UECS3213 / UECS3453 DATA MINING

SESSION: January 2019

TUTORIAL 1

Chapter 1: Introduction to Data Mining

1. Define data mining.

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data such as knowledge rules, constraints, and regularities from data.
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns, trends or relationships.

2. Describe the steps involved in data mining when viewed as a process of *knowledge discovery*.

The steps involved in data mining when viewed as a process of knowledge discovery are as follows:

1) **Data cleaning**, a process that removes or transforms 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
5) **Data mining**, an essential process where intelligent and efficient methods are applied in order to extract patterns
6) **Pattern evaluation**, a process that identifies 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

3. Discuss whether or not each of the following activities is a *data mining task*.

a) Dividing the customers of a company according to their gender.

- No. This is a simple database query.

b) Dividing the customers of a company according to their profitability.

- No. This is an accounting calculation, followed by the application of a threshold.
- However, predicting the profitability of a new customer would be data mining.

c) Computing the total sales of a company.

- No. Again, this is simple accounting.

d) Sorting a student database based on student identification numbers.

- No. Again, this is a simple database query.

e) Predicting the outcomes of tossing a (fair) pair of dice.

- No. Since the die is fair, this is a probability calculation.
- If the die were not fair, and we needed to estimate the probabilities of each outcome from the data, then this is more like the problems considered by data mining.
- However, in this specific case, solutions to this problem were developed by mathematicians a long time ago, and thus, we wouldn't consider it to be data mining.

f) Predicting the future stock price of a company using historical records.

- Yes. We would attempt to create a model that can predict the continuous value of the stock price. This is an example of the area of data mining known as predictive modelling.
- We could use regression for this modelling, although researchers in many fields have developed a wide variety of techniques for predicting time series.

g) Monitoring the heart rate of a patient for abnormalities.

- Yes. We would build a **model** of the normal behavior of heart rate and raise an alarm when an unusual heart behavior occurred. This would involve the area of data mining known as anomaly detection.
- This could also be considered as a classification problem if we had examples of both normal and abnormal heart behavior.

h) Monitoring seismic waves for earthquake activities.

- Yes. In this case, we would build a **model** of different types of seismic wave behavior associated with earthquake activities and raise an alarm when one of these different types of seismic activity was observed.
- This is an example of the area of data mining known as classification.

i) Extracting the frequencies of a sound wave.

- No. This is signal processing.

4. Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as *clustering*, *classification*, *association rule mining*, and *anomaly detection* can be applied.

The following are examples of possible answers.

- **Clustering** can group results with a similar theme and present them to the user in a more concise form, e.g., by reporting the 10 most frequent words in the cluster.
- **Classification** can assign results to pre-defined categories such as "Sports," "Politics," etc.
- **Sequential association analysis** can detect that that certain queries follow certain other queries with a high probability, allowing for more efficient caching.
- **Anomaly detection** techniques can discover unusual patterns of user traffic, e.g., that one subject has suddenly become much more popular. Advertising strategies could be adjusted to take advantage of such developments.

5. Define each of the following data mining functionalities: *characterization, discrimination, association and correlation analysis, classification, prediction, clustering*, and *evolution analysis*.

   Give examples of each data mining functionality, using a real-life database that you are familiar with.

- **Characterization** is a summarization of the general characteristics or features of a target class of data. For example, the characteristics of students can be produced, generating a profile of all the University first year computing science students, which may include such information as a high GPA and large number of courses taken.
- **Discrimination** is 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. For example, the general features of students with high GPA's may be compared with the general features of students with low GPA's. The resulting description could be a general comparative profile of the students such as 75% of the students with high GPA's are fourth-year computing science students while 65% of the students with low GPA's are not.
- **Association** is the discovery of *association rules* showing attribute-value conditions that occur frequently together in a given set of data. For example, a data mining system may find association rules like major(X, "computing science"") ⇒ owns(X, "personal computer") [support = 12%, confidence = 98%] where X is a variable representing a student. The rule indicates that of the students under study, 12% (support) major in computing science and own a personal computer. There is a 98% probability (confidence, or certainty) that a student in this group owns a personal computer.
- **Classification** differs from **prediction** in that the former is to construct a set of models (or functions) that describe and distinguish data class or concepts, whereas the latter is to predict some missing or unavailable, and often numerical, data values. Their similarity is that they are both tools for prediction: Classification is used for predicting the class label of data objects and prediction is typically used for predicting missing numerical data values.
- **Clustering** analyzes data objects without consulting a known class label. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. Each cluster that is formed can be viewed as a class of objects. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

- **Data evolution analysis** describes and models regularities or trends for objects whose behavior *changes over time*. Although this may include characterization, discrimination, association, classification, or clustering of *time-related data*, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

Discussion Questions

6. What is the *difference* between discrimination and classification? Between characterization and clustering? Between classification and prediction? For each of these pairs of tasks, how are they similar?

- **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.
- **Characterization** differs from **clustering** in that the former refers to a summarization of the general characteristics or features of a target class of data while the latter deals with the analysis of data objects without consulting a known class label. This pair of tasks is similar in that they both deal with grouping together objects or data that are related or have high similarity in comparison to one another.
- **Classification** differs from **prediction** in that the former is the process of finding a set of models (or functions) that describe and distinguish data class or concepts while the latter predicts missing or unavailable, and often numerical, data values. This pair of tasks is similar in that they both are tools for prediction: Classification is used for predicting the class label of data objects and prediction is typically used for predicting missing numerical data values.

7. What are the major *challenges* of mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuple data set)?

- One challenge to data mining regarding performance issues is the **efficiency** and **scalability** of data mining algorithms. Data mining algorithms must be efficient and scalable in order to effectively extract information from large amounts of data in databases within predictable and acceptable running times.
- Another challenge is the **parallel**, **distributed**, and **incremental** processing of data mining algorithms. The need for parallel and distributed data mining algorithms has been brought about by the huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods. Due to the high cost of some data mining processes, incremental data mining algorithms incorporate database updates without the need to mine the entire data again from scratch.

The End