Q.2 a. Where is data mining is used? (4)

Answer:

Applications:

1) Military uses
2) Medical field
3) Business Intelligence
4) Intelligence agencies (security purpose in communication and other fields)
5) Data Retrieval (in web documents etc)

b. Describe the steps in data mining. (4)

Answer:

Steps:

1) Develop an understanding of the purpose of the data mining projects.
2) Obtain the data set to be used in the analysis
3) Explore, clean and pre-process the data
4) Reduce the data if necessary.
5) Determine the data mining class (classification, clustering, prediction etc)
6) Choose Data mining technique to be used (regression, neural nets, hierarchical clustering)
7) Use algorithms to perform the task
8) Interpret the result of algorithms.
9) Deploy the model.

c. Explain the different criteria based on which the data mining systems can be categorized. (8)

Answer: Page 29-30

Q.3 a. Explain the following concepts:
(i) Methods for Data transformation
(ii) Various Strategies for Data Reduction (2×4)

Answer:

(i). Data Transformation:

In metadata and data warehouse, a **data transformation** converts a set of data values from the data format of a source data system into the data format of a destination data system.

Data transformation can be divided into two steps:

1. data mapping maps data elements from the source data system to the destination data system and captures any transformation that must occur
2. code generation that creates the actual transformation program

Data element to data element mapping is frequently complicated by complex transformations that require one-to-many and many-to-one transformation rules.

The code generation step takes the data element mapping specification and creates an executable program that can be run on a computer system. Code generation can also create transformation in easy-to-maintain computer languages such as Java or XSLT.

A master data recast is another form of data transformation where the entire database of data values is transformed or recast without extracting the data from the database. All data in a well designed database is directly or indirectly related to a limited set of master database tables by a network of foreign key constraints. Each foreign key constraint is dependent upon a unique database index from the parent database table. Therefore, when the proper master database table is recast with a different unique index, the directly and indirectly related data is also recast or restated. The directly and indirectly related data may also still be viewed in the original form since the original unique index still exists with the master data. Also, the database recast must be done in such a way as to not impact the applications architecture software.

When the data mapping is indirect via a mediating data model, the process is also called **data mediation**.

(ii) 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. Dimension reduction, where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.

3. Data compression, where encoding mechanisms are used to reduce the data set size. The methods used for data compression are wavelet transform and Principal Component Analysis.

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 e.g. regression and log-linear models), or nonparametric methods such as clustering, sampling, and the use of histograms.

5. **Discretization and concept hierarchy generation**, where raw data values for attributes are replaced by ranges or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of abstraction, and are a powerful tool for data mining

b. Explain the following data discretization technique
(i) Entropy-Based Discretization
(ii) Histogram Analysis **(2×4)**

Answer:

(i) Page 89 (ii) Page 88

Q.4 a. Explain how the data warehouse acts as a basis for EIS. **(8)**

Answer:

It is in the EIS environment that the data warehouse operates in its most effective state. The data warehouse is tailor-made for the needs of the EIS analyst. Once the data warehouse has been built, the job of the EIS is infinitely easier than when there is no foundation of data on which the EIS analyst can operate. The figure shows how the data warehouse supports the need for EIS data.

With a data warehouse, the EIS analyst does not have to worry about the following:

1. Searching for the definitive source of data.
2. Creating special extract programs from existing systems.
3. Dealing with unintegrated data.
4. Compiling and linking detailed and summary data and the linkage

between the two.
5. Finding an appropriate time basis of data (finding historical data).
6. Management constantly changing its mind about what needs to be looked at next
In addition, the EIS analyst has a rich supply of summary data available

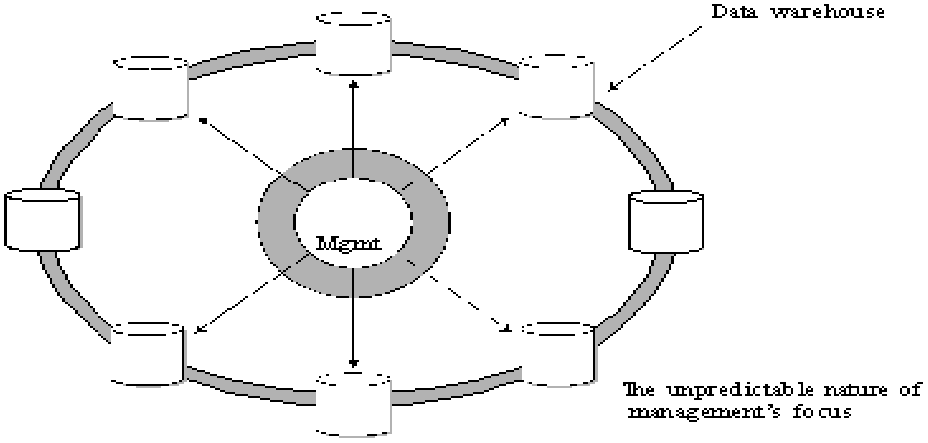


Figure The data warehouse supports management's need for EIS data.

