KARPAGAM ACADEMY OF HIGHER EDUCATION

Coimbatore-641 021

(For the candidates admitted from 2017 onwards)

DEPARTMENT OF CS, CA & IT

SYLLABUS

SUBJECT NAME: DATA MINING **SEMESTER : V**

SUBJECT CODE: 17ITU503B **CLASS: III B. Sc. [IT]**

Instruction Hours / week: L: 4 T: 0 P: 0 **Marks:** Int : **40** Ext : **60** Total: **100**

SCOPE

This course introduce students to the basic concepts and techniques of Data Mining, develop skills of using recent data mining software for solving practical problems, gain experience of doing independent study and research.

OBJECTIVES

- To introduce students to the basic concepts and techniques of Data Mining.
- To develop skills of using recent data mining software for solving practical problems.
- To gain experience of doing independent study and research.
- Possess some knowledge of the concepts and terminology associated with database systems, statistics, and machine learning

UNIT-I

Overview: Predictive and descriptive data mining techniques

UNIT-II

Supervised and unsupervised learning techniques

UNIT-III

Process of knowledge discovery in databases, pre-processing methods

UNIT-IV

Data Mining Techniques: Association Rule Mining, classification and regression techniques, clustering

UNIT-V

Scalability and data management issues in data mining algorithms, measures of interestingness.

Suggested Readings

1. Pang-Ning Tan., Michael Steinbach.,& Vipin Kumar. (2005). Introduction to Data Mining. New Delhi: Pearson Education.
2. Richard Roiger., & Michael Geatz. (2003). Data Mining: A Tutorial Based Primer. New Delhi: Pearson Education.
3. Gupta, G.K. (2006). Introduction to Data Mining with Case Studies. New Delhi: PHI.
4. Soman, K. P., Diwakar Shyam., & Ajay, V. (2006). Insight Into Data Mining: Theory And Practice. New Delhi: PHI.

WEB SITES

1. Thedacs.Com
2. Dwreview.Com
3. Pcai.Com
4. Eruditionhome.Com

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established Under Section 3 of UGC Act 1956)
Pollachi Main Road, Eachanari Post, Coimbatore - 641021
(For the candidates admitted from 2016 onwards)
DEPARTMENT OF CS, CA & IT

SUBJECT : DATA MINING **SEMESTER : V**
SUBJECT CODE: 17ITU503B **CLASS : III B.Sc.(IT)**

LECTURE PLAN

S.NO	LECTURE DURATION (Hour)	TOPICS TO BE COVERED	SUPPORT MATERIALS
		UNIT I	
1	1	**Introduction,** Data Mining, KDD and Data Mining, Classification	S1: 13-18
2	1	Data Mining Cycle, Functionalities	W2
3	1	Data Mining model tasks	W2
4	1	Predictive Data Model Techniques	W2
5	1	Descriptive Data Model Techniques	W2
6	1	Predictive analytics Process	W2
	1	Selecting Modeling Paradigm	W3
8	1	Application of analytics	W3
9	1	**Recapitulation of Unit I** **Discussion of Important Questions**	
		Total no. of Hours Planned for Unit – I	**9 Hrs**
		UNIT – II	
1	1	**Learning,** Machine Learning, Definition for Supervised and unsupervised	W4
2	1	Supervised and unsupervised learning with a Real life Example	W4
3	1	Supervised Machine learning	W5
4	1	Unsupervised Machine learning	W5
5	1	Semi supervised Machine learning	W4
6	1	Analysis of Supervised and Unsupervised Data Mining	W4
7	1	Algorithm choice based on supervised and unsupervised learning	W5
8	1	**Recapitulation of Unit II**	

9	1	**Discussion of Important Questions**	
		Total no. of Hours Planned for Unit – II	**9 Hrs**
UNIT III			
1	1	Knowledge discovery database, KDD Process	S2:9-15
2	1	Data Preprocessing-Overview	S2:47-48
3	1	Data Cleaning, Noisy Data	S2:61-67
4	1	Data Integration and Transformation	S2:67-70
5	1	Data Reduction	S2:72-80
6	1	Data Discretization	S2:80-88
7	1	Concept Hierarchy Generation	S2:88-94
8	1	**Recapitulation of Unit III** **Discussion of Important Questions**	
		Total no. of Hours Planned for Unit - III	**8 Hrs**
UNIT IV			
1	1	Association: Basic concepts, Frequent Item set, Closed Item set and Association Rule	S3:160-166
2	1	Various Kinds of Association Rule	S3:166-170
3	1	Classification: Introduction – Statistical based algorithms	S3:69-83,W3
4	1	Distance based, Decision tree base algorithm	S3:86-96
5	1	Neural network-based, Rule-based algorithm	S3:99-113,W3
6	1	Clustering: Introduction-Similarity and Distance measures, Outliers	S3:119-125
7	1	Hierarchical algorithms, Partitioned algorithms	S3:125-138 S3:170-171
8	1	Parallel & Distributed algorithm	S3:176-180
9	1	Comparing approaches, incremental rules	S3:180-183
10	1	Measuring the quality of rules & Regression	S3:183-189, S3:76-83,W3
11	1	**Recapitulation of Unit IV** **Discussion of Important Questions**	
		Total no. of Hours Planned for Unit - IV	**11 Hrs**
UNIT V			
1	1	Scalability of Data Mining	W3

2	1	Data Mining Issues	W3
3	1	Mining Methodology and user Interaction Issues	W3
4	1	Performance Issues	W3
5	1	Diverse Data Types Issues	W3
6	1	Interesting Measures	W3
7	1	Application of Data Mining in Social sector	W1
8	1	**Recapitulation and discussion of important Questions**	
9	1	**Previous Year ESE Questions Discussion**	
10	1	**Previous Year ESE Questions Discussion**	
11	1	**Previous Year ESE Questions Discussion**	
		Total no. of Hours Planned for Unit - V	**11 Hrs**
		Total Planned Hours	**48 Hrs**

SUPPORTING MATERIAL

1. S1: Data Mining,Pieter Adriaans & Dolf zantinge pearson education.
2. Jaiwaei Han, Micheline Kamber, Jian pei, 2012 Data Mining: Concepts and Techniques, Morgan Kaufmann publishers.
3. Margaret H.Dunham, 2008, 3rd edition, Data mining introductory and advancrd topics, person education.

Websites

W1: http://www.webopedia.com
W2:www.ijarcsse.com/
W3:www.tutorialpoint.com/data-mining
W4:Shodhganya.inflibnet.ac.in
W5:https://en.wikipedia.org/machine learning/

Faculty HoD

Syllabus

Overview: Predictive and descriptive data mining techniques

Overview: **Predictive and Descriptive data mining techniques**

Data Mining:

Data Mining refers to extracting or “mining” knowledge from large amounts of data. The overall goals of the data mining process are extraction of information from large data sets and transform it into some understandable structure for further uses. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods such as neural networks or decision trees. Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data. Finding useful patterns in data is known by different names (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing).

The most important predictive and descriptive data mining techniques by which most of the mining task is performed.

The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data. Finding useful patterns in data is known by different names (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing).

Classification of Data Mining System

Data mining systems can be categorized according to various criteria as follows:

Classification of data mining systems according to the type of data sources mined:

This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

Classification of data mining systems according to the database involved:

This classification based on the data model involved such as relational database, objectoriented database, data warehouse, transactional database, etc.

Classification of data mining systems according to the kind of knowledge discovered:

This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

Classification of data mining systems according to mining techniques used:

This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc.

The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems.

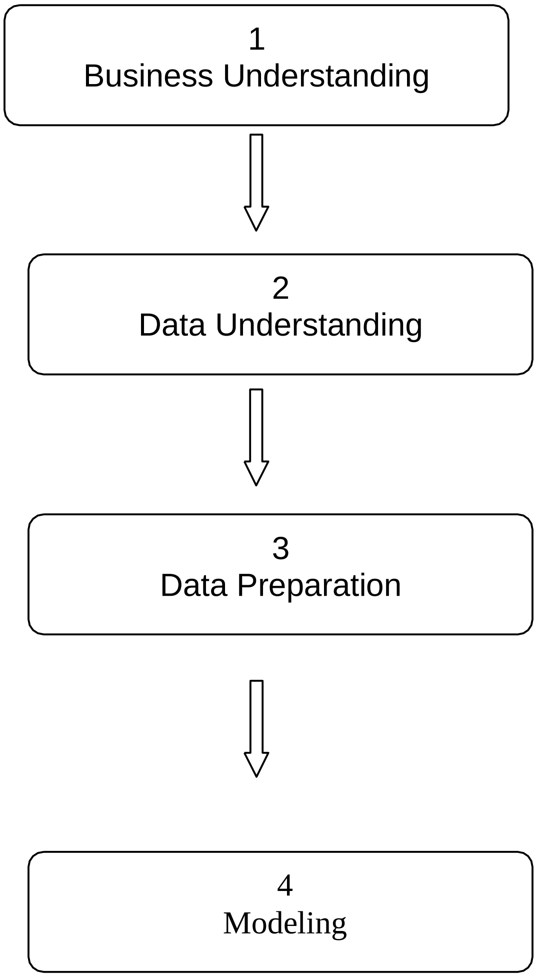
Data Mining Life cycle

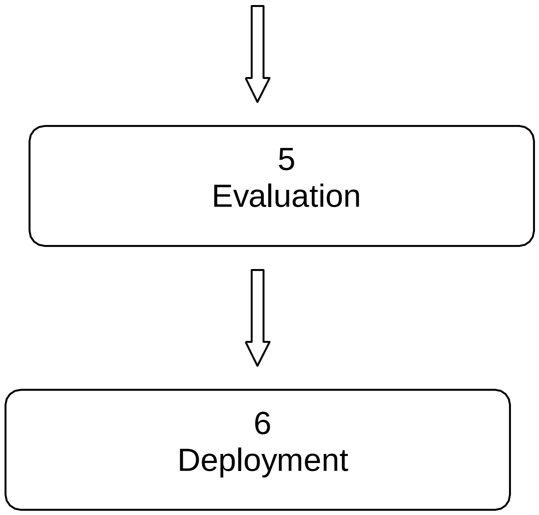
The life cycle of a data mining project consists of six phases [91, 26]. The sequence of the phases is not rigid. Moving back and forth between different phases is always required depending upon the outcome of each phase. The main phases are:

Business Understanding:

This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Figure 3.2: Phases of Data Mining Life Cycle





Data Understanding:

It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Data Preparation:

Covers all activities to construct the final dataset from raw data.

Modeling:

In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

Evaluation:

In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business

Deployment:

The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Data Mining Functionalities

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data. The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list:

Characterization:

It is the summarization of general features of objects in a target class, and produces what is called characteristic rules. The data relevant to a userspecified
class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions. For example, one may wish to characterize the customers of a store who regularly rent more than 30 movies a year. With concept hierarchies on the attributes describing the target class, the attribute oriented induction method can be used to carry out data summarization. With a data cube containing summarization of data, simple OLAP operations fit the purpose of data characterization.

Discrimination:

Data discrimination produces what are called discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. For example, one may wish to compare the general characteristics of the customers who rented more than 30 movies in the last year with those whose rental account is lower than 5. The techniques used for data discrimination are similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures.

Association analysis:

Association analysis studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. This is commonly used for market basket analysis. For example, it could be useful for the manager to know what movies are often rented together or if there is a relationship between renting a certain type of movies and buying popcorn or pop. The discovered association rules are of the form:

$$P \rightarrow Q\ [s, c],$$

where P and Q are conjunctions of attribute value-pairs, and s (support) is the probability that P and Q appear together in a transaction and c (confidence) is the conditional probability that Q appears in a transaction when P is present. For example, RentType(X,“game”) □Age(X,“13-19”)_Buys(X,“pop”)[s=2%, =55%]

The above rule would indicate that 2% of the transactions considered are of customers aged between 13 and 19 who are renting a game and buying a pop, and that there is a certainty of 55% that teenage customers who rent a game also buy pop.

Classification:

It is the organization of data in given classes. Classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the manager of a store could analyze the customers' behavior vis-à-vis their credit, and label accordingly the customers who received credits with three possible labels "safe", "risky" and "very risky". The classification analysis would generate a model that could be used to either accept or reject credit requests in the future.

Prediction:

Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values.

Clustering:

Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).

Outlier analysis:

Outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable.

Evolution and deviation analysis:

Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values. It is common that users do not have a clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore important to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction. This also makes interactivity an important attribute of a data mining system.

Data Mining Models

The data mining models are of two types [146, 70]: Predictive and Descriptive.

Descriptive Models

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Summarization, Association rule, Sequence discovery etc. Clustering is similar to classification except that the groups are not predefined, but are defined by the data alone. It is also referred to as unsupervised learning or segmentation. It is the partitioning or segmentation of the data in to groups or clusters. The clusters are defined by studying the behavior of the data by the domain experts. The term segmentation is used in very specific context; it is a process of partitioning of database into disjoint grouping of similar tuples. Summarization is the technique of presenting the summarize information from the data. The association rule finds the association between the different attributes. Association rule mining is a two step process: Finding all frequent item sets, Generating strong association rules from the frequent item sets. Sequence discovery is a process of finding the sequence patterns in data. This sequence can be used to understand the trend.

Predictive Models

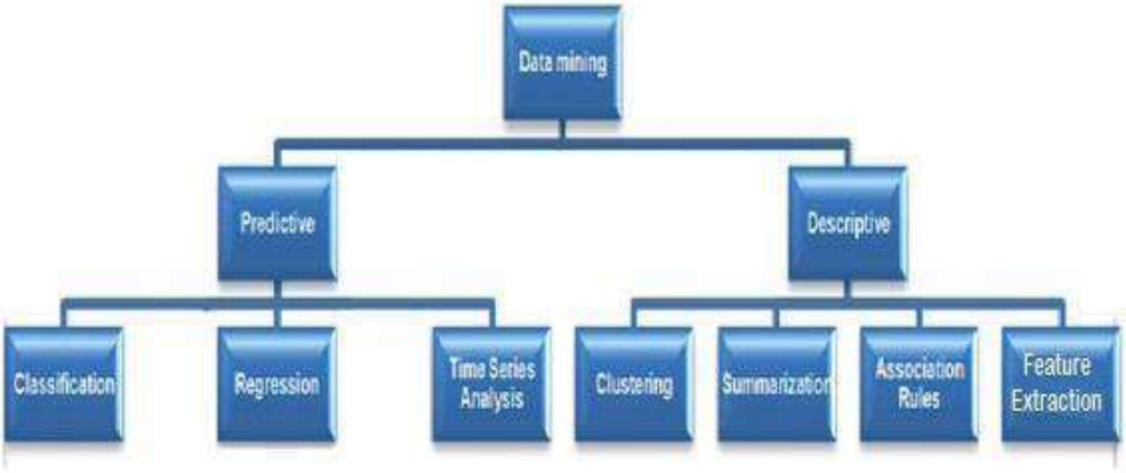
The predictive model makes prediction about unknown data values by using the known values. Ex. Classification, Regression, Time series analysis, Prediction etc. Many of the data mining applications are aimed to predict the future state of the data.

Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefined groups or classes, this is a supervise learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that map data item to real valued prediction variable. In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

DATA MINING TASKS

Data mining satisfy its main goal by identifying valid, potentially useful, and easily understandable correlations and patterns present in existing data. This goal of data mining can be satisfy by modeling it as either Predictive or Descriptive nature. The Predictive model works by making a prediction about values of data, which uses known results found from different datasets. The tasks include in the Predictive data mining model includes classification, prediction, regression and analysis of time series. The Descriptive model mostly identify patterns or relationships in datasets. It serves as a easy way to explore the properties of the data examined earlier and not to predict new properties. The Descriptive model encompasses task to perform as Clustering, Association Rules, Summarizations, and Sequence Analysis. The classification of data mining task as Predictive and Descriptive, along with their own methods is shown in below Fig.

Predictive and Descriptive Data mining models



The descriptive data-mining model is discover patterns in the data and understands the relationships between attributes represented by the data. In contrast, the predictive data-mining modelpredicts the future outcomes based on passed records present in the database or with known answers. The data-mining task is further divided into the following two approaches:

Supervised or directed data mining

The goal of supervised or directed data mining is to use the available data as like predictive data-mining to build a model that describes one particular variable of interest in terms of the rest of the available data. The user selects the target field and directs the computer to determine estimate, classify or predict its value.

Unsupervised or undirected data mining

In unsupervised or undirected data mining however variable is singled out as the target as like the descriptive mining technique. The goal is rather to establish some relationship among all the variables in the data. The user asks the computer to identify patterns in the data that may be significant. Undirected modeling is used to explain those patters and relationships one they have been found.

PREDICTIVE DATA MINING MODEL

The purpose of Predictive mining model is mainly to predict the future outcome than current behavior. The prediction output can be numeric value or in categorized form. The predictive models is the supervised learning functions which predicts the target value. Predictive methods to be utilized in the next phase of empirical study. Use some variables to predict unknown or future values of other variables in addition comparison between these supervised approaches was also conducted to get some insight about the strength and weaknesses of each approach since one of the aims of the study was to determine whether these methods were well suited for extracting the required knowledge. As a result, the predictive method will be able to predict in which cluster does the future student will fall into based on the enrollment information.

A. Classification

Among Predictive data mining technique, Classification model is consider as the best-understood technique of all data mining approaches. The common characteristics of classification tasks are as supervised learning, categories dependent variable and assigning new data to one of a set of well-defined classes. Classification technique is used in customer segmentation, modeling businesses, credit analysis, and many other applications. Consider you have given classes of patients that corresponding to response to medical treatment; a new patient is most likely to respond is identified by the form of treatment suggested to earlier class of patient.

In a classification technique, you typically have historical data called labeled examples and new examples. Each labeled example consists of multiple predictor attributes and one target attribute that is a class label. While unlabeled examples only consist of the predictor attributes. The goal of classification is to construct a model using the data from history and accurately predicts the new class of examples .A classification task begins with build data in database also know as training data for which the target values are known. There are different classification algorithms available that uses their different techniques for finding relations between the predictor attributes values and the target values in the build data. After getting the targeted data, these relations are summarized in a model, so that they can be applied to new cases further with unknown target values for predicting target values.
E.g., classify countries based on climate, or classify cars based on gas mileage.

B. Regression

Statistical Regression is another Predictive data-mining model also known as is a supervised learning technique. This technique analyzes of the dependency of some attribute values, which is dependent upon the values of other attributes mainly present in same item. The development of a model can predict these attribute values for new cases. The difference between regression and classification is that regression deals with numerical/continuous target attributes, whereas classification deals with discrete/categorical target attributes. In other words, if the target attribute contains continuous (floating-point) values, a regression technique is required. The most common form of regression is linear regression, in which a line that best fits the data is calculated, that is, the line that minimizes the average distance of all the points from the line. This line becomes a predictive model when the value of the dependent variable is not known; its value is predicted by the point on the line that corresponds to the values of the independent variables for that record.

C. Time-Series Analysis

Time-series database is a sequence database, consisting sequences of values or events obtained over repeated measurements of time. The values are typically measure at equal time interval such as hourly, daily, weekly. A sequence database is any database that consisting sequences of ordered events, sometimes having concrete notions of time. For example, Web page traversal sequences and customer shopping transaction sequences are sequence data, but they may not be time-series data. The time-series analysis is a prediction application with one or more time-dependent attributes that usually involves prediction of numeric outcomes such as the future price of individual stock. This time-series represents a collection of values obtained from sequential measurements. Time-series data mining gives natural ability to visualize the shape of data. Time series are very long, considered smooth, as subsequent values are within predictable ranges of one another. Time-series are popular in many applications, such as stock market analysis, economic and sales forecasting. It can be useful for observation of natural phenomena like atmosphere, temperature, wind, earthquake, scientific and engineering experiments, and medical treatments.

DESCRIPTIVE DATA MINING MODEL

The second approach for mining data from large datasets is known as Descriptive data mining. It is normally used to generate correlation, frequency, cross tabulation, etc. This Descriptive method can be defined as to discover regularities in the data and to uncover patterns. This is also used to find interesting subgroups in the bulk of data. Some researchers used Descriptive to determine the demographic influence on some particular factors. Descriptive approaches employed in the study were identified, namely the Summarization and Clustering methods. Since the aim of the experiment was to study the patterns and getting some information within the enrollment data, no specific target has been identified. To this end, clusters were generated by Clustering method, and later used as target or output for Predictive approach.

A. Clustering

Clustering is the most important descriptive data mining technique, in which a set of data items is partitioned into a set of classes also called as groups. Clustering is the process of finding natural groups, called as clusters, in a database. The set of data items in a group have similar characteristics. Hence, it is mainly used for finding groups of similar items. It is the identification process for classes for a set of objects whose classes are not known. These objects ware so clustered that their intra class similarities can be maximized and minimized on the criteria defined by attributes of objects. The object is label with their corresponding clusters once their particular class or clusters are decided. For example, between a data set of customers, that have a similar buying behavior can be identified with subgroups of customers.

B. Summarization

Summarization is called as the abstraction or generalization of data. The summarization technique maps data into subsets with simple descriptions. The summarized data set gives general overview of the data with aggregated information. Summarization can scale up to different levels of abstraction and can be viewed from different angles. It is a key data-mining concept involving techniques for finding a compact description of dataset. Summarization approaches are Basically Mean, Standard Deviation, Variance, tabulating, mode, and median. These approaches are often applied for data analysis, data visualization and automated report generation.

Consider example of long distance call, Costumer data related to call can be summarize as total minutes, total spending, total calls, etc. Such important information can be presented to sales manager for analysis of his costumer and his business. The scaling and viewing from different angles can be done for this example as, calling minutes and spending can be totaled along, its period in weeks, months or years. In the same way long distance, call can be summarized as state call, state-to-state call, national, Asia calls, America calls, etc. Further summarized as Domestic and international calls.

C. Association

The Associations or Link Analysis technique are used to discover relationships between attributes and items. In these techniques, the presence of one pattern implies the presence of another pattern i.e. item is related to another in terms of cause-and-effect. This is common in establishing a form of statistical relationships among different interdependent variables of a model. For market basket analysis, Association Rules is a important technique because all possible combinations of potentially interesting product groupings can be explored [11]. Association rules are as if/then statements that help uncover relationships between seemingly unrelated data in a transaction oriented database, relational database or other information repository. In data mining, association rules are useful for analyzing and predicting customer behavior. They also play an important role in shopping basket data analysis, product clustering, catalog design and store layout. This association rules are also build by programmers use to build programs capable of machine learning.

D. Feature Extraction

Feature Extraction extract a new set of features by decomposing the original set of data. This technique describes the data with a number of features far smaller than the number of original attributes. The word feature in the technique are combination of attributes in the data that have special important characteristics of the data [12]. Feature extraction is mostly applicable to latent semantic analysis, data compression, data decomposition and projection, and pattern recognition, etc. Using feature extraction process the speed and effectiveness of supervised learning can also be improved.

For example, feature extraction is used to extract the themes/features of a document collection, where documents are represented by a set of keywords and their frequencies. Each feature is represented by a combination of keywords. The documents can then be expressed from the collection in terms of the discovered themes.

Predictive analytics

Predictive analytics is an area of statistics that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs.

Predictive Analytics Process

1. **Define Project** : Define the project outcomes, deliverable, scope of the effort, business objectives, identify the data sets that are going to be used.
2. **Data Collection** : Data mining for predictive analytics prepares data from multiple sources for analysis. This provides a complete view of customer interactions.
3. **Data Analysis** : Data Analysis is the process of inspecting, cleaning and modelling data with the objective of discovering useful information, arriving at conclusion
4. **Statistics** : Statistical Analysis enables to validate the assumptions, hypothesis and test them using standard statistical models.
5. **Modelling** : Predictive modelling provides the ability to automatically create accurate predictive models about future. There are also options to choose the best solution with multi-modal evaluation.
6. **Deployment** : Predictive model deployment provides the option to deploy the analytical results into everyday decision making process to get results, reports and output by automating the decisions based on the modelling.
7. **Model Monitoring** : Models are managed and monitored to review the model performance to ensure that it is providing the results expected.

In the current literature, concepts ”knowledge discovery”, ”data mining” and ”machine learning” are often used interchangeably. Sometimes the whole KDD process is called data mining or machine learning, or machine learning is considered as a subdiscipline of data mining. To avoid confusion, we follow a systematic division to descriptive and predictive modelling, which matches well with the classical ideas of data mining and machine learning1 . This approach has several advantages:

1. The descriptive and predictive models can often be paired as illustrated in Table 2.1. Thus, descriptive models indicate the suitability of a given predictive model and can guide the search of models (i.e. they help in the selection of the modelling paradigm and the model structure).

2. Descriptive and predictive modelling require different validation techniques. Thus the division gives guidelines, how to validate the results.

