Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 5, Issue. 12, December 2016, pg.103 – 110

Data Mining System and Applications: A Study

Raghavendra Naik[1], N. Deepika[2]

[1]Dept. of Computer Science & Engineering, NHCE, Bangalore, India

[2]Dept. of Computer Science & Engineering, NHCE, Bangalore, India

[1] raaghav.nayak86@gmail.com; [2] deepikajvijay@gmail.com

Abstract - Data mining is the process of discovering and extracting of interesting patterns and knowledge from large amounts of data i.e., knowledge discovery from Data. The most common previous uses of data mining have been to help businesses to gain and maintain a competitive advantage as well as to answer questions, solve problems, or make informed decisions. Some of these industries that have been turning to data mining are engineering, medicine, education, marketing, and more. In this paper a review or survey of alternative uses of data mining in education, medicines, financial and engineering are described. The future of data mining will only grow and expand from where it is currently because more and more technological advances will be made to aid in the data mining process and the increasing need for finding more hidden information in large amounts of data.

Keywords— Data Mining, Data Analysis, Knowledge Discovery, Applications.

I. INTRODUCTION

Data mining is widely used in very different areas. There are a number of commercial data mining systems, tools and products available today and still there are many challenges in this field.

Data Mining is main concerned with the analysis of data and Data Mining tools and techniques are used for finding patterns from the existing data sets. The main objective of Data Mining is to find patterns automatically with least amount of user input and efforts. Data Mining is a powerful tool capable of handling decision making and for forecasting future trends of market and market status. Data Mining tools and techniques can be successfully applied in various fields in various forms. Many Organizations now start using Data Mining as a tool, to deal with the competitive environment for data analysis and evaluate various trends and pattern of market and to produce quick and effective market trend analysis [2].

II. DATA MINING METHODS

The Data mining methods are mainly categorized as follows: Online Analytical Processing (OLAP), Classification, Clustering, Association Rule Mining, Temporal Data Mining, Time Series Analysis, Spatial Mining, Web Mining, Text Mining etc. These methods use different algorithms and different data and types. The data source can be Data-warehouses, Databases, Flat-files or Text-files etc. It uses different algorithms and methodologies like Statistical Algorithms, Decision-tree based algorithms, Nearest-neighbour, Based on Neural networks, Genetic Algorithm based algorithm, Rule-based algorithm, Support Vector Machine etc. The selection of data mining algorithm is mainly based on the required output from the mining process and type of the data or information used for mining process. The skilled persons the specific domains are responsible in data mining algorithm selection [2].

A knowledge discovery process (KDP) involves pre-processing data, selection of a data mining algorithm, and getting the results of processing of mining data. There are lots of different choices for all of these stages, and non-trivial interactions between them. Therefore both new users in data mining and Data mining specialists require help and assistance in Knowledge discovery processes.

The Intelligent Discovery Assistants (IDA) helps users in applying right and required KDPs. The IDA can provide users with three main benefits:

- A methodical list of valid knowledge discovery processes;
- Effective rankings of valid processes by different criteria and condition, which help to choose between the options;
- An infrastructure for sharing knowledge, which leads to network externalities [3].