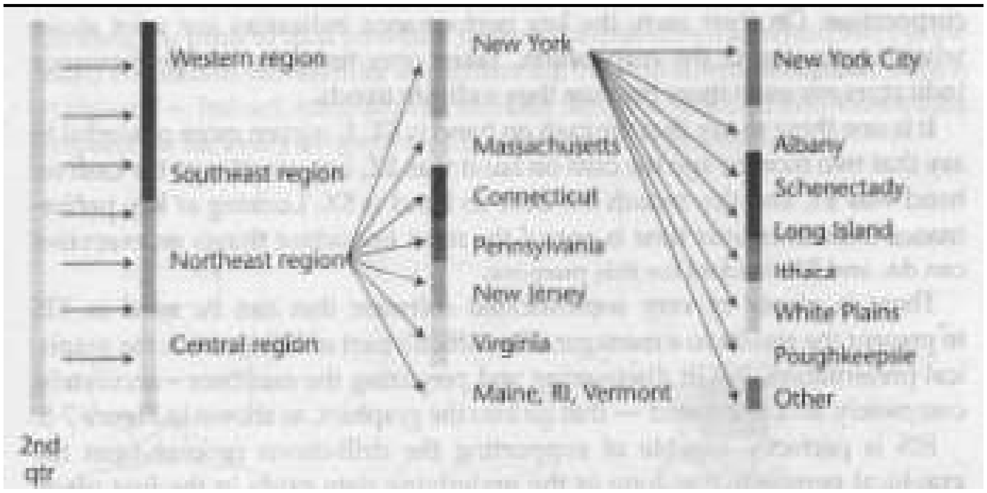
(6 marks)

The important reason why the data warehouse serves the needs of the world of EIS is this: The data warehouse operates at a low level of granularity. The data warehouse contains — for lack of a better word — *atomic data*. The atomic data can be shaped one-way, and then another. When management has a new set of needs for information that has never before been encountered in the corporation, the very detailed data found in the data warehouse sits, waiting to be shaped in a manner suited to management's needs. Because of the granular atomic data that resides in the data warehouse, analysis is flexible and responsive. The detailed data in the data warehouse sits and waits for future unknown needs for information. This is why the data warehouse turns an organization from a reactive stance to a proactive stance

b. Explain drill-down analysis and event mapping in context of EIS. (8)

Answer:

Dilling down refers to the ability to start at a summary number and to break that summary into a successfully finer set of summarizations. By being able to get at the detail beneath a summary number, the manager can get a feel for what is happening, especially where the summary number is surprising. example.



The manager has seen second quarter summary results and wants to explore them further. The manager then looks at the regions that have contributed to the summary analysis. The figures analyzed are those of the Western region, the southern region, the northern region and the the central region.

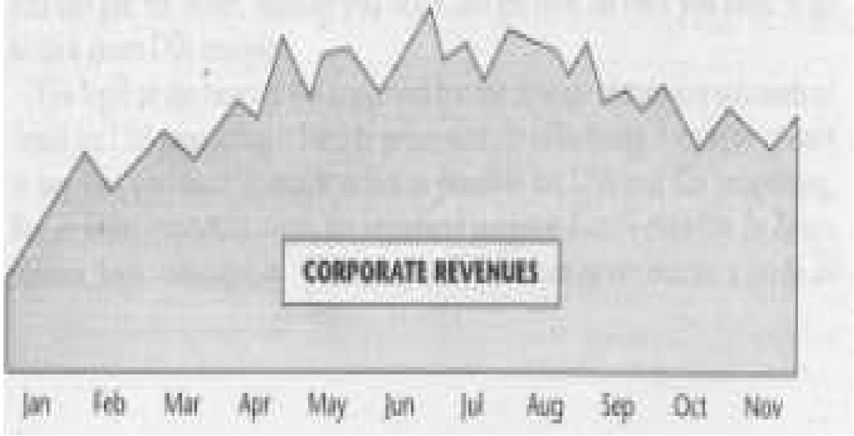
In looking at the numbers of each regions, the manager decides to look more closely at the Northeast region's numbers. The Northeast's numbers are made up of totals from New York, Massachusetts, Connectricts and many more. Of the states, the manager then decided to look more closely at the number for New York.

