1st International Conference on Multidisciplinary Research (ICMR-2018)
NIILM University, Kaithal, Haryana, (India)
4th-5th August 2018 www.conferenceworld.in ISBN:978-93-87793-38-5

A Systematic Review in the area of Data Mining

SANDEEP SINGH, ER. GURPREET SINGH

RESEARCH SCHOLAR ,ASSISTANT PROFESSOR

ABSTRACT

In modern years, data-mining (DM) has turn out to be one of the mainly precious tools for extracting and manipulating data and for beginning outlines in order to construct practical information for decision-making. The breakdowns of structures, metals, or materials in an surroundings are often either a consequence of unawareness or the inability of people to take note of past problems or learning the patterns of previous period occurrences in order to make learned decisions that can forestall future occurrences. Almost all directions of life activities reveal a similar pattern. Whether the activity I finance, banking, marketing, retail sales, production, inhabitants study, employment, person migration, health sector, monitoring of human or machines, science or education, all have ways to proof known in sequence but are handicapped by not having the right tools to use this known information to tackle the reservations of the future In this review paper, we have discussed all the aspects of data mining in the regards to the application areas, future trends and techniques of data mining. The paper also reveals the different algorithms which are used in the field data mining.

I.INTRODUCTION

Data mining [1] now become one of the most progressive and helpful fields for the extraction and manipulation of data to produce useful information. Thousands of businesses are using data mining applications every day in order to

Extract useful information and manipulating and identifying it from the records stored in their data warehouses databases and data repositories. By searching through a large amount of data stored in repositories, data warehouses and corporate databases in which data mining is the process of finding correlations, patterns, trends or relationships. Industrial engineering is a broad field and has many techniques and tools in its problem-solving arsenal. The aim of data mining is to find useful patterns in data. The large availability of rich data is a problem for many enterprises. More specific its work is to extract useful information from these large amounts of data. The challenge is analyzing (often) that large datasets and trying to find trivial (hidden) patterns / relationships. It is harder to retrieve knowledge from several datasets with the growing amount and to find unsuspected relations between data which can be done with the help of data mining which is also known as Knowledge Discovery from Data (KDD). To replace or enhance human intelligence by scanning through massive storehouses of data to discover meaningful new correlations data mining is used [5].

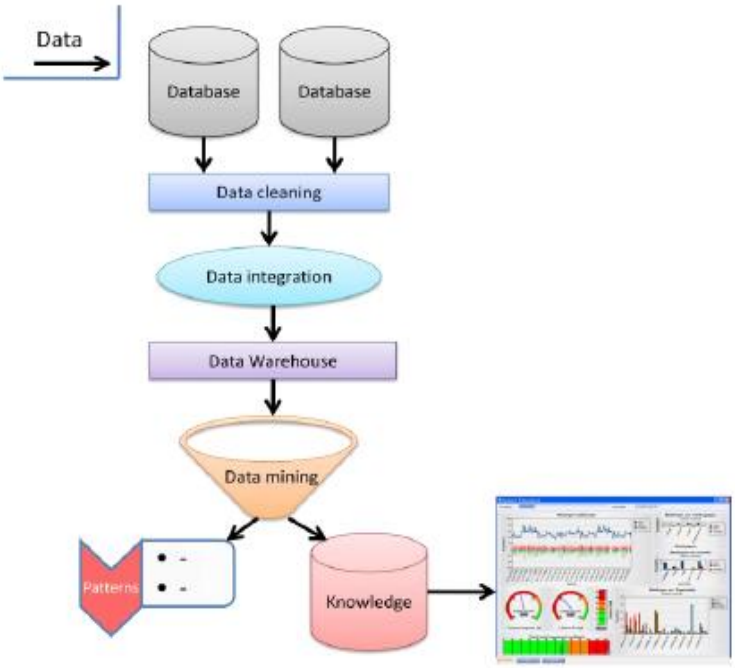

Figure 1. Data mining process[4]

