Data Mining –Introduction

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

Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.

What is Data Mining?

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

Functionalities of Data Mining

Data mining functions are used to define the trends or correlations contained in data mining activities.
In comparison, data mining **activities** can be divided into 2 categories:

1. **Descriptive Data Mining:**
   It includes certain knowledge to understand what is happening within the data without a previous idea. The common data features are highlighted in the data set. For examples: count, average etc.
2. **Predictive Data Mining:**
   It helps developers to provide unlabeled definitions of attributes. Based on previous tests, the software estimates the characteristics that are absent.
   For example: Judging from the findings of a patient's medical examinations that is he suffering from any particular disease.

Data Mining Functionality:

1. Class/Concept Descriptions:

Classes or definitions can be correlated with results. In simplified, descriptive and yet accurate ways, it can be helpful to define individual groups and concepts.
These class or concept definitions are referred to as class/concept descriptions.

- **Data Characterization:**
  This refers to the summary of general characteristics or features of the class that is under the study. For example. To study the characteristics of a software product whose sales increased by 15% two years ago, anyone can collect these type of data related to such products by running SQL queries.
- **Data Discrimination:**
  It compares common features of class which is under study. The output of this process can be represented in many forms. Eg., bar charts, curves and pie charts.

2. Mining Frequent Patterns, Associations, and Correlations:

Frequent patterns are nothing but things that are found to be most common in the data. There are different kinds of frequency that can be observed in the dataset.

- **Frequent item set:**
  This applies to a number of items that can be seen together regularly for eg: milk and sugar.
- **Frequent Subsequence:**
  This refers to the pattern series that often occurs regularly such as purchasing a phone followed by a back cover.
- **Frequent Substructure:**
  It refers to the different kinds of data structures such as trees and graphs that may be combined with the itemset or subsequence.

Association Analysis:

The process involves uncovering the relationship between data and deciding the rules of the association. It is a way of discovering the relationship between various items. for example, it can be used to determine the sales of items that are frequently purchased together.

Correlation Analysis:

Correlation is a mathematical technique that can show whether and how strongly the pairs

of attributes are related to each other. For example, Heighted people tend to have more weight.

Classification of data mining

The information or knowledge extracted so can be used for any of the following applications −

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

Data Mining Applications

Data mining is highly useful in the following domains −

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid

Market Analysis and Management

Listed below are the various fields of market where data mining is used −

- **Customer Profiling** − Data mining helps determine what kind of people buy what kind of products.
- **Identifying Customer Requirements** − Data mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers.
- **Cross Market Analysis** − Data mining performs Association/correlations between product sales.
- **Target Marketing** − Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.
- **Determining Customer purchasing pattern** − Data mining helps in determining customer purchasing pattern.
- **Providing Summary Information** − Data mining provides us various multidimensional summary reports.

Corporate Analysis and Risk Management

Data mining is used in the following fields of the Corporate Sector −

- **Finance Planning and Asset Evaluation** − It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.
- **Resource Planning** − It involves summarizing and comparing the resources and spending.
- **Competition** − It involves monitoring competitors and market directions.

Fraud Detection

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms

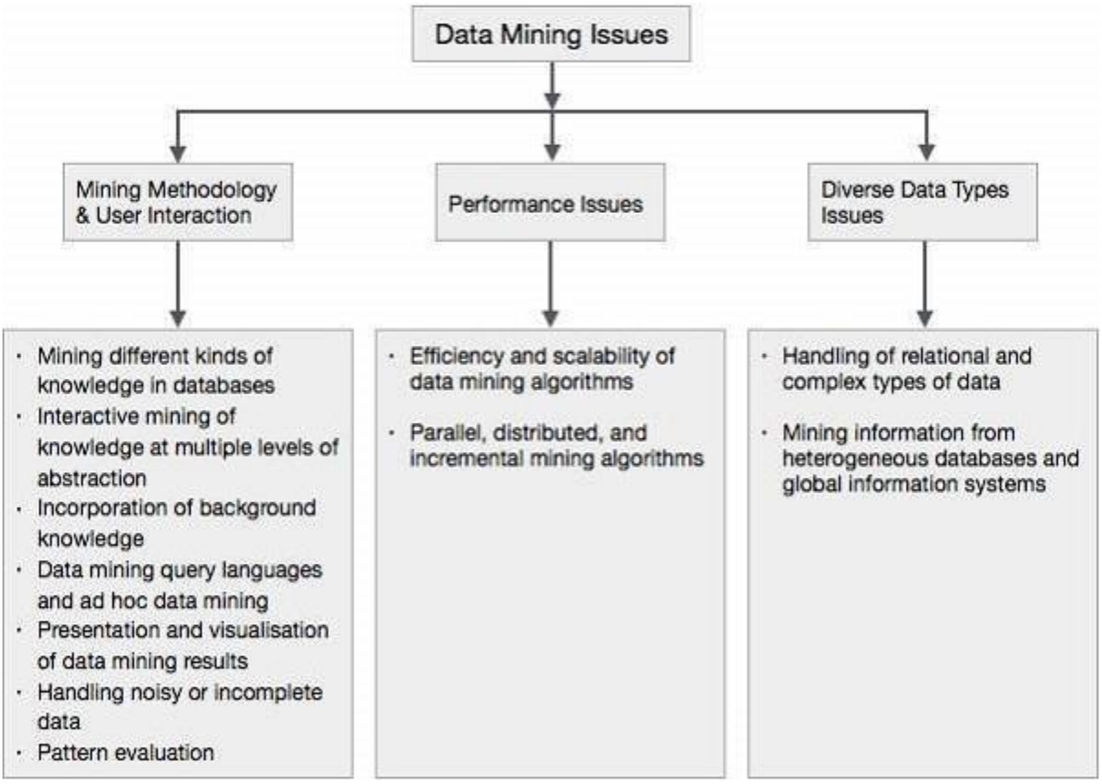
Data Mining - Issues

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create 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.



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.

