CS795/895: Topics in Data Mining and Security
Summer 2021
Course Project
Due: August 1, 2021
Profiling for authentication and authorization
Teams: 1 or 2 students per team

Objective: To apply data mining ideas in a class project

You are provided with the historical login and access data for all 20 users in a department. Based on this data, you are asked to develop a profile for each user. The profile would indicate the login, logoff, and **session time patterns** as well as the **access patterns**. The access pattern consists of statistics regarding---user programs executed, library programs/utilities executed, files accessed for read and update, files created and their size, and printer usage. The statistics for each may consist of start time, duration, resources accessed and the type of operations performed. Here, resources refer to computers (machines), files, network, and printer. These are only example statistics. Use your creativity to come up with other statistics of relevance to express resource access patterns. In addition to the user profiles, determine any association rules with sufficient support.

The profile that you develop is not just a summary of the given pattern. Instead, it should take into account all aspects of modeling we talked about such as avoiding overfits and outliers. In addition, you should use any correlated user patterns and be able to merge user patterns when they are close enough. For example, if you observed that one user accesses files 1-15 and the other user uses files 3-18, they are close enough in terms of file accesses. So you should probably profile them both as accessing file 1-18. Of course, you need to have a justification to do such extensions. It is important to derive additional information from the provided data. For example, given a date you may be able to derive the day and use it in developing the profile. In other words, the profile may consist of given data as well as derived data. For example, the day of the week may be more relevant than the date itself. Similarly, weekend versus weekday may be more relevant than the day itself.

You should clearly document your approach, analysis, assumptions, and final results in your report. It should clearly indicate how the data mining security techniques discussed in the class have been employed to arrive at your answers. There are no restrictions on the type of tools that you can use. But clearly document them in your report.

Finally, you should submit a well structured document that includes all the above details including the final results and conclusions. Your project will be graded on how well you have employed several techniques and how well you were able to analyze the problem to arrive at a fairly generalized modeling of the users.

Following is a sample data that will be provided to you:

User ID	U01-U20
User program ID	UP001-UP500
Library program/utility ID	LP001-LP100
File ID	F0001-F2000
Printer ID	PR1-PR6
E-mail program ID	E1-E5
Host machines	M01-M30
Date is expressed as	MMDDYY
Time is expressed as	HHMMSS

Example login pattern:
1 U01 M01 090508 080010 170040 22 70 12345 12098
(Type 1 record; User U01; Machine M01; Date: 09/05/2008; Login time: 08:00:10; logout time: 17:00:40; Average number of user processes at any time: 22; Maximum number of user processes: 70; Total keyboard characters typed: 12345; CPU use (in seconds) by user processes: 12098;

Example Resource usage pattern:
2 U01 M01 090508 090015 UP007 000230 F0011 R F0017 RW F1800 W PR2 23
(Type 2 record; User U01; Machine M01; Date: 09/05/2008; Start time: 09:00:11; Program: UP007; Execution time: 00:02:30; File: F0011 – R – Read; File: F0017 RW (Read write); F1800 W (write); Printer: PR2 — 23 pages printed)

Example e-mail pattern:
3 U01 M01 090508 100230 E1 jim91@yahoo.com R 10450 0
(Type 3 record; User U01; Machine M01; Date: 09/05/2008; Start time: 10:02:30; E-mail Program: E1; E-mail address: jim91@yahoo.com ; Received (R); Bytes: 10450; Attachments: 0)

3 U01 M01 090508 110430 E2 jane45@gmail.com R 100520 1
(Type 3 record; User U01; Machine M01; Date: 09/05/2008; Start time: 11:04:30; E-mail Program: E2; E-mail address: jane45@gmail.com ; Received (R); Bytes: 100520; Attachments: 1)

3 U01 M01 090508 120630 E2 janet99@gmail.com S 203420 2
(Type 3 record; User U01; Machine M01; Date: 09/05/2008; Start time: 12:06:30; E-mail Program: E2; E-mail address: janet99@gmail.com ; Sent (S); Bytes: 203420; Attachments: 2)

The input file is a text file with 1 record per line:

U01 M01 090508 080010 170040 22 70 12345 12098
U01 M01 090508 090015 UP007 000230 F0011 R F0017 RW F1800 W PR2 23

What to look for? (i) login pattern (ii) program access pattern (iii) file access pattern (iv) printer usage pattern (iv) E-mail pattern (v) Machine usage pattern

By pattern, we mean anything that you think is meaningful and for which data can be collected, statistics computed, and later used to describe the profile of the user.

Definitely, provide a pattern for each user; also try to cluster users together based on some profile characteristics. For example, if two users frequently send/receive e-mail to jane45@gmail.com, then that could be one basis. Make sure to mention the basis on which they are grouped.

Use your imagination and creativity for profiling. Simply collecting statistics for a user is not sufficient. Try to see what is common and develop different groups of users based on some commonalities. Of course, the same user may belong to different groups based on a different characteristics.