UNIT I – DATA MINING BASICS

Introduction

There is huge amount of data available in Information Industry. This data is of no use until converted into useful information. Analysing this huge amount of data and extracting useful information from it is necessary.

The extraction of information is not the only process, need to perform, it also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, now it is ready to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration etc.

Data Mining

Data Mining is defined as extracting information from huge sets of data. In other words, data mining is the procedure of mining knowledge from data.

Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called ***Knowledge Discovery in Database (KDD)***. The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information.

Types of Data can be Mined

Data mining can be performed on the following types of data:

Relational Database:

A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data search ability, reporting, and organization.

Data warehouses:

A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for

analytical purposes and helps in decision- making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing.

Data Repositories:

The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.

Object-Relational Database:

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc. One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.

Transactional Database:

A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

Advantages of Data Mining

- The Data Mining technique enables organizations to obtain knowledge-based data.
- Compared with other statistical data applications, data mining is a cost-efficient.
- Data Mining helps the decision-making process of an organization.
- It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

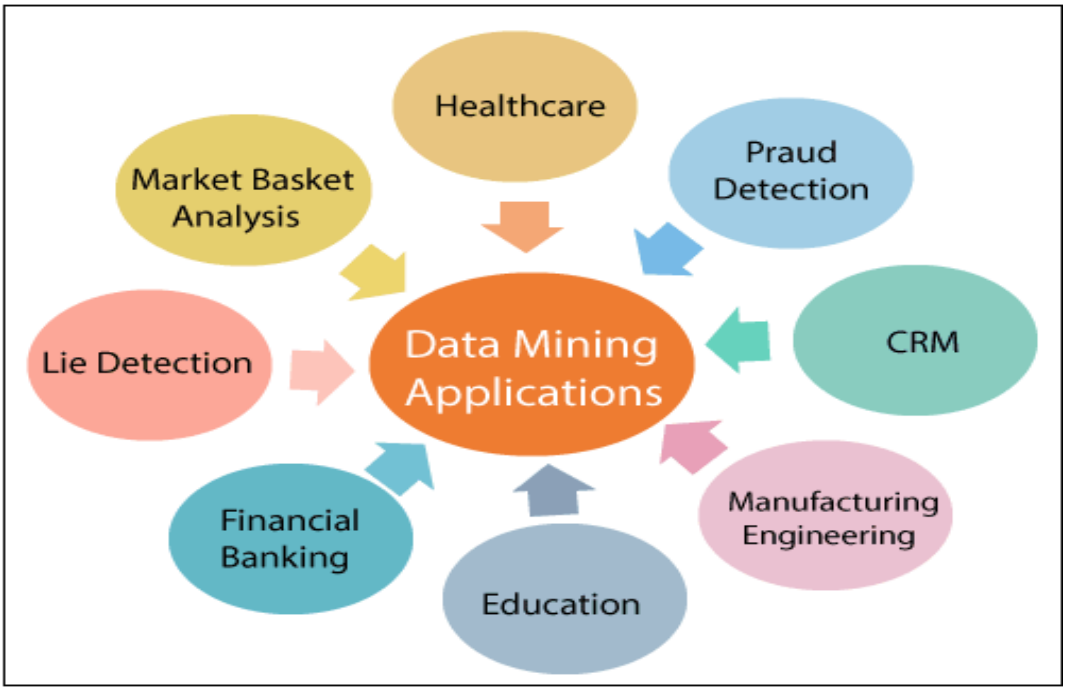
Disadvantages of Data Mining

- There is a probability that the organizations may sell useful data of customers to other organizations for money.
- Many data mining analytics software is difficult to operate and needs advance training to work on.

- Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
- The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

Data Mining Applications

Data Mining is primarily used by organizations with intense consumer demands- Retail, Communication, Financial, marketing company, determine price, consumer preferences, product positioning, and impact on sales, customer satisfaction, and corporate profits. Data mining enables a retailer to use point-of-sale records of customer purchases to develop products and promotions that help the organization to attract the customer.