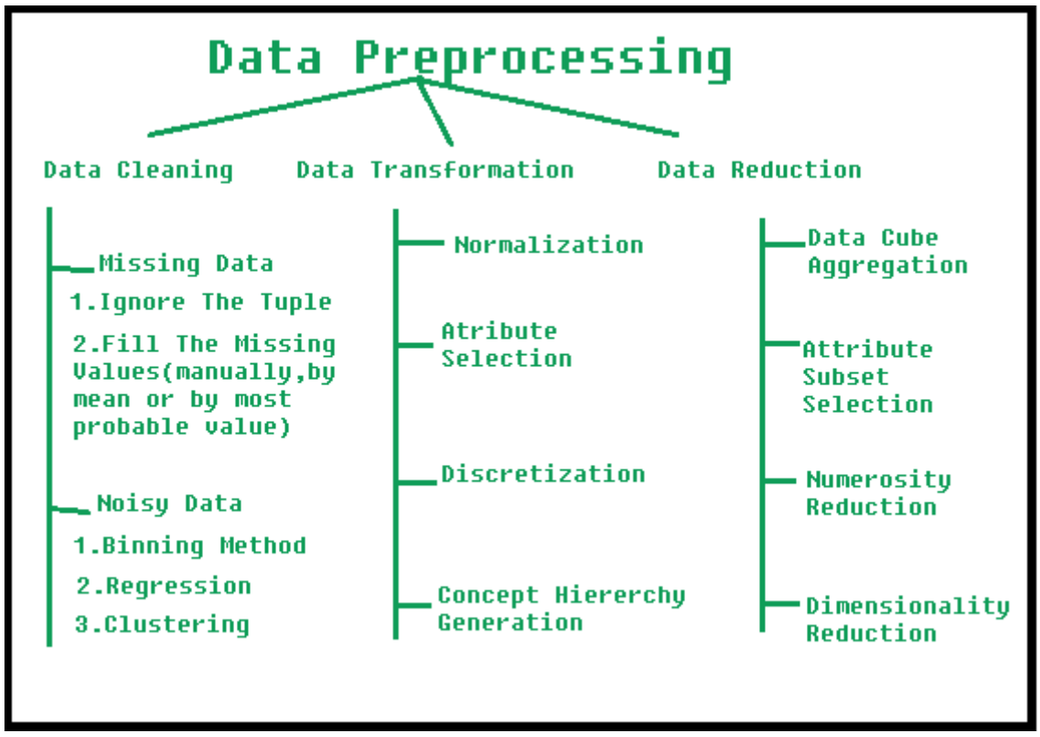
Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

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.

Data Preprocessing in Data Mining

Preprocessing in Data Mining:
Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



Steps Involved in Data Preprocessing:
1. Data Cleaning:
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

- **(a). Missing Data:**
  This situation arises when some data is missing in the data. It can be handled in various ways.
  Some of them are:
  1. **Ignore the tuples:**
     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  2. **Fill the Missing values:**
     There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
- **(b). Noisy Data:**
  Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

  1. **Binning Method:**
     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  2. **Regression:**
     Here data can be made smooth by fitting it to a regression function.The regression used

may be linear (having one independent variable) or multiple (having multiple independent variables).

3. **Clustering:**
   This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation:

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
2. **Attribute Selection:**
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**
   Here attributes are converted from level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. Data Reduction:

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**
   Aggregation operation is applied to data for the construction of the data cube.
2. **Attribute Subset Selection:**
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.
3. **Numerosity Reduction:**
   This enable to store the model of data instead of whole data, for example: Regression Models.
4. **Dimensionality Reduction:**
   This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Componenet Analysis).

Unit II

Data Mining Task Primitives:

Data Mining Task Primitives Each user will have a data mining task in mind, that is, some form of data analysis that he or she would like to have performed. A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths. The data mining primitives specify the following, as illustrated in Figure 1.13.

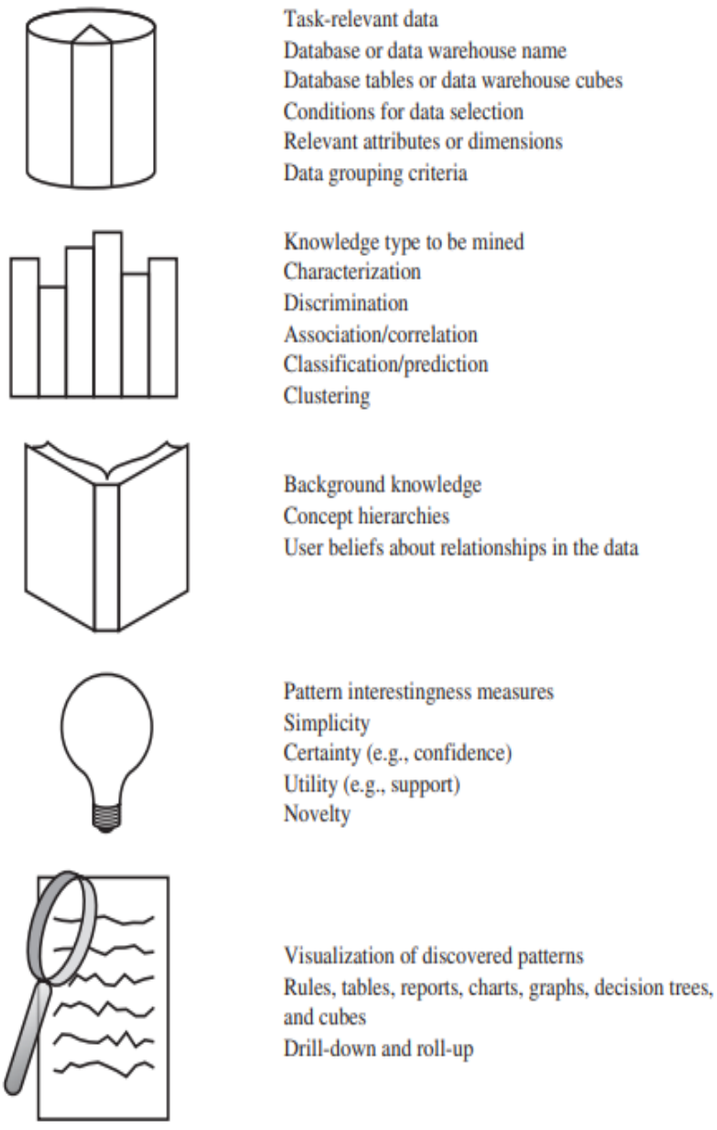
The set of task-relevant data to be mined: This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest (referred to as the relevant attributes or dimensions).

The kind of knowledge to be mined: This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

The background knowledge to be used in the discovery process: This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found. Concept hierarchies are a popular form of background knowledge, which allow data to be mined at multiple levels of abstraction. An example of a concept hierarchy for the attribute (or dimension) age is shown in Figure 1.14. User beliefs regarding relationships in the data are another form of background knowledge.

