UNIT -1

1a. Define data mining?

Ans: Data mining refers to extracting or "mining" knowledge from large amounts of data.

1b. Explain the types of the data can be Mined.

Ans: The most basic forms of data for mining applications are database data , data warehouse data and transactional data .

1)Database Data :

A database system also called a database management system (DBMS), consists of a collection of interrelated data known as a database and a set of software programs to manage and access the data.

A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).

EX:customer (cust ID, name, address, age, occupation, annual income, credit information, category, . . .)

item (item ID, brand, category, type, price, place made, supplier, cost, . .)

2)Data Warehouses:

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site.

Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.

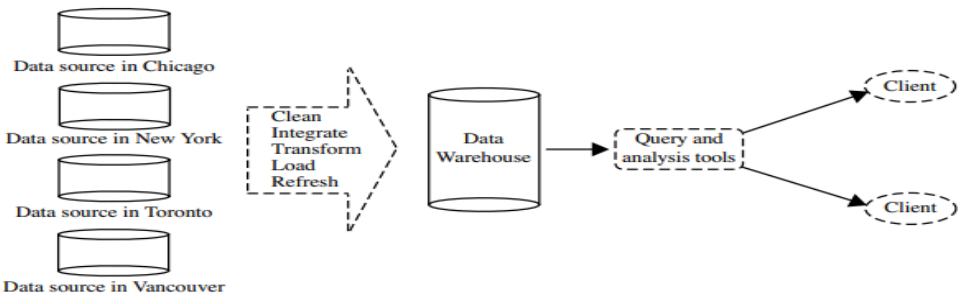


Figure 1.6 Typical framework of a data warehouse for *AllElectronics*.

A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum(sales amount).

A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data.

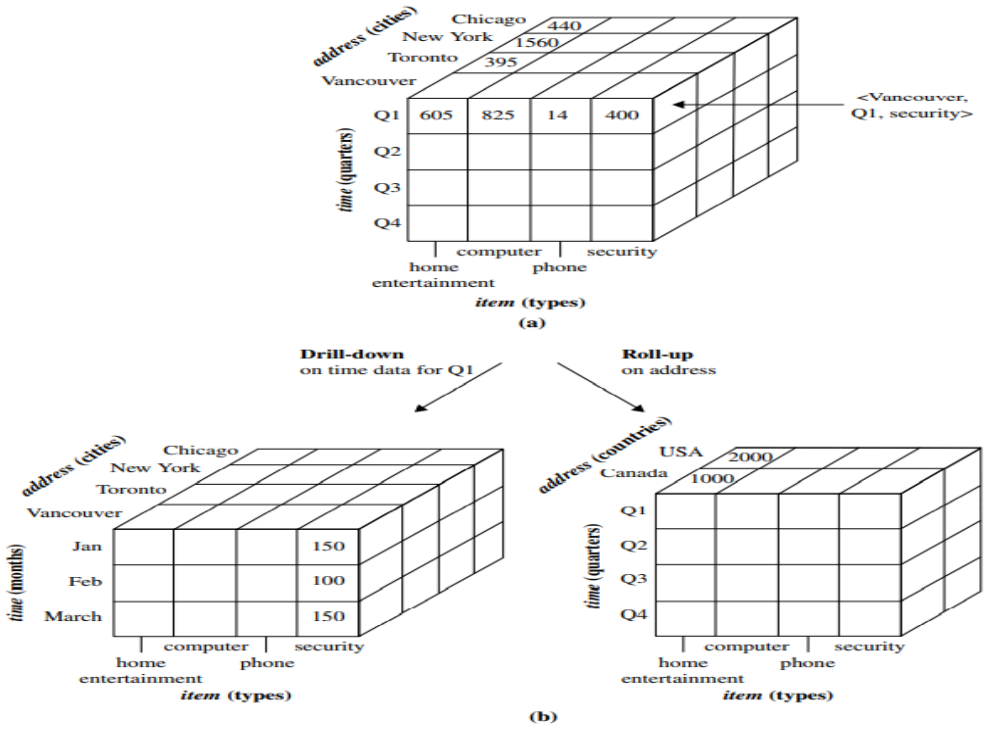


Figure 1.7 A multidimensional data cube, commonly used for data warehousing, (a) showing summa-

3)Transactional Data.

In general, each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page.

A transaction typically includes a unique transaction identity number (trans ID) and a list of the items making up the transaction, such as the items purchased in the transaction.

trans_ID	*list_of_item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

Figure 1.8 Fragment of a transactional database for sales at *AllElectronics*.

1C. Discuss briefly about data cleaning techniques.

Data Cleaning: Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data

a) **Methods for filling missing values:**
   i) **Ignore the tuple:** . This method is not very effective, unless the tuple contains several attributes with missing values
   ii) **Fill in the missing value manually:** In general, this approach is time-consuming and may not be feasible given a large data set with many missing values.
   iii) **Use a global constant to fill in the missing value:** Replace all missing attribute values by the same constant, such as a label like "Unknown" or $-\infty$
   iv) **Use the attribute mean to fill in the missing value:** For example, suppose that the average income of AllElectronics customers is $56,000. Use this value to replace the missing value for income.
   v) **Use the attribute mean for all samples belonging to the same class as the given tuple:** For example, if classifying customers according to credit risk, replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple.
   vi) **Use the most probable value to fill in the missing value:** This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction

b) **Noisy data :**
   Noise is a random error or variance in a measured variable. In order to remove the noise data smoothing techniques are used
   Data Smoothing Techniques:
      i) **Binning:**
         - Binning methods smooth a sorted data value by consulting its neighborhood i.e. the values around it.
         - The sorted values are distributed into a number of "buckets," or bins.
         - **smoothing by bin means**:In this metho each value in a bin is replaced by the mean value of the bin
         - **smoothing by bin medians** : In this method each bin value is replaced by the bin median
         - **smoothing by bin boundaries**: In this method the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Figure 3.2 Binning methods for data smoothing.

ii) Regression:

- Data smoothing can also be done by regression, a technique that conforms data values to a function.
- Linear regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other.
- Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface

iii) Outlier analysis:

- Outliers may be detected by clustering, for example, where similar values are organized into groups, or "clusters."
- Intuitively, values that fall outside of the set of clusters may be considered outliers

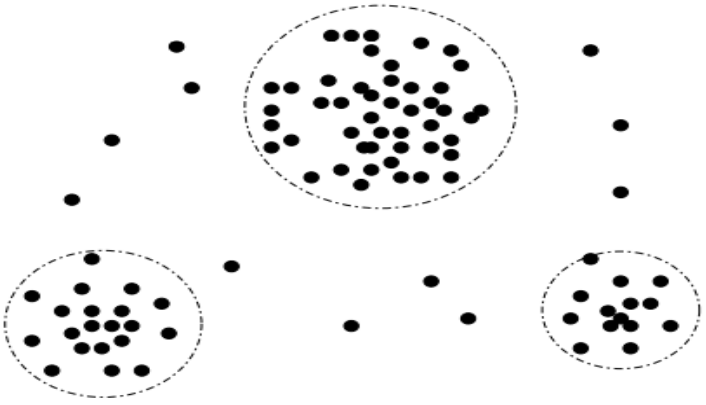


Figure 3.3 A 2-D customer data plot with respect to customer locations in a city, showing three data clusters. Outliers may be detected as values that fall outside of the cluster sets.

2a) Define Data Mining Task primitives.

Ans: 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.

2 b) Explain about Data Mining Functionalities.

Ans: Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks.

- In general, such tasks can be classified into two categories: descriptive and predictive.
- Descriptive mining tasks characterize properties of the data in a target data set.
- Predictive mining tasks perform induction on the current data in order to make predictions.
- There are a number of data mining functionalities.They are

1)Data Characterization and Data Discrimination:

- Data characterization is a summarization of the general characteristics or features of a target class of data.
- Example: A customer relationship manager at AllElectronics may order the following data mining task: Summarize the characteristics of customers who spend more than $5000 a year at AllElectronics. The result is a general profile of these customers, such as that they are 40 to 50 years old, employed, and have excellent credit ratings.
- Data Discrimination is comparison of the target class with one or a set of comparative classes (often called the contrasting classes)
- Example:
- A customer relationship manager at AllElectronics may want to compare two groups of customers—those who shop for computer products regularly (e.g., more than twice a month) and those who rarely shop for such products (e.g., less than three times a year). The resulting description provides a general comparative profile of these customers, such as that 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education, whereas 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree.

2) Mining Frequent Patterns, Associations, and Correlations:

- Frequent patterns, as the name suggests, are patterns that occur frequently in data.
- There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures.
- A frequent itemset typically refers to a set of items that often appear together in a transactional data set—for example, milk and bread, which are frequently bought together in grocery stores by many customers.
- A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern.
- A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences.
- Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

3) Classification and Regression:

- Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.
- The model are derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known).
- The model is used to predict the class label of objects for which the the class label is unknown.
- The derived model may be represented in various forms, such as classification rules (i.e., IF-THEN rules), decision trees, mathematical formulae, or neural networks.

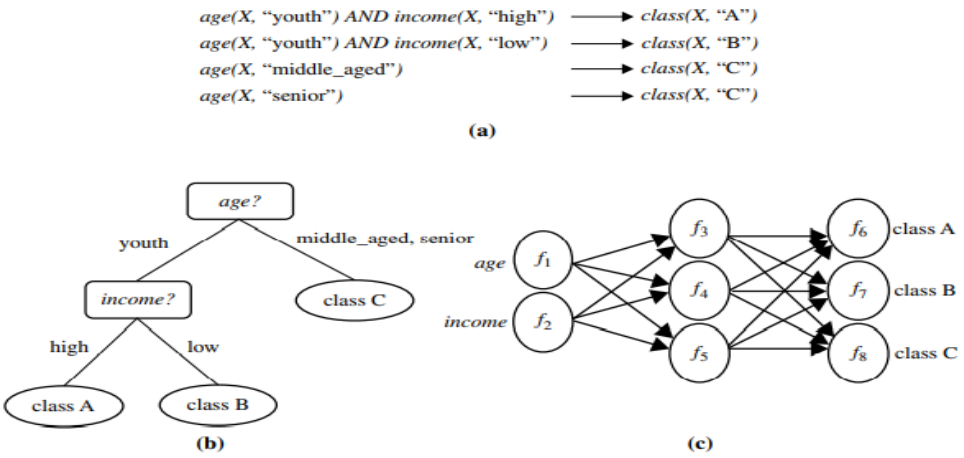


Figure 1.9 A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

-
- Whereas classification predicts categorical (discrete, unordered) labels, regression models continuous-valued functions.
- That is, regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels.

4)Cluster Analysis :

- Clustering can be used to generate class labels for a group of data.
- The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.
  That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather
  dissimilar to objects in other clusters.
- Each cluster so formed can be viewed as a class of objects, from which rules can be derived.

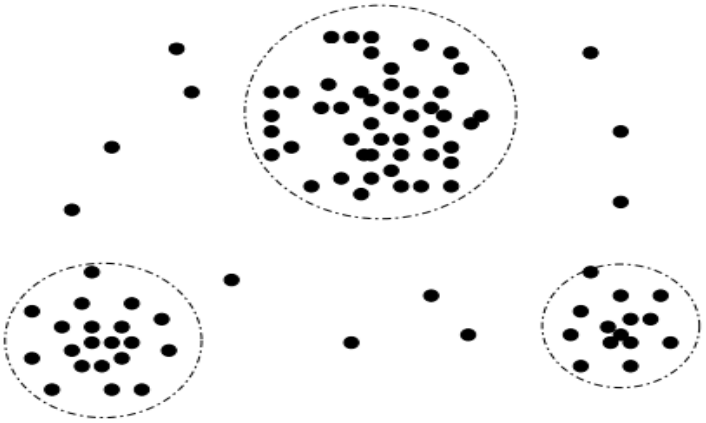


Figure 3.3 A 2-D customer data plot with respect to customer locations in a city, showing three data clusters. Outliers may be detected as values that fall outside of the cluster sets.

