Profiling for Authentication and Authorization

CS 773-TOPICS IN DATA MINING & SECURITY, COURSE PROJECT

Mounika Kompalli (UIN: 01014100)
OLD DOMINION UNIVERSITY

ABSTRACT

Data Mining is the efficient discovery of previously unknown, valid, potentially useful, understandable patterns in large datasets. It is an extension of traditional data analysis and statistical approaches. The objective of data mining is to identify valid novel, potentially useful, and understandable correlations and patterns in existing data. A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm.

This project involved analyzing data and deriving useful user patterns based on the given user information. A dataset was given which contained user login and access data for all 19 users in a Department. A profile was developed for each user based on login pattern, session time patterns and access patterns.

User profiles were carefully analyzed to derive some useful statistics that helped in user modelling. Association Rules along with support were determined. Different aspects of modelling like over fitting and outliers were taken into consideration.

Additional information was derived from the given dataset which helped in identifying different user patterns. The assumptions and analysis of data was done considering different data mining security techniques. Users were grouped together based on similar profile characteristics and conclusions are derived to provide a pattern for each user based on commonalities. Clustering technique was implemented on the set of data (or objects) to convert it into a set of meaningful sub-classes, called clusters.

Finally suitable conclusions are derived about the user data using the data mining techniques.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ACKNOWLEDGEMENT

We would like to take this opportunity to thank Dr. Ravi Mukkamala in providing a helping hand in this project. His valuable guidance, support and supervision all through the course Data mining and Security, are responsible for attaining this project in its present form.

The References and the video links provided in the course material under each module were immensely helpful to understand and apply various data mining techniques.

Name

Mounika Kompalli

Course

CS 773 - Data mining and Security

INTRODUCTION

Login and access data for 19 users belonging to a department are given. A profile has to be developed for each user describing the different statistics with respect to the programs executed, files accessed for read and update, files created and their size, library programs/ utilities executed and printer usage. The statistics for each profile may consist of start time, duration, resources (computers, files, network, and printer) accessed and the type of operations performed.

GIVEN DATA

We are provided with three different types of data which helps in understanding the user's login, resource access and email usage pattern. The first column in the given user data set defines the type of the record.

Type 1 Records – User Login Related data.

The different attributes of the Type 1 Records are:

- Record Type
- User
- Machine
- Date
- Login time
- logout time
- Average number of user processes at any time
- Maximum number of user processes
- Total keyboard characters typed
- CPU use (in seconds) by user processes.

Type 2 Records – User Resource Access related data.

The different attributes of the Type 2 Records are:

- Record Type
- User
- Machine
- Date
- Start time
- Program
- Execution time
- File R – Read/File RW (Read write)/File W (write);
- Printer
- Pages printed.

Type 3 Records – User Email Related data

The different attributes of the Type 3 Records are:

- Record Type
- User
- Machine
- Date
- Start time
- E-mail Program
- E-mail address
- Received (R)/Sent(S)
- Bytes
- Attachments.

ANOMALIES IN THE GIVEN DATA

1) A user can only access the resources after he logs in to the system but for few users resource usage and email usage data is mentioned without the login details. E.g.: User U15
The below picture helps in identifying the anomaly.

Count of c	Column Labels									
Row Labels	M09	M10	M11	M12	M13	M14	M15	M16	M18	Grand Total
1	22	1	1	1	1	5		22	8	61
U09	22									22
U10		1								1
U11			1							1
U12				1						1
U13					1					1
U14						5				5
U15 ?										
U16								22		22
U18									8	8
2	19	3	3	3	3	7	3	13	10	64
U09	19									19
U10		3								3
U11			3							3
U12				3						3
U13					3					3
U14						7				7
U15							3			3
U16								13		13
U18									10	10
3	11		2	2	2	6	2	11	11	47
U09	11									11
U11			2							2
U12				2						2
U13					2					2
U14						6				6
U15							2			2
U16								11		11
U18									11	11
Grand Total	52	4	6	6	6	18	5	46	29	172

Figure 1

2) User U10 has no login information available for M21, M23, M25 but resource and Email usages are mentioned in the data set.

Record Type / Row Labels	Column Labels / M10	M11	M21	M22	M23	M24	M25	Grand Total
1	**1**	**1**	**13**	**7**	**10**	**7**	**21**	**60**
U10	1			1		1		3
U11		1	3	2	2	2	3	13
U12			3	1	2	1	5	12
U13			2	1	2	1	6	12
U14			2	1	2	1	2	8
U15			3	1	2	1	5	12
2	**3**	**3**	**10**	**16**	**2**	**3**	**42**	**79**
U10	3		1		2			6
U11		3		4		3	6	16
U12				3			10	13
U13			9	3			10	22
U14				3			6	9
U15				3			10	13
3		**2**		**4**		**1**	**23**	**30**
U10						1	1	2
U11		2		2			4	8
U12				1			5	6
U13							6	6
U14							2	2
U15				1			5	6
Grand Total	**4**	**6**	**23**	**27**	**12**	**11**	**86**	**169**

Figure 2

DERIVATION OF ADDITIONAL INFORMATION

From the given user data, additional information is derived to analyze and study user patterns. The Type 1 records are studied and carefully analyzed using MS Excel.
Pivot tables are used to filter the attributes and to generate separate profiles for each user.
The type 1 Records have login information and the following additional attributes are derived which helped us to analyze data and study user patterns.
The type 2 Records have Resource Usage Details like the Printer Access, File access and the Program Access related details.
The type 3 Records have the Email Usage details.

Derivation of Additional Information for Type 1 Records to study the login pattern:

1) No. of days a user logged in

This attribute helps us to know the number of days each user logged in for the given month. Using MS Excel filters the number of days is calculated making use of the Date column from the given Type 1 user data.

This is a numerical attribute.

2) Login periods

The Login periods attribute is derived from the login time column of type 1 user dataset. This attribute helps us to figure out all those users who are logging in the same time interval.

A numeric value is assigned for every hour to denote the login periods.

Time Range	Numeric value
00:00:00-00:59:59	1
1:00:00-1:59:59	2
2:00:00-2:59:59	3
3:00:00-3:59:59	4
4:00:00-4:59:59	5
5:00:00-5:59:59	6
6:00:00-6:59:59	7
7:00:00-7:59:59	8
8:00:00-8:59:59	9
9:00:00-9:59:59	10
10:00:00-10:59:59	11
11:00:00-11:59:59	12
12:00:00-12:59:59	13
13:00:00-13:59:59	14
14:00:00-14:59:59	15
15:00:00-15:59:59	16
16:00:00-16:59:59	17
17:00:00-17:59:59	18
18:00:00-18:59:59	19
19:00:00-19:59:59	20
20:00:00-20:59:59	21
21:00:00-21:59:59	22
22:00:00-22:59:59	23
23:00:00-23:59:59	24

Table 1

3) Logout periods

The Logout periods attribute is derived from the logout time column of type 1 user dataset. This attribute helps us to figure out all those users who are logging out the same time interval.

