DATA MINING TECHNIQUES

Course Code	Course Title	Duration (Weeks)	Course Type	L	T	P	C	Hrs. /Wk.
BTCS14F6410	**Data Mining Techniques**	16	SC	3	1	0	4	5

Prerequisites:

Probability and statistics and Database Management systems

Course Objectives:

Objectives of this course are to:

1. Introduce the basics of data mining, data types, similarity and dissimilarity measures
2. Explain association rules and algorithms
3. Describe the classification algorithms for data categorization
4. Illustrate the clustering algorithms for grouping data sets
5. Demonstrate the appropriate data mining techniques for decision making

Course Outcomes:

On successful completion of this course; student shall be able to:

1. Explain the basics of data mining techniques, data types, identify the similarity and dissimilarity between the data sets.
2. Analyze the data sets using the association rules and algorithms
3. Characterize and discriminate data sets with classification methods
4. Employ the clustering methods in real life problems
5. Apply the knowledge for data mining applications

Course Contents

Unit - 1
Introduction: What is Data Mining? Motivating Challenges, The origins of data mining, Data Mining Tasks, Types of Data, Data Quality, Data Preprocessing, Measures of Similarity and Dissimilarity, Data Mining Applications, Visualization.

Unit - 2
Association Analysis: Basic Concepts and Algorithms, Frequent Itemset Generation, Rule Generation, Compact Representation of Frequent Itemsets, Alternative methods for generating Frequent Itemsets, FP GROWTH Algorithm, Evaluation of Association Patterns .

Unit - 3
Classification: Basics, General approach to solve classification problem, Decision Trees, Evaluating the performance of a Classifier, Rule Based Classifiers, Nearest Neighbour Classifiers, Naïve Bayes Classifier

Unit - 4
Clustering: overview, K-means**,** agglomerative hierarchical clustering, DBSCAN, Cluster Evaluation, Characteristics of Data, Clusters and Clustering Algorithms, Prototype Based Clustering.

Recommended Learning Resources:

1. A Pang-Ning Tan, Michael Steinbach and Vipin Kumar, “Introduction to Data Mining”, Pearson Education, 2007.
2. Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques” Second Edition, Elsevier, Reprinted 2008.

References:

1. K.P. Soman, Shyam Diwakar and V. Ajay, “Insight into Data mining Theory and Practice”, Easter Economy Edition, Prentice Hall of India, 2006.
2. G. K. Gupta, “Introduction to Data Mining with Case Studies”, Easter Economy Edition, Prentice Hall of India, 2006.
3. Data Mining and Knowledge Science – Springer.
4. Inderscience,The International Journal of DataMining, Modelling and Management-
5. IEEE, IEEE Transactions on Knowledge and Data Engineering.

UNIT 1: DATA MINING

Introduction: What is Data Mining? Motivating Challenges, The origins of data mining, Data Mining Tasks, Types of Data, Data Quality, Data Preprocessing, Measures of Similarity and Dissimilarity, Data Mining Applications, Visualization.

WHAT IS DATA MINING?

- Data Mining is the process of automatically discovering useful information in large data- repositories.
- DM techniques
  → can be used to search large DB to find useful patterns that might otherwise remain unknown
  → provide capabilities to predict the outcome of future observations

Why do we need Data Mining?
- Conventional database systems provide users with query & reporting tools.
- To some extent the query & reporting tools can assist in answering questions like, where did the largest number of students come from last year?
- But these tools cannot provide any intelligence about why it happened.

Taking an Example of University Database System
- The OLTP system will quickly be able to answer the query like "how many students are enrolled in university"
- The OLAP system using data warehouse will be able to show the trends in students' enrollments (ex: how many students are preferring BCA),
- Data mining will be able to answer where the university should market.

DATA MINING AND KNOWLEDGE DISCOVERY
- Data Mining is an integral part of KDD (Knowledge Discovery in Databases).
  KDD is the overall process of converting raw data into useful information (Figure: 1.1).

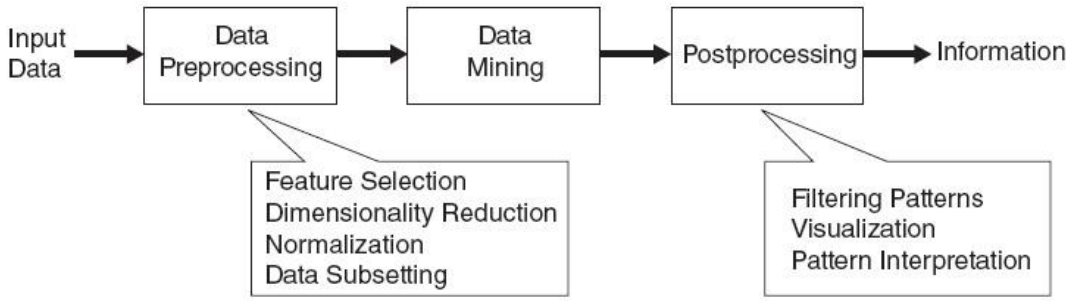


Figure 1.1. The process of knowledge discovery in databases (KDD).

- The input-data is stored in various formats such as flat files, spread sheet or relational tables.
- Purpose of preprocessing: to transform the raw input-data into an appropriate format for subsequent analysis.
- The steps involved in data-preprocessing include
  → combine data from multiple sources
  → clean data to remove noise & duplicate observations, and
  → select records & features that are relevant to the DM task at hand
- Data-preprocessing is perhaps the most time-consuming step in the overall knowledge discovery process.
- "Closing the loop" refers to the process of integrating DM results into decision support systems.
- Such integration requires a postprocessing step. This step ensures that only valid and useful results are incorporated into the decision support system.
- An example of postprocessing is visualization.
  Visualization can be used to explore data and DM results from a variety of viewpoints.

• Statistical measures can also be applied during postprocessing to eliminate bogus DM results.

