Name: A. Shiva Surya Saran
Roll No.: BE-B 20

Data Mining and Warehousing

Assignment 1

Unit 1: Introduction

1. *Describe three challenges to data mining regarding data mining methodology*

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place.

Issues in Data Mining Methodology:

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore, it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

2. Compare OLTP and OLAP.

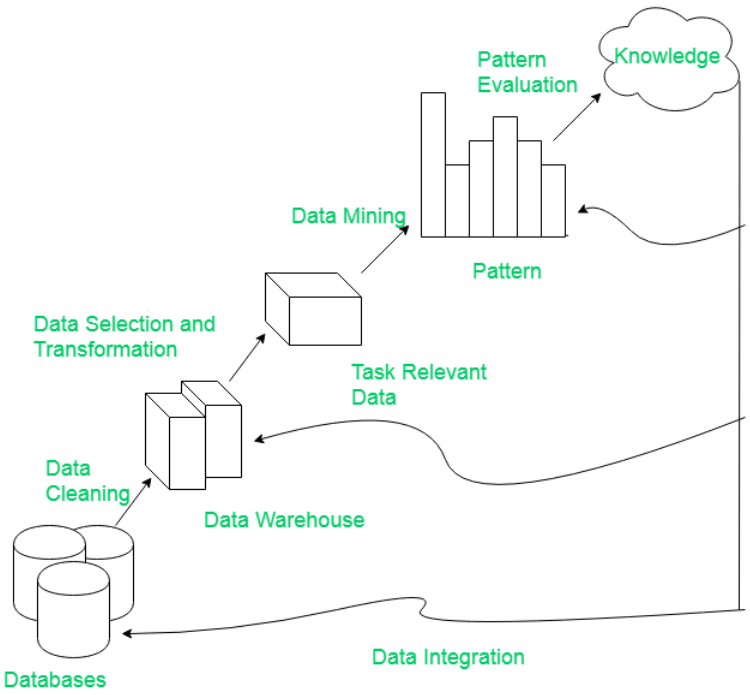
Parameters	Online transaction processing (OLTP)	Online Analytical Processing (OLAP)
Process	It is an online transactional system. It manages database modification.	OLAP is an online analysis and data retrieving process.
Characteristic	It is characterized by large numbers of short online transactions.	It is characterized by a large volume of data.
Functionality	OLTP is an online database modifying system.	OLAP is an online database query management system.
Method	OLTP uses traditional DBMS.	OLAP uses the data warehouse.
Query	Insert, Update, and Delete information from the database.	Mostly select operations
Table	Tables in OLTP database are normalized.	Tables in OLAP database are not normalized.
Source	OLTP and its transactions are the sources of data.	Different OLTP databases become the source of data for OLAP.
Data Integrity	OLTP database must maintain data integrity constraint.	OLAP database does not get frequently modified. Hence, data integrity is not an issue.
Response time	It's response time is in millisecond.	Response time in seconds to minutes.
Data quality	The data in the OLTP database is always detailed and organized.	The data in OLAP process might not be organized.
Usefulness	It helps to control and run fundamental business tasks.	It helps with planning, problem-solving, and decision support.
Operation	Allow read/write operations.	Only read and rarely write.
Audience	It is a market orientated process.	It is a customer orientated process.
Query Type	Queries in this process are standardized and simple.	Complex queries involving aggregations.
Back-up	Complete backup of the data combined with incremental backups.	OLAP only need a backup from time to time. Backup is not important compared to OLTP
Design	DB design is application oriented. Example: Database design changes with industry like Retail, Airline, Banking, etc.	DB design is subject oriented. Example: Database design changes with subjects like sales, marketing, purchasing, etc.
User type	It is used by Data critical users like clerk, DBA & Data Base professionals.	Used by Data knowledge users like workers, managers, and CEO.
Purpose	Designed for real time business operations.	Designed for analysis of business measures by category and attributes.
Performance metric	Transaction throughput is the performance metric	Query throughput is the performance metric.

Number of users	This kind of Database users allows thousands of users.	This kind of Database allows only hundreds of users.
Productivity	It helps to Increase user's self-service and productivity	Help to Increase productivity of the business analysts.
Challenge	Data Warehouses historically have been a development project which may prove costly to build.	An OLAP cube is not an open SQL server data warehouse. Therefore, technical knowledge and experience is essential to manage the OLAP server.
Process	It provides fast result for daily used data.	It ensures that response to the query is quicker consistently.
Characteristic	It is easy to create and maintain.	It lets the user create a view with the help of a spreadsheet.
Style	OLTP is designed to have fast response time, low data redundancy and is normalized.	A data warehouse is created uniquely so that it can integrate different data sources for building a consolidated database

3. With suitable example diagram explain various steps of a KDD process and Briefly explain each step.

Data Mining also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unknown, and potentially useful information from data stored in databases.

