International Journal of Current Engineering and Technology
ISSN 2277 - 4106

Available at http://inpressco.com/category/ijcet

Review Article

A Review on Data Mining: Its Challenges, Issues and Applications

Bhoj Raj Sharma[a*], Daljeet Kaur[a] and Manju[b]

[a]Department of Computer Science, Eternal University, Baru Sahib, Sirmour (H.P)
[b]Computer Science Department,BMJ Group of Colleges,Bathinda,(PB)

Accepted 20 June 2013, Available online 25June 2013, Vol.3, No.2 (June 2013)

Abstract

Data mining is described as a process of discover or extracting interesting knowledge from large amounts of data stored in multiple data sources such as file systems, databases, data warehouses etc. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Powerful systems for collecting data and managing it in large databases are already in place in most large and mid-range companies. However the bottleneck of turning this data into your success is the difficulty of extracting knowledge about the system that you study from the collected data. Data mining and its techniques can be extremely beneficial in many areas such as industry, commerce, government, education, and agriculture, healthcare and so on .data mining tools can analyze massive databases to deliver answers to questions such as, Which clients most likely to respond to my next promotional mailing, and why? Data mining have many advantages but still data mining systems face lot of problems and pitfalls. The purpose of this paper is to discuss Role of data mining, its application and various challenges and issues related to it.

Key words*: Data Mining, Application, challenges,issues, Pros&Cons.*

Introduction

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. Data mining is the analysis of observational data sets to find.

Unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.Data mining is an interdisciplinary field bringing together techniques from Machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases.(Agrawal R *et al*,1994)

The growth in the field of data mining and knowledge discovery has been fastened by a variety of factors:

- The growth in data collection, as exemplified by the supermarket.
- The storing of the data in data warehouses, so that the entire enterprise has access to a reliable current database.
- The availability of increased access to data from Web navigation and intranets.
- The competitive pressure to increase market share in a globalized economy.
- The tremendous growth in computing power and storage capacity.

*Corresponding author: Bhoj Raj Sharma

Data mining functionality

The data mining functionalities and the variety of knowledge discovered. Association analysis (Savasre A. *et al*,1995) is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis. For example, it could be useful for the Our Video Store manager to know what movies are often rented together or if there is a relationship between renting a certain type of movies and buying popcorn or pop. The discovered association rules are of the form: P -> Q [s, c], where P and Q are conjunctions of attribute value-pairs, and s (for support) is the probability that P and Q appear together in a transaction and c (for confidence) is the conditional probability that Q appears in a transaction when P is present. For example, the hypothetic association rules:

RentType(X,game)ANDAge(X,13-19)->Buys(X,pop)[s=2%,c=55%]

would indicate that 2% of the transactions considered are of customers aged between 13 and 19 who are renting a game and buying a pop, and that there is a certainty of 55% that teenage customers who rent a game also buy pop. Classification analysis is the organization of data in

given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the OurVideoStore managers could analyze the customers? behaviours vis-à-vis their credit, and label accordingly the customers who received credits with three possible labels safe, risky and very risky. The classification analysis would generate a model that could be used to either accept or reject credit requests in the future (Yanthy W. *et al*,2009). In case of classification following terms are important:

Accuracy: It gives a measure for the overall accuracy of the classifier:

Accuracy (%) = Number of correctly classified instances * 100 / Number of instances.

Precision and recall: With respect to classifiers:

Precision(X) = Number of correctly classified instances of class X / Number of instances classified as belonging to class X.

Recall(X) = Number of correctly classified instances of class X / Number of instances in class X.

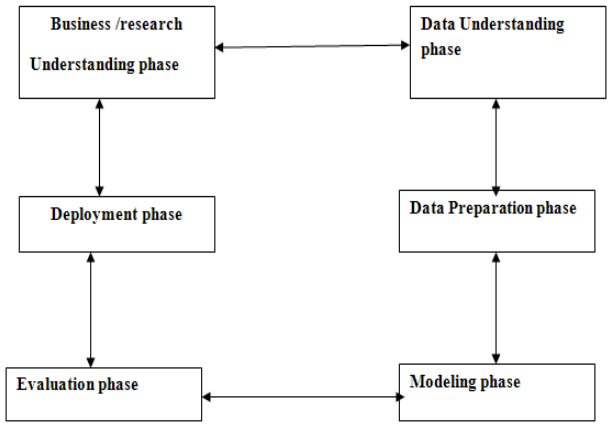
Confusion matrix: Confusion matrices are very useful for evaluating classifiers, as they provide an efficient snapshot of its performance, by displaying the distribution of correct and incorrect instances

It is common that users do not have a clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore important to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction. This also makes interactivity an important attribute of a data mining system.

