Practical No

AIM: Analytical Exercise-I

OBJECTIVE: On completion of this exercise student will able to know about…

- Know about difference between Data Query and Data Mining.
- Understand the depth and importance of data mining.
- Categories the various techniques of data mining.

THEORY:

About Data mining

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration.

Data mining refers to extracting or ―mining‖ knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named ―knowledge mining from data,‖

Many other terms carry a similar or slightly different meaning to data mining, such as *knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.*

Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD.

Data mining as simply an essential step in the process of knowledge discovery

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

Database Queries vs. Data Mining

Virtually all modern, commercial database systems are based on the relational model formalized by Codd in the 60s and 70s (Codd, 1970) and the SQL language (Date,2000) which allows the user to efficiently and effectively manipulate a database. In this model a database table is a representation of a mathematical relation, that is, a set of items that share certain characteristics or attributes. Here, each table column represents an attribute of the relation and each record in the table represents a member of this relation.

In relational databases the tables are usually named after the kind of relation they represent. Figure 1 is an example of a table that represents the set or relation of all the customers of a particular store. In this case the store tracks the total amount of money spent by its customers.

ID	Name	Zip	Sex	Age	Income	Children	Car	Total Spent
5	Peter	05566	M	35	$40,000	2	Mini Van	$250.00
…	…	…		…	…	…	…	…
…	…	…		…	…	…	…	…
22	Maureen	04477	F	26	$55,000	0	Coupe	$50.00

Figure1 A relational database table representing customers of a store.

Relational databases do allow for the manipulation of the tables and the data within them. The most fundamental operation on a database is the ***query***. This operation enables the user to retrieve data from database tables by asserting that the retrieved data needs to fulfill certain criteria. As an example, consider the fact that the store owner might be interested in finding out which customers spent more than $100 at the store. The following query returns all the customers from the above customer table that spent more than $100:

SELECT * FROM CUSTOMER_TABLE WHERE TOTAL_SPENT > $100;

This query returns a list of all instances in the table where the value of the attribute Total Spent is larger than $100.

As this example highlights, queries act as filters that allow the user to select instances from a table based on certain attribute values. It does not matter how large or small the database table is, a query will simply return all the instances from a table that satisfy the attribute value constraints given in the query. This straightforward approach to retrieving data from a database has also a drawback. Assume for a moment that our example store is a large store with tens of thousands of customers(perhaps an online store). Firing the above query against the customer table in the database will most likely produce a result set containing a very large number of customers and not much can be learned from this query except for the fact that a large number of customers spent more than $100 at the store. Our innate analytical capabilities are quickly overwhelmed by large volumes of data.

This is where differences between querying a database and mining a database surface. In contrast to ***a query which simply returns the data that fulfills certain constraints, data mining***

constructs models of the data in question. The models can be viewed as high level summaries of the underlying data and are in most cases more useful than the raw data, since in a business sense they usually represent understandable and actionable items (Berry &Linoff, 2004). Depending on the questions of interest, data mining models can take on very different forms. They include decision trees and decision rules for classification tasks, association rules for market basket analysis, as well as clustering for market segmentation among many other possible models.

To continue our store example, in contrast to a query, a data mining algorithm that constructs decision rules might return the following set of rules for customers that spent more than $100 from the store database:

IF AGE > 35 AND CAR = MINIVAN THEN TOTAL SPENT > $100
OR
IF SEX = M AND ZIP = 05566 THEN TOTAL SPENT > $100

These rules are understandable because they summarize hundreds, possibly thousands, of records in the customer database and it would be difficult to glean this information off the query result. The rules are also actionable. Consider that the first rule tells the store owner that adults over the age of 35 that own a mini van are likely to spend more than $100. Having access to this information allows the store owner to adjust the inventory to cater to this segment of the population, assuming that this represents a desirable cross-section of the customer base. Similar with the second rule, male customers that reside in a certain ZIP code are likely to spend more than $100. Looking at census information for this particular ZIP code the store owner could again adjust the store inventory to also cater to this population segment presumably increasing the attractiveness of the store and thereby increasing sales.

