A SURVEY OF DATA MINING & ITS APPLICATIONS

Pankaj jain
M.Tech Student, Computer Science
Siddhi Vinayak College of Science & Hr.Education, Alwar (Rajasthan)

Abstract- Data mining consists of evolving set of techniques that can be used to extract valuable information and knowledge from massive volumes of data. Data mining research & tools have focused on commercial sector applications. Data and Information or Knowledge has a significant role on human activities. Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information. Due to the importance of extracting knowledge/information from the large data repositories, data mining has become an essential component in various fields of human life. Advancements in Statistics, Machine Learning, Artificial Intelligence, Pattern Recognition and Computation capabilities have evolved the present day's data mining applications and these applications have enriched the various fields of human life including business, education, medical, scientific etc. This paper will focus on issues related to data mining and. Further the paper focuses on some of its current applications.

Keywords – Data, Mining, Knowledge Discovery

I. INTRODUCTION

The field of data mining has evolved from its roots in databases, statistics, artificial intelligence, information theory and algorithms in to a core set of techniques that have been applied to a range of problems. Computational simulation and data acquisition in scientific and engineering domains have made tremendous progress over the past two decades. A mix of advanced algorithms, exponentially increasing computing power and accurate sensing and measurement devices have resulted in more data repositories. Advanced technologies in networks have enabled the communication of large volumes of data across the world. This results in a need of tools & Technologies for effectively analyzing the scientific data sets with the objective of interpreting the underlying physical phenomena. Data mining applications in geology and geophysics have achieved significant success in the areas as weather prediction, mineral prospecting, ecology, modeling etc and finally predicting the earthquakes from satellite maps An interesting aspect of many of these applications is that they combine both spatial and temporal aspects in the data and in the phenomena that is being mined. Data set in these applications comes from both observations and simulation.

II. DATA MINING PROCESS

Generally, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on available data. Data mining can be done on data which are inquantitative, textual, or multimedia forms. Data mining involves some of the following key steps.

a.) Problem definition: The first step is to identify goals. Based on the defined goal, the correct series of tools can be applied to the data to build the corresponding behavioural model.

b.) Data exploration: If the quality of data is not suitable for an accurate model then recommendations on future data collection and storage strategies can be made at this. For analysis, all data needs to be consolidated so that it can be treated consistently.

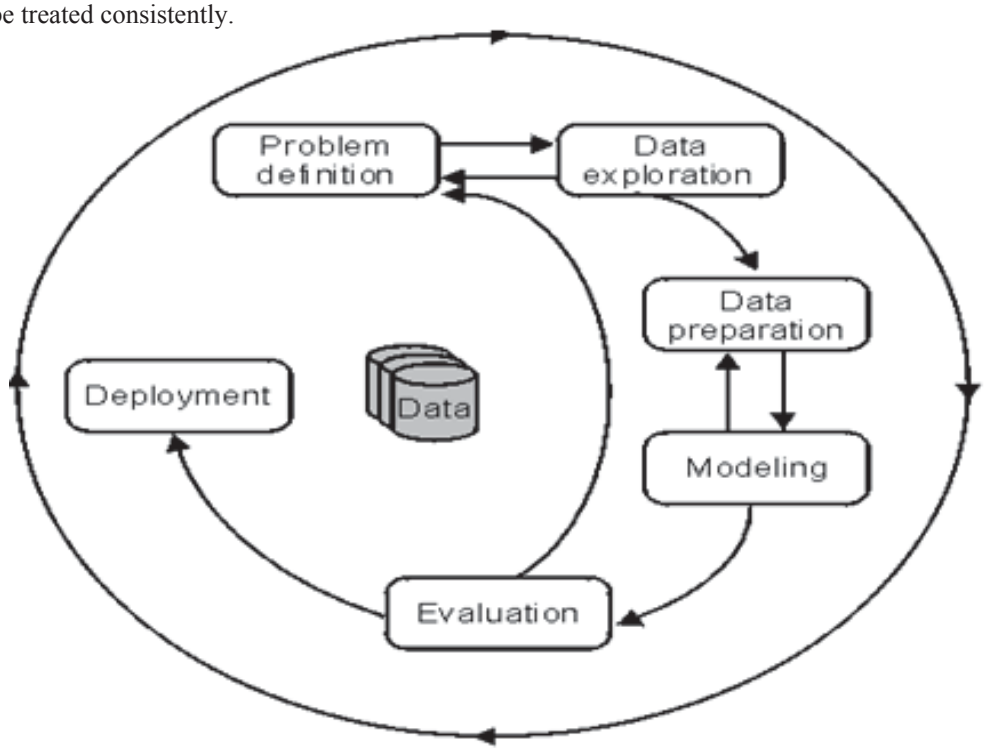

Fig.1. Data Mining Process Representation

c.) Data preparation: The purpose of this step is to clean and transform the data so that missing and invalid values are treated and all known valid values are made consistent for more robust analysis.

d.) Modeling: Based on the data and the desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighbourhoods and clustering but also next generation techniques such as decision trees, networks and rule based algorithms. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analysed.

e.) Evaluation and Deployment: Based on the results of the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