A numeric value is assigned for every hour to denote the logout periods.

Refer to Fig 3 for the numeric values.

This is a numerical attribute.

4) No. of days a user logged in during weekdays.

This attribute is derived from the date column of the given type 1 user data. Pivot tables and filters were used to separate the users who logged in during weekdays.

The number of weekdays each user has logged in is displayed in this attribute.

As the given type 1 user data belongs to September month of the year 2008, all the weekdays of this month are filtered to find out the number of days an individual user logged in to the respective machines.

Weekdays: Monday- Friday

Dates: Sep 1-Sep 5, Sep 8-Sep12, Sep 15- Sep 19, Sep 22-26, Sep 29-Sep 30.

This attribute helps an employer to keep track of the employees who are logging in during weekdays without fail.

This is a numerical attribute.

5) No. of days a user logged in during weekends

This attribute is derived from the date column of the given type 1 user data. Pivot tables and filters were used to separate the users who logged in during weekends

The number of weekends each user has logged in is displayed in this attribute.

As the given type 1 user data belongs to September month of the year 2008, all the weekends of this month are filtered to find out the number of days an individual user logged in to the respective machines

Weekends: Saturday, Sunday

Dates: Sep 6, Sep 7, Sep13, Sep 14, Sep 20, Sep 21, Sep27, Sep28.

This attribute helps an employer to keep track of the employees who are working on weekends to finish their duties.

This is a numerical attribute.

6) No. of times a user logging in during working hours

Considering the fact that whenever a user takes a break during working hours he logs out and logs back into the machine, this particular attributes helps an employer to keep track of the number of breaks each employee is taking during working hours.

This attribute is also used to find out those users who are not logging out of their machines even during breaks or after the working hours. This helps an organization to reduce the power consumption and CPU usage.

This attribute is derived from the date column of the given type 1 user data. Pivot tables and

filters were used to separate the users who logged in multiple times during working hours. Considered Working hours: 09:00:00 to 17:59:59.
This is a numerical attribute.

7) No. of times a user logging in after/before working hours.

This attribute is used to keep track of those users who are logging in before or after the working hours.
This attribute is derived from the date column of the given type 1 user data. Pivot tables and filters were used to separate the users who logged in multiple times during working hours. After/Before working hours: 18:00:00 to 08:59:59.
This is a numerical attribute.

8) No. of times a user logging out during working hours

Considering the fact that whenever a user takes a break during working hours he logs out and logs back into the machine, this particular attributes helps an employer to keep track of the number of breaks each employee is taking during working hours.
This attribute is also used to find out those users who are not logging out of their machines even during breaks or after the working hours. This helps an organization to reduce the power consumption and CPU usage.
This attribute is derived from the date column of the given type 1 user data. Pivot tables and filters were used to separate the users who logged out multiple times during working hours. Considered Working hours: 09:00:00 to 17:59:59.
This is a numerical attribute.

9) No. of times a user logging out before/after working hours

This attribute is used to keep track of those users who are logging out before or after the working hours.
This attribute is derived from the date column of the given type 1 user data. Pivot tables and filters were used to separate the users who logged out multiple times during working hours. After/Before working hours: 18:00:00 to 08:59:59.
This is a numerical attribute.

DERIVATION OF ADDITIONAL INFORMATION TO STUDY THE CPU AND KEYBOARD USAGE (TYPE 1 RECORDS)

1) No. of Keyboard characters typed

This attribute displays the number of keyboard characters typed by each user.
This is a numerical attribute and is used by the employer to keep track of the keyboard usage of each employee.

2) Total keyboard characters typed during weekdays

This attribute displays the number of keyboard characters that are typed during weekdays by each user. This is a numerical attribute

3) Total keyboard characters typed during weekend

This attribute displays the number of keyboard characters that are typed during weekends by each user.
This is a numerical attribute

4) Total CPU use by user processes

This attribute displays the total CPU use by the user processes in seconds.
This is a numerical attribute that denotes the CPU usage with respect to time.

5) Total CPU use during weekdays

This attribute displays the total CPU use by the user processes in seconds during weekdays
This is a numerical attribute that denotes the CPU usage with respect to time.

6) Total CPU use during weekend

This attribute displays the total CPU use by the user processes in seconds during weekends.
This is a numerical attribute that denotes the CPU usage with respect to time.

DERIVATION OF ADDITIONAL INFORMATION FROM TYPE 3 USER DATA TO STUDY THE EMAIL USAGE PATTERN

1) Total No. of emails sent or Received

This attribute helps us to figure out the total number of emails either sent or received by each user.
MS Excel filters were used to separate the total number of emails sent.
This is a numerical attribute.

2) Email Id from which Max No. of emails sent or received

This attribute helps us to figure out the email id from which each user sent or received maximum number of emails.
MS Excel filters were used to separate the total number of emails sent.
This is a numerical attribute.

3) Total size of emails sent/received in Kilo bytes

This attribute helps us to figure out the total size of emails sent or received by each individual user.

This attribute is expressed in terms of Kilo bytes and is a numerical attribute.

This can be used by an employer to figure out the size of data that is exchanged between employees in the form of emails.

4) Total No. of Attachments sent or received

This attribute displays the number of attachments sent or received by each user.

This is a numerical attribute and can be used by the employer to find out the number of attachments that are frequently exchanged between his employees.

5) Emails received or sent during working hours

This attribute displays the number of emails sent or received during working hours. This is a numerical attribute and can be used by the employer to find out the number of emails that are being sent or received by each employee especially during working hours.

Working hours: 9:00:00 to 17:59:59

6) Emails received or sent after/ before working hours

This attribute displays the number of emails sent or received after working hours. This is a numerical attribute and can be used by the employer to find out the number of emails that are being sent or received by each employee after working hours and separating those employees who are working after working hours to complete the assigned tasks.

After /before Working hours: 18:00:00-08:59:59

7) Total No. of emails sent or received during weekdays

This attribute displays the number of emails sent or received during weekdays. This is a numerical attribute and can be used by the employer to find out the number of emails that are being sent or received by each employee during weekdays and to keep track of all those employees who are working and who are not during weekdays.

8) Total No. of emails sent or received during weekends

This attribute displays the number of emails sent or received during weekends. This is a numerical attribute and can be used by the employer to find out the number of emails that are being sent or received by each employee during weekends and to keep track of all those employees who are working even during weekends to finish their tasks.

9) Sent/Received to/from mom@icare.com

This attribute displays the information whether a particular user has sent any emails to the id mom@icare.com.

mom@icare.com is identified to be a personal email and this attribute can be used by the employer to keep track of those employees that are sending personal emails during office hours.