The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in order to obtain good visualization tools for large datasets that could be used to display and manipulate mined knowledge. The major issues related to user interfaces and visualization are screen real-estate, information rendering, and interaction. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels.

Most algorithms assume the data to be noise-free. This is of course a strong assumption. Most datasets contain exceptions, invalid or incomplete information, etc., which may complicate, if not obscure, the analysis process and in many cases compromise the accuracy of the results. As a consequence, data preprocessing (data cleaning and transformation) becomes vital. It is often seen as lost time, but data cleaning, as time-consuming and frustrating as it may be, is one of the most important phases in the knowledge discovery process. Data mining techniques should be able to handle noise in data or incomplete information. More than the size of data, the size of the search space is even more decisive for data mining techniques. The size of the search space is often depending upon the number of dimensions in the domain space. The search space usually grows exponentially when the number of dimensions increases. This is known as the curse of dimensionality. This curse affects so badly the performance of some data mining approaches that it is becoming one of the most urgent issues to solve.

According to CRISP–DM, a given data mining project has a life cycle consisting of six phases, as illustrated in Figure 1.1. Note that the phase sequence is adaptive. That is, the next phase in the sequence often depends on the outcomes associated with the preceding phase. The most significant dependencies between phases are indicated by the arrows. For example, suppose that we are in the modeling phase. Depending on the behavior and characteristics of the model, we may have to return to the data preparation phase for further refinement before moving forward to the model evaluation phase.



Life cycle of data mining

Evolution of Data Mining

The evolution of data mining began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery(Fayyad U *et al*, 1996).

Data mining is ready for application in the business world because it is supported by three technologies that are now sufficiently mature:

Massive data collection

Powerful multiprocessor computers

Data mining algorithms

Commercial databases are growing at rapid rate. The accompanying need for improved computational engines can now be met in a cost effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods. In the evolution from business data to business information, each new step has built upon the previous one. From the user's point of view, the four steps listed below were revolutionary because they allowed new business questions to be answered accurately and quickly.

Data Collection (1960s): Answered questions like What was my total income in the last five years?

Data Access (1980s): Answered business questions like What were unit sales in India last year? Relational Databases (RDBMS), Structured Query Language (SQL), etc. were used for querying and reporting.

Data Warehousing & Decision Support (1990s): These technologies were capable of answering business questions like What were sales last year? The technologies used are On-line analytic processing, multidimensional databases, data warehouse etc.

Data Mining (Emerging Today): Capable of answering questions like How many people will buy black Car next year? Uses advanced algorithms, multiprocessor computers, massive databases, etc.

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments. The data mining is having various challenges and issues which are discussed below(Berry M *et al*,2002).

Data mining challenges

The shift towards intrinsically distributed complex problem solving environments is prompting a range of new data mining research and development problems. These can be classified into the following broad challenges:

Distributed data: The data to be mined is stored in distributed computing environments on heterogeneous platforms. Both for technical and for organizational reasons it is impossible to bring all the data to a centralized place. Consequently, development of algorithms, tools, and services is required that facilitate the mining of distributed data(Mechitov A. et al,2001).

Distributed operations: In future more and more data mining operations and algorithms will be available on the grid. To facilitate seamless integration of these resources into distributed data mining systems for complex problem solving, novel algorithms, tools, grid services and other IT infrastructure need to be developed.

Massive data: Development of algorithms for mining large, massive and high-dimensional data sets (out-of-memory, parallel, and distributed algorithms) is needed.

Complex data types: Increasingly complex data sources, structures, and types (like natural language text, images, time series, multi-relational and object data types etc.) are emerging. Grid-enabled mining of such data will require the development of new methodologies, algorithms, tools, and grid services.