As we have shown, the fundamental difference between database queries and data mining is the fact that in contrast to queries data mining does not return raw data that satisfies certain constraints, but returns models of the data in question. These models are attractive because in general they represent understandable and actionable items. Since no such modeling ever occurs in database queries we do not consider running queries against database tables as data mining, it does not matter how large the tables are.

Data querying is the process of asking questions of data in search of a specific answer. Unlike many forms of search (i.e. Google), queries are normally structured and require specific parameters or code, known as SQL (Structured Query Language). A query could be written to answer questions like, "How many items were sold in Region 2 last month?"

Data mining is the process of sorting through large amounts of data to identify patterns and relationships using statistical algorithms. These relationships may help us to understand which factors affected the outcome of something, or they may be used to predict future outcomes. Data mining might be used to answer questions like, "What factors affected sales in Region 2 last month?" Knowing which factors drive sales in the past could help to predict or make estimates about sales in the future.

Data Mining Task

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and predictive.

1. **Descriptive** mining tasks characterize the general properties of the data in the database.
2. **Predictive** mining tasks perform inference on the current data in order to make predictions.

Common data mining tasks

1. Classification [Predictive]
2. Clustering [Descriptive]
3. Association Rule Discovery [Descriptive]
4. Sequential Pattern Discovery [Descriptive]
5. Regression [Predictive]

DataMining Application

1. Health Care
    a. Drug development – to help uncover less expensive but equally effective drug treatments.
    b. Medical diagnostics – imaging, real-time monitoring (e.g. Predicting women at high risk for emergency C-section).
    c. Insurance claims analysis – identify customers likely to buy new policies; define behavior patterns of risky customers

2. Business and Finance
    a. Banks - to detect which customers are using which products so they can offer the right mix of products and services to better meet
    b. Customer needs – cross sell and up sell.
    c. Credit card companies - to assist in mailing promotional materials to people who are most likely to respond.
    d. Lenders - to determine which applicants are most likely to default on a loan.

3. Sports and Gambling
    a. Sports teams – to analyze data to determine favorable player matchups and call the best plays
    b. Gaming industry - to analyze customer gambling trends at casinos.
    c. Sports Fanatics – to predict which teams will be chosen for tournament berths as well as to predict game winners.

4. Education
    a. Enrollment Management – which students are likely to attend
    b. Retention/Graduation Analysis – which students will remain enrolled after the first year and/or through graduation
    c. Donation Prediction – who is likely to donate and how much might they donate.

5. Other Application Areas
    a. Insurance – pricing, fraud detection, risk analysis
    b. Stock Market – market timing, stock selection, risk analysis
    c. Transportation – performance & network optimization to predict life-cycle costs of road pavement
    d. Telecommunications – churn reduction
    e. Retail – market basket analysis to help determine marketing strategies

EXCERSICE:

1) Define data mining?
2) For each of the following, identify if it is data query or data mining task(s).
    A. To find characteristics that differentiate people who had repeat knee construction from those who needed surgery only once
    B. A stock market analyst has been asked by his client to predict the future prices of 10 stocks 3 months in advance.
    C. When customers visit my website, what products together are they most likely to order?
    D. Develop a profile of credit card customers who are likely to carry an average monthly balance of more than $1,000.00
    E. Determine the number of customers who spent more than $100 last month in a local video store.
    F. In a locality, which group is greater in number – customers holding policies in one or several insurance companies?
    G. Do meaningful attribute relationships exist in a database containing information about credit card customers?
    H. Do single men play more golf than married men?
    I. Determine whether any credit card transaction was made by a customer last week.
3) Visit website www.kdnuggets.com and describe any five software of data mining.
4) Visit http://archive.ics.uci.edu/ml/datasets.html and describe any five datasets.

Type your text