision Tree, Associative Rule

I. INTRODUCTION

Data mining is a set of techniques to discover patterns, associations or, in general, interesting knowledge from large amounts of data. In the last ten to twenty years, as the volumes of stored digital data, the memory capabilities and the computing power have grown, also has the need to take advantage of all that potential.

For instance, in several industries like communications or retail distribution (e.g.: supermarkets) there are huge databases of operational data that have plenty of hidden underlying information. The aim of data mining is to uncover that information and provide the decision makers with the knowledge to make better informed decisions. In an academic environment, as is the case of this thesis, the aim is identical, it is to perform knowledge discovery in a huge database.

Data mining is essentially available as number of commercial systems. Today, data mining is widely used in nearly every industry. For example, financial data analysis is usually systematic as the data is highly reliable. Typical cases of financial data analysis include: loan payment prediction, customer credit policy analysis, classification and clustering of customers for targeted marketing, detection of money laundering, and other financial crimes.

Data mining has a bigger role to play in the retail industry, since it collects data from various sources like sales, customer purchasing history, goods transportation, consumption, and services. In the retail industry it helps in identifying customer behaviours; designing and constructing data warehouses based on the benefits of data mining; multidimensional analysis of sales, customers, products, time and region; effectiveness of sales campaigns; customer retention; product recommendation, and cross-referencing of items.

In the telecommunication industry, data mining helps identify telecommunication patterns, detect

fraudulent activities, improve the quality of services and also make better use of resources.

Data mining has also made significant contributions to biological data analysis like genomics, proteomics, functional genomics, and biomedical research. It helps in analysis by semantic integration of heterogeneous, distributed genomic and proteomic databases; association and path analysis, visualization tools in genetic data analysis, and more.

It also helps in the analysis of large amounts of data from domains such as geosciences, astronomy, and more. Other scientific applications such as climate and ecosystem modeling, chemical engineering, and fluid dynamics all benefit from data mining.

One of the most important tasks in Data Mining is to select the correct data mining technique. Data Mining technique has to be chosen based on the type of business and the type of problem your business faces. A generalized approach has to be used to improve the accuracy and cost-effectiveness of using data mining techniques. There are basically seven main Data Mining techniques which are discussed in this article. There are also a lot of other Data Mining techniques but these seven are considered more frequently used by business people.

- Statistics
- Clustering
- Visualization
- Decision Tree
- Association Rules
- Neural Networks
- Classification

Data mining techniques statistics is a branch of mathematics which relates to the collection and description of data. The statistical technique is not considered as a data mining technique by many analysts. But still, it helps to discover the patterns and build predictive models. For this reason, data analyst should possess some knowledge about the different statistical techniques. In today's world, people have to deal with a large amount of data and derive important patterns from it. Statistics can help you to a greater extent to get answers for questions about their data like

- What are the patterns in their database?
- What is the probability of an event to occur?
- Which patterns are more useful to the business?
- What is the high-level summary that can give you a detailed view of what is there in the database?

Statistics not only answer these questions they help in summarizing the data and count it. It also helps in providing information about the data with ease.

CLUSTERING

Clustering is one of the oldest techniques used in Data Mining. Clustering analysis is the process of identifying data that are similar to each other. This will help to understand the differences and similarities between the data. This is sometimes called segmentation and helps the users to understand what is going on within the database. For example, an insurance company can group its customers based on their income, age, nature of policy and type of claims. This survey's emphasis is on clustering in data mining. Such clustering is characterized by large datasets with many attributes of different types.

Clustering Algorithms

1. Hierarchical Methods
   i. Agglomerative Algorithms
   ii. Divisive Algorithms
2. Partitioning Methods
   i. Relocation Algorithms
      Probabilistic Clustering
   ii. K-medoids Methods
   iii. K-means Methods
   iv. Density-Based Algorithms

Density-Based Connectivity Clustering
Density Functions Clustering

3. Grid-Based Methods
4. Methods Based on Co-Occurrence of Categorical Data
5. Constraint-Based Clustering
6. Clustering Algorithms Used in Machine Learning
    i. Gradient Descent and Artificial Neural Networks
    ii. Evolutionary Methods
7. Scalable Clustering Algorithms
8. Algorithms For High Dimensional Data
    i. Subspace Clustering
    ii. Projection Techniques
    iii. Co-Clustering Techniques

II. VISUALIZATION

Visualization is the most useful technique which is used to discover data patterns. This technique is used at the beginning of the Data Mining process. Many types of research are going on these days to produce an interesting projection of databases, which is called Projection Pursuit. There is a lot of data mining technique which will produce useful patterns for good data. But visualization is a technique which converts Poor data into good data letting different kinds of Data Mining methods to be used in discovering hidden patterns.

A decision tree is a predictive model and the name itself implies that it looks like a tree. In this technique, each branch of the tree is viewed as a classification question and the leaves of the trees are considered as partitions of the dataset related to that particular classification. This technique can be used for exploration analysis, data pre-processing and prediction work.