Data privacy, security, and governance: Automated data mining in distributed environments raises serious issues in terms of data privacy, security, and governance. Grid-based data mining technology will need to address these issues.

User-friendliness: Ultimately a system must hide technological complexity from the user. To facilitate this, new software, tools, and infrastructure development is needed in the areas of grid-supported workflow management, resource identification, allocation, and scheduling, and user interfaces.

Data mining issues

There are many important implementation issues associated with data mining:

Human interaction: Since data mining problems are often not precisely stated, interfaces may be need with both domain and technical experts. Technical experts are used to formulate the queries and assist in interpreting the results. Users are needed to identify training data and desired results.

Over-fitting: When a model is generated that is associated with a given database state, it is desirable that the model also fit future database states. Over-fitting occurs when the model does not fit future states. This may be caused by assumptions that are made about the data or may simply be caused by the small size of the training database. For example, a classification model for a student database may be developed to classify students as an excellent, good, or average. If the training database is quite small, the model might erroneously indicate that an excellent student is anyone who scores more than 90% because there is only one entry in the training database under 90%. In this case, many future students would be erroneously classified as an excellent. Over-fitting can arise under other circumstances as well, even though the data are not changing.

Outliers: There are often many data entries that do not fit nicely into the derived model. This becomes even more of an issue with very large databases. If a model is developed that includes these outliers, then the model may not behave well for data that are not outliers.

Social One: One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences.

Data integrity: Another issue is that of data integrity. Clearly, data analysis can only be as good as the data that is being analyzed. A key implementation challenge is

integrating conflicting or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases. The addresses (or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.

Interpretation of results: Currently, data mining output may require experts to correctly interpret the results, which might otherwise be meaningless to the average database user.

Visualization of results: To easily view and understand the output of data mining algorithms, visualization of the results is helpful.

Large datasets: The massive datasets associated with data mining create problems when applying algorithms designed for small datasets. Many modeling applications grow exponentially on the dataset size and thus are too inefficient for larger datasets. Sampling and parallelization are effective tools to attack this scalability problem.

High dimensionality: A conventional database schema may be composed of many different attributes. The problem here is that not all attributes may be needed to solve a given data mining problem. In fact, the use of some attributes may interfere with the correct completion of a data mining task. The use of other attributes may simply increase the overall complexity and decrease the efficiency of an algorithm. This problem is sometimes referred to as the dimensionality curse, meaning that there are many attributes (dimensions) involved and it is difficult to determine which ones should be used. One solution to this high dimensionality problem is to reduce the number of attributes, which is known as dimensionality reduction. However determining which attributes not needed is not always easy to do.

Multimedia data: Most previous data mining algorithms targeted to traditional data types (numeric, character, text, etc.). The use of multimedia data such as is found in GIS databases complicates or invalidates many proposed algorithms.

Relational or Multidimensional databases: A hotly debated technical issue is whether it is better to set up a relational database structure or a multidimensional one. In a relational structure, data is stored in tables, permitting ad hoc queries. In a multidimensional structure, on the other hand, sets of cubes are arranged in arrays, with subsets created according to category. While multidimensional structures facilitate multidimensional data mining, relational structures thus far have performed better in client/server environments. And, with the explosion of the Internet, the world is becoming one big client/server environment.

Noisy data: Some attribute value might be invalid or incorrect. These values are often corrected before running data mining applications.

Irrelevant data: Some attributes in the database might not be of interest to the data mining task being developed.

Missing data: During the pre-processing phase of knowledge discovery in databases (KDD), missing data may be replaced with estimates. This and other approaches to handling missing data can lead to invalid results in the data mining step.

Changing data: Databases cannot be assumed to be static. However, most data mining algorithms do assume a static database. This requires that the algorithm be completely rerun anytime the database changes.

Application: Determining the intended use for the information obtained from the data mining function is a challenge. Indeed, how business executives can effectively use the output is sometimes considered the more difficult part, not the running of the algorithms themselves. Because the data are of a type that has not previously been known, business practices may have to be modified to determine how to effectively use the information uncovered (Zurada J. *et al*,2005).