3. Descriptive and predictive models have different underlying assumptions (bias) about good models. This is reflected especially by the score functions, which guide the search.

Table : Examples of descriptive and predictive modelling paradigm pairs. The descriptive models reveal the suitability of the corresponding predictive model and guide the search.

Descriptive paradigm	**Predictive paradigm**
Correlation analysis	Linear regression
Associative rules	Probabilistic rules
Clustering	Classification
Episodes	Markov models

Selecting the modelling paradigm

Selecting the modelling paradigm has a critical role in data modelling. An unsuitable modelling paradigm can produce false, unstable or trivial models, even if we use the best available learning algorithms. Often, we have to try several modelling paradigms, and learn several models, before we find the optimal model for the given data. The number of alternative modelling paradigms can be pruned by analyzing the following factors:

ˆ Properties of data.

ˆ The inductive bias in the modelling paradigm.

ˆ The desired robustness and the representational power of the model.

The problem is to find such a modelling paradigm that the properties of the data match the inductive bias of the paradigm and the resulting models are accurate. We recall that in accuracy, we often have to make a compromise between the robustness and the representational power.

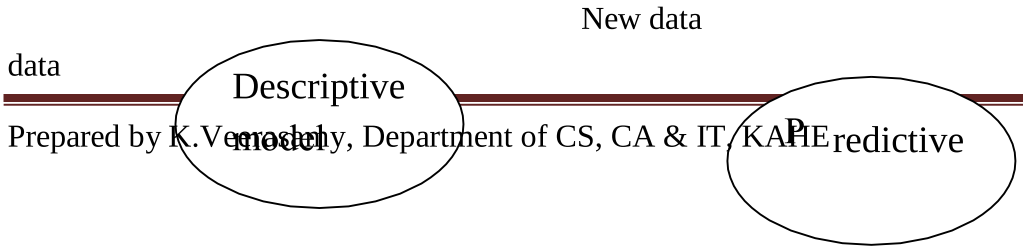
Data properties

The first question is the type and the size of the given data. Some modelling paradigms require only numeric or only categorial data, while others can combine both. Numeric data can always be discretized to categorial, but categorial data cannot be transformed to numeric. The only exception is the transformation of categorial attributes to binary values. However, this transformation increases the number of attributes significantly and the model becomes easily too complex.

In educational domain, the most critical factor of data is the size of the data relative to the model complexity. As a rule of thumb, it is often suggested that we should have at least 5-10 rows of data per each model parameter. The number of model parameters depends on the number of attributes and their domain sizes. Often, we can reduce the number of attributes and/or their domain sizes in the data preprocessing phase. We should select a modelling paradigm, which produces simple models. In practice, we recommend to take the simplest modelling paradigms like linear regression or naive Bayes as a starting point, and analyze whether they can represent the essential features in the data. Only if the data requires higher representational power (e.g. non-linear dependencies, or non-linear class boundaries), more powerful modelling paradigms should be considered.

The representational power is part of data bias in the modelling paradigm. The other assumptions in data bias should be checked as well. For example, our analysis suggests that the educational data is seldom normally distributed. One reason is the large number of outliers – exceptional and unpredictable students. If we want to use a paradigm, which assumes normality, we should first check how large is the deviation from normality and how sensitive the paradigm is to the violation of this assumption.

When the goal is predictive modelling, we recommend to analyze the data properties first by descriptive modelling. According to our view (Figure 3.1), descriptive and predictive tasks are complementary phases of the same modelling process. This view is especially useful in adaptive learning environments, in which the model can be developed through several courses. The existing data is analyzed in the descriptive phase and a desirable modelling paradigm and model family are defined. An initial model is learnt, and applied to new data in the prediction phase. In the same time we can gather new features from users, because the descriptive modelling often reveals also what data we are missing. After the course, the new data is analyzed, and the old model is updated or a new better model is constructed. As a result, our domain knowledge increases and the predictions improve in each cycle.



model

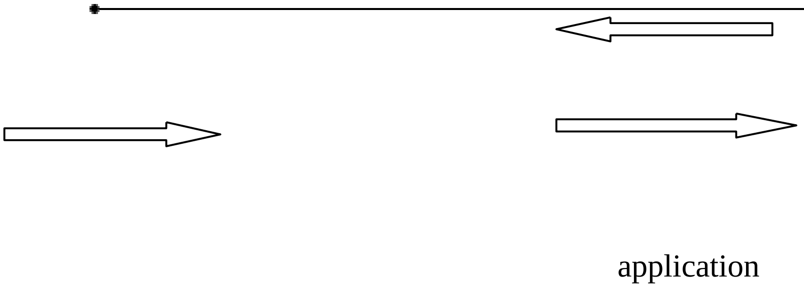


Figure: Iterative process of descriptive and predictive modelling. Descriptive modelling reveals the underlying patterns in the data and guides the selection of the most appropriate modelling paradigm and model family for the predictive modelling. When the predictive model is applied in practice, new data is gathered for new descriptive models.

Applications of Analytics

Predictive (forecasting)
Descriptive (business intelligence and data mining)

Predictive Analytics

Predictive analytics turns data into valuable, actionable information. Predictive analytics uses data to determine the probable future outcome of an event or a likelihood of a situation occurring.

Predictive analytics encompasses a variety of statistical techniques from modeling, machine learning, data mining and game theory that analyze current and historical facts to make predictions about future events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

Three basic cornerstones of predictive analytics are:

Predictive modeling
Decision Analysis and Optimization
Transaction Profiling

An example of using predictive analytics is optimizing customer relationship management systems. They can help enable an organization to analyze all customer data therefore exposing patterns that predict customer behavior.

Another example is for an organization that offers multiple products, predictive analytics can help analyze customers' spending, usage and other behavior, leading to efficient cross sales, or selling additional products to current customers. This directly leads to higher profitability per customer and stronger customer relationships.

An organization must invest in a team of experts (data scientists) and create statistical algorithms for finding and accessing relevant data. The data analytics team works with business leaders to design a strategy for using predictive information.

Descriptive Analytics

Descriptive analytics looks at data and analyzes past events for insight as to how to approach the future. Descriptive analytics looks at past performance and understands that performance by mining historical data to look for the reasons behind past success or failure. Almost all management reporting such as sales, marketing, operations, and finance, uses this type of post-mortem analysis.

Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups. Unlike predictive models that focus on predicting a single customer behavior (such as credit risk), descriptive models identify many different relationships between customers or products. Descriptive models do not rank-order customers by their likelihood of taking a particular action the way predictive models do.

Descriptive models can be used, for example, to categorize customers by their product preferences and life stage. Descriptive modeling tools can be utilized to develop further models that can simulate large number of individualized agents and make predictions.

For example, descriptive analytics examines historical electricity usage data to help plan power needs and allow electric companies to set optimal prices.

Possible Questions

1. Explain the classifications of data mining.
2. Explain the Predictive data mining.
3. Explain Descriptive data mining.
4. Explain Predictive analytics.
5. Explain the data mining paradigm.

Syllabus

Supervised and unsupervised learning techniques

Learning

Data Mining can be defined as the process that starting from apparently unstructured **data** tries to extract knowledge and/or unknown interesting patterns. During this process machine **Learning** algorithms are used.

Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) with data, without being explicitly programmed. The name **machine learning** was coined in 1959 by Arthur Samuel.

Data Mining is a cross-disciplinary field that focuses on discovering properties of data sets.

There are different approaches to discovering properties of data sets. Machine Learning is one of them. Another one is simply looking at the data sets using visualization techniques or Topological Data Analysis

On the other hand **Machine Learning** is a sub-field of data science that focuses on designing algorithms that can learn from and make predictions on the data. Machine learning includes Supervised Learning and Unsupervised Learning methods. Unsupervised methods actually start off from unlabeled data sets, so, in a way, they are directly related to finding out unknown properties in them (e.g. clusters or rules). It is clear then that machine learning can be used for data mining. However, data mining can use other techniques besides or on top of machine learning.

SUPERVISED LEARNING

DEFINITION

Supervised learning is the Data mining task of inferring a function from **labeled training data**.The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the**supervisory signal**). A **supervised learning algorithm** analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for **unseen instances**. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

UNSUPERVISED LEARNING

DEFINITION

In Data mining, the problem of **unsupervised learning** is that of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution.

LET'S LEARN SUPERVISED AND UNSUPERVISED LEARNING WITH A REAL LIFE EXAMPLE

- suppose you had a basket and it is fulled with some different kinds of fruits, your task is to arrange them as groups.

- For understanding let me clear the names of the fruits in our basket.
- We have four types of fruits. They are: **apple, banana, grape and cherry.**

SUPERVISED LEARNING :

- You already learn from your previous work about the physical characters of fruits.
- So arranging the same type of fruits at one place is easy now.
- Your previous work is called as **training data** in data mining.
- so you already learn the things from your train data, this is because of **response variable.**
- Response variable mean just a **decision variable.**
- You can observe response variable below (**FRUIT NAME**) .

NO.	SIZE	COLOR	SHAPE	FRUIT NAME
1	Big	Red	Rounded shape with a depression at the top	Apple
2	Small	Red	Heart-shaped to nearly globular	Cherry

3	Big	Green	Long curving cylinder	Banana
4	Small	Green	Round to oval,Bunch shape Cylindrical	Grape

- Suppose you have taken an new fruit from the basket then you will see the size , color and shape of that particular fruit.
- If size is Big , color is Red , shape is rounded shape with a depression at the top, you will conform the fruit name as apple and you will put in apple group.
- Likewise for other fruits also.
- Job of groping fruits was done and happy ending.
- You can observe in the table that a column was labeled as “**FRUIT NAME**” this is called as response variable.
- If you learn the thing before from training data and then applying that knowledge to the test data(for new fruit), This type of learning is called as **Supervised Learning**.
- **Classification** come under Supervised learning.

UNSUPERVISED LEARNING

- Suppose you had a basket and it is fulled with some different types fruits, your task is to arrange them as groups.

- This time you don’t know any thing about that fruits, honestly saying this is the first time you have seen them.
- so how will you arrange them.
- What will you do first???
- You will take a fruit and you will arrange them by considering physical character of that particular fruit. suppose you have considered color.
- Then you will arrange them on considering base condition as **color.**
- Then the groups will be some thing like this.
- RED COLOR GROUP: apples & cherry fruits.
- GREEN COLOR GROUP: bananas & grapes.
- so now you will take another physical character such as **size** .
- RED COLOR AND BIG SIZE: apple.
- RED COLOR AND SMALL SIZE: cherry fruits.
- GREEN COLOR AND BIG SIZE: bananas.
- GREEN COLOR AND SMALL SIZE: grapes.
- job done happy ending.
- Here you didn’t know learn any thing before ,means no train data and no response variable.
- This type of learning is know unsupervised learning.
- clustering comes under unsupervised learning.

Supervised and Unsupervised Machine Learning Algorithms

Supervised Machine Learning

The majority of practical machine learning uses supervised learning.

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

Supervised learning problems can be further grouped into regression and classification problems.

- **Classification**: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
- **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively.

Some popular examples of supervised machine learning algorithms are:

- Linear regression for regression problems.
- Random forest for classification and regression problems.
- Support vector machines for classification problems.

Unsupervised Machine Learning

Unsupervised learning is where you only have input data (X) and no corresponding output variables.

The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

Unsupervised learning problems can be further grouped into clustering and association problems.

- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Some popular examples of unsupervised learning algorithms are:

- k-means for clustering problems.
- Apriori algorithm for association rule learning problems.

Semi-Supervised Machine Learning

Problems where you have a large amount of input data (X) and only some of the data is labeled (Y) are called semi-supervised learning problems.

These problems sit in between both supervised and unsupervised learning.

A good example is a photo archive where only some of the images are labeled, (e.g. dog, cat, person) and the majority are unlabeled.

Many real world machine learning problems fall into this area. This is because it can be expensive or time-consuming to label data as it may require access to domain experts. Whereas unlabeled data is cheap and easy to collect and store.

You can use unsupervised learning techniques to discover and learn the structure in the input variables.

You can also use supervised learning techniques to make best guess predictions for the unlabeled data, feed that data back into the supervised learning algorithm as training data and use the model to make predictions on new unseen data.

Summary

In this post you learned the difference between supervised, unsupervised and semi-supervised learning. You now know that:

- **Supervised**: All data is labeled and the algorithms learn to predict the output from the input data.
- **Unsupervised**: All data is unlabeled and the algorithms learn to inherent structure from the input data.
- **Semi-supervised**: Some data is labeled but most of it is unlabeled and a mixture of supervised and unsupervised techniques can be used.

Do you have any questions about supervised, unsupervised or semi-supervised learning? Leave a comment and ask your question and I will do my best to answer it.

When To Use Supervised And Unsupervised Data Mining

- Data mining techniques come in two main forms: supervised (also known as predictive or directed) and unsupervised (also known as descriptive or undirected). Both categories encompass functions capable of finding different hidden patterns in large data sets.
- Data analytics tools are placing more emphasis on self service, it’s still useful to know which data mining operation is appropriate for your needs before you begin a data mining operation.

Supervised Data Mining

- Supervised data mining techniques are appropriate when you have a specific target value you'd like to predict about your data. The targets can have two or more possible outcomes, or even be a continuous numeric value (more on that later).
- To use these methods, you ideally have a subset of data points for which this target value is already known. You use that data to build a model of what a typical data point looks like when it has one of the various target values. You then apply that model to data for which that target value is currently unknown. The algorithm identifies the "*new*" data points that match the model of each target value.

Classification

- As a supervised data mining method, classification begins with the method.
- For example:Imagine you're a credit card company and you want to know which customers are likely to default on their payments in the next few years.
- You use the data on customers who have and have not defaulted for extended periods of time as build data (or training data) to generate a classification model. You then run that model on the customers you're curious about. The algorithms will look for customers whose attributes match the attribute patterns of previous defaulters/non-defaulters, and categorize

them according to which group they most closely match. You can then use these groupings as indicators of which customers are most likely to default.

- Similarly, a classification model can have more than two possible values in the target attribute. The values could be anything from the shirt colors they're most likely to buy, the promotional methods they'll respond to (mail, email, phone), or whether or not they'll use a coupon.

Regression

- Regression is similar to classification except that the targeted attribute's values are numeric, rather than categorical. The order or magnitude of the value is significant in some way.
- To reuse the credit card example, if you wanted to know what threshold of debt new customers are likely to accumulate on their credit card, you would use a regression model.
- Simply supply data from current and past customers with their maximum previous debt level as the target value, and a regression model will be built on that training data. Once run on the new customers, the regression model will match attribute values with predicted maximum debt levels and assign the predictions to each customer accordingly.
- This could be used to predict the age of customers with demographic and purchasing data, or to predict the frequency of insurance claims.

Anomaly Detection

- Anomaly detection identifies data points atypical of a given distribution. In other words, it finds the outliers. Though simpler data analysis techniques than full-scale data mining can identify outliers, data mining anomaly detection techniques identify much more subtle attribute patterns and the data points that fail to conform to those patterns.
- Most examples of anomaly detection uses involve fraud detection, such as for insurance or credit card companies.

Unsupervised Data Mining

- Unsupervised data mining does not focus on predetermined attributes, nor does it predict a target value. Rather, unsupervised data mining finds hidden structure and relation among data.

Clustering

- The most open-ended data-mining technique, clustering algorithms, finds and groups data points with natural similarities.
- This is used when there are no obvious natural groupings, in which case the data may be difficult to explore. Clustering the data can reveal groups and categories you were previously unaware of. These new groups may be fit for further data mining operations from which you may discover new correlations.

Association

- Frequently used for market basket analysis, association models identify common co-occurrences among a list of possible events. Market basket analysis is examining all items available in a particular medium, such as the products on store shelves or in a catalogue, and finding the products that are commonly sold together.
- This operation produces association rules. Such a rule could be a statement declaring "*80 percent of people who buy charcoal, hamburger meat, and buns also buy sliced cheese,*" or, in a less "*market basket*" style example, "*90 percent of Detroit citizens who root for the Tigers, the Lions, and the Pistons also favor the Red Wings over other hockey teams*."
- Such rules can be used to personalize the customer experience to promote certain events or actions. This can be accomplished by organizing store shelves with associated items nearby, or by tracking customer movements through a website in real time to present them with relevant product links.

Feature Extraction

- Feature extraction creates new features based on attributes of your data. These new features describe a combination of significant attribute value patterns in your data.
- If violence, heroism, and fast cars were attributes of a movie, then the feature may be "*action*," akin to a genre or a theme. This concept can be used

to extract the themes of a document based on the frequencies of certain key words.

- Representing data points by their features can help compress the data (trading dozens of attributes for one feature), make predictions (data with this feature often has these attributes as well), and recognize patterns. Additionally, features can be used as new attributes, which can improve the efficiency and accuracy of supervised learning techniques (classification, regression, anomaly detection, etc.).
- Knowing your goals and the appropriate techniques to achieve them can help your data mining operations run smoothly and effectively. Different data is appropriate for different insight and understanding what you're asking from your data analystsexpedites the process for everyone.

Algorithms used in unsupervised learning vary, including:

1. **Clustering**. k-means. mixture models. hierarchical **clustering**,
2. Anomaly detection.
3. Neural Networks.
4. Approaches for **learning** latent variable models such as. Expectation–maximization algorithm (EM) **Method** of moments.

Algorithms used in supervised learning vary, including:

The most widely used learning algorithms are:

- Support Vector Machine
  linear regression
- logistic regression
- naive Bayes
- linear discriminant analysis
- decision trees
- k-nearest neighbor algorithm
- Neural Networks (Multilayer perceptron)

Possible Questions

1. Explain Supervised learning algorithm.
2. Explain Unsupervised learning algorithm.
3. Explain supervised and unsupervised with real life example.
4. Explain machine learning algorithm.

Process of knowledge discovery in databases, pre-processing methods

Knowledge Discovery in Database:

The term **Knowledge Discovery in Databases** or KDD for short, refers to the broad **process** of finding **knowledge** in data, and emphasizes the "high-level" application of particular data mining methods. ... The unifying goal of the KDD **process** is to extract **knowledge** from data in the context of large data bases.

KNOWLEDGE DISCOVERY DATABASE

Data mining is the core part of the knowledge discovery process. In this, process may consist of the following steps Data selection, Data cleaning, Data transformation, pattern searching (data mining), finding presentation, finding interpretation and finding evaluation. The data mining and KDD often used interchangeably because Data mining is the key part of KDD process. The term Knowledge Discovery in Databases or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods.

It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large data bases. It does this by using data mining methods(algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required pre-processing, sub sampling, and transformations of that database.

The KDD Process

The knowledge discovery process is iterative and interactive, consisting of nine steps [3]. Note that the process is iterative at each step, meaning that moving back to previous steps may be required .So it is required to understand the process and the different needs and possibilities in each step.

A typical knowledge discovery process is shown in figure 1, and the process is elaborated in each step.

- Developing an understanding of the application domain
- Selecting and creating a data set on which discovery will be performed.
- Preprocessing and cleansing.
- Choosing the appropriate Data Mining task.
- Choosing the Data Mining algorithm.
- Employing the Data Mining algorithm.
- Evaluation.
- Using the discovered knowledge.

The terms knowledge discovery and data mining are distinct. KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step. Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

DATA MINING

Data mining is the process of discovering actionable information from large sets of data [4]. Data mining uses mathematical analysis to derive patterns and trends that exist in data. These patterns and trends can be collected and defined as a data mining model. Mining models can be applied to specific scenarios, such as

Forecasting: Estimating sales, predicting server loads or server downtime

Risk and probability: Choosing the best customers for targeted mailings, determining the probable break-even point for risk scenarios, assigning probabilities to diagnoses or other outcomes

Recommendations: Determining which products are likely to be sold together, generating recommendations

Finding sequences: Analyzing customer selections in a shopping cart, predicting next likely events

Grouping: Separating customers or events into cluster of related items, analyzing and predicting affinities.

Building a mining model is part of a larger process that includes everything from asking questions about the data and creating a model to answer those questions, to deploying the model into a working environment [6]. This process can be defined by using the following basic steps:

- Defining the Problem
- Preparing Data
- Exploring Data
- Building Models
- Exploring and Validating Models

DATA MINING TECHNIQUES

There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns[2].

Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction.

Types of association rules:

Different types of association rules based on

- Types of values handled
- Boolean association rules
- Quantitative association rules
- Levels of abstraction involved
- Single-level association rules
- Multilevel association rules
- Dimensions of data involved

- Single-dimensional association rules
- Multidimensional association rules

Classification

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups.

Classification Techniques

- Regression
- Distance
- Decision Trees
- Rules
- Neural Networks

Clustering

Clustering is the process of organizing objects into groups whose members are similar in some way. The cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

Prediction

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables.

Sequential Patterns

Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period[5].The uncover patterns are used for further business analysis to recognize relationships among data. A sequence is an ordered list of events, denoted < $e_1e_2\ldots e_L$>.

ASPECTS OF DATAMINING

Data Integration: First of all the data are collected and integrated from all the different sources.

Data Selection: We may not all the data we have collected in the first step. In this step we select only those data which we think useful for data mining.

Data Cleaning: The data we have collected are not clean and may contain errors, missing values, noisy or inconsistent data. Therefore we need to apply different techniques to get rid of such anomalies.

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.

Data Mining: Now we are ready to apply data mining techniques on the data to discover the interesting patterns. The techniques like clustering and association analysis are among the many different techniques used for data mining.

Pattern Evaluation and Knowledge Presentation: This step involves visualization, transformation, removing redundant patterns etc from the patterns we generated.

Decisions / Use of Discovered Knowledge: This step helps user to make use of the knowledge acquired to take better decisions.

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

Issues in data mining

A number of issues that need to be addressed by any serious data mining package

- Uncertainty Handling
- Dealing with Missing Values
- Dealing with Noisy data
- Efficiency of algorithms
- Constraining Knowledge Discovered to only Useful or Interesting Knowledge
- Incorporating Domain Knowledge
- Size and Complexity of Data

- Data Selection
- Understandability of Discovered Knowledge: Consistency between Data and Discovered Knowledge

Data mining consists of five major elements

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

DIFFERENT LEVELS OF ANALYSIS

Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

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

Decision trees: Tree-shaped structures that represent sets of decisions. So these decisions generate rules for the classification of a dataset. The Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). The CART and CHAID are decision tree techniques used for classification of a dataset.

Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics t relationships.

Data Preprocess an Overview:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications (like neural networks).

Data Preprocessing

Data goes through a series of steps during preprocessing:

- Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
- Data Integration: Data with different representations are put together and conflicts within the data are resolved.
- Data Transformation: Data is normalized, aggregated and generalized.
- Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.

- Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

Data cleaning

1. Fill in missing values (attribute or class value):
   - Ignore the tuple: usually done when class label is missing.
   - Use the attribute mean (or majority nominal value) to fill in the missing value.
   - Use the attribute mean (or majority nominal value) for all samples belonging to the same class.
   - Predict the missing value by using a learning algorithm: consider the attribute with the missing value as a dependent (class) variable and run a learning algorithm (usually Bayes or decision tree) to predict the missing value.

2. Identify outliers and smooth out noisy data:
   - Binning

        - Sort the attribute values and partition them into bins (see "Unsupervised discretization" below);
        - Then smooth by bin means, bin median, or bin boundaries.
    - Clustering: group values in clusters and then detect and remove outliers (automatic or manual)
    - Regression: smooth by fitting the data into regression functions.

3. Correct inconsistent data: use domain knowledge or expert decision.

Data transformation

1. Normalization:
    - Scaling attribute values to fall within a specified range.
        - Example: to transform V in [min, max] to V' in [0,1], apply V'=(V-Min)/(Max-Min)
    - Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers): V'=(V-Mean)/StDev

2. Aggregation: moving up in the concept hierarchy on numeric attributes.
3. Generalization: moving up in the concept hierarchy on nominal attributes.
4. Attribute construction: replacing or adding new attributes inferred by existing attributes.

Data reduction

1. Reducing the number of attributes
   - Data cube aggregation: applying roll-up, slice or dice operations.
   - Removing irrelevant attributes: attribute selection (filtering and wrapper methods), searching the attribute space (see Lecture 5: Attribute-oriented analysis).
   - Principle component analysis (numeric attributes only): searching for a lower dimensional space that can best represent the data..
2. Reducing the number of attribute values
   - Binning (histograms): reducing the number of attributes by grouping them into intervals (bins).
   - Clustering: grouping values in clusters.

  - Aggregation or generalization

3. Reducing the number of tuples
  - Sampling

Discretization and generating concept hierarchies

1. Unsupervised discretization - class variable is not used.
  - Equal-interval (equiwidth) binning: split the whole range of numbers in intervals with equal size.
  - Equal-frequency (equidepth) binning: use intervals containing equal number of values. 2. Supervised discretization - uses the values of the class variable.