A. Data mining consists of an iterative sequence of the following steps

1) Data cleaning (to remove noise and inconsistent data)
2) Data integration (where multiple data sources may be combined)
3) Data selection (where data relevant to the analysis task are retrieved from the database)
4) Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
5) Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)
6) Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

The concept of data mining, e.g. in large collections of data finding relevant important patterns which is not an emerging technology anymore. For mining data most of the companies have developed software whose application is however far from universal. Software like STATISTICA is used by only bigger companies for appliance on their Business Intelligence (BI). Data mining is becoming more mature, the techniques are highly developed and much research is performed in this area [6].

II.TECHNIQUES OF DATA MINING

Fundamentally, in data mining there is processing of data and identifying patterns and trends in that information so that you can decide or judge. Its principles have been around for many years, but, it become even more prevalent with the advent of big data, recently there are several major data mining *techniques* have been developing and using in data mining projects including *association*, *clustering*, *prediction*, *sequential patterns classification*, and *decision tree[7]*.

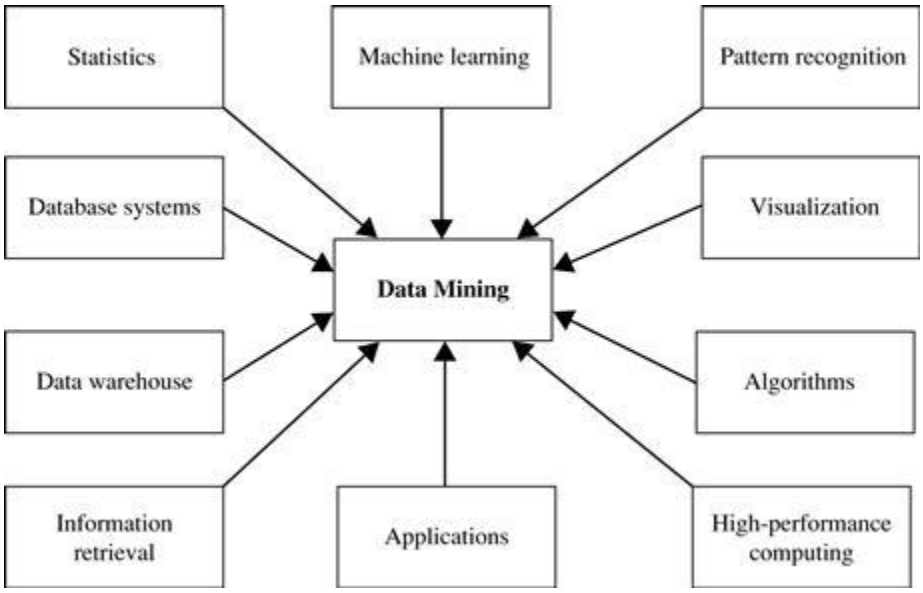

Figure2. Process of Data Mining [2]

1) Association

It is one of the best-known data mining techniques. In association [6], based on a relationship between items in the same transaction a pattern is discovered that's why this is also known as *relation technique*. In *market basket* to identify and analyze a set of products that customers frequently purchase together, association technique is used. To research customer's buying habits, Retailers are using association technique like based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and therefore, to save time of customer and increase sales they can put beers and crisps next to each other.

2) Classification

Classification [9] is based on machine learning and it is a classic data mining technique. Basically, it is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method uses mathematical techniques such as decision trees, neural network, statistics and linear programming, In classification, we develop the software that can learn how to classify the data items into groups. For example,

we can apply classification in the application that "given all records of employees who left the company and also predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.

3) Clustering

Clustering [6] is a data mining technique in which by using the automatic technique it makes a meaningful or useful cluster of objects which have similar characteristics. It firstly defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take example of book management in the library. In a library, there is a wide range of books available on various topics. The challenge is how to keep those books in a way so that readers can take several books on a particular topic without difficulty. We can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name by using the clustering technique. If readers want to grab books in that topic than instead of looking for the entire library they would only have to go to that shelf.