Cost: Finally, there is the issue of cost. While system hardware costs have dropped dramatically within the past years, data mining and data warehousing tend to be self-reinforcing. The more powerful the data mining queries, the greater the utility of the information being gleaned from the data, and the greater the pressure to increase the amount of data being collected and maintained, which increases the pressure for faster, more powerful data mining queries. This increases pressure for larger, faster systems, which are more expensive(Baazaoui Z et al,2005).

Data Mining Applications in Sales/Marketing

Data mining enables businesses to understand the hidden patterns inside historical purchasing transaction data, thus helping in planning and launching new marketing campaigns in prompt and cost effective way(Berry M.J.A. *et al*,1997). The following illustrates several data mining applications in sale and marketing.

Data mining is used for market basket analysis to provide information on what product combinations were purchased together, when they were bought and in what sequence. This information helps businesses promote their most profitable products and maximize the profit. In addition, it encourages customers to purchase related products that they may have been missed or overlooked(Pang-Ning T *et al*,2012).

Retail companies' uses data mining to identify customer's behavior buying patterns.

Data Mining Applications in Banking / Finance

Several data mining techniques e.g., distributed data mining have been researched, modeled and developed to help credit card fraud detection.

Data mining is used to identify customer's loyalty by analyzing the data of customer's purchasing activities such as the data of frequency of purchase in a period of time, total monetary value of all purchases and when was the last purchase. After analyzing those dimensions, the relative measure is generated for each customer. The higher of the score, the more relative loyal the customer is. To help bank to retain credit card customers, data mining is applied. By analyzing the past data, data mining can

help banks predict customers that likely to change their credit card affiliation so they can plan and launch different special offers to retain those customers.

Credit card spending by customer groups can be identified by using data mining.

The hidden correlations between different financial indicators can be discovered by using data mining.

From historical market data, data mining enables to identify stock trading rules.

Data Mining Applications in Health Care and Insurance

The growth of the insurance industry entirely depends on the ability of converting data into the knowledge, information or intelligence about customers, competitors and its markets. Data mining is applied in insurance industry lately but brought tremendous competitive advantages to the companies who have implemented it successfully. (Berson A *et al*,1997)The data mining applications in insurance industry are listed below:

Data mining is applied in claims analysis such as identifying which medical procedures are claimed together (Kusiak A *et al*,2000)

Data mining enables to forecasts which customers will potentially purchase new policies.

Data mining allows insurance companies to detect risky customers' behavior patterns.

Data mining helps detect fraudulent behavior.

Data Mining Applications in Transportation

Data mining helps determine the distribution schedules among warehouses and outlets and analyze loading patterns.

Data Mining Applications in Medicine

Data mining enables to characterize patient activities to see incoming office visits. Data mining helps identify the patterns of successful medical therapies for different illnesses(Antonie M *et al*,2001).

Data mining applications are continuously developing in various industries to provide more hidden knowledge that increases business efficiency and grows businesses *(*Awan MSK *et al*,1999*).*

Pros and cons of Data Mining

Data mining has a lot of pros when using in a specific industry. Besides those pros, data mining also has its own cons e.g., privacy, security and misuse of information. We will examine those **pros and cons of data mining** in different industries in a greater detail.

Pros of Data Mining

Marketing / Retail:-Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers.

Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

Finance / Banking:-Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

Manufacturing:-By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example semi-conductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality(Awan MSK et al,1999).

Governments:-Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities.

Cons of data mining

Privacy Issues:-The concerns about the personal privacy have been increasing enormously recently especially when internet is booming with social networks, e-commerce, forums, blogs. Because of privacy issues, people are afraid of their personal information is collected and used in unethical way that potentially causing them a lot of troubles. Businesses collect information about their customers in many ways for understanding their purchasing behaviors trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time the personal information they own probably is sold to other or leak.

Security issues:-Security is a big issue. Businesses own information about their employees and customers including social security number, birthday, payroll and etc. However how properly this information is taken care is still in questions. There have been a lot of cases that hackers accessed and stole big data of customers from big corporation such as Ford Motor Credit Company, Sony. with so much personal and financial information available, the credit card stolen and identity theft become a big problem.

Misuse of information/inaccurate information:-Information is collected through data mining intended for

the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people. In addition, data mining technique is not perfectly accurate. Therefore if inaccurate information is used for decision-making, it will cause serious consequence.