- Using class boundaries. Three steps:
    - Sort values.
    - Place breakpoints between values belonging to different classes.
    - If too many intervals, merge intervals with equal or similar class distributions.
- Entropy (information)-based discretization. Example:
    - Information in a class distribution:
        - Denote a set of five values occurring in tuples belonging to two classes (+ and -) as [+,+,+,-,-]
        - That is, the first 3 belong to "+" tuples and the last 2 - to "-" tuples
        - Then, Info([+,+,+,-,-]) = -(3/5)*log(3/5)-(2/5)*log(2/5) (logs are base 2)
        - 3/5 and 2/5 are relative frequencies (probabilities)
        - Ignoring the order of the values, we can use the following notation: [3,2] meaning 3 values from one class and 2 - from the other.
        - Then, Info([3,2]) = -(3/5)*log(3/5)-(2/5)*log(2/5)
    - Information in a split (2/5 and 3/5 are weight coefficients):
        - Info([+,+],[+,-,-]) = (2/5)*Info([+,+]) + (3/5)*Info([+,-,-])
        - Or, Info([2,0],[1,2]) = (2/5)*Info([2,0]) + (3/5)*Info([1,2])

- Method:
    - Sort the values;
    - Calculate information in all possible splits;
    - Choose the split that minimizes information;
    - Do not include breakpoints between values belonging to the same class (this will increase information);
    - Apply the same to the resulting intervals until some stopping criterion is satisfied.

3. Generating concept hierarchies: recursively applying partitioning or discretization methods.

Data Discretization and Concept Hierarchy Generation

Data Discretization techniques can be used to divide the range of continuous attribute into intervals.Numerous continuous attribute values are replaced by small interval labels.

This leads to a concise, easy-to-use, knowledge-level representation of mining results.

Top-down discretization

If the process starts by first finding one or a few points (called split points or cut points) to split the entire attribute range, and then repeats this recursively on the resulting intervals, then it is called top-down discretization or splitting.

Bottom-up discretization

If the process starts by considering all of the continuous values as potential split-points, removes some by merging neighborhood values to form intervals, then it is called bottom-up discretization or merging.

Discretization can be performed rapidly on an attribute to provide a hierarchical partitioning of the attribute values, known as a **concept hierarchy**.

Concept hierarchies

Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts with higher-level concepts.

In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides users with the flexibility to view data from different perspectives.

Data mining on a reduced data set means fewer input/output operations and is more efficient than mining on a larger data set.

Because of these benefits, discretization techniques and concept hierarchies are typically applied before data mining, rather than during mining.

Discretization and Concept Hierarchy Generation for Numerical Data

Typical methods

1 Binning

Binning is a top-down splitting technique based on a specified number of bins.Binning is an unsupervised discretization technique.

2 Histogram Analysis

Because histogram analysis does not use class information so it is an unsupervised discretization technique.Histograms partition the values for an attribute into disjoint ranges called buckets.

3 Cluster Analysis

Cluster analysis is a popular data discretization method.A clustering algorithm can be applied to discrete a numerical attribute of A by partitioning the values of A into clusters or groups.

Each initial cluster or partition may be further decomposed into several subcultures, forming a lower level of the hierarchy.

Segmentation by natural partitioning:

Breaking up annual salaries in the range of into ranges like ($50,000-$100,000) are often more desirable than ranges like ($51, 263, 89-$60,765.3) arrived at by cluster analysis. The 3-4-5 rule can be used to segment numeric data into relatively uniform “natural” intervals. In general the rule partitions a give range of data into 3,4,or 5 equinity intervals, recursively level by level based on value range at the most significant digit. The rule can be recursively applied to each interval creating a concept hierarchy for the given numeric attribute.

Discretization and Concept Hierarchy Generation for Categorical Data:

Categorical data are discrete data. Categorical attributes have finite number of distinct values, with no ordering among the values, examples include geographic location, item type and job category. There are several methods for generation of concept hierarchies for categorical data.

Specification of a partial ordering of attributes explicitly at the schema level by experts:

Concept hierarchies for categorical attributes or dimensions typically involve a group of attributes. A user or an expert can easily define concept hierarchy by specifying a partial or total ordering of the attributes at a schema level. A hierarchy can be defined at the schema level such as street < city < province <state < country.

Specification of a portion of a hierarchy by explicit data grouping:

This is identically a manual definition of a portion of a concept hierarchy. In a large database, is unrealistic to define an entire concept hierarchy by explicit value enumeration. However, it is realistic to specify explicit groupings for a small portion of the intermediate-level data.

Specification of a set of attributes but not their partial ordering:

A user may specify a set of attributes forming a concept hierarchy, but omit to specify their partial ordering. The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept hierarchy.

Specification of only of partial set of attributes:

Sometimes a user can be sloppy when defining a hierarchy, or may have only a vague idea about what should be included in a hierarchy. Consequently the user may have included only a small subset of the relevant attributes for the location, the user may have only specified street and city. To handle such partially specified hierarchies, it is important to embed data semantics in the database schema so that attributes with tight semantic connections can be pinned together.

Possible Questions

1. Explain KDD process.
2. Explain about preprocessing steps.
3. Explain about Data cleaning.
4. Explain about different levels of analysis.
5. Explain the issues in data mining.

Data Mining Techniques: Association Rule Mining, classification and regression techniques, clustering

Classification:

Classification Problem

- Given a database D={t1,t2,…,tn} and a set of classes C={C1,…,Cm}, the *Classification Problem* is to define a mapping f:DgC where each ti is assigned to one class.
- Actually divides D into *equivalence classes*.
- *Prediction*is similar, but may be viewed as having infinite number of classes.

Classification Examples

- Teachers classify students' grades as A, B, C, D, or F.
- Identify mushrooms as poisonous or edible.
- Predict when a river will flood.
- Identify individuals with credit risks.
- Speech recognition
- Pattern recognition

Classification Ex: Grading

If x >= 90 then grade =A.
If 80<=x<90 then grade =B.
If 70<=x<80 then grade =C.
If 60<=x<70 then grade =D.
If x<50 then grade =F.

Classification Techniques

Approach:

1. Create specific model by evaluating training data (or using domain experts' knowledge).

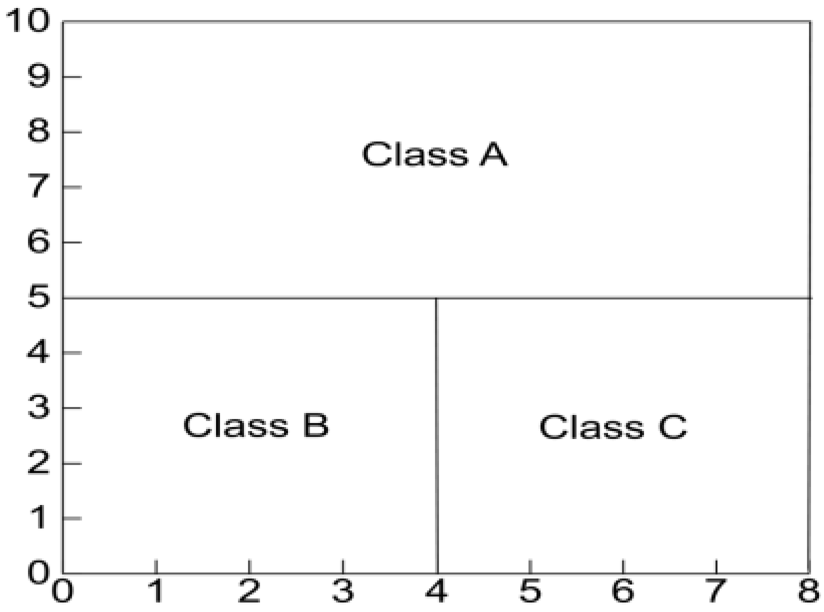
2. Apply model developed to new data.

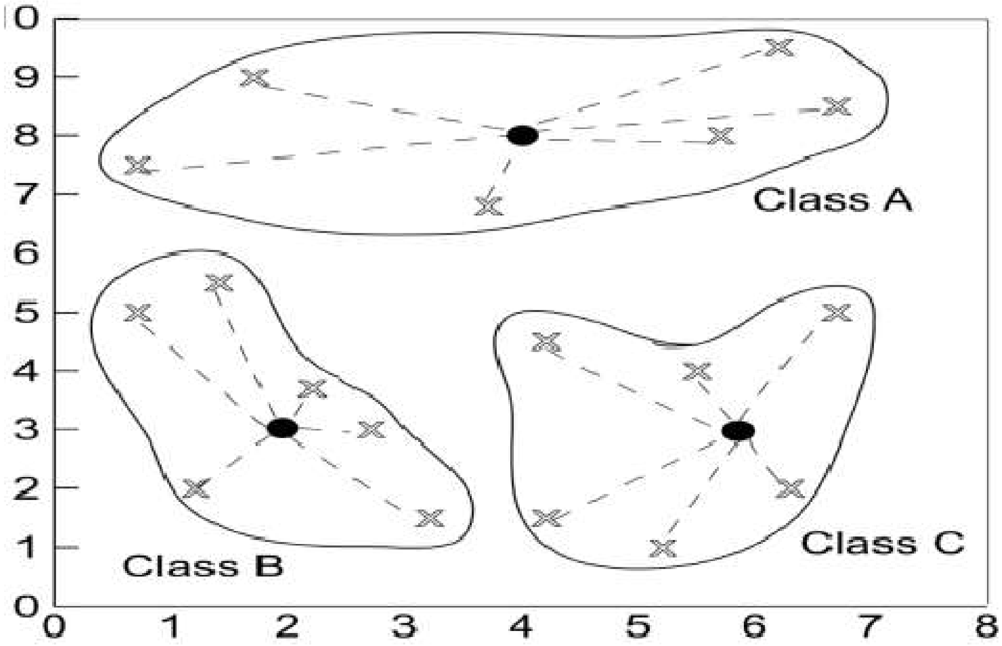
Classes must be predefined

Most common techniques use DTs, NNs based on distances or statistical methods.

Defining Classes

Partitioned based Distance based

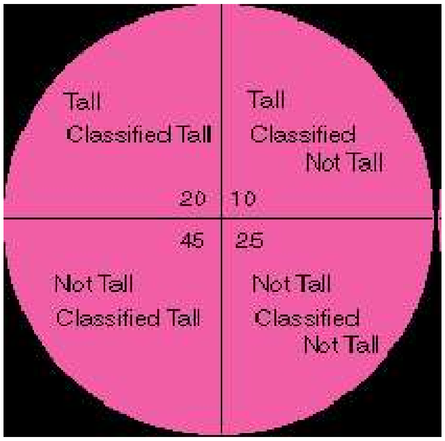




Issues in Classification

- Missing Data
- Ignore
- Replace with assumed value
- Measuring Performance
- Classification accuracy on test data
- Confusion matrix
- OC Curve

Classification Performance



ROC Curve

- Shows the relationship between false positives and true positives
- Information retrieval – percentage of retrieved that are not relevant (fallout)
- Communication – false alarm rates

Regression

- Assume data fits a predefined function
- Determine best values for *regression coefficients* c0,c1,…,cn.
- Assume an error: y = c0+c1x1+…+cnxn+e

$$y_i = c_0 + c_1 x_{1i} + c_i, i = 1, \ldots, k$$

$$L = \sum_{i=1}^{k} c_i^2 = \sum_{i=1}^{k} (y_i - c_0 - c_1 x_{1i})^2$$

Classification Using Regression

- *Division:* Use regression function to divide area into regions.
- *Prediction*: Use regression function to predict a class membership function. Input includes desired class.

Classification Using Distance

- Place items in class to which they are “closest”.
- Must determine distance between an item and a class.
- Classes represented by

– *Centroid:* Central value.

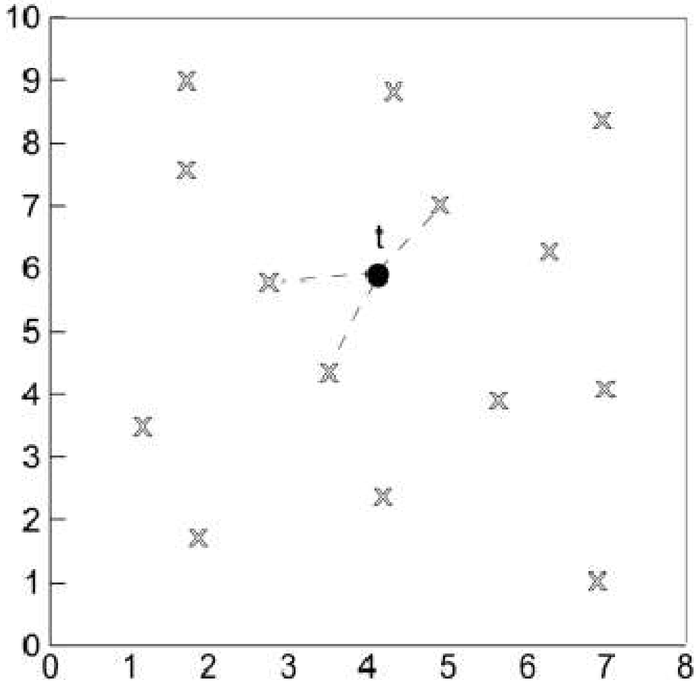
– *Medoid:* Representative point.

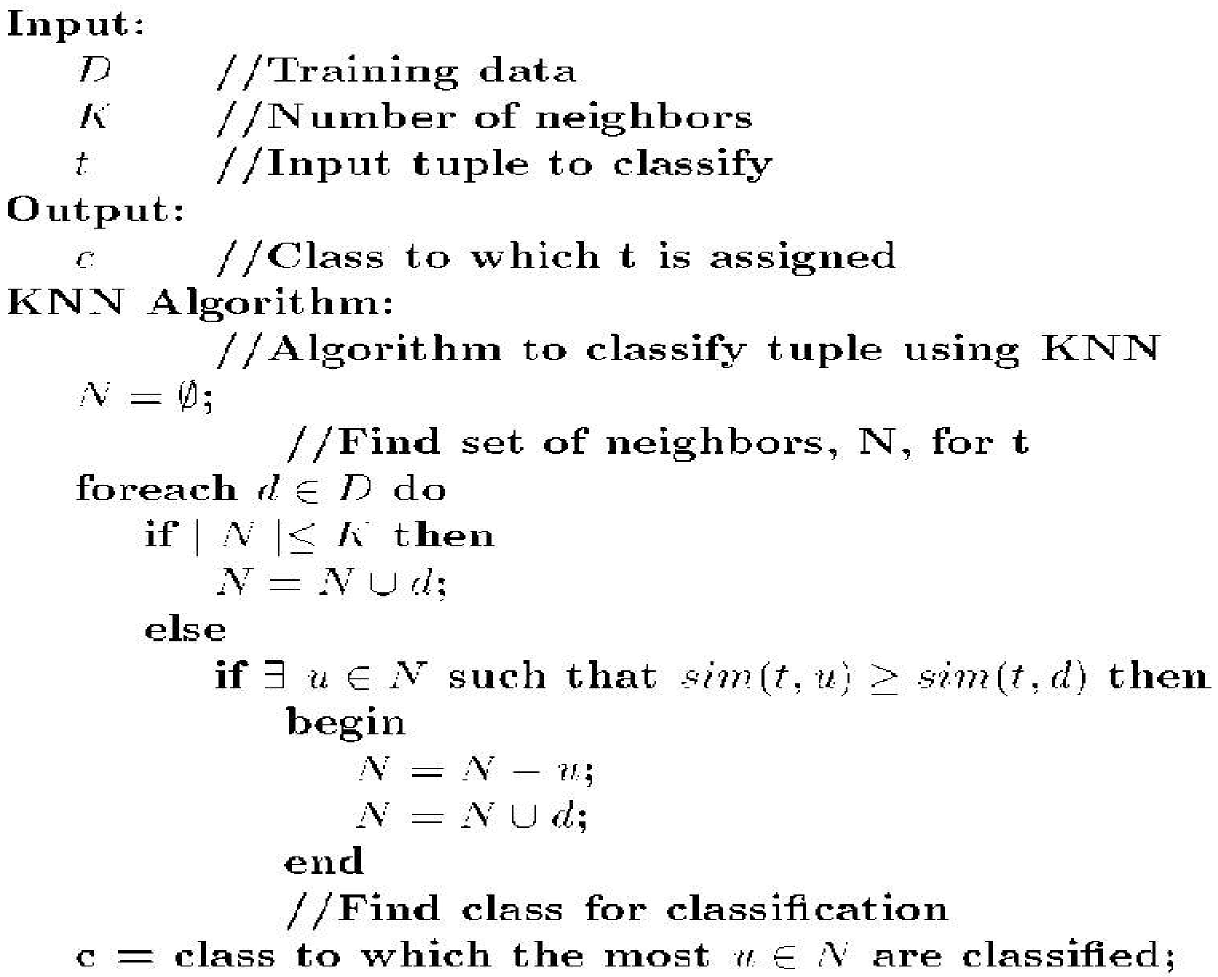
– Individual points

Algorithm: KNN

K Nearest Neighbor (KNN):

- Training set includes classes. Each member with a label of a class. Training data = model.
- Compare new item with all members in training set for distance. Examine K items nearest item to be considered further.
- New item placed in class with the most number of nearest items (among K) belongs.
- O(q) for each tuple to be classified. (Here q is the size of the training set.)





```
Input:
    $D$        //Training data
    $K$        //Number of neighbors
    $t$        //Input tuple to classify
Output:
    $c$        //Class to which t is assigned
KNN Algorithm:
               //Algorithm to classify tuple using KNN
    $N = \emptyset$;
                   //Find set of neighbors, N, for t
    foreach $d \in D$ do
        if $| N | \leq K$ then
            $N = N \cup d$;
        else
            if $\exists\ u \in N$ such that $sim(t, u) \geq sim(t, d)$ then
                begin
                    $N = N - u$;
                    $N = N \cup d$;
                end
                //Find class for classification
    c = class to which the most $u \in N$ are classified;
```

\

Classification Using Decision Trees

- *Partitioning based:* Divide search space into rectangular regions.

- Tuple placed into class based on the region within which it falls.

 DT approaches differ in how the tree is built: *DT Induction*

 Internal nodes associated with attribute and arcs with values for that attribute.

 Algorithms: ID3, C4.5, CART

Given:
– D = {t1, …, tn} where ti=<ti1, …, tih>

– Database schema contains {A1, A2, …, Ah}

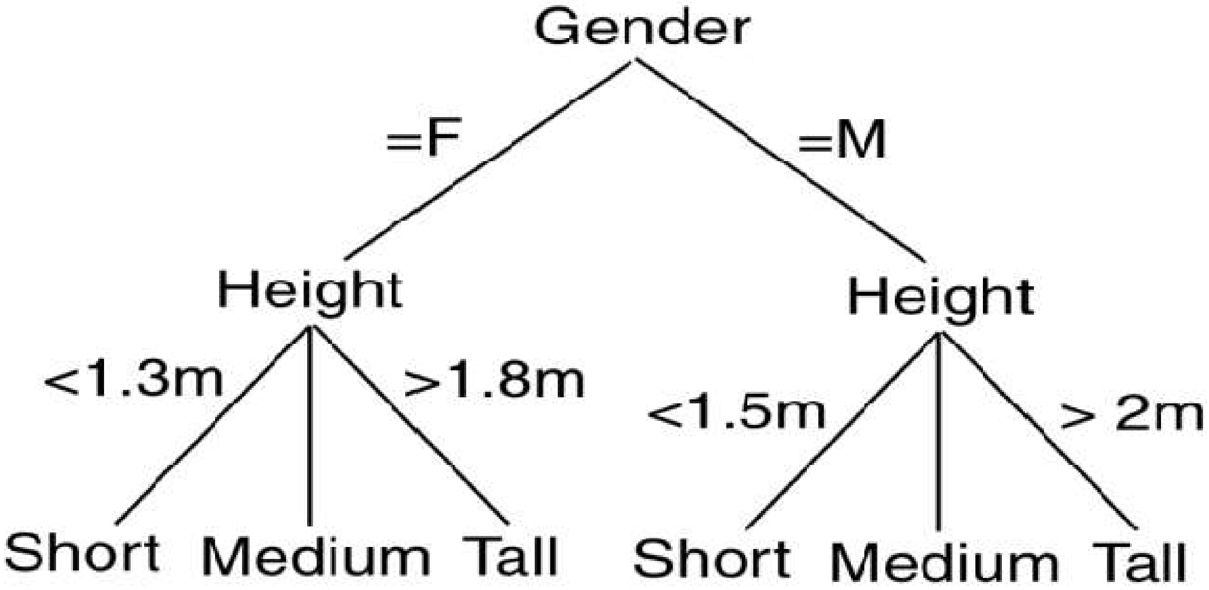
– Classes C={C1, …., Cm}

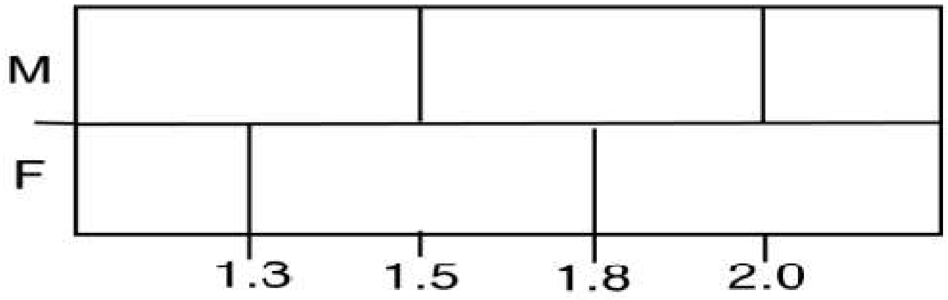
Decision or Classification Tree is a tree associated with D such that
– Each internal node is labeled with attribute, Ai

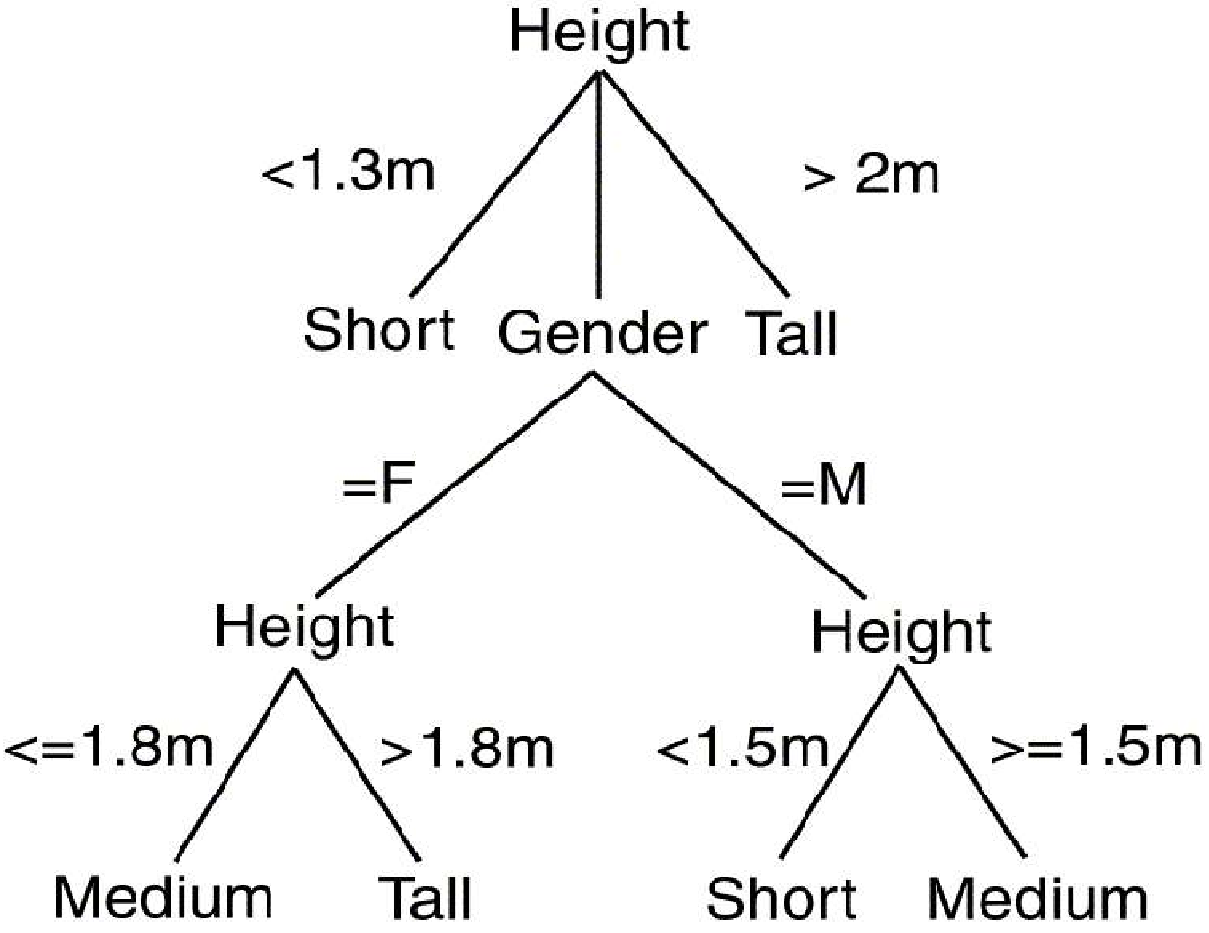
– Each arc is labeled with predicate which can be applied to attribute at parent

