MSIS-DL 435 Session 1
Shaddy Abado
Data Warehousing & Data Mining
Winter 2013

Introduction

In this handout brief overviews of data warehousing and data mining principles are presented. This handout will be followed by a discussion of these principles to enhance the student's understanding of this session's learning objectives.

Reading for this session: follow up

Chapters 1, 2, 3 of the textbook:

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Reading, MA: Addison-Wesley.

What is Data Warehousing?

The two fields of *Data Warehousing* and *Data Mining* are attracting a great deal of attention in recent years due to the large rate of data accumulation from modern computer systems. Data warehousing is a set of tools designed to store large amount of data which is collected from multiple sources. These sources may vary based on the application, such as credit card transactions, web logs, industrial sales, and satellite observations.

A data warehouse is usually defined as "*A subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of decision-making process.*" (W.H. Inmon, in "Building a Data Warehouse, Wiley 1996). *Subject-oriented* means that a data warehouse is organized around one or several subjects of analysis (e.g., patients, products, suppliers). *Integrated* expresses the fact that the data warehouse is usually constructed by integrating data from various heterogeneous sources (i.e., the input data can be stored in a variety of formats such as flat files, spreadsheets, or relational tables)**.** *Time-variant* expresses the fact that the data warehouse keeps historical track of the data (e.g., past ten quarters). Finally, *nonvolatile* indicates that data modification and removal are not allowed in a data warehouse. The only allowed operation is the purging of data that is no longer used.

What is Data Mining?

The data stored in a data warehouse can provide information which can be turned into knowledge by data mining. Data mining is the process of automatically "mining" non-trivial useful information in large data repositories that might otherwise remain unknown. In addition, data mining can be used to predict the outcome of a future observation. For example, data mining techniques can assist retailers which collect up-to-date data records about costumer purchases.

These collected records can be utilized to understand the need of their customers and conduct future market campaigns. Similarly, data mining techniques can also be used in other fields, such as medicine, science, and engineering where they can be used to identify lethal illnesses, analyze organic compounds, and predict electric loads, respectively.

Prior to applying data mining, the data is prepared for mining through an appropriate data *preprocessing*, for example: removing noise and inconsistent data, and combining multiple data sources. After that, depending on the application, the proper data mining technique is applied. To insure that the data mining technique outcome is appropriately integrated into the decision support system, a *postprocessing*, such as visualization, is then required.

In this course, three data mining techniques will be discussed: classification, clustering, and association rules. Prior to providing a briefly definition for each data mining technique, the definitions of various types of data mining variables and different statistical parameters which we will used during this course are provided bellow.

Types of Data Variables and Data Mining Tasks

The complete set of interrelated data available for analysis is called a *dataset*. Datasets are usually stored either in *flat files* or in *relational databases*. Each record in the dataset is called an *instance* or *object*. Each object can be described by a number of *attributes*. The attribute's values are numbers or symbols which are assigned to the attribute. The predicted attribute is called the *class* and is given a special significance.

There are four main types of variables that could be used to measure the properties of the attributes:

- *Nominal variables*: This type of variable is used to put attributes into categories. If numerical values are used, then they have no mathematical interpolation. For example, we might label customers as numbers 1, 2, 3, …, 100. A binary variable is a special case of nominal variables that takes only two possible values. For example: good or bad, 0 or 1, etc.
- *Ordinal variables*: This type of variable is used to arrange attributes in a meaningful order. For example: small, medium, and large.
- *Interval variables*: For this type of variable, the difference between values is meaningful. Two examples of this type of variable are the Fahrenheit and Celsius temperature scales. For these cases, a temperature of 20 is twice as far from the zero value as 10 degrees; however, the zero value does not imply absence of temperature.
- *Ratio variables*: This type of variable is similar to interval variables; however, for this type of variable, the ratio between the variables is meaningful. An example of a ratio variable is the Kelvin temperature scale. Here, the zero point does reflect the absence of the measured temperature.

These four variables can be sorted into two categories of attributes: categorical and continuous attributes. Categorical attributes have only finite, or countable infinite, set of values, and they correspond to nominal and ordinal variables. Examples of categorical attributes are:

Student ID, zip code, geographical location. Continuous attributes have real numbers as attribute values, and they correspond to interval and ratio variables. Examples of continuous attributes are: length, price, temperature. When applying data mining techniques, it is important to choose the analysis technique based on the type of stored variables.

In general, we have two types of data mining tasks. The objective of the first type of data mining task is to predict the value of a particular, not yet been seen, attribute based on the values of other attributes. Data mining tasks of this kind are called *labeled* and are known as *predictive* (supervised) learning. The objective of the second type of data mining is to derive a pattern that summarizes the underlying relationships in data. Data mining tasks that do not have any specially designated attributes are called *unlabelled* and are known as *descriptive* (unsupervised) learning.