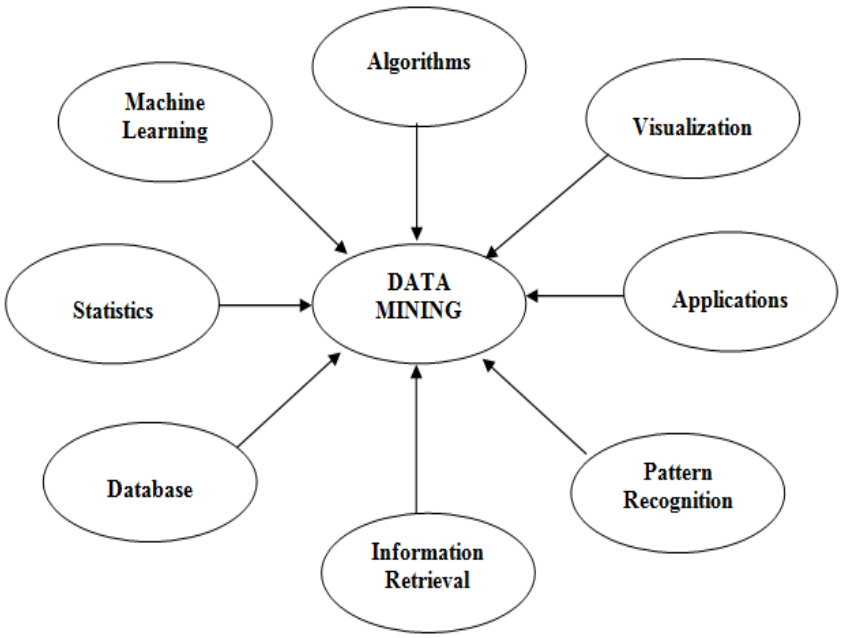

Fig 1: Data Mining Architecture [3]

Several other attempts have been made to automate this process and design of a generalized data mining tool that poses intelligence to select the data and data mining algorithms and up to some extent the knowledge discovery.

III. DATA MINING TASKS

The Data mining techniques are classified into different categories based on the usage and requirement of the Data mining results as follows:

Exploratory Data Analysis: It is simply exploring the data and summarizing their main characteristics. These techniques are interactive and visual.

Descriptive Modelling: A mathematical process of describing real-world events and relationships between factors those are responsible for them. It displays the correlation between the marketing and sales strategies of the company and the measurable results of those operations. This type of modelling used in the organizations where they run based on advertising and marketing targets.

Predictive Modelling: This model permits the value that is predicted from the known values of other variables. In this, a model is created, tested and validated for the best prediction probability of the outcome.

Discovering Patterns and Rules: This method is concern with detecting fraudulent behaviour by determining the data that are differ from others or different types of transactions where the data points are not same.

Retrieval by Content: It is finding pattern similar to the pattern that is a user of it is in interest. These types of data mining tasks are mainly used in text searches and image search data sets.

IV. DATA MINING APPLICATIONS

The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as:

A. Data Mining as Financial Data Analysis

Financial data is mainly collected from banks and from other financial sectors. This financial data is usually reliable, complete and has high quality. Financial data need a systematic method for data analysis. Data Mining plays an important in analysis of financial data. Data Mining follows steps such as data collection and understanding, data refinement, model building and model evaluation and deployment. Primarily, we should identify the problem and what logic or hypothesis can be used to do a successful data analysis and, type and quality of pattern that can use, success measures, selection of data, and applying attribute and relational based methodologies taking financial data. The proper analysis of financial data enables us to better decisions making capabilities according to the market analysis. Some typical cases of financial Data Mining techniques are mentioned below [2]:

1) *Design and construction of data warehouses for multidimensional data analysis and data mining*: Data warehouses mainly need to construct for banking and financial data. As they are of huge data, a multidimensional data analysis methods need to be used for data analysis the general properties of data. If the company's financial officer or any other higher level person or manager want to view the debt and revenue changes by month or year, region, and sector, and other factors, along with maximum, minimum, total, average, and other statistical information. Data warehouses, data cubes, characterization and class comparisons, clustering, and outlier analysis methods are the important ones in financial data analysis and data mining.

2) *Loan payment prediction and customer credit policy analysis:* This is the critical section to do the business of a bank with their customer. Factors related to the risk of loan payments include reason for loan, other loan

history, backup sources, bank account activities, term of the loan, and payment terms to income ratio, education level, residence region, and credit history. The bank may then decide to adjust its loan-granting policy so as to grant loans to those customers whose applications were previously denied but whose profiles show relatively low risks according to the critical factor analysis.

3) *Classification and clustering of customers for targeted marketing:* These methods are used for customer group identification and targeted marketing. We have to classification to identify different factors that may influence a customer's decision regarding banking and factors in which customer will come under specific category of people. Through multi-dimensional clustering techniques, we can identify customer with similar behaviour and can group them and facilitate targeted marketing.

