Data Mining Techniques & Applications

Shabnam Kumari[1], Krishan[2], Reema[3], Sunita Kumari[4]
[1, 2 ,3 ,4] Department of CSE
[1, 2, 3] Sat Kabir Institute of Technology & Management, Bahadurgarh, Haryana, India
[4] G.B Pant Engineering College, Okhla.

Abstract- *this paper illustrates the concept of data mining, how this process is used to find useful patterns from large amount of data. The paper elaborates that Mining process can be applied to any type of data ranging from weather forecasting, electric load prediction, product design, etc. Data mining is the process of discovering potentially useful, interesting, and previously unknown patterns from a large collection of data. By its simplest definition, data mining automates the detections of relevant patterns in database. The paper discusses few of the data mining techniques, algorithms and some of the organizations which have adapted data mining technology to improve their businesses and found excellent results*

Keywords- Knowledge discovery, Data mining Techniques, clustering, decision trees.

I. INTRODUCTION

Data mining refers to extracting or mining the knowledge from large amount of data. Data collection and storage technology has made it possible for organizations to accumulate huge amounts of data at lower cost. Exploiting this stored data, in order to extract useful and actionable information, is the overall goal of the generic activity termed as data mining. The following definition is given:

Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. In [1], the following definition is given:

Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. Data mining is an interdisciplinary subfield of computer science which involves computational process of large data sets' patterns discovery. The goal of this advanced analysis process is to extract information from a data set and transform it into an understandable structure for further use. The methods used are at the juncture of artificial intelligence, machine learning, statistics, database systems and business intelligence.

Data Mining is about solving problems by analyzing data already present in databases [2]. Data mining is also stated as essential process where intelligent methods are applied in order to extract the data patterns.

Data mining consists of five major elements:

- Extract, transform, and load transaction data (ETL) onto the data warehouse system (fig.1.1).
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

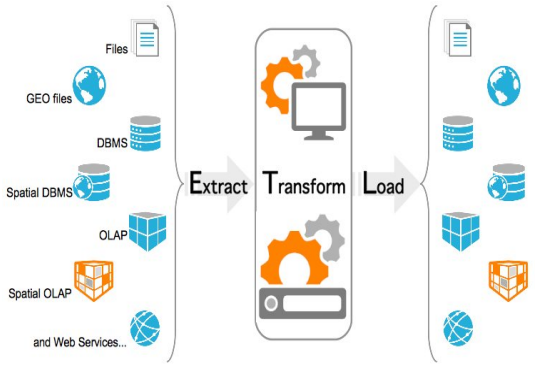

Figure 1. ETL process

The term data mining is alternatively named as "Knowledge mining". It refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [2]. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 1.2) shows data mining as a step in an iterative knowledge discovery process.

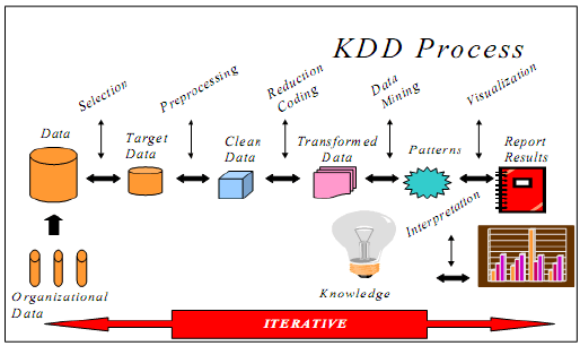

Figure 2. KDD process

The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:

1. Developing an understanding of

- the application domain
- the relevant prior knowledge
- the goals of the end-user

2. Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.

3. Data cleaning and preprocessing.

- Removal of noise or outliers.
- Collecting necessary information to model or account for noise.
- Strategies for handling missing data fields.
- Accounting for time sequence information and known changes.

4. Data reduction and projection.

- Finding useful features to represent the data depending on the goal of the task.
- Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

5. Choosing the data mining task.

- Deciding whether the goal of the KDD process is classification, regression, clustering, etc.

6. Choosing the data mining algorithm(s).

- Selecting method(s) to be used for searching for patterns in the data.
- Deciding which models and parameters may be appropriate.
- Matching a particular data mining method with the overall criteria of the KDD process.

