Data mining: Techniques and Applications

Riya Chauhan[1], Anupama Chauhan[2], Bharti Jha[3], Vinay Jain[4]
[1,2,3,4] *Computer Science and Engineering, Manav Rachna University, Faridabad, India*

***Abstract*- Big data refers to data sets that consist of a large amount of data. Most of the data in these sets are unstructured and complex to understand. The data sets are so large that traditional storage and processing techniques are insufficient to convert it into information or other forms for various requirements. In this paper, we have discussed about various data mining techniques that can be applied on large data sets to get the required results. We have also discussed the various fields where these data mining techniques can be applied to obtain solutions.**

Index Terms- data mining, knowledge discovery, applications, techniques.

I. INTRODUCTION

Data plays a major role in day to day applications and technologies. Everyday, a large amount of data is generated from various sources like micro-blogs, e-commerce sites, social media sites, companies, research results etc. This data is scattered randomly and makes no useful sense until data processing is done and techniques are applied to convert it to useful information. But applying traditional data processing and data analysis techniques on such a huge volume of data is a very lengthy and tedious task. It might not always be possible to convert the data using the traditional methods.

For big data sets we use a computing technique known as Data Mining. Through data mining we can analyze data from different perspectives and classify it into new useful information which is easier to understand and put into use. Data mining helps in formulating a solution for many real life problems where a result has to be obtained from big data sets.

II.DATA MINING STAGES

The data on microblogging sites is very large and classifying it is very difficult.

Only Twitter, for example, contains a huge amount of tweets which increase rapidly within minutes. There are many types of users such as politicians, company employees, professors, businessmen, celebrities, news channels, students, and common man. [1]

Also, these users may belong from different countries. Further, they may post their views in different languages and so on. [2] With increasing users, the diversity increases and there is more variation in the opinions on various topics. Analysis of such type of data can be done through data mining. [3]

Data mining consists of the following stages:

1. Business understanding
2. Data Understanding
3. Data Preparation
4. Data Modelling
5. Data Evaluation
6. Data Deployment

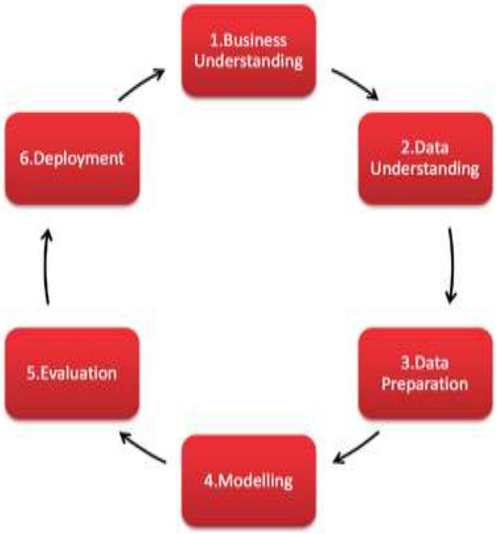

Fig. 1. Stages in Data Mining

• Business understanding

We must first need to examine the requirements and objectives of the project from a business perspective, and then convert this knowledge into data mining problem definition and a plan designed to achieve the objectives.

• Data Understanding

We may not be able to use all the data we have collected in the first step. Therefore, we must select only those data which we think useful for data mining.

•Data Preparation

The data we have collected is not clean and may contain errors, missing values, noisy or inconsistent data. For using the extracted data efficiently, we need to apply different techniques to get desired data.

•Data Modelling

The data even after cleaning is not ready for mining as we need to transform them into forms appropriate for mining. The techniques that are used to achieve this are smoothing, aggregation, normalization and so on. Now we are ready to apply data mining techniques on the data to discover the interesting patterns. Data mining techniques like clustering and association analysis are among the many different techniques that can be used to mine data and conclude results.

•Data Evaluation

This step includes the visualization, transformation, and also removing redundant patterns from the patterns that we have generated from the data.

•Data Deployment

This step helps the user or organization to make use of the knowledge acquired to take better decisions.

III.LITERATURE REVIEW

Data mining techniques have been widely used in the past to solve a variety of problems including health care problems, banking, sales, student performance, bioinformatics, e-learning systems and so on. The following table includes the various research papers related to the applications of data mining in solving problems which have been published in the years 2015-2018.