MOTIVATING CHALLENGES

Scalability

• Nowadays, data-sets with sizes of terabytes or even petabytes are becoming common.
• DM algorithms must be scalable in order to handle these massive data sets.
• Scalability may also require the implementation of novel data structures to access individual records in an efficient manner.
• Scalability can also be improved by developing parallel & distributed algorithms.

High Dimensionality

• Traditional data-analysis technique can only deal with low dimensional data.
• Nowadays, data-sets with hundreds or thousands of attributes are becoming common.
• Data-sets with temporal or spatial components also tend to have high dimensionality.
• The computational complexity increases rapidly as the dimensionality increases.

Heterogeneous and Complex Data

• Traditional analysis methods can deal with homogeneous type of attributes.
• Recent years have also seen the emergence of more complex data-objects.
• DM techniques for complex objects should take into consideration relationships in the data, such as
  → temporal & spatial autocorrelation
  → parent-child relationships between the elements in semi-structured text & XML documents

Data Ownership & Distribution

• Sometimes, the data is geographically distributed among resources belonging to multiple entities.
• Key challenges include:
  1) How to reduce amount of communication needed to perform the distributed computation
  2) How to effectively consolidate the DM results obtained from multiple sources &
  3) How to address data-security issues

Non Traditional Analysis

• The traditional statistical approach is based on a hypothesized and test paradigm.
In other words, a hypothesis is proposed, an experiment is designed to gather the data, and then the data is analyzed with respect to hypothesis.
• Current data analysis tasks often require the generation and evaluation of thousands of hypotheses, and consequently, the development of some DM techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation.

THE ORIGIN OF DATA MINING

• Data mining draws upon ideas from
  → Sampling, estimation, and hypothesis test from statistics
  → Search algorithms, modeling techniques machine learning, learning theories from AI pattern recognition, statistics database systems
• Traditional techniques may be unsuitable due to
  →Enormity of data →High dimensionality of data →Heterogeneous nature of data
• Data mining also had been quickly to adopt ideas from other areas including
  →Optimization →Evolutionary computing →Signal processing →Information theory
• Database systems are needed to provide supports for efficient storage, indexing, query processing.
• The parallel computing and distribute technology are two major data addressing issues in data mining to increase the performance.

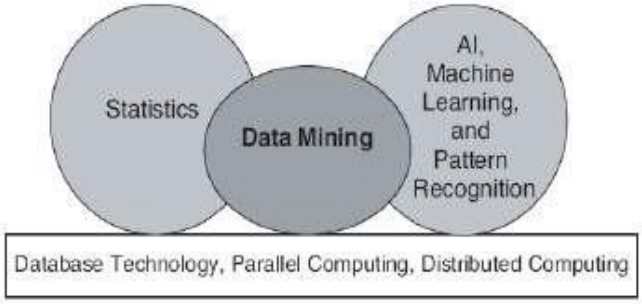


Figure 1.2. Data mining as a confluence of many disciplines.

DATA MINING TASKS

- DM tasks are generally divided into 2 major categories.

Predictive Tasks

- The objective is to predict the value of a particular attribute based on the values of other attributes.
- The attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for making the predication are known as the explanatory or independent variables.

Descriptive Tasks

- The objective is to derive patterns (correlations, trends, clusters, trajectories and anomalies) that summarize the relationships in data.
- Descriptive DM tasks are often exploratory in nature and frequently require postprocessing techniques to validate and explain the results.

Four of the Core Data Mining Tasks

1) Predictive Modeling

- This refers to the task of *building a model for the target variable* as a function of the explanatory variable.
- The goal is to learn a model that minimizes the error between the predicted and true values of the target variable.
- There are 2 types of predictive modeling tasks:

i) Classification: used for discrete target variables

Ex: Web user will make purchase at an online bookstore is a classification task, because the target variable is binary valued.

ii) Regression: used for continuous target variables.

Ex: forecasting the future price of a stock is regression task because price is a continuous values attribute

2) Cluster Analysis

- This seeks to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other than observations that belong to other clusters.
- Clustering has been used

→ to group sets of related customers

→ to find areas of the ocean that have a significant impact on the Earth's climate

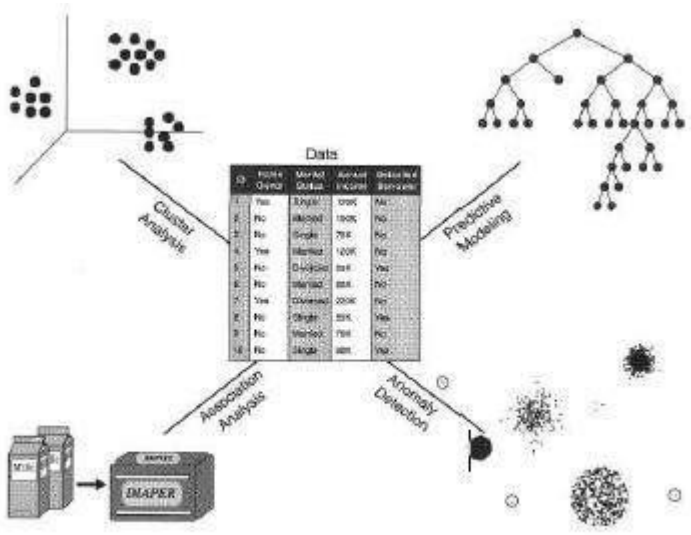


Figure 1.3. Four of the core data mining tasks.

3) Association Analysis

- This is used to discover patterns that describe strongly associated features in the data.
- The goal is to extract the most interesting patterns in an efficient manner.
- Useful applications include

→ finding groups of genes that have related functionality or

→ identifying web pages that are accessed together

- Ex: market based analysis

We may discover the rule that {diapers} -> {Milk}, which suggests that customers who buy diapers also tend to buy milk.

4) Anomaly Detection

- This is the task of identifying observations whose characteristics are significantly different from the rest of the data. Such observations are known as anomalies or outliers
- The goal is
  → to discover the real anomalies and
  → to avoid falsely labeling normal objects as anomalous.
- Applications include the detection of fraud, network intrusions, and unusual patterns of disease.

