ISSN: 2321-7782 (Online)
Volume 3, Issue 9, September 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Recent Trends in Datamining Techniques

Dr. E. Kesavulu Reddy
Assistant Professor
Department of Computer Science
College of Commerce Management & Computer Science
S V University
Tirupati-A.P-India-517502

Abstract: ***Society produces massive amounts of data from different sources like business, science, medicine, economics, sports, web data etc. Tremendous amounts of data are stored in databases, data warehouses and other information repositories. The availability of large datasets and increasing importance of data analysis for scientific discovery is creating a new class of high-end applications. This class of applications includes data mining and scientific data analysis. Data mining is the process of knowledge discovery where knowledge is gained by analyzing the data store in very large repositories, which are analyzed from various perspectives and the result is summarized it into useful information. The process involves an analysis of historic data and based on that analysis to predict the future occurrences or events. Predictive analytics is able to not only deal with continuous changes, but discontinuous changes as well. Classification, prediction, and to some extent, affinity analysis constitute the analytical methods employed in predictive analytics.***

KeyWords : Datamining, Association Rules, Clustering, Artificial Neural networks, Data Constraints, Patterns

I. INTRODUCTION

Knowledge discovery from databases (KDD), also known as data Mining (DM), Datamining (DM) is the extraction of new knowledge from large databases. Many techniques are currently used in this fast emerging field, including statistical analysis and machine learning based approaches. The main challenges in data mining are: • Data mining to deal with huge amounts of data located at different sites The amount of data can easily exceed the terabyte limit; • Data mining is very computationally intensive process involving very large data sets. Usually, it is necessary to partition and distribute the data for parallel processing to achieve acceptable time and space performance;

II. THE SCOPE OF DATAMINING

Data mining derives its name from the similarities between searching for valuable business information in a large database. Data mining technology can generate new business opportunities by providing these capabilities[1].

a) Automated Prediction of Trends and Behaviours

Data mining automates the process of finding predictive information in large databases. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

b) Automated Discovery of Previously unknown Patterns

Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors. The most commonly used techniques in data mining are

» Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

» Decision trees: 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 (CART) and Chi Square Automatic Interaction Detection (CHAID).

» Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

» Nearest neighbour method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbour technique.

» Rule induction: The extraction of useful if-then rules from data based on statistical significance.

III. ROOTS OF DATA MINING

A. Statistics

Without statistics,[2] there would be no data mining, as statistics are the foundation of most technologies on which data mining is built. Statistics embrace concepts such as regression analysis, standard distribution, standard deviation, standard variance, discriminate analysis, cluster analysis, and confidence intervals, all of which are used to study data and data relationships.

B. Artificial Intelligence & Machine Learning

Data mining's second longest family line is artificial intelligence and machine learning. AI is built upon heuristics as opposed to statistics, and attempts to apply human-thought like processing to statistical problems. Machine Learning could be considered as an evolution of AI, because it blends AI heuristics with advanced statistical methods. It let computer programs learn about the data they study and then apply learned knowledge to data.

C. Databases

Huge amount of data needs to be stored in a repository, and that too needs to be managed. Data warehousing also supports OLAP operations to be applied on it, to support decision making.

D. Data Mining Algorithms and 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.

A. 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. Fraud detection and credit risk applications are particularly well suited to this type of analysis. 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[3] .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
- 2.2. Clustering

B. Clustering

It can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. Types of clustering methods[6].

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based method business intelligence etc.

C. Regression analysis

It 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. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. Neural networks too can create both classification and regression models. Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

D. Association rule

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. Association Rule algorithms need to be able to generate rules with confidence values less than one. Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

E. 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.Types of neural networks.

» Back Propagation databases.

IV. Future Trends And Applications

A. Distributed Collective Datamining

Much of the data mining[5] which is being done currently focuses on a database or data warehouse of information which is physically located in one place. However, the situation arises where information may be located in different places, in different physical locations. This is known generally as distributed data mining (DDM). Therefore, the goal is to effectively mine distributed data which is located in heterogeneous sites. Distributed data mining (DDM) is used to offer a different approach to traditional approaches analysis, by using a combination of localized data analysis.

