IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, September 2022
Impact Factor: 6.252

Data Mining and Big Data - A Review

Abhishek R Bhat, Abhishek S V, Akash K Acharya, Amruth P S
Students, Department of Information Science and Engineering
Alva's Institute of Engineering and Technology, Mangalore, Karnataka, India

Abstract: *A tremendous amount of information has been converted into a digital format due to technological advancement, particularly in the last thirty years or more. This has led to the creation of massive data repositories. As information accumulated in these archives, the problem of how to extract useful knowledge from it continued. The problem was solved using data mining as a tool. Data mining is a technique for extracting secret information from huge datasets in order to uncover insightful patterns and rules. It is often seen as a stepping stone to the knowledge discovery process in databases. Nowadays, data mining is a necessary component in practically all facets of human life. The current article offers a review of the data mining literature that is currently available. The concept of data mining as well as its various methodologies are summarized. Some applications, tasks and issues related to it have also been illustrated. In the digital era like today the growth of data in the database is very rapid, all things related to technology have a large contribution to data growth as well as social media, financial technology and scientific data. Therefore, topics such as big data and data mining are topics that are often discussed. Data mining is a method of extracting information through from big data to produce an information pattern or data anomaly.*

Keywords: Data mining; dataset; database; big data.

I. INTRODUCTION

The availability of a sizable amount of data in practically every discipline and the need to extract useful knowledge and information from it served as the primary driving force behind researchers' recent interest in data mining. Applications ranging from small business administration to sophisticated engineering design to scientific research can greatly benefit from the information and knowledge gathered. Data mining is the examination and analysis of enormous data sets with the goal of revealing important patterns and previously undetected laws. The main objective is to combine computer data processing power with human pattern recognition abilities.

Five main components make up data mining:

- Extraction, transformation, and loading of transaction data onto the system of the data warehouse.
- Use a multidimensional database system to store and manage the data.
- Give business analysts and information technology specialists access to data.
- Utilize application software to analyse the data.
- Provide the data in an understandable format, such as a table or graph. [2]

II. DATA ATTRIBUTES

The categories below can be determined by the type of data:

2.1 Nominal Attributes

"Relating to names" is what nominal means. A nominal attribute's values are items' symbols or names. Nominal attributes are sometimes known as categorical attributes since each value reflects a certain category, code, or state. There is no logical order to the values. Values are also known as enumerations in computer science.

2.2 Binary Attributes

A binary attribute is a nominal attribute with only two possible states or categories: 0 or 1, where 1 denotes presence and 0 indicates absence. If the two states of a binary attribute are true and false, the attribute is said to be Boolean.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, September 2022
Impact Factor: 6.252

2.3 Binary Attributes

An attribute with multiple possible values that can be ranked or ordered meaningfully, but whose magnitude between subsequent values is unknown, is known as an ordinal attribute.

III. BIG DATA

Big data is data that has a wider diversity and comes in larger volumes and at a faster rate. The three Vs are a name for this.

3.1 Volume

The volume of data is important. You'll need to process large amounts of low-density, unstructured data when working with big data. This can be unvalued data from sources like Twitter data feeds, clickstreams from websites or mobile apps, or sensor-enabled hardware. This amount of data may reach tens of gigabytes for some corporations. Others might need several hundred petabytes.

3.2 Velocity

Velocity refers to how quickly data is received and (perhaps) used. In contrast to being written to disc, the highest velocity of data often streams straight into memory. Some internet-enabled smart goods function in real time or almost real time, necessitating real-time analysis and decision-making.

3.3 Variety

Variety alludes to the wide range of data types that are accessible. In a relational database, traditional data kinds were organised and easily suited. Data now comes in new unstructured data formats thanks to the growth of big data. Text, audio, and video are examples of semistructured and unstructured data types that require further preprocessing to create meaning and enable metadata. [6]

III. ALGORITHMS USED IN DATA MINING FOR BIG DATA

3.1 Classification Trees

A common method for categorising dependent categorical variables using measurements of one or more predictor factors. As a result, a tree containing nodes and relationships between them that can be interpreted to generate if-then rules is produced.

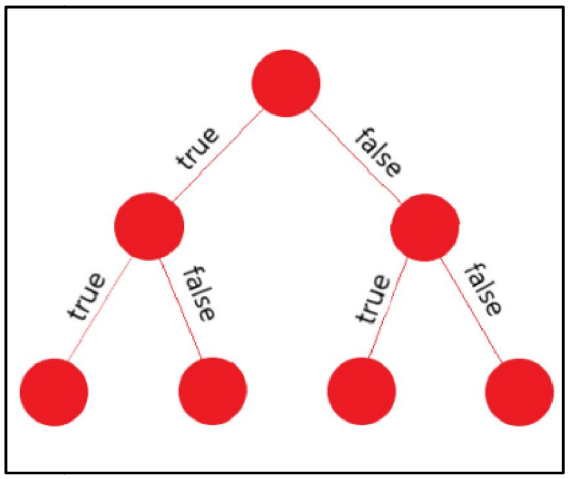

Figure 1: Classification tree

3.2 Logistic Regression

A statistical method that extends the idea of conventional regression to cope with classification. It generates a formula that estimates the likelihood of the event in relation to the independent variables. [3]

