UNIT-2
DATA MINING
&
ITS APPLICATIONS

Prepared by S NOORTAJ

DATA MINING

Data mining is often defined as finding hidden information in a database. Alternatively, it has been called exploratory data analysis, data driven discovery, and deductive learning.

I. Traditional database queries, access a database using a well-defined querystated in a language such as SQL.
II. Data mining access of a database differs from this traditional access in several ways:

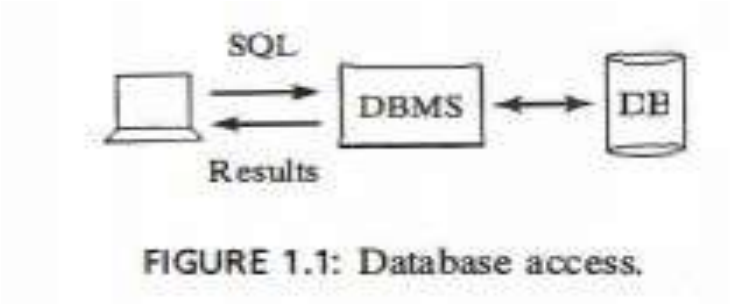


FIGURE 1.1: Database access.

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 needed 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: Over-fitting occurs when the model does not fit future states. **Multimedia data:** Most previous data mining algorithms are targeted to traditional data types (numeric, character, text, etc.). The use of multimedia data such as is found in GIS databases complicates or invalidates manyproposed algorithms.

Missing data: During the preprocessing phase of KDD, missing data may be replaced with estimates. This and other approaches to handling missing datacan lead to invalid results in the data mining step.

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

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

Changing data: Databases cannot be assumed to be static. However, most data mining algorithms do assume a static database.

Outliers: There are often many data entries that do not fit nicely into the derived model.

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.

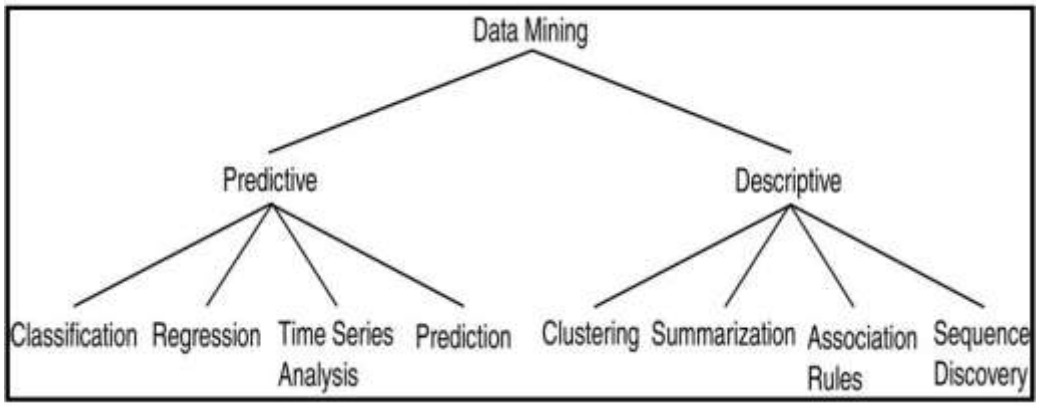
Integration: The KDD process is not currently integrated into normal data processing activities. KDD requests may be treated as special, unusual, or one- time needs. This makes them inefficient, ineffective, and not general enough to be used on an ongoing basis. Integration of data mining functions into traditional DBMS systems is certainly a desirable goal.

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.

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.

BASIC DATA MINING TASKS

the basic outline of tasks shown in below:



I. A predictive model makes a prediction about values of data using known results found from different data. Predictive model data mining tasks include classification, regression, time series analysis, and prediction.

II. A descriptive model identifies patterns or relationships in data. Clustering, summarization, association rules, and sequence discovery are usually viewed as descriptive in nature.

Classification :

Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. Two examples of classification applications are determining whether to make a bank loan and identifying credit risks.