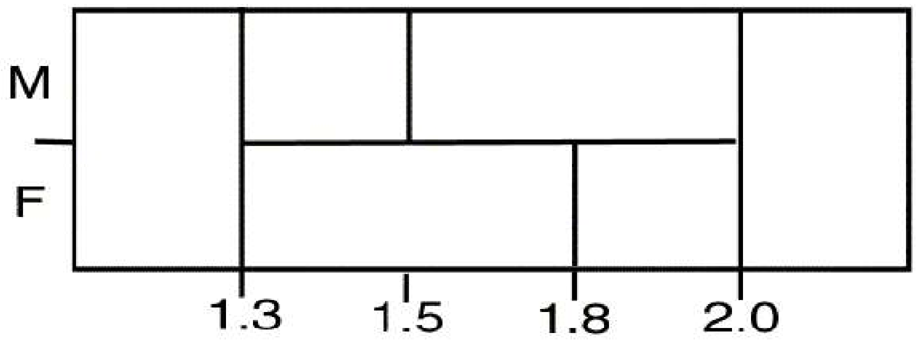
– Each leaf node is labeled with a class, Cj

Comparing DTs


Gender
=F
=M
Height
Height
<1.3m
>1.8m
<1.5m
> 2m
Short
Medium
Tall
Short
Medium
Tall

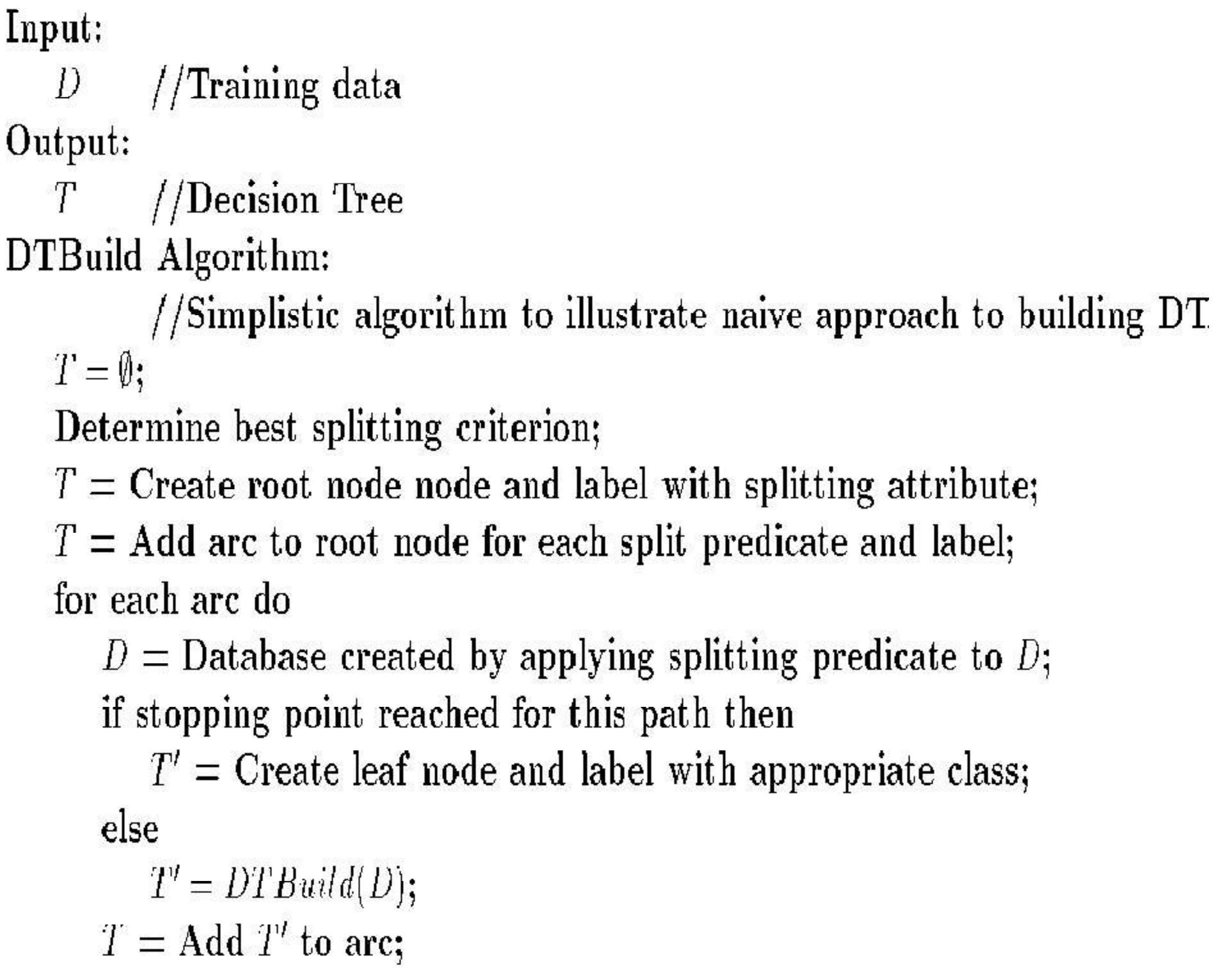

M
F
1.3
1.5
1.8
2.0


Height
<1.3m
> 2m
Short
Gender
Tall
=F
=M
Height
Height
<=1.8m
>1.8m
<1.5m
>=1.5m
Medium
Tall
Short
Medium


M
F
1.3
1.5
1.8
2.0

[Type text]

Algorithm



```
Input:
    $D$      //Training data
Output:
    $T$      //Decision Tree
DTBuild Algorithm:
             //Simplistic algorithm to illustrate naive approach to building DT
    $T = \emptyset$;
    Determine best splitting criterion;
    $T$ = Create root node node and label with splitting attribute;
    $T$ = Add arc to root node for each split predicate and label;
    for each arc do
        $D$ = Database created by applying splitting predicate to $D$;
        if stopping point reached for this path then
            $T'$ = Create leaf node and label with appropriate class;
        else
            $T' = DTBuild(D)$;
        $T$ = Add $T'$ to arc;
```

- **ID3**
- Creates tree using information theory concepts and tries to reduce expected number of comparison..
- ID3 chooses split attribute with the highest information gain:

$$Gain(D,S) = H(D) - \sum_{i=1}^{s} P(D_i)H(D_i)$$

Example

- Starting state entropy:

- 4/15 log(15/4) + 8/15 log(15/8) + 3/15 log(15/3) = 0.4384
- Gain using gender:
  - Female: 3/9 log(9/3)+6/9 log(9/6)=0.2764
  - Male: 1/6 (log 6/1) + 2/6 log(6/2) + 3/6 log(6/3) = 0.4392
  - Weighted sum: (9/15)(0.2764) + (6/15)(0.4392) = 0.34152
  - Gain: 0.4384 – 0.34152 = 0.09688
- Gain using height:

- 0.4384 – (2/15)(0.301) = 0.3983
- Choose height as first splitting attribute

C4.5

– Missing Data

– Continuous Data

– Pruning

– Rules

– GainRatio:

$$GainRatio(D,S) = \frac{Gain(D,S)}{H(\frac{|D_1|}{|D|}, ..., \frac{|D_s|}{|D|})}$$

CART

Create Binary Tree

□ Uses entropy

□ Formula to choose split point, s, for node t:

$$\Phi(s/t) = 2P_L P_R \sum_{j=1}^{m} \mid P(C_j \mid t_L) - P(C_j \mid t_R) \mid$$

PL,PR probability that a tuple in the training set will be on the left or right side of the tree.

CART Example

□ At the start, there are six choices for split point (right branch on equality):

– P(Gender)=2(6/15)(9/15)(2/15 + 4/15 + 3/15)=0.224

– P(1.6) = 0

– P(1.7) = 2(2/15)(13/15)(0 + 8/15 + 3/15) = 0.169

P(1.8) = 2(5/15)(10/15)(4/15 + 6/15 + 3/15) = 0.385

– P(1.9) = 2(9/15)(6/15)(4/15 + 2/15 + 3/15) = 0.256

– P(2.0) = 2(12/15)(3/15)(4/15 + 8/15 + 3/15) = 0.32

□ Split at 1.8

Classification Using Neural Networks

□ Typical NN structure for classification:

– One output node per class

– Output value is class membership function value

□ Supervised learning

□ For each tuple in training set, propagate it through NN. Adjust weights on edges to improve future classification.

□ Algorithms: Propagation, Backpropagation, Gradient Descent

NN Issues

□ Number of source nodes

□ Number of hidden layers

□ Training data

□ Number of sinks

□ Interconnections

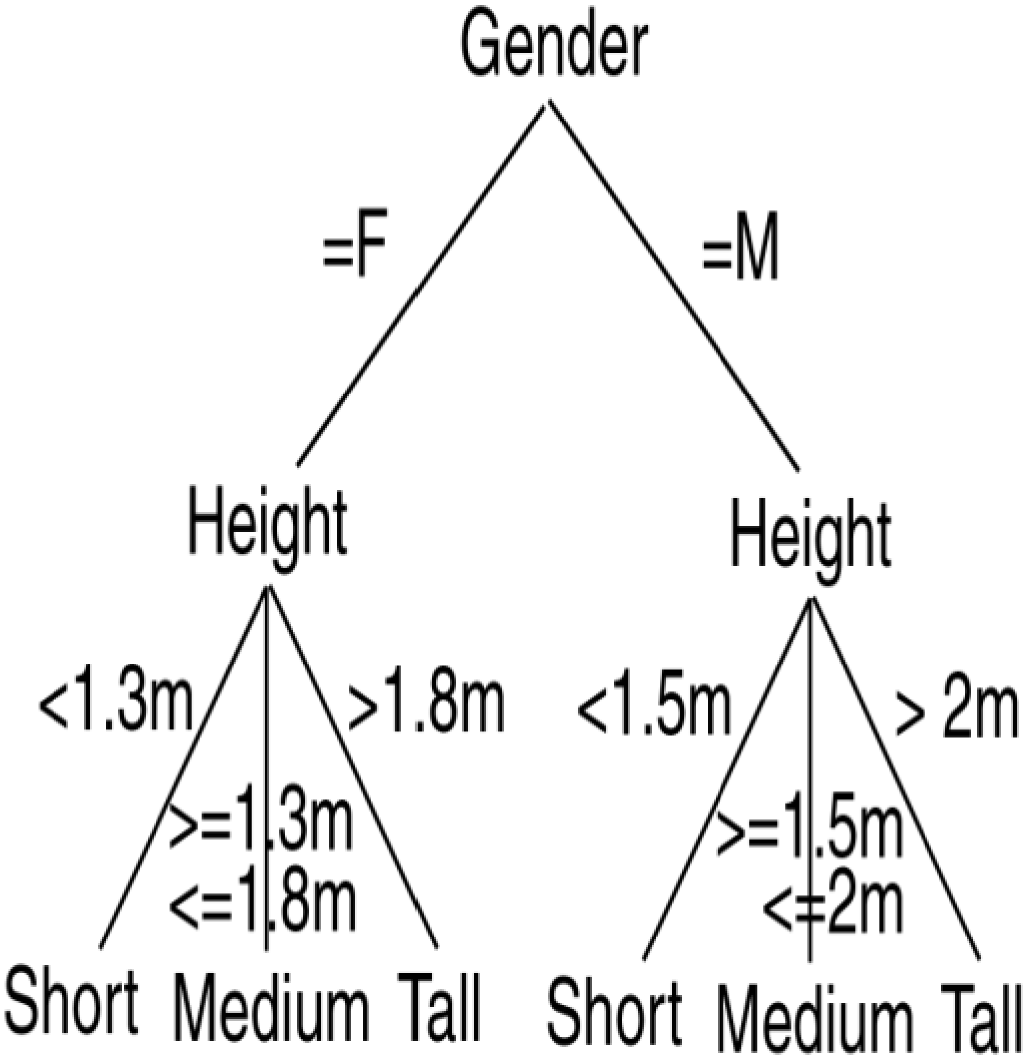
□ Weights

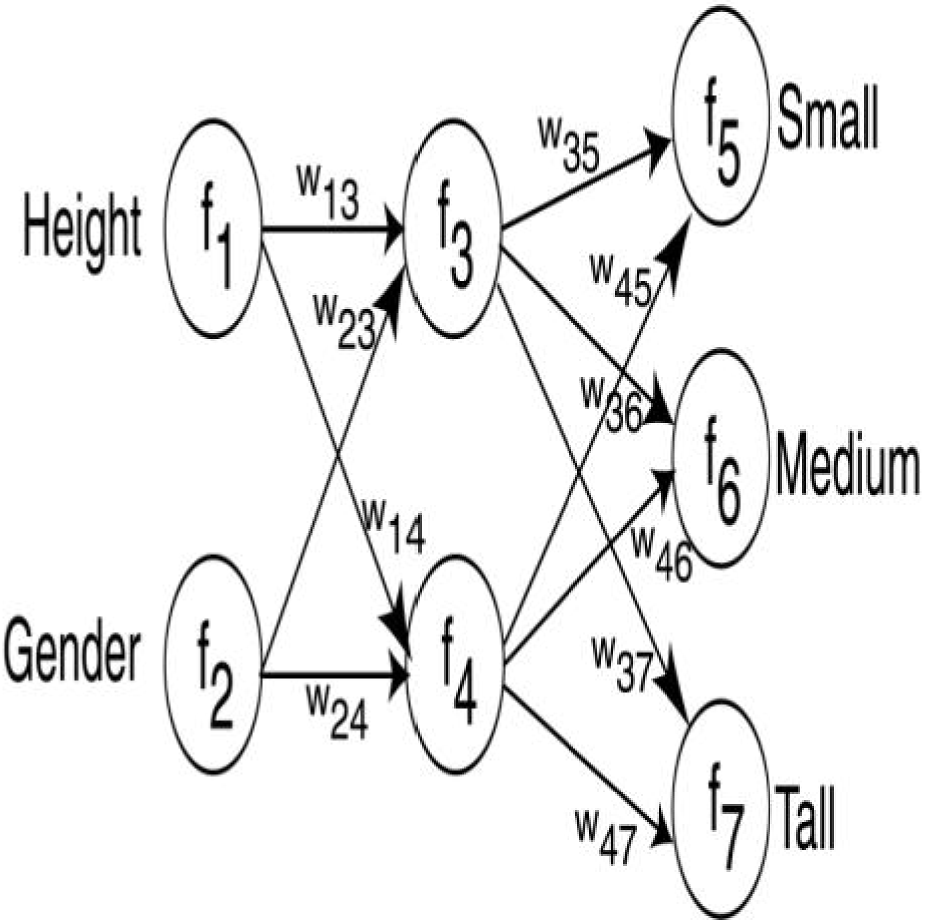
 Activation Functions

 Learning Technique

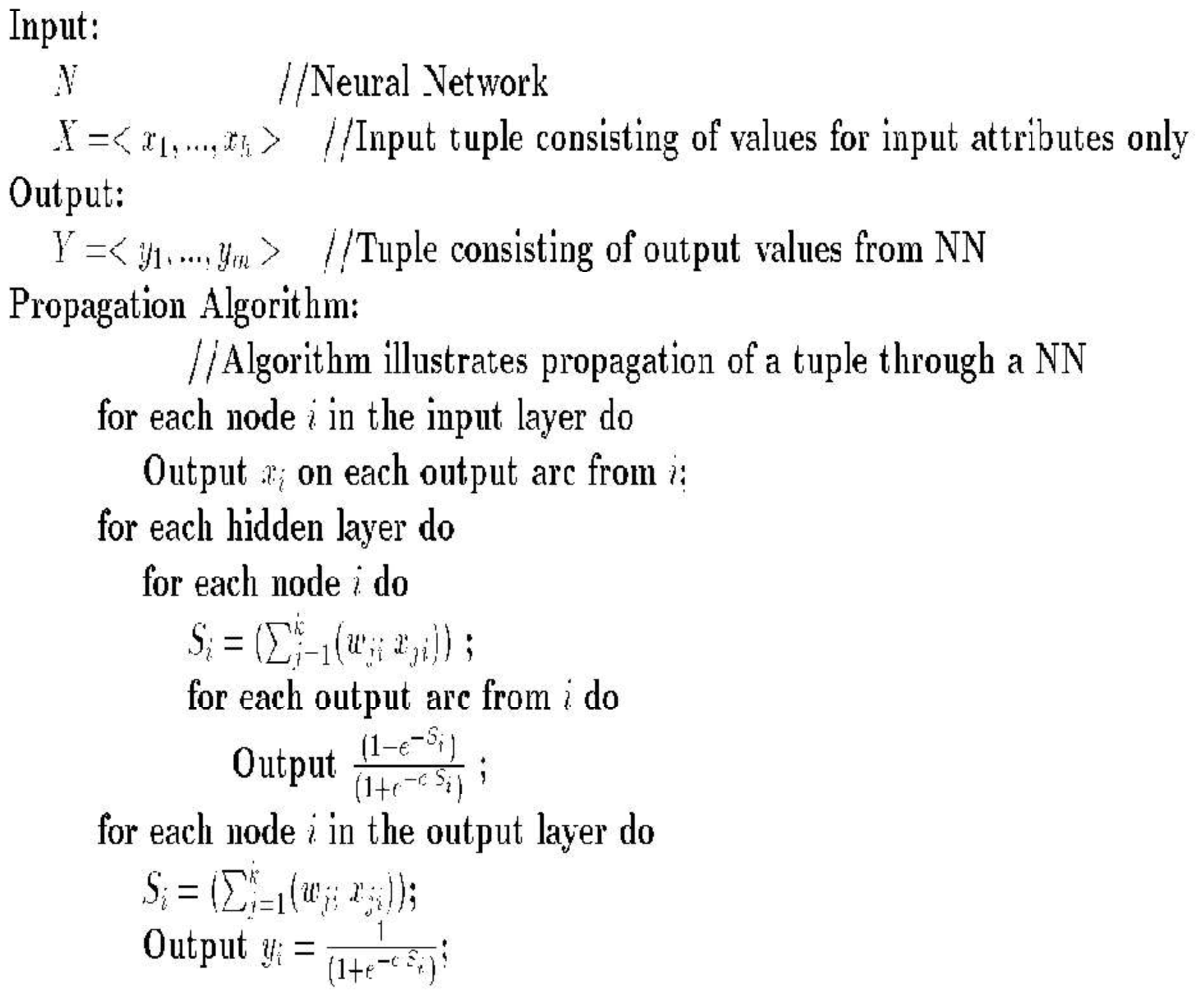
 When to stop learning

Decision Tree vs. Neural Network


Gender
=F
=M
Height
Height
<1.3m
>=1.3m
<=1.8m
>1.8m
<1.5m
>=1.5m
<=2m
> 2m
Short
Medium
Tall
Short
Medium
Tall



Propagation



```
Input:
    $N$                    //Neural Network
    $X = < x_1, ..., x_h >$    //Input tuple consisting of values for input attributes only
Output:
    $Y = < y_1, ..., y_m >$    //Tuple consisting of output values from NN
Propagation Algorithm:
                //Algorithm illustrates propagation of a tuple through a NN
        for each node $i$ in the input layer do
            Output $x_i$ on each output arc from $i$;
        for each hidden layer do
            for each node $i$ do
                $S_i = (\sum_{j-1}^{k}(w_{ji} x_{ji}))$ ;
                for each output arc from $i$ do
                    Output $\frac{(1-e^{-S_i})}{(1+e^{-c S_i})}$ ;
        for each node $i$ in the output layer do
            $S_i = (\sum_{j=1}^{k}(w_{ji} x_{ji}))$;
            Output $y_i = \frac{1}{(1+e^{-c S_i})}$;
```

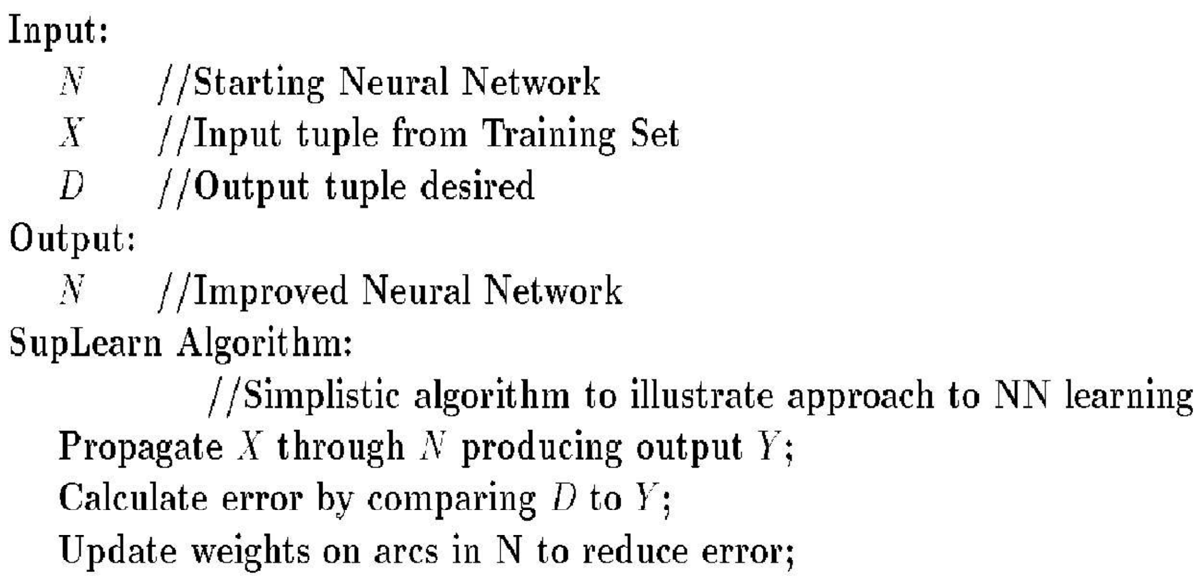
NN Learning

 Adjust weights to perform better with the associated test data.

 Supervised: Use feedback from knowledge of correct classification.

Unsupervised: No knowledge of correct classification needed.