The interestingness measures and thresholds for pattern evaluation: They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures. For example, interestingness measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.

The expected representation for visualizing the discovered patterns: This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes



Data Mining Query Language

The Data Mining Query Language (DMQL) was proposed by Han, Fu, Wang, et al. for the DBMiner data mining system. The Data Mining Query Language is actually based on the Structured Query Language (SQL). Data Mining Query Languages can be designed to support ad hoc and interactive data mining. This DMQL provides commands for specifying primitives. The DMQL can work with databases and data warehouses as well. DMQL can be used to define data mining tasks. Particularly we examine how to define data warehouses and data marts in DMQL.

Syntax for Task-Relevant Data Specification

Here is the syntax of DMQL for specifying task-relevant data −

```
use database database_name
```

or

```
use data warehouse data_warehouse_name
in relevance to att_or_dim_list
from relation(s)/cube(s) [where condition]
order by order_list
group by grouping_list
```

Syntax for Specifying the Kind of Knowledge

Here we will discuss the syntax for Characterization, Discrimination, Association, Classification, and Prediction.

Characterization

The syntax for characterization is −

```
mine characteristics [as pattern_name]
   analyze  {measure(s) }
```

The analyze clause, specifies aggregate measures, such as count, sum, or count%.

For example −

```
Description describing customer purchasing habits.
mine characteristics as customerPurchasing
analyze count%
```

Discrimination

The syntax for Discrimination is −

```
mine comparison [as {pattern_name]}
For {target_class } where  {target condition }
{versus  {contrast_class_i }
where {contrast_condition_i}}
analyze  {measure(s) }
```

For example, a user may define big spenders as customers who purchase items that cost $100 or more on an average; and budget spenders as customers who purchase items at less than $100 on an average. The mining of discriminant descriptions for customers from each of these categories can be specified in the DMQL as −

```
mine comparison as purchaseGroups
for bigSpenders where avg(I.price) ≥$100
versus budgetSpenders where avg(I.price)< $100
analyze count
```

Association

The syntax for Association is−

```
mine associations [ as {pattern_name} ]
{matching {metapattern} }
```

For Example −

```
mine associations as buyingHabits
matching P(X:customer,W) ^ Q(X,Y) ≥ buys(X,Z)
```

where X is key of customer relation; P and Q are predicate variables; and W, Y, and Z are object variables.

Classification

The syntax for Classification is −

```
mine classification [as pattern_name]
analyze classifying_attribute_or_dimension
```

For example, to mine patterns, classifying customer credit rating where the classes are determined by the attribute credit_rating, and mine classification is determined as classifyCustomerCreditRating.

```
analyze credit_rating
```

Prediction

The syntax for prediction is −

```
mine prediction [as pattern_name]
analyze prediction_attribute_or_dimension
{set {attribute_or_dimension_i= value_i}}
```

Syntax for Concept Hierarchy Specification

To specify concept hierarchies, use the following syntax −

```
use hierarchy <hierarchy> for <attribute_or_dimension>
```

We use different syntaxes to define different types of hierarchies such as−

```
-schema hierarchies
define hierarchy time_hierarchy on date as [date,month quarter,year]
```

```
-
set-grouping hierarchies
define hierarchy age_hierarchy for age on customer as
level1: {young, middle_aged, senior} < level0: all
level2: {20, ..., 39} < level1: young
level3: {40, ..., 59} < level1: middle_aged
level4: {60, ..., 89} < level1: senior

-operation-derived hierarchies
define hierarchy age_hierarchy  for age  on customer  as
{age_category(1), ..., age_category(5)}
:= cluster(default, age, 5) < all(age)

-rule-based hierarchies
define hierarchy profit_margin_hierarchy  on item  as
level_1: low_profit_margin < level_0:  all

if (price - cost)< $50
   level_1:  medium-profit_margin < level_0:  all

if ((price - cost) > $50)  and ((price - cost) ≤ $250))
   level_1:  high_profit_margin < level_0:  all
```

Syntax for Interestingness Measures Specification

Interestingness measures and thresholds can be specified by the user with the statement −

```
with <interest_measure_name>  threshold = threshold_value
```

For Example −

```
with support threshold = 0.05
with confidence threshold = 0.7
```

Syntax for Pattern Presentation and Visualization Specification

We have a syntax, which allows users to specify the display of discovered patterns in one or more forms.

```
display as <result_form>
```

For Example −

```
display as table
```

Full Specification of DMQL

As a market manager of a company, you would like to characterize the buying habits of customers who can purchase items priced at no less than $100; with respect to the customer's age, type of item purchased, and the place where the item was purchased. You would like to know the percentage of customers having that characteristic. In particular, you are only interested in purchases made in Canada, and paid with an American Express credit card. You would like to view the resulting descriptions in the form of a table.

```
use database AllElectronics_db
use hierarchy location_hierarchy for B.address
mine characteristics as customerPurchasing
analyze count%
in relevance to C.age,I.type,I.place_made
from customer C, item I, purchase P, items_sold S,  branch B
where I.item_ID = S.item_ID and P.cust_ID = C.cust_ID and
P.method_paid = "AmEx" and B.address = "Canada" and I.price ≥ 100
with noise threshold = 5%
display as table
```

Data Generalization In Data Mining – Summarization Based Characterization

What is Concept Description?

• Descriptive vs. predictive data mining

Descriptive mining: describes concepts or task-relevant data sets in concise, summarily, informative, discriminative forms

Predictive mining: Based on data and analysis, constructs models for the database, and predicts the trend and properties of unknown data

• Concept description: Characterization: provides a concise and succinct summarization of the given collection of data

Comparison: provides descriptions comparing two or more collections of data

Concept Description vs. OLAP

Concept description:

-Can handle complex data types of the attributes and their aggregations .

-A more automated process

OLAP:

– Restricted to a small number of dimension and measure types

– User-controlled process.

Data Generalization and Summarization-based Characterization

Data generalization – A process which abstracts a large set of task-relevant data in a database from a low conceptual level to higher ones.

Approaches:

CONCEPTUAL LEVELS

• Data cube approach(OLAP approach)

• Attribute-oriented induction approach

Characterization: Data Cube Approach:

• Perform computations and store results in data cubes

• Strength:

– An efficient implementation of data generalization

– Computation of various kinds of measures • e.g., count( ), sum( ), average( ), max( )

– Generalization and specialization can be performed on a data cube by roll-up and drill-down •

Limitations:

– handle only dimensions of simple nonnumeric data and measures of simple aggregated numeric values.

