3
UNIT

Data Mining

CONTENTS

PART-1

Overview, Motivation, Definition and Functionalities.

CONCEPT OUTLINE

- Data mining is a process used by organizations to turn raw data into useful information.
- Functionalities of data mining :

1. Characterization
2. Discrimination
3. Classification
4. Outlier analysis
5. Evolution analysis

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.1. **Explain data, information and knowledge.**

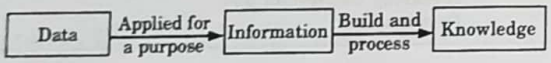
AKTU 2014-15, Marks 05

Answer

Data : Data are raw facts and figures that can be processed or stored by a computer. For example, text, numbers, symbols, etc.

Information : Information is data that has been processed into a form that gives it meaning. For example, analysis of retail of sale data can provide information on which products are selling.

Knowledge : Knowledge is the understanding of rules needed to interpret information. For example, information on retail market sales can be analyzed with promotional efforts to yield knowledge of customer behaviour.



Que 3.2. **What is data mining ? Define the major issues in data mining.**

AKTU 2014-15, Marks 05

OR

Describe challenges to data mining regarding data mining methodology and user interaction issues.

AKTU 2016-17, Marks 10

Answer

Data mining : Data mining is defined as a process used to extract usable data from a larger set of any raw data.

Key features of data mining :

1. Automatic pattern predictions based on trend and behaviour analysis.
2. Prediction based on likely outcomes.
3. Creation of decision oriented information.
4. Focus on large data sets and databases for analysis.
5. Clustering based on groups of facts not previously known.

Major issues in data mining :

1. **Mining methodology and user interaction issues :**
   a. **Mining different kinds of knowledge in databases :** Different users may be interested in different kinds of knowledge.
   b. **Interactive mining of knowledge at multiple levels of abstraction :** It allows users to focus the search for patterns from different angles.
   c. **Incorporation of background knowledge :** Background knowledge is used to guide discovery process and to express the discovered patterns.
   d. **Data mining query languages and adhoc data mining :** Data mining query language should be integrated with data warehouse query language.
   e. **Presentation and visualization of data mining results :** Once the patterns are discovered it needs to be expressed in high level languages.
   f. **Handling noisy or incomplete data :** The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities.
   g. **Pattern evaluation :** The patterns discovered should be interesting because they represent common knowledge.
2. **Performance issues :**
   a. **Efficiency and scalability of data mining algorithms :** To extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
   b. **Parallel, distributed, and incremental mining algorithms :** The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms.
3. **Diverse data types issues :**
   a. Handling of relational and complex types of data.

b. Mining information from heterogeneous databases and global information systems.

Que 3.3. Explain the data mining / knowledge extraction process in detail ? **AKTU 2017-18, Marks 10**

Answer

Knowledge Discovery in Databases (KDD) refers to the process of discovering useful knowledge from data.

Steps involved in the knowledge discovery process are :

1. **Data cleaning :**
   a. Data cleaning is defined as removal of noisy and irrelevant data from collection.
   b. It includes :
      i. Cleaning in case of missing values.
      ii. Cleaning noisy data, where noise is a random or variance error.
      iii. Cleaning with data discrepancy detection and data transformation tools.
2. **Data integration :**
   a. Data integration is defined as heterogeneous data from multiple sources combined in a common source (Data Warehouse).
   b. It includes :
      i. Data integration using data migration tools.
      ii. Data integration using data synchronization tools.
      iii. Data integration using ETL (Extract-Load-Transformation) process.
3. **Data selection :**
   a. Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
   b. It includes :
      i. Data selection using neural network.
      ii. Data selection using decision trees.
      iii. Data selection using Naive Bayes.
      iv. Data selection using clustering, regression, etc.
4. **Data transformation :**
   a. In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
   b. Data transformation is a two step process :
      i. **Data mapping :** Assigning elements from source base to destination to capture transformations.

      ii. **Code generation :** Creation of the actual transformation program.
5. **Data mining :**
   a. Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
   b. It includes :
      i. Transforms task relevant data into patterns.
      ii. Decides purpose of model using classification or characterization.
6. **Pattern evaluation :** Pattern evaluation is defined as an identifying strictly increasing patterns representing knowledge based on given measures.
7. **Knowledge representation :** Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.

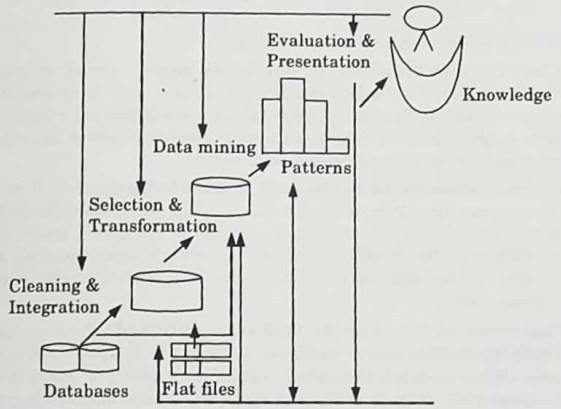


Fig. 3.3.1.

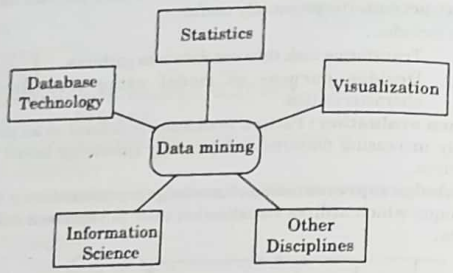
Que 3.4. How data mining systems are classified ? Describe each classification with example. **AKTU 2016-17, Marks 10**

Answer

A data mining system can be classified according to the following criteria :

1. Database technology
2. Statistics
3. Machine learning

4. Information science
5. Visualization
6. Other disciplines



Data mining system can also be classified as :

a. **Classification based on the databases mined :** Database system can be classified according to different criteria such as data models, types of data, etc. For example, if we classify a database according to the data model, then we may have a relational, transactional, object-relational, or data warehouse mining system.

b. **Classification based on the kind of knowledge mined :** It means the data mining system is classified on the basis of functionalities such as characterization, discrimination, association analysis, classification, prediction, outlier analysis, evolution analysis. A comprehensive data mining system usually provides multiple integrated data mining functionalities.

c. **Classification based on the techniques utilized :** We can classify a data mining system according to the kind of techniques used in user autonomous systems, interactive exploratory systems, query-driven systems or the methods of analysis employed such as machine learning, statistics, visualization, pattern recognition, neural networks.

d. **Classification based on the applications adapted :** We can classify a data mining system according to the applications adapted. The applications are as follows : finance, telecommunications, DNA, stock markets, E-mail.

Que 3.5. Explain data mining functionalities.

Answer

Following are the data mining functionalities :

1. **Data characterization :** It is a summarization of the general characteristics or features of a target class of data.

2. **Data discrimination :** It refers to the mapping or classification of a class with some predefined group or class.
3. **Association analysis :** It analyses the set of items that frequently appear together in a transactional dataset.
4. **Classification :** In classification, data are grouped into predefined classes.
5. **Prediction :** It refers to predict some unavailable data values rather than class labels.
6. **Cluster analysis :** Classification and prediction analyze class labeled data objects, where as clustering analyzes data objects without consulting a known class label.
7. **Outlier analysis :** Outliers are data elements that cannot be grouped in a given class or cluster.
8. **Evolution analysis :** Evolution analysis refers to the description and model regularities or trends for objects whose behaviour changes over time.

Que 3.6. Describe the difference between the following approaches for the integration of data mining system with database or data warehouse systems : no coupling, loose coupling and semi tight coupling. **AKTU 2015-16, Marks 7.5**

Answer

If a data mining system is not integrated with a database or a data warehouse system, then there will be no system to communicate with. This scheme is known as the non-coupling scheme.

Various integration schemes are as follows :

a. **No coupling :** In this scheme, the data mining system does not utilize any of the database or data warehouse functions. It fetches the data from a particular source and processes that data using some data mining algorithms.

b. **Loose coupling :** In this scheme, the data mining system may use some of the functions of database and data warehouse system. It fetches the data from the data respiratory and performs data mining on that data.

c. **Semi-tight coupling :** In this scheme, the data mining system is linked with a database or a data warehouse system and efficient implementations of a few data mining primitives can be provided in the database.

d. **Tight coupling :** In this scheme, the data mining system is smoothly integrated into the database or data warehouse system. The data mining subsystem is treated as one functional component of an information system.

PART-2

Data Processing, Form of Data Pre-Processing.

CONCEPT OUTLINE

- Data processing is the conversion of data into usable and desired form.
- Forms of data processing are :

1. Data cleaning
2. Data integration
3. Data transformation
4. Data reduction

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.7. **What are the different forms of data processing ?**

AKTU 2014-15, Marks 05

OR

Explain the data cleaning, data integration and transformation in brief.

AKTU 2014-15, Marks 05

Answer

Different forms of data processing are :

1. **Data cleaning :** Data cleaning is a process to remove the noisy data, clean the data by filling in the missing values and correct the inconsistencies in data.
2. **Data integration :** Data integration is a technique that combines the data from multiple heterogeneous data sources into a coherent data store. Data integration may involve inconsistent data and therefore needs data cleaning.
3. **Data transformation :** In this step, data is transformed or consolidated into forms appropriate for mining, by performing summary or aggregation operations. It involves the following :
   a. **Smoothing :** Smoothing is a process of removing noise from data.
   b. **Aggregation :** Aggregation is a process where summary or aggregation operations are applied to the data.

   c. **Generalization :** In generalization low-level data are replaced with high-level data by using concept hierarchies climbing.
   d. **Normalization :** Normalization scaled attribute data so as to fall within a small specified range, such as 0.0 to 1.0. It is of two types:
      i. **Min-max normalization :** It is a technique that helps to normalize data. It will scale the data between 0 and 1.
      ii. ***z*-score normalization :** Transform the data by converting the values to a common scale with an average of zero and a standard deviation of one.
   e. **Attribute/feature construction :** New attributes constructed from the given ones.
4. **Data reduction :** Data reduction is used to obtain reduced representation of data in small values by maintaining the integrity of original data.

Que 3.8. **Data consolidation is data modeling activity. This statement is true or not ? Justify.**

AKTU 2015-14, Marks 05

Answer

1. The statement is true as data consolidation means transforming data into the forms that are appropriate for mining by performing certain operations.
2. The normal data which we obtain from different data sources is not in suitable form to be stored in data warehouses or for performing data mining operations. So, data is modeled for further activities after performing data consolidation.
3. **Data consolidation involve the following operations :** Refer Q. 3.7, Page 3–8D, Unit-3.

PART-3

Data Cleaning : Missing Values, Noisy Data (Binning, Clustering Regression, Computer and Human Inspection), Inconsistent Data.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.9. **Summarize the smoothing techniques followed in data cleaning process.**

AKTU 2016-17, Marks 10

OR

How to handle noisy data ?

Answer

Noise is a random error or variance in a measured variable.

Following are the data smoothing techniques :

1. **Binning :** It is a technique in which first of all we sort the data and then partition the data into equal frequency bins. For example,

   Price = 4, 8, 15, 21, 21, 24, 25, 28, 34

   a. **Partition into (equal-frequency) bins :**

      Bin a: 4, 8, 15, Bin b: 21, 21, 24, Bin c: 25, 28, 34

   b. **Smoothing by bin means :** In smoothing by bin, each value in a bin is replaced by the mean value of the bin.

      Bin a: 9, 9, 9, Bin b: 22, 22, 22, Bin c: 29, 29, 29

   c. **Smoothing by bin boundaries :** In smoothing by bin boundaries, each bin value is replaced by the closest boundary value.

      Bin a: 4, 4, 15, Bin b: 21, 21, 24, Bin c: 25, 25, 34

2. **Regression :**
   a. Data can be smoothed by fitting the data into a regression functions. Linear regression and multiple linear regression are type of regression.
   b. A regression task begins with a dataset in which the target values are known.
   c. For example, a regression model could be used to predict the value of a house based on location, number of rooms, lot size, and other factors.
3. **Clustering :**
   a. Outliers may be detected by clustering, where similar values are organized into groups, or clusters.

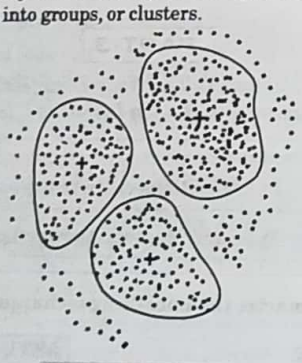


Fig. 3.9.1. Clustering.

   b. Values that fall outside of the set of clusters may be considered outliers.
   c. For example, clustering analysis can be used in area such as market research, pattern recognition, data analysis, and image processing.
4. **Combined computer and human inspection :** The outliers can also be identified with the help of computer and human inspection. The outliers patterns can be informative or garbage. Humans can sort out the garbage patterns.

Que 3.10. **Elaborate the different strategies for data cleaning.**

AKTU 2017-18, Marks 10

Answer

Data is cleaned through processors such as data migration, data scrubbing, and data auditing :

1. **Data migration :**
   a. During data migration, transformation rules are specified (for example, replacing sex by gender) to clean the data.
   b. Transcription errors, incomplete information, and lack of standard formats are also addressed during data migration.
2. **Data scrubbing :**
   a. It involves detecting and removing errors and inconsistencies from data in order to improve the quality of data.
   b. Data scrubbing involves a complex cleaning and mapping process that is the most labor intensive part of building a data warehouse.
   c. During the cleaning process, desired information is filtered out and its quality is maintained for the target system.
3. **Data auditing :**
   a. Data auditing tools make it possible to discover rules and relationships or to signal violation of stated rules by scanning data.
   b. It enhances the systems reliability and makes it possible to prevent, detect, and eliminate data errors, irregularities, and fraud.

Que 3.11. **List the ways to handle the missing values. What do you mean by inconsistent data ?**

Answer

Ways to handle missing values are :

1. **Ignore the tuple :** This is usually done when class label is missing.

2. **Fill in the missing value manually :** This approach is time consuming and may not be feasible with many missing values.
3. **Use a global constant to fill in the missing value :** Replace all the missing attribute values by the same constant.
4. Use the attribute mean to fill in the missing value.
5. **Use the most probable value to fill in the missing value:** This may be determined with regression or decision tree induction.
6. Use the attribute mean for all samples belonging to the same class as the given tuple.

Inconsistent data : Data inconsistency occur when similar data is kept in different formats in two different files, or when matching of data must be done between files. The inconsistency can be recorded in some transactions during data entry or arising from integrating data from different database.

Que 3.12. Explain Chi-square test method. Show using Chi-square test that gender and preferred reading are independent or not from given table. (Given are the observed counts).

	Male	Female	Total
Fiction	250	200	450
Non-Fiction	50	1000	1050
Total	300	1200	1500

AKTU 2015-16, Marks 15

Answer

1. A correlation relationship between two categorical (discrete) attributes, A and B, can be discovered by a χ^2 (Chi-square) test.
2. The χ^2 value (also known as the Pearson χ^2 statistics) is computed as :

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed frequency (*i.e.*, actual count) of the joint event (A_i, B_j) and e_{ij} is the expected frequency of (A_i, B_j), which can be computed as

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{N}$$

where,

N is the number of data instances

count $(A = a_i)$ is the number of instances having value a_i for A

count $(B = b_j)$ is the number of tuples having value b_j for B.

3. The larger the χ^2 value, the more likely the variables are related.
4. The cells that contribute the most of the χ^2 value are those whose actual count is very different from the expected count.

Numerical :

	Male	Female	Total
Fiction	250	200	450
Non-Fiction	50	1000	1050
Total	300	1200	1500

1. Suppose that a group of 1,500 people was surveyed. The gender of each person was noted. Each person was polled as to whether their preferred type of reading material was fiction or non-fiction. Thus, we have two attributes, gender and preferred reading.
2. The observed frequency (or count) of each possible joint event is summarized in the contingency table as shown, where the numbers in parentheses are the expected frequencies.

	Male	Female	Total
Fiction	250 (90)	200 (360)	450
Non-Fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

3. The expected frequency for the cell (male, fiction) is

$$e_{11} = \frac{\text{count(male)} \times \text{count(fiction)}}{N} = \frac{300 \times 450}{1500} = 90$$

and so on.

4. Using equation for χ^2 computation, we get

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 284.44 + 121.90 + 71.41 + 30.48 = 507.93$$

5. For this 2×2 table, the degrees of freedom are $(2-1)(2-1) = 1$. For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.001 significance level is 10.828. Since our computed value is above this, we can reject the hypothesis that gender and preferred reading are independent and conclude that the two attributes are (strongly) correlated for the given group of people.

PART-4

Data reduction : Data Cube Aggregation, Dimensionality Reduction, Data Compression, Numerosity Reduction, Discretization and Concept Hierarchy Generation and Decision Tree.

CONCEPT OUTLINE

- Data reduction is used to obtain a reduced representation of the data set.
- Strategies for data reduction include the following:
  i. Data cube aggregation
  ii. Attribute subset selection
  iii. Dimensionality reduction
  iv. Numerosity reduction
  v. Discretization and concept hierarchy generation
- A decision tree is a technique used in classification, clustering and prediction tasks.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.13. **What do you mean by data reduction ? Discuss different methods for data reduction.**

Answer

Data reduction : Data reduction is the process of minimizing the amount of data that needs to be stored in a data storage environment. Data reduction can increase storage efficiency and reduce costs.

Data reduction strategies :

1. **Data cube aggregation :** Aggregation operations are applied to the data in the construction of a data cube.
2. **Attribute subset selection :** Irrelevant or redundant characteristics or dimensions may be detected and removed.
3. **Dimensionality reduction :** In dimensionality reduction, redundant attributes are detected and removed which reduce the data set size.
4. **Data compression :** Encoding mechanisms are used to reduce the data set size.
5. **Numerosity reduction :** In numerosity reduction, the data are replaced or estimated by alternative or smaller data representations.
6. **Discretization and concept hierarchy generation :** In this, raw data values for attributes are replaced by ranges or higher conceptual levels.

Que 3.14. **Explain methods for attribute subset selection method.**

Answer

Methods for attribute subset selection are :

1. **Stepwise forward selection :** In this method, the best of the original attributes is determine and added to the reduced set.
   For example : Initial attribute set : {A1, A2, A3, A4, A5}
   Initial reduced set : { } = {A1} = {A1, A4}
   Reduced attribute set : {A2, A3, A5}
2. **Stepwise backward elimination :** It removes the worst attribute remaining in the set.
   For example : Initial attribute set : {A1, A2, A3, A4, A5}
   {A1, A3, A4, A5} = {A1, A4, A5}
   Reduced attribute set : {A1, A5}
3. **Combination of forward selection and backward elimination :** This procedure selects the best attribute and removes the worst from remaining attributes.
   For example :
   Initial attribute set : {A1, A2, A3, A4, A5}
   Reduced attribute set in stepwise forward selection : {A2, A3, A5}
   Reduced attribute set in stepwise backward elimination : {A1, A5}
   Reduced attribute set : {A1, A2, A3, A5}
4. **Decision tree induction :** It constructs a flowchart where the best attribute is chosen to partition the data into individual classes.
   For example : Initial attribute set : $\{A_1, A_2, A_3, A_4, A_5, A_6\}$

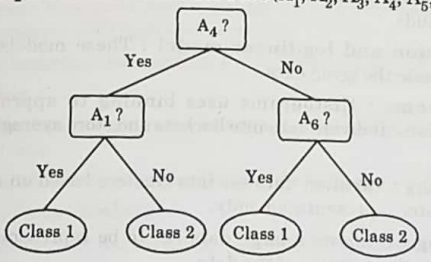


Fig. 3.14.1. Decision tree.

Reduced attribute set : $\{A_1, A_4, A_6\}$

Que 3.15. **Write a short note on dimensionality reduction.**

Answer

1. Data encoding or transformation are applied so as to obtain a reduced or compressed representation of the original data.
2. There are two components of dimensionality reduction :
   a. **Feature selection :** Feature selection is a process of removing features that are not relevant or are redundant.
   b. **Feature extraction :** Feature extraction is a process of transformation of raw data into features suitable for modeling.
3. The various methods used for dimensionality reduction include :
   a. **Wavelet transform :** It is a linear signal processing technique which transforms the data vector into numerically different vector of wavelet coefficients.
   b. **Principal Component Analysis (PCA) :** In this, the data in a higher dimensional space is mapped to data in a lower dimension space. It involves the following steps :
      i. Construct the covariance matrix of the data.
      ii. Compute the eigen vectors of this matrix.
      iii. Eigen vectors corresponding to the largest eigen values are used to reconstruct a large fraction of variance of the original data.