Example 1.4 (Credit Card Fraud Detection).

- A credit card company records the transactions made by every credit card holder, along with personal information such as credit limit, age, annual income, and address.
- Since the number of fraudulent cases is relatively small compared to the number of legitimate transactions, anomaly detection techniques can be applied to build a profile of legitimate transactions for the users.
- When a new transaction arrives, it is compared against the profile of the user.
- If the characteristics of the transaction are very different from the previously created profile, then the transaction is flagged as potentially fraudulent

EXERCISES

1. What is data mining? Explain Data Mining and Knowledge Discovery? (10)
2. What are different challenges that motivated the development of DM? (10)
3. Explain Origins of data mining (5)
4. Discuss the tasks of data mining with suitable examples. (10)
5. Explain Anomaly Detection .Give an Example? (5)
6. Explain Descriptive tasks in detail? (10)
7. Explain Predictive tasks in detail by example? (10)

TYPES OF DATA

WHAT IS A DATA OBJECT?

- A data-set refers to a collection of data-objects and their attributes.
- Other names for a data-object are record, transaction, vector, event, entity, sample or observation.
- Data-objects are described by a number of attributes such as
  → mass of a physical object or
  → time at which an event occurred.
- Other names for an attribute are dimension, variable, field, feature or characteristics.

WHAT IS AN ATTRIBUTE?

- An attribute is a characteristic of an object that may vary, either
  → from one object to another or
  → from one time to another.
- For example, eye color varies from person to person.

Eye color is a symbolic attribute with a small no. of possible values {brown, black, blue, green}.

Example 2.2 (Student Information).

- Often, a data-set is a file, in which the objects are records(or rows) in the file and each field (or column) corresponds to an attribute.
- For example, Table 2.1 shows a data-set that consists of student information.
- Each row corresponds to a student and each column is an attribute that describes some aspect of a student, such as grade point average(GPA) or identification number(ID).

Table 2.1. A sample data set containing student information.

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

PROPERTIES OF ATTRIBUTE VALUES

- The type of an attribute depends on which of the following properties it possesses:
  1) Distinctness: = ≠
  2) Order: < >
  3) Addition: + -
  4) Multiplication: * /
- Nominal attribute: Uses only distinctness.
  Examples: ID numbers, eye color, pin codes
- Ordinal attribute: Uses distinctness & order.
  Examples: Grades in {SC, FC, FCD}
  Shirt sizes in {S, M, L, XL}
- Interval attribute: Uses distinctness, order & addition
  Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- Ratio attribute: Uses all 4 properties
  Examples: temperature in Kelvin, length, time, counts

DIFFERENT TYPES OF ATTRIBUTES

Table 2.2. Different attribute types.

Attribute Type		Description	Examples	Operations
Categorical (Qualitative)	Nominal	The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. $(=, \neq)$	zip codes, employee ID numbers, eye color, gender	mode, entropy, contingency correlation, χ^2 test
	Ordinal	The values of an ordinal attribute provide enough information to order objects. $(<, >)$	hardness of minerals, $\{good, better, best\}$, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric (Quantitative)	Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. $(+, -)$	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
	Ratio	For ratio variables, both differences and ratios are meaningful. $(*, /)$	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Table 2.3. Transformations that define attribute levels.

Attribute Type		Transformation	Comment
Categorical (Qualitative)	Nominal	Any one-to-one mapping, e.g., a permutation of values	If all employee ID numbers are reassigned, it will not make any difference.
	Ordinal	An order-preserving change of values, i.e., $new\_value = f(old\_value)$, where f is a monotonic function.	An attribute encompassing the notion of good, better, best can be represented equally well by the values $\{1, 2, 3\}$ or by $\{0.5, 1, 10\}$.
Numeric (Quantitative)	Interval	$new\_value = a * old\_value + b$, a and b constants.	The Fahrenheit and Celsius temperature scales differ in the location of their zero value and the size of a degree (unit).
	Ratio	$new\_value = a * old\_value$	Length can be measured in meters or feet.

DESCRIBING ATTRIBUTES BY THE NUMBER OF VALUES

1) Discrete

- Has only a finite or countably infinite set of values.
- Examples: pin codes, ID Numbers, or the set of words in a collection of documents.
- Often represented as integer variables.
- Binary attributes are a special case of discrete attributes and assume only 2 values.
  E.g. true/false, yes/no, male/female or 0/1

2) Continuous

- Has real numbers as attribute values.
- Examples: temperature, height, or weight.
- Often represented as floating-point variables.

ASYMMETRIC ATTRIBUTES

- Binary attributes where only non-zero values are important are called asymmetric binary attributes.
- Consider a data-set where each object is a student and each attribute records whether or not a student took a particular course at a university.
- For a specific student, an attribute has a value of 1 if the student took the course associated with that attribute and a value of 0 otherwise.
- Because students take only a small fraction of all available courses, most of the values in such a data-set would be 0.
- Therefore, it is more meaningful and more efficient to focus on the non-zero values.
- This type of attribute is particularly important for association analysis.

TYPES OF DATA SETS

1) Record data
   → Transaction (or Market based data)
   → Data matrix
   → Document data or Sparse data matrix
2) Graph data
   → Data with relationship among objects (World Wide Web)
   → Data with objects that are Graphs (Molecular Structures)
3) Ordered data
   → Sequential data (Temporal data)
   → Sequence data
   → Time series data
   → Spatial data

GENERAL CHARACTERISTICS OF DATA SETS

- Following 3 characteristics apply to many data-sets:

1) Dimensionality

- Dimensionality of a data-set is no. of attributes that the objects in the data-set possess.
- Data with a small number of dimensions tends to be qualitatively different than moderate or high-dimensional data.
- The difficulties associated with analyzing high-dimensional data are sometimes referred to as the curse of dimensionality.
- Because of this, an important motivation in preprocessing data is dimensionality reduction.