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, September 2022
Impact Factor: 6.252

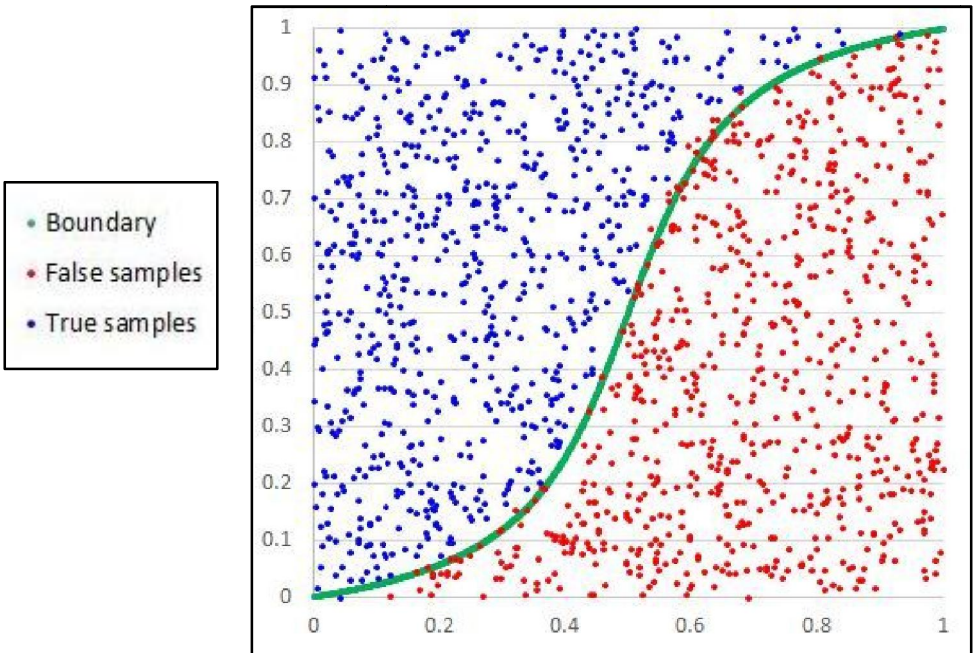

Figure 2: Logical Regression

3.3 Neural Networks

A computer programme inspired by the parallel structure of animal brains. Input nodes, hidden layers, and output nodes make up the network. A weight is put on each unit. The algorithm modifies the weights through a system of trial and error until it satisfies a predetermined stopping criterion. Data is provided to the input node. This has been compared by some to a black-box strategy.

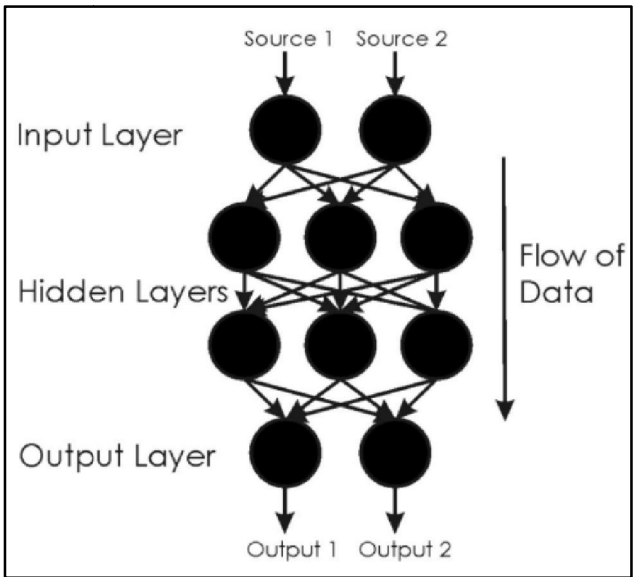

Figure 3: Neural Networks

3.4 Clustering techniques like K-Nearest Neighbours

A method for locating collections of related records. The K-nearest neighbour method determines the separations between points in the historical (training) data and the record. After that, it places this record in a data set under the classification of its closest neighbour. [4]

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, September 2022
Impact Factor: 6.252