In each case , the manager has selected a path going from summary to detail, and then successively lower level. In such a fashion the manager can determine where the troublesome results are.

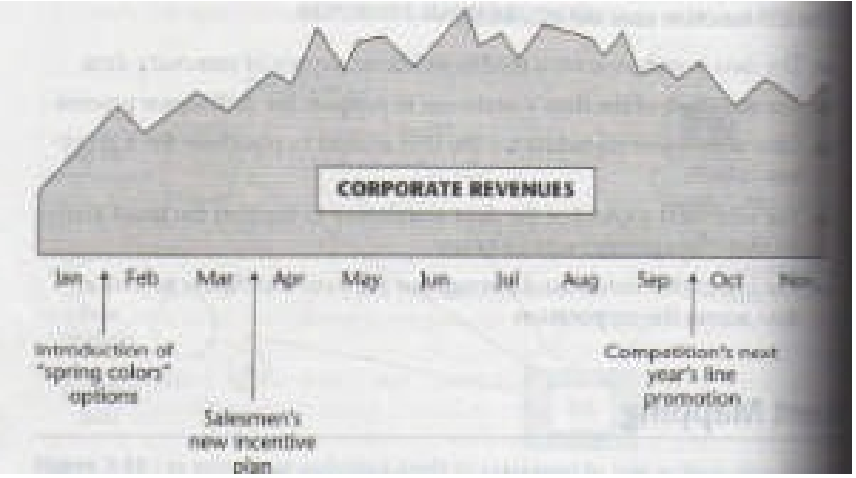
EIS is perfectly capable of supporting the drill down process from the graphical perspective as long as the underlying data exists in the first place. If the data to analyze does not exist, the drill down process becomes very tedious and awkward.

Eventmapping

A useful technique in using data warehouse for EIS processing is event mapping. The simplest way to depict event mapping is to start with a simple trend line.



Shows that corporate revenues have varied by months, as expected. The trend has been calculated from data found in the data warehouse. The trend of revenue s in and of itself is interesting but gives only superficial view of what is going on with the corporation. To enhance the view, events are mapped onto the trend line.



Three notable events have been mapped to the corporate revenue trend line- the introduction of a " spring colors" line of products, the advent of a sales incentives program and the introduction of competition. Now the relationship between corporate revenues and significant events begins to take on a different perspective.

For some sorts of events, event mapping is the only way to measure the results. Some events and activities cannot be measured directly and have to be measured in a correlative fashion. Cost justification and actual cost benefit cannot be measured any other way for some types of events.

Q.5 a. Discuss various OLAP operations in the multidimensional data modal. (7)

Answer:

OLAP operations

In the multidimensional model, data are organized into multiple dimensions. And each dimension contains multiple levels of abstraction defined by concept hierarchies. Using these hierarchies different types of OLAP operations are possible.

1) Roll up (drill-up)

Summarize data by climbing up hierarchy or by dimension reduction. In the given example of Figure 2, the Roll up operation on the dimension Location is given. It shows the Roll up from Cities to Countries and reduces the dimensions. It provides the summarized view of the dimension Location.

2) Drill down (roll down)

Reverse of roll-up from higher level summary to lower level summary or detailed data, or introducing new dimensions.In the given example of Figure 2, the Drill down operation on the time dimension is given. It shows the Drill down from Quarters to Months. It provides the detailed view of the dimension Time.