Ex: An airport security screening station is used to determine: if passengers are potential terrorists or criminals. To do this, the face of each passenger is scanned and its basic pattern (distance between eyes, size and shape of mouth, shape of head, etc.) is identified. This pattern is compared to entries in a database to see if it matches any patterns that are associated with known offenders.

Regression :

Regression is used to map a data item to a real valued prediction variable. In actuality, regression involves the learning of the function that does this mapping. Regression assumes that the target data fit into some known type of function (e.g., linear, logistic,etc.)

Ex: Forecasting the pension, vehicles and accidents, increasing of birth rate and decreasing of food , father age and son age in future

Time Series Analysis:

With time series analysis, the value of an attribute is examined as it varies overtime. The values usually are obtained as evenly spaced time points (daily, weekly, hourly, etc.). A time series plot which is shown in below, is used to visualize the time series.

Ex 1:

Mr. Smith is trying to determine whether to purchase stock from Companies X, Y, or z. For a period of one month he charts the daily stock price for each company. Figure 1.3 shows the time series plot that Mr. Smith has generated. Using this and similar information available from his stockbroker, Mr. Slllllth decides to purchase stock X because it is less volatile while overall showing a slightly larger relative amount of growth than either of the other stocks. As a matter of fact, the stocks . for Y and Z have a similar behavior. The behavior of Y between days 6 and 20 is Identical to that for Z between days 13 and 27.

Ex 2: preparing weather report, changes in share market value ,health care monitoring

Prediction

Many real-world data mining applications can be seen as predicting future data states based on past and current data. Prediction can be viewed as a typeof classification. The difference is that prediction is predicting a future state rather than a current state Prediction applications include flooding, speech recognition, machine learning, and pattern recognition. Although future values may be predicted using time series analysis or regression techniques, other approaches may be used as well.

Ex: health prediction of a heart patient, cancer patient

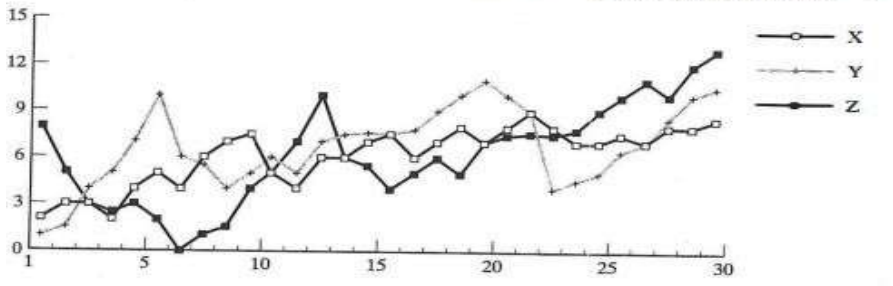
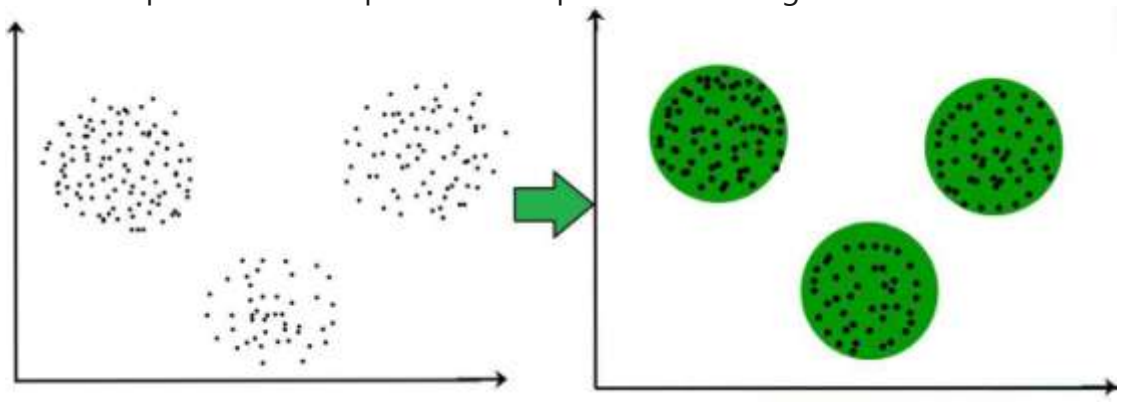