4) *Detection of money laundering, fraud and other financial crimes:* This is one the critical section in the banking and finance sectors to check or detect money laundering and other financial crimes. To achieve this we have to integrate information from different, multiple, heterogeneous databases (e.g., bank transaction databases and federal or state crime history databases, and transaction and CIBIL scores of the customer's all related accounts), as long as they are completely up to the details and history of both personal and transactions. Useful tools include data visualization tools (to display transaction activities using graphs by time and by groups of customers), linkage and information network analysis tools (to identify links among different customers and activities), classification tools (to filter unrelated attributes and rank the highly related ones), clustering tools (to group different cases), outlier analysis tools (to detect unusual amounts of fund transfers or other activities), and sequential pattern analysis tools (to characterize unusual access sequences) [3].

With computerised banking everywhere 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 apparent to managers because the volume data is too large or is generated too quickly to screen by experts. The managers may find these information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer.

B. Data Mining for the Retail Industry

In retailer industry, it needs to collect lot of data or the information to analyse and to take further actions or operations in the industry. So the data needs to be collected from the areas like Sales of the products, Customers and their shopping details and history, transportation methodologies of goods, amount of items consumes histories and services that are often used and rarely used. The amount of data collected are change based on the increasing availability, popularity and other parameters that provided in web-sites i.e., in e-commerce sites. In retailer data manipulation, data also to collect both from offline (from the physical shops) as well as from online (e-commerce websites) if available. Electronic commerce describes the buying and selling of products, services, and information via computer networks including the Internet [1].

Retail data mining can help in identifying customer's way of buying, discover customer shopping patterns and trends, trends of market with the price, profit, customer budget ranges, and quality so that such areas can be improved by the quality of customer service based on customer requirement, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business. By this study, effective analysis needs to be done and then by checking the price of the sale, check if there any need of advertisement in the market and adjustment of price and variety in various goods in order to attract customer. Association analysis also helps to get information in order to promote sales. In this way, various Data Mining tools help in the retail industry.

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 [9] -

- Design and Construction of data warehouses based on the benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

Today, most major chain stores also have web sites where customers can make purchases online. Some businesses, such as Amazon.com (www.amazon.com), Flipkart.com (www.flipkart.com), exists solely online, without any brick-and-mortar (i.e., physical) store locations. Retail data provide a rich source for data mining [3].

C. Data Mining in Healthcare Sectors

In healthcare sectors, data mining and analysis techniques used to predict patterns or sequences from large amount of data. The healthcare industry today generates large amounts of complex data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. Improving the quality of patient care and reducing the healthcare costs are the main goals of many programs. Data mining has helped these programs succeed. The knowledge gained in the data mining methods can be utilized in continues improving and decision making on daily based works in healthcare industries. Data mining applications in healthcare can be grouped as the evaluation into broad categories [4].

1) *Treatment effectiveness*: Data mining applications can develop by studying the different scenarios which are reported earlier in the similar cases to evaluate the effectiveness of medical treatments. Data mining can produce an analysis of which continuous of action proves effective by comparing and contrasting causes, symptoms, and courses of treatments and helps in most probable decision making.

REMIND (Reliable Extraction and Meaningful Inference from Non-structured Data) system integrates the structured and unstructured clinical data in patient records to automatically create high quality structured clinical data. The high quality of structuring allows existing patient records to be mined to support guidelines compliance and to improve patient care [2].

2) *Healthcare management:* Data mining applications can be developed for better identification and tracking of chronic disease states and high-risk patients, design appropriate interventions, investigation and reduce the possibilities to admit to hospitals and get treatments at the earliest stages. Data mining helps to analyse large amount of data and states to search and match for different patterns that could cause a huge impact in the future.

To illustrate a data mining application in healthcare, suppose that as part of its healthcare management program, HealthOrganization (a fictional healthcare organization) is interested in finding out how certain variables are associated with the onset of diabetes. The purpose of this data mining application is to identify high-risk individuals so appropriate messages can be communicated to them.