B. Ubiquitous Datamining (UDM)

UDM introduces additional cost due to communication, computation, security, and other factors. Some of the objectives of UDM are to mine the data while minimizing the cost of ubiquitous presence. Human-computer interaction is another challenging aspect of UDM. Visualizing patterns like classifiers, clusters, associations and others, in portable devices are usually difficult. The small display areas offer serious challenges to interactive data mining environments. Data management in a mobile environment is also a challenging issue. The key issues to consider include theories of UDM, advanced algorithms for mobile and distributed applications, data management issues, mark-up languages, and other data representation techniques; integration with database applications for mobile environments, architectural issues:

C. Hypertext and Hypermedia Data mining

Hypertext and hypermedia data mining can be characterized as mining data which includes text, hyperlinks, text mark-ups, and various other forms of hypermedia information. While the World Wide Web is substantially composed of hypertext and hypermedia elements, there are other kinds of hypertext/hypermedia data sources which are not found on the web. Examples of these include the information found in online catalogues, digital libraries, online information databases, and the like.. Some of the important data mining techniques used for hypertext and hypermedia data mining include classification (supervised learning), clustering (unsupervised learning), semi-structured learning, and social network analysis. In the case of classification, or supervised learning, the process starts off by reviewing training data in which items are marked as being part of a certain class or group. This data is the basis from which the algorithm is trained. Unsupervised learning, or clustering, differs from classification in that classification involved the use of training data, clustering is concerned with the creation of hierarchies of documents based on similarity, and organize the documents based on that hierarchy. Techniques which have been used for unsupervised learning include k-means clustering, agglomerative clustering, random projections, and latent semantic indexing.

D. Multimedia Data mining

Multimedia Data Mining is the mining and analysis of various types of data, including images, video, audio, and animation. The idea of mining data which contains different kinds of information is the main objective of multimedia data mining. As multimedia data mining incorporates the areas of text mining, as well as hypertext/hypermedia mining, these fields are closely related. Multimedia information, because its nature as a large collection of multimedia objects, must be represented

differently from conventional forms of data. One approach is to create a multimedia data cube which can be used to convert multimedia-type data into a form which is suited to analysis using one of the main data mining techniques, but taking into account the unique characteristics of the data. This may include the use of measures and dimensions for texture, shape, color, and related attributes. In essence, it is possible to create a multidimensional spatial database. The basic advantage of audio data mining is that while using a technique such as visual data mining may disclose interesting patterns from observing graphical displays, it does require users to concentrate on watching patterns, which can become monotonous.

E. Spatial and Geographic Data mining

Spatial and geographic data which could contain information about astronomical data, natural resources, or even orbiting satellites and spacecraft which transmit images of earth from out in space. Much of this data is image-oriented, and can represent a great deal of information if properly analyzed and mined. A definition of spatial data mining is as follows: ―the extraction of implicit knowledge, spatial relationships, or other patterns not explicitly stored in spatial databases. spatial data which differentiate it from other kinds include distance and topological information, which can be indexed using multidimensional structures, and required special spatial data access methods, The applications of these would be useful in such fields as remote sensing, medical imaging, navigation, and related uses.

F. Time Series/Sequence Data mining

Another important area in data mining centers on the mining of time series and sequence-based data. Simply put, this involves the mining of a sequence of data, which can either be referenced by time (time-series, such as stock market and production process data), or is simply a sequence of data which is ordered in a sequence. These can include long-term or trend movements, seasonal variations, cyclical variations, and random movements (Han and Kamber, 2001). Sequential pattern mining has as its focus the identification of sequences which occur frequently in a time series or sequence of data. This is particularly useful in the analysis of customers, where certain buying patterns could be identified, such as what might be the likely follow-up purchase to purchasing a certain electronics item or computer

G. Constraint – Based Data mining

This form of data mining incorporates the use of constraints which guides the process. Frequently this is combined with the benefits of multidimensional mining to add greater power to the process. There are several categories of constraints which can be used, each of which has its own characteristics and purpose. These are:

- Knowledge-type constraints: This type of constraint specifies the ―type of knowledge‖ which is to be mined, and is typically specified at the beginning of any data mining query.
- Data constraints: This constraint identifies the data which is to be used in the specific data mining query. Data constraints can be specified in a form similar to that of a SQL query.
- Dimension/level constraints: Information being mined is in the form of a database or multidimensional data warehouse, it is possible to specify constraints which specify the levels or dimensions to be included in the current query.
- Interestingness constraints: It would also be useful to determine what ranges of a particular variable or measures are considered to be particularly interesting and should be included in the query.
- Rule constraints: It is also important to specify the specific rules which should be applied and used for a particular data mining query or application. One application of the constraint-based approach is in the Online Analytical Mining Architecture (OLAM) developed by Han, Lakshamanan, and Ng, 1999.