Que 3.16. **Discuss numerosity reduction in detail.**

Answer

In numerosity reduction, data volume can be reduced by choosing alternative forms of data representation. The various methods used for numerosity reduction include :

a. **Regression and log-linear model :** These models are used to approximate the given data.

b. **Histograms :** Histograms uses binning to approximate data distributions. It divide data into buckets and store average sum for each bucket.

c. **Clustering :** Partition data set into clusters based on similarity and store cluster representation only.

d. **Sampling :** It allows a large data set to be represented by a much smaller random sample of the data.

Que 3.17. **Distinguish between dimensionality reduction and numerosity reduction.** **AKTU 2014-15, Marks 05**

Answer

S. No.	Dimensionality reduction	Numerosity reduction
1.	In dimensionality reduction, data encoding or transformations are applied to obtain a reduced or compressed representation of original data.	In numerosity reduction, data volume is reduced by choosing alternating, smaller forms of data representation.
2.	Methods for dimensionality reduction are : a. Wavelet transforms b. Principal Component Analysis (PCA)	Methods for numerosity reduction are : a. Regression and log-linear model (parametric) b. Histograms, clustering, sampling (non-parametric).
3.	It can be used for removing irrelevant and redundant attributes.	It is merely a representation technique of original data to smaller form.
4.	In this method, some data can be lost which is irrelevant.	In this method, there is no loss of data.

Que 3.18. **Write a short note on concept hierarchy generation for numeric data.**

Answer

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

Concept hierarchy generation for numerical data methods :

1. **Binning :**
   a. Binning is a top-down splitting technique based on a specified number of bins.
   b. Binning is an unsupervised discretization technique.
2. **Histograms analysis :**
   a. Histograms partition the values for an attribute into disjoint ranges called buckets.
   b. Histograms analysis is an unsupervised discretization technique.
3. **Cluster analysis :** It is used to partition the data into clusters or groups.

Que 3.19. **Explain concept hierarchy generation for categorical data.**

AKTU 2014-15, Marks 05

Answer

1. Categorical data are discrete data.
2. Categorical attributes have finite number of distinct values, with no ordering among the values.
3. There are several methods for generation of concept hierarchies for categorical data :

   a. **Specification of a partial ordering of attributes explicitly at the schema level by experts :** Concept hierarchies for categorical attributes or dimensions typically involve a group of attributes. A user or an expert can easily define concept hierarchy by specifying a partial or total ordering of the attributes at a schema level.

   b. **Specification of a portion of a hierarchy by explicit data grouping :** In a large database, it is unrealistic to define an entire concept hierarchy by explicit value enumeration. However, it is realistic to specify explicit groupings for a small portion of the intermediate level data.

   c. **Specification of a set of attributes but not their partial ordering :** A user may specify a set of attributes forming a concept hierarchy, but omit to specify their partial ordering. The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept hierarchy.

   d. **Specification of only of partial set of attributes :** To handle partially specified hierarchies, it is important to embed data semantics in the database schema so that attributes with tight semantic connections can be pinned together.

Que 3.20. **What do you mean by data mining ? Differentiate between data mining technique and data mining strategy.**

AKTU 2013-14, Marks 05

Answer

Data mining : *Refer Q. 3.2, Page 3–2D, Unit-3.*

S. No.	Data mining technique	Data mining strategy
1.	A data mining technique applies a data mining strategy to a set of data.	A data mining strategy outlines an approach for problem solution.
2.	Data mining techniques are of following types : i. Clustering ii. Association rules iii. Statistical regression iv. Neural network v. Rule based techniques	Data mining strategies are of following types : i. Classification ii. Estimation iii. Prediction iv. Unsupervised clustering v. Market-basket analysis

☺☺☺

4 UNIT

Classification and Clustering

CONTENTS

PART-1

Classification : Definition, Data Generalization, Analytical Characterization, Analysis of Attribute Relevance.

CONCEPT OUTLINE

- Classification is a data mining function that assigns items in a collection to target categories or classes.
- Data generalization is the process of creating summary data in an evaluational database.
- Analytical characterization is the analysis of attribute relevance.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.1. **Define the terms data generalization and analytical characterization with example.** **AKTU 2013-14, Marks 05**

Answer

Data generalization :

1. Data generalization summarizes data by replacing relatively low level values with higher level concepts.
2. Data generalization approaches include : Data cube approach and attribute oriented induction approach.
3. Data generalization is a form of descriptive data mining.
4. For example, let us consider the database of XYZ electronics, instead of examining individual customer transactions, sales manager may prefer to view the generalized data to higher levels, such as summarized by customers groups according to regions, income, etc.

Analytical characterization :

1. Analytical characterization performs attribute and dimension relevance analysis in order to filter cut irrelevant or weakly attributes.
2. It is performed to overcome the various limitations of class characterization.
3. For example, employee birth_date, birth_month, birth_year are not relevant to the employee's salary but experience is highly relevant to the salary of employee.

Que 4.2. **Explain data cube approach and attribute oriented approach.** **AKTU 2014-15, Marks 05**

OR

Discuss basic approaches of data generalization.

Answer

There are two basic approaches of data generalization :

1. **Data cube approach :**
   a. It is also known as OLAP approach.
   b. In this approach, computation and results are stored in the data cube.
   c. It is an efficient approach as it is helpful to make the past selling graph.
   d. It uses roll-up and drill-down operations on a data cube.
2. **Attribute oriented induction :**
   a. It is an online data analysis, query oriented and generalization based approach.
   b. In this approach, we perform generalization on the basis of different values of each attributes within the relevant data set. After that, same tuples are merged and their respective counts are accumulated in order to perform aggregation.
   c. Attribute oriented induction approach used two methods :
      i. Attribute removal
      ii. Attribute generalization

PART-2

Mining Class Comparisons.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.3. **Why class comparisons needed in data mining ? Discuss the steps of class comparisons.**

Answer

In many applications, users may not be interested in having a single class description but they need to compare two or more classes that distinguish

target class to its contrasting classes. For example, the three classes : person, address and item are not comparable.

Steps of class comparisons are :

1. Data collection
2. Dimension relevance analysis
3. Synchronous generalization
4. Presentation of the derived comparison

Que 4.4. **What is the role of statistics in data mining ?** **AKTU 2014-15, Marks 05**

Answer

1. Statistics is a component of data mining that provides the tools and analytics techniques for dealing with large amounts of data.
2. It is the science of learning from data and includes everything from collecting and organizing to analyzing and presenting data. Statistics focuses on probabilistic models, specifically inference, using data.
3. Statistics is used in data mining for computing skills required to manage the data and its analysis and in automation of data analysis.
4. Main areas where statistical approach used in data mining are :
   i. Visualization
   ii. Size of data
   iii. Sampling
   iv. Data analysis

Que 4.5. **Explain various measures of central tendency.**

Answer

Measures of central tendency are :

1. **Mean :**
   a. It is a center of the data set.
   b. Let data set X are in values as $x_1, x_2, \ldots\ldots x_n$.

      Mean of data set X is : $X' = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. **Median :**
   a. It is the middle value of the ordered set if the number of values n is an odd number or it is the average of middle two values if n is an even number.
   b. $\text{Median} = L_1 + \left(\frac{n/2(\sum f)l}{f_{\text{medium}}}\right)C$

3. **Mode :**
   a. It is a most frequently occur value from a large data set.
   b. Mode = 3 Median – 2 Mean.
4. **Midrange :** It is the average of the largest and smallest value of data set.

PART-3

Statistical Measures in Large Databases, Statistical-Based Algorithms, Distance-Based Algorithms.

CONCEPT OUTLINE

- Two descriptive statistics are used in statistical measures :
  1. Measuring the central tendency
  2. Measuring the dispersion of data
- Distance-based algorithms are :
  1. Simple approach
  2. *k*-nearest neighbours

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.6. **Discuss various measures of dispersion of data.**

OR

What are the properties of standard deviation and give its formula ?

AKTU 2014-15, Marks 05

Answer

Measures of dispersion of data are :

1. **Range :** The range of the data set is the difference between highest and lowest value.

$$\text{Range} = H - L$$

where H is the highest and L is the lowest value in the data set.

2. **Quartiles :** The first quartile is denoted by $Q1$, is the 25th percentile. The third quartile is denoted by $Q3$, is the 75th percentile. The distance between the 1st and 3rd quartiles is the simple measure of distribution which gives the range covered by the middle half of the data. This distance is called as Interquartile Range (IQR), defined as :

$$IQR = Q3 - Q1$$

3. **Outliers :** Outliers are the values higher/lower than 1.5*IQR.

4. **Boxplot :** Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five number summary as :
   a. Typically, the ends of the box are the quartiles, so that the box length is the Interquartile Range (IQR).
   b. The median is marked by a line within the box.
   c. Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.
5. **Standard deviation and variance :** The standard deviation of a data set gives a measure of how each value in a data set varies from the mean.

The standard deviation of a set of n observations, $x_1, x_2, \ldots x_n$, is given by :

$$\sigma = \sqrt{\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{2}\left[\sum x_i^2 - \frac{1}{n}(\sum x_i)^2\right]}$$

The basic properties of the standard deviation are :

a. Σ measures spread about the mean and should be used only when the mean is chosen as the measure of center.

b. $\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise $\sigma > 0$, the variance is the mean of the squared deviations about the by σ^2. The variance of n observations, $x_1, x_2 \ldots x_n$, is given by :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]$$

Que 4.7. **Draw a box-and-whisker plot for the following data set : 126, 132, 138, 140, 141, 141, 142, 143, 144, 144, 144, 145, 146, 147, 148, 148, 149, 149, 150, 150, 150, 154, 155, 158, 158.**

Also find the outliers. **AKTU 2015-16, Marks 10**

Answer

Given : 126, 132, 138, 140, 141, 141, 142, 143, 144, 144, 144, 145, 146, 147, 148, 148, 149, 149, 150, 150, 150, 154, 155, 158, 158

Since there are 25 data points, the median Q_2 will be = 146

The first half has twelve values, so the median is the average of the middle two :

$$Q_1 = \frac{(141 + 142)}{2} = 141.5$$

The median of the second half is :

$$Q_3 = \frac{(150 + 150)}{2} = 150$$

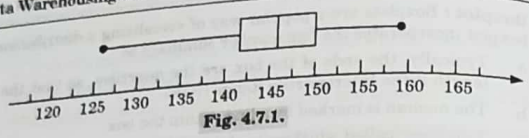


Fig. 4.7.1.