2) Sparsity

- For some data-sets with asymmetric feature, most attribute of an object have values of 0.
- In practical terms, sparsity is an advantage because usually only the non-zero values need to be stored & manipulated.
- This results in significant savings with respect to computation-time and storage.
- Some DM algorithms work well only for sparse data.

3) Resolution

- This is frequently possible to obtain data at different levels of resolution, and often the properties of the data are different at different resolutions.

- Ex: the surface of the earth seems very uneven at a resolution of few meters, but is relatively smooth at a resolution of tens of kilometers.
- The patterns in the data also depend on the level of resolution.
- If the resolution is too fine, a pattern may not be visible or may be buried in noise.

RECORD DATA

- Data-set is a collection of records.
  Each record consists of a fixed set of attributes.
- Every record has the same set of attributes.
- There is no explicit relationship among records or attributes.
- The data is usually stored either
  → in flat files or → in relational databases

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(a) Record data.

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

(b) Transaction data.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

(d) Document-term matrix.

Figure 2.2. Different variations of record data.

TYPES OF RECORD DATA

1) Transaction (Market Basket Data)

- Each transaction consists of a set of items.
- Consider a grocery store.
  The set of products purchased by a customer represents a transaction
  while the individual products represent items.
- This type of data is called market basket data because
  the items in each transaction are the products in a person's "market basket."
- Data can also be viewed as a set of records whose fields are asymmetric attributes.

2) Data Matrix

- An m*n matrix, where there are m rows, one for each object, &
  n columns, one for each attribute.
  This matrix is called a data-matrix.
- Since data-matrix consists of numeric attributes, standard matrix operation can be applied to manipulate the data.

3) Sparse Data Matrix

- This is a special case of a data-matrix.
- The attributes are of the same type and are asymmetric i.e. only non-zero values are important.

Document Data

- A document can be represented as a ‘vector’,
  where each term is a attribute of the vector and

value of each attribute is the no. of times corresponding term occurs in the document.

GRAPH BASED DATA

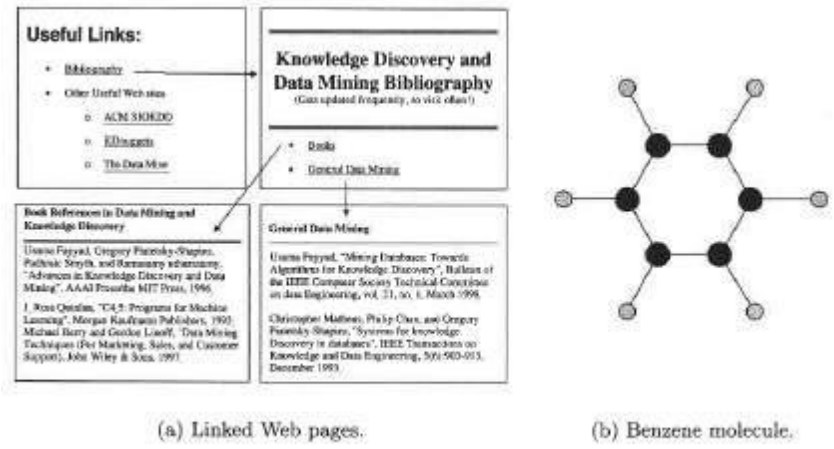
- Sometimes, a graph can be a convenient and powerful representation for data.
- We consider 2 specific cases:

1) Data with Relationships among Objects

- The relationships among objects frequently convey important information.
- In particular, the data-objects are mapped to nodes of the graph, while relationships among objects are captured by link properties such as direction & weight.
- For ex, in web, the links to & from each page provide a great deal of information about the relevance of a web-page to a query, and thus, must also be taken into consideration.

2) Data with Objects that are Graphs

- If the objects contain sub-objects that have relationships, then such objects are frequently represented as graphs.
- For ex, the structure of chemical compounds can be represented by a graph, where nodes are atoms and links between nodes are chemical bonds.



(a) Linked Web pages. (b) Benzene molecule.

Figure 2.3. Different variations of graph data.

ORDERED DATA

Sequential Data (Temporal Data)

- This can be thought of as an extension of record-data, where each record has a time associated with it.
- A time can also be associated with each attribute.
- For example, each record could be the purchase history of a customer, with a listing of items purchased at different times.
- Using this information, it is possible to find patterns such as "people who buy DVD players tend to buy DVDs in the period immediately following the purchase."

Sequence Data

- This consists of a data-set that is a sequence of individual entities, such as a sequence of words or letters.
- This is quite similar to sequential data, except that there are no time stamps; instead, there are positions in an ordered sequence.
- For example, the genetic information of plants and animals can be represented in the form of sequences of nucleotides that are known as genes.

Time Series Data

- This is a special type of sequential data in which a series of measurements are taken over time.
- For example, a financial data-set might contain objects that are time series of the daily prices of various stocks.
- An important aspect of temporal-data is temporal-autocorrelation i.e. if two measurements are close in time, then the values of those measurements are often very

similar.

Spatial Data

- Some objects have spatial attributes, such as positions or areas.
- An example is weather-data (temperature, pressure) that is collected for a variety of geographical location.
- An important aspect of spatial-data is spatial-autocorrelation i.e. objects that are physically close tend to be similar in other ways as well.

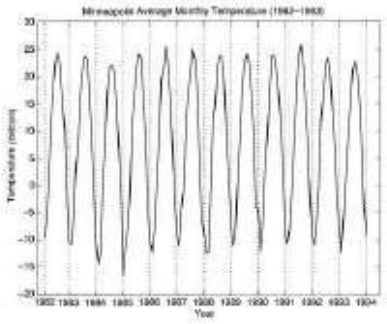
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

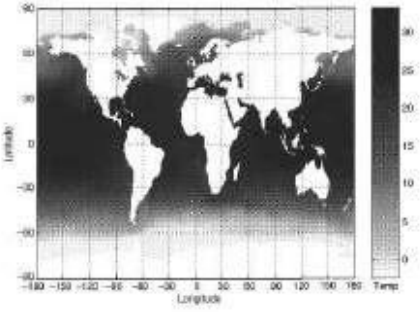
(a) Sequential transaction data.