5)Outlier Analysis

- A data set may contain objects that do not comply with the general behavior or model of the data.
- These data objects are outliers. Many data mining methods discard outliers as noise or exceptions.
- The analysis of outlier data is referred to as outlier analysis or anomaly mining.

2c) Explain Briefly about Data Cube aggregation and attribute subset selection?

Ans:

1. **Data cube aggregation :** where aggregation operations are applied to the data in the construction of a data cube.
    - o For example, a data cube for multidimensional analysis of sales data with respect to annual sales per item type for each AllElectronics branch.
    - o Each cell holds an aggregate data value, corresponding to the data point in multidimensional space.
    - o Concept hierarchies may exist for each attribute, allowing the analysis of data at multiple levels of abstraction.
    - o For example, a hierarchy for branch could allow branches to be grouped into regions, based on their address. Data cubes provide fast access to precomputed, summarized data

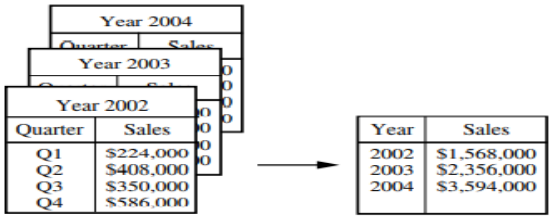


Figure 2.13 Sales data for a given branch of *AllElectronics* for the years 2002 to 2004. On the left, the sales are shown per quarter. On the right, the data are aggregated to provide the annual sales.

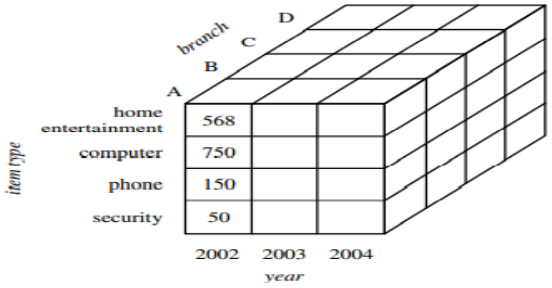


Figure 2.14 A data cube for sales at *AllElectronics*.

Attribute subset selection: where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.

Attribute subset selection include the following techniques:
1.**Stepwise forward selection**: The procedure starts with an empty set of attributes as the reduced set. The best of the original attributes is determined and added to the reduced set. At each subsequent iteration or step, the best of the remaining original attributes is added to the set.
2. **Stepwise backward elimination:** The procedure starts with the full set of attributes. At each step, it removes the worst attribute remaining in the set.
3.**Combination of forward selection and backward elimination:** The stepwise forward selection and backward elimination methods can be combined so that, at each step, the procedure selects the best attribute and removes the worst from among the remaining attributes.
4. **Decision tree induction:** Decision tree induction constructs a flow chart like structure where each internal (nonleaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external (leaf) node denotes a class prediction. At each node, the algorithm chooses the “best” attribute to partition

the data into individual classes\

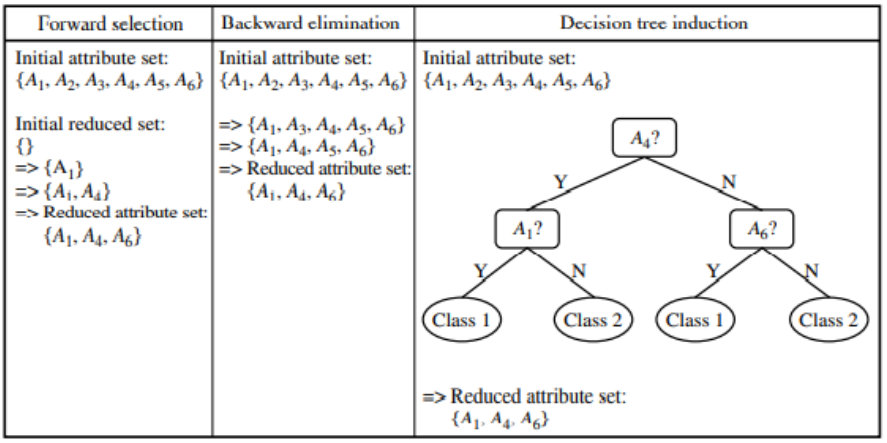


Figure 2.15 Greedy (heuristic) methods for attribute subset selection.

3a) What is data cleaning?

Ans: Data Cleaning is a process to fill in missing values , smooth noisy data while identifying outlier, and correct inconsistencies in the data.

3b) Explain about Classification of Data mining Systems

Ans: Data mining is an interdisciplinary field ,the confluence of a set of disciplines,

including database systems, statistics, machine learning, visualization, and information science.

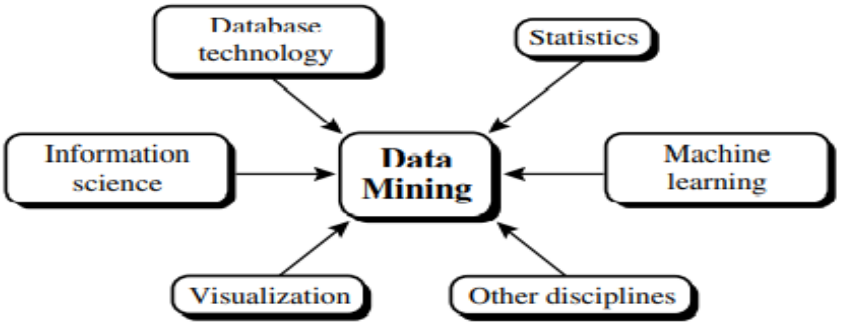


Figure 1.12 Data mining as a confluence of multiple disciplines.

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

i)Classification according to the kinds of databases mined:

- Database systems can be classified according to data models, we may have a relational, transactional, object-relational, or data warehouse mining system.
- Each of which may require its own data mining technique.

ii) Classification according to the kinds of knowledge mined:

Data mining systems can be categorized according to the kinds of knowledge they mine, that is, based on data mining functionalities
such as characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis.

iii)Classification according to the kinds of techniques utilized:

- Data mining systems can be categorized according to the underlying data mining techniques employed.
- These techniques can be described according to the degree of user interaction involved (e.g., autonomous systems, interactive exploratory systems, query-driven systems)

IV) **Classification according to the applications adapted:**

- Data mining systems can also be categorized according to the applications they adapt.
- For example, data mining systems may be tailored specifically for finance, telecommunications, DNA, stock markets, e-mail, and so on.