A dataset exists in the data warehouse of HealthOrganization that contains the following seven variables of particular interest to HealthOrganization: gender, age, body mass index (BMI), waist hip ratio (WHR), smoking status, the number of times a patient exercises per week, and onset of diabetes, which is the target variable, measured by a dichotomous variable indicating whether an individual has tested positive for diabetes. The dataset comprises 262, or 12.78 percent, of positive diabetic cases and 1,778, or 87.22 percent, of negative non-diabetic cases [8].

3) *Customer relationship managements:* Through the customer relationship and frequent interaction with the customers alias patients or their parties will also gains the history or statements about the diseases and their symptoms so that, analysis of this can be use further mining process. Customer interactions can be done through call centres, physicians, offices, inpatient settings, billing departments, and ambulatory care settings.

Miller [8] has suggested that the data mining of patient survey data can help set reasonable expectations about waiting times, reveal possible ways to improve service, and provide knowledge about what patients want from

their healthcare providers. Also, Hallick [8] has suggested that CRM in healthcare can help promote disease education, prevention, and wellness services.

4) *Fraud and abuse*: Data mining application finds out fraud and abuse in the healthcare system by means of identifying unusual and abnormal sequences or methods in hospitals or by physicians. Data mining also prevents and highlights invalid or fraudulent usage of insurance and medical claims. Some countries use Medicaid Fraud and Abuse Detection Systems uses data mining to detect and discover the fraud and abuse and saved millions of dollars.

Using data mining to detect fraud and abuse is the Texas Medicaid Fraud and Abuse Detection System, which recovered $2.2 million and identified 1,400 suspects for investigation in 1998 after operating for less than a year. 27 In recognition of its success, the Texas system has won a national award for outstanding achievement and top honors for innovative use of technology [8].

5) *Pharmaceutical Industry*: Data mining or knowledge discovery from data helps in pharmaceutical firms to discover the patterns in continuous improving in the discovery of quality of drug and better delivery methods. Also to manage their inventories and to develop new product and services.

Pharmaceutical companies can also apply data mining to huge masses of genomic data to predict how a patient's genetic makeup determines his or her response to a drug therapy.

6) *Ranking Hospitals*: Organizations rank hospitals and healthcare plants based on the survey information reported by healthcare providers. Based on the capacity or capability of handling of high risk patients the ranking of hospitals are done and their success rates. The hospitals with high rank are capable of handling high risk patients and lower rank hospitals do not consider the risk factors.

The ranking of hospitals are also done based on the study of usage of advanced equipments and quality and standard and well qualified doctors and survey results of the hospitals. Standardized reporting would also be important for meaningful comparisons across hospitals [6].

Johnson [8] has suggested that, at a higher level, data mining can facilitate comparisons across healthcare groups of things such as practice patterns, resource utilization, length of stay, and costs of different hospitals. Recently,

Sierra Health Services has used data mining extensively to identify areas for quality improvements, including treatment guidelines, disease management groups, and cost management. [8]

D. Data Mining in Science & Engineering

In Science & Engineering sectors, Vast amounts of data have been collected from different domains of science like geosciences, astronomy, meteorology, geology, and biological sciences etc. using sophisticated telescopes, multispectral high-resolution remote satellite sensors, global positioning systems, and new generations of biological data collection and analysis technologies. Huge amount of data are generated and collected due to different types of researches and modelling is created gradually by different peoples on various topics in different areas. Here we look at some of the challenges brought about by emerging scientific applications of data mining. [3].

1) *Data warehouses and data pre-processing*: Data pre-processing and data warehouses are critical for information exchange and data mining [3]. Warehouse creation requires too much time to resolving inconsistent or incompatible of data that are collected from multiple environments and at different time periods. For example, some events in the earth or in space may occur only after few or several years, and previous data on them might not have been collected as systematically and stored as they are today. Methods are also needed for the efficient computation of sophisticated spatial aggregates and the handling of spatial-related data streams.