These are the following areas where data mining is widely used:

Data Mining in Healthcare:

Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs. Data Mining can be used to forecast patients in each category. The procedures ensure that the patients get intensive care at the right place and at the right time. Data mining also enables healthcare insurers to recognize fraud and abuse.

Data Mining in Market Basket Analysis:

Market basket analysis is a modeling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products. This technique may enable the retailer to understand the purchase behavior of a buyer. This data may assist the retailer in understanding the requirements of the buyer and altering the store's layout accordingly.

Data mining in Education:

Education data mining is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from educational Environments. EDM objectives are recognized as affirming student's future learning behavior, studying the impact of educational support, and promoting learning science. An organization can use data mining to make precise decisions and also to predict the results of the student. With the results, the institution can concentrate on what to teach and how to teach.

Data Mining in Manufacturing Engineering:

Knowledge is the best asset possessed by a manufacturing company. Data mining tools can be beneficial to find patterns in a complex manufacturing process. Data mining can be used in system-level designing to obtain the relationships between product architecture, product portfolio, and data needs of the customers. It can also be used to forecast the product development period, cost, and expectations among the other tasks.

Data Mining in CRM (Customer Relationship Management):

Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies. To get a decent relationship with the customer, a business organization needs to collect data and analyze the data. With data mining technologies, the collected data can be used for analytics.

Data Mining in Fraud detection:

Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated. An ideal fraud detection system should protect the data of all the users. Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent. A model is constructed using this data, and the technique is made to identify whether the document is fraudulent or not.

Data Mining in Lie Detection:

Apprehending a criminal is not a big deal, but bringing out the truth from him is a very challenging task. Law enforcement may use data mining techniques to investigate offenses, monitor suspected terrorist communications, etc. This technique includes text mining also, and it seeks meaningful patterns in data, which is usually unstructured text. The information collected from the previous investigations is compared, and a model for lie detection is constructed.

Data Mining Financial Banking:

The Digitalization of the banking system is supposed to generate an enormous amount of data with every new transaction. The data mining technique can help bankers by solving business-related problems in banking and finance by identifying trends, casualties, and correlations in business information and market costs that are not instantly evident to managers or executives because the data volume is too large or are produced too rapidly on the screen by experts. The manager may find these data for better targeting, acquiring, retaining, segmenting, and maintain a profitable customer.

Data Mining vs Query Tools

Query Tools are tools that help analyze the data in a database. They provide query building, query editing, searching, finding, reporting and summarizing functionalities. On the other hand, Data mining is a field in computer science, which deals with the extraction of previously unknown and interesting information from raw data. Data used as the input for the Data mining process usually is stored in databases. Data miners are interested in finding useful relationships between different data elements, which is ultimately profitable for businesses.

Data mining

Data mining is also known as Knowledge Discovery in Data (KDD). As mentioned above, it is a field of computer science, which deals with the extraction of previously unknown and interesting information from raw data. For example, it is currently been used for various applications such as social network analysis, fraud detection and marketing. Data mining usually deals with following four tasks: clustering, classification, regression, and association. Clustering is identifying similar groups from unstructured data. Classification is learning rules that can be applied to new data and will typically include following steps: preprocessing of data, designing modeling, learning/feature selection and Evaluation/validation. Regression is finding functions with minimal error to model data. And association is looking for relationships between variables. Data mining is usually used to answer questions like what are the main products that might help to obtain high profit next year in Wal-Mart?

Query Tools

Query Tools are tools that help to analyze the data in a database. Usually these query tools have a GUI front end with convenient ways to input queries as a set of attributes. Once these inputs are provided the tool generates actual queries made up of the underlying query language used by the database. SQL, T-SQL and PL/SQL are examples of query languages used in many popular databases today. Then, these generated queries are executed against the databases and the results of the queries are presented or reported to the user in an organized and clear manner. Typically, the user does not need to know a database-specific query language to use a Query tool. Key features of Query tools are integrated query builder and editor, summery reports and figures, import and export features and advanced find/search capabilities.

What is the difference between Data mining and Query Tools?