3) Slicing

Look at a subcube to provide more specific information. In the given example of

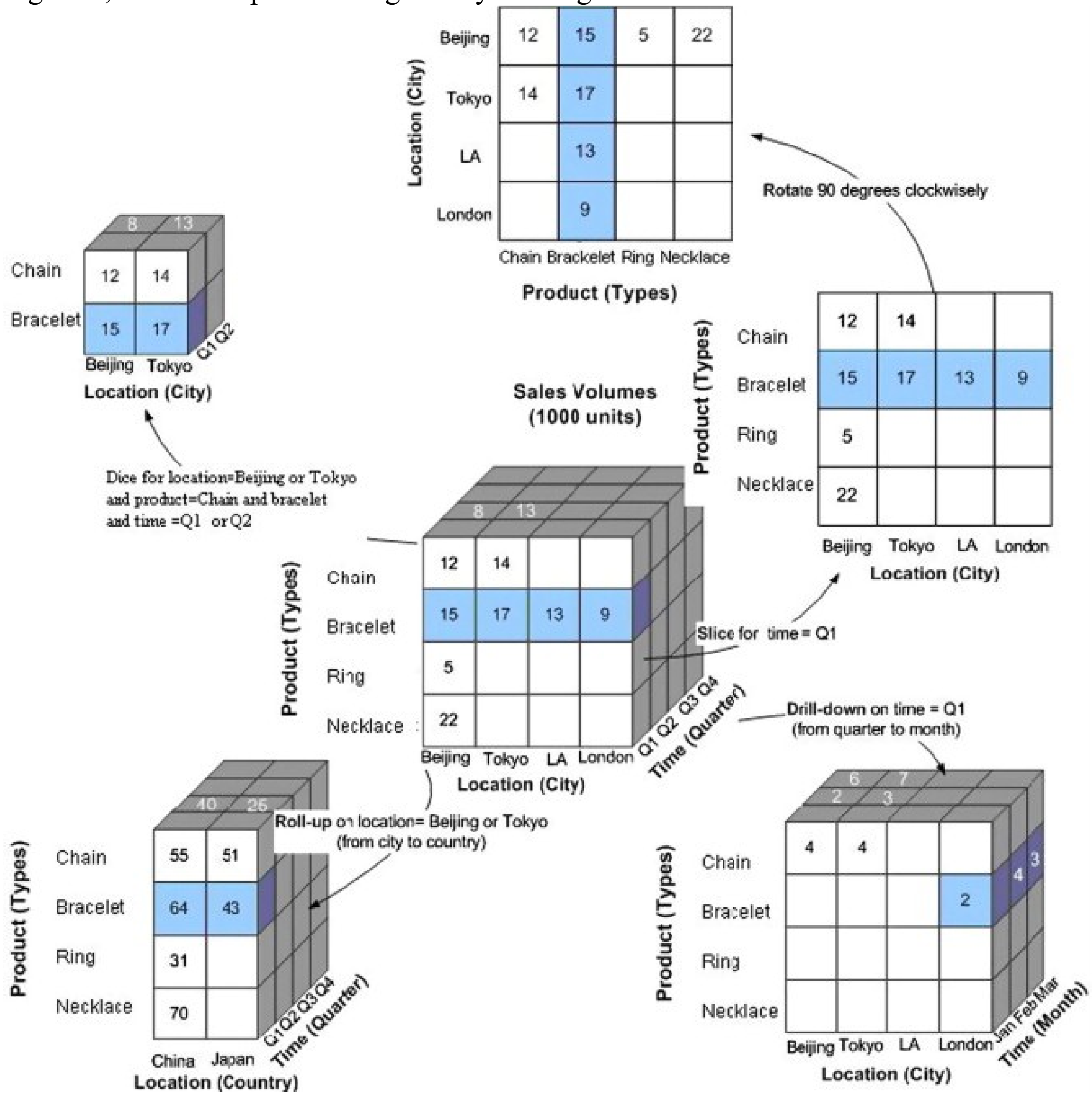
Figure 2, the Slice operation on Time Q1 is given.

4) Dicing

Rotate cube to look at another dimension or provides subcube. It shows the Dice operation for (location=”Toronto” or “Vancouver”) and (time= “Q1”or “Q2”) and (item=”home entertainment” or” computer”).

5) Pivoting (8)

Reorient the cube, visualization, 3D to series of 2D planes. It rotates the cube against axis and reset the whole dimension on another axis. In the given example of Figure 2, the Pivot operation is given by rotating cube.



b. What is generalization and its approaches? (5)

Answer:

It is process that abstracts a large set of task-relevant data in a database from a relatively low conceptual to higher conceptual levels.

