Q2 (a) What is the difference between discrimination and classification?

Answer

Discrimination differs from **classification** in that the former refers to a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes, while the latter is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Discrimination and classification are similar in that they both deal with the analysis of class data objects.

Q2 (b) Describe three challenges to data mining regarding data mining methodology and user interaction issues.

Answer

Challenges to data mining regarding data mining methodology and user interaction issues include the following: mining different kinds of knowledge in databases, interactive mining of knowledge at multiple levels of abstraction, incorporation of background knowledge, data mining query languages and ad hoc data mining, presentation and visualization of data mining results, handling noisy or incomplete data, and pattern evaluation. Below are the descriptions of the first three challenges mentioned:

Mining different kinds of knowledge in databases: Different users are interested in different kinds of knowledge and will require a wide range of data analysis and knowledge discovery tasks such as data characterization, discrimination, association, classification, clustering, trend and deviation analysis, and similarity analysis. Each of these tasks will use the same database in different ways and will require different data mining techniques.

Interactive mining of knowledge at multiple levels of abstraction: Interactive mining, with the use of OLAP operations on a data cube, allows users to focus the search for patterns, providing and refining data mining requests based on returned results. The user can then interactively view the data and discover patterns at multiple granularities and from different angles.

Incorporation of background knowledge: Background knowledge, or information regarding the domain under study such as integrity constraints and deduction rules, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction. This helps to focus and speed up a data mining process or judge the interestingness of discovered patterns.

Q2 (c) Describe why ‘concept hierarchies’ are useful in data mining.

Answer

Concept hierarchies define a sequence of mappings from a set of lower-level concepts to higher-level, more general concepts and can be represented as a set of nodes organized in a tree, in the form of a lattice, or as a partial order. They are useful in data mining because they allow the discovery of knowledge at multiple levels of abstraction and provide the structure on which data can be generalized (rolled-up) or specialized (drilled-down). Together, these operations allow users to view the data from different perspectives, gaining further insight into relationships hidden in the data. Generalizing has the advantage of compressing the data set, and mining on a compressed data set will require fewer I/O operations. This will be more efficient than mining on a large, uncompressed data set.

Q3 (a) Explain the following concepts:
(i) Data transformation
(ii) Data Reduction

Answer

(1) In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following:

- Smoothing, which works to remove noise from the data. Such techniques include binning, regression, and clustering.
- Aggregation, where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to computer monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.
- Generalization of the data, where low-level or “primitive” (raw) data are replaced by higher-level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higher-level concepts, like city or country. Similarly, values for numerical attributes, like age, may be mapped to higher-level concepts, like youth, middle-aged, and senior.
- Normalization, where the attribute data are scaled so as to fall within a small specified range, such as – 1.0 to 1.0, or 0.0 to 1.0.

Attribute construction (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.

(2) Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

Strategies for data reduction include the following:

1. **Data cube aggregation**, where aggregation operations are applied to the data in the construction of a data cube.

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

3. **Dimensionality reduction**, where encoding mechanisms are used to reduce the data set size.

4. **Numerosity reduction**, where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering sampling, and the use of histograms.

Discretization and concept hierarchy generation, where raw data values for attributes are replaced by ranges or higher conceptual levels.

Q3 (b) In real world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

Answer
The various methods for handling the problem of missing values in data tuples include:

(a) Ignoring the tuple: This is usually done when the class label is missing (assuming the mining task involves classification or description). This method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

(b) Manually filling in the missing value: In general, this approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined.

(c) Using 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 $-\infty$. If missing values are replaced by, say, “Unknown,” then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common — that of “Unknown.” Hence, although this method is simple, it is not recommended.

(d) Using a measure of central tendency for the attribute, such as the mean (for symmetric numeric data), the median (for asymmetric numeric data), or the mode (for
nominal data): For example, suppose that the average income of All Electronics customers is $28,000 and that the data are symmetric. Use this value to replace any missing values for income.

(e) Using the attribute mean for numeric (quantitative) values or attribute mode for nominal values, 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. If the data are numeric and skewed, use the median value.