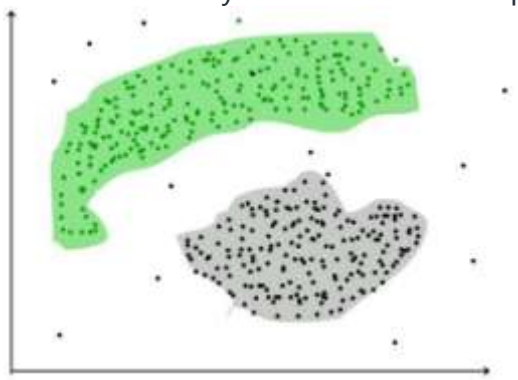
FIGURE 1.3: Time series plots.

Clustering

Clustering is similar to classification except that the groups are not predefined, but rather defined by the data alone. Clustering is alternatively referred to as unsupervised learning or segmentation. It can be thought of as partitioning or segmenting the data into groups that might or might not be disjointed. The clustering is usually accomplished by determining the similarity among the data on predefined attributes. The most similar data are grouped into clusters. Since the clusters are not predefined, a domain expert is often required to interpret the meaning of the created clusters.



It is not necessary for clusters to be spherical. Such as :



Summarization

Summarization maps data into subsets with associated simple descriptions. Summarization is also called characterization or generalization. It extracts or derivesrepresentative information about the database. This may be accomplished by actually retrieving portions of the data. Alternatively, summary type information (such as the mean of some numeric attribute) can be derived from the data. The summarization succinctly characterizes thecontents of the database.

Association Rules

An association rule is a model that identifies specific types of data associations. These associations are often used in the retail sales community to identify items that are frequently purchased together. Link analysis, alternatively referred to as affinity analysis or association, refers to the data mining task of uncovering relationships among data. The best example of this type of application is to determine association rules.

Here the data analyzed consist of information about what items a customer purchases. Associations are also used in many other applications such as predicting the failure of telecommunication switches.

Ex: Market basket analysis

Sequence Discovery

Sequential analysis or sequence discovery is used to determine sequential patterns in data. These patterns are based on a time sequence of actions. These patterns are similar to associations in that data (or events) are found to be related, but the relationship is based on time. Unlike a market basket analysis, which requires the items to be purchased at the same time, in sequence discovery the items are purchased over time in some order. sA similar type of discovery can be seen in the sequence within which data are purchased.

Ex: For example, most people who purchase CD players may be found to purchase CDs within one week. As we will see, temporal association rulesreally fall into this category.

TECHNOLOGIES USED IN DATA MINING

Data mining has incorporated many techniques from other domain fields like machine learning, statistics, information retrieval, data warehouse and data base, decision support system. Since it is a highly application-driven domain, the interdisciplinary nature is typically very significant. Research and development in data mining and its applications prove quite useful in implementing it. We will see major technologies utilized in data mining.

Several techniques used in the development of data mining methods. Some of them are mentioned below:

1. Statistics:
   - It uses the mathematical analysis to express representations, model and summarize empirical data or real world observations.
   - Statistical analysis involves the collection of methods, applicable to large amount of data to conclude and report the trend.
2. Machine learning
   - Arthur Samuel defined machine learning as a field of study that gives computers the ability to learn without being programmed.
   - When the new data is entered in the computer, algorithms help the data to grow or change due to machine learning.
   - In machine learning, an algorithm is constructed to predict the data from the available database (Predictive analysis).
   - It is related to computational statistics.