This is a Boolean attribute.

DERIVATION OF ADDITIONAL INFORMATION FROM TYPE 2 USER DATA TO STUDY THE RESOURCE USAGE

Some additional attributes were derived to study the resource usage pattern of users.

They are listed below:

Attributes to study the program access pattern.

1) Total No. of programs executed

This attribute displays the total number of programs that are executed by individual user.

This attribute is a numerical attribute and is used to highlight the workload that is shared between different users.

2) Types of programs executed

From the given type 3 user dataset different types of programs executed by the users are identified and listed under this attribute.

Different type of programs executed:

UP: User Program

LP: Library Program

3) Programs executed on weekdays

This attribute displays the number of programs that are executed during weekdays.

This is a numerical attribute and is used by the employer to keep track of the program execution in terms of quantity.

4) Programs executed during working hours

This attribute displays the number of programs that are executed during working hours.

This attribute is a numerical attribute and helps the employer to keep track of those employees who are working on programs during working hours.

5) Programs executed after working hours

This attribute displays the number of programs that are executed after working hours. This attribute is a numerical attribute and helps the employer to keep track of those employees who are working on programs after working hours.

6) Programs executed on weekends

This attribute displays the number of programs that are executed during weekdays.
This is a numerical attribute and is used by the employer to keep track of the program execution in terms of quantity.

7) Total Execution time

This attribute displays the total execution time for each user.

This is a numerical attribute and is used by the employer to keep track of those employees who are utilizing the working hours in executing the programs.

8) Type of program executed Max no. of times

This attribute displays the type of programs that are executed maximum number of times.

This is a numerical attribute and is used to figure out the type of programs on which the users spend maximum time.

Type of programs: LP / UP

Attributes to study the printer usage pattern

1) No. of pages printed

This attribute displays the number of pages printed by each user.

This is a numerical attribute and is used by the employer to keep track of those employees who are printing maximum number of pages.

2) No. of pages printed on weekdays

This attribute displays the number of printers that are used during weekdays by each user.

This is a numerical attribute and is used by the employer to keep track of the number of pages that are printed on weekdays.

3) No. of pages printed on weekend

This attribute displays the number of pages that are printed on weekends.

This is a numerical attribute and is used by the employer to keep track of the number of pages that will be needed by the organization during weekends.

4) No. of pages printed during working hours

This attribute displays the number of pages that are printed during working hours by each user.

This is a numerical attribute and is used by the employer to keep track of the number of pages that are needed in the organization during working hours.

5) No. of pages printed after working hours

This attribute displays the number of pages that are printed after working hours by each user.

This is a numerical attribute and is used by the employer to keep track of the number of pages that are needed in the organization after working hours.

Additional Attributes to study the file access pattern.

1) No. of files opened with Read access

This attribute displays the total number of files that are opened with Read permission by each user.

This is a numerical attribute and is used by the employer to keep track of those employees who are having the access to Read the files in the organization.

2) No. of files opened with Read-write access

This attribute displays the total number of files that are opened with Read permission by each user.

This is a numerical attribute and is used by the employer to keep track of those employees who are having the access to Read and make changes to the files in the organization.

3) Total no. of files accessed

This attributes displays the total number of files that are accessed by each user.

This is a numerical attribute and is the sum of attributes No. of files opened with Read access and No. of files opened with Read-write access.

4) Total no. of files opened during working hours

This attribute displays the total number of files that are opened during working hours.

This is a numerical attribute and is used by the employer to keep track of the number of files that are being accessed during working hours.

5) Total no. of files opened after working hours

This attribute displays the total number of files that are opened after working hours.

This is a numerical attribute and is used by the employer to keep track of the number of files that are being accessed after working hours.

6) Total no. of files accessed on weekdays

This attribute displays the total number of files that are opened during weekdays.

This is a numerical attribute and is used by the employer to keep track of the number of files that are being accessed by the employees during weekdays.

7) Total no. of files accessed on weekend

This attribute displays the total number of files that are opened on weekends.

This is a numerical attribute and is used by the employer to keep track of the number of files that are being accessed by the employees on weekends.

DERIVATION OF ADDITIONAL ATTRIBUTES TO STUDY THE MACHINE USAGE PATTERN

1) Shared ?

This attribute helps us to identify whether the machine is shared between two users or used by a single user.

This is a Boolean attribute and the possible values are Yes and No.

This attribute can be used by the employer to keep track of those machines that are used by multiple users.

2) Total time Used (In mins)

This attribute displays the total time each machine is used by the users.

This is a numerical attribute that displays the time in minutes and is used by the employer to keep track of those machines that are used for large intervals of time.

3) Frequently used by

This attribute displays the User Id of the user who uses the machine frequently.

There are some systems that are exclusively used by a single user and some systems that are used by multiple users.

This is a nominal attribute and is used to identify those users that use a particular system frequently.

4) Machine used during weekdays

This attribute helps us to identify those machines that are used during weekdays.

This is a Boolean attribute and the possible values are Yes or No.

This attribute helps the employer to be informed about those machines that are used during weekdays.

5) Machine used during weekends

This attribute helps us to identify those machines that are used during weekends.

This is a Boolean attribute and the possible values are Yes or No.

This attribute helps the employer to be informed about those machines that are used during weekends.

USER PROFILING

USER PROFILING BASED ON LOGIN PATTERNS

User	No. of days a user logged in	Login_period	Logout_period	No. of days a user logged in during weekdays	No. of days a user logged in during weekends	No. of times a user logging in during working hours	No. of times a user logging in before/after working hours	No. of times a user logging out during working hours	No. of times a user logging out before/after working hours
U01	22	9	18	22	0	0	22	22	0
U02	22	More than once	More than once	22	0	7	15	12	10
U03	22	9	18	22	0	0	22	22	0
U04	22	More than once	More than once	22	0	6	16	12	10
U05	22	9	18	22	0	0	22	22	0
U06	22	More than once	More than once	22	0	6	16	12	10
U07	22	9	18	22	0	0	22	22	0
U08	22	More than once	More than once	22	0	6	16	12	10
U09	22	9	18	22	0	0	22	22	0
U10	22	More than once	More than once	22	0	6	16	12	10
U11	22	9	18	22	0	0	22	22	0
U12	22	More than once	More than once	22	0	8	14	20	2
U13	22	More than once	More than once	22	0	8	14	20	2
U14	22	More than once	More than once	22	0	8	14	20	2
U15	21	More than once	More than once	21	0	8	13	20	1
U16	22	More than once	More than once	22	0	8	14	20	2
U17	22	More than once	More than once	22	0	8	14	20	2
U18	8	More than once	More than once	1	7	1	7	7	1
U19	8	More than once	More than once	1	7	1	7	7	1