(f) Using the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

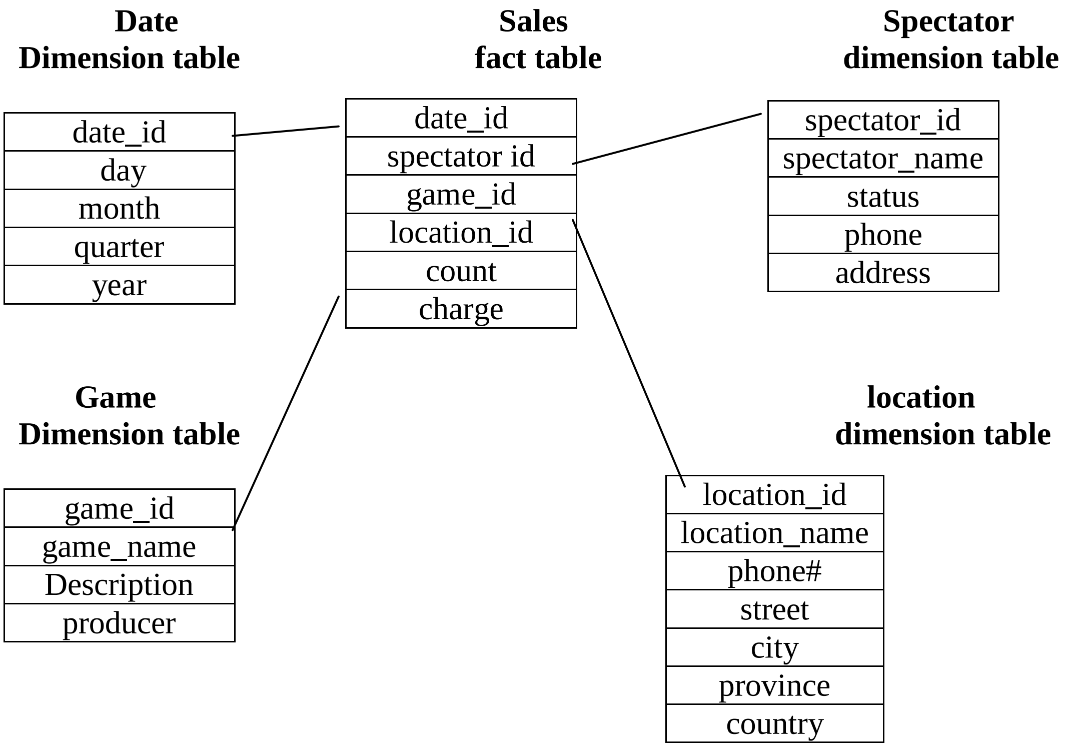
Q4 (a) Suppose that a data warehouse consists of the four dimensions: date, spectator, location, and game, and the two measures: count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

(i) Draw a star schema diagram for the data warehouse

(ii) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM Place in 2010?

Answer

(1) Draw a star schema diagram for the data warehouse.
A star schema is shown here



Date Dimension table
date_id
day
month
quarter
year

Sales fact table
date_id
spectator id
game_id
location_id
count
charge

Spectator dimension table
spectator_id
spectator_name
status
phone
address

Game Dimension table
game_id
game_name
Description
producer

location dimension table
location_id
location_name
phone#
street
city
province
country

(2) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of
CS courses for each Big-University student.
The specific OLAP operations to be performed are:
• Roll-up on course from course id to department.
• Roll-up on semester from semester id to all.
• Slice for course="CS".

Q4 (b) In data warehouse technology, a multiple dimensional view can be implemented by a relational database technique (ROLAP), or by a multidimensional database technique (MOLAP), or by a hybrid database technique (HOLAP).
(i) Briefly describe each implementation technique.
(ii) For each technique, explain how each of the following functions may be implemented:

- **The generation of a data warehouse (including aggregation)**
- **Roll-up**
- **Drill-down**

Incremental updating

Answer
(a) Briefly describe each implementation technique.
A ROLAP technique for implementing a multiple dimensional view consists of intermediate servers that stand in between a relational back-end server and client front-end tools, thereby using a relational or extended-relational DBMS to store and manage warehouse data, and OLAP middleware to support missing pieces. A MOLAP implementation technique consists of servers, which support multidimensional views of data through array-based multidimensional storage engines that map multidimensional views directly to data cube array structures. A HOLAP implementation approach combines ROLAP and MOLAP technology, which means that large volumes of detailed data and some very low level aggregations can be stored in a relational database, while some high level aggregations are kept in a separate MOLAP store.

