Applied Science and Engineering Journal for Advanced Research *Peer Reviewed and Refereed Journal*
ISSN (Online): 2583-2468
Volume-2 Issue-4 || July 2023 || PP. 13-17 DOI: 10.54741/asejar.2.4.3

Data Clustering and Techniques with Weights Over Data Stream

Abhishek Verma
Department of Computer Science & Applications, S D College, Ambal Cantt, India

Corresponding Author: abhishek7685@gmail.com

Received: 25-06-2023	Revised: 09-07-2023	Accepted: 22-07-2023

ABSTRACT
Data mining refers to extract and identify useful information from large sets of data. This term is really a misnomer. Thus, data mining should be named as knowledge mining which rely stress on mining from vast sets of data . An enormous quantity of data is present in the information industry. This data is meaningless until it is converted into useful form of information or help the industries in their business. It is essential to analyze this plenty of data and extract the valuable information from it. In data mining, extraction of information is not only the process to be performed it also involves various other process such as cleaning, integration, data transformation, data mining, pattern evaluation and presentation. When all these processes are completed one will be able to use this valuable information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.(Dhaka et al.,2018)
This paper introduces the significance use of data mining techniques such as clustering, association-rules, sequential pattern, statistics analysis, characteristics rules and so on can be used to find out the useful knowledge. Finally, various tools has been explained in this paper.

Keywords: *data mining, data mining methods, tools & techniques*

I. INTRODUCTION TO DATA MINING

Collection of data is the prerequisite for generation of information. The nature of data is not always specific; it could be simple text documents, numerical figures or could be more complex information such as spatial data, hypertext documents and multimedia data. But simply data retrieval cannot serve the purpose alone, the data has to summarized, analyzed and information has to be extracted. When the data is enormous it becomes impossible to do the above mentioned tasks manually, hence the need for powerful tool for analysis and interpretation of such huge volume of data stored in repositories like online databases and for the extraction of required information paves in. All these requirements are met by "Data Mining".

Kantardzic & Mehmed (2003) has described data mining as a bridge between applied statistics, artificial intelligence and data base management to discover hidden patterns in stored large data set more efficiently using feasible algorithms.

Deshpande & Thakre, 2010 stated data mining as a tool for the extraction of hidden predictive information from large databases and stated it is as a powerful technology possessing great potential which would help organizations in different sectors to extract fruitful information from the huge lot of data generated in this information technology era.

Mena, Jesús (2011) in their book has mentioned the origin of the term "data mining". The origin of term data mining in the database community dates back to somewhere around 1990. In 1980s, HNC, a San Diego-based company trademarked a phrase "database mining"™, for pitching their Database Mining Workstation. Being trademarked, researchers had to use an alternate to term and so the term "data mining" came into existence.

Han et al. 2012 defined data mining as ''data mining is a process of discovering or extracting interesting patterns, associations, changes, anomalies and significant structures from large amounts of data which is stored in multiple data sources such as file systems, databases, data warehouses or other information repositories.'' The authors have stated various functions of data mining such as a summarization, characterization and discrimination, association, clustering, classification, outlier analysis, regression and trend analysis, etc.

In yet another study Arora and Gupta, 2017 stated data mining to be an essential step in knowledge discovery in databases which have potential to be used for discovery of useful unknown patterns from large repository of data making use of various functionalities, techniques and algorithms.

Thus, data mining represents a useful method for the analysis of huge data to draw out important information which assist in decision making in many spheres of businesses.

ISSN (Online): 2583-2468
Volume-2 Issue-4 || July 2023 || PP. 13-17 DOI: 10.54741/asejar.2.4.3

II. DATA MINING METHODS

Many authors have described different methods deployed in data mining which are described below:

a) Summarization

Chen et al, 1996 have reported that based on concept hierarchy, summarization method results into a comprehensive summary of the data.

In another study Hung et al, 2015 reports that summarization method is carried out using aggregation leading to different levels of abstraction and thus could be viewed and analyzed from many angles. Summarization method helps in extraction of different kinds of patterns that could be extracted using attributed oriented induction approach and data cube approach.

b) Characterization and Discrimination

Bhatnagar et al, 2015 have described characterization and discrimination method used in data mining. Characterization generally comprises of summarization of data based concept hierarchy for generation of characterization rules as output whereas discrimination is a method to identify the varieties among various data sets and generation of discriminant rules as output.

c) Classification

Liao & Triantaphyllou, 2007 has reported a vast variety of classification algorithms also known as classifiers which has also been proposed in earlier literature. Some popular classification algorithms as reported by the author are rough set approach, genetic algorithms, semi-supervised learning, fuzzy sets and active learning.