Query tools can be used to easily build and input queries to databases. Query tools make it very easy to build queries without even having to learn a database-specific query language. On the other hand, Data Mining is a technique or a concept in computer science, which deals with extracting useful and previously unknown information from raw data. Most of the times, these raw data are stored in very large databases. Therefore Data miners can use the existing functionalities of Query Tools to preprocess raw data before the Data mining process. However, the main difference between Data mining techniques and using Query tools is that, in order to use Query tools the users need to know exactly what they are looking for, while data mining is used mostly when the user has a vague idea about what they are looking for.

Machine learning

Machine learning is related to the development and designing of a machine that can learn itself from a specified set of data to obtain a desirable result without it being explicitly coded. Hence Machine learning implies 'a machine which learns on its own.

Arthur Samuel invented the term Machine learning an American pioneer in the area of *computer gaming* and *artificial intelligence* in **1959**. He said that "**it gives computers the ability to learn without being explicitly programmed**."

Machine learning is a technique that creates complex algorithms for large data processing and provides outcomes to its users. It utilizes complex programs that can learn through experience and make predictions. The algorithms are enhanced by themselves by frequent input of training data. The aim of machine learning is to understand information and build models from data that can be understood and used by humans.

Machine learning algorithms are divided into:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi supervised Learning

Supervised Machine Learning:

Supervised learning is commonly used in real world applications, such as face and speech recognition, products or movie recommendations, and sales forecasting. Supervised learning can be further classified into two types - **Regression** and **Classification**.

Regression trains on and predicts a continuous-valued response, for example predicting real estate prices.

Classification attempts to find the appropriate class label, such as analyzing positive/negative sentiment, male and female persons, benign and malignant tumors, secure and unsecure loans etc.

In supervised learning, learning data comes with description, labels, targets or desired outputs and the objective is to find a general rule that maps inputs to outputs. This kind of learning data is called **labeled data**. The learned rule is then used to label new data with unknown outputs.

Unsupervised Machine Learning:

Unsupervised learning does not depend on trained data sets to predict the results, but it utilizes direct techniques such as clustering and association in order to predict the results.

Unsupervised learning is used to detect anomalies, outliers, such as fraud or defective equipment, or to group customers with similar behaviors for a sales campaign. It is the opposite of supervised learning. There is no labeled data here.

When learning data contains only some indications without any description or labels, it is up to the algorithm to find the structure of the underlying data, to discover hidden patterns, or to determine how to describe the data. This kind of learning data is called **unlabeled data**.

Reinforcement learning:

Here learning data gives feedback so that the system adjusts to dynamic conditions in order to achieve a certain objective. The system evaluates its performance based on the

feedback responses and reacts accordingly. The best known instances include self-driving cars and chess master algorithm AlphaGo.

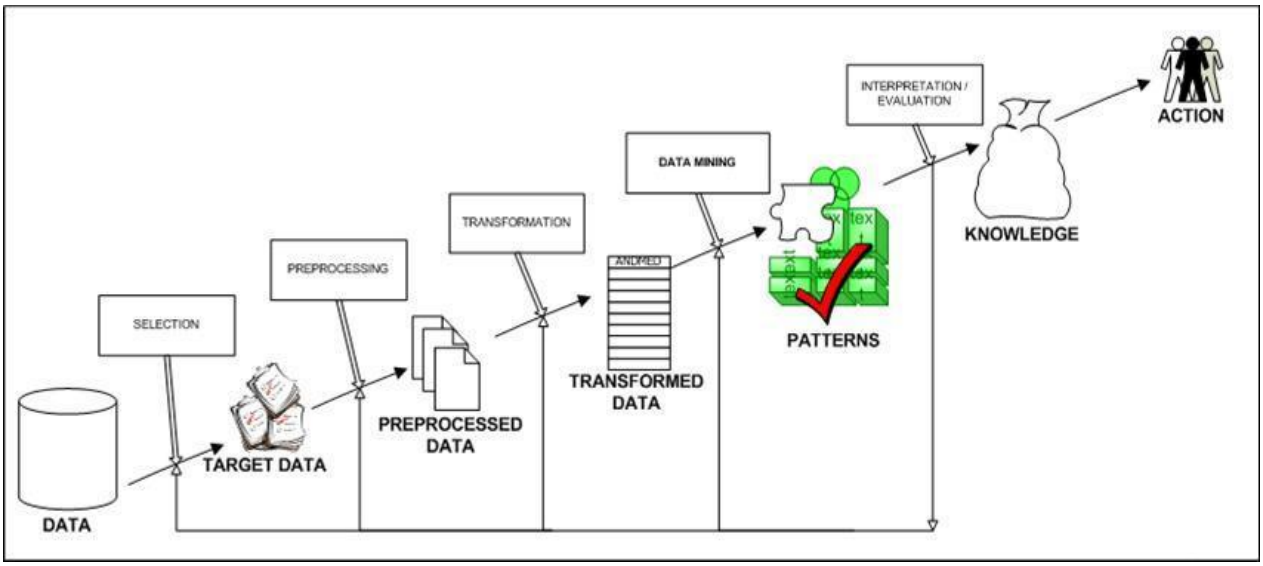
Semi-supervised learning:

If some learning samples are labeled, but some other are not labeled, then it is semi-supervised learning. It makes use of a large amount of **unlabeled data for training** and a small amount of **labeled data for testing**. Semi-supervised learning is applied in cases where it is expensive to acquire a fully labeled dataset while more practical to label a small subset.

Data Mining Vs Machine Learning

Factors	Data Mining	Machine Learning
Origin	Traditional databases with unstructured data.	It has an existing algorithm and data.
Meaning	Extracting information from a huge amount of data.	Introduce new Information from data as well as previous experience.
History	In 1930, it was known as knowledge discovery in databases(KDD).	The first program, i.e., Samuel's checker playing program, was established in 1950.
Responsibility	Data Mining is used to obtain the rules from the existing data.	Machine learning teaches the computer, how to learn and comprehend the rules.
Applications	In compare to machine learning, data mining can produce outcomes on the lesser volume of data. It is also used in cluster analysis.	It needs a large amount of data to obtain accurate results. It has various applications, used in web search, spam filter, credit scoring, computer design, etc.
Nature	It involves human interference more towards the manual.	It is automated, once designed and implemented, there is no need for human effort.
Scope	Applied in the limited fields.	It can be used in a vast area.

Steps in Data Mining

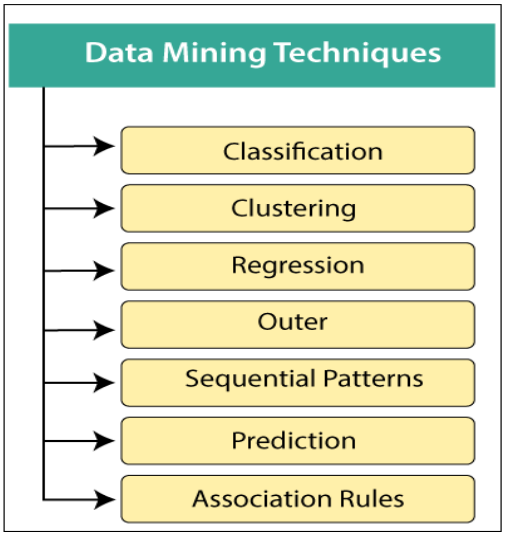


There are various steps that are involved in mining data as shown in the picture.

1. **Data Integration:** First of all the data are collected and integrated from all the different sources.
2. **Data Selection:** We may not all the data we have collected in the first step. So in this step we select only those data which we think useful for data mining.
3. **Data Cleaning:** The data we have collected are not clean and may contain errors, missing values, noisy or inconsistent data. So we need to apply different techniques to get rid of such anomalies.
4. **Data Transformation:** The data even after cleaning are not ready for mining as we need to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.
5. **Data Mining:** Now we are ready to apply data mining techniques on the data to discover the interesting patterns. Techniques like clustering and association analysis are among the many different techniques used for data mining.
6. **Pattern Evaluation and Knowledge Presentation:** This step involves visualization, transformation, removing redundant patterns etc from the patterns we generated.
7. **Decisions / Use of Discovered Knowledge:** This step helps user to make use of the knowledge acquired to take better decisions.