H. Phenomenal Data Mining

Phenomenal data mining is not a term for a data mining project that went extremely well. Instead, it focuses on the relationships between data and the phenomena which are inferred from the data . One example of this is that using receipts from cash supermarket purchases, it is possible to identify various aspects of the customers who are making these purchases.

V. TRENDS IN DATA MINING

A. Historical Trends

Data mining application era was perceived in early 1980s principally focused on single tasks driven by research tools. Data mining is helpful in various disciplines like Data Base Management Systems (DBMS), Artificial Intelligence (AI), Machine Learning (ML) and Statistics. Historical trends of data mining are explained as follows [4]: Data mining algorithm work best with the numerical data especially collected from a single data base and various data mining techniques have developed for flat files, traditional and relational database where the data is mostly represented in the tabular form. Afterwards, with the convergence of Statistics and Machine Learning pave way to the evolution of various algorithms to mine the non numerical data and relational data bases. Development in fourth generation programming language influenced much in the field of data mining and various related computing techniques. Initially, most of the algorithms engaged to work only on statistical techniques. Various computing techniques such as AI, ML and pattern reorganization evolved to do the data mining tasks in ease manner. Various data mining techniques like Induction, Compression, approximation and other algorithms developed to mine the large volume of heterogeneous data stored in the data warehouse.

B. Current Trends

Advancement in data mining with various integrations and implications of the methods and techniques have formed the present data

C. Future Trends

Data mining has been acquiring noteworthy amount of importance in recent years and it has a strong industrial impact. Future of data mining companies would be promising in the coming years based on this observation. A huge amount of data gets agitate in the research, medical, corporate and media industries as it becomes great for anybody involves in gathering useful information. Increasing technology and future application areas always creates new challenges and opportunities for data mining. Advance data mining techniques can be developed and used by R& D and other information rich companies to discover useful patterns that can help in research or business development to ensure the growth and development of the companies. Future data mining technologies involve standardization of data mining languages; predictive analysis, advanced text mining, Semantic and image mining are discussed as follows [20]:

A. Standardization of Data Mining Languages

Different syntaxes are used in various data mining tools, hence standardized syntaxes needs to be developed in order to make convenient coding for the users. Standardization of interaction language and flexible user interaction has to be much concentrated by the data mining applications [4].

B. Predictive Analysis

In earlier days of data mining whereby assumptions about structure of data were unheard where as now a days, data is put through algorithms based on certain attributes such as trends, relations and patterns and predictions are thereby projected. This paves way for significant increase in decision making capabilities especially in business process. For instance, predicting customer behaviors with the help of mathematical modeling and statistical analysis, their spending habits on their credit cards

can be determined and credit point allotted accordingly. This kind of predictive analysis can create huge impact in the near future and business can propagate in well manner based on such predictions [20].

C. Advanced Text Mining

In earlier times, text mining was only performed on structured data. But, majority of unstructured data are available in the form of memos, emails, surveys, notes, chats, whitepapers, forums, presentation, etc. It can be tapped and accessed using data mining services. Vast amount of information can be gathered using such text mining techniques and this can be used effectively for the business purpose. This is taking data mining a step further from earlier times [20].

D. Semantic and Image Mining

Semantic and image mining will take a predominant stage in future as researchers will be able to find hidden meaning in data and document using artificial intelligence and structural analysis software. Images can be searched for identifying patterns and the information derived can be used for various scientific and business advancements. Plenty of opportunities will be opened through the data mining services offered by various professional data mining companies [20].

Comparison Fast, Current and Future Trends in Data mining

	Algorithms/ Techniques Employed	**Data Formats**	**Computing Resources**
Past	Statistical, Machine Learning Techniques	Numerical data and structured data stored in traditional databases	Evolution of 4G PL and various related techniques
Current	Statistical, Machine Learning, Artificial Intelligence, Pattern Reorganization Techniques	Heterogeneous data formats includes structured, semi structured and unstructured data	High speed networks, High end storage devices and Parallel, Distributed computing etc…
Future	Soft Computing techniques like Fuzzy logic, Neural Networks and Genetic Programming	Complex data objects includes high dimensional, high speed data streams, sequence, noise in the time series, graph, Multi instance objects, Multi represented objects and temporal data etc…	Multi-agent technologies and Cloud

VI. Advantages Of Data Mining In Various Applications

Advantages of using data mining in various applications such as Banking, Manufacturing and production, marketing, health care etc., are as follows[7]:

A. Banking:

Data mining supports banking sector in the process of searching a large database to discover previously unknown patterns; automate the process of finding predictive information. Data mining helps to forecast levels of bad loans and fraudulent credit cards use, predicting credit card spending by new customers and predicting the kinds of customer best respond to new loan offered by the backs.

