Data Mining

Dipti Lodwal[1], Karan Singh Bamniya[2]

[1]Librarian, Govt. Nehru P G College Agar Malwa
[2]Guest Faculty (Librarian), Govt. College, Bidwal

Introduction

Generally, Data Mining (Sometimes called data or Knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information–information that can be used to increase revenue, cuts, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, Data Mining is the process of finding correlation or patterns among dozens of filed in large relation databases. Although Data Mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of super market scanner data and analyze market research report for years. Continues innovation in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis.

Data Mining has been definite as “The nontrivial extraction of implicit, previously unknown, and potentially useful information from data”. Data Mining uses machine learning, statistical and visualization techniques to discover and present knowledge in a form. This knowledge is easily comprehensible to humans. Data Mining can also be “the nontrivial extraction of implicit, previously unknown, and potentially useful information from data” and “The science of extracting useful information from large data sets or databases”. Although it is usually used in relation to analysis of data, data mining, like artificial intelligence, is an umbrella term and can be used in a wide range of contexts. It is usually associate with identifying trends. Data Mining involves the process of analyzing data to show patterns or relationship; sorting through large amounts of data; and picking out pieces of relative information or patterns that occur.

Data Mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potently to help organizations focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing business to make proactive, knowledge-driven decision. Data mining tools can answer business question that traditionally were too time consuming to resolve. Data mining tools scour identify hidden patterns and find predictive information.

Need of Data Mining

The goal of data mining in precise are prediction and description. Prediction involves using some variables or fields in the database of interest, and description focuses on finding human-interpretable patterns describing the data. The goal of prediction and description can be achieved using a verity of particular data-mining methods like.

- Classification

- Clustering
- Association rules
- Sequential patterns
- Regression
- Summarization
- Change and Deviation Detection

Most organization already collects and refine massive quantities of data. The important feature is that data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources and these techniques can be integrated with news products and systems as they are brought on-line. Data mining tools can analyze massive database to deliver answers to questions when implemented on high performance client/server or parallel processing computers.

Essentials of Data Mining

Data mining techniques are the result of a long process of research and product development. Data mining takes the evolutionary process beyond retrospective data access and navigation to prospective information delivery. Data mining is ready for application in the organization due to by three technologies that are given below:

- Massive data collection
- Powerful multiprocessor computers
- Data mining techniques algorithms

How does Data Mining work?

Data Mining provides link between separate transaction and analytical systems. Data Mining software analyzes relationships and patterns in stored transaction data which is based on open-ended user queries. Several types of analytical software are available such as statistical, machine learning, and neural networks. Generally, any of four types of relationship are sought

- **Classes:** Stored data is used to locate data in predermined groups.
- **Cluster :** data items are grouped according to logical relationships or consumer preferences.
- **Associations:** data can be mined to identify associations.
- **Sequential patterns:** data is mined to anticipate behavior patterns and trends.

- **Elements of Data Mining**

  Data mining consists of fore major elements:

  - Extract, transform, and load transaction data onto the data warehouse system.
  - Store and manage the data in a multidimensional database system.
  - Provide data access to business analysis and information technology professionals.
  - Analyze the data in a useful format, such as a graph or table.

- **Techniques in Data Mining**

The most commonly used techniques in data mining and different levels of analysis are as following:

I. **Artificial neural networks:** these are non-linear predictive models that learn through training and resemble biological neural networks in structure.

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

III. **Decision trees:** these are tree-shaped structures that represents sets of decisions. These decisions generate rules for the classification of a database. Specific decisions tree methods include Classification and Regression Tree (CART) and Chi Square Automatic Interaction Detection (CSAID). They provides a set of rules that user can apply to a new (unclassified) dataset to predict which recodes will have a given outcome.

IV. **Nearest neighbor method:** this technique classified each recorded in a dataset based on a combination of a classes of the k record(S) most similar to it in a historical dataset (where k 1). Sometimes it is called the k-nearest neighbor techniques.