4) Prediction

The prediction [9] is one of a data mining techniques that discover the relationship between independent variables and between dependent and independent variables. For instance, it can be used to predict profit for the future in the sale and if we consider the sale is an independent variable, profit could be a dependent variable. Then we can draw a fitted regression curve that is used for profit prediction based on the historical sale and profit data.

5) Sequential Patterns

Sequential patterns [10] analysis is one of data mining techniques that discovers or identify similar patterns, regular events or trends in transaction data over a business period. In sales, businesses can identify a set of items that customers buy together at different times in a year with historical transaction data. Based on their purchasing frequency in the past, businesses can use this information to recommend customers buy it with better deals.

6) Decision trees

It is one of the most common used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple condition or question that has multiple answers. Each answer then leads to a set of condition or question that help us determine the data than based on it we can make the final decision. For example, we determine whether or not to play by the use of following decision trees.

III.PROBLEM STATEMENT

Data mining [3, 4] not only engage a set of solutions, technologies or systems, but also contains a structured process in which human interaction is important. Humans decide if they justify further study and exploration or if the patterns discovered have some relevance to the problem at hand. With this in mind the needs and interests of specific businesses data mining approaches have been integrated. Data mining techniques have many applications which can be used in many different fields. They range from the biomedical and DNA analysis to financial analysis, and fraud detection. They can also be used to track cross-selling products and customer preferences. Every day new applications have being found. It is very important that the data is correctly prepared and collected for its specific applications in order for data mining techniques to provide the intended result the full exploitation of all available data should be there. If there is no existing technique that matches, users must manipulate available ones in that way by which will find the best fit. With so many choices in the market, users need help in deciding the various tools offered by many vendors in the market. Additionally, data mining applications continue to be developed. There are, however, few that support decision-making in industrial engineering. Thus, applications of data mining in fields such as process control, quality run, human factors, material handling and preservation and reliability in production systems should be studied and addressed in more detail.

IV.OBJECTIVES FOR RESEARCH ON DATA MINING

To address the lack of industrial engineering applications and guidelines for using data mining in existing applications, this research proposes to do the following [7]:

- Develop a appropriate methodology in industrial engineering for the application of data mining.
- Analyze and compare the different and successfully applied techniques and tasks which are used in data mining.
- Identify the main advantages and disadvantages of data mining techniques and tools used for the application in industrial engineering.
- Identify possible problem issues or areas for the application of data mining in industrial engineering.
- In the field of industrial engineering identify possible applications of data mining.

In order to accomplish these goals, existing approaches to data mining and current data mining applications are analyzed and reviewed. The results are then used to develop a proposed methodology for applying data mining to the informational needs of industrial engineering.

V.APPLICATIONS OF DATA MINING

Data mining is widely used in diverse areas. There are number of commercial data mining system available today and yet there are many challenges in this field.

1) Future Healthcare

Data mining holds great potential to improve health systems [11]. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches such as machine learning, soft computing, multi-dimensional databases, data visualization and statistics. By using mining we can predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help in to detect fraud and abuse in healthcare insurers.

2) Market Basket Analysis

Market basket analysis [13] is a modeling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This may allow the retailer to understand the purchase behavior of a buyer and help it to know the buyer's needs so that they can change the store's layout accordingly. Using differential analysis the comparison of results between different stores between customers in different demographic groups can be done.

3) Education

There is a new emerging field, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of educational data mining are identified as predicting students future learning behavior, studying the effects of educational support, and advancing scientific knowledge about learning. To take accurate decisions data mining can be used by an institution and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. According to that learning pattern of the students can be captured and used to develop techniques to teach them to get the better results.