Misuse of information/inaccurate information:- Information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people.In addition, data mining technique is not perfectly accurate. Therefore if inaccurate information is used for decision-making, it will cause serious consequence(Jain B.A et al,1997).

Conclusion

Data Mining brings a lot of benefits to businesses, society, governments as well as individual. Howeverprivacy, security and misuse of information are the major problems if they are not addressed and resolved properly. Moreover,sincethe data mining process is systematic, it offers enterprises/government the ability to discover hidden patterns in their data-patterns that can help them understand customer behavior and market trends.

In this paper, the concept of data mining, role of data mining its major challenges, issues and application have been focused which help in business strategy formulations, decision making and analysis to the business, society and governments.

References

Agrawal R., Shrikant R., (1994),Fast algorithms for mining association rule. In the *proceeding of 20th international conference on VLDB*, pp487-499.

Savasre A., Omienciski E., and Navathe S.,(1995), An efficient algorithm for mining association rules in large databases. In the *proceeding of 21st international conference on VLDB*, pp. 432-444.

Yanthy W., Sekiya T., Yamaguchi K., (2009), Mining Interesting Rules by Association and Classification Algorithms. In the *proceeding of International Conference on Frontier of Computer Science and Technology*, pp. 177182.

Fayyad U., Piatetsky-Shapiro G., and Smyth P. (1996),Knowledge Discovery and Data Minning: Towards a Unifying Framework, Proc, 2nd *Int. Conf. on Knowledge Discovery and Data Mining, Portland*, OR, pp. 82-88.

Zurada J., and Subhash L.(2004), Application of Data Mining Methods for Bad Debt Recovery in the Healthcare Industry, in *Proceedings of the 6th International Conference on Database and Information Systems, J. Barzdins (Ed.)*, vol. 672, pp. 207-217.

Berson A., Smith S. and Thearling K. Building Data Mining Applications for CRM, *McGraw-Hill Professional.*

Berry M.J.A., and Linoff, G.S.(1999), Data Mining Techniques for Marketing, Sales, and Customer Support, *John Wiley & Sons, Inc*,

Jain B.A., and Nag B.N.(1997), Performance Evaluation of Neural Network Models, *Journal of Management Information Systems*, vol. 14, No. 2, 201-216.

Malik ShahzadKaleemAwan, Mian Muhammad Awais (1999), Data Mining-Redefining the Boundaries, *IEEE Computer Society*

Incorporating Data Mining Models into Business Classes: Some methodological and Practical Considerations by Alexander Mechitov, University of Montevallo; Helen Moshkovich, University of Montevallo; David Olson, University of Nebraska published in Journal of Informatics Education Research Mechitov, Moshkovich, & Olson 2001.

Berry M. and G. Linoff (2002) Mastering Data Mining, New York: *John Wiley & Sons.*

Tan Pang-Ning, Steinbach M., Vipin Kumar (2012), Introduction to Data Mining, Pearson Education, *New International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, Vol.2, No.3.

Baazaoui Z., H., Faiz S., and Ben Ghezala, H. (2005), A Framework for Data Mining Based Multi-Agent: An Application to Spatial Data, volume 5, ISSN 1307-6884, *Proceedings of World Academy of Science, Engineering and Technology*,Communication and Information Sciences of the Faculty of Humanities,

Antonie M. L., Zaiane O. R.,Coman A. (2001), Application of Data Mining Techniques for Medical Image Classification, *Proceedings of the Second International Workshop on Multimedia Data Mining* MDM/KDD 2001) in conjunction with ACM SIGKDD conference, San Francisco.

Kusiak A., Kernstine K.H., Kern J.A., McLaughlin, K.A., and Tseng, T.L. (2000), Data Mining: Medical And Engineering Case Studies. *Proceedings of the Industrial Engineering Research Conference*, Cleveland, Ohio, pp. 1-7.

Luis R., Redol J., Simoes D., Horta N., Data Warehousing and Data Mining System Applied to E- Learning, Proceedings of the II International Conference on Multimedia and Information & Communication Technologies in Education, Badajoz, Spain, December 3-6th 2003.