2) Mining complex data types: Scientific data sets are heterogeneous in nature. They typically involve semi-structured and unstructured data, such as multimedia data and geo referenced stream data, as well as data with sophisticated, deeply hidden semantics (e.g., genomic and proteomic data) [3]. Different biological processes involve different sets of genes acting together in precisely regulated patterns and their structures are different based on animal and their body and parental genes. Thus, to understand a biological process we need to identify the participating genes and their regulators. This requires a lot of research and lot of data needs to be collected and different samples are created and tested and different researches in various conditions may show different behaviours and if earlier documents are not available, need to create them from the beginning. [3].

3) *Data mining in computer science:* Data mining in computer science can be used to help monitor system status, improve system performance, isolate software bugs, detect software plagiarism, analyze computer system faults, uncover network intrusions, and recognize system malfunctions [3].

Data mining in computer software and system engineering can operate either on static or dynamic data, depending on whether the system dumps traces beforehand for post analysis or if it must react in real time to handle online data. Various fields in the computer engineering are now made available in internet websites. Existing data analysis will help to improve computer programming and to overcome the difficulties and to connect computer based techniques with other fields like space, biology etc. Development and researches are continuous going on in this domain, which integrate and extend methods from machine learning, data mining, software/system engineering, pattern recognition, and statistics. Data mining in computer science is an active and rich domain for data miners.

V. LIMITATIONS

Data mining applications are greatly benefit in all the industries. But it has its own limitations. It is difficult to get the proper and complete healthcare data. Health data are complex and are different from once situations to another as they are collected from various sources like reports from the laboratories, from the discussion of patient or from the doctor's reviews. Hence, the data have to be collected and integrated before data mining can be done. Data mining process is bit complicated than structuring of other information technology as it needs a lot of information to collect and analyse. There may occur some problems which arise due to missing, corrupted, inconsistent or non-standardized data such as pieces of information recorded in different formats in different data sources. There will be ethical, legal and social issues, such as data ownership and privacy issues, related to data. Also in all the cases the successful application of data mining requires proper knowledge of the domain areas as well as in data mining methodologies and tools.

VI. CONCLUSIONS

Data Mining can be used in various fields like retail industry, Telecommunication industry etc. In Retail industry Data mining helps in identifying customer behaviour, shopping patterns and distribution policies etc. Data Mining also helps to identify fraud activities and also helps to better use of resources and improves the quality of services. Data mining tools helps greatly to study DNA analysis and to find various patterns and functions.

Most of the previous studies on data mining applications in various fields use the variety of data types range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain.

We can also conclude that there is no single data mining techniques which gives us the complete result for all type of data or applications [7]. Performance and cost of the data mining/processing or analysing techniques depends on the amount of data or dataset that we have taken for doing the research or experiment.

REFERENCES

[1] Simmi Bagga, Dr. G. N. Singh, Applications of Data Mining, "International Journal for Science and Emerging Technologies with Latest Trends",1(1):19-23(2012).

[2] Mr. S. P. Deshpande and Dr. V. M. Thakare, DATA MINING SYSTEM AND APPLICATIONS: A REVIEW, International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, (September 2010).

[3] Jiawei Han, Micheline Kamber, Jian Pei, "DATA MINING-Concepts and Techniques", 3rd Edition, ISBN:978-0-12-381479-1, MK Publications.

[4] M. Durairaj, V. Ranjani, Data Mining Applications In Healthcare Sector: A Study, International Journal of Scientific & Technology Research, Volume 2, Issue 10, October 2013, ISSN:2277-8616.

[5] Mary K. Obenshain, MAT, "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, Vol. 25, No. 8 August 2004, pp. 690-695.

[6] Salim Diwani, Suzan Mishol, Daniel S.Kayange, Dina Machuve and Anael Sam, "Overview Applications of Data Mining In Health Care: The Case Study of Arusha Region", International Journal of Computational Engineering Research, Vol:03, Issue: 8.

[7] Divya Tomar and Sonali Agarwal, "A survey on Data Mining approaches for Healthcare":International Journal of Bio-Science and Bio-Technology, Vol.5, No.5(2013), pp.241-266

[8] Hian Chye Koh and Gerald Tan, Data Mining Applications in Healthcare, Journal of Healthcare Information Management — Vol. 19, No. 2

[9] https://www.tutorialspoint.com/data_mining/dm_applications_trends.html