Statistical Parameters

A number of statistical parameters will be used in this course to summarize and characterize the datasets. A summary of these parameters is provided bellow. For categorical attributes, two main statistical parameters are usually defined: *frequency* and *mode*. Given a categorical attribute u, which can take the set of values $\{u_1, u_2, \dots, u_N\}$, the frequency is defined as:

$$Frequency(u_i) = \frac{Numbecr\ of\ objects\ with\ attribute\ value\ u_i}{N}$$

And the Mode is the u_i vale that has the heights frequency.

Similarly, given an ordinal or a continuous attribute x which can take the set of values $\{x_1, x_2, \dots, x_N\}$, we can define the p^{th} *percentile* x_p as the value of x such that $p\%$ of the observed values of x are less than x_p.

For continuous attributes, the location of the set of values $\{x_1, x_2, \dots, x_N\}$ can be measured by calculating the *mean* or *median* (the average of the middle two values) which are defined as follows:

Mean

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Median

$$Median(x) = \begin{Bmatrix} x_{r+1} & if\ N\ is\ odd \\ \frac{1}{2}(x_r + x_{r+1}) & if\ N\ is\ even \end{Bmatrix}$$

In addition to the measuring location, for continuous attributes we can also measure the dispersion or spread of a set of values $\{x_1, x_2, \dots, x_N\}$ by calculating the *range* (difference between the largest and smallest values) or *variance*. These two spread measures are defined as follows:

Range

$$Range(x) = max(x) - min(x)$$

Variance

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

What is Classification?

Classification is one of the most commonly used data mining tasks. It is a predictive data mining task, where given a set of records (training record), its goal is to build a model which assignees a class to previously unseen records (test records) as accurately as possible. For example, a bank would like to predict the probability of default for consumer loan applications based on previous records of borrowers. Table 1 shows a typical example of this record which contains car ownership, marital status, and annual income. Classification techniques will be further discussed in sessions 2 and 3.

Car Owner	Marital Status	Annual Income	Defaulted Borrower
Yes	Married	100K	No
No	Single	90K	Yes
No	Divorced	75K	No
Yes	Single	150K	Yes
…….	…….	…….	…….
No	Married	80K	Yes
Yes	Married	95K	???

Table1

What is Association?

Given a set of records, each of which contain some number of items from a given collection, association data mining tasks generate association rules which predict occurrence of a specific item based on occurrences of other items. Association is a descriptive data mining task. A common application of association is called '*market basket analysis*'. An example of this type of association is shown in Table 2 which illustrates point-of-sale data collected at the checkout counter of a grocery store. Here, the association rules can be applied to discover items that are frequently bought together and plan future market campaigns. Association techniques will be further discussed in sessions 4 and 5.

Transaction	Items
1	Milk, Water, Newspaper
2	Milk, Newspaper
3	Water, Tea
4	Milk, Newspaper, Tea, Coffee
…….	…….
100	Milk, Newspaper, Tea

Rules Discovered
If {Milk} → {Newspaper}
If { Milk, Newspaper } → { Tea }

Table 2

What is Clustering?

Give a set of records, each having a set of attributes, clustering techniques find groups (clusters) of items such that the data points in one cluster are more similar to one another, and data points in separate clusters are less similar to one another. For example, a hospital might group patients based on blood pressure, age, or geographical location (Figure 1). Clustering techniques are descriptive data mining tasks, and will be further discussed in sessions 6 and 7.

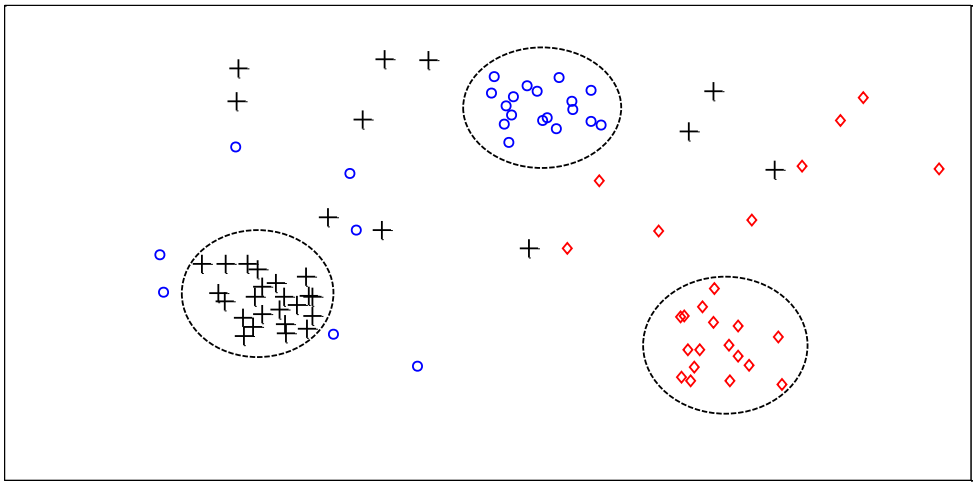

Figure 1

References and Further Readings

- Tan, P., Steinbach, M., & Kumar, V. (2005). "Introduction to data mining."
- Han, J., & Kamber M., (2006). "Data Mining: Concepts and Techniques."
- Bramer M. (2007). "Principles of Data Mining."