– Lack of intelligent analysis, can’t tell which dimensions should be used and what levels should the generalization reach

Attribute-Oriented Induction

• Proposed in 1989 (KDD ‘89 workshop)

• Not confined to categorical data nor particular measures.

• How it is done?

– Collect the task-relevant data(initial relation) using a relational database query

– Perform generalization by attribute removal or attribute generalization.

– Apply aggregation by merging identical, generalized tuples and accumulating their respective counts.

– Interactive presentation with users.

Initial Relation

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
…	…	…	…	…	…	…	…
Removed	Retained	Sci,Eng, Bus	Country	Age range	City	Removed	Excl, VG,..

Prime Generalized Relation

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
…	…	…	…	…	…	…

Gender \ Birth_Region	Canada	Foreign	Total
M	16	14	30
F	10	22	32
Total	26	36	62

See Principles See Algorithm See Implementation See Analytical Characterization

Analytical Characterization : Analysis of Attribute Relevance

Introduction

“What if am not sure which attribute to include or class characterization and class comparison ? I may end up specifying too many attributes, which could slow down the: system considerably .” Measures of attribute relevance analysis can be used to help identify irrelevant or weakly relevant attributes that can be excluded from the concept description process. The incorporation of this pre-processing step into class characterization or comparison is referred to as analytical characterization or analytical comparison, respectively . This section describes a general method of attribute relevance analysis and its integration with attribute-oriented induction.

The first limitation of class characterization for multidimensional data analysis in Data warehouses and OLAP tools is the handling of complex objects . The second Limitation is the lack of an automated generalization process: the user must explicitly Tell the system which dimension should be included in the class characterization and to How high a level each dimension should be generalized . Actually , the user must specify each step of generalization or specification on any dimension.

Usually , it is not difficult for a user to instruct a data mining system regarding how high level each dimension should be generalized . For example , users can set attributegeneralization thresholds for this , or specify which level a given dimension should reach ,such as with the command “generalize dimension location to the country level”. Even without explicit user instruction , a default value such as 2 to 8 can be set by the data mining system , which would allow each dimension to be generalized to a level that contains only 2 to 8 distinct values. If the user is not satisfied with the current level of generalization, she can specify dimensions on which drill-down or roll-up operations should be applied.

It is nontrivial, howesver, for users to determine which dimensions should be included in the analysis of class characteristics. Data relations often contain 50 to 100 attributes , and a user may have little knowledge

regarding which attributes or dimensions should be selected for effective data mining. A user may include too few attributes in the analysis, causing the resulting mined descriptions to be incomplete. On the other hand, a user may introduce too many attributes for analysis (e.g. , by indicating “in relevance to *”, which includes all the attributes in the specified relations).

Methods should be introduced to perform attribute (or dimension)relevance Analysis in order to filter out statistically irrelevant or weakly relevant attributes, and retain or even rank the most relevant attributes for the descriptive mining task at hand. Class characterization that includes the analysis of attribute/dimesnsion relevance is called analytical characterization. Class comparison that includes such analysis is called analytical comparison.

Intuitively, an attribute or dimension is considered highly relevant with respect to a Given class if it is likely that the values of the attribute or dimension may be used to Distinguish the class from others. For example, it is unlikely that the color of an Automobile can be used to distinguish expensive from cheap cars, but the model , make, style, and number of cylinders are likely to be more relevant attributes. Moreover, even within the same dimension, different levels of concepts may have dramatically different powers for distinguishing a class from others.

For example, in the birth_date dimension, birth_day and birth_month are unlikely to be relevant to the salary of employees. However, the birth_decade (i.e. , age interval) may be highly relevant to the salary of employees. This implies that the analysis of dimension relevance should be performed at multi-levels of abstraction, and only the most relevant levels of a dimension should be included in the analysis. Above we said that attribute/ dimension relevance is evaluated based on the ability of the attribute/ dimension to distinguish objects of a class from others. When mining a class comparison (or discrimination), the target class and the contrasting classes are Explicitly given in the mining query. The relevance analysis should be performed by Comparison of these classes, as we shall see below. However, when mining class Characteristics, there is only one class to be characterized. That is, no contrasting class is specified. It is therefore not obvious what the contrasting class should be for use in of comparable data in the database that excludes the set of data to be characterized. For example, to characterize graduate students, the contrasting class can be composed of the set of undergraduate students.

Methods of Attribute Relevance Analysis:

There have been many studies in machine learning, statistics, fuzzy and rough set Theories, and so on , on attribute relevance analysis. The general idea behind attribute Relevance analysis is to compute some measure that is used to quantify the relevance of an attribute with respect to a given class or concept. Such measures include information gain, the Gini index, uncertainity, and correlation coefficients. Here we introduce a method that integrates an information gain analysis technique With a dimension-based data analysis method. The resulting method removes the less informative attributes, collecting the more informative ones for use in concept description analysis.

Data Collection:

Collect data for both the target class and the contrasting class by query processing. For class comparison, the user in the data-mining query provides both the target class and the contrasting class. For class characterization, the target class is the class to be characterized, whereas the contrasting class is the set of comparable data that are not in the target class.

Preliminary relevance analysis using conservative AOI:

This step identifies a Set of dimensions and attributes on which the selected relevance measure is to be Applied. Since different levels of a dimension may have dramatically different Relevance with respect to a given class, each attribute defining the conceptual levels of the dimension should be included in the relevance analysis in principle. Attribute-oriented induction (AOI)can be used to perform some preliminary relevance analysis on the data by removing or generalizing attributes having a very large number of distinct values (such as name and phone#). Such attributes are unlikely to be found useful for concept description. To be conservative , the AOI performed here should employ attribute generalization thresholds that are set reasonably large so as to allow more (but not all)attributes to be considered in further relevance analysis by the selected measure (Step 3 below). The relation obtained by such an application of AOI is called the candidate relation of the mining task.

Remove irrelevant and weakly attributes using the selected relevance analysis measure:

Evaluate each attribute in the candidate relation using the selected relevance analysis measure. The relevance measure used in this step may be built into the data mining system or provided by the user. For example, the information gain measure described above may be used. The attributes are then sorted(i.e., ranked)according to their computed relevance to the data mining task. Attributes that are not relevant or are weakly relevant to the task are then removed. A threshold may be set to define “weakly relevant.” This step results in an initial Target class working relation and an initial contrasting class working relation.