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

(b) Genomic sequence data.



(c) Temperature time series.



(d) Spatial temperature data.

Figure 2.4. Different variations of ordered data.

DATA QUALITY

Data mining focuses on
(1) the detection and correction of data quality problems and
(2) the use of algorithms that can tolerate poor data quality.

The first step, detection and correction, is often called **data cleaning**.

Measurement and Data Collection Issues

It is unrealistic to expect that data will be perfect. There may be problems due to human error, limitations of measuring devices, or flaws in the data collection process. Values or even entire data objects may be missing. In other cases, there may be spurious or duplicate objects; i.e., multiple data objects that all correspond to a single "real" object.

For example, there might be two different records for a person who has recently lived at two different addresses. Even if all the data is present and "looks fine," there may be inconsistencies-a person has a height of 2 meters, but weighs only 2 kilograms.

The **term measurement error** refers to any problem resulting from the measurement process.

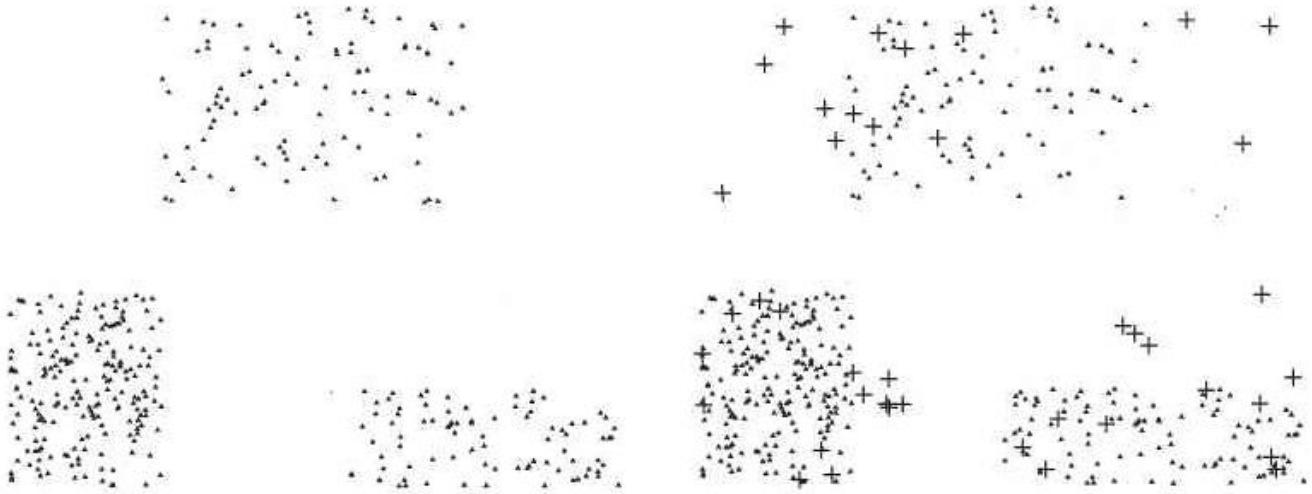
A common problem is that the value recorded differs from the true value to some extent. For continuous attributes, the numerical difference of the measured and true value is called the **error.** The term data collection error refers to errors such as omitting data objects or attribute values, or inappropriately including a data object.

Noise and Artifacts

Noise is the random component of a measurement error. It may involve the distortion of a value or the addition of spurious objects. Figure 2.5 shows a time series before and after it has been disrupted by random noise

a) Time series. (b) Time series with noise.

Figure2. 5. Noise in a times series context



(a) Three groups of points. (b) With noise points (+) added.

Figure2 .6. Noise in a spatial context

Data errors may be the result of a more deterministic phenomenon, such as a streak in the same place on a set of photographs. Such deterministic distortions of the data are often referred to as **artifacts**.

Definition 2.3 (Precision). The closeness of repeated measurements(of the same quantity) to one another.

Definition 2.4 (Bias). A systematic quantity being measured.

Definition 2.5 (Accuracy). The closeness of measurements to the true value of the quantity being measured.

Precision is often measured by the standard deviation of a set of values, while bias is measured by taking the difference between the mean of the set of values and the known value of the quantity being measured.

The more general term, **accuracy**, to refer to the degree of measurement error in data.there is no specific formula for accuracy in terms of these two quantities.

Outliers

Outliers are either (1) data objects that, in some sense, have characteristics that are different from most of the other data objects in the data set, or
(2) values of an attribute that are unusual with respect to the typical valuesfor that attribute. Alternatively, we can speak of anomalous objects or values

Missing Values
It is not unusual for an object to be missing one or more attribute values.In some cases, the information was not collected; e.g., some people decline to give their age or weight.

1) Eliminate Data Objects or Attributes A simple and effective strategy is to eliminate objects with missing values.

2) Estimate Missing Values Sometimes missing data can be reliably estimated.

3) Ignore the Missing Value during Analysis Many data mining approaches can be modified to ignore missing values. For example, suppose that objects are being clustered and the similarity between pairs of data objects needs to be calculated.
4) Inconsistent Values Data can contain inconsistent values. Consider an address field, where both a zip code and city are listed, but the specified zip code area is not contained in that city.

1) Duplicate Data A data set may include data objects that are duplicates, or almost duplicates, of one another

DATA PREPROCESSING

- Data preprocessing is a broad area and consists of a number of different strategies and techniques that are interrelated in complex way.
- Different data processing techniques are:
  1. Aggregation
  2. Sampling
  3. Dimensionality reduction
  4. Feature subset selection
  5. Feature creation
  6. Discretization and binarization
  7. Variable transformation

AGGREGATION

- This refers to combining 2 or more attributes (or objects) into a single attribute (orobject).
  For example, merging daily sales figures to obtain monthly sales figures
- Motivations for aggregation:
  1) Data reduction: The smaller data-sets require
     → less memory → less processing time.
     Because of aggregation, more expensive algorithm can be used.
  2) Aggregation can act as a *change of scale* by providing a high-level view of the data instead of a low-level view. E.g. Cities aggregated into districts, states, countries, etc
  3) The behavior of groups of objects is often *more stable* than that of individual objects.
- Disadvantage: The potential loss of interesting details.

