DATA MINING TECHNIQUES

AUTHOR: NUWAMANYA EZRA MATAGAIZI

DESIGNATION: STUDENT

GUIDED BY: DR. PRIYA SWAMINARAYAN

DEPARTMENT OF COMPUTER APPLICATIONS,

PARUL UNIVERSITY, VADODARA – GUJARAT, INDIA.

ABSTRACT.

In a generation revolutionized and anchored by information technology, data is primarily part of the core fundamentals of every innovation and development. This data is undoubtedly in big chunks referred to as big data and as the name suggests, these are extremely huge and complex data sets that may be analyzed computationally to reveal patterns, trends and associations in analytical applications.

The assignment of adequate processing and utilization of such large unstructured, semi-structured and structured data sets could no longer be efficiently dealt with by traditional data processing application software and hence the emerging of data mining.

Data mining, a practice that critically examines big data for information or knowledge extraction was deployed to substitute the unyielding traditional methods. A result oriented data mining process links to use of well-defined data mining techniques. The most popular data mining techniques include; Association, Clustering, Classification, Decision Tree, Sequential Patterns, Prediction etc.

This research seminar paper aims at comparing three (3) of these prominent techniques (Association, Clustering and Classification) in attempt to discover why, how, where and when they are used in data mining!

Keywords: ***Data mining, Data mining techniques, Association rule, Clustering, Classification.***

I. INTRODUCTION

Data mining techniques are the already established methodologies used in the implementation of data mining during the Knowledge Discovery in Databases.

Early methods of identifying patterns in data include Bayes' Theorem that traces back in the 1700s and the Regression Analysis of the 1800s. The increasing power of computer technology has dramatically increased data collection and manipulation abilities. As data sets have grown in size and complexity, better data and knowledge extraction techniques have been invented for example Clustering, Association rules, Classification etc.

Data mining is the process of applying these techniques with the intention of uncovering the hidden patterns in large data sets. It bridges the gap from hidden statistics and AI to management by exploiting the way data is stored and indexed within the databases to execute the particular learning and discovery algorithms more efficiently allowing such methods to always be applied to larger data sets.

II. LITERATURE REVIEW

DATA MINING

This is a field of data science that examines large pre-existing databases to generate hidden knowledge and patterns. It is a vast process that involves sorting and selection through large data sets to identify patterns and build relationships to solve problems through data

analysis. Data mining takes into note three (3) things which are; new, correct, and useful information. It (data mining) is done on selected and preprocessed data after which knowledge is presented from pattern evaluation.

DATA MINING TECHNIQUES

Data mining techniques are the tools or methodologies used in extraction of patterns and knowledge from large volumes of numerous forms of data sets. These techniques focus on using specific machine learning and statistical sub models to predict the future and discover the patterns among data.

To mine data with natural language text, structured and unstructured content, it makes sense to fuse data mining techniques with methods of information retrieval and natural language processing.

Data mining techniques also cover the scope of data uncertainty, noise and incompleteness among others.

ASSOCIATION RULE

The Association rules also known as Market Basket Analysis is a data mining function that discovers the probability of the co-occurrence of items in a collection. The relationship between certain co-occurring elements is referred to as Association rules.

Following the original definition by Agrawal, Imielinski and Swami, the problem of Association rule mining is defined as;

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a group(set) of *n* binary attributes called *items*.

Let $D = \{t_1, t_2, \ldots, t_m\}$ be a group(set) of transactions called the *database*

Each *transaction* in D has a unique transaction ID and contains a subset of items in I.

A *rule* is interpreted as an insinuation of the form:

$X \rightarrow Y$, where X, Y are subsets of I.

According to Agrawal, Imielinski and Swami, a rule is defined only between a set and a single item, $X \rightarrow i_j$ for $i_j \in I$.

Every rule is composed of 2 different sets of items, also known as *itemsets*, X and Y, where X is called *antecedent* or left-hand-side (LHS) and Y *consequent* or right-hand-side (RHS)

Algorithms in Association rule technique.

i. **Apriori Algorithm.**

The Apriori algorithm takes on the BFS (Breadth First Search) strategy to add up the support of itemsets and uses a candidate generation function which utilizes and exploits the downward closure property of support.

ii. **Eclat Algorithm.**

Eclat stands for Equivalence Class Transformation. It is a depth-first search algorithm, based on set intersection. It is suitable for both sequential also as parallel execution with locality-enhancing properties.

iii. FP-growth algorithm.

FP stands for Frequent Pattern. In the first pass, the algorithm counts occurrence of items (attribute-value pairs) in the dataset and stores them in 'header table'. In the subsequent pass, it sets up the FP-tree structure by inserting instances. Items in each instance need to be sorted by descending order of their frequency within the dataset, in order that the tree is often processed quickly. Items in each instance that don't meet minimum coverage threshold are discarded. If many instances share most frequent items, FP-tree provides high compression on the brink of tree root.

