International Journal of Advanced Technology in Engineering and Science
Vol. No. 10, Issue No. 06, June 2022
www.ijates.com
ijates
ISSN 2348 - 7550

Survey Paper on "Data Mining: Techniques and Applications"

Mrs. Ashwini J. Sawakhande1, Ms.Ashvini A. Todkar2

1(Department Of Computer Engineering, ATS's SBGI, Miraj, India)

2(Department Of Computer Engineering, ATS's SBGI, Miraj, India)

Abstract

Data mining is the process of uncovering patterns and finding anomalies and relationships in large datasets that can be used to make predictions about future trends. The main purpose of data mining is to extract valuable information from available data. Data mining is considered an interdisciplinary field that joins the techniques of computer science and statistics. Note that the term "data mining" is a misnomer. It is primarily concerned with discovering patterns and anomalies within datasets, but it is not related to the extraction of the data itself. Data mining is a process which finds useful patterns from large amount of data. This Paper discuss few techniques and applications of data mining.

Key words: *Datamining, Patterns, Dataset, anomalies*

I. Introduction

1.1 Definition of Data Mining

In the context of computer science, "Data Mining" can be referred to as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. It is basically the process carried out for the extraction of useful information from a bulk of data or data warehouses

Nowadays, data mining is used in almost all places where a large amount of data is stored and processed. For example, banks typically use 'data mining' to find out their prospective customers who could be interested in credit cards, personal loans, or insurance as well. Since banks have the transaction details and detailed profiles of their customers, they analyse all this data and try to find out patterns that help them predict that certain customers could be interested in personal loans, etc. Mining of data means the process of discovering interesting patterns and knowledge from large amounts of data. The development of Information Technology has generated large amount of databases and huge data in various areas.

Basically, Data mining has been integrated with many other techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, etc. to gather more information about the data and to helps predict hidden patterns, future trends, and behaviours and allows businesses to make decisions. Technically, data mining is the computational process of analysing data from different perspectives, dimensions, angles and categorizing/summarizing it into meaningful information.

The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

II. Knowledge Discovery Process

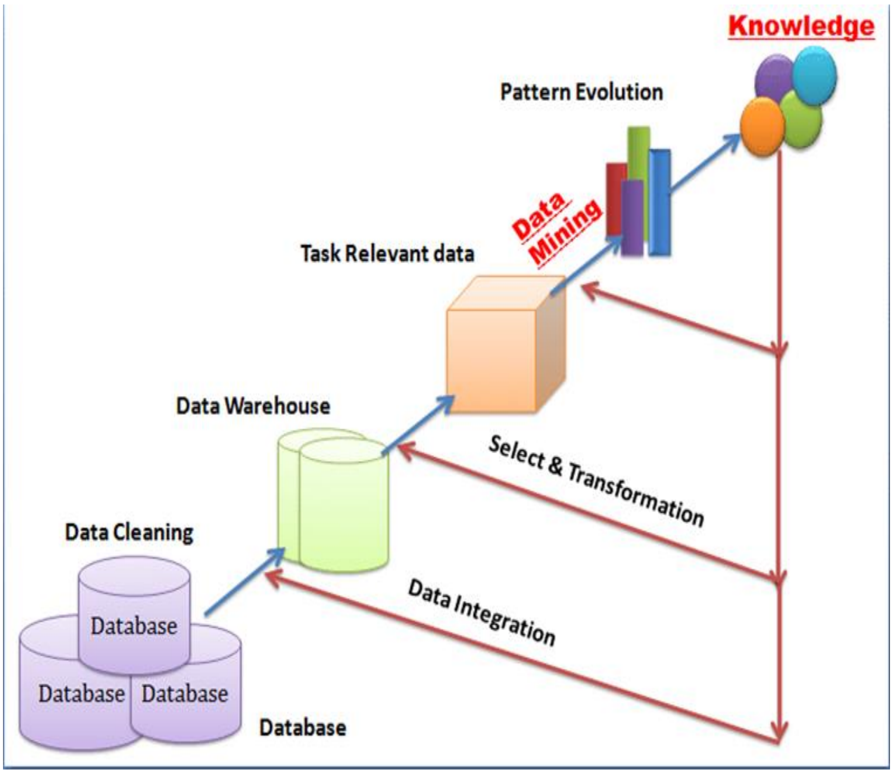

Fig1. Knowledge Discovery Process

Knowledge Discovery process(KDD) is the systematic process of identifying valid, practical, and understandable patterns in massive and complicated data sets. The base of the KDD method is data mining, which involves the inference of algorithms that analyse the data, build the model, and discover previously unknown patterns.