Table 2

USER PROFILING BASED ON CPU USAGE AND KEYBOARD USAGE

User	Machine	No. of Keyboard characters typed	Total keyboard characters typed during weekdays	Total keyboard characters typed during weekend	Total CPU use by user processes	Total CPU use during weekdays	Total CPU use during weekend
U01	M01	246995	246995	0	241844	241844	0
U02	M02	246995	246995	0	241844	241844	0
U03	M03	246995	246995	0	241844	241844	0
U04	M04	246995	246995	0	241844	241844	0
U05	M05	246995	246995	0	241844	241844	0
U06	M06	246995	246995	0	241844	241844	0
U07	M07	246995	246995	0	241844	241844	0
U08	M08	246995	246995	0	241844	241844	0
U09	M09	246995	246995	0	241844	241844	0
U10	Multiple	246995	246995	0	241844	241844	0
U11	Multiple	246995	246995	0	241844	241844	0
U12	Multiple	246995	246995	0	241844	241844	0
U13	Multiple	246995	246995	0	241844	241844	0
U14	Multiple	246995	246995	0	241844	241844	0
U15	Multiple	234650	234650	0	229746	229746	0
U16	M16	246995	246995	0	241844	241844	0
U17	M19	246995	246995	0	241844	241844	0
U18	M18	87820	12345	75475	84650	13400	71250
U19	M19	87820	12345	75475	84650	13400	71250

Table 3

USER PROFILING BASED ON MACHINE USAGE

Machine	Shared ?	Total time Used (minutes)	Frequently used by	Machine used during weekdays	Machine used during weekends	Total number of keyboard characters typed	Total CPU usage by user processes
M01	No	11965	U01	Yes	No	246995	241844
M02	No	10418	U02	Yes	No	246995	241844
M03	No	11965	U03	Yes	No	246995	241844
M04	No	10343	U04	Yes	No	246995	241844
M05	No	11952	U05	Yes	No	246995	241844
M06	No	10343	U06	Yes	No	246995	241844
M07	No	11965	U07	Yes	No	246995	241844
M08	Yes	13134	Multiple Users	Yes	No	316250	309310
M09	No	11955	U09	Yes	No	246995	241844

M10	No	601	U10	Yes	No	12345	12098
M11	No	540	U11	Yes	No	12345	12098
M12	No	240	U12	Yes	No	12345	12098
M13	No	240	U13	Yes	No	12345	12098
M14	No	2820	U14	Yes	No	42565	40458
M16	No	9800	U16	Yes	No	246995	241844
M18	No	4162	U18	Yes	Yes	87820	84650
M19	Yes	13962	Multiple Users	Yes	Yes	334815	326494
M21	Yes	7410	Multiple Users	Yes	No	120360	110844
M22	Yes	3854	Multiple Users	Yes	No	93940	82700
M23	Yes	4860	Multiple Users	Yes	No	149000	146660
M24	Yes	1744	Multiple Users	Yes	No	80390	85586
M25	Yes	12620	Multiple Users	Yes	No	230925	228112
M26	Yes	5417	Multiple Users	Yes	No	166905	172976
M27	Yes	3333	Multiple Users	Yes	No	49135	51116
M28	Yes	6675	Multiple Users	Yes	No	171215	150490
M29	Yes	1945	Multiple Users	Yes	No	105940	104756
M30	Yes	6178	Multiple Users	Yes	No	140615	149410

Table 4

USER PROFILING BASED ON RESOURCE USAGE

PROFILING BASED ON PRINTER USAGE

User	Machine	Printer used	No. of pages printed	No. of pages printed on weekday	No. of pages printed on weekend	No. of pages printed during working hours	No. of pages printed after working hours
U01	M01	PR1	142	142	0	122	20
U02	M02	PR1	160	160	0	140	20
U03	M03	PR1	142	142	0	122	20
U04	M04	PR1	160	160	0	140	20
U05	M05	PR2	142	142	0	122	20
U06	M06	PR2	160	160	0	140	20

U07	M07	PR2	142	142	0	122	20
U08	M08	PR2	160	160	0	140	20
U09	M09	PR2	142	142	0	122	20
U10	Multiple	PR2	160	160	0	140	20
U11	Multiple	PR3	142	142	0	122	20
U12	Multiple	PR3	142	142	0	95	47
U13	Multiple	PR4	202	202	0	155	47
U14	Multiple	PR4	142	142	0	95	47
U15	Multiple	PR4	142	142	0	95	47
U16	M16	PR4	92	92	0	92	0
U17	M19	PR6	92	92	0	92	0
U18	M18	PR5	56	0	56	56	0
U19	M19	PR6	56	0	56	56	0

Table 5

PROFILING BASED ON PROGRAM ACCESS

User	Machine	Total No. of programs executed	Types of programs executed	Programs executed on weekdays	Programs executed during working hours	Programs executed after working hours	Programs executed on weekend	Total Execution time	Type of program executed Max no. of times
U01	M01	19	3	19	18	1	1	11910	LP
U02	M02	28	5	28	27	1	1	28310	UP
U03	M03	19	3	19	18	1	1	11910	LP
U04	M04	28	5	28	27	1	1	28310	UP
U05	M05	19	3	19	18	1	1	11910	UP
U06	M06	28	5	28	27	1	1	28310	UP
U07	M07	19	3	19	18	1	1	11910	LP
U08	M08	28	5	28	27	1	1	28310	UP
U09	M09	19	3	19	18	1	1	11910	LP
U10	Multiple	28	7	28	27	1	1	28310	UP
U11	Multiple	19	3	19	18	1	1	11910	UP
U12	Multiple	19	3	19	15	4	4	11910	UP
U13	Multiple	28	6	28	24	4	4	18300	UP/LP
U14	Multiple	19	3	19	15	4	4	11910	UP
U15	Multiple	19	3	19	15	4	4	11910	UP
U16	M16	13	6	13	13	0	0	8550	UP/LP
U17	M19	13	6	13	13	0	0	8550	LP
U18	M18	10	10	0	13	0	10	6720	UP
U19	M19	10	6	0	13	0	10	6720	LP

Table 6

PROFILING BASED ON FILE ACCESS

User	Machine	No. of files opened with Read access	No. of files opened with Read-write access	Total no. of files accessed	Total no. of files opened during working hours	Total no. of files opened after working hours	Total no. of files accessed on weekdays	Total no. of files accessed on weekend
U01	M01	11	8	19	18	1	19	0
U02	M02	16	12	28	27	1	28	0
U03	M03	11	8	19	18	1	19	0
U04	M04	16	12	28	27	1	28	0
U05	M05	3	16	19	18	1	19	0
U06	M06	7	21	28	27	1	28	0
U07	M07	3	16	19	18	1	19	0
U08	M08	7	21	28	27	1	28	0
U09	M09	3	16	19	18	1	19	0
U10	Multiple	7	21	28	27	1	28	0
U11	Multiple	19	0	19	18	1	19	0
U12	Multiple	19	0	19	15	4	19	0
U13	Multiple	25	3	28	24	4	28	0
U14	Multiple	19	0	19	15	4	19	0
U15	Multiple	19	0	19	15	4	19	0
U16	M16	12	1	13	13	0	13	0
U17	M19	12	1	13	13	0	13	0
U18	M18	9	1	10	10	0	0	10
U19	M19	9	1	10	10	0	0	10