7. Data mining.

- Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.

KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step. Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process [3].

II. DATA MINING TECHNIQUES

There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns etc., are used for knowledge discovery from databases[4].

1. Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit. Applications: market basket data analysis, cross-marketing, catalog design, loss-leader analysis, etc.

Types of association rules: Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Different types of association rules based on

Types of values handled

- Boolean association rules
- Quantitative association rules

Levels of abstraction involved

- Single-level association rules
- Multilevel association rules

Dimensions of data involved

- Single-dimensional association rules
- Multidimensional association rules

2. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Applications: Fraud detection and credit risk applications are particularly well suited to this type of analysis. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics[5].

For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay". And then we can ask our data mining software to classify the employees into each group. Classification Techniques

- Regression
- Distance
- Decision Trees
- Rules
- Neural Networks

This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis.

The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

3. Clustering

Clustering is "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clustersAlso we can define Clustering as the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering. Types of clustering method:

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

We can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without irritate. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

4. Prediction

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. In data mining independent variables are attributes already known and response variables are what we want to predict unfortunately, many real-world problems are not simply prediction For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict

because they may depend on complex interactions of multiple predictor variables[6]. Therefore, more complex techniques (e.g., decision trees) may be necessary to forecast future values. For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

5. Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

Types of neural networks

- Back Propagation

6. Regression

Attempts to find a function which models the data with the least error. Regression is the oldest and most well-known statistical technique that the data mining community utilizes. Basically, regression takes a numerical dataset and develops a mathematical formula that fits the data. When you're ready to use the results to predict future behavior, you simply take your new data, plug it into the developed formula and you've got a prediction! The major limitation of this technique is that it only works well with continuous quantitative data (like weight, speed or age). If you're working with categorical data where order is not significant (like color, name or gender) you're better off choosing another technique.
Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line ($y = mx + b$) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation[7].

III. APPLICATION OF DATA MINING

1. Market Basket Analysis

Market basket analysis is a modelling 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 technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

2. Education

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

3. Manufacturing Engineering

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

4. CRM

Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyse the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

5. Fraud Detection

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

6. Lie Detection

Apprehending a criminal is easy whereas bringing out the truth from him is difficult. Law enforcement can use mining techniques to investigate crimes, monitor communication of suspected terrorists. This filed includes text mining also. This process seeks to find meaningful patterns in data which is usually unstructured text. The data sample collected from previous investigations are compared and a model for lie detection is created. With this model processes can be created according to the necessity.

7. 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. Market is always about retaining the customers. Data mining allows to find a segment of customers based on vulnerability and the business could offer them with special offers and enhance satisfaction.

8. Financial Banking

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.

9. Corporate Surveillance

Corporate surveillance is the monitoring of a person or group's behaviour by a corporation. The data collected is most often used for marketing purposes or sold to other corporations, but is also regularly shared with government agencies. It can be used by the business to tailor their products desirable by their customers. The data can be used for direct marketing purposes, such as the targeted advertisements on Google and Yahoo, where ads are targeted to the user of the search engine by analyzing their search history and emails.

10. Research Analysis

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

11. Criminal Investigation

Criminology is a process that aims to identify crime characteristics. Actually crime analysis includes exploring and detecting crimes and their relationships with criminals. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Text based crime reports can be converted into word processing files. These information can be used to perform crime matching process.

12. Bio Informatics

Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction

13. 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 −

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

- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

14. 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 −

- 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.

15. 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 −

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

16. 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 −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

17. 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 −

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

18. 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 −

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

IV. CHALLENGES OF DATA MINING

1. Noisy and Incomplete Data

Data mining is the process of extracting information from large volumes of data. The real-world data is heterogeneous, incomplete and noisy. Data in large quantities normally will be inaccurate or unreliable. These problems could be due to errors of the instruments that measure the data or because of human errors. Suppose a retail chain collects the email id of customers who spend more than $200 and the billing staff enters the details into their system[10]. The person might make spelling mistakes while entering the email id which results in incorrect data. Even some customers might not be ready to disclose their email id which results in incomplete data. The data even could get altered due to system or human errors. All these result in noisy and incomplete data which makes the data mining really challenging.