Steps Involved in KDD Process:



i. 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.

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

iii. Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
   - Data selection using Neural network.
   - Data selection using Decision Trees.
   - Data selection using Naive Bayes.
   - Data selection using Clustering, Regression, etc.

iv. 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:
   - Data Mapping: Assigning elements from source base to destination to capture transformations.
   - Code generation: Creation of the actual transformation program.

v. Data Mining: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
   - Transforms task relevant data into patterns.
   - Decides purpose of model using classification or characterization.

vi. Pattern Evaluation: Pattern Evaluation is defined as as identifying strictly increasing patterns representing knowledge based on given measures.
   - Find interestingness score of each pattern.
   - Uses summarization and Visualization to make data understandable by user.

vii. Knowledge representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
   a. Generate reports.
   b. Generate tables.
   c. Generate discriminant rules, classification rules, characterization rules, etc.

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

4. What is data pre-processing? Explain the different steps in data Pre-processing

Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Steps Involved in Data Pre-processing:

1. **Data Cleaning**
   The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.
   a. *Missing Data*
      This situation arises when some data is missing in the data. It can be handled in various ways like ignoring the tuples and filling the missing values.
   b. Noisy Data
      Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can by Binning Methods, Regression and Clustering

2. **Data Transformation**
   This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:
   a. Normalization:
      It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

   b. Attribute Selection:
      In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

   c. Discretization:
      This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

   d. Concept Hierarchy Generation:
      Here attributes are converted from level to higher level in hierarchy. For Example-The attribute "city" can be converted to "country".

3. **Data Reduction:**
   Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs. The various steps to data reduction are:
   a. Data Cube Aggregation:
      Aggregation operation is applied to data for the construction of the data cube.

b. **Attribute Subset Selection:**
The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.

c. **Numerosity Reduction:**
This enable to store the model of data instead of whole data, for example: Regression Models.

d. **Dimensionality Reduction:**
This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Componenet Analysis).

5. What is predictive and descriptive data mining. Explain in brief any two task for both predictive and descriptive data mining with one example.

Descriptive mining:
This term is basically used to produce correlation, cross-tabulation, frequency etc. These technologies are used to determine the similarities in the data and to find existing patterns. One more application of descriptive analysis is to develop the captivating subgroups in the major part of the data available. This analytics emphasis on the summarization and transformation of the data into meaningful information for reporting and monitoring.

Predictive Data Mining:
The main goal of this mining is to say something about future results not of current behaviour. It uses the supervised learning functions which are used to predict the target value. The methods come under this type of mining category are called classification, time-series analysis and regression. Modelling of data is the necessity of the predictive analysis, and it works by utilizing a few variables of the present to predict the future not known data values for other variables.

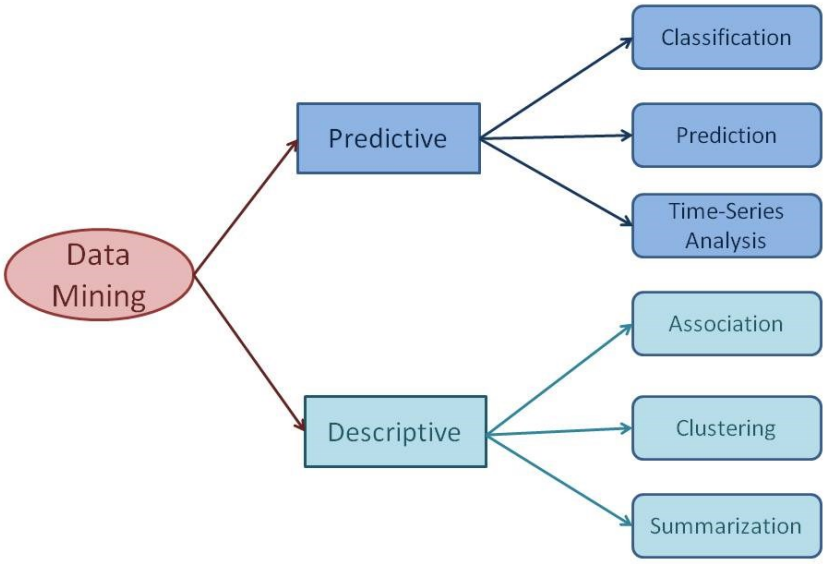


Fig. Different Data Mining Tasks

a) Classification
Classification derives a model to determine the class of an object based on its attributes. A collection of records will be available, each record with a set of attributes. One of the attributes will be class attribute and the goal of classification task is assigning a class attribute to new set of records as accurately as possible.
Classification can be used in direct marketing, that is to reduce marketing costs by targeting a set of customers who are likely to buy a new product. Using the available data, it is possible to know which customers purchased similar products and who did not purchase in the past. Hence, {purchase, don't purchase} decision forms the class attribute in this case.