TABLE I
Research Papers Related to Data Mining And its Applications: 2015-2018

S.No.	Name of paper published	Techniques used
1.	Agrawal, S., & Agrawal, J. (2015). Survey on anomaly detection using data mining techniques.[7]	Clustering based, classification based anomaly detection techniques, hybrid approaches.
2.	Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection.[12]	CRISP-DM Model, Association Rules and Fuzzy Association Rules, Anomaly Detection and Hybrid Detection, Bayesian Network
3.	Chavan, S., Jadhav, A., Suryagandh, P., & Sharma, N. (2018). Data Mining Techniques to Improve Customer Relationships Management.[14]	K-means clustering Algorithm, Customer Retention Program, Integrating CRM and Data Mining technologies
4.	Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015).Data mining for the internet of things: literature review and challenges.[9]	Clustering, Classification, Association, Bayesian Network, Decision Trees, Time Series Analysis, Challenges and Open Research Issues in IoT and Big Data Era
5.	Gera, M., & Goel, S. (2015). Data mining-techniques, methods and algorithms:A review on tools and their validity.[6]	Data Mining Techniques, Tools for Data Mining Techniques and their features, Orange, WEKA, SCaVis, Apache Mahout, R Software Environment
6.	Jothi, N., & Husain, W. (2015). Data mining in healthcare–a review. [11]	Data mining models, Linear discriminant analysis (LDA), Swarm Intelligence, Logistic regression (LR)
7.	Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining.[13]	Classification, association,clustering,prediction, outlier analysis, association rule mining, ARM-Predictor Algorithm
8.	Lim, K. C., Selamat, A.,	Issues with current post evaluation analysis,

	Alias, R. A., Zabil, M. H. M., Puteh, F., & Mohamed, F. (2018). Measuring the Feasibility of Clustering Techniques on Usability Performance Data. [15]	clustering algorithms HCA and K-means cluster population
9..	Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques.[10]	Decision trees, neural networks, Naive Bayes, K-Nearest Neighbour, Support Vector Machine
10.	S. Moro. Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Systems with Applications.[8]	A recent literature review on application of mining techniques in Banking domain, latent Dirichlet allocation.

IV.DATA MINING TECHNIQUES

There are various data mining techniques that can be used to analyze different formats and types of data. Some of these techniques are association, classification, clustering, regression, etc. We have discussed these data mining techniques below.

- Association:

Association is a data mining technique that examines and finds out the probability of the co-occurrence of elements/items in a collection or group. The relationships between co-occurring items are known as association rules. Association rules are commonly used to analyze the sales transactions. [6]

- Classification:

Classification is one of the data mining techniques which is based on machine learning. [6] It simply classifies data into groups. Basically, classification is used to classify each item belonging to a particular data set into the predefined set of classes or groups. Classification method makes efficient use of mathematical techniques such as decision trees, linear programming, neural network, and so on. In classification, we develop the software that can classify the data items into groups.

Further, there are four major methods of classification: Decision tree induction, rule – based classification, classification by back propagation and lazy learners.

- Clustering:

Clustering is a data mining technique which involves grouping of a specific set of items on the basis of their qualities and characteristics, and accumulating them on the basis of their similarities.

There are two types of clustering: hard clustering and soft clustering. [6]

Hard clustering: In hard clustering, same item can only belong to single cluster.

Soft clustering: In this clustering, same item can belong to different clusters.

Some of the major clustering algorithms are:

1.Centroid-based
2.Distributed-based
3.Connectivity-based
4.Density-based

- Regression:

Regression is a data mining strategy used to anticipate a scope of numeric qualities (called consistent qualities), given a specific dataset. For instance, regression may be utilized to foresee the cost of an item or administration, given different factors.

Regression is utilized over numerous ventures for business and advertising arranging, money related determining, natural demonstrating and examination of patterns.

V.DATA MINING APPLICATIONS

Data mining techniques are used by many companies to focus on the consumers. These companies may include retail, marketing, financial, communication, brand etc to analyze their consumers' preferences.

Through data mining, they can determine the pricing of products, the interest of consumers in buying a particular product, customer feedback and satisfaction and the company's profits as well.

Data mining has many applications out of which a few have been listed below.

- Healthcare

Data mining has an incredible potential to enhance medicinal services frameworks. It utilizes the data to analyze the most efficient practices that can improvise care and reduce costs. Many researchers make use of various data mining approaches like big databases, soft computing, machine learning, pattern generation and statistics. Mining can also be used to foresee the amount of patients in each category. For this purpose, numerous procedures and methods have been produced which ensure that the patients get the required care at the ideal place and at the ideal time.

- Market Basket Analysis

Market basket analysis is one of the approaches that depends on a hypothesis which expresses that if we purchase a specific group or type of items we are considerably more likely to purchase some other type of items. This procedure may enable the retailer to analyze the purchase inclinations of a customer. This data might also help the retailer to understand the buyer's needs and keep up the store accordingly.

- Education

There is another developing field, called Educational Data Mining. The objectives are distinguished as anticipating the students' future learning choices and interests, examining the impacts of educational help, and progressing logical information about learning. Data mining can be utilized by an association to make accurate choices and furthermore to foresee the results of students. With the outcomes the institution can center around what to teach or impart and how to educate also. [20]

- Customer Relationship Management

Customer Relationship Management is related to getting customers, additionally enhancing customers' faithfulness and improving customer centered systems. To keep up an appropriate relationship with a customer a business has to gather data and analyze the information that it contains. Here, date mining has a significant role. With data mining techniques the gathered date can be utilized for many kinds of analysis. a considerable amount of money has been lost because of frauds.