3c) Explain Briefly about Data Transformation Techniques.

Ans: In this preprocessing step, the data are transformed or consolidated so that the resulting mining process may be more efficient, and the patterns found may be easier to understand.

- **Data Transformation Strategies:**
  - **i) Data Smoothing:**
    - **a)Binning:**
      - Binning methods smooth a sorted data value by consulting its neighborhood i.e. the values around it.
      - The sorted values are distributed into a number of "buckets," or bins.
      - **smoothing by bin means**:In this metho each value in a bin is replaced by the mean value of the bin
      - **smoothing by bin medians** : In this method each bin value is replaced by the bin median
      - **smoothing by bin boundaries**: In this method the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Figure 3.2 Binning methods for data smoothing.

b) Regression:

- Data smoothing can also be done by regression, a technique that conforms data values to a function.
- Linear regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other.
- Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface

c)Outlier analysis:

- Outliers may be detected by clustering, for example, where similar values are organized into groups, or "clusters."
- Intuitively, values that fall outside of the set of clusters may be considered outliers

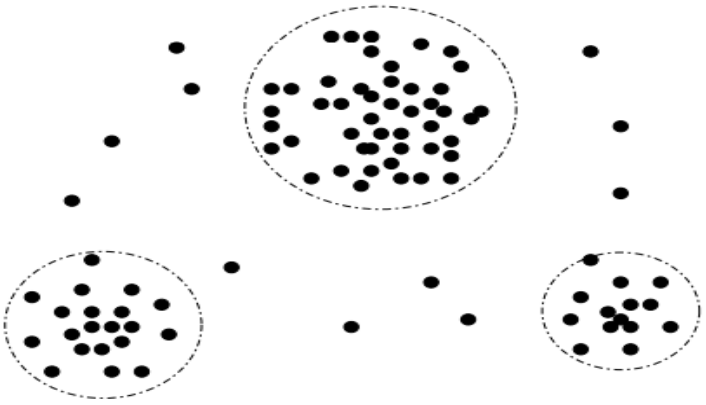


Figure 3.3 A 2-D customer data plot with respect to customer locations in a city, showing three data clusters. Outliers may be detected as values that fall outside of the cluster sets.

ii) Attribute construction : where new attributes are constructed and added from the given set of attributes to help the mining process.

iii) Aggregation: where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts.

iv) Normalization : where the attribute data are scaled so as to fall within a smaller range, such as −1.0 to 1.0, or 0.0 to 1.0.

There are many methods for data normalization . Some are

a) Min-max normalization:

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A. \quad (3.8)$$

Suppose that minA and maxA are the minimum and maximum values of an attribute, A.

Min-max normalization maps a value, vi , of A to v'_i in the range [new_minA, new_maxA]

Example 3.4 **Min-max normalization.** Suppose that the minimum and maximum values for the attribute *income* are \$12,000 and \$98,000, respectively. We would like to map *income* to the range [0.0,1.0]. By min-max normalization, a value of \$73,600 for *income* is transformed to $\frac{73,600-12,000}{98,000-12,000}(1.0-0)+0=0.716$. ∎

b) z-score normalization (or zero-mean normalization):

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}, \quad (3.9)$$

where $\bar{A}$ and σ_A are the mean and standard deviation, respectively, of attribute *A*. The mean and standard deviation were discussed in Section 2.2, where $\bar{A} = \frac{1}{n}(v_1 + v_2 + \cdots +$

Example 3.5 **z-score normalization.** Suppose that the mean and standard deviation of the values for the attribute *income* are \$54,000 and \$16,000, respectively. With z-score normalization, a value of \$73,600 for *income* is transformed to $\frac{73,600-54,000}{16,000} = 1.225$. ∎

C) Normalization by decimal scaling : normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value, v_i , of A is normalized to v'_i by computing.

$$v'_i = \frac{v_i}{10^j}, \tag{3.12}$$

where j is the smallest integer such that $max(|v'_i|) < 1$.

Example 3.6 **Decimal scaling.** Suppose that the recorded values of A range from −986 to 917. The maximum absolute value of A is 986. To normalize by decimal scaling, we therefore divide each value by 1000 (i.e., $j = 3$) so that −986 normalizes to −0.986 and 917 normalizes to 0.917. ■

4a) Define dimensionality reduction

- **Ans:** In dimensionality reduction, data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data.

4b) Describe the various phases in knowledge discovery process with a neat diagram.

Ans:

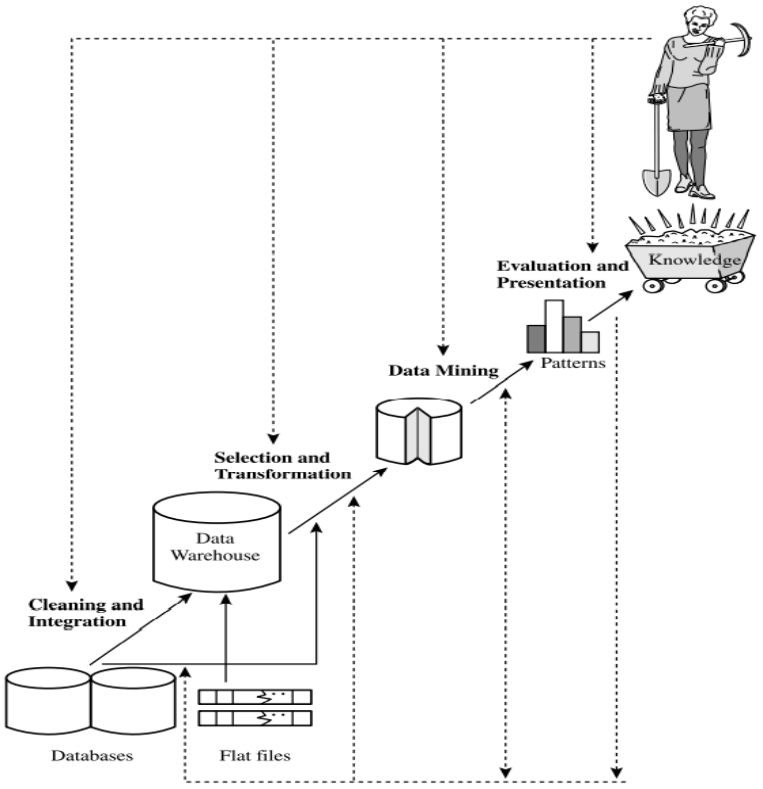


Figure 1.4 Data mining as a step in the process of knowledge discovery.