Generate the concept description using AOI:

Perform AOI using a less Conservative set of attribute generalization thresholds. If the descriptive mining Task is class characterization, only the initial target class working relation is included here. If the descriptive mining task is class comparison, both the initial target class working relation and the initial contrasting class working relation are included. The complexity of this procedure is the induction process is perfomed twice, that Is, in preliminary relevance analysis (Step 2)and on the initial working relation (Step4). The statistics used in attribute relevance analysis with the selected measure (Step 3) may be collected during the scanning of the database in Step 2

Mining Class Comparisons: Discriminating Between Different Classes

Introduction: In many applications, users may not be interested in having a single class (or concept) described or characterized, but rather would prefer to mine a description that compares or distinguishes one class (or concept) from other comparable classes (or concepts).Class discrimination or comparison (hereafter referred to as class comparison) mines descriptions that distinguish a target class from its contrasting classes. Notice that the target and contrasting classes must be comparable in the sense that they share similar dimensions and attributes. For example, the three classes, person, address, and item, are not comparable.

However, the sales in the last three years are comparable classes, and so are computer science students versus physics students. Our discussions on class characterization in the previous sections handle multilevel data summarization and characterization in a single class. The techniques developed can be extended to handle class comparison across several comparable classes. For example, the attribute generalization process described for class characterization can be modified so that the generalization is performed synchronously among all the classes compared. This allows the attributes in all of the classes to be generalized to the same levels of abstraction. Suppose, for instance, that we are given the All Electronics data for sales in 2003 and sales in 2004 and would like to compare these two classes. Consider the dimension location with abstractions at the city, province or state, and country levels. Each class of data should be generalized to the same location level. That is, they are synchronously all generalized to either the city level, or the province or state level, or the country level. Ideally, this is more useful than comparing, say, the sales in Vancouver in 2003 with the sales in the United States in 2004 (i.e., where each set of sales data is generalized to a different level). The users, however, should have the option to overwrite such an automated, synchronous comparison

with their own choices, when preferred.

“How is class comparison performed?” In general, the procedure is as follows:

1. Data collection: The set of relevant data in the database is collected by query processing and is partitioned respectively into a target class and one or a set of contrasting class(es).

2. Dimension relevance analysis: If there are many dimensions, then dimension relevance analysis should be performed on these classes to select only the highly relevant dimensions for further analysis. Correlation or entropy-based measures can be used for this step (Chapter 2).

3. Synchronous generalization: Generalization is performed on the target class to the level controlled by a user- or expert-specified dimension threshold, which results in a prime target class relation. The concepts in the contrasting class(es) are generalized to the same level as those in the prime target class relation, forming the prime contrasting class(es) relation.

4. Presentation of the derived comparison: The resulting class comparison description can be visualized in the form of tables, graphs, and rules. This presentation usually includes a “contrasting” measure such as count% (percentage count) that reflects the comparison between the target and contrasting classes. The user can adjust the comparison description by applying drill-down, roll-up, and other OLAP operations to the target and contrasting classes, as desired.

The above discussion outlines a general algorithm for mining comparisons in databases. In comparison with characterization, the above algorithm involves synchronous generalization of the target class with the contrasting classes, so that classes are simultaneously

Example

Task - Compare graduate and undergraduate students using the discriminant rule.

for this, the DMQL query would be.

use University_Database
mine comparison as “graduate_students vs_undergraduate_students”
in relevance to name, gender, program, birth_place, birth_date, residence, phone_no, GPA
for “graduate_students”
where status in “graduate”
versus “undergraduate_students”
where status in “undergraduate”
analyze count%
from student

Now from this, we can formulate that

- **attributes** = name, gender, program, birth_place, birth_date, residence, phone_no, and GPA.
- **Gen(ai)** = concept hierarchies on attributes ai.
- **Ui** = attribute analytical thresholds for attributes ai.
- **Ti** = attribute generalization thresholds for attributes ai.
- **R** = attribute relevance threshold.

1. Data collection -Understanding **Target** and **Contrasting** classes.

2. Attribute relevance analysis - It is used to **remove attributes** name, gender, program, phone_no.

3. Synchronous generalization - It is controlled by user-specified dimension thresholds, a prime target, and contrasting class(es) relations/cuboids.

Initial target class working relation (graduate student)

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
...	...	...	...	...	...	...	...

Initial contrasting class working relation (graduate student)

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Bob Schumann	M	Chem	Calagary, Alt, Canada	10-1-78	2642 Halifax St, Burnaby	294-4291	2.96
Ammy. Eau	F	Bio	Golden, BC, Canada	30-3-76	463 Sunset Cres, Vancouer	681-5417	3.52
...	...	...	...	...	...	...	...

4. Drill down, roll up and other OLAP operations on target and contrasting classes to adjust levels of abstractions of resulting description.

Prime generalized relation for the target class: Graduate students

Major	Age_range	Gpa	Count%
Science	20-25	Good	5.53%
Science	26-30	Good	2.32%
Science	Over_30	Very_good	5.86%
…	…	…	…
Business	Over_30	Excellent	4.68%

Prime generalized relation for the contrasting class: Undergraduate students

Major	Age_range	Gpa	Count%
Science	15-20	Fair	5.53%
Science	15-20	Good	4.53%
…	…	…	…
Science	26-30	Good	5.02%
…	…	…	…
Business	Over_30	Excellent	0.68%

Unit 3

Association Rule Mining

ssociation rule mining is a procedure which aims to observe frequently occurring patterns, correlations, or associations from datasets found in various kinds of databases such as relational databases, transactional databases, and other forms of repositories.

An association rule has 2 parts:

- **an antecedent (if) and**
- **a consequent (then)**

An antecedent is something that's found in data, and a consequent is an item that is found in combination with the antecedent. Have a look at this rule for instance:

"If a customer buys bread, he's 70% likely of buying milk."

In the above association rule, bread is the antecedent and milk is the consequent. Simply put, it can be understood as a retail store's association rule to target their customers better. If the above rule is a result of a thorough analysis of some data sets, it can be used to not only improve customer service but also improve the company's revenue.

Association rules are created by thoroughly analyzing data and looking for frequent if/then patterns. Then, depending on the following two parameters, the important relationships are observed:

1. **Support**: Support indicates how frequently the if/then relationship appears in the database.
2. **Confidence**: Confidence tells about the number of times these relationships have been found to be true.