B. Manufacturing And Production

Data mining helps to predict the machine failures and finding key factors that control optimization of manufacturing capacity.

3) Marketing: Data mining facilitates marketing sector by classifying customer demographic that can be used to predict which customer will respond to a mailing or buy a particular product and it is very much helpful in growth of business.

C. Health-Care: Data mining

Supports a lot in health care sector. It supports health care sector by correlating demographics of patients with critical illnesses, developing better insights on symptoms and their causes and learning how to provide proper treatments

D. Insurance

Data mining assist insurance sector in predicting fraudulent claims and medical coverage cost, classifying the important factors that affect medical coverage and predicting the customers' pattern which customer will buy new policies.

E. Law

Law enforcement is helped by data mining by monitoring the behaviour patterns of the criminals. Tracking crime pattern, locations and criminal behaviours, identifying various attributes to data mining, assist in solving criminal cases.

F. Government and Defense

Data mining helps to forecast the cost of moving military equipment and predicting resource consumption. Apart from that it assists in testing strategies for potential military engagements and improving homeland security by mining data from many sources.

G. Brokerage And Securities Trading

Data mining assists in predicting the change in bond prices and forecasting the range of stock fluctuation determining when to buy or sell stocks.

H. Computer hardware and software

Predicting disk-failures and potential security violations can be done by data mining.

I. Airlines

It supports in checking the feasibility of adding routes to increase the business profit and to decrease the loss by capturing data on where passengers are flying and the ultimate destination of passengers.

VII. THE DISADVANTAGES OF DATA MINING

A. Privacy Issues

One of the disadvantages is a personal privacy issue. In recent years, with the boom of internet, the concerns about privacy have increased tremendously. Because of this privacy concern, individuals like internet users, employees, customers are afraid that unknown person may have access to their personal information and then use that information in an unethical way and this may cause harm to them. Although, several laws have protected the users to sell or trade personal information between different organisation, selling personal information have occurred [2][7].

B. Security Issues

Another biggest disadvantage is security issue which is always a major concern in information technology. Companies have a lot of personal information about the employees and customers including social security number, birthdates, payroll etc., and it is also available in online. But, they do not have sufficient security systems in place to protect this information. They have been a lot of cases where hackers access and stole personal data of customers [2][7].

C. Misuse of Information/Inaccurate information

Trends obtain from the data mining intended to be used for business or some ethical purpose. However it may be misused for other unethical purpose. Unethical businesses or Individual may use the information to take advantage of vulnerable people

or to discriminate against a certain group of people. Apart from that, data mining techniques is not cent percent accurate one. Thus mistakes may happen which can have serious consequence [8].

D. Challenges of Data Mining

There are many challenges faced by the data mining and these challenges of data mining are pointed as follows [2][9][10]:

» Scalability

» Complex and Heterogeneous Data

» Network Setting

» Data Quality

» Data Ownership and Distribution

» Dimensionality

» Privacy preservation

» Streaming Data

VIII. CONCLUSION

Data mining will be considered one of the most important frontiers and one of the most promising interdisciplinary developments in Information technology. Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology. Data mining is useful for both public and private sectors for finding patterns, forecasting, discovering knowledge in different domains such as finance, marketing, banking, insurance, health care and retailing. Data mining is commonly used in these domains to increase the sales, to reduce the cost and enhance research to reduce costs, enhance research.

References

1. Data Mining Concepts and Techniques – Jiawei Han & Micheline Kamber
2. Heikki, Mannila, ―Data mining: machine learning, statistics, and databases‖, Statistics and Scientific Data Management, pp. 2-9. 1996.
3. Knowledge Discovery in Databases, AAAI Press / the MIT Press, Massachusetts Institute of Technology. ISBN 0– 26256097–6. MIT1996.
4. Chakrabarti, van den Berg, and Dom. ―Distributed Hypertext Resource Discovery through Examples, ―Proceedings of the 25thVLDB (International Conference on Very Large Data Bases), Edinburgh Scotland, 1999.
5. Han, J. and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001.
6. Han, J., M. Kamber, and A. K. H. Tung, "Spatial Clustering Methods in Data Mining: A Survey", H. Miller and J. Han (eds.), Geographic Data Mining and Knowledge Discovery, Taylor and Francis, 2001.