KDD Process steps:

1) **Data cleaning:** Data cleaning process used to remove noise and inconsistent data
2) **Data integration :** In this process multiple data sources may be combined
3) Data selection : In this process data relevant to the analysis task are retrieved from the database
4) **Data transformation :** In this process data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5) **Data mining :** it is an essential process where intelligent methods are applied in order to extract data patterns
6) **Pattern evaluation :** In this process to identify the truly interesting patterns representing knowledge based on some interestingness measures
7) **Knowledge presentation:** In this process where visualization and knowledge representation techniques are used to present the mined knowledge to the user

4c) What are the typical methods for Data Discretization and Concept Hierarchy Generation?

Ans Data Discretization and Concept Hierarchy Generation:

- Data discretization: Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals.
- Interval labels can then be used to replace actual data values.
- If the discretization process uses class information, then we say it is supervised discretization.
- Otherwise, it is unsupervised. If the process starts by first finding one or a few points (called split points or cut points) to split the entire attribute range, and then repeats this recursively on the resulting intervals

- **Concept hierarchy :**A concept hierarchy for a given numerical attribute defines a discretization of the attribute.
- Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute age) with higher-level concepts (such as youth, middle-aged, or senior).

Discretization and Concept Hierarchy Generation for Numerical Data

- **The following are methods for Data Discretization and Concept Hierarchy Generation:**

i) Binning:

- Binning methods smooth a sorted data value by consulting its neighborhood i.e. the values around it.
- The sorted values are distributed into a number of "buckets," or bins.
- **smoothing by bin means**:In this metho each value in a bin is replaced by the mean value of the bin
- **smoothing by bin medians** : In this method each bin value is replaced by the bin median
- **smoothing by bin boundaries**: In this method the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Figure 3.2 Binning methods for data smoothing.

ii) Histogram Analysis:

- A histogram for an attribute, A, partitions the data distribution of A into disjoint subsets, or buckets.
- There are several partitioning rules, including the following:
- Equal-width: In an equal-width histogram, the width of each bucket range is uniform (such as the width of $10 for the bucket)
- Equal-frequency (or equidepth): In an equal-frequency histogram, the buckets are created so that, roughly, the frequency of each bucket is constant.
- V-Optimal: If we consider all of the possible histograms for a given number of buckets, the V-Optimal histogram is the one with the least variance. Histogram variance is a weighted sum of the original values that each bucket represents, where bucket weight is equal to the number of values in the bucket.

- MaxDiff: In a MaxDiff histogram, we consider the difference between each pair of adjacent values. A bucket boundary is established between each pair for pairs having the β−1 largest differences, where β is the user-specified number of buckets.
-

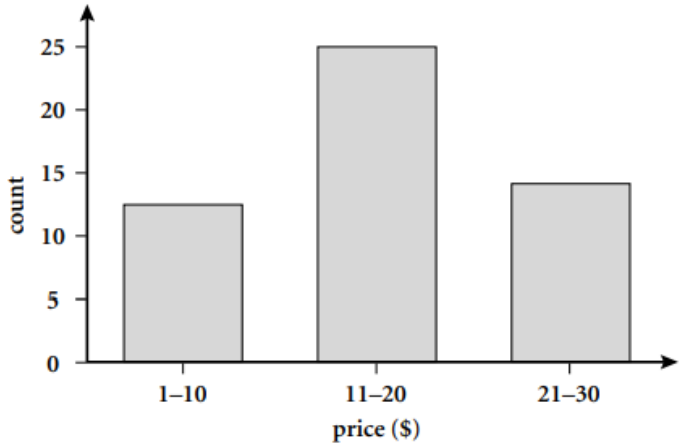


Figure 2.19 An equal-width histogram for *price*, where values are aggregated so that each bucket has a uniform width of $10.

iii)Cluster Analysis:

- Clustering techniques consider data tuples as objects.
- They partition the objects into groups or clusters, so that objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters.
- Similarity is commonly defined in terms of how "close" the objects are in space, based on a distance function.
- The "quality" of a cluster may be represented by its diameter, the maximum distance between any two objects in the cluster.
- Centroid distance is an alternative measure of cluster quality and is defined as the average distance of each cluster object from the cluster centroid

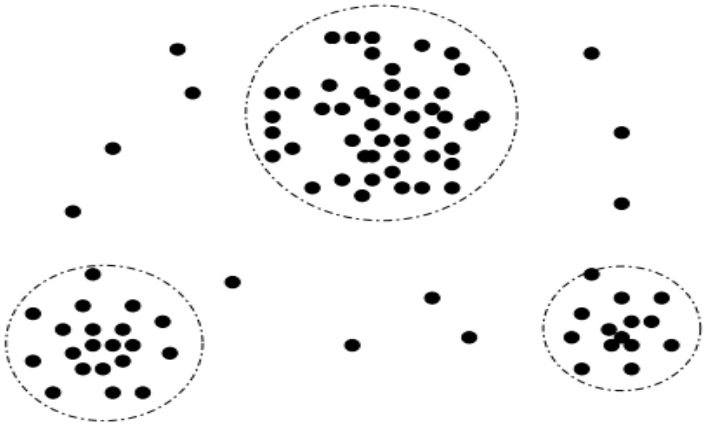


Figure 3.3 A 2-D customer data plot with respect to customer locations in a city, showing three data clusters. Outliers may be detected as values that fall outside of the cluster sets.