4) Manufacturing Engineering

Knowledge is the best asset a manufacturing enterprise would maintain. To discover patterns in complex manufacturing process data mining tools can be very useful. To extract the relationships between product architecture, product portfolio, and customer needs data in system-level designing data mining can be used. It can also be used to predict the product development cost, span time and dependencies among other tasks.

5) CRM

Customer Relationship Management is all about acquiring and retaining customers, also implementing customer focused strategies and improving customers' loyalty. To maintain a proper relationship with a customer a

business need to collect data and analyzed the information. This is where data mining plays its part with the help of which collected data can be used for analysis. Instead of being confused where to focus to retain customer, the applicant for the solution get filtered results.

6) Fraud Detection

To the action of frauds billions [9] of dollars have been lost. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns to turn data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system work is that the information of all the users should be protected. A supervised method includes collection of sample records which are classified to fraudulent or non-fraudulent. To identify whether the record is fraudulent or not a model is built using this data and the algorithm is made.

7) Intrusion Detection

Any action that will compromise the integrity and intimacy of a resource is an intrusion. The protective measures to avoid an intrusion includes user authentication, avoid programming errors, and information protection. By adding a level of focus to anomaly detection data mining can help to improve the intrusion detection. It helps an analyst to distinguish an activity from common everyday network activity. Data which is more relevant to the problem can be extracted by the help of data.

8) Lie Detection

Catching a criminal is easy whereas bringing out the truth from him is difficult. To investigate crimes, monitor communication of suspected terrorist's law enforcement can use mining techniques. This filed includes text mining also. From unstructured text the meaningful patterns in data can be extracted by using text mining. A model for lie detection is created [10] by comparing the data samples collected from previous investigations. Processes can be created according to the necessity with this model.

9) Customer Segmentation

Traditional market research may help us to segment customers but data mining goes in deep and increases market effectiveness. Data mining aids in aligning the customers into a distinct segment and can tailor the needs according to the customers. The main aid of market is always about retaining the customers. Based on vulnerability data mining allows to find a segment of customers and to enhance satisfaction the business could offer them with special offers.

10) Financial Banking

With computerized banking [2] everywhere, so with new transactions huge amount of data is supposed to be generated. By finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts data mining can contribute to solving business problems in banking and finance. The managers may

find these information for better segmenting, acquiring, targeting, retaining and maintaining a profitable customer.

11) Corporate Surveillance

Corporate surveillance is the monitoring of a person or group's behavior by a corporation. The data collected is mostly used for marketing purposes or sold to other corporations and it is also regularly shared with government agencies. The business person can make their products desirable by their customers with the use of this. The targeted advertisements on Google and Yahoo used that data for direct marketing purposes, such as, where ads are targeted to the user of the search engine by analyzing their search history and emails [11].

12) Research Analysis

History shows that we have witnessed revolutionary changes in research. Data cleaning, data pre-processing and integration of databases can be done with the help of data mining. Any similar data from the database can be find by the researchers that might bring any change in the research. The correlation between any activities and identification of any co-occurring sequences can be known. Data visualization and visual data mining provide us with a clear view of the data.

13) Criminal Investigation

Criminology [9] is a process that aims to identify crime characteristics. Exploration and detection of crimes and their relationships with criminals can be analyzed by crime analyses. Criminology have become an appropriate field for applying data mining techniques because of the high volume of crime datasets and also the complexity of relationships between these kinds of data. Crime matching process can be performed by converting text based crime reports into word processing files.

14) Bio Informatics

Data mining approaches [7] seem ideally adapted for Bioinformatics, since it is data-rich. Useful knowledge can be extracted from massive datasets gathered in biology with the help of mining biological data, and also in other related life sciences areas such as medicine and neuroscience. Applications of data mining to bioinformatics include protein function inference, disease treatment optimization, gene finding, disease diagnosis, disease prognosis, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

15) E-commerce

Perhaps E-commerce sites are one of the most well known examples of Data Mining and Analytics come from. Data Mining and Business Intelligence used by many E-commerce companies to offer cross-sells and up-sells through their websites. Amazon is One of the most famous of these, who use sophisticated mining techniques to drive there, this functionality is liked by People who viewed that product.