The Four Types of Machine Learning Are

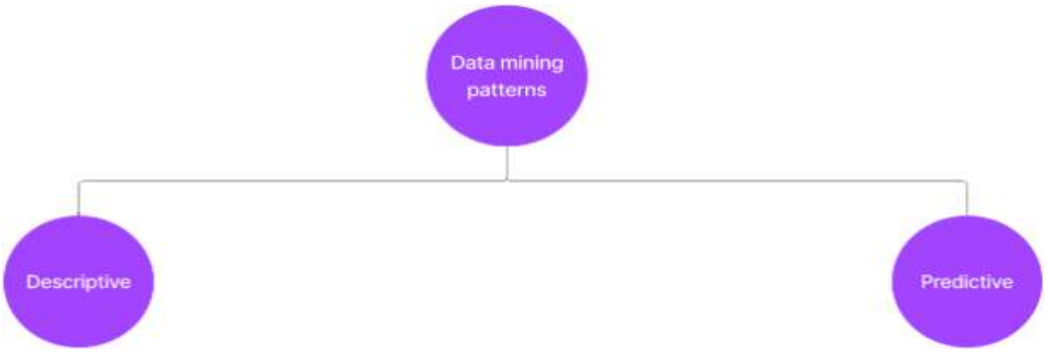
1. Supervised learning
   - It is based on the classification.
   - It is also called as inductive learning. In this method, the desired outputs are included in the training dataset.

2. Unsupervised learning
   - ❖ Unsupervised learning is based on clustering.
   - ❖ Clusters are formed on the basis of similarity measures and desired outputs are not included in the training dataset.
3. Semi-supervised learning
   - ❖ Semi-supervised learning includes some desired outputs to the training dataset to generate the appropriate functions. This method generally avoids the large number of labeled examples (i.e. desired outputs).
4. Active learning
   - ❖ Active learning is a powerful approach in analyzing the data efficiently.
   - ❖ The algorithm is designed in such a way that, the desired output should be decided by the algorithm itself (the user plays important role in this type).

3. Information retrieval
   - ➢ Information deals with uncertain representations of the semantics of objects (text, images).
   - ➢ For example: Finding relevant information from a large document. of objects text, images).
4. Database systems and data warehouse
   - ➢ Databases are used for the purpose of recording the data as well as data warehousing.
   - ➢ Online Transactional Processing (OLTP) uses databases for day to day transaction purpose.
   - ➢ To remove the redundant data and save the storage space, data is normalized and stored in the form of tables.
   - ➢ Entity-Relational modeling techniques are used for relational database management system design.
   - ➢ Data warehouses are used to store historical data which helps to take strategical decision for business.
   - ➢ It is used for online analytical processing (OALP), which helps to analyze the data.
5. Decision Support System
   - ➢ Decision support system is a category of information system. It is very useful in decision making for organizations.
   - ➢ It is an interactive software based system which helps decision makers to extract useful information from the data, documents to make the decision.

PATTERNS CAN BE MINED IN DATA MINING?

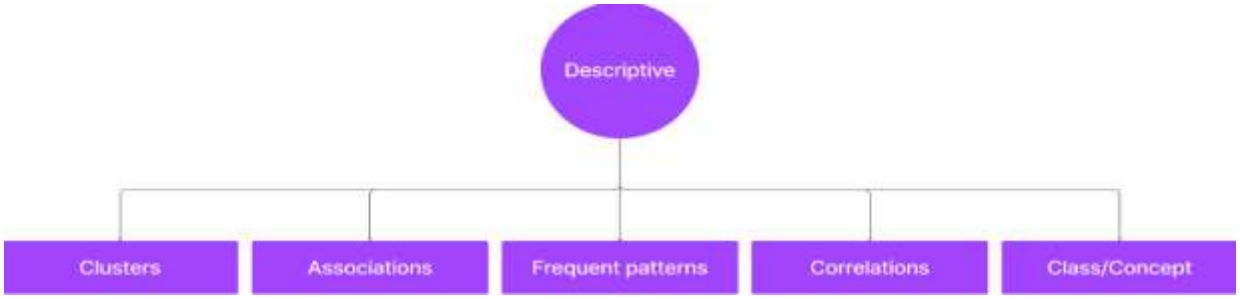
Based on the data functionalities, patterns can be further classified into two categories.