iv) Discretization by Intuitive Partitioning:

- Numerical ranges partitioned into relatively uniform, easy-to-read intervals that appear intuitive or "natural."
- For example, annual salaries broken into ranges like ($50,000, $60,000] are often more desirable than ranges like ($51,263.98, $60,872.34]
- The 3-4-5 rule can be used to segment numerical data into relatively uniform, naturalseeming intervals.
- In general, the rule partitions a given range of data into 3, 4, or 5 relatively equal-width intervals, recursively and level by level, based on the value range at the most significant digit.
- The rule is as follows:
- If an interval covers 3, 6, 7, or 9 distinct values at the most significant digit, then partition the range into 3 intervals
- If it covers 2, 4, or 8 distinct values at the most significant digit, then partition the range into 4 equal-width intervals.
- If it covers 1, 5, or 10 distinct values at the most significant digit, then partition the range into 5 equal-width intervals

Concept Hierarchy Generation for Nominal Data(Categorical Data):

- Categorical data are discrete data.
- Categorical attributes have a finite (but possibly large) number of distinct values, with no ordering among the values.
- Examples include geographic location, job category, and item type.
- **There are several methods for the generation of concept hierarchies for categorical data.**
- **i) Specification of a partial ordering of attributes explicitly at the schema level by users or experts:**

- Concept hierarchies for nominal attributes or dimensions typically involve a group of attributes.
- A user or expert can easily define a concept hierarchy by specifying a partial or total ordering of the attributes at the schema level.
- For example, suppose that a relational database contains the following group of attributes: street, city, province or state, and country
- A hierarchy can be defined by specifying the total ordering among these attributes at the schema level such as street < city < province or state < country.
- **ii) Specification of a portion of a hierarchy by explicit data grouping:**
- This is essentially the manual definition of a portion of a concept hierarchy
- In a large database, it is unrealistic to define an entire concept hierarchy by explicit value enumeration.
- On the contrary, we can easily specify explicit groupings for a small portion of intermediate-level data.
- For example, after specifying that province and country , form a hierarchy at the schema level, a user could define some intermediate levels manually, such as "{Alberta, Saskatchewan, Manitoba} ⊂ prairies Canada" and "{British Columbia, prairies Canada} ⊂ Western Canada."
- iii) **Specification of a set of attributes, but not of their partial ordering:**
- A user may specify a set of attributes forming a concept hierarchy, but omit to explicitly state their partial ordering.
- The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept hierarchy.
- **iv)Specification of only a partial set of attributes:**
- The user may have included only a small subset of the relevant attributes in the hierarchy specification.
- For example, instead of including all of the hierarchically relevant attributes for location, the user may have specified only street and city

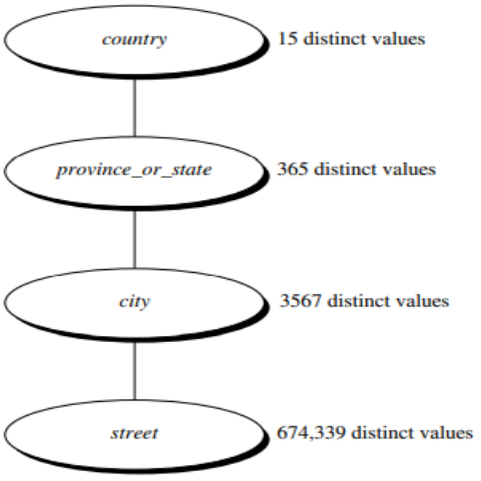


Figure 3.13 Automatic generation of a schema concept hierarchy based on the number of distinct attribute values.

5a) What is Data Binaryzation?

Ans: Binarization is the process of transforming data features of any entity into vectors of binary numbers

5b) Explain about Data Mining Task primitives.

Ans: Data Mining Task primitives:

- 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.
- The data mining primitives specify the following:

i) 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
- ii)**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.
- iii) **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

iv)**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

v)**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

5c) Explain briefly about Dimensionality Reduction.

Ans: Dimensionality Reduction

- In dimensionality reduction, data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data.
- If the original data can be reconstructed from the compressed data without any loss of information, the data reduction is called lossless.

There are two popular and effective methods of lossy dimensionality reduction: wavelet transforms and principal components analysis

i)Wavelet Transforms

- The discrete wavelet transform (DWT) is a linear signal processing technique that, when applied to a data vector X, transforms it to a numerically different vector, Xi , of wavelet coefficients.
- The two vectors are of the same length. When applying this technique to data reduction, we consider each tuple as an n-dimensional data vector, that is, X = (x1,x2,...,xn), depicting n measurements made on the tuple from n database attributes.

- A compressed approximation of the data can be retained by storing only a small fraction of the strongest of the wavelet coefficients.

- The technique also works to remove noise without smoothing out the main features of the data, making it effective for data cleaning as well
- There are several families of DWTs. Popular wavelet transforms include the Haar-2, Daubechies-4, and Daubechies-6 transforms.
- The method is as follows:
  a) The length, L, of the input data vector must be an integer power of 2. This condition can be met by padding the data vector with zeros as necessary ($L \geq n$).
  b) Each transform involves applying two functions. The first applies some data smoothing, such as a sum or weighted average. The second performs a weighted difference, which acts to bring out the detailed features of the data.
  The two functions are applied to pairs of data points in X, that is, to all pairs of measurements (x2i ,x2i+1). This results in two sets of data of length L/2. In general, these represent a smoothed or low-frequency version of the input data and the high frequency content of it, respectively.
  c)The two functions are recursively applied to the sets of data obtained in the previous loop, until the resulting data sets obtained are of length 2.

d)Selected values from the data sets obtained in the above iterations are designated the wavelet coefficients of the transformed data