Table 7

USER PROFILING BASED ON EMAIL ACCESS

User	Machine	Email program	Sent or Received	Total No. of emails sent or Received	Email Id from which Max No. of emails sent or Received	Total size of emails sent/received in Kilo bytes	Total No. of Attachments sent or received	Emails received or sent during working hours	Emails received or sent after working hours	Total No. of emails sent or received during weekdays	Total No. of emails sent or received during weekends	Sent/Received to/from mom@icare.com
U01	M01	E1	Sent	11	jones@pqr.com	5061	10	9	1	11	0	No

U02	M02	E1	Sent	12	jones@pqr.com	5065	10	9	1	12	0	Yes
U03	M03	E1	Sent	11	jones@pqr.com	5061	10	9	1	11	0	No
U04	M04	E1	Sent	12	jones@pqr.com	5065	10	9	1	12	0	Yes
U05	M05	E1	Sent	11	jones@pqr.com	5061	10	9	1	11	0	No
U06	M06	E1	Received	12	smith@abc.org	5065	10	9	1	12	0	Yes
U07	M07	E1	Sent	11	smith@abc.org	5061	10	9	1	11	0	No
U08	M08	E1	Received	12	smith@abc.org	5065	10	9	1	12	0	Yes
U09	M09	E3	Sent	11	smith@abc.org	5061	10	9	1	11	0	No
U10	Multiple	E3	Received	12	smith@abc.org	5065	10	9	1	12	0	Yes
U11	Multiple	E1	Sent	11	xyz@sai.org	5061	10	9	1	11	0	No
U12	Multiple	E1	Sent	11	xyz@sai.org	5061	10	6	4	11	0	No
U13	Multiple	E1	Sent	11	xyz@sai.org	5061	10	6	4	11	0	No
U14	Multiple	E1	Sent	11	xyz@sai.org	5061	10	6	4	11	0	No
U15	Multiple	E1	Received	11	bob@xyz.com	5061	10	6	4	11	0	No
U16	M16	E1	Received	11	bob@xyz.com	5061	10	6	4	11	0	No
U17	M19	E4	Received	11	bob@xyz.com	5061	10	6	4	11	0	No
U18	M18	E5	Received	11	bob@xyz.com	5061	10	6	4	2	9	No
U19	M19	E4	Received	11	bob@xyz.com	5061	10	6	4	2	9	No

Table 8

OVERFITTING AND OUTLIERS

Arguably, the most important safeguard in building predictive models is complexity regularization to avoid over fitting the data. When models are overfit, their accuracy is lower on new data that wasn't seen during training, and therefore when these models are deployed, they will disappoint, sometimes even leading decision makers to believe that predictive modeling "doesn't work".

One way modelers reduce the likelihood of overfit is to apply the principle of Occam 's razor, where if two models exhibit the same accuracy, we will prefer the simpler model because it is more likely to generalize well. By simpler, we must keep in mind that we prefer models that *behave* more simply rather than models that just appear to be simpler because they have fewer terms.

OUTLIERS

Outliers are also referred to as abnormalities, deviants, or anomalies in the data mining and statistics literature. In most applications, the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves in an unusual way, it results in the creation of outliers. Therefore, an outlier often contains useful information about abnormal characteristics of the systems an entities, which impact the data generation process. The recognition of such unusual characteristics provides useful application-specific insights.

ASSOCIATION RULES

1) If

Emails received or sent during working hours = 10
Emails received or sent after working hours =1
Then, Total size of emails sent/received in Kilo bytes = 5061.

Coverage = 6
Support = 6
Accuracy =100%

2) If

Emails received or sent during working hours = 11
Emails received or sent after working hours =1
Then, Total size of emails sent/received in Kilo bytes = 5065.

Coverage = 5
Support = 5
Accuracy =100%

3) If

Types of programs executed = 3
Programs executed during working hours = 18
Programs executed after working hours = 1
Type of program executed Max no. of times= LP
Then, Total Execution time = 11910

Coverage = 4
Support = 4
Accuracy = 100%

4) If

Types of programs executed = 3
Programs executed during working hours = 18
Programs executed after working hours = 1
Then, Total Execution time = 11910,

Coverage = 5
Support = 5
Accuracy = 100%

5) If

Types of programs executed = 5
Programs executed during working hours = 27
Programs executed after working hours = 1
Type of program executed Max no. of times=UP
Then, Total Execution time = 28310

Coverage = 5
Support = 5
Accuracy = 100%

6) If,

Login_period=9
Logout_period=18
Then, No. of days a user logged in=22

Coverage = 6
Support = 6
Accuracy = 100%

7) If,

No. of times a user logging in during working hours=6
No. of times a user logging in before/after working hours=16
Then, Login_period="More than once".
i.e. the user is logging in more than one interval of time.

Coverage = 4
Support = 4
Accuracy = 100%

8) If,

No. of times a user logging out during working hours=12
No. of times a user logging out before/after working hours=10
Then, Logout_period="More than once".
i.e. the user is logging out more than one interval of time.

Coverage = 4
Support = 4
Accuracy = 100%

9) If,

Total number of keyboard characters typed=246995
Total CPU usage by user processes=241844
Then, Total time Used (minutes) =11965

Coverage = 9
Support = 3
Accuracy = 33.33%
As the accuracy is falling under 75%, this cannot be made as a Rule.

10) If,

Total No. of emails sent or Received=12
Total size of emails sent/received in Kilo bytes=5065
Then, Sent/Received to/from mom@icare.com= Yes

Coverage = 5
Support = 5
Accuracy = 100%

11) If,

Total No. of emails sent or Received=11
Total size of emails sent/received in Kilo bytes=5061
Then, Sent/Received to/from mom@icare.com= No

Coverage = 14
Support = 14
Accuracy = 100%

12) If,

Total No. of emails sent or Received=11
Total size of emails sent/received in Kilo bytes=5061
Then, Email Id from which Max No. of emails sent or Received= xyz@sai.org

Coverage = 14
Support = 4
Accuracy = 28.57%

As the accuracy is falling under 75%, this cannot be made as a Rule.

13) If,

Email program='E1'
Email Id from which Max No. of emails sent or Received= jones@pqr.com
Then, Total size of emails sent/received in Kilo bytes =5061

Coverage = 5
Support = 3
Accuracy = 60%

As the accuracy is falling under 75%, this cannot be made as a Rule.
So Let us try Pruning of any one of the attribute from the above Rule,
Let's remove the attribute "Email Id from which Max No. of emails sent or Received"
The Rule after pruning is as follows.