Descriptive patterns

It deals with the general characteristics and converts them into relevant and helpful information.

Descriptive patterns can be divided into the following patterns:



- **Class/concept description**: Data entries are associated with labels or classes. For instance, in a library, the classes of items for borrowed items include books and research journals, and customers' concepts include registered members and not registered members. These types of descriptions are class or concept descriptions.

- **Frequent patterns**: These are data points that occur more often in the dataset. There are many kinds of recurring patterns, such as frequent items, frequent subsequence, and frequent sub-structure.

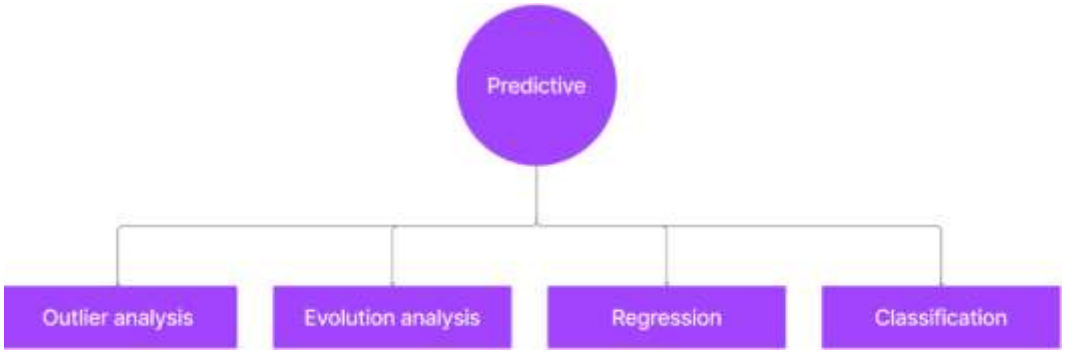
- **Associations**: It shows the relationships between data and pre-defined association rules. For instance, a shopkeeper makes an association rule that 70% of the time, when a football is sold, a kit is bought alongside. These two items can be combined together to make an association.

- **Correlations**: This is performed to find the statistical correlations between two data points to find if they have positive, negative, or no effect.

- **Clusters**: This is the formation of a group of similar data points. Each point in the collection is somewhat similar but very different from other members of different groups.

Predictive patterns

It predicts future values by analyzing the data patterns and their outcomes based on the previous data. It also helps us find missing values in the data.Predictive patterns can be categorized into the following patterns.



- **Classification**: It helps predict the label of unknown data points with the help of known data points**.** For instance, if we have a dataset of X-rays of cancer patients, then the possible labels would be **cancer patient** and **not cancer patient**. These classes can be obtained by data characterizations or by data discrimination.
- **Regression**: Unlike classification, regression is used to find the missing numeric values from the dataset. It is also used to predict future numeric values as well. For instance, we can find the behavior of the next year's sales based on the past twenty years' sales by finding the relation betweenthe data.
- **Outlier analysis**: Not all data points in the dataset need to follow the same behavior. Data points that don't follow the usual behavior are called outliers. Analysis of these outliers is called outlier analysis. These outliers are not considered while working on the data.
- **Evolution analysis**: As the name suggests, those data points change their behavior and trends with time.

DATA MINING - APPLICATIONS & TRENDS

Data mining is widely used in diverse areas. There are a number of commercial data mining system available today and yet there are many challenges in this field. In this tutorial, we will discuss the applications and the trend of data mining

Data Mining Applications

Here is the list of areas where data mining is widely used −

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

Financial Data Analysis

The financial data in 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 construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web. Data mining in retail industry helps 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 examples of data mining in the retail industry −

- Design and Construction of data warehouses based on the benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason

why data mining is become very important to help and understand the business.

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services −

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.

- Use of visualization tools in telecommunication data analysis.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following arethe aspects in which data mining contributes for biological data analysis −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiplenucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- 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. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications −

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

Intrusion Detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this 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 prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection −

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

TRENDS IN DATA MINING

Data mining concepts are still evolving and here are the latest trends that weget to see in this field −

- Application Exploration.
- Scalable and interactive data mining methods.
- Integration of data mining with database systems, data warehouse systems and

web database systems.