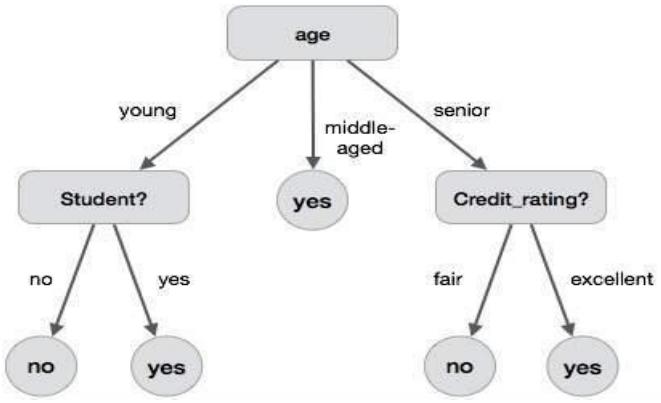

Figure 2. The decision tree

The decision tree can be considered as a segmentation of the original dataset where segmentation is done for a particular reason. Each data that comes under a segment has some similarities in their information being predicted. Decision trees provide results that can be easily understood by the user.

Neural Network

Neural Network is another important technique used by people these days. This technique is most often used in the starting stages of the data mining technology. The artificial neural network was formed out of the community of Artificial intelligence.

Neural networks are very easy to use as they are automated to a particular extent and because of this the user is not expected to have much knowledge about the work or database. But to make the neural network work efficiently you need to know

- How the nodes are connected?
- How many processing units to be used?
- When should the training process be stopped?

There are two main parts of this technique – the node and the link

- **The node** – which freely matches to the neuron in the human brain
- **The link** – which freely matches to the connections between the neurons in the human brain.

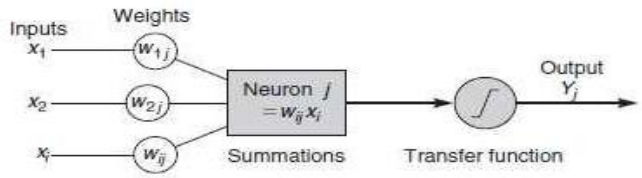

Figure 2. Processing in an Artificial Neuron

This technique helps to find the association between two or more items. It helps to know the relations between the different variables in databases. It discovers the hidden patterns in the data sets which is used to identify the variables and the frequent occurrence of different variables that appear with the highest frequencies.

Association rule

Association rule offers two major information

- Support – Hoe often is the rule applied?
- Confidence – How often the rule is correct?

This technique follows a two-step process

- Find all the frequently occurring data sets
- Create strong association rules from the frequent data sets

There are three types of association rule. They are

- Multilevel Association Rule
- Multidimensional Association Rule

Data mining techniques classification is the most commonly used data mining technique which contains a set of pre-classified samples to create a model which can classify the large set of data. This technique helps in deriving important information about data and metadata (data about data). This technique is closely related to the cluster analysis technique and it uses the decision tree or neural network system. There are two main processes involved in this technique

- **Learning** – In this process the data are analyzed by the classification algorithm
- **Classification** – In this process, the data is used to measure the precision of the classification rules

There are different types of classification models. They are as follows

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Association Rule

III. CONCLUSION

There are several classification techniques in data mining and each and every technique has its advantage and disadvantage. Decision tree classifiers, Bayesian classifiers, classification by back propagation, support vector machines, these techniques are eager learners they use training tuples to construct a generalization model. Some of than are lazy learner like nearest-neighbor classifiers and case-based reasoning. These store training tuples in pattern space and wait until presented with a test tuple before performing generalizatio

IV. ACKNOWLEDMENT

The author is grateful to Department of Computer of Science, Saraswati Higher Education and Technical College of Engineering, Varanasi for conducting entire work.

V. REFERENCES

[1] Cabena, Peter; Hadjnian, Pablo; Stadler, Rolf; Verhees, Jaap; Zanasi, Alessandro (1997); Discovering Data Mining: From Concept to Implementation, Prentice Hall, ISBN 0-13-743980-6

[2] M.S. Chen, J. Han, P.S. Yu (1996) "Data mining: an overview from a database perspective". Knowledge and data Engineering, IEEE Transactions on 8 (6), 866–883

[3] Feldman, Ronen; Sanger, James (2007); The Text Mining Handbook, Cambridge University Press, ISBN 978-0-521-83657-9

[4] Guo, Yike; and Grossman, Robert (editors) (1999); High Performance Data Mining: Scaling Algorithms, Applications and Systems, Kluwer Academic Publishers

[5] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.

[6] Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome (2001); The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, ISBN 0-387-95284-5

[7] Liu, Bing (2007, 2011); Web Data Mining: Exploring Hyperlinks, Contents and Usage Data.

Cite this article as :

Alok Kumar Awasthi, "Study of Data Mining Techniques and Its Types", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 3, pp. 529-533, May-June 2019.

Journal URL : http://ijsrcseit.com/CSEIT1953167