$Q_1 = 141.5$

$Q_3 = 150$

Interquartile range = 150 – 141.5 = 8.5

An outlier is any data point that is more than 1.5 times the *IQR* from either end of the box.

i.e., $8.5 \times 1.5 = 12.75$

At upper end, outlier is any data point more than

$150 + 12.75 = 162.75$

There are no data points larger than 162.75 so there are no outliers at the upper end.

At the lower end an outlier is any data point less than

$141.5 - 12.75 = 128.75$

Data point 126 is less than 128.75 therefore it is an outlier.

Que 4.8. **Given the following set of values {1, 3, 9, 15, 20), determine the Jack knife estimate for both the mean and standard deviation of the mean.** **AKTU 2013-14, Marks 05**

Answer

Given : $n = 5$,

$x_i = \{1, 3, 9, 15, 20\}$

$x_1 = 1, \quad x_2 = 3, \quad x_3 = 9, \quad x_4 = 15, \quad x_5 = 20$

Mean, $\bar{x} = \dfrac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$

$= \dfrac{1+3+9+15+20}{5} = 9.6$

By ignoring x_1, $\theta_1 = \dfrac{x_2 + x_3 + x_4 + x_5}{4}$

$= \dfrac{3+9+15+20}{4} = 11.75$

By ignoring x_2, $\theta_2 = \dfrac{x_1 + x_3 + x_4 + x_5}{4}$

$= \dfrac{1+9+15+20}{4} = 11.25$

By ignoring x_3, $\theta_3 = \dfrac{x_1 + x_2 + x_4 + x_5}{4}$

$= \dfrac{1+3+15+20}{4} = 9.75$

By ignoring x_4, $\theta_4 = \dfrac{x_1 + x_2 + x_3 + x_5}{4}$

$= \dfrac{1+3+9+20}{4} = 8.25$

By ignoring x_5, $\theta_5 = \dfrac{x_1 + x_2 + x_3 + x_4}{4}$

$= \dfrac{1+3+9+15}{4} = 7$

$\hat{\theta} = \dfrac{\theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5}{5}$

$= \dfrac{11.75+11.25+9.75+8.25+7}{5} = 9.6$

Jack knife estimate for mean is given by :

$= n(\bar{x}) - (n - 1)\hat{\theta}$

$= 5(9.6) - 4(9.6)$

$= 9.6$

Standard deviation :

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{N}}$$

$$\sigma = \sqrt{\frac{1}{5}[1-9.6)^2 + (3-9.6)^2 + (9-9.6)^2 + (15-9.6)^2 + (20-9.6)^2]}$$

$$\sigma = \sqrt{\frac{1}{5} \times 255.2} = \sqrt{51.04} = 7.144$$

By ignoring x_1 :

$$\sigma_1 = \sqrt{\frac{1}{4}[(3-9.6)^2 + (9-9.6)^2 + (15-9.6)^2 + (20-9.6)^2]}$$

$$\sigma_1 = \sqrt{\frac{1}{4} \times 181.24} = \sqrt{45.31} = 6.73$$

By ignoring x_2 :

$$\sigma_2 = \sqrt{\frac{1}{4}[(1-9.6)^2 + (9-9.6)^2 + (15-9.6)^2 + (20-9.6)^2]}$$

$$\sigma_2 = \sqrt{\frac{1}{4} \times 211.64} = \sqrt{52.91} = 7.27$$

By ignoring x_3 :

$$\sigma_3 = \sqrt{\frac{1}{4}[(1-9.6)^2 + (3-9.6)^2 + (15-9.6)^2 + (20-9.6)^2]}$$

$$\sigma_3 = \sqrt{\frac{1}{4} \times 254.84} = \sqrt{63.71} = 7.98$$

By ignoring x_4 :

$$\sigma_4 = \sqrt{\frac{1}{4}[(1-9.6)^2 + (3-9.6)^2 + (9-9.6)^2 + (20-9.6)^2]}$$

$$\sigma_4 = \sqrt{\frac{1}{4} \times 226.04} = \sqrt{56.51} = 7.51$$

By ignoring x_5 :

$$\sigma_5 = \sqrt{\frac{1}{4}[(1-9.6)^2 + (3-9.6)^2 + (9-9.6)^2 + (15-9.6)^2]}$$

$$\sigma_5 = \sqrt{\frac{1}{4} \times 147.04} = \sqrt{36.76} = 6.06$$

$$\hat{\sigma} = \frac{\sigma_1 + \sigma_2 + \sigma_3 + \sigma_4 + \sigma_5}{5}$$

$$\hat{\sigma} = \frac{6.73 + 7.27 + 7.98 + 7.51 + 6.06}{5}$$

$$\hat{\sigma} = 7.11$$

Jack knife estimate for standard deviation is given by :

$$= n(\sigma) - (n-1)\hat{\sigma}$$
$$= 5(7.144) - (5-1)(7.11)$$
$$= 35.72 - 28.44 = 7.28$$

Que 4.9. **Write short notes on :**

i. **Quartiles**
ii. **Histograms**
iii. **Scatter plots**

AKTU 2014-15, Marks 05

OR

Explain the various graphs for statistical class description.

Answer

Different types of graphs are :

1. **Histogram :** In this, we partition the data distribution of an attribute into disjoint sets but the width of each subset should be uniform. Each subset is drawn by a rectangle whose height is equal to the count of the subset.
2. **Scatter plots :** This graphical method is used for determining the existence of any relationship, pattern between two numerical attributes. In this method, every pair of value considered as a pair of coordinates in an algebraic sense and plotted as points in the plane.

3. **LOESS curve :** LOESS is locally estimated scatterplot smoothing. It adds smooth curve to existing scatterplot to provide better perception of the pattern of dependence.
4. **Quartile plots :** A quartile plot is a simple and effective way to have a first look at a univariate data distribution. First, it displays all of the data for the given attribute. Second, it plots quartile information. The mechanism used in this step is slightly different from the percentile computation.
5. **Q-Q (Quartile-Quartile) plot :** A quartile-quartile plot graphs the quartiles of one univariate distribution against the corresponding quartiles of another. It is a powerful visualization tool that allows the user to view whether there is a shift in going from one distribution to another.

Que 4.10. **Write a short note on Bayesian classification.**

AKTU 2013-14, Marks 05

Answer

1. Bayesian classifiers are the statistical classifiers.
2. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.
3. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.
4. Bayesian classification is based on Bayesian theorem.

Bayesian theorem : The purpose of Bayesian theorem is to predict the class label for a given tuple. Let X be a data tuple. In Bayesian terms, X is considered "evidence." Let H be some hypothesis, such as that the data tuple X belongs to a specified class C. There are two types of probabilities :

1. Posterior Probability $[P(H/X)]$
2. Prior Probability $[P(H)]$

where X is data tuple and H is some hypothesis. According to Bayes theorem,

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

Que 4.11. **Write a short note on Naïve Bayes classifiers.**

Answer

1. A Naive Bayes classifier uses probability theory to classify data. Naive Bayes is also known as simple Bayes or independence Bayes.
2. Naive Bayes is a kind of classifier which uses the Bayes theorem.
3. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class.

Answer

1. The genetic algorithm is derived from natural evolution. In genetic algorithm, first of all, the initial population is created.
2. This initial population consists of randomly generated rules. We can represent each rule by a string of bits.
3. For example, in a given training set, the samples are described by two boolean attributes such as A1 and A2. And this given training set contains two classes such as C1 and C2.
4. We can encode the rule IF A1 AND NOT A2 THEN C2 into a bit string 100. In this bit representation, the two leftmost bits represent the attribute A1 and A2, respectively.

Advantages of genetic algorithm :

1. Does not require any derivative information.
2. Faster and more efficient as compared to the traditional methods.
3. Has very good parallel capabilities.

Disadvantage of genetic algorithm :

1. Genetic algorithms are not suited for all problems, especially problems which are simple and for which derivative information is available.
2. Fitness value is calculated repeatedly which might be computationally expensive for some problems.

☺☺☺

5 UNIT

Data Visualization and Overall Perspective

CONTENTS

CONCEPT OUTLINE

- A neural network usually involves a large number of processors operating in parallel and arranged in tiers.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.33. What do you mean by neural network ?

Answer

1. An artificial neural network, often called a neural network, is a mathematical model based on biological neural networks.
2. A neural network consists of an interconnected group of artificial neurons, and its information system.
3. Neural networks are used to model complex relationships between inputs and outputs or to find patterns in data.
4. Neural network method is used for classification, clustering, feature mining, prediction and pattern recognition.
5. The neural network model can be broadly divided into the following three types :
   a. Feed-forward network
   b. Feedback network
   c. Self-organization networks
6. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining.

Que 4.34. Explain multilayer feed-back neural network. Differentiate between feed-forward and feedback system.

AKTU 2014-15, Marks 10

Answer

Multilayer feed-back neural network :

1. In feedback network (recurrent network) there is at least one feedback loop.
2. A recurrent network may consist of a single layer of neurons with each neuron feeding its output signal back to the inputs of all the other neuron.

3. Feedback loops, has a profound impact on learning capability of the network and on its performance.

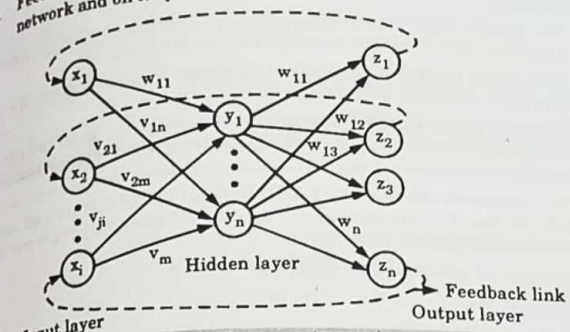


Fig. 4.34.1. A recurrent neural network.