V. **Rule induction:** the extraction of useful if-term rules from data based on statistical significance.

VI. **Data visualization:** the visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

- **Using Data Mining Tools**

Data mining software are analyzed into two groups: data mining tools and data mining applications. Data mining tools provide a number of techniques that can be applied to any business problems. Data mining applications, on the other hand, embed techniques inside an application customized to address a specific business problem. Both data mining tools and data mining applications are valuable. Organizations are using data mining application tools and data mining applications together in an integrated environment for predictive analytics. Data mining tools are used to ensure flexibility and the greatest accuracy possible. It is found that, data mining tools increase the effectiveness of data mining applications. Data mining tools deliver in-depth techniques as well as flexibility to use combinations of technique to improve predictive accuracy.

- **The scope of Data Mining**

Data mining derives its name from the similarities between searching for valuable business information in a large database – It requires either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Scope of data mining is as following:

- **Automated prediction of trends and behaviors:**

Data mining automates the process of finding predictive information in large databases. Extensive hands-on analysis can now be answered directly from the data-quickly. For example, in targeted marketing, it uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings.

- **Automated discovery of previously unknown patterns:**

Data mining tools sweep through databases and identify previously hidden pattern in one step. The pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

- **Importance:**

Data mining techniques can yield the benefits of automation on exciting software and hardware platforms, and these can be implemented on new system as existing platform are upgraded and new protect developed. Due to being implanted on high performance parallel processing systems, data mining tools can analyze massive database 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 database, in turn, yield improved predictions. It is further explained as:

1. **More columns:** analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Though variables that are discarded because they seem unimportant may carry information about unknown patterns.
2. **More rows:** the larger samples yield lower estimation errors and variance, and they also allow users to make inferences about small but important segments of a population.

It is found that data mining uncovers patterns play a critical role in decision making because they reveal areas for process improvement. Using data mining, organizations can increase the profitability of their interactions with customers, detect fraud, and improve risk management.

- **Architecture for Data Mining**

Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytical data warehousing can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Following figure illustrates architecture for advanced analyze data warehouse.

Sale & Marketing of data warehouse

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact couples with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database system: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allows the user to analyze the data as they want to view their business- summarizing by product line, region, and other key perspectives of their business. The data mining server must be integrated with the data warehouse and the OLAP server to embed ROI-focused (Region of Internet) Business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decision and apply them to future decisions.

This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users business models directly to the warehouse and returned a proactive analysis of the most relevant information. These results enhance metadata layer that represent a distilled view of the data. Reporting, visualization, and other analysis tools can than be applies to plan future actions and confirm the impact of those plants.

- **Data Mining and Classification**

Data mining creates classification models by examining already classified data (cases) and inductively finding a productive pattern. These existing cases may come from an historical database, such as people who have already undergone a particular medical treatment or moved to a new long distance service. They may come from an experiment in which a sample of the entire database is tasted in the real world and the results used to create a classier. Sometime an export classifies a sample of the database, and this classification is then used to create the model which will be applied to the entire database.

The use of classification algorithms begins whit a training set of per-classified example transactions. 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 require for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. The approach effects the explanation capability of the system. Once an affective classifier is developed, it is used in a predictive mode to classify new records into there same predefined classes.

- **Data Mining and Clustering**

Clustering approaches adder's regimentation problems. These approaches assign records with a large number of attributes into a relatively small set of groups or "segments" this assignment process is performed automatically by Clustering algorithms that identify the distinguishing characteristics of the dataset and then partition the new dimensional space defined by the dataset attributes along natural cleaving boundaries. This is no need to identify the groupings desired or the attributes that should be used to segment the dataset.

Clustering is often one of the first steps in data mining analysis. It identifies groups are related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analysis using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired out comes. For example, the buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign.

Clustering divides a database into different groups. The goal of clustering is to find groups that are very different from each other, and whose members are very similar to each other.

One should not confuse Clustering with segmentation. Segmentation refers to the general problem of identifying groups that have common characteristics. Clustering is a way to segment data into groups that are not previously defined, whereas classification is way to segment data by assigning it to groups that are already defined.