Once the class attribute is assigned, demographic and lifestyle information of customers who purchased similar products can be collected and promotion mails can be sent to them directly.

b) Prediction

Prediction task predicts the possible values of missing or future data. Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest. For example, a model can predict the income of an employee based on education, experience and other demographic factors like place of stay, gender etc. Also prediction analysis is used in different areas including medical diagnosis, fraud detection etc.

c) Association

Association discovers the association or connection among a set of items. Association identifies the relationships between objects. Association analysis is used for commodity management, advertising, catalogue design, direct marketing etc. A retailer can identify the products that normally customers purchase together or even find the customers who respond to the promotion of same kind of products.

e) Clustering

Clustering is used to identify data objects that are similar to one another. The similarity can be decided based on a number of factors like purchase behavior, responsiveness to certain actions, geographical locations and so on. For example, an insurance company can cluster its customers based on age, residence, income etc. This group information will be helpful to understand the customers better and hence provide better customized services.

Q6. Discuss whether each of the following activities is a data mining task.

i. Computing the total sales of a company
   This activity is not a data mining task because the total sales can be computed by using simple calculations.

ii. Predicting the future stock price of a company using historical records.
   This activity is a data mining task. Historical records of stock price can be used to create a predictive model called regression, one of the predictive modeling tasks that is used for continuous variables.

iii. Predicting the outcomes of tossing a pair of dice.
   This activity is not a data mining task because predicting the outcome of tossing a fair pair of dice is a probability calculation, which doesn't have to deal with large amount of data or use complicate calculations or techniques.

Q7. Following data for the attribute age:
13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40, 45,46,52,70

a) Use smoothing by bin means to smooth these data, using a bin depth of 3.

N = 27
Bin size = 3

Bin#	Values	Smoothing by Bin Means
1	13, 15, 16	14.67, 14.67, 14.67
2	16, 19, 20	18.33, 18.33, 18.33
3	20, 21, 22	21, 21, 21
4	22, 25, 25	24, 24, 24
5	25, 25, 30	26.67, 26.67, 26.67
6	33, 33, 35	33.67, 33.67, 33.67
7	35, 35, 35	35, 35, 35
8	36, 40, 45	40.33, 40.33, 40.33
9	46, 52, 70	56, 56, 56

b) How might u determine outliers in the data?

Outliers in data can be identified in several ways.

- By dividing the data into equi-width histograms and identifying the outlying histograms.
- By clustering the data into groups. Any data that do not fall in a group can be taken as outliers.
- In general, fit a model to the data. Any data points that deviate significantly (based on some threshold) from the model can be considered outliers.

c) What other methods are there for data smoothing?

Other methods that can be used for data smoothing include alternate forms of binning such as smoothing by bin medians or smoothing by bin boundaries. Alternatively, equal-width bins can be used to implement any of the forms of binning, where the interval range of values in each bin is constant. Methods other than binning include using regression techniques to smooth the data by fi setting it to a function such as through linear or multiple regression.

Q8. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem

The various methods for handling the problem of missing values in data tuples include:

a) Ignoring the tuple: This is usually done when the value is missing. This method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.
b) Manually filling in the missing value: In general, this approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined.
c) Using a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like \Unknown," Hence, although this method is simple, it is not recommended.
d) Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values: For example, suppose that the average income of All Electronics customers is $28,000. Use this value to replace any missing values for income.
e) Using the most probable value to fill in the missing value.

Q9. Consider the following group of data 200, 300, 400, 600,1000.

i. Use the min-max normalization to transform value 600 onto the range [0.0,1.0].
Min = 0
Max = 1

V’ = 600
Min_{Value} = 200
Max_{Value} = 1000

Min Max = [V’ - Min_{Value} (Max - Min) / Max_{Value} - Min_{Value}] + Min

Min Max = [600 – 200 (1-0) / 1000 – 200] + 0

Min Max = 0.5

ii. Use the decimal scaling to transform value 600.
V’ = V/10

V’ = 600/1000
V’ = 0.6

0.6 is the Normalized value after Decimal scaling.

Q10. What are the OLAP operations.

