DATA MINING TECHNIQUES AND APPLICATIONS

Nitesh Kumar Jangir
MTECH (CSE), Shekhawati Institute of Engineering and Technology, Sikar

***Abstract:* There is a huge amount of data available in the Information Industry. This data is of no use until it is transformed into useful information. It is compulsory to analyze this huge amount of data and extract useful information from it.**

The amount of data being generated and stored is increasing exponentially, due in large part to the continuing advances in computer technology. This presents tremendous opportunities for those who can unseal the information embedded within this data, but also introduces new challenges.

Data mining also requires other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are above, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, and Science Exploration.

***Indexed Terms:* Data Mining Techniques, data mining algorithms, data mining applications.**

I. INTRODUCTION TO DATA MINING

The development of Information Technology has produced large amount of databases and huge data in various areas. The research in databases and information technology has given increase to an approach to store and manipulate this precious data for further decision making.

Data mining is a process of extraction of useful information and patterns from large data. It is also known as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

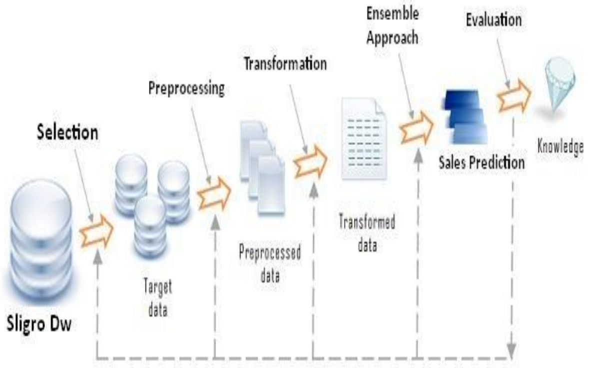
The relationship between "patterns in data" and "knowledge" will be discussed presently. Although not published explicitly in this definition, it is understood that the process must be at least partially automated, relying heavily on specialized computer algorithms that search for patterns in the data.

It is important to point out that there is some equivocation about the term "data mining", which is in large part purposeful. This term originally mentioned to the algorithmic step in the data mining process, which initially was known as the Knowledge Discovery in Databases (KDD) process.

Data Mining is defined as extracting information from large amount of data. In other words, we can say that data mining is the concept of mining knowledge from data.

The information or knowledge extracted so can be used for different types of applications −

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration



II. DATA MINING ALGORITHMS & TECHNIQUES

Different types of 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.

2.1 Classification

Classification is the most usually 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. This concept frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification.

In Learning the training data are examined by classification algorithm. In classification test data are used to approximate the accuracy of the classification rules. If the accuracy is acceptable then the rules can be applied to the new data tuples.

The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then convert 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

Clustering can be used in identification of similar classes of objects. By using clustering techniques we can stimulate 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 differentiating 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 same functionality

Requirements of Clustering in Data Mining
The following points specifies light on why clustering is required in data mining −

- Scalability − we require highly scalable clustering algorithms to deal with large databases.
- Ability to deal with different kinds of attributes − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- Discovery of clusters with attribute shape − the clustering algorithm should be efficient of detecting clusters of arbitrary shape.
- High dimensionality − the clustering algorithm should not only be capable to handle low-dimensional data but also the high dimensional space.
- Ability to deal with noisy data − Databases involves noisy, missing or erroneous data. Some algorithms are sensitive to such data and may cause to poor quality clusters.
- Interpretability − the clustering results should be interpretable, comprehensible, and acceptable.

Clustering Methods

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

2.3. Predication
Regression technique can be adapted for predication. Regression analysis can be used to design the relationship between one or more independent variables and dependent variables. In data mining independent variables attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction.

The same model types can frequently be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to construct both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks also can create both classification and regression models.

Types of regression methods
Linear Regression
Multivariate Linear Regression
Nonlinear Regression
Multivariate Nonlinear Regression

2.4. Association rule
Association and correlation is usually to find frequent item set findings among large amount of data sets. This type of discovering helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis.

Types of association rule

- Multilevel association rule
- Multidimensional association
- Quantitative association rule

2.5. Neural networks
Neural network is the collection of connected input/output units and each connection has a weight present with it. During the training 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 capability 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. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully concerned in many industries. Neural networks are prime at identifying patterns or trends in data.

III. DATA MINING APPLICATIONS

Data mining is a comparatively new technology that has not fully matured. Besides this, there are a number of industries that are already using it on a regular basis. Some of these organizations encompass retail stores, hospitals, banks, and insurance companies. Most of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools.

Data mining can be used to search patterns and connections that would otherwise be difficult to find. This technology is favor with many businesses because it allows them to learn more about their customers and make smart marketing decisions. Here is overview of business problems and solutions found by using data mining technology.

- ➢ Financial Data Analysis
- ➢ Retail Industry
- ➢ Telecommunication Industry
- ➢ Biological Data Analysis
- ➢ Other Scientific Applications
- ➢ Intrusion Detection

Financial Data Analysis: The financial data in the banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

- Design and creation of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit scheme analysis.
- Taxonomy and clustering of customers for targeted marketing.
- Diagnosis of money laundering and other financial crimes.

Retail Industry: Data mining in retail industry supports in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of applications of data mining in the retail industry −

- Design and Construction of data warehouses related to the advantages of data mining.
- Multidimensional survey of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.

Telecommunication Industry: Data mining in telecommunication industry supports in identifying the telecommunication patterns, make better use of resource, and improve quality of service. Here is the list of instances for which data mining improves telecommunication services −

- Multidimensional Analysis of data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Utilize of visualization tools in telecommunication data analysis.

Biological Data Analysis: In present times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining plays a very important role in Bioinformatics. Following are the features in which data mining contributes for biological data analysis −

- Semantic incorporation of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis different types of nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein routes.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Other Scientific Applications: The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Large amount of data have been collected from scientific domains such as geosciences, astronomy, etc.

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

Intrusion Detection: Intrusion introduce to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In the world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network provoked intrusion detection to become a critical component of network administration.

- Evolution of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and design discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.

IV. CONCLUSION

Data mining has importance about finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in discovering the patterns to decide upon the future trends in businesses to grow.

Data mining has vast 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 encouraging interdisciplinary developments in Information Technology.

REFERENCES

[1] "Introduction to data mining" by Tan, Steinbach & Kumar (2006)
[2] Data Mining: Concepts and Techniques, Third Edition by Han, Kamber & Pei (2013)
[3] Data Mining and Analysis Fundamental Concepts and Algorithms by Zaki & Meira (2014)
[4] Data Mining: The Textbook by Aggarwal (2015)
[5] "The Elements of Statistical Learning" by Freidman et al (2009)