NN Supervised Learning



```
Input:
    N       //Starting Neural Network
    X       //Input tuple from Training Set
    D       //Output tuple desired
Output:
    N       //Improved Neural Network
SupLearn Algorithm:
                //Simplistic algorithm to illustrate approach to NN learning
    Propagate X through N producing output Y;
    Calculate error by comparing D to Y;
    Update weights on arcs in N to reduce error;
```

Supervised Learning

□ Possible error values assuming output from node i is yi but should be di:

□ Change weights on arcs based on estimated error

$$| y_i - d_i |$$

$$\frac{(y_i - d_i)^2}{2}$$

$$\sum_{i=1}^{m} \frac{(y_i - d_i)^2}{m}$$

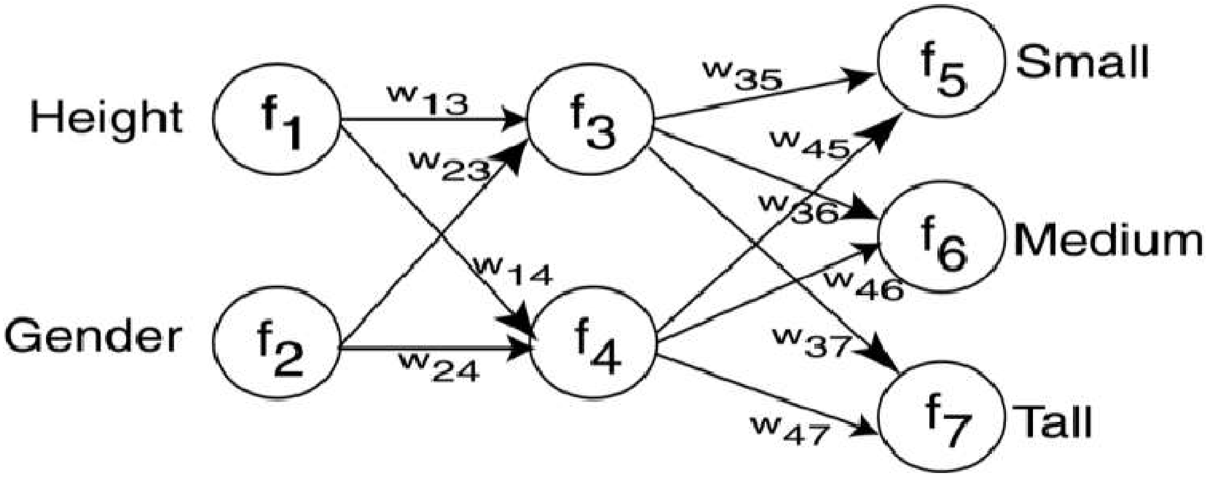
NN Backpropagation

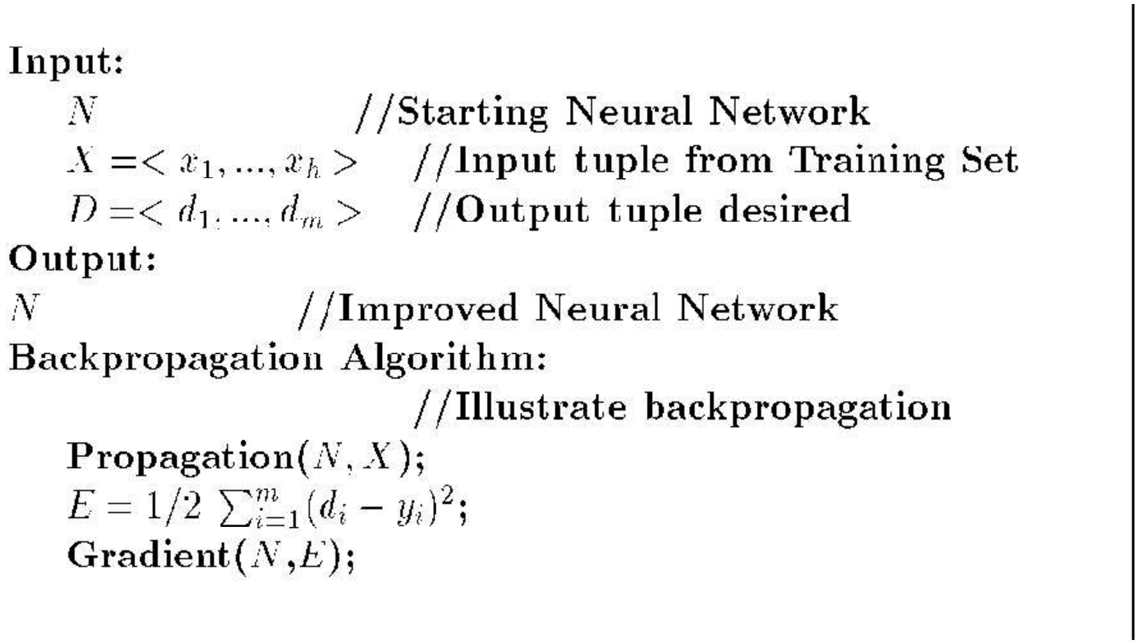
□ Propagate changes to weights backward from output layer to input layer.

□ ***Delta Rule:*** r wij= c xij (dj– yj)

□ ***Gradient Descent:*** technique to modify the weights in the graph.

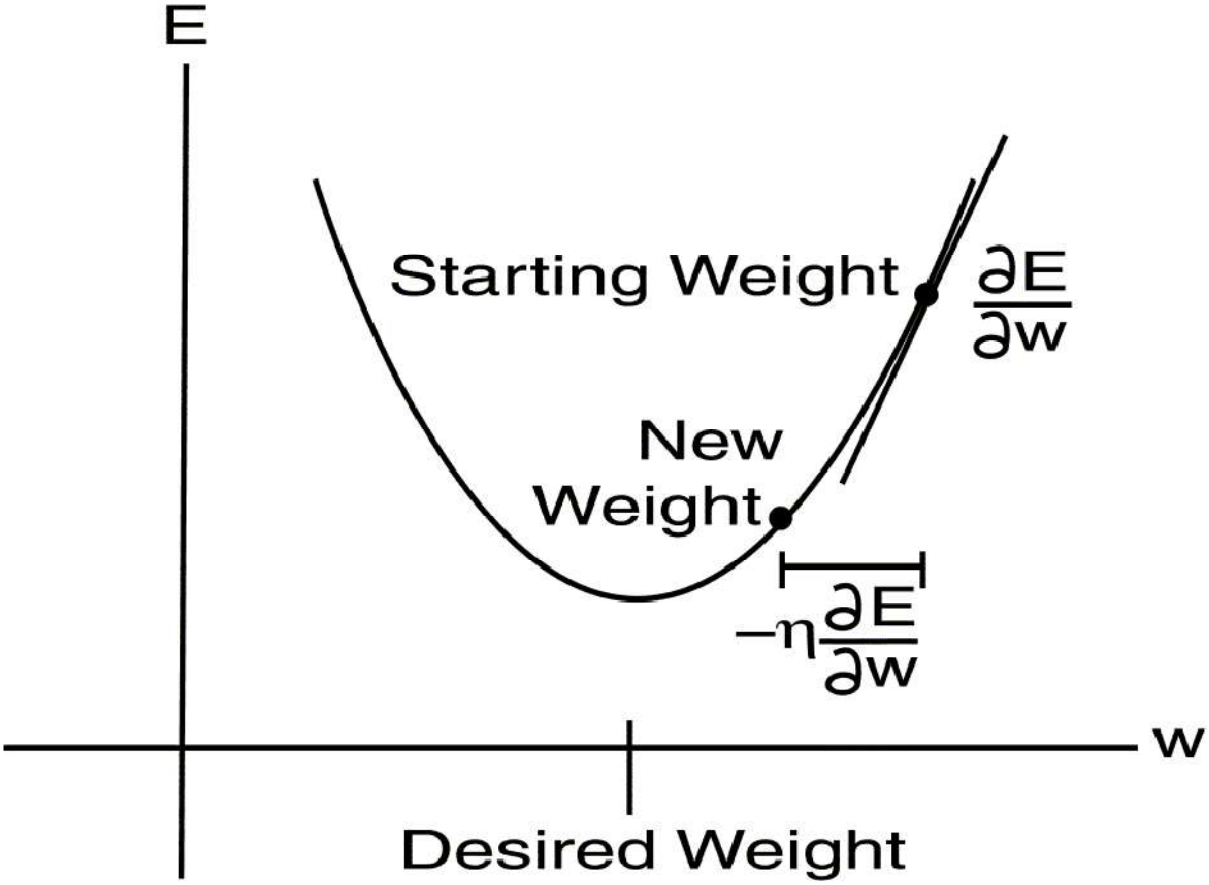
Backpropagation



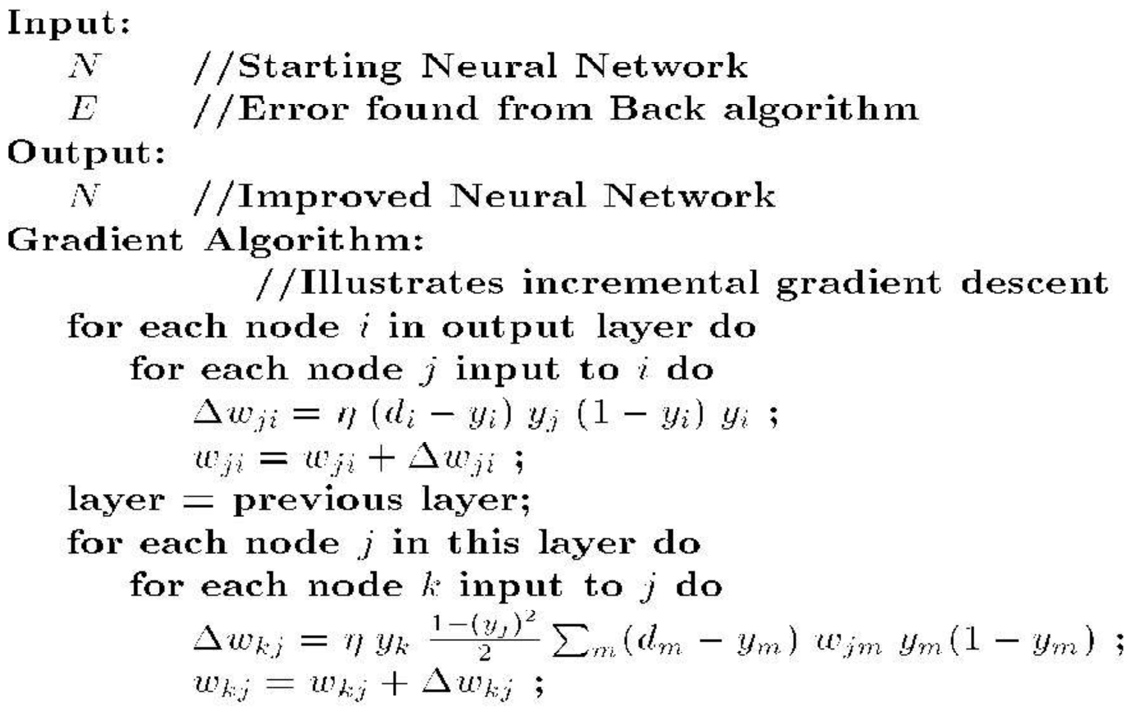


```
Input:
    $N$                     //Starting Neural Network
    $X = < x_1, ..., x_h >$     //Input tuple from Training Set
    $D = < d_1, ..., d_m >$     //Output tuple desired
Output:
$N$                     //Improved Neural Network
Backpropagation Algorithm:
                        //Illustrate backpropagation
    Propagation($N$, $X$);
    $E = 1/2 \sum_{i=1}^{m} (d_i - y_i)^2$;
    Gradient($N$,$E$);
```

Gradient Descent


E
Starting Weight
$\frac{\partial E}{\partial w}$
New Weight
$-\eta\frac{\partial E}{\partial w}$
w
Desired Weight

Gradient descent algorithm

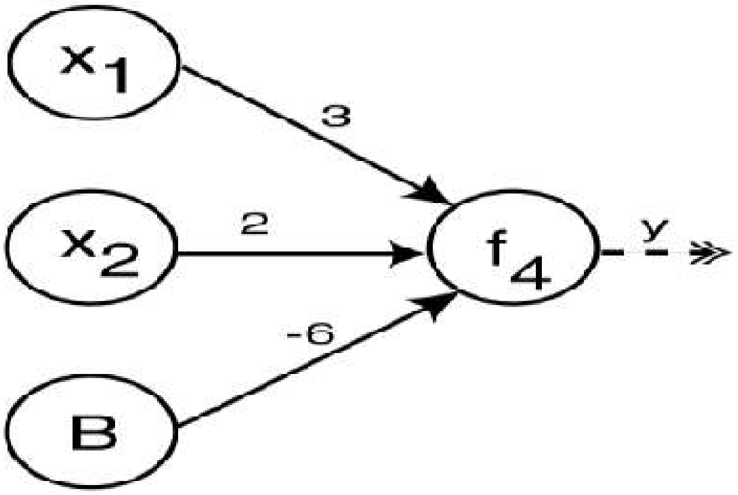


```
Input:
    $N$    //Starting Neural Network
    $E$    //Error found from Back algorithm
Output:
    $N$    //Improved Neural Network
Gradient Algorithm:
            //Illustrates incremental gradient descent
    for each node $i$ in output layer do
        for each node $j$ input to $i$ do
            $\Delta w_{ji} = \eta\, (d_i - y_i)\, y_j\, (1 - y_i)\, y_i$ ;
            $w_{ji} = w_{ji} + \Delta w_{ji}$ ;
    layer = previous layer;
    for each node $j$ in this layer do
        for each node $k$ input to $j$ do
            $\Delta w_{kj} = \eta\, y_k\, \frac{1-(y_j)^2}{2} \sum_m (d_m - y_m)\, w_{jm}\, y_m(1 - y_m)$ ;
            $w_{kj} = w_{kj} + \Delta w_{kj}$ ;
```

Types of NNs

- Different NN structures used for different problems.
- Perceptron
- Self Organizing Feature Map
- Radial Basis Function Network

Perceptron



Self Organizing Feature Map (SOFM)
 Competitive Unsupervised Learning

 Observe how neurons work in brain:
– Firing impacts firing of those near

– Neurons far apart inhibit each other

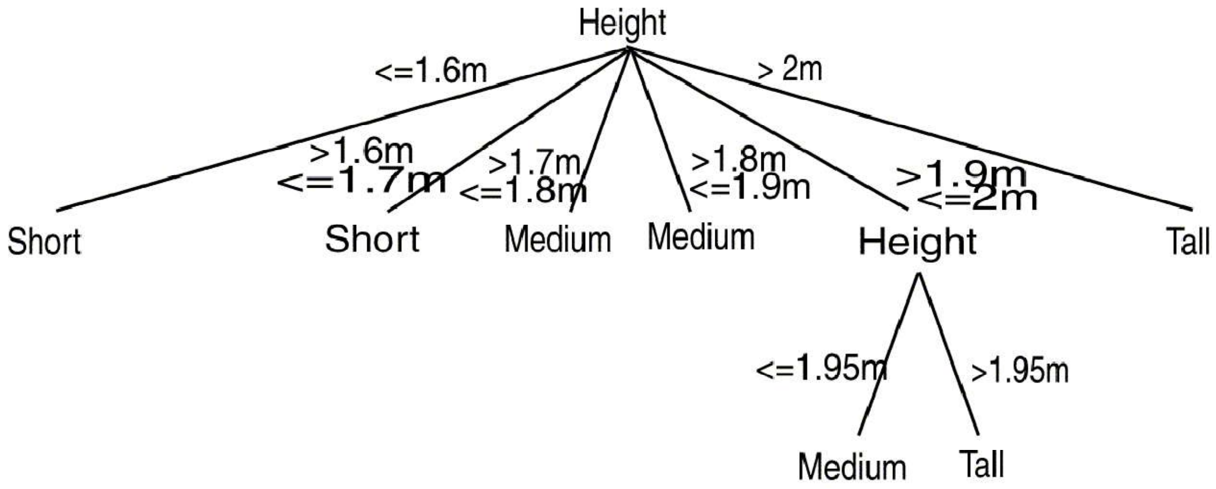
– Neurons have specific nonoverlapping tasks

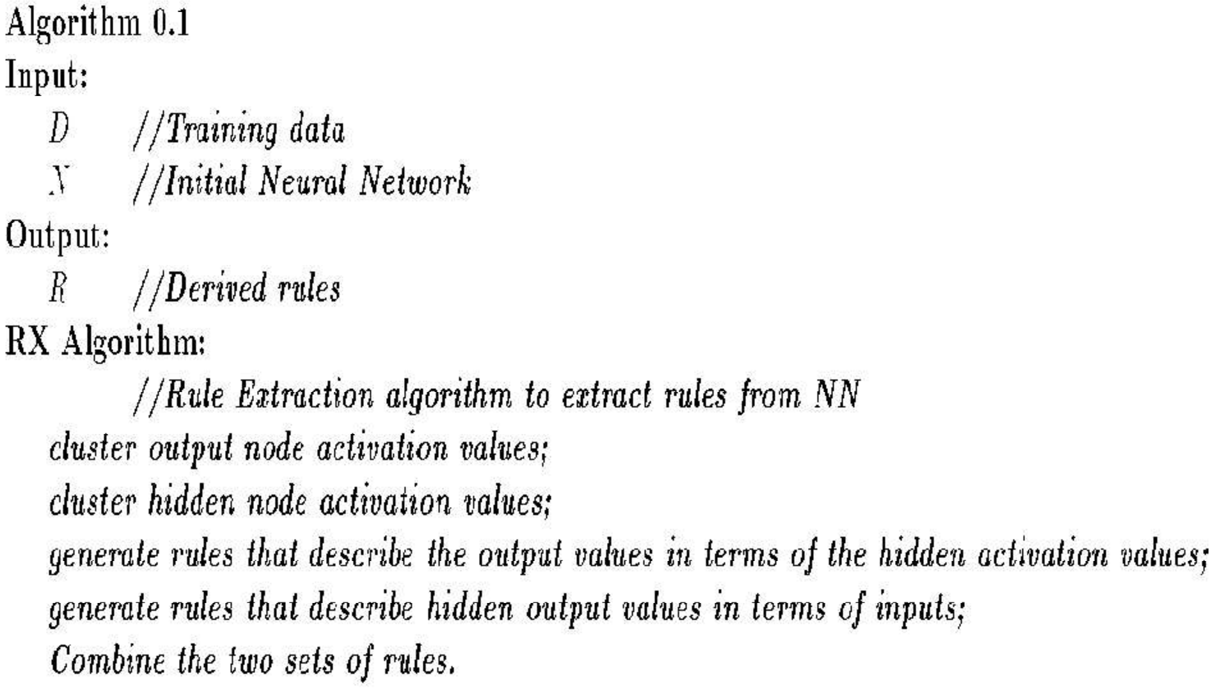
Classification Using Rules
 Perform classification using If-Then rules

 Classification Rule: r = <a,c>
o Antecedent, Consequent
 May generate from from other techniques (DT, NN) or generate directly.

 Algorithms: Gen, RX, 1R, PRISM

Generating Rules from DTs





```
Algorithm 0.1
Input:
    D      //Training data
    N      //Initial Neural Network
Output:
    R      //Derived rules
RX Algorithm:
           //Rule Extraction algorithm to extract rules from NN
    cluster output node activation values;
    cluster hidden node activation values;
    generate rules that describe the output values in terms of the hidden activation values;
    generate rules that describe hidden output values in terms of inputs;
    Combine the two sets of rules.
```

1R

 An easy way to find very simple classification rules from a set of instances.

 1 level DT

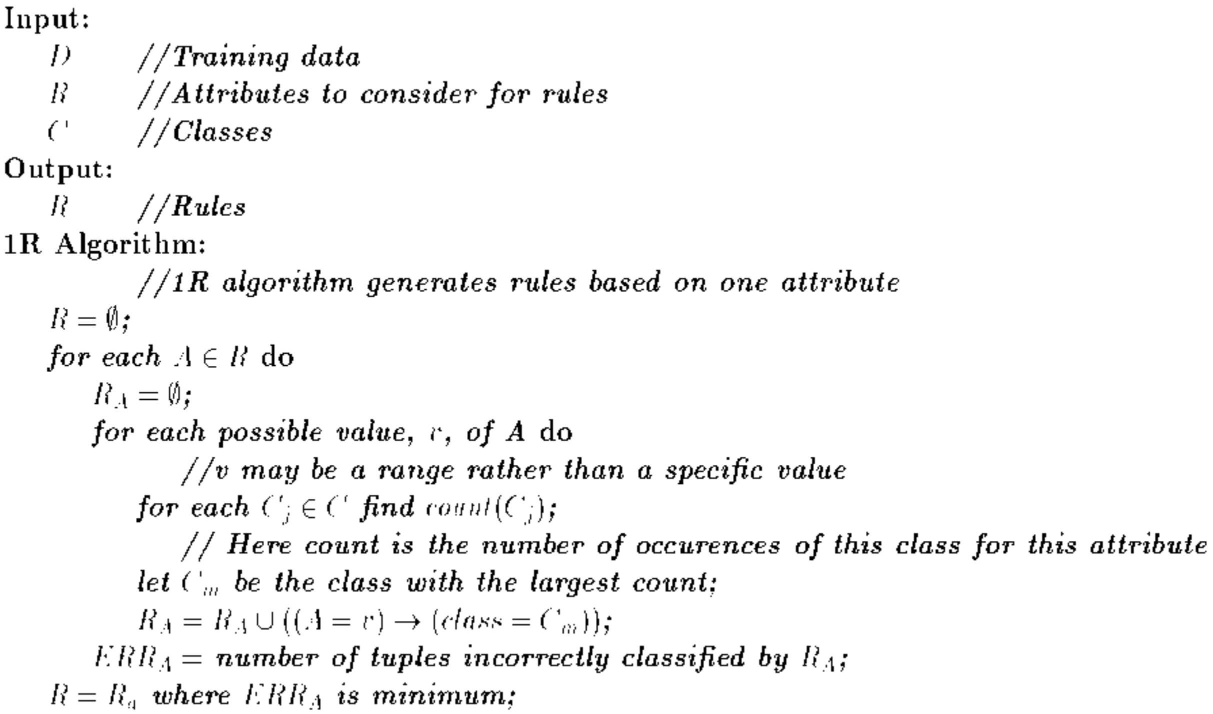
 When to use it – always try the simplest thing first

1R informal description
 For each attribute
– For each value of that attribute, make a rule:
» Count how often each class appears, find the most frequent class, make the rule assign that class to this attribute-value

» Calculate the error rate of the rules for each attribute

– Choose one rule with the smallest total error rate among all attribute value-based rules

1R Algorithm



```
Input:
    $D$      //Training data
    $R$      //Attributes to consider for rules
    $C$      //Classes
Output:
    $R$      //Rules
1R Algorithm:
             //1R algorithm generates rules based on one attribute
    $R = \emptyset$;
    for each $A \in R$ do
        $R_A = \emptyset$;
        for each possible value, $v$, of $A$ do
                //v may be a range rather than a specific value
            for each $C_j \in C$ find $count(C_j)$;
                // Here count is the number of occurences of this class for this attribute
            let $C_m$ be the class with the largest count;
            $R_A = R_A \cup ((A = v) \rightarrow (class = C_m))$;
        $ERR_A$ = number of tuples incorrectly classified by $R_A$;
    $R = R_a$ where $ERR_A$ is minimum;
```

Clustering

- ***Segment*** customer database based on similar buying patterns.
- Group houses in a town into neighborhoods based on similar features. Identify new plant species

□ Identify similar Web usage

patterns □ Clustering vSupport

Confidence

□ Interest

□ Conviction

□ Chi Squared

Test

Clustering vs.

Classification

□ No prior knowledge

– Number of

clusters –Meaning

of clusters

□ Unsupervised learning

Clustering Issues

□ Outlier

handling

□ Dynamic data

□ Interpreting results

□ Evaluating
results

□ Number of
clusters

□ Data to be used

□ Scalability

Impact of Outliers on Clustering

Clustering Problem

□ Given a database D={t1,t2,…,tn} of tuples and an integer value k, the ***Clustering Problem*** is to define a mapping f:Dg{1,..,k} where each ti is assigned to one cluster Kj, 1<=j<=k.

□ A ***Cluster***, Kj, contains precisely those tuples mapped to it.

□ Unlike classification problem, clusters are not known a priori.

Types of Clustering

□ ***Hierarchical*** – Nested set of clusters created. □ ***Partitional***– One set of clusters created.

□ ***Incremental*** – Each element handle ***Simultaneous*** – All elements handled together.

□ ***Overlapping/Non-overlapping***

Cluster Parameters

$$centroid = C_m = \frac{\sum_{i=1}^{N}(t_{mi})}{N}$$

$$radius = R_m = \sqrt{\frac{\sum_{i=1}^{N}(t_{mi} - C_m)^2}{N}}$$

$$diameter = D_m = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(t_{mi} - t_{mj})^2}{(N)(N-1)}}$$

$$centroid = C_m = \frac{\sum_{i=1}^{N}(t_{mi})}{N}$$

$$radius = R_m = \sqrt{\frac{\sum_{i=1}^{N}(t_{mi} - C_m)^2}{N}}$$

$$diameter = D_m = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(t_{mi} - t_{mj})^2}{(N)(N-1)}}$$

Distance Between Clusters

□ ***Single Link***: smallest distance between points

□ ***Complete Link:*** largest distance between points □ ***Average Link:*** average distance between points □ ***Centroid:*** distance between centroids

Hierarchical Clustering

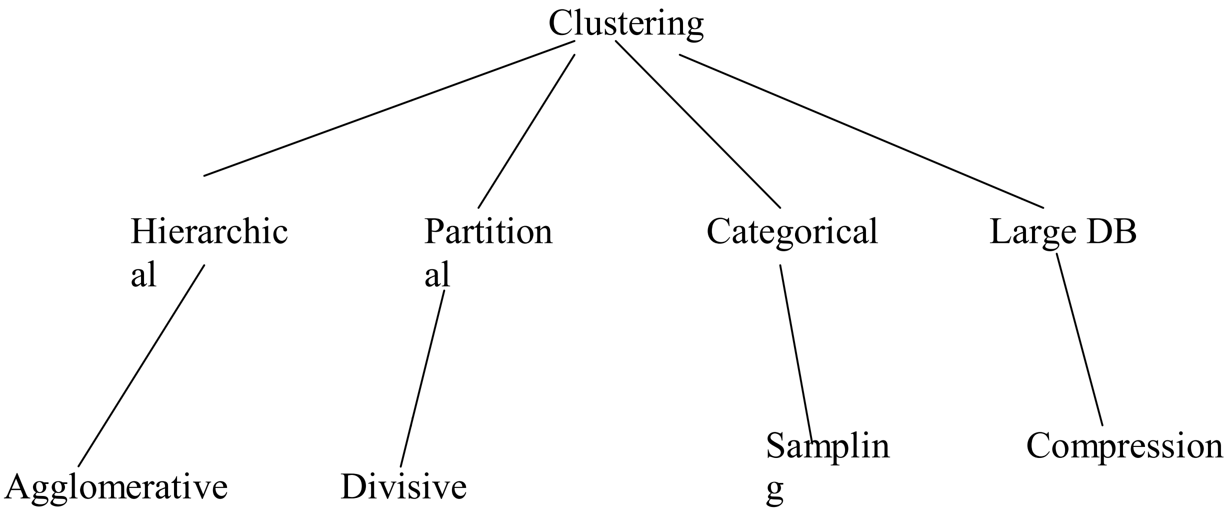
□ Clusters are created in levels actually creating sets of clusters at each level. □ ***Agglomerative (compare with Merge sort)***

_ _ _

Initially each item in its own cluster

Iteratively clusters are merged together

Bottom Up



□ ***Divisive (compare with Bubble sort)*** – Initially all items in one cluster

– Large clusters are successively divided – Top Down

Hierarchical Algorithms □ Single Link

□ MST Single Link □

Complete Link

□ Average Link

Dendrogram

□ ***Dendrogram:*** a tree data structure which illustrates hierarchical clustering techniques. □ Each level shows clusters for that level.