SAMPLING

- This is a method used for selecting a subset of the data-objects to be analyzed.
- This is often used for both
  → preliminary investigation of the data → final data analysis
- Q: Why sampling?
  Ans: Obtaining & processing the entire set of "data of interest" is too expensive or time consuming.
- Sampling can reduce data-size to the point where better & more expensive algorithm can be used.
- Key principle for effective sampling: Using a sample will work almost as well as using entire data-set, if the *sample is representative*.

Sampling Methods

1) Simple Random Sampling
- There is an equal probability of selecting any particular object.
- There are 2 variations on random sampling:

i) Sampling without Replacement
- As each object is selected, it is removed from the population.

ii) Sampling with Replacement
- Objects are not removed from the population as they are selected for the sample.
- The same object can be picked up more than once.

• When the population consists of different types(or number) of objects, simple random sampling can fail to adequately represent those types of objects that are less frequent.

2) Stratified Sampling

• This starts with pre-specified groups of objects.

• In the simplest version, equal numbers of objects are drawn from each group even though the groups are of different sizes.

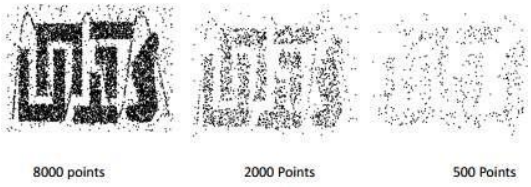
• In another variation, the number of objects drawn from each group is proportional to the size of that group.

3) Progressive Sampling

• If proper sample-size is difficult to determine then progressive sampling can be used.

• This method starts with a small sample, and
then increases the sample-size until a sample of sufficient size has been obtained.

• This method requires a way to evaluate the sample to judge if it is large enough.



DIMENSIONALITY REDUCTION

• Key benefit: many DM algorithms work better if the dimensionality is lower.

Purpose

• May help to eliminate irrelevant features or reduce noise.
• Can lead to a more understandable model (which can be easily visualized).
• Reduce amount of time and memory required by DM algorithms.
• Avoid curse of dimensionality.

The Curse of Dimensionality

• Data-analysis becomes significantly harder as the dimensionality of the data increases.

• For classification, this can mean that there are not enough data-objects to allow the creation of a model that reliably assigns a class to all possible objects.

• For clustering, the definitions of density and the distance between points (which are critical for clustering) become less meaningful.

• As a result, we get
→ reduced classification accuracy &
→ poor quality clusters.

FEATURE SUBSET SELECTION

• Another way to reduce the dimensionality is to use only a subset of the features.

• This might seem that such approach would lose information, this is not the case if redundant and irrelevant features are present.

1) *Redundant features* duplicate much or all of the information contained in one or more other attributes.
For example: purchase price of a product and the amount of sales tax paid.
2) *Irrelevant features* contain almost no useful information for the DM task at hand.
For example: students' ID numbers are irrelevant to the task of predicting students' grade point averages.

Techniques for Feature Selection

1) Embedded approaches: Feature selection occurs naturally as part of DM algorithm. Specifically, during the operation of the DM algorithm, the algorithm itself decides which attributes to use and which to ignore.
2) Filter approaches: Features are selected before the DM algorithm is run.
3) Wrapper approaches: Use DM algorithm as a black box to find best subset of attributes.

An Architecture for Feature Subset Selection

• The feature selection process is viewed as consisting of 4 parts:
1) A measure of evaluating a subset,
2) A search strategy that controls the generation of a new subset of features,
3) A stopping criterion and
4) A validation procedure.

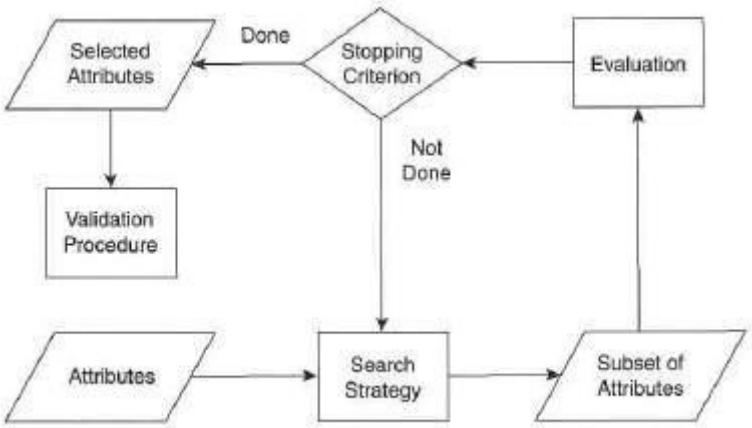


Figure 2.11. Flowchart of a feature subset selection process.

DISCRETIZATION AND BINARIZATION

- Some DM algorithms (especially classification algorithms) require that the data be in the form of categorical attributes.
- Algorithms that find association patterns require that the data be in the form of binary attributes.
- Transforming continuous attributes into a categorical attribute is called *discretization.*

And transforming continuous & discrete attributes into binary attributes is called as *binarization.*

Binarization

- A simple technique to binarize a categorical attribute is the following: If there are m categorical values, then uniquely assign each original value to an integer in interval [0,m-1].
- Next, convert each of these m integers to a binary number.
- Since n=[log2(m)] binary digits are required to represent these integers, represent these binary numbers using 'n' binary attributes.