III. DATA MINING TECHNIQUES

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. The various data mining algorithms & techniques are:

a) Statistic
b) Classification
c) Clustering
d) Predication
e) Association rule
f) Neural networks
g) Decision Trees
h) Outlier analysis
i) Trend and evolution analysis

a) Statistics:

- Data cleaning i.e. the removal of erroneous or irrelevant data known as outliers.
- EDA Exploratory data analysis e.g. frequency counts histograms.
- Attribute redefinition e.g. bodies mass index.
- Data analysis is a measure of association and their relationships between attributes interestingness of rules, classification, prediction etc.

b) **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. 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.

c) **Clustering:** Clustering is a process of grouping similar data. The data which is not part of clustering are called as outliers. Clustering can be used as preprocessing approach for attribute subset selection and classification.

d) **Predication:** Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict.

e) **Association rule:** Mining association rules finds the interesting correlation relationship among large databases.Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis.

f) **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. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

g) **Decision Trees:** Decision tree is tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees.

h) **Outlier analysis:** A data object that is irrelevant to general behavior of the data ,it can be considered as an exception but is quite useful in fraud detection in rare events analysis.

i) Trend and evolution analysis-

- Trend and deviation: regression analysis
- Sequential pattern mining, periodicity analysis
- Similarity-based analysis

IV. MAJOR ISSUE IN DATA MINING

T he following are the major issues related with the usage of the data mining functionalities.

a.) Mining methodology and user interaction: Mining different kinds of knowledge in databases.Interactive mining of knowledge at multiple Levels of abstraction.

b.) Performance Scalability: Efficiency and scalability of data mining algorithms.Parallel distributed and incremental mining methods.

c.) Other Issues in Data Mining: Issues relating to the diversity of data types. Handling relational and complex types of data Mining information from heterogeneous databases and global information systems (WWW). Issues related to applications and social impacts. Application of discovered knowledge. Domain-specific data mining tools. Intelligent query answering. protection of data security, integrity, and privacy. Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem.

V. APPLICATIONS OF DATA MINING

Data mining techniques have been applied successfully in many areas from business to science to sports.

a.) Business applications: Many organizations now employ data mining as a secreat weapon to keep in pace or gain a competitive edge .Data mining has been used in advertising, CRM (Customer Relationship management), investments, manufacturing, sports/entertainment, telecom, e-Commerce, targeted marketing, health care, etc.

b.) Science applications: Data mining techniques have been used in astronomy, bioinformatics, drug discovery and many more.

c.) Data Mining in Bioinformatics: Applications of data mining to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction. For example, microarray technologies are used to predict a patient's outcome. On the basis of patients' genotypic microarray data, their survival time and risk of tumor metastasis or recurrence can be estimated. Machine learning can be used for peptide identification through mass spectroscopy. Correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. An efficient scoring algorithm that considers the correlative information in a tunable and comprehensive manner is highly desirable.

d.) Other application: Data mining has been successfully used for various other application such as Web: search engines, Government, law enforcement, profiling tax cheaters, anti-terror, credit approval, etc. Putting it together Data Mining is the step in the process of knowledge discovery in databases, that inputs predominantly cleaned, transformed data, searches the data using algorithms, and outputs patterns and relationships to the interpretation/evaluation step of the KDD process.

VI. CONCLUSION

In this paper we briefly reviewed the various data mining trends from its inception. This review would be helpful to researchers to focus on the various issues of data mining. This paper will focus on issues related to data mining and. Further the paper focuses on some of its current applications.

REFRENCES

[1] El-Sebakhy A. Emad, Faisal Abed Kanaan, Helmy T., Azzedin F. and Al-Suhaim F., "Evaluation of breast cancer tumor classification with unconstrained functional networks classifier," *Computer Systems and Applications, IEEE International Conference*, 2006, pp. 281 – 287.

[2] Han J. and Kamber M., *Data Mining: Concepts and Techniques*, 2nd ed., San Francisco, Morgan Kauffmann Publishers,2001.

[3] Dataminig Concepts and Techniques- Jiawei Han, Micheline Kamber.

[4] Yang, Qiang. *Data Mining and Bioinformatics: Some Challenges*, http://www.cse.ust.hk/~qyangC.S. Lu, H.Y.M Liao, "Multipurpose watermarking for image authentication and protection," *IEEE Transaction on Image Processing*, vol. 10, pp. 1579-1592, Oct. 2001.

[5] K.R.Venugopal, K.G. Srinivasa and L.M. Patnaik soft computing for data Mining ApplicationP. Tay and J. Havlicek, "Image Watermarking Using Wavelets", in *Proceedings of the 2002 IEEE*, pp. II.258 – II.261, 2002.

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

[7] C.Brunk, J.Kelly & Rkohai "Mineset An integrate system for data access, Visual Data Mining & Analytical Data Mining", proceeding of the 3rd conference on KDD 1997

[8] http://aaaipress.org/Papers/KDD/1996/KDD96-034.pdf

[9] http://www.slideshare.net/gtzi/a-datamineit-case-study-analyzing-earthquakes-presentation

[10] http://www.cs.uwec.edu/MICS/papers/mics2010_submission_32.pdf

[11] U . K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.