Department of Computing

Data Mining & Data Warehousing

Group Project (3-5 members) – confirm your group on or before *October 3, 2015*

Due: 23:00 Sunday, **November 22, 2015**

Project Objectives:

This project is aimed for students to go through the complete data analytics process (*data mining for business analysis*) using some real datasets. You can use either existing tools to do a case study (*as indicated by Task A below*) or create your own programs for the same purpose (*Task B below*).

Project Description:

In this group project, students can choose one of the following suggested datasets:

i) Weblog Data,

ii) Stock Data,

iii) Data set from UCI Machine Learning Repository, or

iv) Data set from an online (official) data storage site, e.g. DATA.GOV.HK

You can also choose one of the two tasks. In Task A, you can do a case study based on your choice of dataset. In Task B, you can write programs to do data mining yourself.

Tasks:

A) *Case study* – Assuming you are a group of data analysts and need to find some interesting rules to the company which produced this data. You are expected to undertake all the processes of the KDD (*Knowledge Discovery in Databases*) model to extract interesting patterns (*rules*) from the data sets. And then, you are expected to make business related suggestion(s) from your findings, i.e. how to make use of the pattern(s) / interesting rule(s) found from the data set.

Since the data mining process is an iterative one, and you are expected to go through this process a number of times with different methods and settings to eventually find something interesting (if any), it is important for you to record the methods you used in each iteration and the settings you used. Therefore, the final report for this project should also present the process you have gone through in the project and the rationale for certain settings. For example, if you have used a particular data pre-processing technique, you should explain why you choose that technique(in other words, how is it relevant to the specific data you are using); and/or why attributes used are related to the data mining model; and/or why the data mining model is being used is the best one; etc.

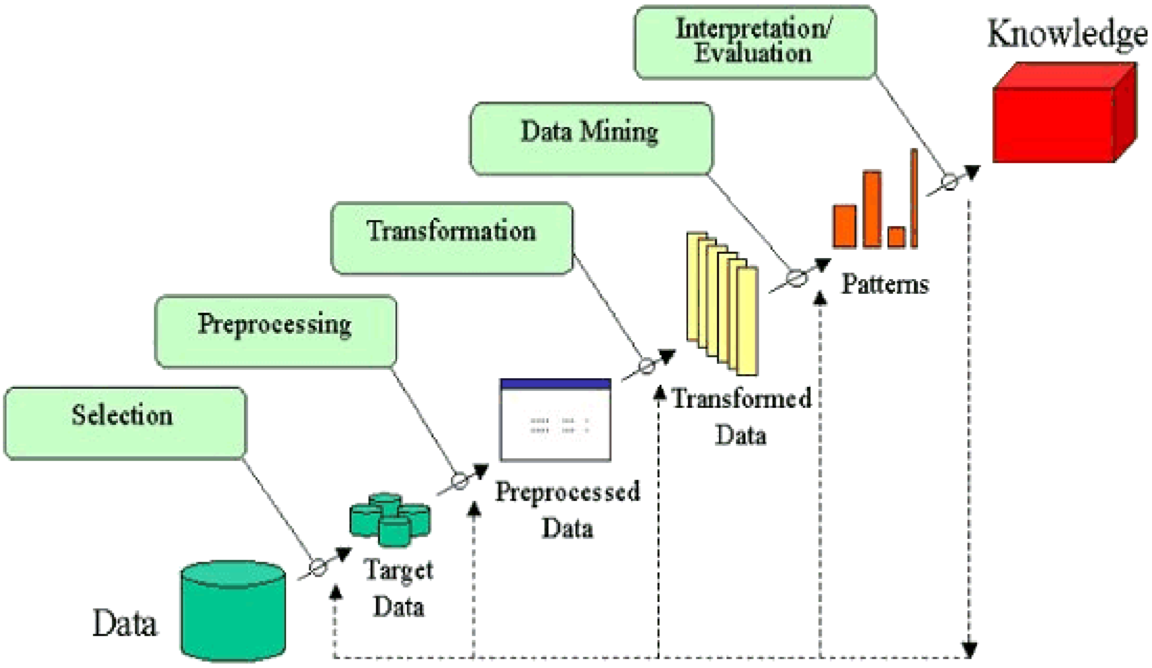

Figure 1 – Processes of Knowledge Discovery in Database

B) *Programming* – develop a data mining application to find out interesting patterns (rules) from the given data sets. You can use a programming language which you are familiar with (for example, JAVA, C++, Python, etc). Although your main focus is on the development of an application, you are also required to undertake the different steps in the process of the KDD relevant to your choice of data. In particular, you should express your understanding of data mining algorithms/principles and to implement those essential data mining techniques for extracting interesting patterns (rules) from your chosen dataset. Also, the application should provide functions to help users to examine the generated model. In general, the correctness (accuracy) of the model can be measured.

Datasets:

A) "Microsoft Anonymous Web Data" – publicly available from http://kdd.ics.uci.edu/databases/msweb/msweb.html

B) Stock data – There are numbers of links/websites that you can use to download the historical prices, for example Hang Seng Index from Yahoo.

C) Others. If you want to use some other data sets, check the link (UCI Machine Learning Repository) http://archive.ics.uci.edu/ml/ and select one for which the data domain is of interest to you.

D) Or, you can use the dataset provided by your company or found from government organizations, or others, for example DATA.GOV.HK. You can consult Alvin, if you are not sure your dataset is reasonable for project use.

Presentation / Demonstration

1) If you choose to do Task A, an 8 to 15 minutes presentation is required.

2) If you choose to do Task B, a 5 to 10 minutes demonstration of the application will be arranged.

3) Students need to attend not only their own group's presentation, but also, some other groups' presentation/demo. The presentation/demonstration attendance will be arranged randomly. Schedule of presentation/demonstration will be made available once all the groups are formed and your choice of tasks are confirmed.

Submission Requirement

1) For Task A, a detail project report (around 3000 words) in Word format is needed. And, a presentation file (either hard copy or soft copy) is also required on or before the presentation.

2) For Task B, a detail project report (around 2000 words) and a user manual in Word format are required. In addition, the source code file(s) should be submitted on or before the demonstration.

Assessment

For Task A:

Details of the KDD being processed	40%
Innovative Ideas Introduced	10%
Presentation	20%
Report and Analysis	30%

For Task B:

Correctness of Development	40%
Innovative Ideas Introduced	10%
Demonstration	20%
Report and Analysis	30%