Difference between feed-forward and feedback system :

S.No.	Feed-forward system	Feedback system
1.	Feed-forward system allows signals to travel one way only from input to output.	Feedback system allows signal travelling in both directions by introducing loops in the network.
2.	There is no feedback.	There is a feedback.
3.	Feed-forward tends to be straight forward network that associate input with output.	Computations derived from earlier input are feedback into the network.
4.	Feed-forward are static *i.e.*, in feed-forward, state does not changes.	Feedback are dynamic *i.e.*, in feedback, state changes continuously until equilibrium is reached.

Que 4.35. Write a short note on genetic algorithm.

AKTU 2013-14, Marks 05

OR

Describe the role of genetic algorithm in data mining.

AKTU 2014-15, Marks 05

C_{k+1} = Apriori-gen(L_k)
until $C_{k+1} = \phi$;

Que 4.31. **Find frequent patterns under the association rules by using Apriori algorithm for the following transactional database :**

TID	T100	T200	T300	T400	T500
Items	M,O,N,K,E,Y	D,O,N,K,E,Y	M,A,K,E	M,U,C,K,Y	C,O,O,K,I,E

Let minimum support = 60 % and minimum confidence = 80 %

AKTU 2017-18, Marks 10

Answer

Using Apriori, we successively generate the sets C_k of candidate k-itemsets, and then verify these for minsup, obtaining the sets L_k of frequent k-itemsets. In the database D, this leads to :

C_1 =

Item	Count
A	A
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	3
U	1
Y	3

L_1 =

Item	Count
E	4
K	5
M	3
O	3
Y	3

C_2 =

Item	Count
EK	4
EM	2
EO	3
EY	2
KM	3
KO	3
KY	3
MO	1
MY	2
OY	2

L_2 =

Item	Count
EK	4
EO	3
KM	3
KO	3
KY	3

$C_3 = L_3$ =

Item	Count
EKO	3

Association rules :

$\{KO\} \Rightarrow \{E\}$
$\{EO\} \Rightarrow \{K\}$

PART-9

Basic Algorithms, Parallel and Distributed Algorithms.

CONCEPT OUTLINE

- Data parallelism and task parallelism are the parallel and distributed algorithms.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.32. **Explain parallel and distributed algorithms.**

Answer

Parallel and distributed algorithms : Parallel or distributed and algorithms strive to parallelize either the data, known as data parallelism, or the candidates, referred to as task parallelism.

Data parallelism :

1. One data parallelism algorithm is the Count Distribution Algorithm (CDA).
2. The database is divided into p partitions, one for each processor.
3. Each processor counts the candidates for its data and then broadcasts its counts to all other processors.
4. Each processor then determines the global counts.
5. These counts are used to determine the large itemsets and to generate the candidates for the next scan.

Task parallelism :

1. The Data Distribution Algorithm (DDA) demonstrates task parallelism.
2. Here the candidates as well as the database are partitioned amount the processors.
3. Each processor in parallel counts the candidates given to it using its local database partition.
4. Then each processor broadcasts its database partition to all other processors.
5. Each processor then uses this to obtain a global count for its data and broadcasts this count to all other processors.
6. Each processor then can determine globally large itemsets and generate the next candidates.
7. These candidates then are divided among processors for the next scan.

PART-10

Neural Network Approach.

PART-8

Association Rules : Introduction, Large Item Sets.

CONCEPT OUTLINE

- Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.29. **What do you mean by association rule mining ?**

Answer

1. Association rule mining is a procedure which is meant to find frequent patterns, correlations, or associations from data sets found in various kinds of databases.
2. The first step in association analysis is the enumeration of itemsets. An itemset is any combination of two or more items in a transaction.
3. The main applications of association rule mining are :
   a. Basket data analysis
   b. Cross marketing
   c. Catalog design
4. Association rule generation is usually split up into two separate steps:
   a. First, minimum support is applied to find all frequent itemsets in a database.
   b. Second, these frequent itemsets and the minimum confidence constraint are used to form rules.
5. The selection of association rules is based on two values called support and confidence. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

Que 4.30. **What is association rule mining ? Explain the Apriori algorithm to find the frequent itemsets.**

AKTU 2014-15, Marks 05

Answer

Association rule mining : Refer Q. 4.29, Page 4–31D, Unit-4.

Apriori algorithm :

a. The Apriori algorithm is used in association rule mining which uses the property of large itemset *i.e.*, any subset of a large itemset must be large.

b. Apriori uses bottom-up approach, where frequent subsets are extended one item at a time.

c. The entire algorithm can be divided into two steps :

Step 1 : Apply minimum support to find all the frequent sets with k items in a database.

Step 2 : Use the self-join rule to find the frequent sets with $k + 1$ items with the help of frequent k-itemsets. Repeat this process from $k = 1$ to the point when we are unable to apply the self-join rule.

Algorithm :

Input :

I // Itemsets
D // Database of transactions
s // Support

Output :

L // Large itemsets

Apriori algorithm :

$k = 0$; // k is used as the scan number.
$L = \phi$;
$C_1 = I$; // Initial candidates are set to be the items.
repeat
$k = k + 1$;
$L_k = \phi$;
for each $I_i \in C_k$ do
$c_i = 0$; // Initial counts for each itemset are 0.
for each $t_j \in D$ do
for each $I_j \in C_k$ do
if $I_i \in t_j$ then
$c_i = c_i + 1$;
for each $I_i \in C_k$ do
if $c_i \geq (s \times | D |)$ do
$L_k = L_k \cup I_i$;
$L = L \cup L_k$;

DBSCAN algorithm :

$k = 0$; // Initially there are no clusters.
for $i = 1$ to n do
if t_i is not in a cluster, then
$X = \{t_j \mid t_j$ is density-reachable from $t_i\}$;
if X is a valid cluster, then
$k = k + 1$;
$K_k = X$;

The expected time complexity of DBSCAN is $O(n \log n)$.

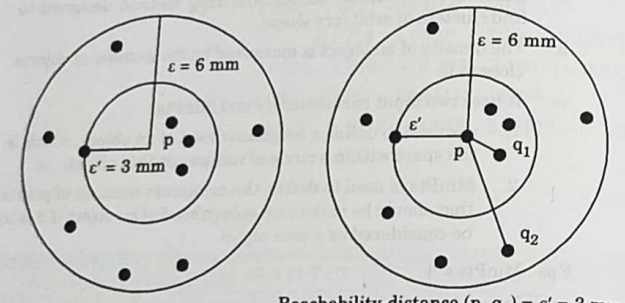
b. OPTICS :

i. OPTICS is a variation of DBSCAN which was designed to surmount issues occurs in DBSCAN.

ii. OPTICS does not explicitly produce a data set clustering.

iii. It instead gives us cluster ordering such that objects which are in a denser cluster are closer in a list.

iv. OPTICS stores two additional attributes *i.e.*, Core-distance and reachability distances, which are used to derive the ordering such that clusters with higher density will be finished first.



Core-distance of p

Reachability-distance (p, q_1) = ε' = 3 mm
Reachability-distance (p, q_2) = d(p, q_2)

Fig. 4.27.1.

Que 4.28. Discuss in detail various grid-based methods.

OR

Explain STING in detail. **AKTU 2017-18, Marks 05**

Answer

1. In this, the objects together form a grid.
2. The object space is quantized into finite number of cells that form a grid structure.
3. Grid-based methods are of two types :

a. STING (Statistical Information Grid) :

i. STING works with numerical attributes.

ii. Information's such as mean, maximum and minimum are pre-computed and stored in rectangular cells.

iii. Parameters at the higher level cells are drawn from the parameters of the bottom level cells.

iv. For each cell, there are attribute independent parameters and attribute dependent parameters.

Algorithm :

Input :

T // Tree
q // Query

Output :

R // Regions of relevant cells

STING algorithm :

$i = 1$
repeat
for each node in level i do
determine if this cell is relevant to q and mark as such;
$i = i + 1$
until all layers in the tree have been visited;
identify neighbouring cells of relevant cells to create regions of cells;

b. CLIQUE :

i. CLIQUE (Clustering in QUEst) is a bottom-up subspace clustering algorithm that constructs static grids.

ii. It uses apriori approach to reduce the search space.

iii. CLIQUE is a density and grid based and find out the clusters by taking density threshold and number of grids as input parameters.

iv. Steps in CLIQUE are :

1. The dimension space is partitioned into no overlapping units called cells.
2. Identify the dense and sparse cells.
3. Use the dense cells to assemble the clusters.
4. Starting with an arbitrary dense cell, we find the maximal region of all connected dense cells in all dimensions.
5. Repeat step 4 until all cells are covered.

$$\left.\begin{aligned} d(M_1, X_1) &= (1.5^2 + 1.42^2)^{1/2} = 2.066 \\ d(M_2, X_1) &= (2.875^2 + 1^2)^{1/2} = 3.044 \end{aligned}\right\} X_1 \in C_2$$

$$\left.\begin{aligned} d(M_1, X_2) &= (1.5^2 + 0.58^2)^{1/2} = 1.608 \\ d(M_2, X_2) &= (2.875^2 + 1^2)^{1/2} = 3.044 \end{aligned}\right\} X_2 \in C_1$$

$$\left.\begin{aligned} d(M_1, X_3) &= (1.5^2 + 0.83^2)^{1/2} = 1.714 \\ d(M_2, X_3) &= (2.875^2 + 1.25^2)^{1/2} = 3.135 \end{aligned}\right\} X_3 \in C_1$$

$$\left.\begin{aligned} d(M_1, X_4) &= (1.5^2 + 0.83^2)^{1/2} = 1.714 \\ d(M_2, X_4) &= (2.875^2 + 1.25^2)^{1/2} = 3.135 \end{aligned}\right\} X_4 \in C_1$$

$$\left.\begin{aligned} d(M_1, X_5) &= (1.5^2 + 1.67^2)^{1/2} = 2.245 \\ d(M_2, X_5) &= (2.875^2 + 1.25^2)^{1/2} = 3.135 \end{aligned}\right\} X_5 \in C_2$$

Above calculation is based on Euclidean distance formula,

$$d(X_i, X_j) = \sum_{k=1}^{m} (X_{ik} - X_{jk})^{1/2}$$

New clusters $C_1 = \{X_2, X_3, X_4\}$ and $C_2 = \{X_1, X_5\}$ have new centroids.

$$M_1 = (1.679, 3.105)$$

$$M_2 = (2.156, 3.089)$$

The corresponding within-cluster variations

$$e_1^2 = [(1.608 - 1.679)^2 + (3.044 - 3.105)^2] + [(1.714 - 1.679)^2 + (3.135 - 3.105)^2] + [(1.714 - 1.679)^2 + (3.135 - 3.105)^2] = 0.0129$$

$$e_2^2 = [(2.066 - 2.156)^2 + (3.044 - 3.089)^2] + [(2.245 - 2.156)^2 + (3.135 - 3.089)^2] = 0.0201$$

So, the total square error is

$$E^2 = e_1^2 + e_2^2 = 0.033$$

PART-7

Density-Based Methods : DBSCAN, OPTICS, Grid-Based Methods: STING, CLIQUE Model-Based Method : Statistical Approach.