- Standardization of data mining query language.
- Visual data mining.
- New methods for mining complex types of data.
- Biological data mining.
- Data mining and software engineering.
- Web mining.
- Distributed data mining.
- Real time data mining.
- Multi database data mining.
- Privacy protection and information security in data mining.

KDD PROCESS

KDD (Knowledge Discovery in Databases) is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets. The KDD process is an iterative process and it requires multiple iterations of the above steps to extract accurate knowledge from the data. The following steps are included in KDD process:

Data Cleaning

Data cleaning is defined as removal of noisy and irrelevant data from collection.

- ✓ Cleaning in case of Missing values.
- ✓ Cleaning noisy data, where noise is a random or variance error.
- ✓ Cleaning with Data discrepancy detection and Data transformation tools.

Data Integration

Data integration is defined as heterogeneous data from multiple sources combined in a common source (Datawarehouse). Data integration using Data Migration tools, Data Synchronization tools and ETL(Extract-Load-Transformation) process.

Data Selection

Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection. For this we can use Neural network, Decision Trees, Naive bayes, Clustering, and Regression methods.

Data Transformation

Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure. Data Transformation is a two step process:

1. **Data Mapping**: Assigning elements from source base to destination to capture transformations.

2. **Code generation**: Creation of the actual transformation program.

Data Mining

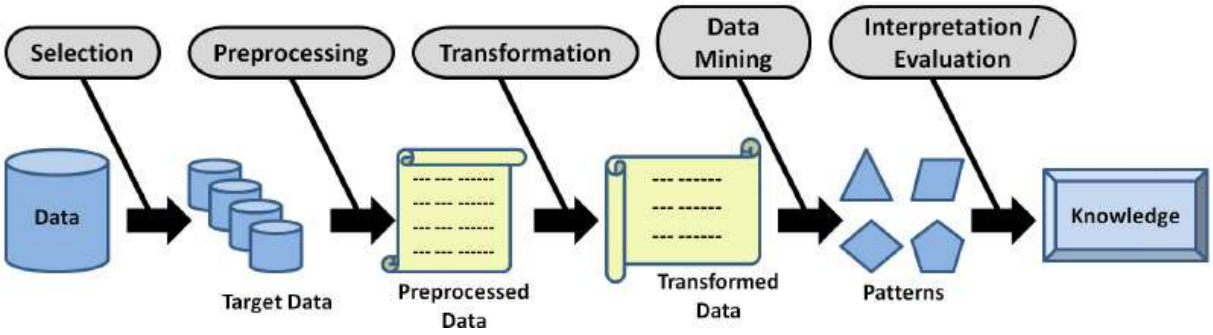
Data mining is defined as techniques that are applied to extract patterns potentially useful. It transforms task relevant data into patterns, and decides purpose of model using classification or characterization.

Pattern Evaluation

Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures. It find interestingness score of each pattern, and uses summarization and Visualization to make data understandable by user.

Knowledge Representation:

This involves presenting the results in a way that is meaningful and can be used to make decisions.



Note: KDD is an iterative process where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results. Preprocessing of databases consists of Data cleaning and Data Integration.

Advantages of KDD

1. **Improves decision-making:** KDD provides valuable insights and knowledge that can help organizations make better decisions.
2. **Increased efficiency:** KDD automates repetitive and time-consuming tasks and makes the data ready for analysis, which saves time and money.
3. **Better customer service:** KDD helps organizations gain a better understanding of their customers' needs and preferences, which can help them provide better customer service.
4. **Fraud detection:** KDD can be used to detect fraudulent activities by identifying patterns and anomalies in the data that may indicate fraud.
5. **Predictive modeling:** KDD can be used to build predictive models that can forecast future trends and patterns.

Disadvantages of KDD

1. **Privacy concerns:** KDD can raise privacy concerns as it involves collecting and analyzing large amounts of data, which can include sensitive information about individuals.
2. **Complexity:** KDD can be a complex process that requires specialized skills and knowledge to implement and interpret the results.