List of steps involved in the knowledge discovery process

2.2 Data Cleaning − In this step, the noise and inconsistent data is removed.

2.3 Data Integration − In this step, multiple data sources are combined.

2.4 Data Selection − In this step, data relevant to the analysis task are retrieved from the database.

2.5 Data Transformation − In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

2.6 Data Mining − In this step, intelligent methods are applied in order to extract data patterns.

2.7 Pattern Evaluation − In this step, data patterns are evaluated.

2.8 Knowledge Presentation- In this step, knowledge is presented.

III. Data Mining Techniques

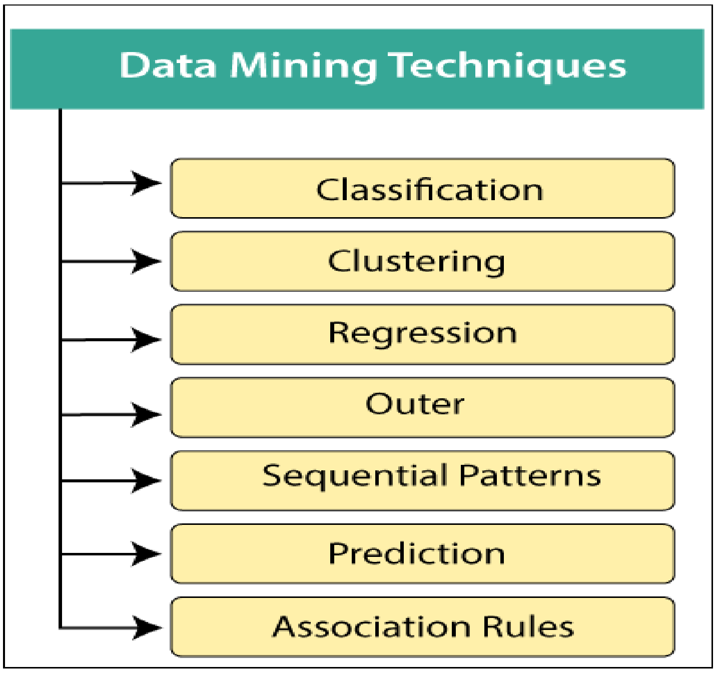

Fig 2: Datamining Techniques

3.1 Classification:

This technique is used to obtain important and relevant information about data and metadata. This data mining technique helps to classify data in different classes.

Data mining techniques can be classified by different criteria, as follows:

3.1.1. Classification of Data mining frameworks as per the type of data sources mined:

This classification is as per the type of data handled. For example, multimedia, spatial data, text data, time-series data, World Wide Web, and so on.

3.1.2. Classification of data mining frameworks as per the database involved:

This classification based on the data model involved. For example, Object-oriented database, transactional database, relational database, and so on.

3.1.3. Classification of data mining frameworks as per the kind of knowledge discovered:

This classification depends on the types of knowledge discovered or data mining functionalities. For example, discrimination, classification, clustering, characterization, etc. some frameworks tend to be extensive frameworks offering a few data mining functionalities together.

3.1.4. Classification of data mining frameworks according to data mining techniques used:

This classification is as per the data analysis approach utilized, such as neural networks, machine learning, genetic algorithms, visualization, statistics, data warehouse-oriented or database-oriented, etc.

3.1.5. The classification can also take into account, the level of user interaction involved in the data mining procedure, such as query-driven systems, autonomous systems, or interactive exploratory systems.

3.2 Clustering:

Clustering is a division of information into groups of connected objects. Describing the data by a few clusters mainly loses certain confine details, but accomplishes improvement. It models data by its clusters. Data modeling puts clustering from a historical point of view rooted in statistics, mathematics, and numerical analysis. From a machine learning point of view, clusters relate to hidden patterns, the search for clusters is unsupervised learning, and the subsequent framework represents a data concept. From a practical point of view, clustering plays an extraordinary job in data mining applications. For example, scientific data exploration, text mining, information retrieval, spatial database applications, CRM, Web analysis, computational biology, medical diagnostics, and much more. In other words, we can say that Clustering analysis is a data mining technique to identify similar data. This technique helps to recognize the differences and similarities between the data. Clustering is very similar to the classification, but it involves grouping chunks of data together based on their similarities.

3.3 Regression:

Regression analysis is the data mining process is used to identify and analyse the relationship between variables because of the presence of the other factor. It is used to define the probability of the specific variable. Regression, primarily a form of planning and modelling. For example, we might use it to project certain costs, depending on other factors such as availability, consumer demand, and competition. Primarily it gives the exact relationship between two or more variables in the given data set.

3.4. Association Rules:

This data mining technique helps to discover a link between two or more items. It finds a hidden pattern in the data set. Association rules are if-then statements that support to show the probability of interactions between data items within large data sets in different types of databases. Association rule mining has several applications and is commonly used to help sales correlations in data or medical data sets.The way the algorithm works is that you have various data, for example, a list of grocery items that you have been buying for the last six months. It calculates a percentage of items being purchased together.

These are three major measurements technique:

3.4.1 Lift:

This measurement technique measures the accuracy of the confidence over how often item B is purchased.

(Confidence) / (item B)/ (Entire dataset)

3.4.2 Support:

This measurement technique measures how often multiple items are purchased and compared it to the overall dataset.

(Item A + Item B) / (Entire dataset)

3.4.3 Confidence:

This measurement technique measures how often item B is purchased when item A is purchased as well.