(b) For each technique, explain how each of the following functions may be implemented:
i. The generation of a data warehouse (including aggregation)
ROLAP: Using a ROLAP server, the generation of a data warehouse can be implemented by a relational or extended-relational DBMS using summary fact tables. The fact tables can store aggregated data and the data at the abstraction levels indicated by the join keys in the schema for the given data cube.
MOLAP: In generating a data warehouse, the MOLAP technique uses multidimensional array structures to store data and multiway array aggregation to compute the data cubes.
HOLAP: The HOLAP technique typically uses a relational database to store the data and some low level aggregations, and then uses a MOLAP to store higher-level aggregations.

ii. Roll-up ROLAP: To roll-up on a dimension using the summary fact table, we look for the record in the table that contains a generalization on the desired dimension. For example, to roll-up the date dimension from day to month, select the record for which the day field contains the special value all. The value of the measure field, dollars sold, for example, given in this record will contain the subtotal for the desired roll-up.

MOLAP: To perform a roll-up in a data cube, simply climb up the concept hierarchy for the desired dimension. For example, one could roll-up on the location dimension from city to country, which is more general.

HOLAP: The roll-up using the HOLAP technique will be similar to either ROLAP or MOLAP,

Depending on the techniques used in the implementation of the corresponding dimensions.

iii. Drill-down

ROLAP: To drill-down on a dimension using the summary fact table, we look for the record in the table that contains a generalization on the desired dimension. For example, to drill-down on the location dimension from country to province or state, select the record for which only the next lowest field in the concept hierarchy for location contains the special value all. In this case, the city field should contain the value all. The value of the measure field, dollars sold, for example, given in this record will contain the subtotal for the desired drill-down.

MOLAP: To perform a drill-down in a data cube, simply step down the concept hierarchy for the desired dimension. For example, one could drill-down on the date dimension from month

to day in order to group the data by day rather than by month.

HOLAP: The drill-down using the HOLAP technique is similar either to ROLAP or MOLAP

depending on the techniques used in the implementation of the corresponding dimensions.

iv. Incremental updating

To perform incremental updating, check whether the corresponding tuple is in the summary fact table. If not, insert it into the summary table and propagate the result up. Otherwise, update the value and propagate the result up.

MOLAP: To perform incremental updating, check whether the corresponding cell is in the

MOLAP cuboid. If not, insert it into the cuboid and propagate the result up. Otherwise, update the value and propagate the result up.

HOLAP: similar either to ROLAP or MOLAP depending on the techniques used in the implementation of the corresponding dimensions.

(c) Which implementation techniques do you prefer, and why?

HOLAP is often preferred since it integrates the strength of both ROLAP and MOLAP methods

and avoids their shortcomings—if the cube is quite dense, MOLAP is often preferred. Also, if the data are sparse and the dimensionality is high, there will be too many cells (due to exponential growth) and, in this case, it is often desirable to compute iceberg cubes instead of materializing the complete cubes.

Q5 (a) Explain Multiway Array Aggregation for full cube computation.

Answer

The Multiway Array Aggregation (or simply MultiWay) method computes a full data cube by using a multidimensional array as its basic data structure. It is a typical MOLAP approach that uses direct array addressing, where dimension values are accessed via the position or index of their corresponding array locations. Hence, MultiWay cannot perform any value-based reordering as an optimization technique. A different approach is developed for the array-based cube construction, as follows:

1. Partition the array into chunks. A **chunk** is a subcube that is small enough to fit into the memory available for cube computation. **Chunking** is a method for dividing an *n*-dimensional array into small *n*-dimensional chunks, where each chunk is stored as an object on disk. The chunks are compressed so as to remove wasted space resulting from empty array cells (i.e., cells that do not contain any valid data, whose cell count is zero). For instance, “chunked + offset” can be used as a cell addressing mechanism to compress a sparse array structure and when searching for cells within a chunk. Such a compression technique is powerful enough to handle sparse cubes, both on disk and in memory.