So, in a given transaction with multiple items, Association Rule Mining primarily tries to find the rules that govern how or why such products/items are often bought together. For example, peanut butter and jelly are frequently purchased together because a lot of people like to make PB&J sandwiches.

A Beginner's Guide to Data Science and Its Applications

Association Rule Mining is sometimes referred to as "Market Basket Analysis", as it was the first application area of association mining. The aim is to discover associations of items occurring together more often than you'd expect from randomly sampling all the possibilities. The classic anecdote of Beer and Diaper will help in understanding this better.

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is Market Based Analysis.

Market Based Analysis is one of the key techniques used by large relations to show associations between items.It allows retailers to identify relationships between the items that people buy together frequently.

Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

TID	ITEMS
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

	Beer	Bread	Milk	Diapel	Eggs	Coke
T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1

Before we start defining the rule, let us first see the basic definition

Support COUNT(σ)– Frequency of occurrence of a itemset.
Here σ({Milk, Bread, Diaper})=2

Frequent Itemset – An itemset whose support is greater than or equal to minsup threshold.
Association Rule – An implication expression of the form X -> Y, where X and Y are any 2 itemsets.
Example: {Milk, Diaper}->{Beer}

Rule Evaluation Metrics –

- **Support(s) –**
  The number of transactions that include items in the {X} and {Y} parts of the rule as a percentage of the total number of transaction.It is a measure of how frequently the collection of items occur together as a percentage of all transactions.
- **Support = σ (X+Y) ÷ total –**
  It is interpreted as fraction of transactions that contain both X and Y.
- **Confidence(c) –**
  It is the ratio of the no of transactions that includes all items in {B} as well as the no of transactions that includes all items in {A} to the no of transactions that includes all items in {A}.
- **Conf(X=>Y) = Supp(XUY) ÷ Supp(X) –**
  It measures how often each item in Y appears in transactions that contains items in X also.
- **Lift(l) –**
  The lift of the rule X=>Y is the confidence of the rule divided by the expected confidence, assuming that the itemsets X and Y are independent of each other.The expected confidence is the confidence divided by the frequency of {Y}.
- **Lift(X=>Y) = Conf(X=>Y) ÷ Supp(Y) –**
  Lift value near 1 indicates X and Y almost often appear together as expected, greater than 1

means they appear together more than expected and less than 1 means they appear less than expected.Greater lift values indicate stronger association.

Example – From the above table, {Milk, Diaper}=>{Beer}

s= σ({Milk, Diaper, Beer}) / |T|

= 2/5

= 0.4

c= σ(Milk, Diaper, Beer) / (Milk, Diaper)

= 2/3

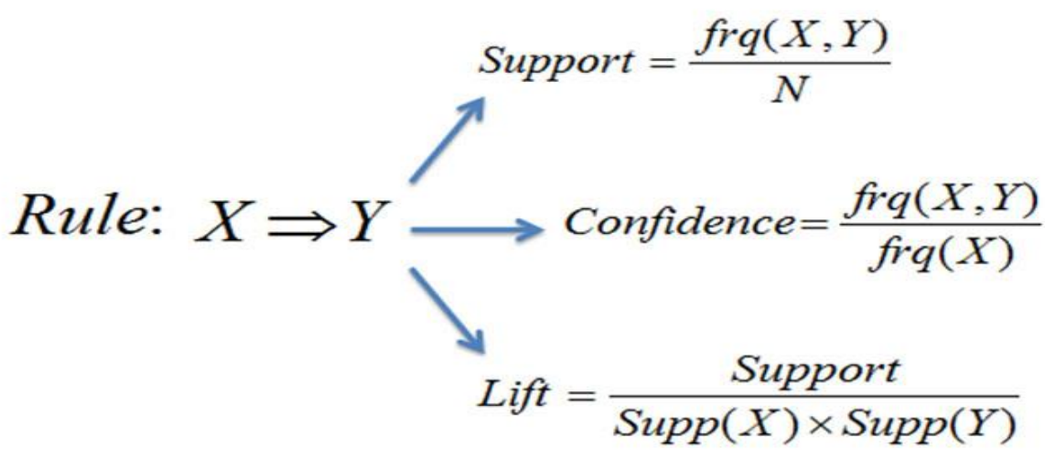
= 0.67

l= Supp({Milk, Diaper, Beer}) / Supp({Milk, Diaper})*Supp({Beer})

= 0.4/(0.6*0.6)

= 1.11

The Association rule is very useful in analyzing datasets. The data is collected using bar-code scanners in supermarkets. Such databases consists of a large number of transaction records which list all items bought by a customer on a single purchase. So the manager could know if certain groups of items are consistently purchased together and use this data for adjusting store layouts, cross-selling, promotions based on statistics.



Mining single-dimensional Boolean association rules from transactional databases and multilevel Association rule

What is association mining?

Association mining aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items or objects in transaction databases, relational database or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control, cross-marketing, catalog design, loss-leader analysis, clustering, classification, etc.

Examples:

Rule form: "Body⇒head *support, confidence+". Buys

(x,"diapers") ⇒buys(x, "beers") *0.5%, 60%+

major(x, "cs") ^ takes(x, "DB") ⇒grade(x, "A") *1%, 75%+

Association rule: basic concepts:

- Given: (1) database of transaction, (2) each transaction is a list of items (purchased by a customer in visit)
- Find: all rules that correlate the presence of one set of items with that of another set of items.
  - E.g., 98% of people who purchase tires and auto accessories also get automotive services done.
  - E.g., Market Basket Analysis

    This process analyzes customer buying habits by finding associations between the different items that customers place in their "Shopping Baskets". The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customer.

- **Applications**
  - ***⇒**maintenance agreement (what the store should do to boost maintenance agreement sales)
  - Home electronics ⇒* (what other products should the store stocks up?)
  - Attached mailing in direct marketing
  - Detecting "ping-pong" ing of patients, faulty "collisions"

RULE Measures: supports and confidence

Support: percentage of transaction in D that contain AUB.

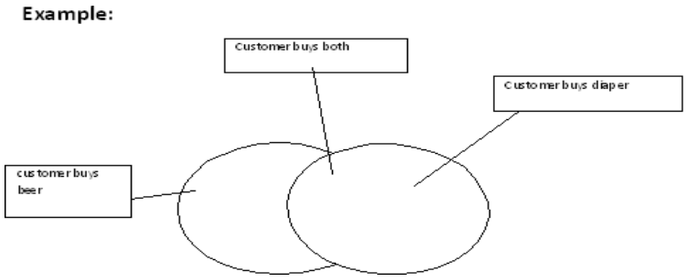
Confidence: percentage of transaction in D containing A that also contains B.