CONCEPT OUTLINE

- Density-based methods are of two types :
  1. DBSCAN
  2. OPTICS
- Grid-based methods are of two types :
  1. STING
  2. CLIQUE

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.27. Explain density-based methods.

OR

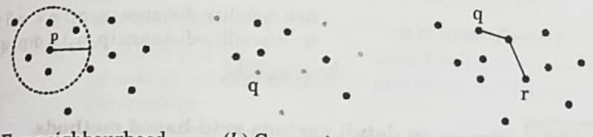
Explain the density-based clustering method based on connected regions with sufficiently high density (DBSCAN).

AKTU 2014-15, Marks 10

Answer

1. Density-based method is based on the notion of density.
2. The basic idea is to continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold, *i.e.*, for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.
3. It is of two types :
   - **a. DBSCAN :**
     - i. Density-based spatial clustering of applications with noise.
     - ii. DBSCAN is a density-based clustering method designed to find clusters of arbitrary shape.
     - ii. The density of an object is measured by the number of objects close to it.
     - iv. It uses two input parameters; ε and MinPts.
       1. ε is used to define ε neighbourhood of an object, which is the space within a circle of radius ε at that object.
       2. MinPts is used to define the minimum number of points that should be in the ε neighbourhood of an object if it is to be considered as a core object.

Eps MinPts = 4



(*a*) Eps-neighbourhood (*b*) Core points (*c*) Density reachable

Fig. 4.27.1. DBSCAN example.

Algorithm :

Input :

$D = \{t_1, t_2, ..., t_n\}$ // Set of elements

MinPts // Number of points in cluster

Eps // Maximum distance for density measure

Output :

$K = \{K_1, K_2, ..., K_k\}$ // Set of clusters

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

a. seed1 = A1 = (2, 10), seed2 = A4 = (5, 8), seed3 = A7 = (1, 2)

epoch1 – start :

A1 :

d(A1, seed1) = 0 as A1 is seed1

d(A1, seed2) = $\sqrt{13} > 0$

d(A1, seed3) = $\sqrt{65} > 0$

→ A1 ∈ cluster 1

A3 :

d(A3, seed1) = $\sqrt{36}$ = 6

d(A3, seed2) = $\sqrt{25}$ = 5 ← smaller

d(A3, seed3) = $\sqrt{53}$ = 7.28

→ A3 ∈ cluster 2

B2 :

d(B2, seed1) = $\sqrt{50}$ = 7.07

d(B2, seed2) = $\sqrt{13}$ = 3.60 ← smaller

d(B2, seed3) = $\sqrt{45}$ = 6.70

→ B2 ∈ cluster 2

C1 :

d(C1, seed1) = $\sqrt{65}$ >0

d(C1, seed2) = $\sqrt{52}$ >0

d(C1, seed3) = 0 as A7 is seed3

→ C1 ∈ cluster 3

A2 :

d(A2, seed1) = $\sqrt{25}$ = 5

d(A2, seed2) = $\sqrt{18}$ = 4.24

d(A2, seed3) = $\sqrt{10}$ = 3.16 ← smaller

→ A2 ∈ cluster 3

B1 :

d(B1, seed1) = $\sqrt{13}$

d(B1, seed2) = 0 as B1 is seed2

d(B1, seed3) = $\sqrt{52}$ > 0

→ B1 ∈ cluster 2

B3 :

d(B3, seed1) = $\sqrt{52}$ = 7.21

d(B3, seed2) = $\sqrt{17}$ = 4.12 ← smaller

d(B3, seed3) = $\sqrt{29}$ = 5.380

→ B3 ∈ cluster 2

C2 :

d(C2, seed1) = $\sqrt{5}$

d(C2, seed2) = $\sqrt{2}$ ← smaller

d(C2, seed3) = $\sqrt{58}$

→ C2 ∈ cluster 2

end of epoch1

new clusters : 1 : {A1}, {A3, A4, A5, A6, A8} 3: {A2, A7}

b. Centers of the new clusters :

C1 = (2, 10), C2 = ((8 + 5 + 7 + 6 + 4)/5 = (6, 6), C3 = ((2 + 1)/2, (5 + 2)/2) = (1.5, 3.5)

Que 4.26. Consider five points $\{X_1, X_2, X_3, X_4, X_5\}$ with the following coordinates as a two dimensional sample for clustering : X_1 = (0, 2.25); X_2 = (0,0.25); X_3 = (1.25,0); X_4 = (4.5,0); X_5 = (4.5, 2.5); Illustrate the k-means partitioning algorithm (clustering algorithm) using the above data set. **AKTU 2016-17, Marks 15**

Answer

Given : $X_1 = (0, 2.25)$, $X_2 = (0, 0.25)$, $X_3 = (1.25, 0)$, $X_4 = (4.5, 0)$, $X_5 = (4.5, 2.5)$

Since, coordinates is a two-dimensional sample for clustering.

Initially, clusters are formed from random distribution of samples :

$$C_1 = \{X_1, X_2, X_4\} \text{ and } C_2 = \{X_3, X_5\}$$

Since, $$\text{centroid } (M_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ik}$$

So, the centroid for these two clusters are :

$$M_1 = \{(0 + 0 + 4.5)/3, (2.25 + 0.25 + 0)/3\} = \{1.5, 0.83\}$$

$$M_2 = \{(1.25 + 4.5)/2, (0 + 2.5)/2\} = \{2.875, 1.25\}$$

The error within-cluster variation

$$e_k^2 = \sum_{i=1}^{n_k} (X_{ik} - M_k)^2$$

Within-cluster variation after initial random distribution of sample, are

$$e_1^2 = [(0 - 1.5)^2 + (2.25 - 0.83)^2] + [(0 - 1.5)^2 + (0.25 - 0.83)^2] + [(4.5 - 1.5)^2 + (0 - 0.83)^2]$$

$$= 16.5417$$

$$e_2^2 = [(1.25 - 2.875)^2 + (0 - 1.25)^2] + [(4.5 - 2.875)^2 + (2.5 - 1.25)^2]$$

$$= 8.40625$$

The total square error is

$$E^2 = e_1^2 + e_2^2 = 16.5417 + 8.40625 = 24.94795$$

Depending on a minimum distance from centroids μ_1 and μ_2, the new redistribution of samples inside clusters will be

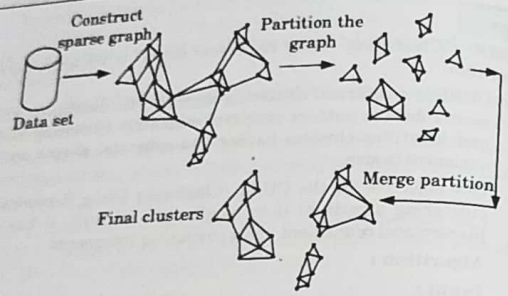


Fig. 4.23.1. Frameworks of Chameleon.

Que 4.24. **What do you mean by partitional clustering ? Also, explain *k*-mean clustering algorithm.**

Answer

Partitional clustering :

1. Non-hierarchical or partitional clustering creates the clusters in one step as opposed to several steps.
2. Only one set of clusters is created, although several different sets of clusters may be created internally within the various algorithms.
3. Since only one set of clusters is output, the user must input the desired, *k* of clusters.
4. One common measure is a squared distance from each point to the centroid for the associated cluster :

$$\sum_{m=1}^{k} \sum_{t_{mi} \in Km} dis\,(C_m, t_{mi})^2$$

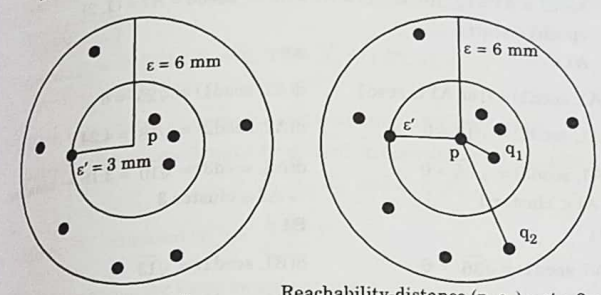
5. A problem with partitional algorithm is that they suffer from a combinatorial explosion due to the number of possible solutions.

***k*-mean clustering :**

1. In this algorithm, the data is divided into a pre-specified *k* clusters.
2. The result obtained is such that there is at least one item in each cluster all of which does not overlap and are non-hierarchical.
3. OPTICS computes an augmented cluster ordering for automatic and interactive cluster analysis.
4. This ordering represents the density-based clustering structure obtained from a wide range of parameter settings.
5. The cluster ordering can be used to extract basic clustering information as well as provide the intrinsic clustering structure.

Two values need to be stored for each object, core-distance and reachability-distance.

1. **Core-distance :** The core-distance of an object p is the smallest ε value that makes $\{p\}$ a core object. If p is not a core object, the core-distance of p is undefined.
2. **Reachability-distance :** The reachability-distance of an object q with respect to another object p is the greater value of the core-distance of p and the Euclidean distance between p and q. If p is not a core object, the reachability-distance between p and q is undefined.



Core-distance of p

Reachability-distance (p, q_1) = ε' = 3 mm
Reachability-distance (p, q_2) = d(p, q_2)

Fig. 4.24.1.

Que 4.25. **Write the *k*-mean algorithm. Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are : A1 (2,10), A2 (2, 5), A3 (8, 4) B1 (5, 8), B2 (7, 5), B3 (6, 4), C1 (1, 2), C2 (4, 9)**

The distance function is Euclidian distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the *k*-means algorithm to show only the three cluster centers after the first round of execution. **AKTU 2017-18, Marks 10**

Answer

***k*-mean algorithm :** Refer Q. 4.24, Page 4–23D, Unit-4.

Numerical : The Euclidean distances between the given points are in the following matrix :

in which all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) :

1. It is a scalable clustering method designed for very large data sets. In this, only one scan of data is necessary.
2. It is based on the notation of CF (Clustering Feature) tree. CF tree is a height balanced tree that stores the clustering features for a hierarchical clustering.
3. Cluster of data points is represented by a triple of numbers (N, LS, SS) where N = Number of items in the sub cluster, LS = Linear sum of the points, SS = Sum of the squared of the points.