CLUSTERING

This is a data mining technique that groups sets of data items or objects based on their characteristics and partitions them according to their similarities. Clusters or Partitions are done to form groups of data elements which are very similar to others with in the same group but very dissimilar to elements of the other groups. Clusterings can be roughly distinguished as;

i. **Hard clustering:** Each object belongs to a cluster or not.
ii. **Soft clustering (fuzzy clustering):** Each object belongs to each cluster to a certain degree (i.e. a likelihood of belonging to a cluster)

There are also finer distinctions possible, for example;

i. **Strict partitioning clustering:** Each object belongs to exactly one cluster.
ii. **Strict partitioning clustering with outliers:** Objects can also belong to no cluster, and are considered outliers.
iii. **Overlapping clustering:** It is also known as alternative clustering, multi-view clustering. Objects may appear in more than one cluster; usually involving hard clusters.
iv. **Hierarchical clustering:** Objects that belong to a child cluster also belong to the parent cluster.
v. **Subspace clustering:** While an overlapping clustering, within a uniquely defined subspace, clusters are not expected to overlap.

CLASSIFICATION

Classification is a data mining technique that assigns an object to a certain class relating to its similarities to the already defined and classified objects. It is a predictive method of data mining where new data samples are classified with the known predefined classes. This methodology immensely works with decision trees, linear programming, statistics and mathematical techniques.

In the terminology of machine learning, classification is taken into account as an instance of supervised learning, i.e. learning where a training set of correctly identified observations is obtainable. Often, the individual observations are analyzed into a group of quantifiable properties, known variously as explanatory variables or features.

These properties might variously be categorical (for example; 'A', 'B', 'AB' or 'O', for blood type), ordinal (for example; 'large', 'medium' or 'small'), integer-valued (for example the number of occurrences of a specific word in an email) or real-valued (for example a measurement of blood pressure). Other classifiers work by a comparison of observations to previous observations by means of a similarity or distance function.

Classification algorithms

i. Linear Classifier.

In the field of machine learning, the goal of statistical classification is to use an object's characteristics to spot which class (or group) it belongs to. A linear classifier attains this by making a classification decision hinging on the value of a linear combination of the characteristics.

An object's characteristics are also referred to as feature values and are typically presented to the machine in a vector known as a feature vector. Such classifiers work well for practical problems like document classification.

ii. Decision tree.

Decision tree learning implements a decision tree as a predictive model to travel from observations about an item represented with in the branches to conclusions about the item's target value represented in the leaves.

III. FUTURE WORKS

Standardization of data mining languages

There are various data mining tools with different syntaxes, hence it is to be standardized to make it convenient to the users. Data mining applications has got to concentrate more in standardization of interaction languages and versatile user interactions.

Data Preprocessing

To identify useful novel patterns in distributed, large, complex and temporal data, data processing techniques has got to evolve in various stages. The present techniques and algorithms of knowledge preprocessing stage aren't up to the mark compared with its significance find out the novel patterns of knowledge. In future there is a profound need of data mining applications with well-organized data preprocessing techniques.

Complex object of data

Data mining shall utterly penetrate in all spheres of human life; the presently available data processing and mining techniques are restricted to mining the normal forms of data only, and in future there is a potentiality of data mining techniques for complex data objects like; high dimensional, high speed data streams, sequence, noise in the time series, graph, multi-instance objects, multi-represented objects and temporal data.

IV. REFERENCES

i. *Microsoft Academic Search: Most cited data mining articles Retrieved: 04/02/2019*

ii. *R. Ng and J. Han. "Efficient and Effective clustering method for spatial data mining" In: Proceedings of the 20th VLDB Conference. Retrieved: 28/01/2019*

iii. *Arabie, P. "Comparing partitions". Journal of Classification. Retrieved:23/01/2019*

iv. *"KDD-2000 Workshop on Text Mining – Call for Papers" cs.cmu.edu Retrieved: 05/01/2019*

v. *Ma, Y.; Guo, Y.; Tian X.; Ghanem, M.(2011). "Distributed Clustering-Based Aggregation Algorithm for Spatial Correlated Sensor Networks". IEEE Sensors Journal 11 (3): 641. Retrieves: 03/02/2019*

vi. *Z. K. Baker and V.K. Prasanna. 2005. Efficient Parallel Data Mining with Apriori Algorithm on FPGAs. Retrieved: 10/02/2019*

vii. *Ian H. Witten; Eibe Frank; Mark A. Hall (30th January 2011). Data Mining: Practical Machine Learning Tools and Techniques (3 ed.). Retrieved: 07/02/2019*