2 approaches for Generalization

1) Data cube approach
2) Attribute-oriented induction approach : These method collets the task-relevant data using a relational database query and then perform generalization based on the examination in the relevant set of data.

c. Define Iceberg Cube and Shell Cube. (4)

Answer:

Q.6 a. What is Data Classification Process? How it is differ than Predication? (8)

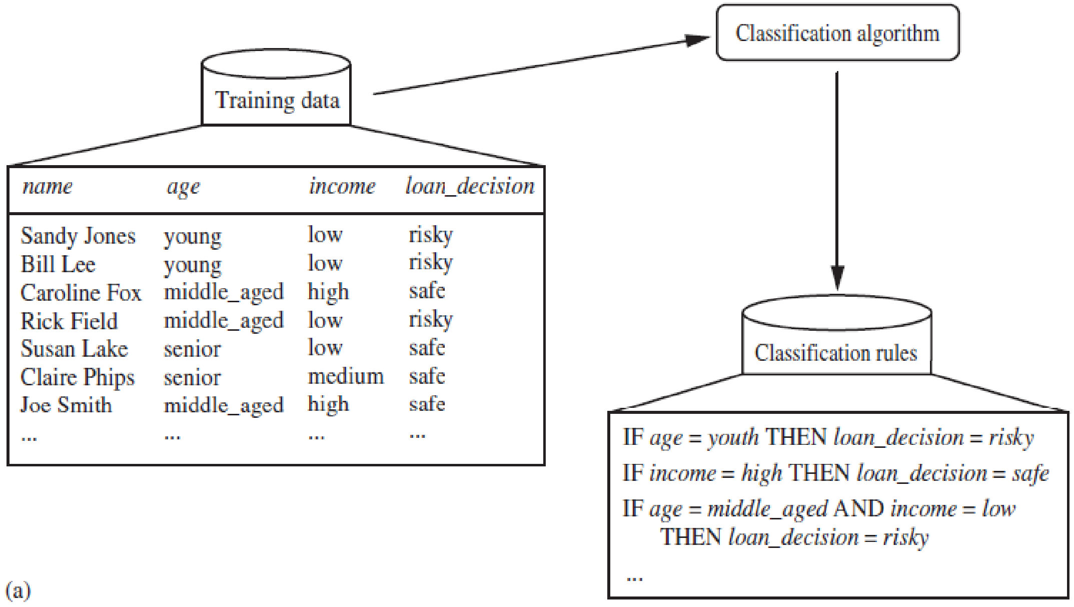
Answer:

Data Classification Process

Figure 6.1 The data classification process**:**

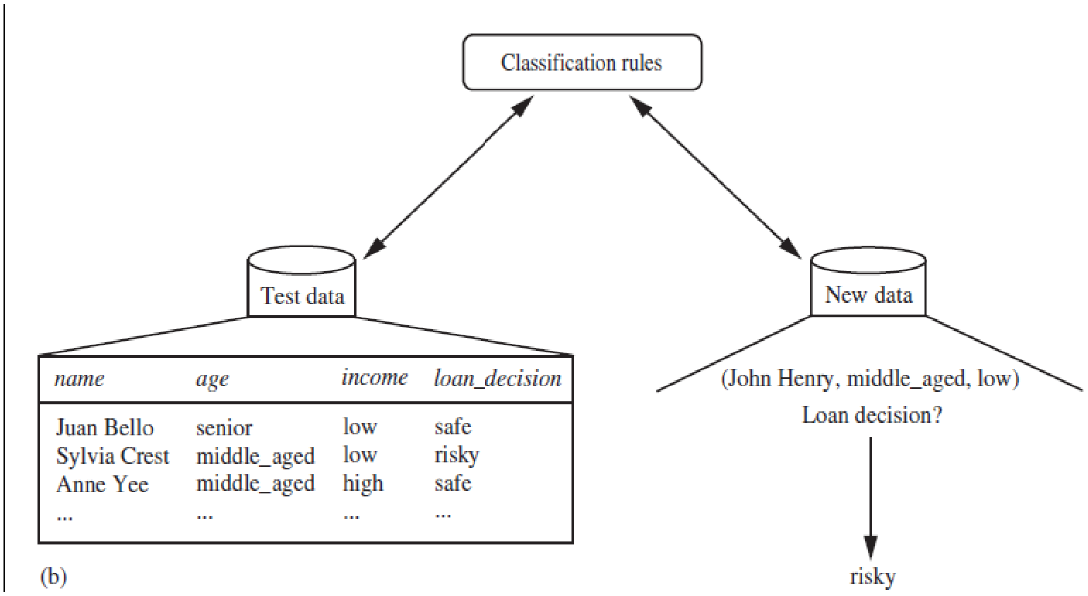
(a) ***Learning***: Training data are analyzed by a classification algorithm.Here, the class label attribute is *loan decision*, and the learned model or classifier is represented in the formof classification rules.