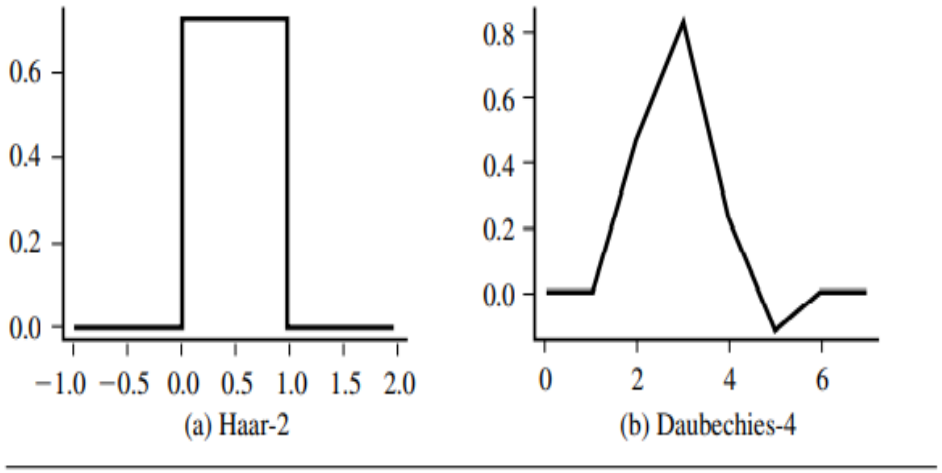


Figure 2.16 Examples of wavelet families. The number next to a wavelet name is the number of *vanishing*

ii) Principal Components Analysis

- Principal components analysis, or PCA , searches for k n-dimensional orthogonal vectors that can best be used to represent the data, where $k \leq n$.
- The original data are thus projected onto a much smaller space, resulting in dimensionality reduction.

The basic procedure for PCA is as follows:

- The input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains
- PCA computes k orthonormal vectors that provide a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to the others.

These vectors are referred to as the principal components. The input data are a linear combination of the principal components.

- The principal components are sorted in order of decreasing "significance" or strength. The principal components essentially serve as a new set of axes for the data, providing important information about variance. That is, the sorted axes are such that the first axis shows the most variance among the data, the second axis shows the next highest variance, and so on.
- Because the components are sorted according to decreasing order of "significance," the size of the data can be reduced by eliminating the weaker components, that is, those with low variance. Using the strongest principal components, it should be possible to reconstruct a good approximation of the original data.

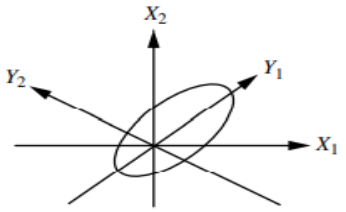


Figure 2.17 Principal components analysis. Y_1 and Y_2 are the first two principal components for the given data.

6a) What is Data Preprocessing?

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

6b) Explain about data integration.

- **Ans:** Data integration combines data from sources into coherent data store**.**

- **Data integration issues:**
  i) Entity Identification Problem:
  - How can equivalent real-world entities from multiple data sources be matched up?This is referred to as the entity identification problem.

-For example, how can the data analyst or the computer be sure that customer id in one database and cust number in another refer to the same attribute?

-Examples of metadata for each attribute include the name, meaning, data type, and range of values permitted for the attribute, and null rules for handling blank, zero or null values .

-Such metadata can be used to help avoid errors in schema integration.

ii) Redundancy and Correlation Analysis:

-An attribute may be redundant if it can be "derived" from another attribute or set of attributes.

-Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set

-Some redundancies can be detected by correlation analysis. Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data.

-For nominal data, we use the χ 2 (chi-square) test. For numeric attributes, we can use the correlation coefficient and covariance

iii) Tuple Duplication:

-In addition to detecting redundancies between attributes, duplication should also be detected at the tuple level .

-The use of denormalized tables is another source of data redundancy.

-Inconsistencies often arise between various duplicates, due to inaccurate data entry or updating some but not all data occurrences.

iv) Data Value Conflict Detection and Resolution:

-Data integration also involves the detection and resolution of data value conflicts.

-For example, for the same real-world entity, attribute values from different sources may differ.

-For a hotel chain, the price of rooms in different cities may involve not only different currencies but also different services (e.g., free breakfast) and taxes.

6c) Explain about Numerosity Reduction.

Ans:

The numerosity reduction techniques are

i)Regression and Log-Linear Models

- Regression and log-linear models can be used to approximate the given data.
- linear regression, the data are modeled to fit a straight line,with the equation y = wx+b
- In the context of data mining, x and y are numerical database attributes. The coefficients, wand b (called regression coefficients), specify the slope of the line and the Y-intercept, respectively.
- Log-linear models approximate discrete multidimensional probability distributions.

- Given a set of tuples in n dimensions (e.g., described by n attributes), we can consider each tuple as a point in an n-dimensional space.

ii)Histograms

- A histogram for an attribute, A, partitions the data distribution of A into disjoint subsets, or buckets.
- There are several partitioning rules, including the following:
- Equal-width: In an equal-width histogram, the width of each bucket range is uniform (such as the width of $10 for the bucket)
- Equal-frequency (or equidepth): In an equal-frequency histogram, the buckets are created so that, roughly, the frequency of each bucket is constant.
- V-Optimal: If we consider all of the possible histograms for a given number of buckets, the V-Optimal histogram is the one with the least variance. Histogram variance is a weighted sum of the original values that each bucket represents, where bucket weight is equal to the number of values in the bucket.
- MaxDiff: In a MaxDiff histogram, we consider the difference between each pair of adjacent values. A bucket boundary is established between each pair for pairs having the $\beta-1$ largest differences, where β is the user-specified number of buckets.

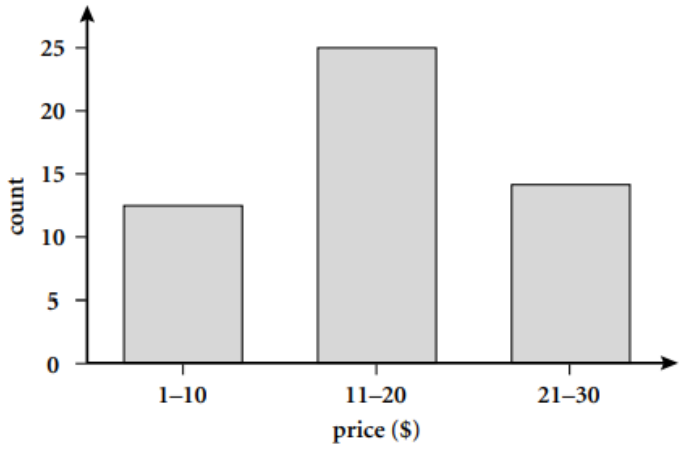


Figure 2.19 An equal-width histogram for *price*, where values are aggregated so that each bucket has a uniform width of $10.

iii)Clustering

- Clustering techniques consider data tuples as objects.
- They partition the objects into groups or clusters, so that objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters.