VI.FUTURE TRENDS IN DATA MINING

The most widely [6] used methods to extract data from different sources and organize them for better usage is data mining. In spite of having different commercial systems for data mining, when they are actually implemented it at that time a lot of challenges come up. With rapid evolution in the field of data mining, companies are expected to stay abreast with all the new developments.

Important Future Trends in Data Mining

Now businesses catching up with the others which have been slow in adopting the process of data mining Critical business decisions is widely used in extracting important information through the process of data mining. In the coming decade, we can expect data mining to become as omnipresent as some of the more prevalent technologies used today. For the future some of the key data mining trends include :-

1. Multimedia Data Mining

Because of the growing ability to capture useful data accurately multimedia data mining is one of the latest methods which is catching up. It involves the extraction of data from different kinds of multimedia sources such as text, hypertext, images, audio video etc. and the data is converted into a numerical representation in different formats. In clustering and classifications, performing similarity checks, and also to identify associations this method can be used.

2. Ubiquitous Data Mining

To get information about individuals this method involves the mining of data from mobile devices. In spite of having several challenges in this type such as complexity, privacy, cost, etc. this method has a lot of opportunities to be enormous in various industries especially in studying human-computer interactions.

3. Distributed Data Mining

It involves the mining of huge amount of information stored in different company locations or at different organizations that's why this type of data mining is gaining popularity. Extracting data from different locations and provide proper insights and reports based upon them can be done by using highly sophisticated algorithms.

4. Spatial and Geographic Data Mining

For extracting information from environmental, astronomical, and geographical data which also includes images taken from outer space this is new trending type of data mining. Distance and topology which is mainly used in geographic information systems and other navigation applications various aspects can be reveal by this type of data mining.

5. Time Series and Sequence Data Mining

The primary application of this type of data mining is study of cyclical and seasonal trends. This practice is also helpful in analyzing even random events which occur outside the normal series of events. This method is mainly being use by retail companies to access customer's buying patterns and their behaviors.

VII.ALGORITHMS FOR DATA MINING

An *algorithm* in data mining (or machine learning) is a set of heuristics and calculations that creates a model from data. The algorithm first analyzes the data you provide, looking for specific types of patterns or trend. to create a model. For a specific analytical task choosing the best algorithm to use can be a challenge. Although you can use different algorithms to perform the same business task, each algorithm produces a different result and more than one type of result can also be produced by algorithms.

Choosing an Algorithm by Type

- One or more discrete variables is predicted by **classification algorithms**, based on the other attributes in the dataset.
- One or more continuous numeric variables are predicted by **regression algorithms** such as profit or loss, based on other attributes in the dataset.
- Data is divided into groups, or clusters by **segmentation algorithms** that have similar properties.
- Correlations between different attributes in a dataset are find out by **association algorithms.** Creating association rules is the most common application of this kind of algorithm, which can be used in a market basket analysis.
- Frequent sequences or episodes in data is summarize by **sequence analysis algorithms,** such as a series of clicks in a web site, or a series of log events preceding machine maintenance.

Most useful algorithms for Data Mining [5,7]

1. **C4.5:** It constructs a classifier in the form of a decision tree. In order to do this, a set of data representing things that are already classified is given to C4.5.
2. **K-Means:** k groups are created by k-means from a set of objects so that the members of a group are more similar. For exploring a dataset it's a popular cluster analysis technique which is a family of algorithms

designed to form groups such that the group members are more similar versus non-group members. In the world of cluster analysis clusters and groups are synonymous.

3. **Support Vector Machines**: It makes a hyperplane to classify data into two classes. SVM performs a similar task like C4.5 except at a high-level, but SVM doesn't use decision trees at all.