– Leaf – individual
clusters –Root – one
cluster

 A cluster at level i is the union of its children clusters at level i+1.

Levels of Clustering

Agglomerative Algorithm

d) Two Clusters e) One Cluster

Agglomerative Example

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

Single Link

□View all items with links (distances) between them. Finds maximal connected components in this graph.

□ Two clusters are merged if there is at least one edge which connects them.

□ Uses threshold distances at each level.

- Could be agglomerative or divisive.

MST Single Link Algorithm

Single Link Clustering

a) Single Link b) Complete Link b) Average Link

Partitional Clustering

□ Nonhierarchical

□ Creates clusters in one step as opposed to several steps.

□ Since only one set of clusters is output, the user normally has to input the desired number of clusters, k.

□ Usually deals with static sets.

Partitional
Algorithms □ MST
□ Squared
Error □ K-
Means

□ Nearest Neighbor

□ PAM

□ BEA

□ GA

MST Algorithm

Squared Error

$$se_{K_i} = \sum_{j=1}^{m} \|t_{ij} - C_k\|^2$$

$$se_K = \sum_{j=1}^{k} se_{K_j}$$

- **Minimized squared error**

Squared Error Algorithm

K-Means

- Initial set of clusters randomly chosen.

 Iteratively, items are moved among sets of clusters until the desired set is reached. High degree of similarity among elements in a cluster is obtained.

 Given a cluster Ki={ti1,ti2,…,tim}, the ***cluster mean*** is m = (1/m)(ti1 + … + tim)

K-Means Example

 Given: {2,4,10,12,3,20,30,11,25}, k=2

 Randomly assign means (seeds): m1=3,m2=4

 K1={2,3}, K2={4,10,12,20,30,11,25},
m1=2.5,m2=16
K1={2,3,4},K2={10,12,20,30,11,25},
m1=3,m2=18

 K1={2,3,4,10},K2={12,20,30,11,25},
m1=4.75,m2=19.6
K1={2,3,4,10,11,12},K2={20,30,25},
m1=7,m2=25

 Stop as the clusters with these means are the same. (until the centroids do not change)

K-Means Algorithm

Nearest Neighbor

Items are iteratively merged into the existing clusters that
are closest. □ Incremental

□ Threshold, t, used to determine if items are added to existing clusters or a new cluster is created.

PAM

□ ***Partitioning Around Medoids (PAM) (K-***

Medoids) □ Handles outliers well.

□ Ordering of input does not impact

results. □ Does not scale well.

□ Each cluster represented by one item, called the

medoid. □ Initial set of k medoids randomly

chosen.

PAM Cost Calculation

□ At each step in algorithm, medoids are changed if the overall cost is improved.

□ Cjih – cost change for an item tj associated with swapping medoidti with non-medoidth.

PAM Algorithm

BEA

□ Bond Energy Algorithm

□ Database design (physical and

logical) □ Vertical

fragmentation

□ Determine affinity (bond) between attributes based on

common usage. □ Algorithm outline:

1. Create affinity matrix
2. Convert to BOND matrix
3. Create regions of close bonding

Genetic Algorithm Example

(cross over, mutation, fitness function)

Association Rule

Definitions □ ***Set of items:*** I={I1,I2,…,Im}

□ ***Transactions:*** D={t1,t2, …, tn}, tj□□I □ ***Itemset:*** {Ii1,Ii2, …, Iik} □□I

□ ***Support of an itemset:*** Percentage of transactions which contain that itemset.

□ ***Large (Frequent) itemset:***Itemset whose number of occurrences is above a threshold.

I = { Beer, Bread, Jelly, Milk, PeanutButter}

Support of {Bread,PeanutButter} is 60%

Association Rule Problem

□ Given a set of items I={I1,I2,…,Im} and a database of transactions D={t1,t2, …, tn} where ti={Ii1,Ii2, …, Iik} and Iij□□I, the ***Association Rule Problem*** is to identify all association rules X □□Y with a minimum (lower bound) support and confidence.

□ Link Analysis

□ ***NOTE:*** Support of X □□Y is same as support of X □□Y.

Association Rule

Techniques 1. Find

Large Itemsets.

2. Generate rules from frequent itemsets.
3. Importance is measured by two features called support and confidence
4. Algorithms are mostly based on smart ways to reduce the number of itemsets to be counted to identify large itemsets
5. Data structure during counting: trie or hash tree

Apriori

□ ***Large Itemset Property:***

Any subset of a large itemset is large. □ Contrapositive:

If an itemset is not large, none of its supersets are large. **Large Itemset Property**

Apriori Algorithm

1. C1 = Itemsets of size one in I;

2. Determine all large itemsets of size

1, L1; 3. i = 1;

4. Repeat

5. i = i +

1;

6. Ci = Apriori-Gen(Li-

1); 7. Count Ci to

determine Li;

8. until no more large itemsets found;

Apriori-Gen

□ Generate candidates of size i+1 from large itemsets of size i.

□Approach used: join large itemsets of size i if they
agree on i-1 □May also prune candidates who have
subsets that are not large.

AprioriAdv/Dis

Advantages:

– Easily

parallelized –

Easy to

implement.

 Disadvantages:

- – Assumes transaction database is memory resident.
- – Requires up to m database scans.

Sampling

□ Large databases

□ Sample the database and apply Apriori to the sample.

□ ***Potentially Large Itemsets (PL):*** Large

itemsets from sample □ ***Negative Border***

(BD^-):

– Generalization of Apriori-Gen applied to itemsets of varying sizes.

– Minimal set of itemsets which are not in PL, but whose

subsets are all in PL. **Sampling Algorithm**

1. Ds = sample of Database D;
2. PL = Large itemsets in Ds using smalls (any support

values less than s); 3. C = PL □□BD^-(PL);

4. Count C in Database using s;
5. ML = large

itemsets in BD^-(PL); 6.

If ML = □□then done

7. else C = repeated

application of BD^-; 8.

Count C in Database;

Sampling

Adv/Dis

adv □

***Adva
ntages:***

– Reduces number of database scans to one in the best
case and two in worst. – Scales better.

□ ***Disadvantages:***

–

□ Divide database into
partitions D^1,D^2,…,D^p □
Apply Apriori to each
partition

□ Any large itemset must be large in at least one partition.

Partitioning Algorithm

1. Divide D into partitions
D^1,D^2,…,D^p; 2. For I = 1
to p do
3. L^i =
Apriori(D^i);
4. C =
L^1□□…
□L^p;

5. Count C on D

to generate L;

Partitioning

Adv/Disadv

□ ***Advantages:***

– Adapts to available

main memory – Easily

parallelized

– Maximum number of

database scans is two. □

Disadvantages:

– May have many candidates

during second scan. **Parallelizing AR**

Algorithms

□ Based

on Apriori

□

Techni

ques

differ:

– What is counted at each site

– How data (transactions)

are distributed □ Data

Parallelism

– Data partitioned

– Count

Distribution Algorithm □

Task Parallelism

– Data and

candidates partitioned

– Data Distribution

Algorithm

Comparison of AR Techniques

Incremental Association Rules

□ Generate ARs in a dynamic database.

□ Problem: algorithms assume

static database □ Objective:

– Know large itemsets for D

– Find large itemsets

for D □□{D D} □ Must be

large in either D or D D

□ Save Li

and counts

Note on ARs

□ Many applications outside market

basket data analysis – Prediction

(telecom switch failure)

– Web usage mining

- Many different types of association rules – Temporal – Spatial – Causal

Advanced AR Techniques

- Generalized Association Rules
- Multiple-Level Association Rules
- Quantitative Association Rules

 Using multiple
minimum supports
Correlation Rules

Measuring
Quality of
Rules
Support
Confide
nce
In
terest
 Conviction
 Chi Squared Test

Possible Questions

1. Explain large itemset algorithm.
2. Explain apriori-gen algorithm.
3. Explain sampling algorithm.
4. Explain incremental rules.
5. How will you measure the quality of rules?
6. Explain classification problem.
7. Explain K-nearest neighbor algorithm
8. Explain distance algorithm
9. Discuss issues in decision tree algorithms.
10. Explain back propagation.

Scalability and data management issues in data mining algorithms, measures of interestingness.

Scalability of Data Mining

The scalability of data mining techniques is very important due to the rapid growth in the size of databases. Usage of decision tree classifiers has become an effective classification model. The main objective of this research is to study the existing Scalable Decision Tree Classifiers and analyze them to find the best algorithm. Our aim is also to design a new SDTC algorithm and compare it analytically with the existing ones.

Data that are now available in any field of research poses new problems for data mining and knowledge discovery methods. Due to this huge amount of data, most of the current data mining algorithms are inapplicable to many real-world problems. Data mining algorithms become ineffective when the problem size becomes very large. In many cases, the demands of the algorithm in terms of the running time are very large, and mining methods cannot be applied when the problem grows. This aspect is closely related to the time complexity of the method. A second problem is linked with performance; although the method might be applicable, the size of the search space prevents an efficient execution, and the resulting solutions are unsatisfactory. Two approaches have been used to deal with this problem: scaling up data mining algorithms and data reduction. However, because data reduction is a data mining task itself, this technique also suffers from scalability problems. Thus, for many problems, especially when dealing with very large datasets, the only way to deal with the aforementioned problems is to scale up the data mining algorithm. Many efforts have been made to obtain methods that can be used to scale up existing data mining algorithms. In this paper, we review the methods that have been used to address the problem of scalability. We focus on general ideas, rather than specific implementations, that can be used to provide a general view of the current approaches for scaling up data mining methods. A taxonomy of the algorithms is proposed, and many examples of different tasks are presented. Among the different techniques used for data mining, we will pay special attention to evolutionary methods, because these methods have been used very successfully in many data mining tasks.

Data Mining Very Large Data Sets (Databases): Scalability of Statistica Data Miner

An important issue in data mining is how the various techniques for exploratory (EDA), visual, and particularly predictive data mining (see Concepts in Data Mining) perform when applied to extremely large data sets. It is not uncommon in many domains of application to deal with data sets in the multiple gigabyte range, with tens of millions of observations. Analyzing data sets of this size will require some planning to avoid unnecessary performance bottlenecks, and inappropriate analytic choices. For example, using advanced neural network techniques to analyze all of 20 million observations is simply inappropriate, because a) in some cases it could take as long as several days to complete even using a designated supercomputer-class mainframe, and b) the same information can be quickly extracted from the data by applying an appropriate sub-sampling method first, and then analyzing a reasonable subset of the input data.

Statistica Data Miner uses a number of technologies specifically developed to optimize processing large data sets, and it is designed to handle even largest scale computational problems based on very large databases. However, in order to take best advantage of the computational power of Statistica Data Miner, a number of issues still need to be considered when planning data mining projects for very large data sets. The following paragraphs discuss various strategies and (unique) tools available in Statistica Data Miner that you can use to analyze and build models for huge source data.

Connecting to Data

Most likely, data sets that are very large (in the gigabyte range) will reside on a (remote) server, and it is not practical or even desirable to copy those data onto a designated computer for data mining. As an alternative, you can use the facilities for Streaming Database Connector to connect to the data. These unique tools enable you to select the variables (fields) of interest from the database, and many subsequent analyses (including many graphs) will be able to process those data in a single pass through all observations in the database without the need to create a local (on your computer or server) copy of the data. See Select a New Data Source and Streaming Database Connector for additional details.

Random Sub-Sampling

We cannot stress enough the importance and utility of random sub-sampling. For example, by properly sampling only 100 observations (from millions of observations) you can compute a very reliable estimate of the mean. One of the rules of statistical sampling that is often not intuitively understood by untrained observers is the fact that the reliability and validity of results depend, among many other things, on the size of a random sample, and not on the size of the population from which it is taken. In other words, the mean estimated from 100 randomly sampled observations is as accurate (i.e., falls within the same confidence limits) regardless of whether the sample was taken from 1000 cases or 100 billion cases. Put another way, given a certain (reasonable) degree of accuracy required, there is absolutely no need to process and include all observations in the final computations (for estimating the mean, fitting models, etc.).

Statistica Data Miner contains nodes in the Data Cleaning and Filtering folder to draw a random sample from the original input data (database connection). Note that Statistica employs a very high-quality (validated using the DIEHARD suite of tests) random number generator algorithm that ensures that the selection of observations will not be biased.

Statistica Power Analysis

For detailed planning of sub-sampling in predictive data mining, you can also use the **Statistica Power Analysis** facilities, which can provide very valuable information regarding the relationship between sample sizes, effect sizes (that would be of interest to you), and statistical power (to detect effects) given different statistical techniques. Power analysis methods have been popular in applied and survey research for a number of years, but have not yet been popularized in the area of data mining, even though these methods can be potentially extremely useful here, in particular in the context of extremely large data sets.

Algorithms for Incremental (vs. Non-Incremental) Learning

Statistica Data Miner contains a large selection of (learning) algorithms for regression and classification problems. These algorithms can be divided into those that require one or perhaps two complete passes through the input data, and those that require iterative multiple access to the data to complete the estimation. The former type of algorithms are also sometimes referred to as incremental learning algorithms, because they will complete the computations necessary to fit the respective models by processing one case at a time, each time refining the solution; then, when all cases have been processed, only few additional computations are necessary to produce the final results. Non-incremental learning algorithms are those that need to process all observations in each iteration of an iterative procedure for refining a final solution. Obviously, incremental learning algorithms are usually much faster than non-incremental algorithms, and for extremely large data sets, non-incremental algorithms may not be applicable at all (without first sub-sampling).

However, as explained in the previous paragraphs on Random Sub-Sampling, in most, if not all cases, it is not useful anyway to process every single observations in a very large database; it simply is a waste of data processing resources and time, and by carefully planning and sub-sampling, the same information can be extracted in a much shorter time, or much more information can be extracted in the same amount of time that it would require to include all observations in the analyses.

Incremental algorithms

Statistica Data Miner includes several extremely powerful, efficient, and fast algorithms for regression and classification that will analyze the data in a single pass through all (or a sub-sample of) observations: For example, Statistica GDA is an extension of the general linear model to classification problems (General Discriminant Function Analysis models). This method (unique to Statistica Data Miner) is the fastest algorithm for classification available and does not require that the source data be copied to a local computer or server (and huge databases can be processed in-place). This method yields outstanding accuracy for predictive classification in most cases; various options are available for this technique to request best subset or stepwise selections of predictor effects. The implementation of stepwise and best-subset selection of predictors for regression problems using the general linear model (GRM, GLM) is equally unique. It also is an incremental learning algorithm that will perform stepwise and best-subset selection of predictor effects, etc. (categorical/class variables will be moved in/out of models as multiple-degree-of-freedom effects).

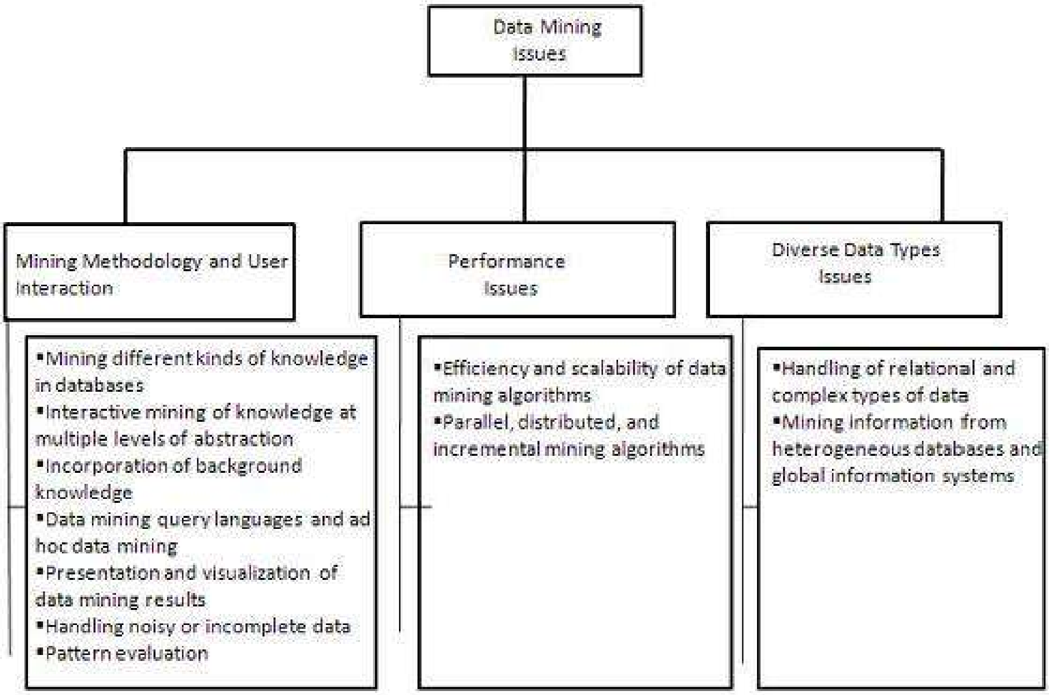
Major Issues in Data Mining

Data mining is a dynamic and fast-expanding field with great strengths. In this section, we briefly outline the **major issues** in **data mining** research, partitioning them into five groups: **mining** methodology, user interaction, efficiency and scalability, diversity of **data**types, and **data mining** and society.

Data mining is not that easy. The algorithm used are very complex. The data is not available at one place it needs to be integrated form the various heterogeneous data sources. These factors also creates some issues. Here in this tutorial we will discuss the major issues regarding:

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.


Data Mining Issues
Mining Methodology and User Interaction
Performance Issues
Diverse Data Types Issues
▪Mining different kinds of knowledge in databases
▪Interactive mining of knowledge at multiple levels of abstraction
▪Incorporation of background knowledge
▪Data mining query languages and ad hoc data mining
▪Presentation and visualization of data mining results
▪Handling noisy or incomplete data
▪Pattern evaluation
▪Efficiency and scalability of data mining algorithms
▪Parallel, distributed, and incremental mining algorithms
▪Handling of relational and complex types of data
▪Mining information from heterogeneous databases and global information systems

Mining Methodology and User Interaction Issues

It refers to the following kind of issues:

- **Mining different kinds of knowledge in databases. -** The need of different users is not the same. And Different user may be in interested in different kind of knowledge. Therefore it is necessary for data mining to cover 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 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 level 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, visual representations. This representations should be easily understandable by the users.
- **Handling noisy or incomplete data.** - The data cleaning methods are required that can handle the noise, incomplete objects while mining the data regularities. If data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation.** - It refers to interestingness of the problem. The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

It refers to the following issues:

- **Efficiency and scalability of data mining algorithms.** - In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms.** - The factors such as huge size of databases, wide distribution of data,and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithm divide the data into partitions which is further processed parallel. Then the results from the partitions is merged. The incremental algorithms, updates databases without having mine the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data.** - The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems.** - The data is available at different data sources on LAN or WAN. These data source may be

structured, semi structured or unstructured. Therefore mining knowledge from them adds challenges to data mining.

Mining Methodology and User Interaction Issues

It refers to the following kinds of issues −

- **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.

Performance Issues

There can be performance-related issues such as follows −

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

Interesting Measures

Interestingness measures play an important role in data mining, regardless of the kind of patterns being mined. These measures are intended for selecting and ranking patterns according to their potential interest to the user. Good measures also allow the time and space costs of the mining process to be reduced. This survey reviews the interestingness measures for rules and summaries, classifies them from several perspectives, compares their properties, identifies their roles in the data mining process, gives strategies for selecting appropriate measures for applications, and identifies opportunities for future research in this area. Measuring the interestingness of discovered patterns is an active and important area of data mining research. Although much work has been conducted in this area, so far there is no widespread agreement on a formal definition of interestingness in this context. Based on the diversity of definitions presented to-date, interestingness is perhap best treated as a broad concept that emphasizes *conciseness*, *coverage*, *reliability*, *peculiarity*, *diversity*, *novelty*, *surprisingness*, *utility*, and *actionability*. These nine specific criteria are used to determine whether or not a pattern is interesting. They are described as follows.

Conciseness. A pattern is concise if it contains relatively few attribute-value pairs, while a set of patterns is concise if it contains relatively few patterns. A concise pattern or set of patterns is relatively easy to understand and remember and thus is addedmore easily to the user's knowledge (set of beliefs). Accordingly, much research has been conducted to find a "minimum set of patterns," using properties such as monotonicity and confidence invariance.

Generality/Coverage. A pattern is *general* if it covers a relatively large subset of a dataset. Generality (or coverage) measures the comprehensiveness of a pattern, that is, the fraction of all records in the dataset that matches the pattern. If a pattern characterizes more information in the dataset, it tends to be more interesting . Frequent itemsets are the most studied general patterns in the data mining literature. An itemset is a set of items, such as some items from a grocery basket. An itemset is *frequent* if its *support*, the fraction of records in the dataset containing the itemset, is above a given threshold best known algorithm for finding frequent itemsets is the Apriori algorithm.

Some generality measures can form the bases for pruning strategies; for example, the support measure is used in the Apriori algorithm as the basis for pruning itemsets. For classification rules, gave an empirical evaluation showing how generality affects classification results. Generality frequently coincides with conciseness because concise patterns tend to have greater coverage.

Reliability. A pattern is reliable if the relationship described by the pattern occurs in a high percentage of applicable cases. For example, a classification rule is reliable if its predictions are highly accurate, and an association rule is reliable if it has high confidence. Many measures from probability, statistics, and information retrieval have been proposed to measure the reliability of association rules.

Peculiarity. A pattern is peculiar if it is far away from other discovered patterns according to some distance measure. Peculiar patterns are generated from peculiar data (or outliers), which are relatively few in number and significantly different from the rest of the data. Peculiar patterns may be unknown to the user, hence interesting.

Interestingness Measures for Data Mining:

Diversity. A pattern is diverse if its elements differ significantly from each other, while a set of patterns is diverse if the patterns in the set differ significantly from each other. Diversity is a common factor for measuring the interestingness of summaries. According to a simple point of view, a summary can be considered diverse if its probability distribution is far from the uniform distribution. A diverse summary may be interesting because in the absence of any relevant knowledge, a user commonly assumes that the uniform distribution will hold in a summary.

According to this reasoning, the more diverse the summary is, the more interesting it is. We are unaware of any existing research on using diversity to measure the interestingness of classification or association rules.

Novelty. A pattern is novel to a person if he or she did not know it before and is not able to infer it from other known patterns. No known data mining system represents everything that a user knows, and thus, novelty cannot be measured explicitly with reference to the user's knowledge. Similarly, no known data mining system represents what the user does not know, and therefore, novelty cannot be measured explicitly with reference to the user's ignorance. Instead, novelty is detected by having the user either explicitly identify a pattern as novel or notice that a pattern cannot be deduced from and does not contradict previously discovered patterns.

Surprisingness. A pattern is surprising (or *unexpected*) if it contradicts a person's existing knowledge or expectations. A pattern that is an exception to a more general pattern which has already been discovered can also be considered surprising. Surprising patterns are interesting because they identify failings in
previous knowledge and may suggest an aspect of the data that needs further study.

The difference between surprisingness and novelty is that a novel pattern is new and not contradicted by any pattern already known to the user, while a surprising pattern contradicts the user's previous knowledge or expectations.

Utility. A pattern is of utility if its use by a person contributes to reaching a goal. Different people may have divergent goals concerning the knowledge that can be extracted from a dataset. For example, one person may be interested in finding all sales with high profit in a transaction dataset, while another may be interested in finding all transactions with large increases in gross sales. This kind of interestingness is based on user-defined utility functions in addition to the raw data .