- Similarity is commonly defined in terms of how “close” the objects are in space, based on a distance function.
- The “quality” of a cluster may be represented by its diameter, the maximum distance between any two objects in the cluster.
- Centroid distance is an alternative measure of cluster quality and is defined as the average distance of each cluster object from the cluster centroid

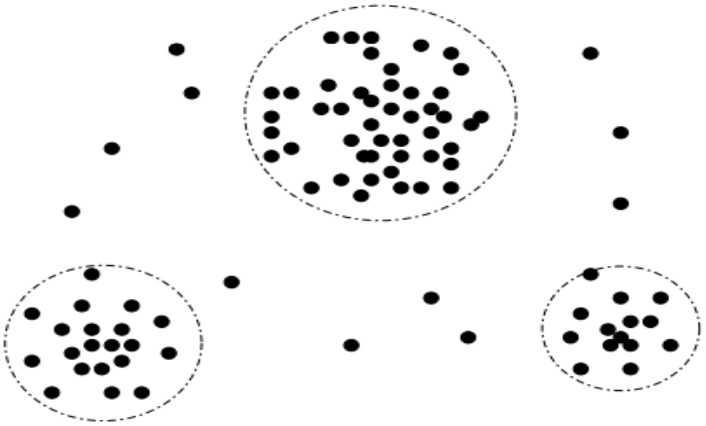
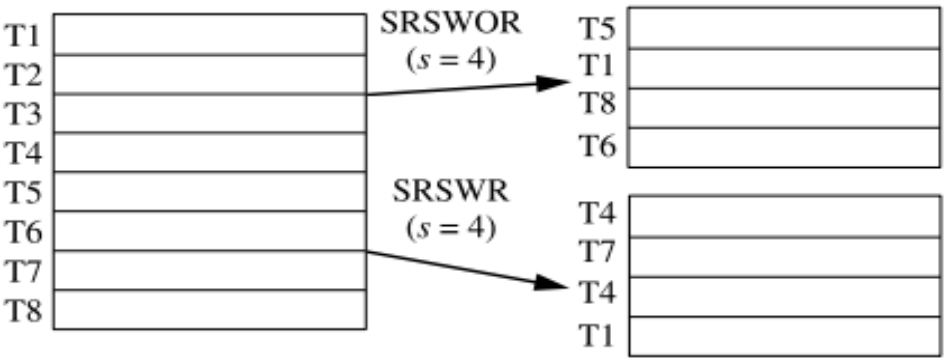


Figure 3.3 A 2-D customer data plot with respect to customer locations in a city, showing three data clusters. Outliers may be detected as values that fall outside of the cluster sets.

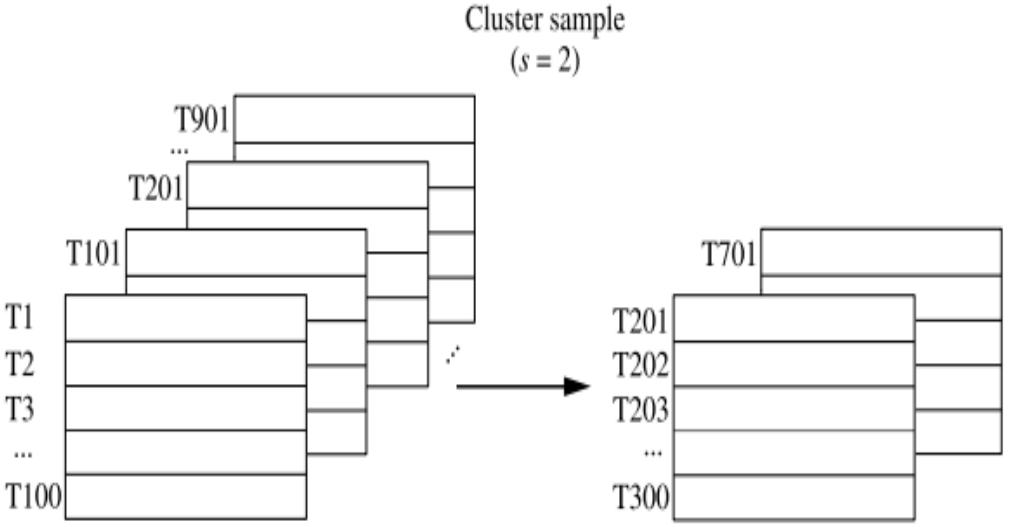
iv)Sampling

- Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random sample of the data.
- Suppose that a large data set, D, contains N tuples. Let’s look at the most common ways that we could sample D for data reduction
- **Simple random sample without replacement (SRSWOR) of size s**: This is created by drawing s of the N tuples from D (s < N), where the probability of drawing any tuple in D is 1/N, that is, all tuples are equally likely to be sampled.
- **Simple random sample with replacement (SRSWR) of size s:** This is similar to SRSWOR, except that each time a tuple is drawn from D, it is recorded and then replaced. That is, after a tuple is drawn, it is placed back in D so that it may be drawn again



-

- **Cluster sample:** If the tuples in D are grouped into M mutually disjoint "clusters," then an SRS of s clusters can be obtained, where s < M. For example, tuples in a database are usually retrieved a page at a time, so that each page can be considered a cluster. A reduced data representation can be obtained by applying, say, SRSWOR to the pages, resulting in a cluster sample of the tuples



- **Stratified sample**: If D is divided into mutually disjoint parts called strata, a stratified sample of D is generated by obtaining an SRS at each stratum.

Stratified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

Figure 2.21 Sampling can be used for data reduction.