4. **Apriori:** association rules is applied to a database containing a large number of transactions which are learned by apriori algorithm. For learning correlations and relations among variables in database, association rule learning is a data mining technique.

5. **Expectation-maximization EM:** In data mining, For knowledge discovery EM is generally used as a clustering algorithm (like k-means). In statistics, the EM algorithm iterates and optimizes the likelihood of seeing observed data while estimating the parameters of a statistical model with unobserved variables.

6. **PageRank:** Designed to determine the relative importance of some object linked within a network of objects and it is a link analysis algorithm. looking to explore the associations (a.k.a. links) among objects It's a type of network analysis. Google's search engine is the most prevalent example of PageRank. Although their search engine doesn't solely rely on PageRank, it's one of the measures Google uses to determine a web page's importance.
7. **AdaBoost:** It constructs a classifier and is a boosting algorithm. As you probably remember a bunch of data is taken by classifier and attempts to predict or classify which class a new data element belongs to.
8. **kNN:** kNN, or k-Nearest Neighbors, is a classification algorithm. However, As it's a lazy learner that's why it differs from the classifiers previously described. kNN has an easy time when all neighbors are the same class. The intuition is that, a new data point likely falls in the same class, if all the neighbors agree.
9. **Naive Bayes:** Naive Bayes is not a single algorithm, but a family of classification algorithms that share one common assumption. All other features given the class is independent of every feature of the data being classified..

10. **CART:** CART stands for classification and regression trees. It is a decision tree learning technique that outputs either classification or regression trees. Like C4.5, CART is a classifier. A classification tree is a type of decision tree. The output of a classification tree is a class.

REFERENCES

[1] L. Vaughan and Y. Chen, "Data mining from web search queries: A comparison of google trends and baidu index", Journal of the Association for Information Science and Technology, vol. 66, no. 1, (2015), pp. 13-22.

[2] L. E. Barrera, A. B. Montes-Servín, L. A. Ramírez-Tirado, F. Salinas-Parra, J. L. Bañales-Méndez, M. Sandoval-Ríos and Ó. Arrieta, "Cytokine profile determined by data-mining analysis set into clusters of

non-small-cell lung cancer patients according to prognosis", Annals of Oncology, vol. 26, no. 2, (2015), pp. 428-435.

[3] Diti Gupta, Abhishek Singh Chauhan, Mining Association Rules from Infrequent Itemsets: A Survey, International Journal of Innovative Research in Science, Engineering and Technology(IJIRSET),Vol.2,Issue 10,2013.

[4] T. Karthikeyan and N. Ravikumar, A Survey on Association Rule Mining International Journal of Advance Research in
Computer and Communication Engineering (IJARCCE)Vol. 3, Issue 1, January 2014.

[5] B. Samei, H. Li, F. Keshtkar, V. Rus and A. C. Graesser, "Context-based speech act classification in intelligent tutoring systems", In Intelligent Tutoring Systems, Springer International Publishing, (2014), pp. 236-241.

[6] Lemnaru, C., Firte, A., Potolea, R., "Static and Dynamic User Type Identification in Adaptive E-learning with Unsupervised Methods", *Proceedings of the 2011 IEEE 7th International Conference on Intelligent Computer Communication and Processing*, pp. 11-18, 2011.

[7] M.A. Hall and G. Holmes. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transaction on Knowledge and Data Engineering*, 15(3):in press, May/June 2003.

[8] Pinky Saikiya Datta (2010) Prediction of rainfall using data mining technique over Assam. Indian Journal of Computer Science and Engineering Vol. 5 (2), 2014; 85-90.

[9] Lior Rokach and Oded Maimon,"Data Mining with Decision Trees: Theory and Applications(Series in Machine Perception and Artificial Intelligence)", ISBN: 981-2771-719, World Scientific Publishing Company, 2008.

[10] Umamaheswari. K, S. Niraimathi "A Study on Student Data Analysis Using Data Mining Techniques" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013.

[11] D Ramesh , B Vishnu Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013.