Actionability/Applicability. A pattern is actionable (or *applicable*) in some domain if it enables decision making about future actions in this domain. Actionability is sometimes associated with a pattern selection strategy. So far, no general method for measuring actionability has been devised. Existing measures depend on the applications.

Data Mining Application Fields and Studies Done In These Fields

Data mining has application fields in many branches such as banking, stock exchange, marketing management, retail sales, signal processing, insurance, telecommunication, electronic commerce, health, medicine, biology, genetics, industry, construction, education, intelligence, science, and engineering [5, 10, 11]. Various businesses apply data mining to critical business processes to gain competitive advantage and help businesses grow. In this study, data mining applications in the main sectors covering the mentioned sectors were researched and explained with examples of how and for what purpose data mining was used.

Banking and Finance Sector

Data mining is mostly used in the banking and finance sectors to determine what, when, and why the customer profile prefers. At the same time, it is also used in these fields to find appropriate solutions for the right demand creation and presentation the right time demands. Furthermore, it is also used in

□financial forecasting,

□estimation of stock prices,

□management of new investments,

□determination of investment portfolio,

□formation of marketing strategies,

□making risk analyses,

□making the right choice in terms of human resources forbusiness, credit and credit card fraud estimates,

□credit limit determination and fee management.

Data mining applications are used extensively toimprove the performance of some core business processes in the banking sector . Some banks, for example Garanti Bank in Turkey, use data mining methods to have information about the behavior models of customers and offer appropriate and successful promotions by examining the relationship between customer 'credit cards' selectivity and their character.

Education Sector

It has been observed that data mining has been used in many studies in the education sector including:

□determining the status of students' pass and fail,

□factors affecting the success of the students enrolling atthe university,

□creating the preference of university department,determining the factors that influence the preferenceorder of new enrolled students,

□choosing a profession according to the demographicand personal characteristics ,

□preventing students from failing and determining thefactors that affect success,

□determine the relationships between the type of schoolin which students graduate and their universitydepartments evaluating the study activities ofdistance education students ,

□determining the profiles and preferences of studentsentering the university entrance exam ,

□determining the relationship between academic successand participation in extracurricular activities ofuniversity students,

□determining the relationships between the socio-economic level of students and the level of academiclearning ,

□determining whether there is a relationship betweenstudent entry scores and school achievement.

These usage areas in educational sector help teachers tomanage their classes, to understand their students' learning, and to provide proactive feedback to learners.

2.2.1.3 Telecommunication Sector

Data Mining can be used in the telecommunication sector:

□to predict mobile user movements in thecommunications sector,

□to determine the future movements of mobile users,

□to detect frauds,

□to reduce much of human-based analysis,

□to determine the factors that influence customers to callmore at certain times,

□to determine user templates for social network usage,

□to identify new prospects using demographic data ,

□to identify the characteristics of customers who needspecial action as suspension or deactivation ,

□to prevent customer loss.

In order to prevent customer loss, telecommunicationorganizations can eliminate this problem by resorting to strategies for developing strategies, low cost and effective campaigns.

A good example of the use of data mining in this sector, regarding customer loss, is Verizon, America's largest wireless communications provider. In the study, Verizon has resorted to data mining methods to identify customers whom they are likely to lose and the factors that cause customer loss.

2.2.1.4 Health Sector

Data mining techniques and application tools are more valuable for health sector. Data mining applications are used extensively to reduce the complexity of the healthcare data transactions’ study in the health sector . Data Mining is used in the health sector:

□to diagnose the disease,

□to determine the treatment method to be applied to thedisease,

□to estimate the resource use and patient numbers inhospitals,

□to set the success of treatment methods applied in thehospital,

□to classify the patient data according to factors such asage, gender, race and treatment,

□to determine the high risk factors in surgeries,

□to prevent corruption in hospital expenditures.

One of the best examples of data mining studies in thehealth field is the one conducted at the San Francisco Heart Institute. In this study, some data such as the patient's history, laboratory data, and other medical data which are obtained from the patients to improve patient outcomes and reduce patient's hospital stay, were converted to information through data mining methods.

2.2.1.5 Public Sector

Data Mining is often used to predict public safety and security problems in the public sector. Data mining techniques offer open opportunities for the public sector to optimize decisions. These decisions are based on general trends extracted from past experience and historical data [25]. Apart from that, it is necessary:

□to determine the tax related corruption,

□to predict the impact of changes in the tax system onthe budget,

□to determine waste and prevent damage caused bywaste, estimate population,

□to forecast the weather, determine new jobopportunities,

□to measure performance of employees, manage thebusiness processes,

□to classify public expenditures, plan the correct use ofresources,

□to forecast the future of public investment, analyze thedata in defense industry,

□to determine which offenders are likely to commitcrimes in terms of safety.

One of the most important examples in this area is E-Government application in Turkey. Faster feedback is being received as data mining methods are used in conjunction with the web page re-arrangement according to the behavior of the user in the past by determining the simultaneous access of the information and the order in which the web pages on E-Government are visited.

2.2.1.6 Construction Sector

Data Mining is used in the construction sector in construction, project management, hydraulics, occupational health and safety applications, analysis of earthquake data, groundwork studies and many other areas. In view of the studies carried out in this context, it has been found that studies have been done:

□to create the information classification scheme inproject documents,

□to determine the tax related corruption,

□to predict the impact of changes in the tax system onthe budget,

□to determine waste and prevent damage caused bywaste,

□to estimate the cost of highway construction ,

□to estimate population, forecast the weather, determinenew job opportunities,

□to estimate the compressive strength of the cementproduct ,

□to measure performance of employees, manage thebusiness processes,

□to define the characteristics of occupational accidents inthe construction industry ,

□to classify public expenditures, plan the correct use ofresources,

□to measure worker productivity ,

□to determine the concrete compressive strength ,

□to forecast the future of public investment, analyze thedata in defense industry,

□to determine the relationship of leadership-motivationbetween the chief and the worker ,

□to determine which offenders are likely to commitcrimes in terms of safety,

□to determine the location of data mining method inconstruction management .

One of the most important examples in this area is E-Government application in Turkey. Faster feedback is being received as data mining methods are used in conjunction with the web page re-arrangement according to the behavior of the user in the past by determining the simultaneous access of the information and the order in which the web pages on E-Government are visited.

2.2.1.7 Engineering and Science Sector

Large quantities of data have been collected from scientific fields such as astronomy, bioinformatics, computing, criminal science, engineering, geosciences, mathematics, software etc. Data Mining provides many benefits in engineering and science sector including:

□managing the process of the software used in firms,

□reducing the amount of tasks,

□increasing the speed of the Software Development LifeCycle,

□saving time and effort,

□providing competitive advantage to the organizationswith the predicted analysis,

□improving the manufacturing process ,

□biological literature analysis,

□remote sensing,

□soil quality analysis ,

□detecting crime pattern,

□real-time feature extraction for turbulent flow analysis,

□obtaining good quality seed ,

□evolving new crop breeds ,

□classifying the astronomical objects,

□ecosystem modeling,

□discovering the relationships for best utilization of thecold storages and use of canal water ,

□classifying the sequences in bioinformatics.

These usage areas in engineering and science sectorhelp users to improve system performance of software used, to provide insight into many parts of used engineering software development processes, and to plan the future decision making process.

Possible Questions

1. Explain scalability in large dataset.
2. Explain major issues in data mining.
3. Explain about Interesting Measures.
4. Explain the application of data mining in education sector.

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established Under Section 3 of UGC Act 1956)
Coimbatore – 641 021

INFORMATION TECHNOLOGY

Fifth Semester
FIRST INTERNAL EXAMINATION – JULY 2019

DATA MINING

Class: **III B.Sc. IT (A&B)** **Duration:** 2 Hours
Date & Session:24.7.2017 **Maximum :** 50
Subject Code : 16ITU503B

PART-A (20 X 1 = 20 Marks)
Answer ALL the Questions

1. Data mining is __________________
a) The actual discovery phase of a knowledge discovery process
b) The stage of selecting the right data for a KDD process
c) A collection of data in support of management
d) The analysis of data

2. The removal of noise and inconsistent data in data mining is called as ______________.
a) Data integration **b) Data Cleaning** c) Data modeling d) Data evalution

3. The process of constructing the final dataset from raw data is called as data _________.
a) Understanding b) Evaluation **c) Preparation** d) Deployment

4. Which of the following is not involve in data mining?
a) Knowledge extraction b)Data archaeology
c) Data exploration **d) Data transformation**

5. The study of frequency of items occuring together in transactional databases is called _____________.
a) Association b) Clustering c) Classification d) Regression

6. The principle of maximizing the similarity between objects in a same class is called as ________________ similarity.
a) inter b) data c) class **d) intra**

7. The process of analyzing the current and past states of the attributes and prediction of its future state is called as ____________.

a) Descriptive **b) Predictive** c) Pattern d) Segmentation

8. Classification is ______________ type of learning.
   a) Unsupervised b) Relationship **c) Supervised** d) Domain

9. The process of inspecting, cleaning and modelling data with the objective of discovery useful information is known as __________.
   a) Data Project b) Data Collection **c) Data Analysis** d) Statistics

10. A statistical methodology that is most often used for numeric prediction is called as ________________.
    a) Statistical **b) Regression** c) Training set d) Class

11. Task of inferring a model from labeled training data is called _______________ learning.
    a) Unsupervised **b) Supervised** c) Reinforcement d) Machine

12. A subdivision of a set of examples into a number of classes is called as ____________.
    a) Time series b) Summarization c) Prediction **d) Classification**

13. The classification is according to the data handled is called as classification under ___________
    a) Type of data b) Database c) Knoweldge d) Techniques

14. Among the following which is not a phase in life cycle of data mining.
    a) Data understanding b) Data preparation **c) Correlation** d) Evaluation
15. The ____________is a summarization of the general characteristics or features of a target class of data.
    a) Characterization b) Classification c) Discrimination d) Selection

16. Strategic value of data mining is ______________.
    a) Cost - sensitive b) Work - sensitive
    c) Time - sensitive d) Technical sensitive

17. The full form of KDD is _________.
    a) Knowledge Database **b) Knowledge Discovery Database**
    c) Knowledge Data House d) Knowledge Data Definition
18. The output of KDD is ____________.
    a) Data **b) Information** c) Query d) Labels
19. The ____________ is an essential process where intelligent methods are applied to extract data patterns.
    a) Data warehousing **b) Data mining** c) Text mining d) Data selection
20. Data mining can also applied to other forms such as ________________.
    a) Data streams b) Sequence data c) Networked data d) Spatial data

PART-B (3 X 2 = 6 Marks)
(Answer ALL the Questions)

21. Define Data mining. What are the other names of data mining?

Data Mining refers to extracting or "mining" knowledge from large amounts of data. The overall goals of the data mining process is extraction of information from large data sets and transform it into some understandable structure for further uses. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Different names of data mining are knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing.

22. Write a short note on classification techniques.

Classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the manager of a store could analyze the customers' behavior vis-à-vis their credit, and label accordingly the customers who received credits with three possible labels "safe", "risky" and "very risky". The classification analysis would generate a model that could be used to either accept or reject credit requests in the future.

23. Define supervised and unsupervised learning in data mining.

Definition for Supervised data mining:

Supervised learning is the Data mining task of inferring a function from **labeled training data**. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the**supervisory signal**). A **supervised learning algorithm** analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for **unseen instances**. This requires the learning algorithm to generalize from the training data to unseen situations in a “reasonable” way.

Definition for UnSupervised data mining:

In Data mining, the problem of **unsupervised learning** is that of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. The goal is rather to establish some relationship among all the variables in the data.

PART-C (3 X 8 = 24 Marks)
(Answer ALL the Questions)

24. a) Explain about the classification and life cycle of data mining.

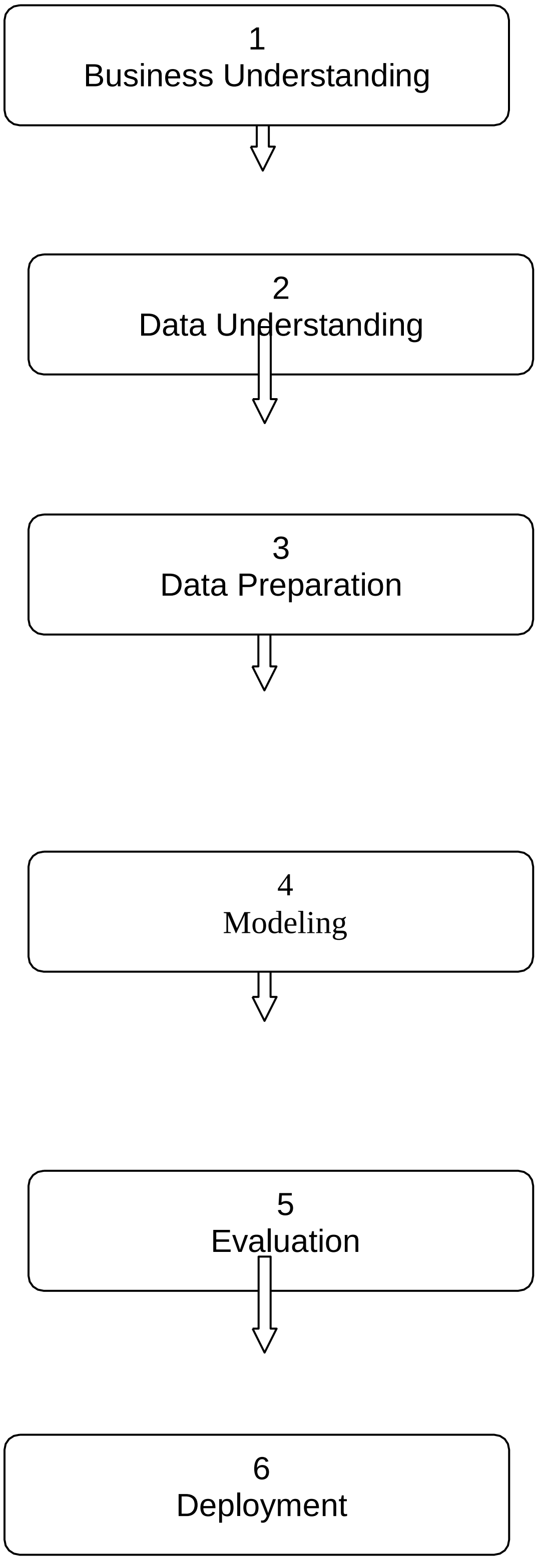
Data Mining Life cycle

The life cycle of a data mining project consists of six phases [91, 26]. The sequence of the phases is not rigid. Moving back and forth between different phases is always required depending upon the outcome of each phase. The main phases are:

Business Understanding:

This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Figure 3.2: Phases of Data Mining Life Cycle


1
Business Understanding
2
Data Understanding
3
Data Preparation
4
Modeling
5
Evaluation
6
Deployment

Data Understanding:

It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Data Preparation:

Covers all activities to construct the final dataset from raw data.

Modeling:

In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

Evaluation:

In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business

Deployment:

The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

b) Describe briefly about the functionalities of data mining.

Data Mining Functionalities

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data. The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list:

Characterization:

It is the summarization of general features of objects in a target class, and produces what is called characteristic rules. The data relevant to a user specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions. For example, one may wish to characterize the customers of a store who regularly rent more than 30 movies a year. With concept hierarchies on the attributes describing the target class, the attribute oriented induction method can be used to carry out data summarization. With a data cube containing summarization of data, simple OLAP operations fit the purpose of data characterization.

Discrimination:

Data discrimination produces what are called discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. For example, one may wish to compare the general characteristics of the customers who rented more than 30 movies in the last year with those whose rental account is lower than 5. The techniques used for data discrimination are similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures.

Association analysis:

Association analysis studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. This is commonly used for market basket analysis. For example, it could be useful for the manager to know what movies are often rented together or if there is a relationship between renting a certain type of movies and buying popcorn or pop. The discovered association rules are of the form:

$$P \rightarrow Q \ [s, c],$$

where P and Q are conjunctions of attribute value-pairs, and s (support) is the probability that P and Q appear together in a transaction and c (confidence) is the conditional probability that Q appears in a transaction when P is present. For example, RentType(X,“game”)∧ Age(X,“13-19”)_Buys(X,“pop”)[s=2%, =55%]

The above rule would indicate that 2% of the transactions considered are of customers aged between 13 and 19 who are renting a game and buying a pop, and that there is a certainty of 55% that teenage customers who rent a game also buy pop.

Classification:

It is the organization of data in given classes. Classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the manager of a store could analyze the customers’ behavior vis-à-vis their credit, and label accordingly the customers who received credits with three possible labels “safe”, “risky” and “very risky”. The classification analysis would generate a model that could be used to either accept or reject credit requests in the future.

Prediction:

Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values.

Clustering:

Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).

Outlier analysis:

Outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable.

Evolution and deviation analysis:

Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values. It is common that users do not have a clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore important to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction. This also makes interactivity an important attribute of a data mining system.

25. a) Briefly explain the techniques followed in Descriptive data mining model.

Descriptive Models

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Summarization, Association rule, Sequence discovery etc. Clustering is similar to classification except that the groups are not predefined, but are defined by the data alone. It is also referred to as unsupervised learning or segmentation. It is the partitioning or segmentation of the data in to groups or clusters. The clusters are defined by studying the behavior of the data by the domain experts. The term segmentation is used in very specific context; it is a process of partitioning of database into disjoint grouping of similar tuples. Summarization is the technique of presenting the summarize information from the data. The association rule finds the association between the different attributes. Association rule mining is a two step process: Finding all frequent item sets, Generating strong association rules from the frequent item sets. Sequence discovery is a process of finding the sequence patterns in data. This sequence can be used to understand the trend.

A. Clustering

Clustering is the most important descriptive data mining technique, in which a set of data items is partitioned into a set of classes also called as groups. Clustering is the process of finding natural groups, called as clusters, in a database. The set of data items in a group have similar characteristics. Hence, it is mainly used for finding groups of similar items. It is the identification process for classes for a set of objects whose classes are not known. These objects ware so clustered that their intra class similarities can be maximized and minimized on the criteria defined by attributes of objects. The object is label with their corresponding clusters once their particular class or clusters are decided. For example, between a data set of customers, that have a similar buying behavior can be identified with subgroups of customers.

B. Summarization

Summarization is called as the abstraction or generalization of data. The summarization technique maps data into subsets with simple descriptions. The summarized data set gives general overview of the data with aggregated information. Summarization can scale up to different levels of abstraction and can be viewed from different angles. It is a key data-mining concept involving techniques for finding a compact description of dataset. Summarization approaches are Basically Mean, Standard Deviation, Variance, tabulating, mode, and median. These approaches are often applied for data analysis, data visualization and automated report generation.

Consider example of long distance call, Costumer data related to call can be summarize as total minutes, total spending, total calls, etc. Such important information can be presented to sales manager for analysis of his costumer and his business. The scaling and viewing from different angles can be done for this example as, calling minutes and spending can be totaled along, its period in weeks, months or years. In the same way long distance, call can be summarized as state call, state-to-state call, national, Asia calls, America calls, etc. Further summarized as Domestic and international calls.

C. Association

The Associations or Link Analysis technique are used to discover relationships between attributes and items. In these techniques, the presence of one pattern implies the presence of another pattern i.e. item is related to another in terms of cause-and-effect. This is common in establishing a form of statistical relationships among different interdependent variables of a model. For market basket analysis, Association Rules is a important technique because all possible combinations of potentially interesting product groupings can be explored [11]. Association rules are as if/then statements that help uncover relationships between seemingly unrelated data in a transaction oriented database, relational database or other information repository. In data mining, association rules are useful for analyzing and predicting customer behavior. They also play an important role in shopping basket data analysis, product clustering, catalog design and store layout. This association rules are also build by programmers use to build programs capable of machine learning.

D. Feature Extraction

Feature Extraction extract a new set of features by decomposing the original set of data. This technique describes the data with a number of features far smaller than the number of original attributes. The word feature in the technique are combination of attributes in the data that have special important characteristics of the data [12]. Feature extraction is mostly applicable to latent semantic analysis, data compression, data decomposition and projection, and pattern recognition, etc. Using feature extraction process the speed and effectiveness of supervised learning can also be improved.

For example, feature extraction is used to extract the themes/features of a document collection, where documents are represented by a set of keywords and their frequencies. Each feature is represented by a combination of keywords. The documents can then be expressed from the collection in terms of the discovered themes.

b) Explain the applications of analytic process in data mining.

Predictive analytics

Predictive analytics is an area of statistics that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs.

Predictive Analytics Process

1. **Define Project** : Define the project outcomes, deliverable, scope of the effort, business objectives, identify the data sets that are going to be used.
2. **Data Collection** : Data mining for predictive analytics prepares data from multiple sources for analysis. This provides a complete view of customer interactions.

3. **Data Analysis** : Data Analysis is the process of inspecting, cleaning and modelling data with the objective of discovering useful information, arriving at conclusion
4. **Statistics** : Statistical Analysis enables to validate the assumptions, hypothesis and test them using standard statistical models.
5. **Modelling** : Predictive modelling provides the ability to automatically create accurate predictive models about future. There are also options to choose the best solution with multi-modal evaluation.
6. **Deployment** : Predictive model deployment provides the option to deploy the analytical results into everyday decision making process to get results, reports and output by automating the decisions based on the modelling.
7. **Model Monitoring** : Models are managed and monitored to review the model performance to ensure that it is providing the results expected.

In the current literature, concepts ”knowledge discovery”, ”data mining” and ”machine learning” are often used interchangeably. Sometimes the whole KDD process is called data mining or machine learning, or machine learning is considered as a subdiscipline of data mining. To avoid confusion, we follow a systematic division to descriptive and predictive modelling, which matches well with the classical ideas of data mining and machine learning1 . This approach has several advantages:

1. The descriptive and predictive models can often be paired as illustrated in Table 2.1. Thus, descriptive models indicate the suitability of a given predictive model and can guide the search of models (i.e. they help in the selection of the modelling paradigm and the model structure).

2. Descriptive and predictive modelling require different validation techniques. Thus the division gives guidelines, how to validate the results.

3. Descriptive and predictive models have different underlying assumptions (bias) about good models. This is reflected especially by the score functions, which guide the search.

26. a) Write briefly about supervised and unsupervised machine learning algorithms.

Supervised and Unsupervised Machine Learning Algorithms

Supervised Machine Learning

The majority of practical machine learning uses supervised learning.

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

Supervised learning problems can be further grouped into regression and classification problems.

- **Classification**: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
- **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively.

Some popular examples of supervised machine learning algorithms are:

- Linear regression for regression problems.
- Random forest for classification and regression problems.
- Support vector machines for classification problems.

Unsupervised Machine Learning

Unsupervised learning is where you only have input data (X) and no corresponding output variables.

The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

Unsupervised learning problems can be further grouped into clustering and association problems.

- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Some popular examples of unsupervised learning algorithms are:

- k-means for clustering problems.
- Apriori algorithm for association rule learning problems.

Semi-Supervised Machine Learning

Problems where you have a large amount of input data (X) and only some of the data is labeled (Y) are called semi-supervised learning problems.

These problems sit in between both supervised and unsupervised learning.

A good example is a photo archive where only some of the images are labeled, (e.g. dog, cat, person) and the majority are unlabeled.

Many real world machine learning problems fall into this area. This is because it can be expensive or time-consuming to label data as it may require access to domain experts. Whereas unlabeled data is cheap and easy to collect and store.

You can use unsupervised learning techniques to discover and learn the structure in the input variables.

You can also use supervised learning techniques to make best guess predictions for the unlabeled data, feed that data back into the supervised learning algorithm as training data and use the model to make predictions on new unseen data.

Summary

In this post you learned the difference between supervised, unsupervised and semi-supervised learning. You now know that:

Supervised: All data is labeled and the algorithms learn to predict the output from the input data.

Unsupervised: All data is unlabeled and the algorithms learn to inherent structure from the input data.

Semi-supervised: Some data is labeled but most of it is unlabeled and a mixture of supervised and unsupervised techniques can be used.

Do you have any questions about supervised, unsupervised or semi-supervised learning? Leave a comment and ask your question and I will do my best to answer it.

b) Illustrate the major issues in supervised learning methods.

Data mining techniques come in two main forms: supervised (also known as predictive or directed) and unsupervised (also known as descriptive or undirected). Both categories encompass functions capable of finding different hidden patterns in large data sets.

Data analytics tools are placing more emphasis on self-service, it’s still useful to know which data mining operation is appropriate for your needs before you begin a data mining operation.

Supervised Data Mining

Supervised data mining techniques are appropriate when you have a specific target value you’d like to predict about your data. The targets can have two or more possible outcomes, or even be a continuous numeric value (more on that later).