14) If,

Email program='E1'
Then, Total size of emails sent/received in Kilo bytes =5061

Coverage = 14
Support = 10
Accuracy = 71.4%

After Pruning, we have considerably increased the Accuracy percentage but it is still falling below the threshold 75 %,
So, this cannot be made as a Rule.

15) If,

Sent or Received= 'Sent'
Email Id from which Max No. of emails sent or Received='jones@pqr.com'
Then, Total No. of emails sent or Received=11

Coverage = 5
Support = 3
Accuracy = 60%

As the accuracy is falling under 75%, this cannot be made as a Rule.
So Let us try Pruning of any one of the attribute from the above Rule,
Let's remove the attribute "Email Id from which Max No. of emails sent or Received"
The Rule after pruning is as follows.

16) If,

Sent or Received= 'Sent'
Then, Total No. of emails sent or Received=11

Coverage = 11
Support = 9
Accuracy = 81.81%

The Accuracy has considerably increased after pruning.
As, the accuracy is greater than 75 % .This can be now considered as a Rule.

17) If,

Email Id from which Max No. of emails sent or Received='bob@xyz.com'
Sent or Received='Received'
Then, Total No. of emails sent or Received=11

Coverage = 5
Support = 5
Accuracy = 100%

CLUSTERING

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters.

Help users understand the natural grouping or structure in a data set.

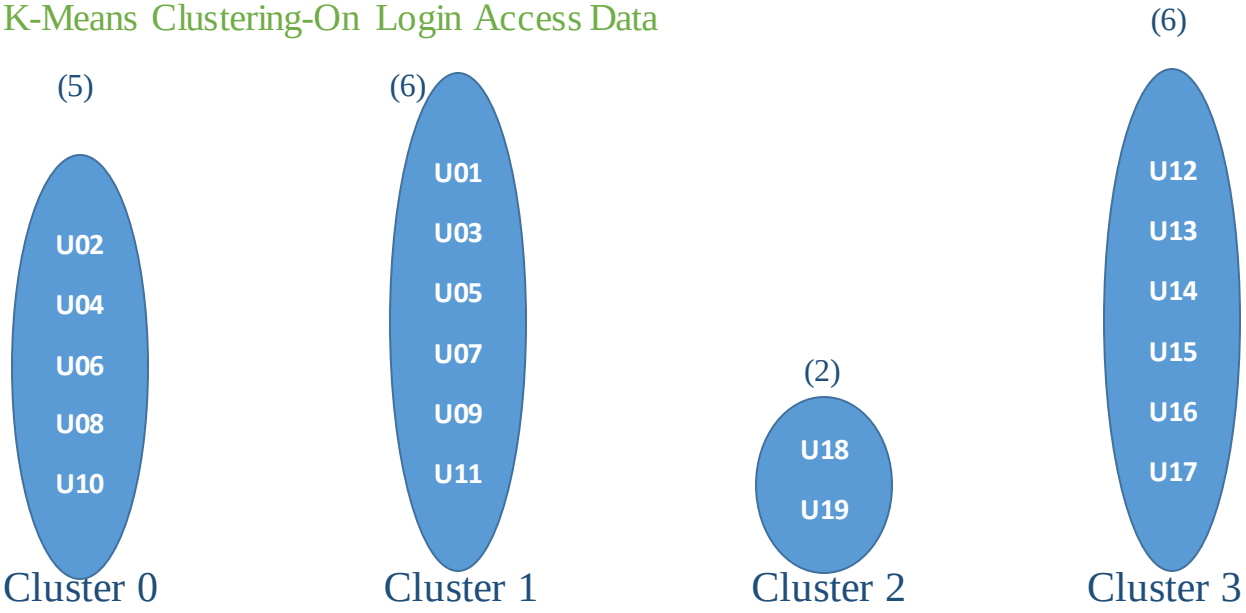
Clustering: unsupervised classification: no predefined classes. Used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms.

We have used K-Means Clustering technique to partition the given user data set into a meaningful clusters.

K-MEANS CLUSTERING TECHNIQUE

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori.
The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.
The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step
After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

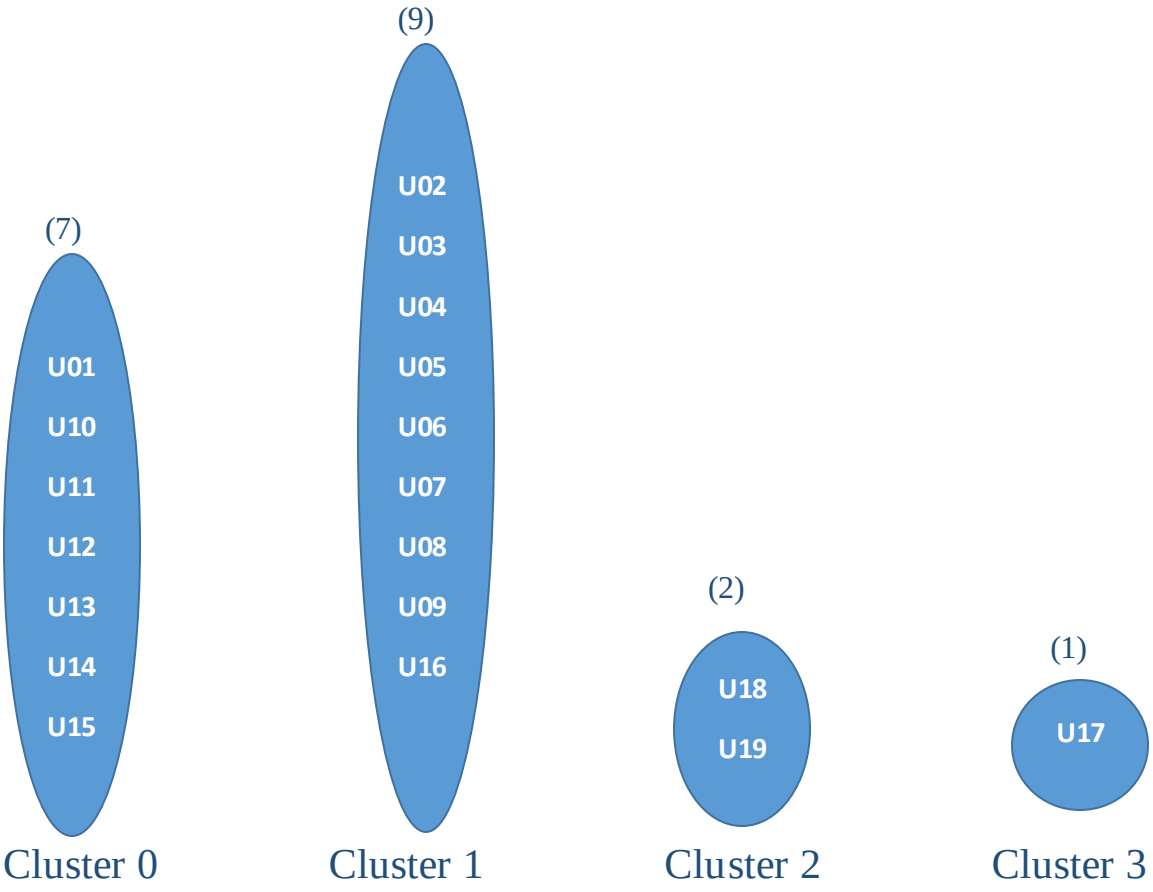
K-Means Clustering-On Login Access Data




```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 15.034233938019653
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2          3
                             (19)        (5)        (6)        (2)        (6)
=============================================================================
User                          U01        U02        U01        U18        U12
logindaysnum              20.4737         22         22          8    21.8333
Loginperiod              Multiple   Multiple          9   Multiple   Multiple
Logoutperiod             Multiple   Multiple         18   Multiple   Multiple
weekdaysloginnum          19.7368         22         22          1    21.8333
weekendloginnum            0.7368          0          0          7          0
workhrsloginnum            4.2632        6.2          0          1          8
aftworkhrsloginnum        16.2105       15.8         22          7    13.8333
workhrslogoutnum          17.1579         12         22          7         20
aftworkhrslogoutnum        3.3158         10          0          1     1.8333



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       5 ( 26%)
1       6 ( 32%)
2       2 ( 11%)
3       6 ( 32%)
```