Support (A ⇒B) = p (A ∪B)

Confidence (A ⇒B) = P (B/A).

Rules that satisfy both a minimum supports threshold (min_sup) and a minimum confidence threshold (min_conf) are called **strong**

Example:



Transaction id	Items Bought
2000	A, B, C
1000	A, C
4000	A, D
5000	B, E, F

LET minimum, support 50%, and minimum confidence 50%, we have

A ⇒C (50%,66.6%)

C ⇒A (50%, 100%)

In general, association rules mining can be viewed as a two-step process:

1. Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, min_sup.
2. Generate strong association rules from the frequent itemsets:
   By definition, these rules must satisfy minimum support and minimum confidence.

Classification of association rules mining:

- **Based on the level of abstraction involved in the rules set:**
  - **Single level association rules** refer items or attribute at only one level.
    Buys (X, "computer") ⇒buys(X, "HP printer")
  - **Multi-level association rules** reference items or attribute at different levels of abstraction.
    Buys(X, "laptop computer") ⇒ buys(X, "HP printer")

- **Based on the number of data dimensions involved in the rules:**
  - **Single dimensional Association rule** is an association rule in which items or attribute reference only one dimension.
    Buys (X, "computer") ⇒ buys (X,"antivirus software")
  - **Multidimensional association rule** reference two or more dimensions age (X, "30….39")^income(X, "42k…48k")⇒buys(X, "high resolution TV")
- **Based on the types of the values handled in rule:**
  - **Boolean association rule** involve associations between the presence and absence of items.
    buys (X, "SQLServer") ^ buys (X, "DMBook") ⇒ buys(X, "DBMiner")
  - **Quantitative association rule** describe association between quantitative items or attributes.
    Age (X, "30…39") ^ income(X, "42k…49k") ⇒ buys(X, "PC")

Data Mining - Classification & Prediction

There are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends. These two forms are as follows −

- Classification
- Prediction

Classification models predict categorical class labels; and prediction models predict continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

What is classification?

Following are the examples of cases where the data analysis task is Classification −

- A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are risky or which are safe.
- A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

In both of the above examples, a model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data.

What is prediction?

Following are the examples of cases where the data analysis task is Prediction −

Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bothered to predict a numeric value. Therefore the data analysis task is an example of numeric prediction. In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value.

Note − Regression analysis is a statistical methodology that is most often used for numeric prediction.

How Does Classification Works?

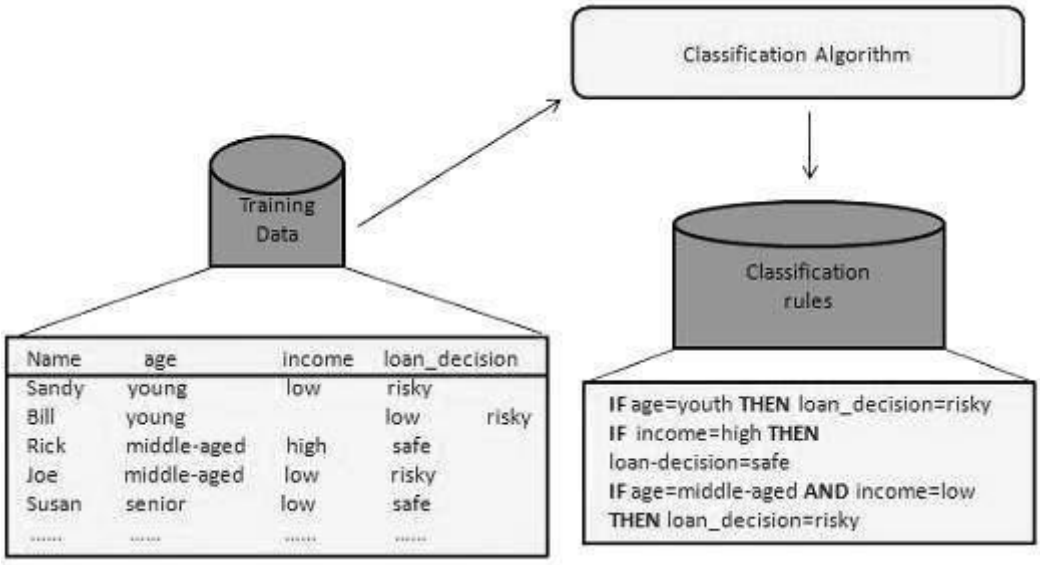
With the help of the bank loan application that we have discussed above, let us understand the working of classification. The Data Classification process includes two steps −

- Building the Classifier or Model
- Using Classifier for Classification

Building the Classifier or Model

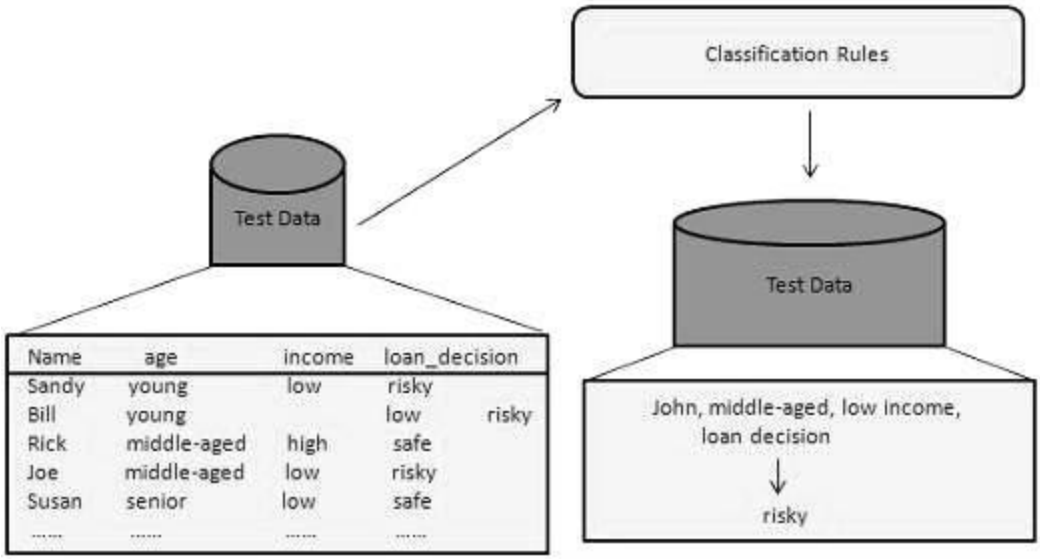
- This step is the learning step or the learning phase.
- In this step the classification algorithms build the classifier.
- The classifier is built from the training set made up of database tuples and their associated class labels.

- Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.



Using Classifier for Classification

In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.



Classification and Prediction Issues

The major issue is preparing the data for Classification and Prediction. Preparing the data involves the following activities −

- **Data Cleaning** − Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.

- **Relevance Analysis** − Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.
- **Data Transformation and reduction** − The data can be transformed by any of the following methods.
    - **Normalization** − The data is transformed using normalization. Normalization involves scaling all values for given attribute in order to make them fall within a small specified range. Normalization is used when in the learning step, the neural networks or the methods involving measurements are used.
    - **Generalization** − The data can also be transformed by generalizing it to the higher concept. For this purpose we can use the concept hierarchies.

Note − Data can also be reduced by some other methods such as wavelet transformation, binning, histogram analysis, and clustering.

Comparison of Classification and Prediction Methods

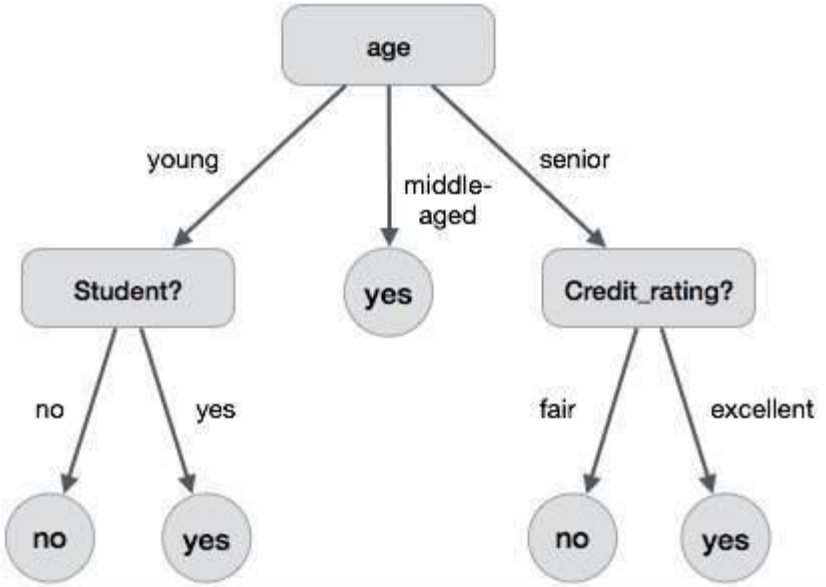
Here is the criteria for comparing the methods of Classification and Prediction −

- **Accuracy** − Accuracy of classifier refers to the ability of classifier. It predict the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.
- **Speed** − This refers to the computational cost in generating and using the classifier or predictor.
- **Robustness** − It refers to the ability of classifier or predictor to make correct predictions from given noisy data.
- **Scalability** − Scalability refers to the ability to construct the classifier or predictor efficiently; given large amount of data.
- **Interpretability** − It refers to what extent the classifier or predictor understand

Data Mining - Decision Tree Induction

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

The following decision tree is for the concept buy_computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node represents a class.



The benefits of having a decision tree are as follows −

- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast.

Tree Pruning

Tree pruning is performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex.

Tree Pruning Approaches

There are two approaches to prune a tree −

- **Pre-pruning** − The tree is pruned by halting its construction early.
- **Post-pruning** - This approach removes a sub-tree from a fully grown tree.

Cost Complexity

The cost complexity is measured by the following two parameters −

- Number of leaves in the tree, and
- Error rate of the tree

Data Mining - Bayesian Classification

Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

Baye's Theorem

Bayes' Theorem is named after Thomas Bayes. There are two types of probabilities −

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

where X is data tuple and H is some hypothesis.

According to Bayes' Theorem,

$$P(H/X)= P(X/H)P(H) / P(X)$$

Bayesian Belief Network

Bayesian Belief Networks specify joint conditional probability distributions. They are also known as Belief Networks, Bayesian Networks, or Probabilistic Networks.

- A Belief Network allows class conditional independencies to be defined between subsets of variables.
- It provides a graphical model of causal relationship on which learning can be performed.
- We can use a trained Bayesian Network for classification.

There are two components that define a Bayesian Belief Network −

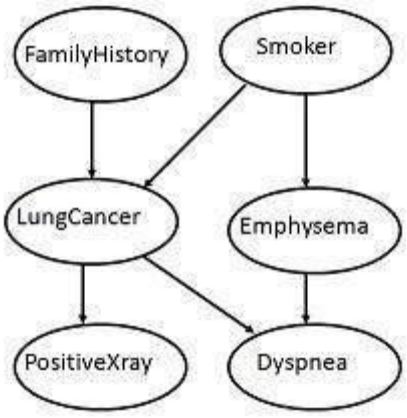
- Directed acyclic graph
- A set of conditional probability tables

Directed Acyclic Graph

- Each node in a directed acyclic graph represents a random variable.
- These variable may be discrete or continuous valued.
- These variables may correspond to the actual attribute given in the data.

Directed Acyclic Graph Representation

The following diagram shows a directed acyclic graph for six Boolean variables.



The arc in the diagram allows representation of causal knowledge. For example, lung cancer is influenced by a person's family history of lung cancer, as well as whether or not the person is a smoker. It is worth noting that the variable PositiveXray is independent of whether the

patient has a family history of lung cancer or that the patient is a smoker, given that we know the patient has lung cancer.

Conditional Probability Table

The conditional probability table for the values of the variable LungCancer (LC) showing each possible combination of the values of its parent nodes, FamilyHistory (FH), and Smoker (S) is as follows −

	FH,S	FH,-S	-FH,S	-FH,S
LC	0.8	0.5	0.7	0.1
-LC	0.2	0.5	0.3	0.9

Definition *Predictive Data Mining* means?

Predictive data mining is data mining that is done for the purpose of using business intelligence or other data to forecast or predict trends. This type of data mining can help business leaders make better decisions and can add value to the efforts of the analytics team.