2. Compute aggregates by visiting (i.e., accessing the values at) cube cells. The order in which cells are visited can be optimized so as to minimize the number of times that each cell must be revisited, thereby reducing memory access and storage costs. The trick is to exploit this ordering so that partial aggregates can be computed simultaneously, and any unnecessary revisitng of cells is avoided.

    Because this chunking technique involves “overlapping” some of the aggregation computations, it is referred to as **multiway array aggregation**. It performs **simultaneous aggregation** –that is, it computes aggregations simultaneously on multiple dimensions.

Q5 (b) Discuss how to support quality drill down although some low level cells may contain empty or too less data for reliable analysis.

Answer

The best way to solve this small sample size problem is to simply get more data. There are two possible methods to expand the query and get more data to boost the confidence. They both expand the original query in the data cube, just in different directions.

The first one is Intra-Cuboid Query Expansion. In the intra-cuboid case, the expansion occurs by

looking at nearby cells in the same cuboid as the queried cell. Dimensions which are uncorrelated or weakly correlated with the measure value (i.e., the value to be predicted) are the best candidates for expansion. Next we should select semantically similar values within those dimension(s) to minimize the risk of altering the final result.

The second is Inter-Cuboid Query Expansion. Here the expansion occurs by looking to a more general cell. And the strategy is similar: correlated dimensions are not allowed in inter-cuboid expansions.

Q6 (a) Give a short example to show that items in a strong association rule may actually be negatively correlated.

Answer

Consider the following table:

	A	A	∑row
B	65	35	100
B	40	10	50
∑col	105	35	150

Let the minimum support be 40%. Let the minimum confidence be 60%. A ⇒B is a strong rule because it satisfies minimum support and minimum confidence with a support of 65/150 = 43.3% and a confidence of 65/100 = 61.9%. However, the correlation between A and B is corrA;B = 0:433 0:700×0:667 = 0.928, which is less than 1, meaning that the occurrence of A is negatively correlated with the occurrence of B.

Q6 (b) Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?

Answer

The decision tree built may overfit the training data. There could be too many branches, some of which may reflect anomalies in the training data due to noise or outliers. Tree pruning addresses this issue of overfitting the data by removing the least reliable branches (using statistical measures). This generally results in a more compact and reliable decision tree that is faster and more accurate in its classification of data.

The drawback of using a separate set of tuples to evaluate pruning is that it may not be representative of the training tuples used to create the original decision tree. If the separate set of tuples are skewed, then using them to evaluate the pruned tree would not be a good indicator of the pruned tree's classification accuracy. Furthermore, using a separate set of tuples to evaluate pruning means there are less tuples to use for creation and testing of the tree. While this is considered a drawback in machine learning, it may not be so in data mining due to the availability of larger data sets.

Q7 (a) Why is naive Bayesian classification called "naive"? Briefly outline the major ideas of naive Bayesian classification.

Answer

Naive Bayesian classification is called naive because it assumes class conditional independence.

That is, the effect of an attribute value on a given class is independent of the values of the other attributes.

This assumption is made to reduce computational costs, and hence is considered “na¨ıve”. The major idea behind na¨ıve Bayesian classification is to try and classify data by maximizing P(X|Ci)P(Ci) (where i is an index of the class) using the Bayes’ theorem of posterior probability. In general:

• We are given a set of unknown data tuples, where each tuple is represented by an n-dimensional vector, X = (x1, x2. . . xn) depicting n measurements made on the tuple from n attributes, respectively A1,A2, ..,An. We are also given a set of m classes, C1,C2, . . .Cm.

• Using Bayes theorem, the na¨ıve Bayesian classifier calculates the posterior probability of each class conditioned on X. X is assigned the class label of the class with the maximum posterior probability conditioned on X. Therefore, we try to maximize P(Ci|X) = P(X|Ci)P(Ci)/P(X).

However, since P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1) = P(C2) = · · · = P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). The class prior probabilities may be estimated by P(Ci) = si
s , where si is the number of training tuples of class Ci, and s is the total number of training tuples.