Figure 3

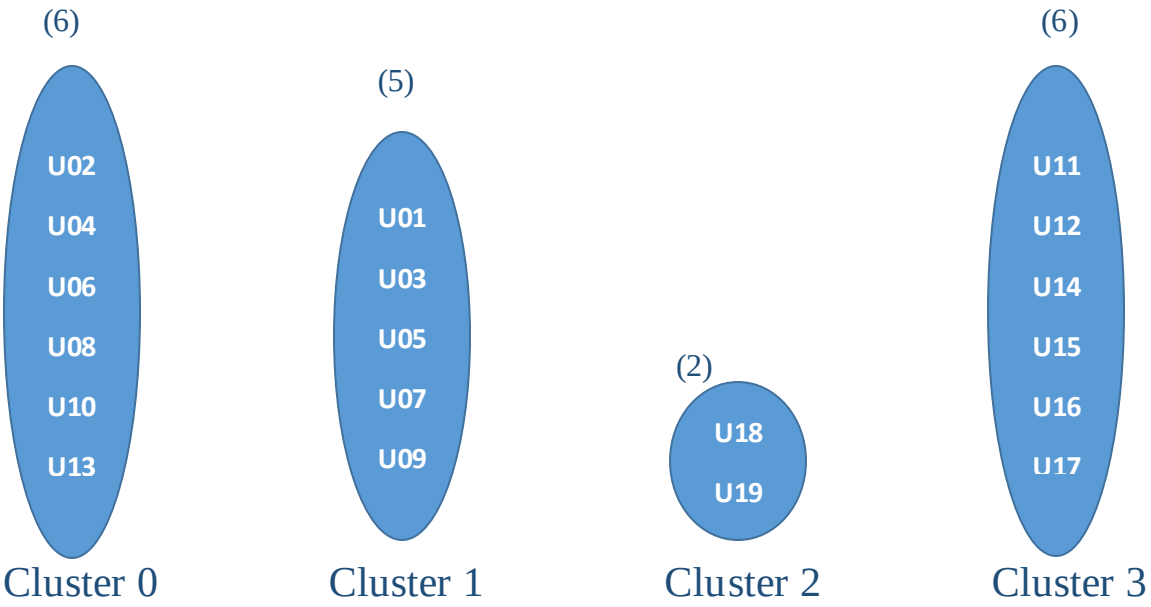
K-Means Clustering-On CPU Usage and Keyboard Usage Data




```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 25.01500905116785
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2          3
                             (19)        (7)        (9)        (2)        (1)
=============================================================================
User                          U01        U01        U02        U18        U17
Machine                  Multiple   Multiple        M02        M18        M19
CharTyped                  229590 245231.4286     246995      87820     246995
CharTypedweekday       221645.2632 245231.4286     246995      12345     246995
CharTypedweekend         7944.7368          0          0      75475          0
CPUusedProc            224660.5263 240115.7143     241844      84650     241844
TotalCPUweekday        217160.5263 240115.7143     241844      13400     241844
TotalCPUweekend               7500          0          0      71250          0



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       7 ( 37%)
1       9 ( 47%)
2       2 ( 11%)
3       1 (  5%)
```