(Item A + Item B)/ (Item A)

3.5 Outer detection:

This type of data mining technique relates to the observation of data items in the data set, which do not match an expected pattern or expected behavior. This technique may be used in various domains like intrusion, detection, fraud detection, etc. It is also known as Outlier Analysis or Outilier mining. The outlier is a data point that diverges too much from the rest of the dataset. The majority of the real-world datasets have an outlier. Outlier detection plays a significant role in the data mining field. Outlier detection is valuable in numerous fields like network interruption identification, credit or debit card fraud detection, detecting outlying in wireless sensor network data, etc.

3.6 Sequential Patterns:

The sequential pattern is a data mining technique specialized for evaluating sequential data to discover sequential patterns. It comprises of finding interesting subsequence's in a set of sequences, where the stake of a sequence can be measured in terms of different criteria like length, occurrence frequency, etc. In other words, this technique of data mining helps to discover or recognize similar patterns in transaction data over some time.

3.7 Prediction:

Prediction used a combination of other data mining techniques such as trends, clustering, classification, etc. It analyses past events or instances in the right sequence to predict a future event.

IV.Data Mining Applications

List of areas where data mining is widely used −

4.1 Financial Data Analysis
4.2 Retail Industry
4.3 Telecommunication Industry
4.4 Biological Data Analysis
4.5 Other Scientific Applications
4.6 Intrusion Detection

4.1 Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

4.1.1 Design and construction of data warehouses for multidimensional data analysis and data mining.

4.1.2 Loan payment prediction and customer credit policy analysis.

4.1.3 Classification and clustering of customers for targeted marketing.

4.1.4 Detection of money laundering and other financial crimes.

4.2 Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry −

4.2.1 Design and Construction of data warehouses based on the benefits of data mining.

4.2.2 Multidimensional analysis of sales, customers, products, time and region.

4.2.3 Analysis of effectiveness of sales campaigns.

4.2.4 Customer Retention.

4.2.5 Product recommendation and cross-referencing of items.

4.3 Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services −

4.3.1 Multidimensional Analysis of Telecommunication data.

4.3.2 Fraudulent pattern analysis.

4.3.3 Identification of unusual patterns.

4.3.4 Multidimensional association and sequential patterns analysis.

4.3.5 Mobile Telecommunication services.

4.3.6 Use of visualization tools in telecommunication data analysis.

4.4 Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −

4.4.1 Semantic integration of heterogeneous, distributed genomic and proteomic databases.

4.4.2 Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.

4.4.3 Discovery of structural patterns and analysis of genetic networks and protein pathways.

4.4.4 Association and path analysis.

4.4.5 Visualization tools in genetic data analysis.

4.5 Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications −

4.5.1 Data Warehouses and data preprocessing.

4.5.2 Graph-based mining.

4.5.3 Visualization and domain specific knowledge.

4.6 Intrusion Detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection −

4.6.1 Development of data mining algorithm for intrusion detection.

4.6.2 Association and correlation analysis, aggregation to help select and build discriminating attributes.

4.6.3 Analysis of Stream data.

4.6.4 Distributed data mining.

4.6.5 Visualization and query tools.

V. Conclusion

Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology.The use of data mining in enrolment management is a fairly new development. Current data mining is done primarily on simple numeric and categorical data. In the future, data mining will include more complex data types. In addition, for any model that has been designed, further refinement is possible by examining other variables and their relationships. Research in data mining will result in new methods to determine the most interesting characteristics in the data. As models are developed and implemented, they can be used as a tool in enrolment management.

VI. ACKNOWLEDGEMENT

There have been many contributors for this to take shape and authors are thankful to each of them.

REFERENCES:

[1].Lei xu, Chunxio Jiang, Jian Wang, Jian Yuan, Yong Ren. Information Security in Big Data: Privacy and Data Mining, DOI 10.1109/ACCESS.2014.2362522,IEEE Access.

[2]. Xiaohua Hu, ‘DB-Reduction: A data pre-processing algorithm for data mining applications’,2003

[3]. J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques. Morgan kaufmann, 2006.

[4]. Mrs. Bharati M. Ramageri. Data Mining Techniques and Applications, Indian Journal of Computer Science and Engineering VOL. 1 No. 4 301-305.

[5]. TipawanSilwattananusarn and Assoc.Prof. Dr.KulthidaTuamsuk, Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012

[6]. Dr.Anubhav Kumar, Dr. Arvind K Sharma, Comprehensive Analysis of Data Mining Techniques and Trends for Knowledge Management System, IJRECE VOL. 4 ISSUE 4 OCT.-DEC. 2016

[7]. Ms Tripti Chopra, Dr. Shine David, Data Mining and its Efficacy in Knowledge Management with respect to HRIS Application, International Journal of Research granthaalayah a knowledge repository, chopra et. Al., vol 4(Iss.9): September,2016

[8]. KotiNeha, M Yogi Reddy, A Study On Applications Of Data Mining, International Journal Of Scientific & Technology Research Volume 9, Issue 02, February 2020