(b) ***Classification***: Test data are used to estimate the accuracy of the classificationrules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples.



name	*age*	*income*	*loan_decision*
Sandy Jones	young	low	risky
Bill Lee	young	low	risky
Caroline Fox	middle_aged	high	safe
Rick Field	middle_aged	low	risky
Susan Lake	senior	low	safe
Claire Phips	senior	medium	safe
Joe Smith	middle_aged	high	safe
...	...	...	...

IF *age* = *youth* THEN *loan_decision* = *risky*
IF *income* = *high* THEN *loan_decision* = *safe*
IF *age* = *middle_aged* AND *income* = *low*
THEN *loan_decision* = *risky*
...



name	*age*	*income*	*loan_decision*
Juan Bello	senior	low	safe
Sylvia Crest	middle_aged	low	risky
Anne Yee	middle_aged	high	safe
...	...	...	...

Difference b/w classification and prediction:

Classification	Prediction
1. Identify catagories	1. Identify numeric values
2. Discrete and unordered	2.ordered and continuous.
3. Done For prediction	3.main task of data mining

we instead wanted to predict the amount (in dollars)->PREDICATE that would be “safe”(-> CLASSIFICATION) for the bank to loan an applicant.

b. Why is tree pruning useful in decision tree induction? (2)

Answer:

Once tree is constructed, some modification to the tree might be needed to improve the performance of the tree during classification phase.

The pruning phase might remove redundant comparisons or remove subtrees to achieve better performance.

c. Prove that all nonempty subsets of a frequent itemset must also be frequent. (6)

Answer: Page 235

Q.7 a. Why is naive Bayesian classification called “naive”? Briefly outline the major ideas of naive Bayesian classification. (8)

Answer:

Naive Bayesian Classification (NBC) is referred to as naive since it makes the assumption that each of its inputs are independent of each other, an assumption which rarely holds true, and hence the word naive. Research has however shown that even though this assumption is often false, the technique still performs well, and hence NBC is seen as a simple yet powerful tool in the world of classification and machine learning.

Naïve Bayesian Classification

The naïve Bayesian classifier, or simple Bayesian classifier, works as follows:

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, $\boldsymbol{X} = (x1, x2, \ldots, xn)$, depicting n measurements made on the tuple from n attributes, respectively, $A1, A2, \ldots, An$.

2. Suppose that there are m classes, $C1, C2, \ldots, Cm$. Given a tuple, $\boldsymbol{X}$, the classifier will predict that $\boldsymbol{X}$ belongs to the class having the highest posterior probability, conditioned on $\boldsymbol{X}$. That is, the naïve Bayesian classifier predicts that tuple $\boldsymbol{X}$ belongs to the class Ci if and only if

$P(Ci|\boldsymbol{X}) > P(Cj|\boldsymbol{X})$ for $1 \le j \le m$, j 6= i.

Thus we maximize $P(Ci|\boldsymbol{X})$. The classCi for which $P(Ci|\boldsymbol{X})$ is maximized is called the *maximum posteriori hypothesis*. By Bayes' theorem (Equation (6.10)),

$P(Ci|\boldsymbol{X}) = P(\boldsymbol{X}|Ci)P(Ci) \,/\, P(\boldsymbol{X})$ (6.11)

3. As $P(\boldsymbol{X})$ is constant for all classes, only $P(\boldsymbol{X}|Ci)P(Ci)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C1) = P(C2) = \cdots = P(Cm)$, and we would therefore maximize $P(\boldsymbol{X}|Ci)$. Otherwise, we maximize $P(\boldsymbol{X}|Ci)P(Ci)$. Note that the class prior probabilities may be estimated by $P(Ci)=|Ci,D|/|D|$,where $|Ci,D|$ is the number of training tuples of class Ci in D.

4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(\boldsymbol{X}|Ci)$. In order to reduce computation in evaluating $P(\boldsymbol{X}|Ci)$, the naive 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). Thus,

$P(\boldsymbol{X}|Ci) = n\prod k = P(xk|Ci)$ (6.12)

$= P(x1|Ci) \times P(x2|Ci) \times \cdots \times P(xn|Ci)$.