• In order to reduce computation in evaluating P(X|Ci), the na¨ıve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple, i.e., that there are no dependence relationships among the attributes.

– If Ak is a categorical attribute then P(xk|Ci) is equal to the number of training tuples in Ci that have xk as the value for that attribute, divided by the total number of training tuples in Ci.

– If Ak is a continuous attribute then P(xk|Ci) can be calculated using a Gaussian density function.

Q7 (b) What are ensemble methods? Explain in detail with the help of algorithm.

Answer Page Number 366- 367 of Textbook

Q8 (a) Briefly describe and give examples of each of the following approaches to clustering: partitioning methods, hierarchical methods, density-based methods, grid-based methods and model based methods.

Answer

Clustering is the process of grouping data into classes, or clusters, so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

There are several approaches to clustering methods.

Partitioning methods: Given a databases of n objects or data tuples, a partitioning methods constructs k partitions of data, where each partition represents a cluster and k ≤ n. Given k, the number of partitions to construct, it creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The general criterion of a good partitioning is that objects in the same cluster are "close" or related to each other, whereas objects of different clusters are "far apart". For example, the k-means method is one of the partitioning methods commonly used.

Hierarchical methods: A hierarchical method creates a hierarchical decomposition of the given set of data objects. It can be either agglomerative or divisive. The agglomerative (bottom-up) approach starts with each object forming a separate group. It successively merges the objects close to one another, until all of the groups are merged into one, or until a termination condition holds. The divisive (top down) approach starts with all objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster or until a termination condition holds. BIRCH is an example of integration of hierarchical method with distance-based clustering.

Density-based methods: This method is based on density such as density-connected points. The
main idea is to continue growing a given cluster as long as the density in its "neighborhood" exceeds some threshold. That is, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. This method can be used to filter out noise and discover clusters of arbitrary shape. DBSCAN is a typical example of density-based clustering method.

Grid-based methods: Such methods quantize the object space into a finite number of cells that
form a grid structure. All of the clustering operations are performed on the grid structure. The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space.

STING is an example of a grid-based method.

Model-based methods: This approach hypothesizes a model for each of the clusters and finds the best fit of the data to the given model. It locates the clusters by constructing a density function that reflects the spatial distribution of the data points. It also leads to a way of automatically determining the number of clusters based on standard statistics, taking "noise" or outliers into account and thus yielding robust clustering methods. COBWEB is an example of a statistical approach of a model-based method.

Q8 (b) Describe any two of the following clustering algorithms in terms of the following criteria:

(i) shapes of clusters that can be determined;
(ii) input parameters that must be specified; and
(iii) limitations

- **k-means**
- **k-medoids**
- **CLARA**
  BIRCH

Answer

(a) k-means
1. Compact clouds (clusters of non-convex shape cannot be determined);
2. Number of clusters;
3. Sensitive to noise and outliers, works good on small data sets only.

(b) k-medoids
1. Compact clouds (clusters of non-convex shape cannot be determined);
2. Number of clusters;
3. Small data sets (not scalable).

(c) CLARA
1. Convex-shaped clusters;
2. Number of clusters;
3. Sensitive to the selection of initial samples.

(d) BIRCH
1. Spherical in shape clusters;
2. N d-dimensional data points;
3. Resulting clusters may be of unnatural shape.

Q9 (b) What are the differences between visual data mining and data visualization?

Answer

Data visualization is to present data in a visual way so that the data contents can be easily comprehensible by human users. Pattern or knowledge visualization is to present patterns and knowledge in a visual way so that one can easily understand the regularities, patterns, outliers, and other forms of knowledge stored in or discovered from large data sets. Visual data mining is the process of data mining that presents data and patterns in a visually appealing and interactive way so that a user can easily participate in the data mining process.

Q9 (c) Describe a situation in which you feel that data mining can infringe on your privacy.

Answer

Another example involves one's Supermarket Club Card. Having access to your card, Supermarket has the potential, without your knowledge, to study your shopping patterns based on your transactions made with the Club Card. What kind of drugs you purchased during specific periods of time is personal information that you may not wish to be used or sold.

Text Book

Data Mining, Concepts and Techniques, Jiawei Han and Micheline Kamber, Elsevier, Second Edition.