- Fraud Detection

Conventional strategies for fraud discovery are tedious and complex. Data mining gives significant and meaningful patterns and transforming data to information. Knowledge can be termed as any information that is legitimate and helpful. An efficient fraud detection framework ought to protect the data as well as information of all the clients. An administered technique incorporates accumulation of test records. These records are then characterized as fraudulent or non-fraudulent. A model is built utilizing this data and the calculation is made to recognize whether the record is fraudulent or not.

- Financial Banking

With computerized banking at most of the places, a huge amount of data is supposed to be generated with new transactions. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately accessible to managers because the volume data is too large or is generated too quickly to analyze by experts. The managers may find these information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer.

- Corporate Surveillance

Corporate Surveillance is the checking of a person or group's behavior by a corporation. The data gathered is frequently utilized for advertising purposes or sold to different companies, but on the other hand is consistently imparted to government agencies. It can be utilized by the business to tailor their items liked by their clients. The data can be utilized for coordinate promoting purposes, for example, focus on commercials on Google and Yahoo, where advertisements are focused to the client of the web search engine by analyzing their search history and emails.

- Research Analysis

We have seen progressive changes in research. Data mining is useful in data cleansing, data pre-handling and combination of databases. The researchers can find a lot of similar data from the database that might bring any change in the researches that have been previously conducted. Distinguishing proof of any co-happening groupings and the relationship between any exercises can be known. Data perception and visual data mining give us a reasonable perspective of the data.

- Criminal Investigation

Criminology is a procedure that intends to recognize criminal characteristics. In reality crime investigation incorporates investigating and distinguishing crimes and their associations with criminals. Text based crime reports can be changed over into word preparing records. This data can be utilized to perform crime coordinating procedure.

- Bio Informatics

Data Mining approaches seem, to be ideally suited for Bioinformatics [4], since it is data-rich. Mining normal data expels important learning from large datasets related to science, and to other related life sciences zones, for instance, medication and neuroscience. Uses of data mining in bioinformatics incorporate quality discovering, protein work surmising, malady determination, sickness visualization, ailment treatment enhancement, protein and quality cooperation arrange reproduction, data purging, and protein sub-cell area expectation.[4][5]

REFERENCES

[1] Eman M.G. Younis, “ Sentiment Analysis And Text Mining For Social Media Microblogs Using Open Source Tools: An Empirical Study”, International Journal Of Computer Applications(0975 – 8887), February 2015, Volume 112 – No. 5

[2] Manu Krishna Bhardwaj, Brajesh Kumar, “Opinion Mining Of Social Media Data Using Machine Learning Techniques‖”, International Journal of Scientific Engineering and Applied Science (IJSEAS) ,Volume-2, Issue-5, May 2016, ISSN: 2395-3470.

[3] Patil Monali S1, Kankal Sandip, “A Concise Survey on Text Data Mining”, International Journal of Advanced Research in Computer and Communication Engineering,Vol. 3, Issue 9, September 2014.

[4] Sushmita Mitra, Tinku Acharya “ Data Mining Multimedia, Soft Computing, and Bioinformatics”.

[5] Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C., & Greene, C. S.(2015). Recent advances and emerging applications in text and data mining for biomedical discovery. Briefings in bioinformatics, 17(1), 33-42.

[6] Gera, M., & Goel, S. (2015). Data mining-techniques, methods and algorithms:A review on tools and their validity. International Journal of Computer Applications, 113(18).

[7] Agrawal, S., & Agrawal, J. (2015). Survey on anomaly detection using data mining techniques. Procedia Computer Science, 60, 708-713.

[8] Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Systems with Applications, 42(3), 1314-1324.

[9] Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015).Data mining for the internet of things: literature review and challenges. International Journal of Distributed Sensor Networks, 11(8), 431047.

[10] Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. Procedia Computer Science, 72, 414-422.

[11] Jothi, N., & Husain, W. (2015). Data mining in healthcare–a review. Procedia Computer Science, 72, 306-313.

[12] Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153-1176.

[13] Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining. Procedia computer science,85, 78-85.

[14] Chavan, S., Jadhav, A., Suryagandh, P., & Sharma, N. (2018). Data Mining Techniques to Improve Customer Relationships Management.

[15] Lim, K. C., Selamat, A., Alias, R. A., Zabil, M. H. M., Puteh, F., & Mohamed, F. (2018). Measuring the Feasibility of Clustering Techniques on Usability Performance Data. Indian Journal of Science and Technology, 11(4).