**EEEE 380 L-10: Data Mining**

**Lab Assignment # 1**

There are two parts to this assignment. The first part is a fun search activity, where you are asked to determine the top 10 data mining inventions/applications. In the second part, you get hands-on experience with data pre-processing.

**Part I – Top 10 Data Mining Applications/Inventions**

For this part of the assignment, you are asked to do a research on the Internet, and determine what you believe are the top 10 data mining inventions or applications. You will need to include your justification for your choice. Additional information needs to be included so that we can later combine all students’ votes. Use the provided template “top 10 template.xls” to fill the necessary information.

**Deliverable:** Updated excel sheet template with your information. Name the file with your name: Last name First Name.xls

**Part II - Pre-processing of Streaming Data from a Desktop Activity Monitor**

1. **Introduction**

Assume you have a Desktop Activity Monitor as an application tool that runs in the background of your desktop and collects information about the user’s activity every second. Some of the features collected are: the application currently in use, the opened applications, the mouse and keyboard usage, etc… The data pulled on a second by second basis is logged to an excel file. This experiment aims at developing a software tool that will automate the pre-processing of the data according to some given criteria.

1. **Objectives**

* Practice and get experience with Data Pre-processing, which will include: Integration (reading from Excel), Feature Engineering by conversion from original units into more meaningful units and identifying useful features for data mining.

1. **Materials and Data**

* You are given the dataset entitled appmon\_1.xls
* You are free to use any programming language: Matlab, c#, c++, Java,..

1. **Details about input data**

Table 1 lists the names of the all the columns in the input excel file and explains the meaning of each.

Table 1 – Features Collected in appmon\_1.xls

|  |  |
| --- | --- |
| Columns | Details |
| mSec from start | Millisecond since the start of the tool |
| last\_input\_ms | Millisecond that passed since last input from user |
| focus\_app\_name | Name of the application in focus |
| focus\_app\_title | Window Title of the application in focus |
| cpu\_intensive\_processes | Most CPU consuming process |
| pc\_locked\_or\_screensaver | if PC is locked or not (0 if not locked, 1 otherwise) |
| opened\_windows | Set of opened windows, delimited by '|' |
| mouseclicks | Number of mouse clicks since last measurement |
| keystrokes | Number of Keystrokes since last measurement |
| mousemoves | Number of mouse moves since last measurement |
| mousewheelmoves | Number of mouse wheel moves since last measurement |

* 1. **Data pre-processing**

Develop a software tool using a programming language of your choice that automates the following data pre-processing steps. Given the input data, the program should be able read the input excel file and then for every row:

**Perform data clean up:**

1. Set *mouseclicks[i]* to 0 if it is an outlier (Value > 200)
2. Set *keystrokes*[i] to 0 if it is an outlier (Value > 100)

**Perform Data Transformation and feature extraction:**

1. Extract the actual time at row i knowing that the experiment started at 08:07:50 AM and *mSec from start[i]* reflects the number of milliseconds since that time.
2. Extract a Window Switch feature that is set to 1 whenever *focus\_app\_name[i]!= focus\_app\_name[i-1] and focus\_app\_title[i]!= focus\_app\_title[i-1]*, and that is set to 0 otherwise.
3. Extract the number of opened windows, by parsing *opened\_windows[i]*

**Perform Data reduction and feature extraction:**

1. Extract a (discretization) categorization of *mousemoves[i]* according to the following categories:
   1. No Move, if value is equal to 0
   2. Slow, if value is greater than 0 and less than 36
   3. Moderate, if value is greater or equal to 36 and less than 55
   4. Fast, if value is greater or equal to 55
2. Unused features from the original set shall be deleted.

**Save the pre-processed data:**

1. Your program should save the new data in appmon\_1\_out.xls.
2. **Deliverables**

In order to accomplish this lab experiment, you will have to deliver:

1. An excel file with the preprocessed data
2. A well-documented code