Table 2.5. Conversion of a categorical attribute to three binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Discretizction of continuous attributes

- Discretization is typically applied to attributes that are used in classification or association analysis.
- In general, the best discretization depends on
  - → the algorithm being used, as well as
  - → the other attributes being considered
- Transformation of a continuous attribute to a categorical attribute involves two subtasks:
  - → deciding how many categories to have and
  - → determining how to map the values of the continuous attribute to these categories

VARIABLE TRANSFORMATION

- This refers to a transformation that is applied to all the values of a variable.
- Ex: converting a floating point value to an absolute value.
- Two types are:

1) Simple Functions

- A simple mathematical function is applied to each value individually.
- If x is a variable, then examples of transformations include e^x, 1/x, log(x), sin(x).

2) Normalization (or Standardization)

- The goal is to make an entire set of values have a particular property.
- A traditional example is that of "standardizing a variable" in statistics.
- If $\bar{x}$ is the mean of the attribute values and s_x is their standard deviation, then the transformation $x'=(x-\bar{x})/s_x$ creates a new variable that has a mean of 0 and a standard deviation of 1.

MEASURE OF SIMILARITY AND DISSIMILARITY

- Similarity & dissimilarity are important because they are used by a no. of DM techniques such as clustering, classification & anomaly detection.
- *Proximity* is used to refer to either similarity or dissimilarity.
- The *similarity* between 2 objects is a numerical measure of degree to which the 2 objects are alike.
- Consequently, similarities are higher for pairs of objects that are more alike.
- Similarities are usually non-negative and are often between 0(no similarity) and 1(complete similarity).
- The *dissimilarity* between 2 objects is a numerical measure of the degree to which the 2 objects are different.
- Dissimilarities are lower for more similar pairs of objects.
- The term distance is used as a synonym for dissimilarity.
- Dissimilarities sometimes fall in the interval [0,1] but is also common for them to range from 0 to infinity.

Table 2.7. Similarity and dissimilarity for simple attributes

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - \min d}{\max d - \min d}$

DISSIMILARITIES BETWEEN DATA OBJECTS

Distances

- The Euclidean distance, d, between 2 points, x and y, in one-,two-,three- or higher-dimensional space, is given by the following familiar formula:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}, \qquad (2.1)$$

where n=number of dimensions

x_k and y_k are respectively the k^{th} attributes of x and y.

- The Euclidean distance measure given in equation 2.1 is generalized by the Minkowski distance metric given by

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}, \qquad (2.2)$$

where r=parameter.

- The following are the three most common examples of minkowski distance:

r=1. City block(Manhattan L_1 norm) distance.
A common example is the Hamming distance, which is the number of bits that are different between two objects that have only binary attributes ie between two binary vectors.
r=2. Euclidean distance (L_2 norm)
$r=\infty$. Supremum(L_∞ or L_{max} norm) distance. This is the maximum difference between any attribute of the objects. Distance is defined by

$$d(\mathbf{x},\mathbf{y}) = \lim_{r\to\infty}\left(\sum_{k=1}^{n}|x_k - y_k|^r\right)^{1/r}. \qquad (2.3)$$

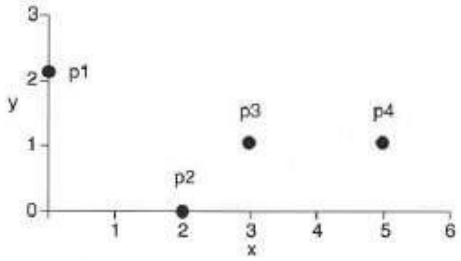


Figure 2.15. Four two-dimensional points.

Table 2.8. x and y coordinates of four points.

point	x coordinate	y coordinate
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Table 2.9. Euclidean distance matrix for Table 2.8.

	p1	p2	p3	p4
p1	0.0	2.8	3.2	5.1
p2	2.8	0.0	1.4	3.2
p3	3.2	1.4	0.0	2.0
p4	5.1	3.2	2.0	0.0

Table 2.10. L_1 distance matrix for Table 2.8.

L_1	p1	p2	p3	p4
p1	0.0	4.0	4.0	6.0
p2	4.0	0.0	2.0	4.0
p3	4.0	2.0	0.0	2.0
p4	6.0	4.0	2.0	0.0

Table 2.11. L_∞ distance matrix for Table 2.8.

L_∞	p1	p2	p3	p4
p1	0.0	2.0	3.0	5.0
p2	2.0	0.0	1.0	3.0
p3	3.0	1.0	0.0	2.0
p4	5.0	3.0	2.0	0.0

- If d(x,y) is the distance between two points, x and y, then the following properties hold
  1) Positivity
     d(x,x)>=0 for all x and y d(x,y)=0 only if x=y
  2) Symmetry
     d(x,y)=d(y,x) for all x and y.
  3) Triangle inequality
     d(x,z)<=d(x,y)+d(y,z) for all points x,y and z.
- Measures that satisfy all three properties are known as *metrics.*

SIMILARITIES BETWEEN DATA OBJECTS

- For similarities, the triangle inequality typically does not hold, but symmetry positivity typically do.
- If s(x,y) is the similarity between points x and y, then the typical properties of similarities are the following
  1) s(x,y)=1 only if x=y
  2) s(x,y)=s(y,x) for all x and y. (Symmetry)
- For ex, cosine and Jaccard similarity.

EXAMPLES OF PROXIMITY MEASURES

- Similarity measures between objects that contain only binary attributes are called similarity coefficients.
- Typically, they have values between 0 and 1.
- A value of 1 indicates that the two objects are completely similar,
  while a value of 0 indicates that the objects are not at all similar.
- Let x and y be 2 objects that consist of n binary attributes.
- Comparison of 2 objects, ie, 2 binary vectors, leads to the following four quantities (frequencies):
  f00=the number of attributes where x is 0 and y is 0.
  f01=the number of attributes where x is 0 and y is 1.
  f10=the number of attributes where x is 1 and y is 0.
  f11=the number of attributes where x is 1 and y is 1.

SIMPLE MATCHING COEFFICIENT

- One commonly used similarity coefficient is the SMC, which is defined as

$$SMC = \frac{\text{number of matching attribute values}}{\text{number of attributes}} = \frac{f_{11}+f_{00}}{f_{01}+f_{10}+f_{11}+f_{00}}. \qquad (2.5)$$

- This measure counts both presences and absences equally.

