CHAPTER 2

Classification and prediction

2.1 Classification and prediction

2.2 Cluster Analysis

2.3 Outlier Analysis

2.4 Evolution Analysis

2.5 Pattern Interestingness Measures

2.6 Classification of Data Mining Systems

2.7 Major Issues in Data Mining:

2.8 Review questions

2.9 References

2. Classification and prediction

2.1 Classification and prediction

We treat many things as a group of things e.g., staff, students' etc. In order to define a class (a group of entities) a set of models that define and distinguish data classes or concept are delineated together. Using this class we get the ability to predict whether any new model belongs to this class or not i.e. a data model for whose class value is unknown can be predicted based on classification rules. There are various ways to apply rules e.g., classification (if-then) rules, decision trees, mathematical formula or neural networks. Classification can be used to predict the class label of the data object. However, classification is most useful in predicting certain missing values or unavailable data within a class. Normally, when classification is used to predict missing values in numeric data this is referred to as prediction. Data values prediction is more useful over class label assignment to an unknown object. This helps in making trend analysis based on available data.

2.2 Cluster Analysis

Clustering, analysis a data set without consulting a known class label. Class labels are not present in the training data, as they are not known to begin with. Clustering is used to divide a data set into classes (by generating labels for them) using the principle of maximizing the intra class similarity and minimizing inter class similarity. Within the data set clusters are formed so that objects which are similar are grouped together and objects that are very different fall into other clusters. Once the data derived for any object inclusion. Clustering, thus also facilitates taxonomy formation i.e. organization of observed objects into a hierarchy of classes that group similar things together.

2.3 Outlier Analysis

A data set may contain objects, which do not comply with the general behavior of the model of the set. These data objects are termed as **Outliers.** Outliers may be identified using statistical tests that

assume a distribution or a probability model for the data. Also, distance measures or deviation based methods identify an outlier by examining the main characteristics of the defining group.

Outlier analysis may be used to uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts of items for a given account number in comparison with the regular charges that are incurred on the same amount.

2.4 Evolution Analysis

Evolution Analysis describes and models trends and regularities that occur in data where behavior changes over a period of time.

Example

Suppose major stock market data (time series data) of several years are available for the Hong Kong stock exchange and you would like to invest in pharmaceutical companies. A data mining study of the stocks in the stock exchange may show evolution regularities for overall stocks of some particular pharmaceutical companies.

2.5 Pattern Interestingness Measures

Data mining systems have potential for generating millions of patterns or rules. The question is how do you identify patterns that are interesting? Typically only a small fraction of the patterns generated would be of use or of interest to particular user. What exactly makes a pattern interesting? An interesting pattern may be:

1. it is easily understood by human beings.
2. valid on a new or test data with some degree of certainty.
3. potentially useful
4. novel

A pattern is interesting if it validates a hypothesis that the user wishes to confirm. Interesting pattern represents knowledge.

Subjective interestingness measures are based on user beliefs of the data. These are based on the structure of the patterns interesting if they are unexpected or offer strategic information, which the user can act upon. Patterns are also interesting when a user's hypothesis or hunch is confirmed.

Several objective measures of interestingness of patterns exist. These are based on the structure of the pattern and the statistics underlying them. An objective measure of association rules of the form x=>y is the rule support i.e. the percentage of transactions form a transaction base which a given rule satisfies. This is the probability of P (XUY), where XUY is a transaction that contains both X and Y. Another objective measure for the association rules is confidence which assesses the degree of certainty of the association. This is the conditional probability P(X|Y).More formally support and confidence are defined as

Support (X=>Y)=P(XUY)

Confidence (X=>Y) = P (Y|X)

Although objective measures help identify interesting patterns they are insufficient unless otherwise combined with the subjective interestingness.

Completeness refers to the case where all possible interesting patterns are generated by data mining.

Can a data mining system be made such that it generates only interesting patterns? This would highly describe since it introduces efficiency of the data mining system. However, this kind of optimization remains a key challenging issue in data mining.

2.6 Classification of Data Mining Systems

Kinds of database mined:

Different systems may themselves be classified into different categories. So the king of database mined may give rise to a category of the data mining system.

Kinds of Knowledge Mined:

Different types of mining process such as characterization, discrimination, association, classification, clustering or outlier analysis may produce different types of Knowledge. A comprehensive data mining system usually provides different functionalities. Moreover, data mining systems may be classified based on the levels of abstraction or granularity that they deal with.

Kinds of Techniques Utilized:

Categorization of data mining systems may be done according to the underlying data mining technique that they employ. Techniques may be differentiated based on the level of user interaction is involved or methods of data analysis employed(data base oriented, data warehouse oriented, machine learning, statistics, visualization, neural networks, and so on).A sophisticated data mining system will offer many techniques for data mining.

Applications adapted:

Data mining systems can also be distinguished based on the applications they adapt. There may be specific data mining systems for telecommunications, DNA, stock markets etc.

2.7 Major Issues in Data Mining:

Mining methodology and user interaction issues:

These reflect the kinds of knowledge mined the ability to mine knowledge at multiple granularities, the use of ad hoc mining and knowledge visualization.

Mining different kinds of knowledge from databases: Since different users may be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, and association, classification, clustering, trend and deviation analysis. There tasks may use the same database in different ways and require the development of numerous data mining techniques.

Interactive mining at different levels of knowledge abstraction: It is very difficult to predict what sort of knowledge is available within the database; therefore the data mining process should be interactive. Interactive mining allows users to focus on search for patterns; providing and refining data mining systems.

Incorporation of background knowledge: Background knowledge or information regarding a domain under study, may be used to guide the discovery process and allow patterns to be expressed in precise terms and at different levels of abstraction. Domain knowledge can help judge the interestingness of discovered patterns.

Data mining query languages and ad hoc data mining: SQL allows for ad hoc queries. In a similar way data mining query languages need to be developed to enable users to describe ad hoc data mining tasks.

Presentation and visualization of data mining tasks: Discovered knowledge must be represented in high level languages, visual representations or other expressive forms. Knowledge representation techniques like, trees, tables, and charts, cross tabs and matrices and curves.

Handling noisy and incomplete data: Data may contain noise, exceptional cases or incomplete data objects. These may cause confusion while mining for regularities causing the knowledge model to over fit the data. Cleaning methods and analysis methods are required to handle noisy data and outlier methods for discovery of exceptional cases. Some times the accuracy of the data mining patterns discovered may be poor due to noise.

Pattern Evaluation: A data mining system may uncover thousands of patterns .Only few of the patterns may be interesting to the user, representing uncommon knowledge with novelty. The search space may be reduced using subjective measures that estimate the value of a pattern.

2.8 Review questions

1) Explian about classification of Data Mininig?
2) Explain about Design Issues of Data Mininig?
3) EXplain about pattern interesting measures?
4) Explain about Outlier Analysis?

2.9 References

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition