International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182, Print ISSN: 2349-5979, Impact Factor: RJIF 5.72
www.allsubjectjournal.com
Volume 4; Issue 5; May 2017; Page No. 03-06

A brief study and analysis of data mining techniques

M. Porkizhi
Ph.D., Research Scholar, Rathinam College of Arts & Science, Coimbatore, Tamil Nadu, India

Abstract
Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It plays a significant role on human activities and has become an essential component in various fields of human life. Data mining is greatly inspired by advancements in Statistics, Machine Learning, Artificial Intelligence, Pattern Recognition and Computation capabilities. It is the knowledge discovery process which analyzes the large volumes of data from various perspectives and summarizes it into useful information. This paper focuses the major study of data mining, its scopes, its tools and techniques, its applications and advantages/disadvantages.

Keywords: data mining, tools, techniques, applications

Introduction
Data mining is the computational process of discovering patterns in large data sets. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Data mining is most often applied to extraction of useful knowledge from business data however it is also useful in some scientific applications where this more empirical approach complements traditional data analysis. It is an essential ingredient in the more general process of Knowledge Discovery in Databases (KDD). Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. The emergence is due to the growth in data warehouses and the realization that this mass of operational data has the potential to be exploited as an extension of Business Intelligence. In Earlier data Mining used similar manual approaches to review data and provide business projections for many years. Changes in data mining techniques, however, have enabled organizations to collect, analyze, and access data in new ways.

Objectives of data mining
Data mining techniques are employed with two main objectives:
To improve our understanding of the relevant factors and their relationships, including the possible discovery of non-obvious features in the data that may suggest better formulations of the physical models.

- To induce models solely from the data so that dynamical simulations might be compared to them and that they may also have utility, offering (short term) predictive power.

Types of data mining
The hypertext and hypermedia data is a collection of data from online catalogues, digital libraries, and online information data bases which include hyperlinks, text markups and other forms of data. Web mining is the application of data mining to discover the patterns from the Web. The important data mining technique used for hypertext and hypermedia data are Classification (supervised learning), Clustering (unsupervised learning) [2, 3].

Ubiquitous data mining
The advent of laptops, palmtops, cell phones, and wearable computer devices with increasing computational capacity and proliferation of all these devices is leading to the emergence of ubiquitous computing paradigm. The Ubiquitous computing environments are subsequently giving rise to a new class of applications termed Ubiquitous Data Mining (UDM). UDM is the process of analysis of data for extracting useful knowledge from the data of ubiquitous computing. Traditional data mining techniques that are drawn from the combination of ML and Statistics are presently employed in ubiquitous data mining.

Multimedia data mining
The multimedia data includes images, video, audio, and animation. The data mining techniques that are applied on multimedia data are rule based decision tree classification algorithms like Artificial Neural Networks, Instance-based learning algorithms, Support Vector Machines, also association rule mining, clustering methods.

Spatial data mining
The spatial data includes astronomical data, satellite data and space craft data. Some of the data mining techniques and data structures which are used when analyzing spatial and related types of data include the use of spatial warehouses, spatial data cubes, spatial OLAP, and spatial clustering methods.

Time series data mining
A time series is a sequence of data points, measured typically at successive times spaced at uniform time intervals. Typical examples include stock prices, currency exchange rates, the

volume of product sales, biomedical measurements, weather data, etc, collected over monotonically increasing time.

Scope and Nature of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality data mining technology can generate new business opportunities by providing these capabilities:

• Automated prediction of trends and behaviors.

Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly.

• Automated discovery of previously unknown patterns.

Data mining tools sweep through databases and identify previously hidden patterns in one step.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

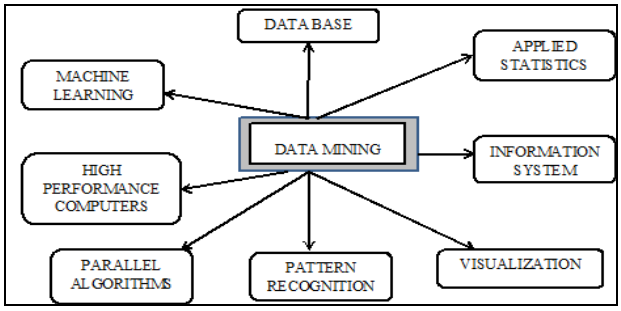

Fig 1: Scope of data mining.

'Is data mining as useful in science as in commerce?' is an important question as far as the nature of data mining is to be made clear. Certainly data mining in science has much in common with that for business data. One difference though is that there is a lot of existing scientific theory and knowledge hence there is less chance of knowledge emerging purely from data however empirical results can be valuable in science (especially where it borders on engineering) as in suggesting causality relationships or for modeling complex phenomena. Another difference is that in commerce rules are soft sociological or cultural and assume consistent behavior.

Tools and phases of data mining

Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. To enumerate a precise list of data mining tool characteristics is quite difficult since the tools governing the process of data mining are not standardized. They are not specific and most of the times various approaches and tools result in data mining and it generates families of In spite of the lack of precise standards, we may conclude that data mining is subject to four phases viz.,

- Data preparation
- Data analysis and classification
- Knowledge acquisition
- Prognosis

They are followed one after the other that is in sequence and are clear from the following figure

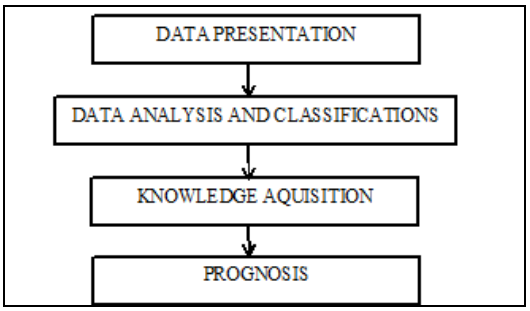

Fig 2: Tools and Phases of data mining.

In the data preparation phase the main data sets to be used by the data mining operation are identified and cleansed of any impurities as data in the data warehouse. As the data warehouses are already integrated and filtered, data warehouse usually is the target set for data mining operation. Most data mining tools can be classified into one of three categories: traditional data mining tools, dashboards, and text-mining tools. Below is a description of each [9].

Traditional Data Mining Tools

Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques. Some of these tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database. The majority are available in both Windows and UNIX versions, although some specialize in one operating system only. In addition, while some may concentrate on one database type, most will be able to handle any data using online analytical processing or a similar technology.

Dashboards: Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen — often in the form of a chart or table — enabling the user to see how the business is performing. Historical data also can be referenced, enabling the user to see where things have changed (e.g., increase in sales from the same period last year). This functionality makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.

Text-mining Tools

The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text — from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications. Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database). Capturing these inputs can provide organizations with a wealth of information that can be mined to discover trends, concepts, and attitudes.

Data Mining Techniques and Its Application

In addition to using a particular data mining tool, internal auditors can choose from a variety of data mining techniques. The most commonly used techniques include artificial neural networks, decision trees, and the nearest-neighbor method [4]. Each of these techniques analyzes data in different ways:

Artificial neural networks are non-linear, predictive models that learn through training. Although they are powerful predictive modelling techniques, some of the power comes at the expense of ease of use and deployment. One area where auditors can easily use them is when reviewing records to identify fraud and fraud-like actions. Because of their complexity, they are better employed in situations where they can be used and reused, such as reviewing credit card transactions every month to check for anomalies.

Decision trees are tree-shaped structures that represent decision sets. These decisions generate rules, which then are used to classify data. Decision trees are the favoured technique for building understandable models. Auditors can use them to assess, for example, whether the organization is using an appropriate cost-effective marketing strategy that is based on the assigned value of the customer, such as profit.

The nearest-neighbour method classifies dataset records based on similar data in a historical dataset. Auditors can use this approach to define a document that is interesting to them and ask the system to search for similar items. The techniques used for data mining can be broadly categorized into three types as in following figure.

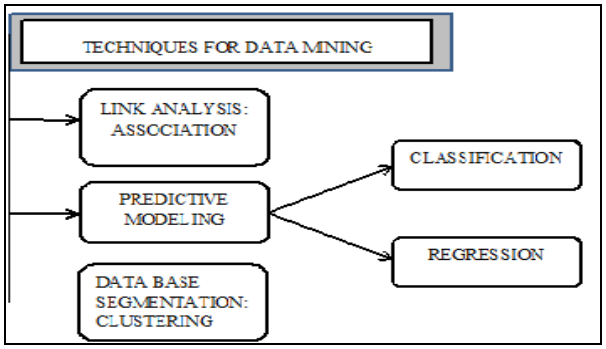

Fig 3: Techniques for data mining.

How Data Mining Works

'How exactly is data mining able to tell you important things that you didn't know or what is going to happen next?' is important consideration. The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where you know the answer and then applying it to another situation that you don't. For instance if you were looking for a sunken galleon on the high seas the first thing you might do is to research the times when treasure had been found by others in the past. You might note that these ships often tend to be found off the coasts and that there are certain characteristics to the ocean currents and certain routes that have likely been taken by the ship's captains in that era. You note these similarities and build a model that includes the characteristics that are common to the locations of these sunken treasures. With these models in hand you sail off looking for treasure where your model indicates it most likely might be given a similar situation in the past. Hopefully if you've got a good model you find your treasure.

Conclusion

The field of data mining has been greatly influenced by the development of fourth generation programming languages and computing techniques. Data mining evolved with various computing techniques like AI, ML and Pattern Reorganization. Various data mining techniques (Induction, Compression and Approximation) and algorithms developed to mine the large volumes of heterogeneous data stored in the data warehouses. The term is of utmost importance in present scenario where every business is taking benefits from the concept.

References

1. Heikki, Mannila. Data mining: machine learning, statistics, and databases, IEEE, 1996.
2. Piatetsky-Shapiro, Gregory. The Data-Mining Industry Coming of Age. IEEE Intelligent Systems, 2000.
3. Salmin, Sultana *et al*. Ubiquitous Secretary: A Ubiquitous Computing Application Based on Web Services Architecture, International Journal of Multimedia and Ubiquitous Engineering. 2009, 4(4).
4. Hsu J. Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st Century, the Proceedings of the 19th Annual Conference for Information Systems Educators, 2002. ISSN: 1542-7382.
5. Baker ZK, Prasanna VK. Efficient Parallel Data Mining with the Apriori Algorithm on FPGAs. In Submitted to the IEEE International Parallel and Distributed Processing Symposium (IPDPS '05), 2005.
6. Jing He. Advances in Data Mining: History and Future, Third international Symposium on Information Technology, 2009. Application, 978-0-7695-3859-4/09 IEEE 2009 DOI 10.1109/IITA.
7. Han, J, Kamber M. Data mining: Concepts and techniques .Morgan-Kaufman Series of Data Management Systems. San Diego: Academic Press, 2001.
8. http://en.wikipedia.org/wiki/Data_mining
9. http://www.statsoft.com/textbook/stdatmin.html

10. Kieffer JC, Yang EH. Grammar-Based Codes: A New Class of Universal Lossless Source Codes, IEEE Transactions on Information Theory. 2000; 46:737-754.
11. John Haggerty, Qi Shi and Madjid Merabti, Statistical Signatures for Early Detection of Flooding Denial-Of service Attacks, Springer Boston, 2006.
12. Giovanni Vigna, Sumit Gwalani, Kavitha Srinivasan, Elizabeth M. Belding- Royer and Richard A. Kemmerer , An Intrusion Detection Tool for AODVbased Ad hoc Wireless Networks, IEEE Computer Society Washington, DC, USA , 2004.
13. Shukor Abd Razak, Steven Furnell, Nathan Clarke, and Phillip Brooke, A Two-Tier Intrusion Detection System for Mobile Ad Hoc Networks – A Friend Approach, Springer-Verlag Berlin Heidelberg, 2006.
14. Eduardo Mosqueira-Rey, Amparo Alonso-Betanzos, Belen Baldonedo Del Rio, and Jesus Lago Pineiro, A Misuse Detection Agent for Intrusion Detection in a Multi-agent Architecture. Springer-Verlag Berlin Heidelberg, 2007.
15. Magnus Almgren, Ulf Lindqvist, and Erland Jonsson, A Multi-Sensor Model to Improve Automated Attack Detection, Springer-Verlag Berlin Heidelberg, 2008.

.