5. In order to predict the class label of $\boldsymbol{X}$, $P(\boldsymbol{X}|Ci)P(Ci)$ is evaluated for each class Ci. The classifier predicts that the class label of tuple $\boldsymbol{X}$ is the class Ci if and only if

$P(\boldsymbol{X}|Ci)P(Ci) > P(\boldsymbol{X}|Cj)P(Cj)$ for $1 \le j \le m$, j 6= i. (6.15)

In other words, the predicted class label is the class Ci for which $P(\boldsymbol{X}|Ci)P(Ci)$ is the maximum.

b. What are the technology changes in bringing the system-of-record data into the data warehouse? (8)

Answer:

A short list of the technological challenges includes the following:

1. **A change in DBMS** — The system of record is in one DBMS, and the data warehouse is in another DBMS.
2. **A change in operating systems** — The system of record is in one operating system, and the data warehouse is in another operating system.
3. **The need to merge data from different DBMSs and operating systems** — The system of record spans more than one DBMS and/or operating system. System-of-record data must be pulled from multiple DBMSs and multiple operating systems and must be merged in a meaningful way.
4. **The capture of the Web-based data in the Web logs** — Once captured, how can the data be freed for use within the data warehouse?
5. **A change in basic data formats** — Data in one environment is stored in ASCII, data in the data warehouse is stored in EBCDIC, and so forth.

Another important technological issue that sometimes must be addressed is the volume of data. In some cases, huge volumes of data will be generated in the legacy environment. Specialized techniques may be needed to enter them into the data warehouse. For example, clickstream data found in the Web logs needs to be preprocessed before it can be used effectively in the data ware- house environment.

There are other issues. In some cases, the data flowing into the data warehouse must be cleansed. In other cases, the data must be summarized. A host of issues relate to the mechanics of the bringing of data from the legacy environment into the data warehouse environment.

Q.8 a. What do you mean by cluster analysis? What are the typical requirements of clustering in data mining? (7)

Answer:

The process of grouping a set of physical or abstract objects into classes of *similar* objects is called clustering. A cluster is a collection of data objects that are *similar* to one another within the same cluster and are *dissimilar* to the objects in other clusters. By automated clustering, we can identify dense and sparse regions in object space and, therefore, discover overall distribution pat- terns and interesting correlations among data attributes. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing.

The following are typical requirements of clustering in data mining:

i. **Scalability:** Many clustering algorithms work well on small data sets containing fewer than several hundred data objects; however, a large database may contain millions of objects. Clustering on a *sample* of a given large data set may lead to biased results. Highly scalable clustering algorithms are needed.

ii. **Ability to deal with different types of attributes:** Many algorithms are designed to cluster interval-based (numerical) data. However, applications may require cluster- ing other types of data, such as binary, categorical (nominal), and ordinal data, or mixtures of these data types.

iii. **Discovery of clusters with arbitrary shape:** Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures. Algorithms based on such distance measures tend to find spherical clusters with similar size and density. However, a cluster could be of any shape. It is important to develop algorithms that can detect clusters of arbitrary shape.

iv. **Minimal requirements for domain knowledge to determine input parameters:** Many clustering algorithms require users to input certain parameters in cluster analysis (such as the number of desired clusters). The clustering results can be quite sensi- tive to input parameters. Parameters are often difficult to determine, especially for data sets containing high-dimensional objects. This not only burdens users, but it also makes the quality of clustering difficult to control.

v. **Ability to deal with noisy data:** Most real-world databases contain outliers or missing, unknown, or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.

vi. **Incremental clustering and insensitivity to the order of input records**: Some clus- tering algorithms cannot incorporate newly inserted data (i.e., database updates) into existing clustering structures and, instead, must determine a new clustering from scratch. Some clustering algorithms are sensitive to the order of input data. That is, given a set of data objects, such an algorithm may return dramatically different clusterings depending on the order of presentation of the input objects. It is important to develop incremental clustering algorithms and algorithms that are insensitive to the order of input.

vii. **High dimensionality:** A database or a data warehouse can contain several dimensions or attributes. Many clustering algorithms are good at handling low-dimensional data, involving only two to three dimensions. Human eyes are good at judging the quality of clustering for up to three dimensions. Finding clusters of data objects in highdimensional space is challenging, especially considering that such data can be sparse and highly skewed.

viii. **Constraint-based clustering:** Real-world applications may need to perform clustering under various kinds of constraints. Suppose that your job is to choose the locations for a given number of new automatic banking machines (ATMs) in a city. To decide upon this, you may cluster households while considering constraints such as the city's rivers and highway networks, and the type and number of customers per cluster. A challenging task is to find groups of data with good

clustering behavior that satisfy specified constraints.
ix. **Interpretability and usability:** Users expect clustering results to be interpretable, com- prehensible, and usable. That is, clustering may need to be tied to specific semantic interpretations and applications. It is important to study how an application goal may influence the selection of clustering features and methods.