Han et al, 2012 has defined classification as the process to classify new observation based on the predetermined classes, i.e. supervised learning using a classification algorithm to forecast classes of the data.

d) Clustering (or Cluster Analysis*)*

Algergawy et al. 2011, has reported Clustering as an efficient method in data mining which helps in segmenting or partitioning of observations or data into subsets termed as groups or clusters. Like classification, clustering classifies the similar data objects into the same group but unlike classification, the class labels are unknown. This technique serves its function not just in data mining but also in image segmentation, statistics, pattern recognition, information retrieval, object recognition, bioinformatics etc.

Table 1: Many clustering approaches have been suggested by many researchers some of them are summarized

Clustering Methods	Author and Year
Parameter free method using minimum description length approach	Mampaey & Vreekan 2011
Parallelized hierarchical clustering approach	Wang, 2011
Gene expression data clustering approach based on z-score measure	Das et al, 2011
Fully automatic clustering algorithm for high dimensional categorical data	Bouguessa 2013
Nature inspired swarm based Intelligent Water Drops—K-Means (IWD-KM) algorithm	Shah- Hosseini 2013
Voronoi diagram based clustering algorithm for artificial as well as biological data	Sawant and Shah, 2013
Domain knowledge based density-based clustering	Jin et al, 2014
Algorithm for clustering large-scale data sets based on the unique combination of matrix decomposition and low-rank matrix approximation named as exemplar-based low-rank sparse matrix decomposition (EMD)	Wang, 2014
Bisect K-means clustering algorithm	Abuaiadah, 2015
A three-phased cluster ensemble method based on discriminant analysis	Bhatnagar et al, 2015

e) Outlier Analysis

Outliers are the data objects which are generally discarded by data mining methods as noise because they differ from the general behavior of the data. But it has been reported by Han et al., 2012 that these outliers may sometimes have more information when compared to other data objects and hence mark their importance in fraud detection, anomaly detection,

ISSN (Online): 2583-2468
Volume-2 Issue-4 || July 2023 || PP. 13-17 DOI: 10.54741/asejar.2.4.3

intrusion detection etc. The outlier detection methods has been further classified as statistical methods, clustering-based methods, classification-based methods, deviation-based methods supervised, semi-supervised and unsupervised methods and proximity-based methods.

Other than this for generalization and unification of statistical outliers, Angiulli and Fassetti, 2013 investigated gradient outlier factor. In yet another study by Campello et al. 2015, a detailed method for density-based outlier detection was investigated.

f) Association Analysis

Association analysis identifies all the frequent items in the data sets that follow a strong association rule that satisfies a minimum support and confidence thresholds. Ceglar & Roddick, 2006 has classified association analysis algorithms as classical algorithms, condensed representation algorithms and incomplete set algorithms.

Some popular association mining algorithms as mentioned by the author are Apriori which is based on confined candidate generation, FP-growth based on conditional pattern base without candidate generation and Eclat (equivalence class transformation) based on data transformation and candidate generation.

g) Regression and Trend Analysis (or Evolution Analysis)

Tan et al, 2009 has reported that the prediction of the value of attribute based on regression techniques using historical time series plot is another important method in data mining. Regression and Trend analysis which is also termed as evolution analysis is vital method in discovery of interesting patterns in object's evolution and matching of the objects' changing trends.

III. DATA MINING TECHNIQUES

There are numerous data mining techniques that have been investigated by researchers so far. The major ones are reported below:

a) Statistical Approaches

Data mining in general has an inherent correlation with statistics. Many statistical analysis tools has been reported by Jackson, 2002 which are widely used in data mining such as discriminant analysis, Bayesian network, factor analysis, cluster analysis, regression analysis, correlation analysis, etc.

b) Machine Learning

Padhraic, 2002 has mentioned the emerging need for automation in knowledge discovery from databases to improve accuracy and efficiency of the data mining process which has led to emergence of data mining algorithms in context to machine learning. Thus these techniques are very useful in many sectors like education and transportation.

RSA (Rough Set Analysis) and DNA (Dependency Network Analysis) have been suggested by Gengshen and Guenther, 2014 as important technique in machine learning.

c) Artificial Neural Networks

Liao et al, 2012 has stated Artificial Neural Network (ANN) to be a system of artificial neurons or nodes and electrical signaling which has resemblance to biological neural network. ANN represents knowledge as neurons which are layered set of interconnected processors and are often used to solve critical research problems, spatial environmental data analysis, and also play an important role in modern operations research tool.

d) Database Systems and Data Warehouses

Chandra and Gupta, 2018 reported the use of database-oriented and data warehouse-oriented approaches of data mining techniques to handle huge data sets so as to increase the effectiveness as well as scalability of data mining task. The multidimensional nature of data structure in data warehouse technique has also been reported by the author for its exceptional use in multidimensional data mining.

e) Genetic Algorithms

Jain et al, 2013 has defined genetic algorithm as a concept based on natural biological evolution which includes various processes like reproduction, selection, mutation, and survival of the fittest. But this technique is often considered confusing as there is no statistical measure and thus making it difficult for the user to understand the logic behind a particular solution.