There are five basic analytical operations that can be performed on an OLAP cube:

1. **Drill Down**
   In drill-down operation, the less detailed data is converted into highly detailed data. It can be done by:
   - Moving down in the concept hierarchy
   - Adding a new dimension

2. **Roll Up**
   It is just opposite of the drill-down operation. It performs aggregation on the OLAP cube. It can be done by:
   - Climbing up in the concept hierarchy
   - Reducing the dimensions

3. **Dice**
   It selects a sub-cube from the OLAP cube by selecting two or more dimensions.

4. **Slice**
   It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the cube given in the overview section, Slice is performed on the dimension Time = “Q1”.

5. **Pivot**
   It is also known as *rotation* operation as it rotates the current view to get a new view of the representation. In the sub-cube obtained after the slice operation, performing pivot operation gives a new view of it.

Q11. Explain the following

1. **Normalization**
   Normalization is used to scale the data of an attribute so that it falls in a smaller range, such as -1.0 to 1.0 or 0.0 to 1.0. It is generally useful for classification algorithms.

   It is used to scale the data of an attribute so that it falls in a smaller range, such as -1.0 to 1.0 or 0.0 to 1.0. It is generally useful for classification algorithms.

   Methods of Data Normalization –
   a. Decimal Scaling
   b. Min-Max Normalization
   c. z-Score Normalization (zero-mean Normalization)

2. **Min-max normalizations**
   In this technique of data normalization, linear transformation is performed on the original data. Minimum and maximum value from data is fetched and each value is replaced according to the following formula.

$$\mathbf{v' = \frac{v - min(A)}{max(A) - min(A)}(new\_max(A) - new\_min(A)) + new\_min(A)}$$

Where A is the attribute data,
Min(A), Max(A) are the minimum and maximum absolute value of A respectively.
v' is the new value of each entry in data.
v is the old value of each entry in data.
new_max(A), new_min(A) is the max and min value of the range(i.e boundary value of range required) respectively.

3. **Z-score normalization**
   In this technique, values are normalized based on mean and standard deviation of the data A. The formula used is:

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

v', v is the new and old of each entry in data respectively. σA, A is the standard deviation and mean of A respectively.

4. **Decimal Scaling**
   It involves the data transformation by dragging the decimal points of values of feature F. The movement of decimals is very dependent on the absolute value of the maximum. A value of feature F is transformed to by calculating:

$$v' = \frac{v}{10^j}.$$

In this formula, j is the lowest integer while Max (|) < 1.
This technique entails the transformation of the decimal points of the values according to the absolute value of the maximum. It follows that the means of the normalized data will always be between 0 and 1.

Q12. Differentiate between: -

a) Supervised learning and unsupervised learning

Parameters	Supervised machine learning technique	Unsupervised machine learning technique
Process	In a supervised learning model, input and output variables will be given.	In unsupervised learning model, only input data will be given
Input Data	Algorithms are trained using labeled data.	Algorithms are used against data which is not labeled
Algorithms Used	Support vector machine, Neural network, Linear and logistics regression, random forest, and Classification trees.	Unsupervised algorithms can be divided into different categories: like Cluster algorithms, K-means, Hierarchical clustering, etc.
Computational Complexity	Supervised learning is a simpler method.	Unsupervised learning is computationally complex
Use of Data	Supervised learning model uses training data to learn a link between the input and the outputs.	Unsupervised learning does not use output data.
Accuracy of Results	Highly accurate and trustworthy method.	Less accurate and trustworthy method.
Real Time Learning	Learning method takes place offline.	Learning method takes place in real time.
Number of Classes	Number of classes is known.	Number of classes is not known.
Main Drawback	Classifying big data can be a real challenge in Supervised Learning.	You cannot get precise information regarding data sorting, and the output as data used in unsupervised learning is labeled and not known.

b) classification and regression

Parameter	Classification	Regression
Basic	Mapping Function is used for mapping of values to predefined classes.	Mapping Function is used for mapping of values to continuous output.
Involves prediction of	Discrete values	Continuous values
Nature of the predicted data	Unordered	Ordered
Method of calculation	by measuring accuracy	by measurement of root mean square error
Example Algorithms	Decision tree, logistic regression, etc.	Regression tree (Random forest), Linear regression, etc.

c) ***Descriptive and predictive data mining task***

Parameter	Descriptive Data Mining	Predictive Data Mining
Basic	It determines, what happened in the past by analyzing stored data.	It determines, what can happen in the future with the help past data analysis.
Preciseness	It provides accurate data.	It produces results does not ensure accuracy.
Practical analysis methods	Standard reporting, query/drill down and ad-hoc reporting.	Predictive modelling, forecasting, simulation and alerts.
Require	It requires data aggregation and data mining	It requires statistics and forecasting methods
Type of approach	Reactive approach	Proactive approach
Describe	Describes the characteristics of the data in a target data set.	Carry out the induction over the current and past data so that predictions can be made.