3. **Unintended consequences:** KDD can lead to unintended consequences, such as bias or discrimination, if the data or models are not properly understood or used.
4. **Data Quality:** KDD process heavily depends on the quality of data, if data is not accurate or consistent, the results can be misleading
5. **High cost:** KDD can be an expensive process, requiring significant investments in hardware, software, and personnel.
6. **Overfitting:** KDD process can lead to overfitting, which is a common problem in machine learning where a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new unseen data.

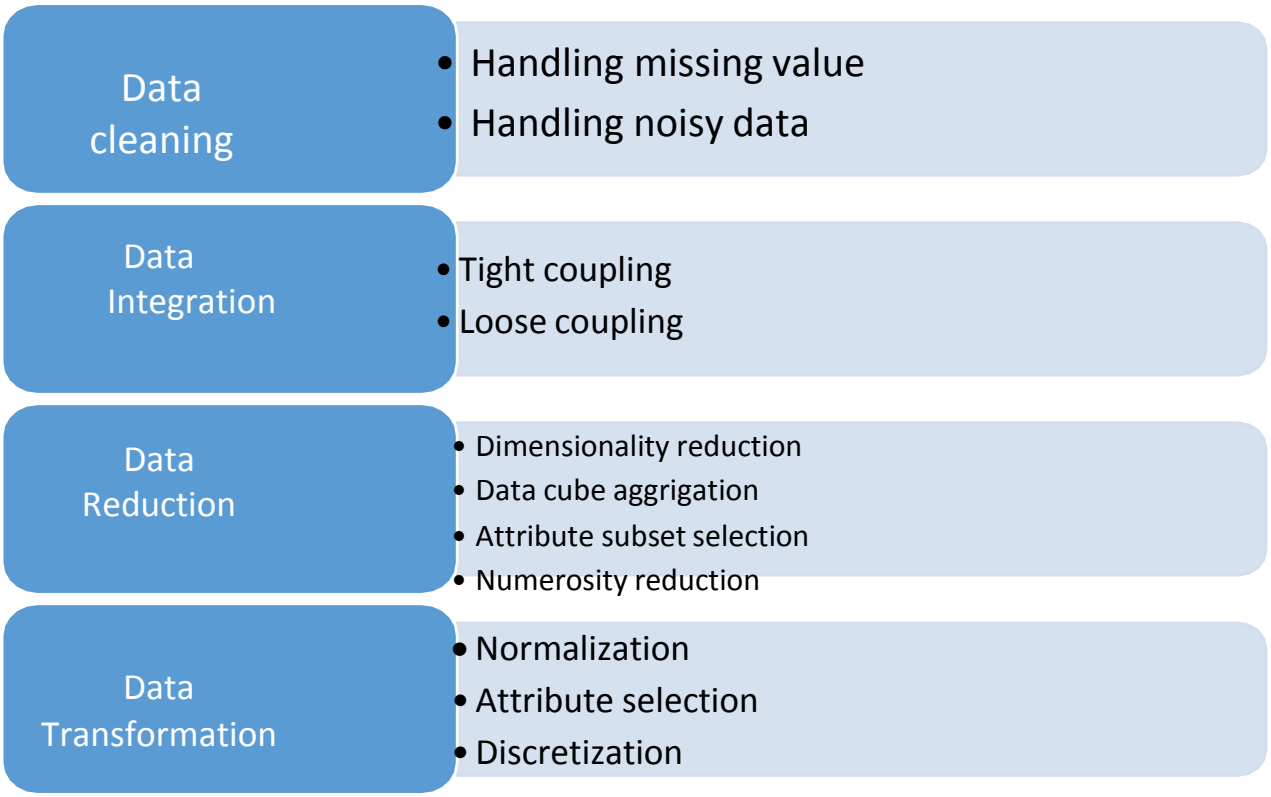
Difference between KDD and Data Mining

Parameter	KDD	Data Mining
Definition	KDD refers to a process of identifying valid, novel, potentially useful, and ultimately understandable patterns and relationships in data.	Data Mining refers to a process of extracting useful and valuable information or patterns from large data sets.
Objective	To find useful knowledge from data.	To extract useful information from data.
Techniques Used	Data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge representation and visualization.	Association rules, classification, clustering, regression, decision trees, neural networks, and dimensionality reduction.
Output	Structured information, such as rules and models, that can be used to make decisions or predictions.	Patterns, associations, or insights that can be used to improve decision-making or understanding.
Focus	Focus is on the discovery of useful knowledge, rather than simply finding patterns in data.	Data mining focus is on the discovery of patterns or relationships in data.

Parameter	KDD	Data Mining
Role of domain expertise	Domain expertise is important in KDD, as it helps in defining the goals of the process, choosing appropriate data, and interpreting the results.	Domain expertise is less critical in data mining, as the algorithms are designed to identify patterns without relying on prior knowledge.

DATA PREPROCESSING IN DATA MINING

Data preprocessing is a data mining technique which is used to transform or converting the raw data in a useful and efficient format.



DATA CLEANING

Real-world data tend to be incomplete, noisy, and inconsistent. *Data cleaning* (or *data cleansing*) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

Missing Values :

- ✧ In place of missing values we can replace with "NA" for raw data
- ✧ replace with "mean values" for normal distribution
- ✧ replace with "median values" for non- normal distribution.
- ✧ Sometimes replace with most probable values.Missing values can be filled in 2 ways.
  - Fill in the missing value manually: used for small data set
  - Fill in the missing value automatically: used for large data set

Noisy Data

Noise is a random error or variance in a measured variable. Let's look at the following data smoothing techniques:

- ◆ *Binning*: Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it. The sorted values are distributed into a number of "buckets," or *bins*. Because binning methods consult the neighborhood of values, they perform *local* smoothing.
  - ✧ In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.
  - ✧ smoothing by bin medians
  - ✧ In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the *bin boundaries*. Each bin value is then replaced by the closest boundary value.

- ◆ *Regression:* Data can be smoothed by fitting the data to a function, such as with regression. *Linear regression* involves finding the "best" line to fit two attributes (or variables), so that one attribute can be used to predict the other. *Multiple linear regression* is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.
- ◆ *Clustering:* Outliers may be detected by clustering, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers.

DATA INTEGRATION

It is likely that your data analysis task will involve *data integration*, which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, orflat files.

Data integration it can be done in 2 ways :

1. Tight coupling
2. Loose coupling

- ❖ *Tight coupling:* data is combined together into a physical location.

DS_A A + DS_B B= C (physical location)

Individual data sources A and B .Integrated in C. After integration in C physical location you cant access separately in DS_A and DS_B

- *Loose coupling:* The data is actually not integrated. Only interface is created and data is combined through the if and also accessed through that interfaces.

Data remains in actual database only.

DATA REDUCTION:

The data set will likely be huge! Complex data analysis and mining on huge amounts of data can take a long time, making such analysis impractical or infeasible.

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume.

If the volume of the data is very high so the performance also will be low.the volume of data is reduce to make analysis easier.if the data is reduce you can do in two ways.

1. Losy: some of the data will be lost.
2. Lossless: data never be lost . online PDF compress are the best example. Methods for data reduction:

- *Data cube aggregation*: Where aggregation operations are applied to the data in the construction of a data cube.
- *Attribute subset selection:* Where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.
- *Dimensional reduction:* Where encoding mechanisms are used to reduce the data set size.
- *Numerosity reduction:* Where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or non parametric methods such as clustering, sampling, and the use of histograms.

DATA TRANSFORMATION:

Data transformation in data preprocessing is an essential step in the data mining process. It forms an integral part of data mining. Data is transformed into appropriate form suitable for mining process.

Methods in data transformation:

- Normalization: Calculating/creating scaling values for given data for mining process.
- Attribute selection: Generating new attributes from older once.
- Discretization: Preparing the interval values form the given raw data