BIRCH clustering algorithm :

a. **Phase 1 :** Build the CF Tree. Load the data into memory by building a cluster-feature tree. Optionally, condense this initial CF tree into a smaller CF.

b. **Phase 2 :** Global Clustering. Apply an existing clustering algorithm on the leaves of the CF tree. Optionally, refine these clusters.

4. BIRCH is sometimes referred to as two-step clustering, because of the two phases.

Que 4.22. Write a short note on hierarchical and non-hierarchical clustering. **AKTU 2013-14, Marks 05**

Answer

1. Non-hierarchical or partitional clustering is faster than hierarchical clustering.
2. Hierarchical clustering requires only a similarity measure, while partitional clustering requires stronger assumptions such as number of clusters and the initial centers.
3. Hierarchical clustering does not require any input parameters, while partitional clustering algorithms require the number of clusters to start running.
4. Hierarchical clustering returns a much more meaningful and subjective division of clusters but partitional clustering results in exactly k clusters.
5. Hierarchical clustering algorithms are more suitable for categorical data as long as a similarity measure can be defined accordingly.

Que 4.23. Write a short note on CURE and Chameleon algorithm.

Answer

1. **CURE :** (Clustering Using Representatives) is an agglomerative algorithm.

a. CURE is an efficient clustering algorithm for databases, which is more robust to outliers compared with other clustering methods, and identifies clusters having non-spherical shapes and wide variances in size.

b. One objective for the CURE (Clustering Using Representative) clustering algorithm is to handle outliers well. It has both a hierarchical component and a portioning component.

c. **Algorithm :**

Input :

$D = \{t_1, t_2, \ldots, t_n\}$ // Set of elements

k //Desired number of clusters

Output :

Q //Heap containing one entry for each cluster

CURE algorithm :

T = build(D);

Q = heapify(D); // Initially build heap with one entry per item;

repeat
u = min(Q) ;
delete (Q, u. close) ;
w = merge(u, v) ;
delete(T, u);
delete(T, v);
insert(T, w);
for each $x \in Q$ do
x . close = find closest cluster to x;
if x is closest to w, then
w. close = x;
insert (Q, w);
until number of nodes in Q is k;

2. **Chameleon :**

a. Chameleon is clustering using dynamic modeling.

b. Chameleon is a hierarchical clustering that uses dynamic modeling to determine the similarity between pairs of clusters.

c. The key feature of Chameleon's agglomerative hierarchical clustering algorithm is that it determines the pair of similar sub-clusters by taking into accounts both the inter-connectivity as well as the closeness (proximity) of the clusters.

Answer

1. Clustering is the process of making a group of abstract objects into classes of similar objects.
2. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
3. Clustering methods can be classified into the following categories :
   i. Partitioning method
   ii. Hierarchical method
   iii. Density-based method
   iv. Grid-based method
   v. Model-based method

Que 4.20. **Explain the different data types used in cluster analysis.**

AKTU 2014-15, Marks 05

OR

Describe the types of data that often occur in cluster analysis and briefly explain how to preprocess that data for clustering.

AKTU 2015-16, Marks 10

Answer

Data types used in cluster analysis are :

1. **Interval scaled variables :** Interval scaled variables are continuous measurements of roughly linear scale. Typical examples include weight and height, latitude and longitude coordinates (for example, when clustering houses), and weather temperature.
2. **Binary variables :** A binary variables has only two states : 0 and 1, where 0 means that the variable is absent, and 1 means that it is present. Given the variable smoker describing a patient, for instance, 1 indicates that the patient smokes, while 0 indicates that the patient does not.
3. **Categorical variables :** A categorical variable is a generation of the binary variable in that it can take on more than two states. For example, map colour is a categorical variable that may have, say, five states : red, yellow, green, pink and blue.
4. **Ordinal variables :** Ordinal variables are very useful for registering subjective assessments of qualities that cannot be measured objectively. For example, professional ranks are often enumerated in a sequential order, such as assistant, associate end full for professors.
5. **Ratio scaled variables :** A ratio scaled variable makes a positive measurement on a non-linear scale, such as an exponential scale, approximately following the formula :

Ae^{Bt} or Ae^{-Bt}

where A and B are positive constant, and t typically represents time. Common examples include the growth of a bacteria population or the decay of a radioactive element.

6. **Variables of mixed type :** Data sets may contain all types of variables such as : symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio.

To preprocess data for clustering : Refer Q. 3.7, Page 3–8D, Unit-3.

PART-6

Hierarchical and Partitional Algorithms, Hierarchical Clustering : CURE and Chameleon.

CONCEPT OUTLINE

- Types of hierarchical clustering :
  1. CURE
  2. Chameleon

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.21. **What is hierarchical method for clustering ? Explain BIRCH method.**

AKTU 2015-16, Marks 10

AKTU 2017-18, Marks 10

Answer

Hierarchical method for clustering :

1. Hierarchical clustering creates hierarchy of clusters on the data set.
2. This hierarchical tree shows levels of clustering with each level having a larger number of smaller clusters. CURE and Chameleon are the examples of hierarchical clustering.
3. **There are two approaches of hierarchical algorithm :**
   a. **Agglomerative clustering :** Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters. Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
   b. **Divisive clustering :** It starts with a single cluster containing all objects, and then successively splits resulting clusters until only clusters of individual objects remain. This is a "top-down" approach

$$\text{Gini}(S) = 1 - \sum_i p_i^2$$

Que 4.17. **Compute the decision rules by deriving a decision tree classifier and information gain as selection measure for the given database in table. 4.13.1.**
Given : Gain (age) = 0.246, Gain (student) = 0.151 and Gain (credit rating) = 0.048

AKTU 2017-18, Marks 10

Answer

Given : Gain (age) = 0.246, Gain (student) = 0.151 and Gain (credit rating) = 0.048

Age has the highest gain, therefore it is used as the decision attribute in the root node. Since, age has three possible values, the root node has three branches :

Gain (youth, student) = 0.970
Gain (youth, income) = 0.570
Gain (youth, credit_rating) = 0.019

The above calculations show that the attribute student shows the highest gain. Therefore, it should be used as the next decision node for the branch youth. This process is repeated until all data are classified perfectly or no attribute is left for the child nodes.

The corresponding rules are :

1. If age = youth and student = no then buys_computer = no
2. If age = youth and student = yes then buys_computer = yes
3. If age = middle-age then buys_computer = yes
4. If age = senior and credit_rating = excellent then buys_computer = yes
5. If age = senior and credit_rating = fair then buys_computer = no

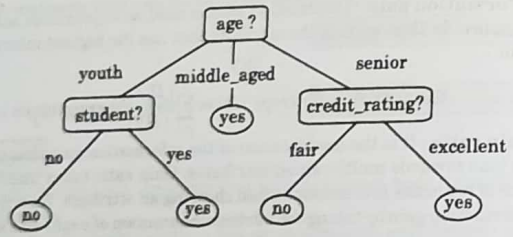


Fig. 4.17.1.

Que 4.18. **Discuss issues that are important to consider when employing a decision tree – based classification algorithm. Explain the decision tree induction algorithm with appropriate examples. Discuss the disadvantages of this approach ? What is over fitting, and how can it be prevented for decision trees ?**

AKTU 2016-17, Marks 10

Answer

Issues to be consider when employing a decision tree :

1. Choosing splitting attributes
2. Ordering of splitting attributes
3. Splits
4. Tree structure
5. Stopping criteria
6. Training data
7. Pruning

Decision tree induction algorithm with example : Refer Q. 4.16, Page 4–15D and Q. 4.17, Page 4–17D; Unit-4.

Disadvantages of decision tree :

1. They do not easily handle continuous data.
2. Handling missing data is difficult.
3. Difficult to use when we have smooth boundaries.

Overfitting : Since the decision tree is constructed from the training data, overfitting may occur. This can be overcome via tree pruning.

PART-5

Clustering : Introduction, Similarity and Distance Measures.

CONCEPT OUTLINE

- Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities.
- Clustering is of two types :
  1. Hierarchical clustering 2. Partitional clustering

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.19. **What do you mean by clustering ? What are the different types of clustering ?**

CONCEPT OUTLINE

Decision tree-based algorithms are :

1. ID3
2. C4.5
3. CART
4. Scalable DT Techniques

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.15. How are decision trees useful in data mining ? Explain.

AKTU 2012-14, Marks 10

Answer

1. Decision trees are used for processing a large amount of data and thus find use in data mining.
2. The decision tree approach is most useful in classification problems. With this technique, a tree is constructed to model the classification process.
3. Decision tree can handle high dimensional data and easily understand by humans.
4. The learning and classification steps of decision tree induction are simple and fast.
5. Decision tree induction algorithm has been useful in many application areas like medicine, manufacturing and production, financial analysis, astronomy, etc.

Que 4.16. Write the algorithm of decision tree induction. What are the methods that can be used for selecting the splitting criteria ?

AKTU 2015-16, Marks 10

OR

Write a short note on gain ratio.

AKTU 2017-18, Marks 2.5

Answer

1. A decision tree is a structure that includes a root node, branches, and leaf nodes.
2. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.
3. Once the tree is built, it is applied to each tuple in the database and results in a classification for tuple. There are two basic steps in the technique: building the tree and applying the tree to the database.