JACCARD COEFFICIENT

- Jaccard coefficient is frequently used to handle objects consisting of asymmetric binary attributes.
- The jaccard coefficient is given by the following equation:

$$J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 matches}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}. \quad (2.6)$$

Example 2.17 (The SMC and Jaccard Similarity Coefficients). To illustrate the difference between these two similarity measures, we calculate SMC and J for the following two binary vectors.

$\mathbf{x} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$
$\mathbf{y} = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)$

$f_{01} = 2$ the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 1
$f_{10} = 1$ the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 0
$f_{00} = 7$ the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 0
$f_{11} = 0$ the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 1

$SMC = \frac{f_{11}+f_{00}}{f_{01}+f_{10}+f_{11}+f_{00}} = \frac{0+7}{2+1+0+7} = 0.7$

$J = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} = \frac{0}{2+1+0} = 0$ ■

COSINE SIMILARITY

- Documents are often represented as vectors, where each attribute represents the frequency with which a particular term (or word) occurs in the document.
- This is more complicated, since certain common words are ignored and various processing techniques are used to account for
  - → different forms of the same word
  - → differing document lengths and
  - → different word frequencies.
- The cosine similarity is one of the most common measure of document similarity.
- If x and y are two document vectors, then

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}, \quad (2.7)$$

where $\cdot$ indicates the vector dot product, $\mathbf{x} \cdot \mathbf{y} = \sum_{k=1}^{n} x_k y_k$, and $\|\mathbf{x}\|$ is the length of vector $\mathbf{x}$, $\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}}$.

Example 2.18 (Cosine Similarity of Two Document Vectors). This example calculates the cosine similarity for the following two data objects, which might represent document vectors:

$\mathbf{x} = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$
$\mathbf{y} = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$

$\mathbf{x} \cdot \mathbf{y} = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$
$\|\mathbf{x}\| = \sqrt{3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0} = 6.48$
$\|\mathbf{y}\| = \sqrt{1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2} = 2.24$
$\cos(\mathbf{x}, \mathbf{y}) = 0.31$

- As indicates by figure 2.16, cosine similarity really is a measure of the angle between x and y.
- Thus, if the cosine similarity is 1,the angle between x and y is 0',and x and y are the same except for magnitude(length).
- If cosine similarity is 0, then the angle between x and y is 90' and they do not share any terms.

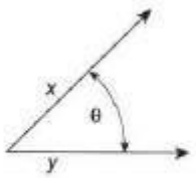


Figure 2.16. Geometric illustration of the cosine measure.

EXTENDED JACCARD COEFFICIENT(TANIMOTO COEFFICIENT)

- This can be used for document data.
- this coefficient is defined by following equation

$$EJ(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x}\cdot\mathbf{y}}. \qquad (2.9)$$

ISSUES IN PROXIMITY CALCULATION

1) How to handle the case in which attributes have different scales and/or are correlated.
2) How to calculate proximity between objects that are composed of different types of attributes e.g. quantitative and qualitative.
3) How to handle proximity calculation when attributes have different weights.

COMBINING SIMILARITIES FOR HETEROGENEOUS ATTRIBUTES

- A general approach is needed when the attributes are of different types.
- One straightforward approach is to compute the similarity between each attribute separately and then combine these similarities using a method that results in a similarity between 0 and 1.
- Typically, the overall similarity is defined as the average of all the individual attribute similarities.

DATA MINING APPLICATIONS

Prediction & Description

- Data mining may be used to answer questions like
  → "would this customer buy a product" or
  → "is this customer likely to leave?"
- DM techniques may also be used for sales forecasting and analysis.

Relationship Marketing

- Customers have a lifetime value, not just the value of a single sale.
- Data mining can help
  → in analyzing customer profiles and improving direct marketing plans
  → in identifying critical issues that determine client loyalty and
  → in improving customer retention

Customer Profiling

- This is the process of using the relevant and available information
  → to describe the characteristics of a group of customers
  → to identify their discriminators from ordinary consumers and
  → to identify drivers for their purchasing decisions
- This can help an enterprise identify its most valuable customers so that the enterprise may differentiate their needs and values.

Outliers Identification & Detecting Fraud

- For this, examples include:
  → identifying unusual expense claims by staff
  → identifying anomalies in expenditure between similar units of an enterprise
  → identifying fraud involving credit cards

Customer Segmentation

- This is a way to assess & view individuals in market based on their status & needs.
- Data mining may be used
  → to understand & predict customer behavior and profitability
  → to develop new products & services and
  → to effectively market new offerings

Web site Design & Promotion

- Web mining may be used to discover how users navigate a web site and the results can help in improving the site design.

- Web mining may also be used in cross-selling by suggesting to a web customer, items that he may be interested in.

EXERCISES

1. Explain Data set. Give an Example? (5)
2. Explain 4 types of attributes by giving appropriate example? (10)
3. With example, explain
   i) Continious ii) Discrete iii) Asymmetric Attributes. (10)
4. Explain general characteristics of data sets. (6)
5. Explain record data & its types. (10)
6. Explain graph based data. (6)
7. Explain ordered data & its types. (10)
8. Explain shortly any five data pre-processing approaches. (10)
9. What is sampling? Explain simple random sampling vs. stratified sampling vs. progressive sampling. (10)
10. Write a short note on the following: (15)
    i) Dimensionality reduction ii)Variable transformation iii)Feature selection
11. Distinguish between
    i) SMC & Jaccard coefficient ii) Discretization & binarization (6)
12. Explain various distances used for measuring dissimilarity between objects. (6)
13. Consider the following 2 binary vectors X=(1,0,0,0,0,0,0,0,0,0)
    Y=(0,0,0,0,0,0,1,0,0,1)
    Find i)hamming distance ii)SMC iii)Jaccard coefficient (4)
14. List out issues in proximity calculation. (4)
15. List any 5 applications of data mining. (8)