- Application areas
- Business intelligence
- Business performance management
- Discovery science
- Loyalty card
- Cheminformatics
- Quantitative structure-activity relationship
- Bioinformatics

- **Working of Data Mining**

How exactly is data mining able to tell one important thing that someone didn't know or what is going to happen next? 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 someone knows the answer and then applying it to another situation that some time don't Hopefully, if someone has got a good model, he finds his treasure. This act of model building is thus something that people have been doing for a long time, certainly before the advent of computers or data mining technology. What happens on computers, however, is not much different than the way people build models. Computers are loaded up with lots of information about a variety of situations where an answer is known and then the data mining software on the computer must run through that data and distill the characteristics of the data that should go into the model. Once the model is built it can then be used in similar situations where one doesn't know the answer.

- **Advances in Data Mining**

Recent advances have led to the newest and hottest trends in data mining-text mining and web mining. These two data mining technologies open a rich vein of customer data in the from of textual comments from survey research and log files from web servers, which were previously unusable. Applying data mining to these data adds a richness and depth to the patterns already uncovered through the data mining efforts.

- **Text Mining**

Text mining enables organization to explore the "unstructured" information contained in text in much the someway that data mining explores tabular or "structured" data. Through text mining, one can uncover hidden patterns, relationships, and trends in text. As a result, he gains greater insight from articles, reports, surveys, call center notes fields, e-mail, RSS (Rich Site Summary) feeds such as blogs and news feeds, and other types of text documents.

In addition, software enables to build on the mining efforts with products that help to incorporate text in predictive data mining models-a solution that is called Predictive Text Analytics provides a way for the organization to combine structured and unstructured information in the same models. This enables to draw more reliable conclusions and take more effective action.

By using Predictive Text Analytics, the organization can:

- Uncover concepts and relationship among concepts in text that could be too costly-or even impossible-to detect with other methods.
- Improved the "lift" or accuracy or predictive models and, by doing so make the models more effective.
- Deploy results both to the people who make decisions and to automated system that makes recommendations.

- **Web Mining**

Traditional web analytics are often to provide the actionable information decision makers need to successfully manage their online business, predictive web analytics solutions help companies close the online insight gaps by combining the leading technologies for predictive analytics and enterprise web analytics-along with software more than 35 year analytical experience, predictive web analytic solution deliver application modules for online insight through an enterprise architecture. These application modules adders critical online business goals like increasing conversion rates with predictive analysis capabilities that range from automated visitor segmentation to predictive search engine marketing intelligence. Predictive intelligence requires little interpretation and can be used to immediately improve online customer interpretation. For beyond "Number of Visits" by providing an actual list of the visitors most likely to convert. This online insight is both predictive in that the helps organizations to anticipates customer needs and prescriptive in delivering plain-English recommendations on how to meet them.

References

1. Frawley W, Piatetsky-Shapiro G. and Matheus C. Knowledge Discovery in Databases: An overview. Al Magazine, fall 1992,pp.213-228.
2. Hand D, Mannila H, and Smyth P. Principal of Data Mining. MIT Press, Cambridge, MA, 2001.
3. Fred Schwed, and Jr. Where are the customers' Yachts" ISBN0471119792 (1940).
4. Menzies, Y. and Hu, Y. Data mining for very busy people. IEEE Computer, Octuber2003, p.18-25.
5. http://www.spsss.com/data_mining/ Searched on 18/11/2022. 6. http://www.the-data-mine.com Searched on 18/11/2022.
6. www.kdnuggets.com Searched on 20/11/2022.
7. Caudill, M & Butler, C. (1990) Naturally Intelligence Systems. Cambridge: MIT Press.

8. Chelton, M.K. (2003). Readers advisory 101: crash course in RA: Common mistakes librarian make and how to avoid them, Library Journal, 11/1/2003 http://www.libraryjournal.com/articles.CA329318.html. Searched on 21/11/2022.
9. Lloyd- Williams, M. & S. (1996) "A Neural Network Approach to Analyzing HealthCare Information", Topics in Health Information Management, 17(2), 26-33.
10. .http://databases.about.com/od/datamining/g/datamining.htm Searched on 21/11/2022.
11. http://databases.about.com/od/datamining/g/clustering.htm Searched on 22/11/2011.