2. Distributed Data

Real world data is usually stored on different platforms in distributed computing environments. It could be in databases, individual systems, or even on the Internet. It is practically very difficult to bring all the data to a centralized data repository mainly due to organizational and technical reasons. For example, different regional offices might be having their own servers to store their data whereas it will not be feasible to store all the data (millions of terabytes) from all the offices in a central server. So, data mining demands the development of tools and algorithms that enable mining of distributed data.

3. Complex Data

Real world data is really heterogeneous and it could be multimedia data including images, audio and video, complex data, temporal data, spatial data, time series, natural language text and so on. It is really difficult to handle these different kinds of data and extract required information. Most of the times, new tools and methodologies would have to be developed to extract relevant information.

4. Performance

The performance of the data mining system mainly depends on the efficiency of algorithms and techniques used. If the algorithms and techniques designed are not up to the mark, then it will affect the performance of the data mining process adversely.

5. Incorporation of Background Knowledge

If background knowledge can be incorporated, more reliable and accurate data mining solutions can be found. Descriptive tasks can come up with more useful findings and predictive tasks can make more accurate predictions. But collecting and incorporating background knowledge is a complex process.

6. Data Visualization

Data visualization is a very importance process in data mining because it is the main process that displays the output in a presentable manner to the user. The information extracted should convey the exact meaning of what it actually intends to convey. But many times, it is really difficult to represent the information in an accurate and easy-to-understand way to the end user. The input data and output information being really complex, very effective and successful data visualization techniques need to be applied to make it successful.

7. Data Privacy and Security

Data mining normally leads to serious issues in terms of data security, privacy and governance. For example, when a retailer analyzes the purchase details, it reveals information about buying habits and preferences of customers without their permission

V. CONCLUSION

Data mining is a “decision support” process in which we search for patterns of information in data. Data mining technology is an application oriented technology. It not only is a simple search, query and transfer on the particular database, but also analyzes, integrates and reasons these data to guide the solution of practical problems and find the relation between events, and even to predict future activities through using the existing data. Data mining techniques such as classification, clustering, prediction, association and sequential patterns etc it helps in finding the patterns to decide upon the future trends in businesses to grow. However privacy, security and misuse of information are the big problem if it is not address correctly.

VI. CONCLUSION

I would like to thank my guide Ms. Shabnam Kumari for her indispensible ideas and continuous support, encouragement, advice and understanding me through my difficult times and keeping up my enthusiasm, encouraging me and for showing great interest in my dissertation work, this work could not finished without her valuable comments and inspiring guidance.

REFERENCES

[1] Xingquan Zhu, Ian Davidson, "Knowledge Discovery and Data Mining: Challenges and Realities", ISBN 978-1-59904-252, Hershey, New York, 2007.

[2] Joseph, Zernik, "Data Mining as a Civic Duty – Online Public Prisoners Registration Systems", International Journal on Social Media: Monitoring, Measurement, Mining, vol. - 1, no.-1, pp. 84-96, September2010.

[3] Dr. Lokanatha C. Reddy, A Review on Data mining from Past to the Future, International Journal of Computer Applications (0975 – 8887) Volume 15– No.7, February 2011 [2]. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth,From Data Mining to Knowledge Discovery in Databases, AI Magazine Volume 17 Number 3 (1996)

[4] http://www.slideshare.net/Annie05/sequential-pattern-discovery-presentation

[5] http://dataminingtools.net/wiki/introduction_to_data_mining.php

[6] http://www.dataminingtechniques.net

[7] http://www.slideshare.net/huongcokho/data-mining-concepts

[8] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf Retrieved 2008-12-17..

[9] "Data mining and ware housing". Electronics Computer Technology (ICECT), 2011 3rd International Conference on Volume:1, Publication Year: 2011 , Page(s): 1 – 5

[10] "The applied research on data mining in the financial analysis of university with the analysis of college students „arrears as an example"

[11] Chen Hongfei; Wang Xiaoyan; Business Management and Electronic Information (BMEI), 2011 International Conference on Volume:2 Digital Object Identifier: 10.1109/ICBMEI.2011.5917992 Publication Year: 2011 , Page(s): 633 - 636