Data Mining Techniques

In recent data mining projects, various major data mining techniques have been developed and used, including association, classification, clustering, prediction, sequential patterns, and regression.



1. **Classification:**

This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes. Data mining techniques can be classified by different criteria, as follows:

i. **Classification of Data mining frameworks as per the type of data sources mined:** This classification is as per the type of data handled. For example, multimedia, spatial data, text data, time-series data, World Wide Web, and so on..

ii. **Classification of data mining frameworks as per the database involved:** This classification based on the data model involved. For example. Object-oriented database, transactional database, relational database, and so on..

iii. **Classification of data mining frameworks as per the kind of knowledge discovered:**
This classification depends on the types of knowledge discovered or data mining functionalities. For example, discrimination, classification, clustering, characterization, etc. some frameworks tend to be extensive frameworks offering a few data mining functionalities together..

iv. **Classification of data mining frameworks according to data mining techniques used:**
This classification is as per the data analysis approach utilized, such as neural networks, machine learning, genetic algorithms, visualization, statistics, data warehouse-oriented or database-oriented, etc.

2. Clustering:

Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data. It models data by its clusters. Data modeling puts clustering from a historical point of view rooted in statistics, mathematics, and numerical analysis. Clustering is very similar to the classification, but it involves grouping chunks of data together based on their similarities.

3. Regression:

Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables. It is used to define the probability of the specific variable. Regression, primarily a form of planning and modeling

4. Association Rules:

This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set. Association rules are if-then statements that support to show the probability of interactions between data items within large data sets in different types of databases. Association rule mining has several applications and is commonly used to help sales correlations in data or medical data sets.

The way the algorithm works is that you have various data, For example, a list of grocery items that you have been buying for the last six months. It calculates a percentage of items being purchased together.

These are three major measurements technique:

- **Lift:**
  This measurement technique measures the accuracy of the confidence over how often item B is purchased.
  (Confidence) / (item B)/ (Entire dataset)
- **Support:**
  This measurement technique measures how often multiple items are purchased and compared it to the overall dataset.
  (Item A + Item B) / (Entire dataset)
- **Confidence:**
  This measurement technique measures how often item B is purchased when item A is purchased as well.
  (Item A + Item B)/ (Item A)

5. Outer detection:

This type of data mining technique refers to observation of data items in the dataset which do not match an expected pattern or expected behavior. This technique can be used in

a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

The majority of the real-world datasets have an outlier. Outlier detection plays a significant role in the data mining field. Outlier detection is valuable in numerous fields like network interruption identification, credit or debit card fraud detection, detecting outlying in wireless sensor network data, etc.

6. Sequential Patterns:

This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period. The sequential pattern is a data mining technique specialized for **evaluating sequential data** to discover sequential patterns. It comprises of finding interesting subsequences in a set of sequences, where the stake of a sequence can be measured in terms of different criteria like length, occurrence frequency, etc.

7. Prediction:

Prediction has used a combination of the other data mining techniques like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.

What is Data Mining?

Data mining is considered the process of extracting useful information from a vast amount of data. It's used to discover new, accurate, and useful patterns in the data, looking for meaning and relevant information for the organization or individual who needs it. It's a tool used by humans.

What is Machine Learning?

On the other hand, machine learning is the process of discovering algorithms that have improved courtesy of experience derived from data. It's the design, study, and development of algorithms that permit machines to learn without human intervention. It's a tool to make machines smarter, eliminating the human element (but not eliminating humans themselves; that would be wrong).

References

1. Jiawei Han, MichelineKamber, "Data Mining: Concepts and Techniques", Morgan Kaufman Publishers.
2. URL: https://www.tutorialspoint.com/data_mining/
3. URL: https://www.javatpoint.com/data-mining