Decision tree induction algorithm :

1. Create a node N;
2. If tuples in D are all of the same class, C then
3. Return N as a leaf node labeled with the class C;
4. If attribute-list is empty then
5. Return N as a leaf node labeled with the majority class in D; // majority voting
6. Apply Attribute_selection _method D, attribute_list, to find the "best" splitting-criterion;
7. Label node N with splitting criterion;
8. If splitting-attribute is discrete-valued and multiway splits allowed then // not restricted to binary trees
9. Attribute_list ← attribute list – splitting-attribute; // remove splitting_attribute
10. For each outcome j of splitting-criterion
    // partition the tuples and grow subtrees for each partition
11. Let D_j be the set of data tuples in D satisfying outcome j; II a partition
12. If D_j is empty then
13. Attach a leaf labeled with the majority class in D to node N;
14. Else attach the node returned by generate decision_tree (D_j, attribute _list) to node N;

    endfor

15. Return N;

Methods used for selecting the splitting-criterion:

1. **Information gain :** Information gain is used as an attribute selection measure. In this, we pick the attribute that has the highest information gain.

$$\text{Gain}(D, A) = \text{Entropy}(D) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{Entropy}(D_j)$$

2. **Gain ratio :** It is the modification of the information gain that reduces its bias towards multi-valued attributes. Gain ratio takes number and size of branches into account when choosing an attribute. It corrects the information gain by taking the intrinsic information of a split into account.

$$GR(S, A) = \frac{\text{Gain}(S, A)}{\text{IntI}(S, A)}$$

3. **Gini index :** The gini index is used in CART. It considers a binary split for each attribute. When considering a binary split, we compute a weighted sum of the impurity of each resulting partition.

Answer

1. Laplacian correction is a technique used for avoiding zero probability values.
2. When training set is large enough that adding one to each count will make negligible difference in estimated probability (avoiding zero probability value) we use Laplacian correction.
3. If we have q counts to which we each add one, then we must remember to add q to the corresponding denominator used in the probability calculation.

Numerical :

X = (age = senior, income = medium, student = no, credit rating = fair)

We need to maximize $P(X|C_i)P(C_i)$, for $i = 1, 2$. $P(C_i)$, the priori probability of each class, can be estimated based on the training tuples :

P(buys_computer = yes) = 9/14 = 0.643

P(buys_computer = no) = 5/14 = 0.357

To compute $P(X|C_i)$, for i = 1, 2, we compute the following conditional probabilities:

P(age=senior | buys_computer=yes)= 3/9= 0.333

P(age=senior | buys_computer=no)=2/5=0.400

P(income=medium | buys_computer=yes)= 4/9 = 0.444

P(income=medium | buys_computer=no)= 2/5 =0.400

P(student=no | buys_computer=yes)=3/9=0.333

P(student=no | buys_computer=no)=4/5=0.800

P(credit_rating=fair | buys_computer=yes)=6/9=0.667

P(credit_rating=fair | buys_computer=no)=2/5=0.400

Using the above probabilities :

$P(X|$buys_computer=yes$)$ = P(age=senior | buys_computer=yes) × P(income=medium | buys_computer=yes) × P(student=no | buys_computer=yes) × P(credit_rating=fair | buys_computer=yes) = 0.033

$P(X|$buys_computer=no$)$= P(age=senior | buys_computer=no) ×P(income=medium | buys_computer=no)× P(student=no | buys_computer=no) ×P(credit_rating=fair | buys_computer=no) = 0.051

Compute $P(X|C_i)P(C_i)$ for each class :

$P(X|$buys_computer=yes$)$ × P(buys_computer=yes)=0.033 ×0.643=0.021

$P(X|$buys_computer=no$)$ × P(buys_computer=no)=0.051 ×0.357=0.018

The Bayesian Classifier predicts buys_computer=yes for tuple X.

Que 4.14. Explain distance-based algorithms in detail.

Answer

1. Distance-based algorithms are non-parametric methods that can be used for classification.
2. These algorithms classify objects by the dissimilarity between them as measured by distance functions.
3. There are two types of distance-based algorithm :
   a. **Simple approach :** It assumes that each class is represented by its center or centroid. The new item is placed in the class with the largest similarity value.
   b. ***k*-nearest neighbour :** The KNN scheme requires not only training set but also the desired classification for each item. When a classification is to be made for a new item, its distance to each item in the training set must be determined. Only the k closest entries in the training set are considered. The new item is then placed in the class that contains the most items for this set of k closest items.

Algorithm :

Input :

T // Training data

K // Number of neighbours

t // Input tuple to classify

Output :

c // class to which t is assigned

KNN algorithm :

// Algorithm to classify tuple using KNN

$N = \phi$;

// Find set of neighbours, N, for t

for each $d \in T$ do

if $|N| \leq K$, then

$N = N \cup \{d\}$;

else

if $\exists x \in N$ such that

sim $(t, u) \leq$ sim (t, d), then

begin

$N = N - \{u\}$;

$N = N \cup \{d\}$;

end

// Find class for classification

c = class to which the most $u \in N$ are classified.

PART-4

Decision Tree-Based Algorithms.

4. The class with the highest probability is considered as the most likely class. This is also known as Maximum A Posteriori (MAP).
5. Naive Bayes classifier assumes that all the features are unrelated to each other.

For example : A fruit may be considered to be an apple if it is red, round, and about 5 in diameter. Even if these features depend on each other or upon the existence of the other features, a Naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Que 4.12. **Classify the tuple X = (Colour = 'RED', Type = 'SUV' Origin = 'DOMESTIC') using Naive Bayesian classification. Training data is given in the following table where class label is (STOLEN).**

Colour	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Yellow	SUV	Imported	No
Red	Sports	Domestic	Yes

AKTU 2015-16, Marks 15

Answer

Colour			Type			Origin		
	Yes	No		Yes	No		Yes	No
Red	4	2	Sports	4	2	Domestic	3	3
Yellow	1	3	SUV	1	3	Imported	2	2

Stolen	
Yes	No
5	5

Colour			Type			Origin		
	Yes	No		Yes	No		Yes	No
Red	4/5	2/5	Sports	4/5	2/5	Domestic	3/5	3/5
Yellow	1/5	3/5	SUV	1/5	3/5	Imported	2/5	2/5

Stolen	
Yes	No
1/2	1/2

Likelihood of yes = $\frac{4}{5} \times \frac{1}{5} \times \frac{3}{5} \times \frac{1}{2} = \frac{6}{125} = 0.048$

Likelihood of no = $\frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{1}{2} = \frac{9}{125} = 0.072$

Therefore the prediction is no.

Que 4.13. **What is Laplacian correction in Bayesian classifier ? Compute the class of the four following tuple by using Bayesian classification for given database in table. X = (age = senior, credit rating = fair, income= medium, student= no).**

Table 4.13.1.

Age	Income	Student	Credit rating	Class : buys computer
youth	high	No	Fair	No
youth	high	No	Excellent	No
middle aged	high	No	Fair	Yes
senior	medium	No	Fair	Yes
senior	low	Yes	Fair	Yes
senior	low	Yes	Excellent	No
middle aged	low	Yes	Excellent	Yes
youth	medium	No	Fair	No
youth	low	Yes	Fair	Yes
senior	medium	Yes	Fair	Yes
youth	medium	Yes	Excellent	Yes
middle aged	medium	No	Excellent	Yes
middle aged	high	Yes	Fair	Yes
senior	medium	No	Excellent	No

AKTU 2017-18, Marks 10

$$\sigma_3 = \sqrt{\frac{1}{4} \times 254.84} = \sqrt{63.71} = 7.98$$

By ignoring x_4 :

$$\sigma_4 = \sqrt{\frac{1}{4}[(1-9.6)^2 + (3-9.6)^2 + (9-9.6)^2 + (20-9.6)^2]}$$

$$\sigma_4 = \sqrt{\frac{1}{4} \times 226.04} = \sqrt{56.51} = 7.51$$

By ignoring x_5 :

$$\sigma_5 = \sqrt{\frac{1}{4}[(1-9.6)^2 + (3-9.6)^2 + (9-9.6)^2 + (15-9.6)^2]}$$

$$\sigma_5 = \sqrt{\frac{1}{4} \times 147.04} = \sqrt{36.76} = 6.06$$

$$\hat{\sigma} = \frac{\sigma_1 + \sigma_2 + \sigma_3 + \sigma_4 + \sigma_5}{5}$$

$$\hat{\sigma} = \frac{6.73 + 7.27 + 7.98 + 7.51 + 6.06}{5}$$

$$\hat{\sigma} = 7.11$$

Jack knife estimate for standard deviation is given by :

$$= n(\sigma) - (n-1)\hat{\sigma}$$
$$= 5(7.144) - (5-1)(7.11)$$
$$= 35.72 - 28.44 = 7.28$$

Que 4.9. **Write short notes on :**

i. **Quartiles**
ii. **Histograms**
iii. **Scatter plots**

AKTU 2014-15, Marks 05

OR

Explain the various graphs for statistical class description.

Answer

Different types of graphs are :

1. **Histogram :** In this, we partition the data distribution of an attribute into disjoint sets but the width of each subset should be uniform. Each subset is drawn by a rectangle whose height is equal to the count of the subset.
2. **Scatter plots :** This graphical method is used for determining the existence of any relationship, pattern between two numerical attributes. In this method, every pair of value considered as a pair of coordinates in an algebraic sense and plotted as points in the plane.

3. **LOESS curve :** LOESS is locally estimated scatterplot smoothing. It adds smooth curve to existing scatterplot to provide better perception of the pattern of dependence.
4. **Quartile plots :** A quartile plot is a simple and effective way to have a first look at a univariate data distribution. First, it displays all of the data for the given attribute. Second, it plots quartile information. The mechanism used in this step is slightly different from the percentile computation.
5. **Q-Q (Quartile-Quartile) plot :** A quartile-quartile plot graphs the quartiles of one univariate distribution against the corresponding quartiles of another. It is a powerful visualization tool that allows the user to view whether there is a shift in going from one distribution to another.

Que 4.10. **Write a short note on Bayesian classification.**

AKTU 2013-14, Marks 05

Answer

1. Bayesian classifiers are the statistical classifiers.
2. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.
3. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.
4. Bayesian classification is based on Bayesian theorem.

Bayesian theorem : The purpose of Bayesian theorem is to predict the class label for a given tuple. Let X be a data tuple. In Bayesian terms, X is considered "evidence." Let H be some hypothesis, such as that the data tuple X belongs to a specified class C. There are two types of probabilities :

1. Posterior Probability $[P(H/X)]$
2. Prior Probability $[P(H)]$

where X is data tuple and H is some hypothesis. According to Bayes theorem,

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

Que 4.11. **Write a short note on Naïve Bayes classifiers.**

Answer

1. A Naive Bayes classifier uses probability theory to classify data. Naive Bayes is also known as simple Bayes or independence Bayes.
2. Naive Bayes is a kind of classifier which uses the Bayes theorem.
3. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class.