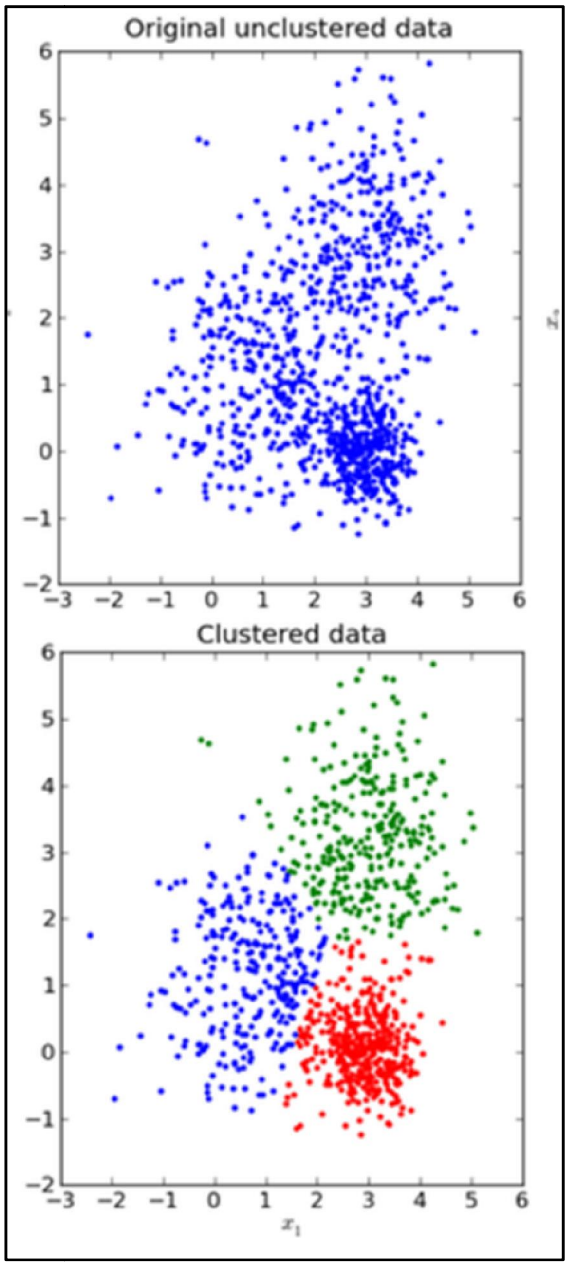

Figure 4: Clustering techniques like K-nearest neighbours

IV. CHALLENGES AND ISSUES

4.1 Challenges

Researchers and developers have several requirements and significant hurdles in order to do efficient and effective data mining in massive databases.

Data mining methods, user engagement, performance and scalability, as well as the processing of a wide range of data kinds, are the challenges at hand. The investigation of data mining applications and their societal effects is one of the other topics. [5]

4.2 Issues

A. Poorness of Information

- The situation has been defined as being data rich but information poor due to the volume of data and the requirement for strong data analysis tools.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, September 2022
Impact Factor: 6.252

- Large data repositories act as "data tombs" for the data they collect.
- databases with infrequently used archives [7]

B. Decision Making

- Important judgments are frequently made using a decision maker's gut instinct rather than the knowledge-rich data contained in data repositories.
- The decision-maker lacks the tools necessary to access the useful information concealed in the massive volumes of data.

C. Data Entry

- System knowledge bases are frequently filled out manually by users or subject matter experts.
- Unfortunately, this method is very time-consuming, expensive, and subject to biases and errors.

D. Bad Nomenclature

- The expression is a misnomer.
- Instead of being called rock or sand mining, the extraction of gold from rocks or sand is referred to as gold mining.
- The more proper name for data mining would have been "knowledge mining from data," which is regrettably a bit lengthy. [6]

V. CONCLUSION

The complexity of big data will continue to rise for individuals who struggle with it as a result of the extremely rapid growth of both big data and data mining research. As a result, we draw the conclusion that data mining techniques still have room for improvement. Due to the numerous issues that still exist and are experienced, the potential is still very much a possibility.

Today, big data is the focus of all IT professionals, engineers, and researchers. Big data is a term used to describe massive amounts of complicated data collections. Numerous scholars suggested various system models and big data strategies to address the challenges brought by big data. In order to use data mining to address the issue of large data, the high-performance computing paradigm is necessary. We come to the conclusion that there is still room for advancement in data mining methods and methodologies. In this essay, we discuss the numerous problems and obstacles that big data face and offer solutions.

REFERENCES

[1] Anusha Prem, & P. Jayanthi. (2016). BIG DATA SOURCES AND DATA MINING. International Education and Research Journal (IERJ), 2(4). Retrieved from http://ierj.in/journal/index.php/ierj/article/view/243

[2] Zhao, Kaidi& Liu, Bing &Tirpak, T.M. & Xiao, Weimin. (2005). A visual data mining framework for convenient identification of useful knowledge. 8 pp.. 10.1109/ICDM.2005.16.

[3] Fayyad, U.M., Gregory, P.S., Padhraic, S.: From Data Mining to Knowledge Discovery: an Overview. In: Advances in Knowledge Discovery and Data Mining, pp. 1–36. AAAI Press, Menlo Park (1996)

[4] Bharati, M. &Ramageri, Bharati. (2010). Data mining techniques and applications. Indian Journal of Computer Science and Engineering. 1.

[5] S. K. Khatri, "Intrusion detection using Data Mining," 2014 Conference on IT in Business, Industry and Government (CSIBIG), 2014, pp. 1-2, doi: 10.1109/CSIBIG.2014.7056926.

[6] Che, Dunren& Safran, Mejdl& Peng, Zhiyong. (2013). From Big Data to Big Data Mining: Challenges, Issues, and Opportunities. Proc Of DASFAA-BDMA Workshop. 7827. 1-15. 10.1007/978-3-642-40270-8_1.

[7] X. Wu, X. Zhu, G. -Q. Wu and W. Ding, "Data mining with big data," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014, doi: 10.1109/TKDE.2013.109