ISSN (Online): 2583-2468
Volume-2 Issue-4 || July 2023 || PP. 13-17 DOI: 10.54741/asejar.2.4.3

f) Fuzzy Sets

Lotfi Zadeh in 1965 has proposed the concept of Fuzzy set theory as "the degree of membership based on the possibility value calculated with the help of membership function".

Hullermeier, 2011 stated that fuzzy set theory is marking a great value in data mining, machine learning, and many other related fields especially in classification and cluster analysis.

In 2012, Edward and Olgierd proposed combined technique including machine learning and Fuzzy set theory for enhancing the data mining process for various applications.

g) Visualization

Yahia 2010 stated visualization as a very important data mining technique as it helps in translation of data into objects which could be displayed in 2D or 3D space such as area, line, point etc.

Campello et al.2015 also have recognized the usefulness of this technique in data mining and have presented a framework for density-based estimates for visualization.

4. Data Mining Tools

Ranga & Bansal, 2006 conducted a study describing the specialisation, technical specifications & features of the six widely used open source data mining tool which could help in choosing and selecting appropriate tool for the purpose.

Table 2: The features of the six data mining tools mentioned in the study are summerised

S.No.	Tool Name	Release date	Category	Operating system & Language	Website
1.	RAPID MINER	2006	Statistical analysis, data mining, predictive analytics	Cross platform; Language Independent	www.rapidminer.com
2.	ORANGE	2009	Machine learning, Data mining, Data visualization	Cross Platform; Python C++, C	www.orange.biolab.si
3.	KNIME	2004	Enterprise Reporting, Business Intelligence, Data mining	Linux, OS X, Windows; Java	www.knime.org
4.	WEKA	1993	Machine Learning	Cross Platform; Java	www.cs.waikato.ac.nz/-ml/weka
5.	KEEL	2004	Machine Learning	Cross Platform; Java	www.sci2s.ugr.cs/keel
6.	R	1997	Statistical Computing	Cross Platform; C, Fortran and R	www.r-project.org

IV. CONCLUSION

In this paper revision of literature of Data Mining is presented. This study gives the idea about various Data Mining Techniques, different methods, different processes and data mining tools. In future we tend to review and compare various Data Mining algorithm's used in Data Mining.

REFERENCES

1. R. Dhaka, & A. Kumar (2018). Need and application of data mining. *International Journal of Innovations & Advancement in Computer Science*, *7*(4), 166-169.
2. Kantardzic, Mehmed. (2003). *Data mining: Concepts, models, methods, and algorithms*. John Wiley & Sons. ISBN 978-0-471-22852-3.
3. Mena, Jesús. (2011). *Machine learning forensics for law enforcement, security, and intelligence*. Boca Raton, FL: CRC Press (Taylor & Francis Group). ISBN 978-1-4398-6069.

ISSN (Online): 2583-2468
Volume-2 Issue-4 || July 2023 || PP. 13-17 DOI: 10.54741/asejar.2.4.3

4. Deshpande, Shrinivas & Thakare, V. M. (2010). Data mining system and applications: A review. *International Journal of Distributed and Parallel systems*. doi: 1.10.5121/ijdps.2010.1103.
5. Han J, Kamber M, & Pei J (2012). *Data mining concepts and techniques*. (3rd ed.). Elsevier, Netherlands.
6. Arora RK, & Gupta MK (2017). e-Governance using data warehousing and data mining. *Int J Comput Appl, 169*(8), 28–31.
7. Chen M, Han J, & Yu PS. (1996). Data mining: an overview from a database perspective. *IEEE Trans Knowl Data Eng, 8*(6), 866–883.
8. Hung LN, Thu TNT, & Nguyen GC. (2015). An efficient algorithmin mining frequent itemsets with weights over data stream using tree data structure. *IJ Intell Syst Appl, 12*, 23–31.
9. Bhatnagar V, Ahuja S, & Kaur S. (2015). Discriminant analysis based cluster ensemble. *Int J Data Min Model Manag, 7*(2), 83–107.
10. Liao TW, & Triantaphyllou E. (2007). Recent advances in data mining of enterprise data: algorithms and applications. *World Scientific Publishing, Singapore*, 111–145.
11. Algergawy A, Mesiti M, Nayak R, & Saake G. (2011). XML data clustering: an overview. *ACM Comput Surv, 43*(4), 1–25.

.