To use these methods, you ideally have a subset of data points for which this target value is already known. You use that data to build a model of what a typical data point looks like when it has one of the various target values. You then apply that model to data for which that target value is currently unknown. The algorithm identifies the “*new*” data points that match the model of each target value.

Classification

As a supervised data mining method, classification begins with the method.

For example: Imagine you’re a credit card company and you want to know which customers are likely to default on their payments in the next few years.

You use the data on customers who have and have not defaulted for extended periods of time as build data (or training data) to generate a classification model. You then run that model on the customers you’re curious about. The algorithms will look for customers whose attributes match the attribute patterns of previous defaulters/non-defaulters, and categorize them according to which group they most closely match. You can then use these groupings as indicators of which customers are most likely to default.

Similarly, a classification model can have more than two possible values in the target attribute. The values could be anything from the shirt colors they're most likely to buy, the promotional methods they'll respond to (mail, email, phone), or whether or not they'll use a coupon.

Regression

Regression is similar to classification except that the targeted attribute's values are numeric, rather than categorical. The order or magnitude of the value is significant in some way.

To reuse the credit card example, if you wanted to know what threshold of debt new customers are likely to accumulate on their credit card, you would use a regression model.

Simply supply data from current and past customers with their maximum previous debt level as the target value, and a regression model will be built on that training data. Once run on the new customers, the regression model will match attribute values with predicted maximum debt levels and assign the predictions to each customer accordingly.

This could be used to predict the age of customers with demographic and purchasing data, or to predict the frequency of insurance claims.

Anomaly Detection

Anomaly detection identifies data points atypical of a given distribution. In other words, it finds the outliers. Though simpler data analysis techniques than full-scale data mining can identify outliers, data mining anomaly detection techniques identify much more subtle attribute patterns and the data points that fail to conform to those patterns.

Most examples of anomaly detection uses involve fraud detection, such as for insurance or credit card companies.

Reg.No __________________

[17ITU503B]

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established Under Section 3 of UGC Act 1956)
Coimbatore – 641 021

INFORMATION TECHNOLOGY

Fifth Semester
SECOND INTERNAL EXAMINATION – AUGUST 2019

DATA MINING

Class: III B.Sc. IT **Duration :** 2 Hours
Date & Session : .8.2017 **Maximum :**50 Marks

PART-A (20 * 1 = 20 Marks)
Answer ALL the Questions

1. Regression is a ________ based algorithm
   a) **Statistical** b) Rule c) Dynamic Support system d) none
2. Classification frequently performed by simply applying ________
   a) Data b) Information c) **Knowledge of the data** d) Report
3. The__________ approximates discrete multidimensional probability distributions
   a) linear regression b) multiple regression c) **log-linear regression** d) aggregation
4. A binary variable is _______ if the outcomes of the states are not equally important
   a) symmetric b) **asymmetric** c) different d) none
5. Which algorithm comes under Hierarchical methods ________
   a) K-means b) K-mediods c) **Categorical** d) Ordinal
6. A __________ is a heuristic for selecting the splitting criterion that best separates a given data partition
   a) partition rule b) decision rule c) **ID3** d) Neural network
7. The cost complexity pruning algorithm used in _________
   a) **CART** b) ID3 c) Triangular d) Rectangle
8. The _________ data is mostly dependent on the external resources for a high percentage of information.
   a) production b) Internal c) Archived d) **External**
9. The _________ function is used to extract information from other data sources.
   a) **Data Extraction** b) Data Transformation c) Data loading d) None
10. In expert systems the knowledge of domain is represented in terms of ________.
    a) **If-then rules** b) Planners c) Samples d) Charts
11. The ___________ is used exclusively for the discovery stage of the KDD process.
    a) OLAP b) Cleaning c) Enriching d) **Data mining**
12. If you know exactly what you are looking for the use ____________.
    a) genetic algorithm b) **SQL** c) neural network d) clustering
13. Data mart is ____________.
    a) **departmental** b) corporate c) organized d) none

14. There are _____________ types of spatial data cube.
a) one b) two c) **three** d) four

15. A ______________ contains only numeric data.
a) spatial measure b) **numeric measure** c) text measure d) none

16. A ________________ contains a collection of pointers to spatial objects.
a) **spatial measure** b) numeric measure c) text measure d) measure

17. BIRCH is Balanced iterative reducing and ___ using hierarchies.
a) Combining b) **Clustering** c) Correlative d) Conclusive

18. Algorithm used in database design as to how to physically place data on disk is ________.
a) Cluster b) BOND c) **Bond energy algorithm** d) Classification

19. Iterative merging of items into existing clusters that are closed is __________ algorithm.
a) **Nearest neighbor** b) Spanning tree c) Decision tree d)Neural network

20. Association rules are frequently used by ______________.
a) Medical representative b) **Retail store** c) Bank d) Manufacturer

PART-B (3 * 2 = 6 Marks) (Answer ALL the Questions)

21. Mention the unsupervised machine learning algorithms.
22. Define KDD Process.
23. What are the issues in Data mining?

PART-C (3 * 8 = 24 Marks) (Answer ALL the Questions)

24. a) Explain about the supervised learning method with a real time example.
b) Write about the aspects of data mining.
25. a) Briefly explain about the steps in Preprocessing technique.
b) Explain about the application Data Cleaning.
26. a) Write briefly about Data Reduction in preprocessing.
b) Illustrate the concept of Discretization and concept hierarchies.

Reg.No __________________

[17ITU503B]

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)
(Established Under Section 3 of UGC Act 1956)
Coimbatore – 641 021

INFORMATION TECHNOLOGY

Fifth Semester
THIRD INTERNAL EXAMINATION – OCTOBER 2019

DATA MINING

Class: III B.Sc. IT **Duration : **2 Hours
Date & Session :06.10.2018 &FN **Maximum :** 50 Marks

PART-A (20 * 1 = 20 Marks)
Answer ALL the Questions

1. The object oriented data modl inherits the essential concepts of _____________ database.
a) Multimedia **b) Object oriented** c) Spatial d) None
2. The ____________ gives the relationship and patterns between data elements.
a) KDD b) Data c) Objects d) Mining
3. The _______ is used to describe the whole process of extraction of knowledge from data.
a) Data **b) Datamining** c) KDD d) Algorithm
4. The ___________ contains the visual map of data sets.
a) **Kohonen map** b) Neural network c) K-nearest d) KNN
5. A __________ network not only has input and output nodes but also hidden nodes.
a) Genetic algorithm **b) Back propagation** c) KNN d) Neural
6. In _______ information can be easily retrived from the database using query tools.
a) **Shallow** b) Multidimensional c) Hidden d) Deep
7. The __________ is a interaction between technology and nature.
a) **Genetic algorithm** b) Neural network c) K-nearest d) None
8. A ____________ consists of a simple three layered network.
a) Responders b) photoreceptors c) perceptrons d) **None**
9. Data can be classified into _____________ categories.
a) **2** b) 3 c) 4 d) 5
10. A respository of subjectively, selected and adapted operational data __________.
a) **Warehouse** b) Mining c) Information d) Record
11. Online Transaction ______________.
a) **Processing** b) Procedure c) Publicity d) Production
12. Chief model for a multidimensional data warehouse is _____________.
a) **Star** b) Cyclic c) Snowflake d) Network

13. Aggregation of data at different levels of hierarchies in a given dimension _____________.
    a) **Snow flake** b) Star c) Network d) Cyclic

14. The __________ is an open OLAP that uses third party software tools.
    a) **Power play** b) Power cubes c) Power stake d) Pointer cube

15. The __________ is data about data.
    a) **Meta data** b) Information c) Tuples d) Warehouse

16. The _______ table is a large control table in a dimensional design that has a multipart key.
    a) **Fact** b) Hierarchical c) dimensional d) Relational

17. The ________ operation system software that can be used to host a data warehouse.
    a) **UNIX** b) LINUX c) Android d) Java

18. The ____________ data base product specifically optimized for data warehouse.
    a) **Red warehouse** b) Sybase c) Oracle d) DBMS

19. GISTNIC is General Information services _________ Informatics centre.
    a) National b) Tamilnadu c) Transaction **d) Terminology**

20. World bank developed live database in 1995.
    a) **LDB** b) OLAP c) OLTP d) GISTNIC

PART-B (3 * 2 = 6 Marks)
(Answer ALL the Questions)

21. Write about the Classification problem in data mining.

Classification Problem

- Given a database D={t1,t2,…,tn} and a set of classes C={C1,…,Cm}, the *Classification Problem* is to define a mapping f:DgC where each ti is assigned to one class.
- Actually divides D into *equivalence classes*.
- *Prediction*is similar, but may be viewed as having infinite number of classes.

Classification Examples

- Teachers classify students’ grades as A, B, C, D, or F.
- Identify mushrooms as poisonous or edible.
- Predict when a river will flood.
- Identify individuals with credit risks.
- Speech recognition
- Pattern recognition

22. What is Regression?

Regression

Regression is a data mining technique used to predict a range of numeric values (also called *continuous values*), given a particular dataset. For example, regression might be used to predict the cost of a product or service, given other variables.

Regression is used across multiple industries for business and marketing planning, financial forecasting, environmental modeling and analysis of trends.

- Assume data fits a predefined function
- Determine best values for *regression coefficients* c0,c1,…,cn.
- Assume an error: y = c0+c1x1+…+cnxn+e.

23. Define Statistica Data Miner.

Statistica Data Miner uses a number of technologies specifically developed to optimize processing large data sets, and it is designed to handle even largest scale computational problems based on very large databases.

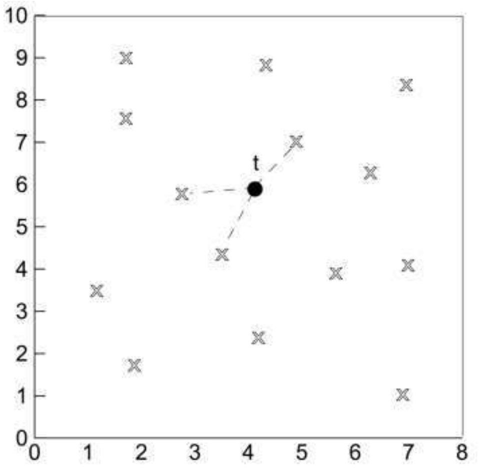
Statistica Data Miner contains nodes in the Data Cleaning and Filtering folder to draw a random sample from the original input data (database connection).

PART-C (3 * 8 = 24 Marks)
(Answer ALL the Questions)

24. a) **Explain about the K Nearest Neighbor (KNN) algorithm in classification.**

K Nearest Neighbor (KNN):

- Training set includes classes. Each member with a label of a class. Training data = model.
- Compare new item with all members in training set for distance. Examine K items nearest item to be considered further.
- New item placed in class with the most number of nearest items (among K) belongs.
- O(q) for each tuple to be classified. (Here q is the size of the training set.)



```
Input:
    $D$        //Training data
    $K$        //Number of neighbors
    $t$        //Input tuple to classify
Output:
    $c$        //Class to which t is assigned
KNN Algorithm:
               //Algorithm to classify tuple using KNN
    $N = \emptyset$;
                   //Find set of neighbors, N, for t
    foreach $d \in D$ do
        if $| N | \leq K$ then
            $N = N \cup d$;
        else
            if $\exists\ u \in N$ such that $sim(t, u) \geq sim(t, d)$ then
                begin
                    $N = N - u$;
                    $N = N \cup d$;
                end
                   //Find class for classification
    c = class to which the most $u \in N$ are classified;
```

\

(or)

b) Illustrate the concept of classification using Decision tree.

Classification Using Decision Trees

- *Partitioning based:* Divide search space into rectangular regions.
- Tuple placed into class based on the region within which it falls.
- DT approaches differ in how the tree is built: *DT Induction*
- Internal nodes associated with attribute and arcs with values for that attribute.
- Algorithms: ID3, C4.5, CART

Given:
– D = {t1, …, tn} where ti=<ti1, …, tih>

– Database schema contains {A1, A2, …, Ah}

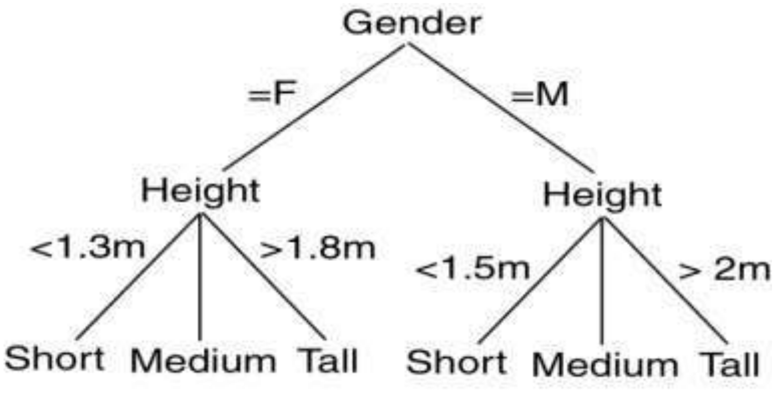
– Classes C={C1, …., Cm}

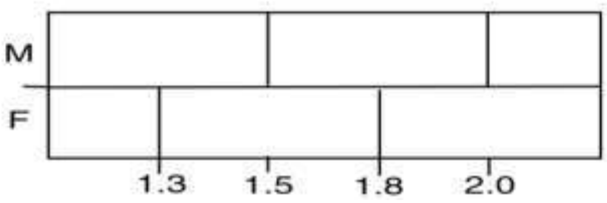
Decision or Classification Tree is a tree associated with D such that
– Each internal node is labeled with attribute, Ai

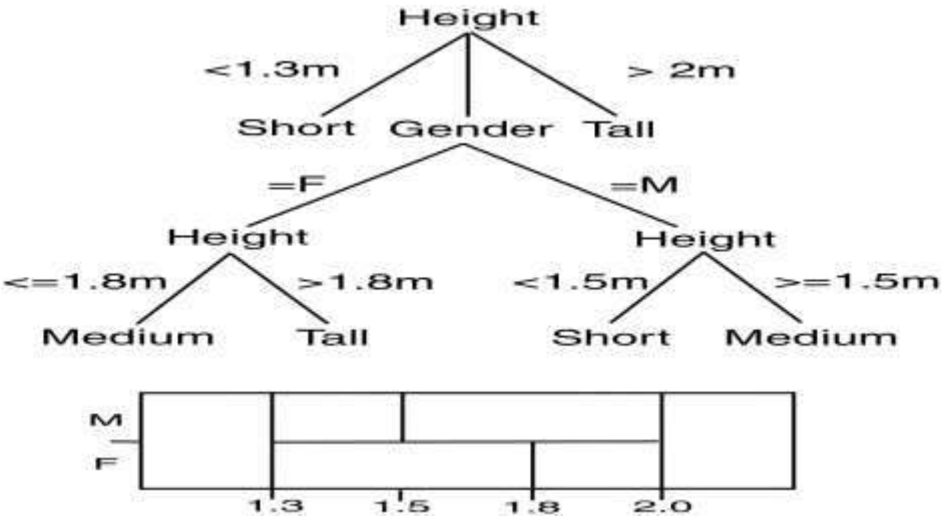
– Each arc is labeled with predicate which can be applied to attribute at parent

– Each leaf node is labeled with a class, Cj

Comparing DTs







Algorithm

```
Input:
    $D$        //Training data
Output:
    $T$        //Decision Tree
DTBuild Algorithm:
            //Simplistic algorithm to illustrate naive approach to building DT
    $T = \emptyset$;
    Determine best splitting criterion;
    $T$ = Create root node node and label with splitting attribute;
    $T$ = Add arc to root node for each split predicate and label;
    for each arc do
        $D$ = Database created by applying splitting predicate to $D$;
        if stopping point reached for this path then
            $T'$ = Create leaf node and label with appropriate class;
        else
            $T' = DTBuild(D)$;
        $T$ = Add $T'$ to arc;
```

25. a) **Briefly explain about the Hierarchical clustering algorithm.**

In **data mining** and statistics, **hierarchical clustering** (also called **hierarchical cluster** analysis or HCA) is a method of **cluster** analysis which seeks to build a**hierarchy** of **clusters**.

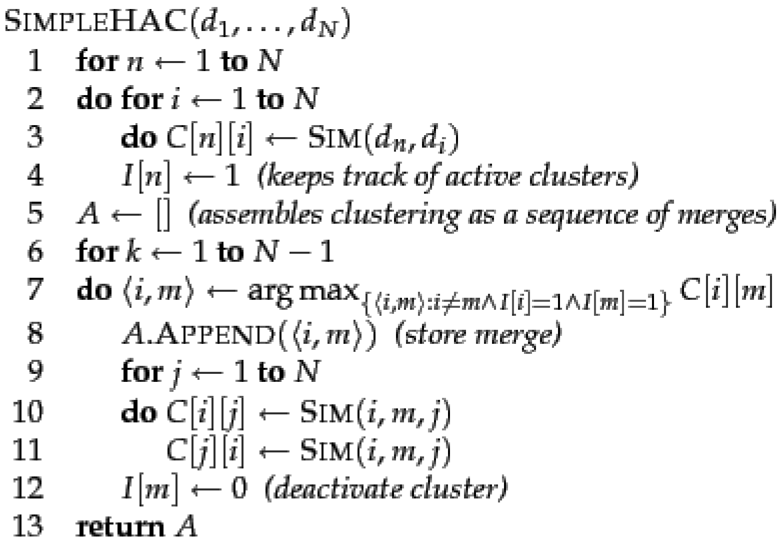
One algorithm is designed for determining the partial order on a given set of nominal attributes. The resulting partial order is a **useful** guide for users to finalize the**concept hierarchy** for their particular **data mining** tasks. ... An encoding method is presented and its properties are studies.

Hierarchical agglomerative clustering

Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or *agglomerate*) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called *hierarchical agglomerative clustering* or *HAC* . Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual documents are reached.

	A	**B**	**C**	**D**	**E**
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

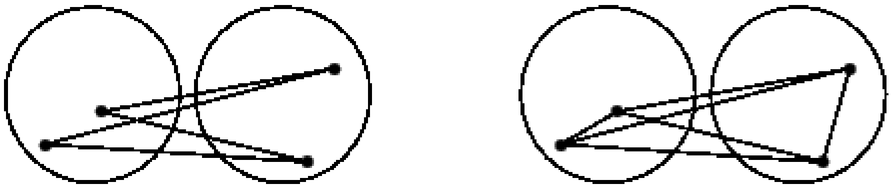
- we can also prespecify the number of clusters $\overset{K}{\_}$ and select the cutting point that produces $\overset{K}{\_}$ clusters.



```
SimpleHAC($d_1, \ldots, d_N$)
for $n \leftarrow 1$ to $N$
do for $i \leftarrow 1$ to $N$
    do $C[n][i] \leftarrow \text{Sim}(d_n, d_i)$
    $I[n] \leftarrow 1$  (keeps track of active clusters)
$A \leftarrow []$  (assembles clustering as a sequence of merges)
for $k \leftarrow 1$ to $N-1$
do $\langle i, m \rangle \leftarrow \arg\max_{\{\langle i,m \rangle : i \neq m \wedge I[i]=1 \wedge I[m]=1\}} C[i][m]$
    $A$.Append($\langle i, m \rangle$)  (store merge)
    for $j \leftarrow 1$ to $N$
    do $C[i][j] \leftarrow \text{Sim}(i, m, j)$
        $C[j][i] \leftarrow \text{Sim}(i, m, j)$
    $I[m] \leftarrow 0$  (deactivate cluster)
return $A$
```

Figure 17.2: A simple, but inefficient HAC algorithm.

(a) single-link: maximum similarity (b) complete-link: minimum similarity

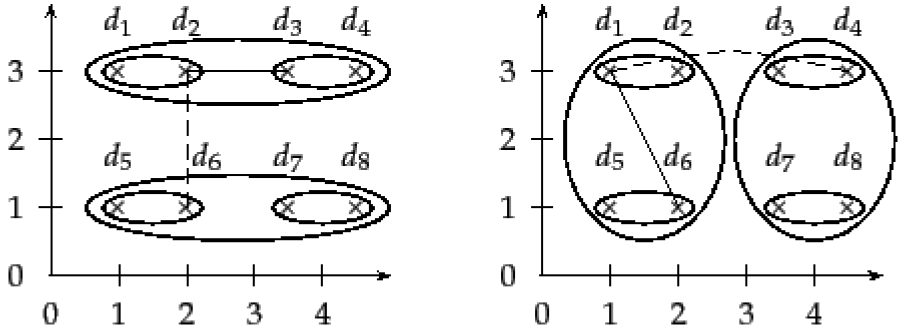


(c) centroid: average inter-similarity (d) group-average: average of all similarities

▶ **Figure 17.1** The different notions of cluster similarity used by the four HAC algorithms. An *inter-similarity* is a similarity between two documents from different clusters.

A simple, naive HAC algorithm is shown in Figure 17.2 . We first compute the $N \times N$ similarity matrix C. The algorithm then executes $N-1$ steps of merging the currently most similar clusters. In each iteration, the two most similar clusters are merged and the rows and columns of the merged cluster i in C are updated. The clustering is stored as a list of merges in A. I indicates which clusters are still available to be merged. The function SIM (i,m,j) computes the similarity of cluster j with the merge of clusters i and m. For some HAC algorithms, SIM (i,m,j) is simply a function of $C[j][i]$ and $C[j][m]$, for example, the maximum of these two values for single-link.

We will now refine this algorithm for the different similarity measures of single-link and complete-link clustering and group-average and centroid clustering. The merge criteria of these four variants of HAC are shown in Figure 17.3 .



► **Figure 17.2** A single-link (left) and complete-link (right) clustering of eight documents. The ellipses correspond to successive clustering stages. Left: The single-link similarity of the two upper two-point clusters is the similarity of d_2 and d_3 (solid line), which is greater than the single-link similarity of the two left two-point clusters (dashed line). Right: The complete-link similarity of the two upper two-point clusters is the similarity of d_1 and d_4 (dashed line), which is smaller than the complete-link similarity of the two left two-point clusters (solid line).

(or)

b) Describe about Apiori algorithm.

The **Apriori Algorithm** is an influential **algorithm** for **mining** frequent itemsets for boolean association rules. • **Apriori** uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation, and groups of candidates are tested against the **data**.

Apriori is an **algorithm** for frequent item set **mining** and **association rule** learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

Example

Assume that a large supermarket tracks sales data by stock-keeping unit (SKU) for each item: each item, such as "butter" or "bread", is identified by a numerical SKU. The supermarket has a database of transactions where each transaction is a set of SKUs that were bought together.

Let the database of transactions consist of following itemsets:

Itemsets
{1,2,3,4}

{1,2,4}
{1,2}
{2,3,4}
{2,3}
{3,4}
{2,4}

We will use Apriori to determine the frequent item sets of this database. To do this, we will say that an item set is frequent if it appears in at least 3 transactions of the database: the value 3 is the *support threshold*.

The first step of Apriori is to count up the number of occurrences, called the support, of each member item separately. By scanning the database for the first time, we obtain the following result

Item	**Support**
{1}	3
{2}	6
{3}	4
{4}	5

All the itemsets of size 1 have a support of at least 3, so they are all frequent.

26. a) **Write briefly about Mining Methodology and User Interaction Issues.**

Mining Methodology and User Interaction Issues

It refers to the following kind of issues:

- **Mining different kinds of knowledge in databases.** - The need of different users is not the same. And Different user may be in interested in different kind of knowledge. Therefore it is necessary for data mining to cover 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 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 level 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, visual representations. This representations should be easily understandable by the users.
- **Handling noisy or incomplete data.** - The data cleaning methods are required that can handle the noise, incomplete objects while mining the data regularities. If data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation.** - It refers to interestingness of the problem. The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

It refers to the following issues:

- **Efficiency and scalability of data mining algorithms.** - In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

Parallel, distributed, and incremental mining algorithms. - The factors such as huge size of databases, wide distribution of data,and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithm divide the data into partitions which is further processed parallel. Then the results from the partitions is merged. The incremental algorithms, updates databases without having mine the data again from scratch.

(or)

b) Discuss about the application of Data Mining in Education sector.

Education Sector

It has been observed that data mining has been used in many studies in the education sector including:

[redacted]

factors affecting the success of the students enrolling atthe university,

[redacted]

preferenceorder of new enrolled students,

phicand personal characteristics ,

universitydepartments evaluating the study activities ofdistance education students ,

entrance exam ,

[redacted]

activities ofuniversity students,

-economic level of students and the level of academiclearning ,

[redacted]termining whether there is a relationship betweenstudent entry scores and school achievement.

These usage areas in educational sector help teachers tomanage their classes, to understand their students' learning, and to provide proactive feedback to learners.