b. Explain the PAM (partition around medoids) algorithm and its phases. (6)

Answer:
PAM stands for "partition around medoids". The algorithm is intended to find a sequence of objects called *medoids* that are centrally located in clusters. Objects that are tentatively defined as medoids are placed into a set S of *selected objects*.
If O is the set of objects that the set U = O − S is the set of *unselected objects*. *The goal of the algorithm is to minimize the average dissimilarity of objects to their closest selected object*. Equivalently, we can minimize the sum of the dissimilarities between object and their closest selected object.
In a general analysis the algorithm proceed this way:
Build phase:
1. Choose k entities to become the medoids, or in case these entities were provided use the them as the medoids;
2. Calculate the dissimilarity matrix if it was not informed;
3. Assign every entity to its closest medoid;
Swap phase:
4. For each cluster search if any of the entities of the cluster lower the average dissimilarity coefficient, if it does select the entity that lower the most this coefficient as the medoid for this cluster;
5. If at least the medoid from one cluster has changed go to (3), else end the algorithm.
As was said the PAM algorithm works with a matrix of dissimilarity, and to calculate this matrix the algorithm can use two metrics the first one is the euclidean, that are the root sumof-squares of differences, while the second one is the manhattan distance that are the sum of absolute distances.

c. Define outlier mining. (3)

Answer:
Outlier mining: In statistics, an **outlier** is an observation that is numerically distant from the rest of the data. Outliers can occur by chance in any distribution, but they are often indicative either of measurement error or that the population has a heavy-tailed distribution.
Naive interpretation of statistics derived from data sets that include outliers may be misleading.
For example, if one is calculating the average temperature of 10 objects in a room, and nine of them are between 20 and 25 degrees Celsius, but an oven is at 175 °C, the median of the data will be between 20 and 25 °C but the mean temperature will be between 35.5 and 40 °C. In this case, the median better reflects t he temperature of a randomly sampled

object than the mean; naively interpreting the mean as "a typical sample", equivalent to the median, is incorrect. As illustrated in this case, outliers may be indicative of data points that belong to a different population than the rest of the sample set.

Q.9 a. Explain how data mining used in retail industry and in biomedical field. (8)

Answer:

data mining in retail industry

 Design and construction of data warehouses based on the benefits of data mining
 Multidimensional analysis of sales, customers, products, timeand region
 Analysis of the effectiveness of sales campaigns
 Customer retention-analysis of customer loyalty
 Purchase recommendation and cross-reference of item

data mining in biomedical field

 Semantic integration of heterogeneous, distributed genome databases
 Similarity search and comparison among DNA sequences
 Association analysis: identification of co-occuring gene sequences
 Path analysis: linking genes to different stages of disease development
 Visualization tools and genetic data analysis

b. What are mining? Explain the techniques in web mining. (8)

Answer:

Web Mining: Web mining - is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are **Web usage mining**, **Web content mining** and **Web structure mining**.

A. Web Content Mining Web content mining targets the knowledge discovery, in which the main objects are the traditional collections of multimedia documents such as images, video, and audio, which are embedded in or linked to the web pages.

It is also quite different from Data mining because Web data are mainly semi-structured and/or unstructured, while Data mining deals primarily with structured data. Web content mining is also different from Text mining because of the semi-structure nature of the Web, while Text mining focuses on unstructured texts.

B. Web Structure Mining

Web Structure Mining focuses on analysis of the link structure of the web and one of its purposes is to identify more preferable documents. The different objects are linked in some way. The intuition is that a hyperlink from document A to document B implies that the author of document. A thinks document B contains worthwhile information. Web structure mining helps in discovering similarities between web sites or discovering important sites for a particular topic or discipline or in discovering web communities.

C. Web Usage Mining

Web Usage Mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW. Web usage mining, discover user navigation patterns from web data, tries to discovery the useful information from the secondary data derived from the interactions of the users while surfing on the Web. Web usage mining collects the data from Web log records to discover user access patterns of web pages. There are several available research projects and commercial tools that analyze those patterns for different purposes. The insight knowledge could be utilized in personalization, system improvement, site modification, business intelligence and usage characterization.

TEXT BOOK

Data Mining, Concepts and Techniques, Jiawei Han and Micheline Kamber, Elsevier, Second Edition, 2006