Figure 4

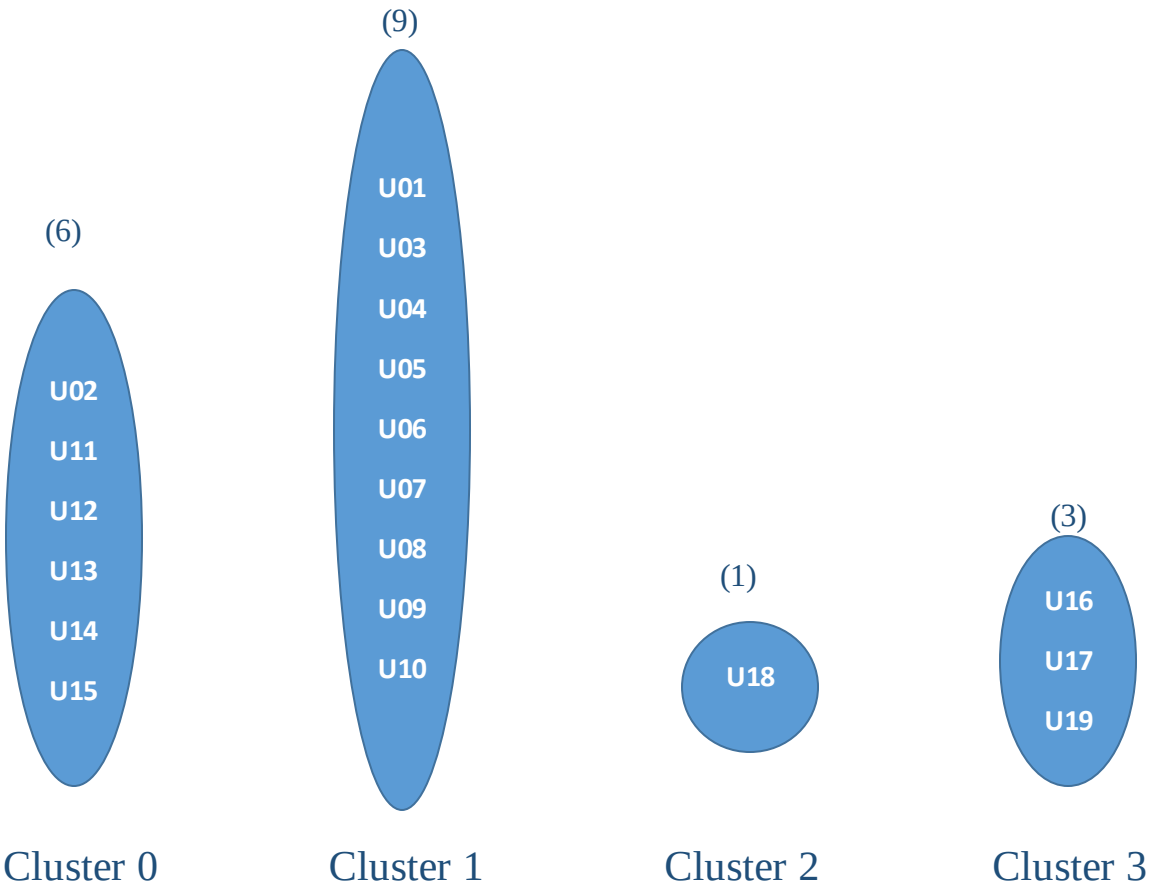
K-Means Clustering-On File Access Data




```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 29.70533749744699
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2          3
                             (19)        (6)        (5)        (2)        (6)
=============================================================================
User                         U01        U02        U01        U18        U11
Machine                 Multiple   Multiple        M01        M18   Multiple
filesrdaccessnum         11.9474         13        6.2          9    16.6667
filesrdwtaccessnum        8.3158         15       12.8          1     0.3333
totalfiles               20.2632         28         19         10         17
fileaccessworkhrs        18.8421       26.5         18         10    14.8333
fileaccessaftworkhrs      1.4211        1.5          1          0     2.1667
fileaccessweekday        19.2105         28         19          0         17
fileaccessweekend         1.0526          0          0         10          0



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       6 ( 32%)
1       5 ( 26%)
2       2 ( 11%)
3       6 ( 32%)
```

Figure 5

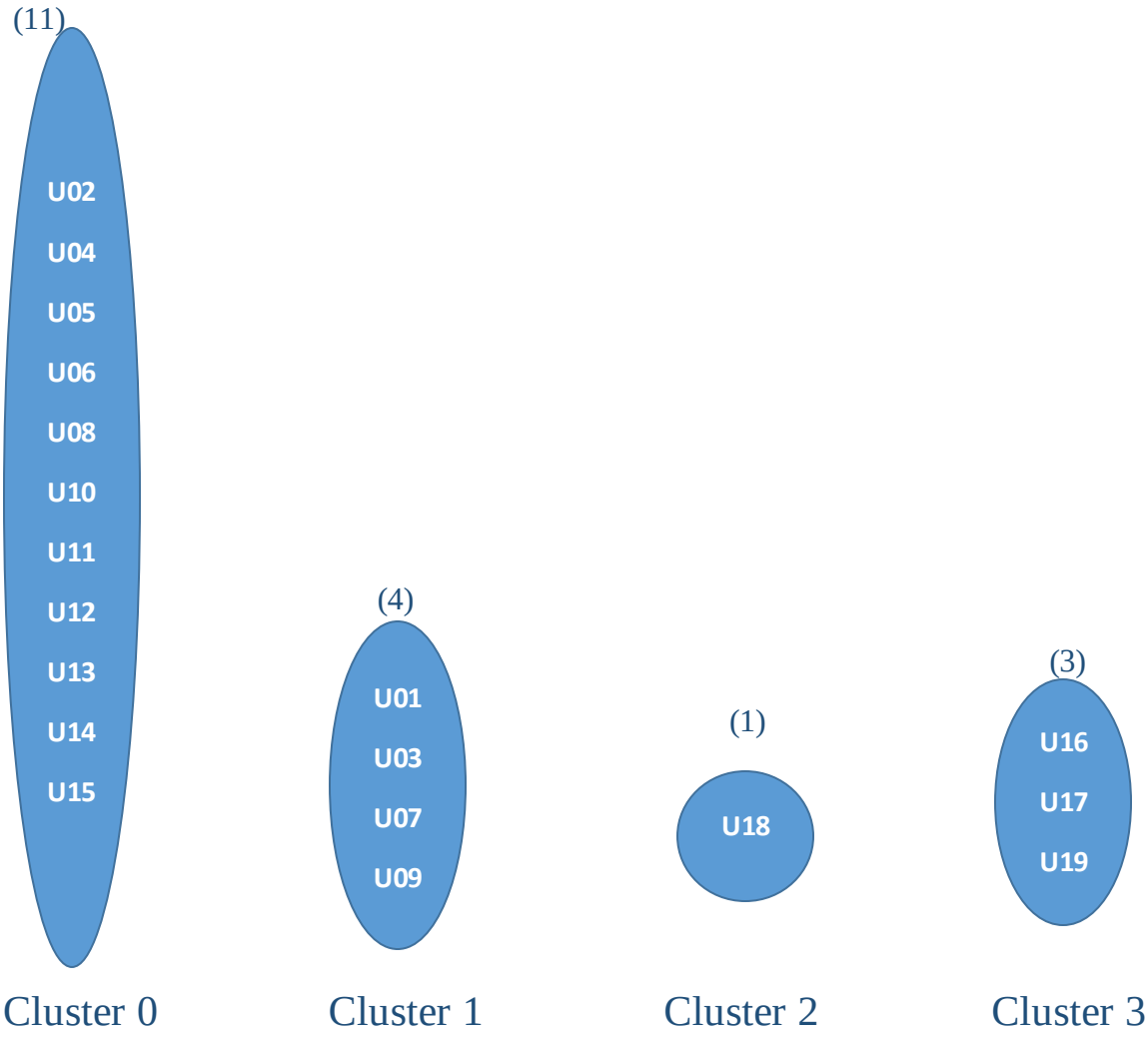
K-Means Clustering-On Printer access Data



```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 34.05838234338559
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute             Full Data         0         1         2         3
                           (19)        (6)       (9)       (1)       (3)
=============================================================================
User                        U01       U02       U01       U18       U16
Machine                Multiple  Multiple       M01       M18       M19
printer                     PR2       PR4       PR2       PR5       PR6
pagesprinted           135.5789       155       150        56        80
weekdayprintpage       129.6842       155       150         0   61.3333
weekendprintpage         5.8947         0         0        56   18.6667
workhrprintpage        114.1053       117       130        56        80
aftworkhrprintpage      21.4737        38        20         0         0



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       6 ( 32%)
1       9 ( 47%)
2       1 (  5%)
3       3 ( 16%)
```

Figure 6

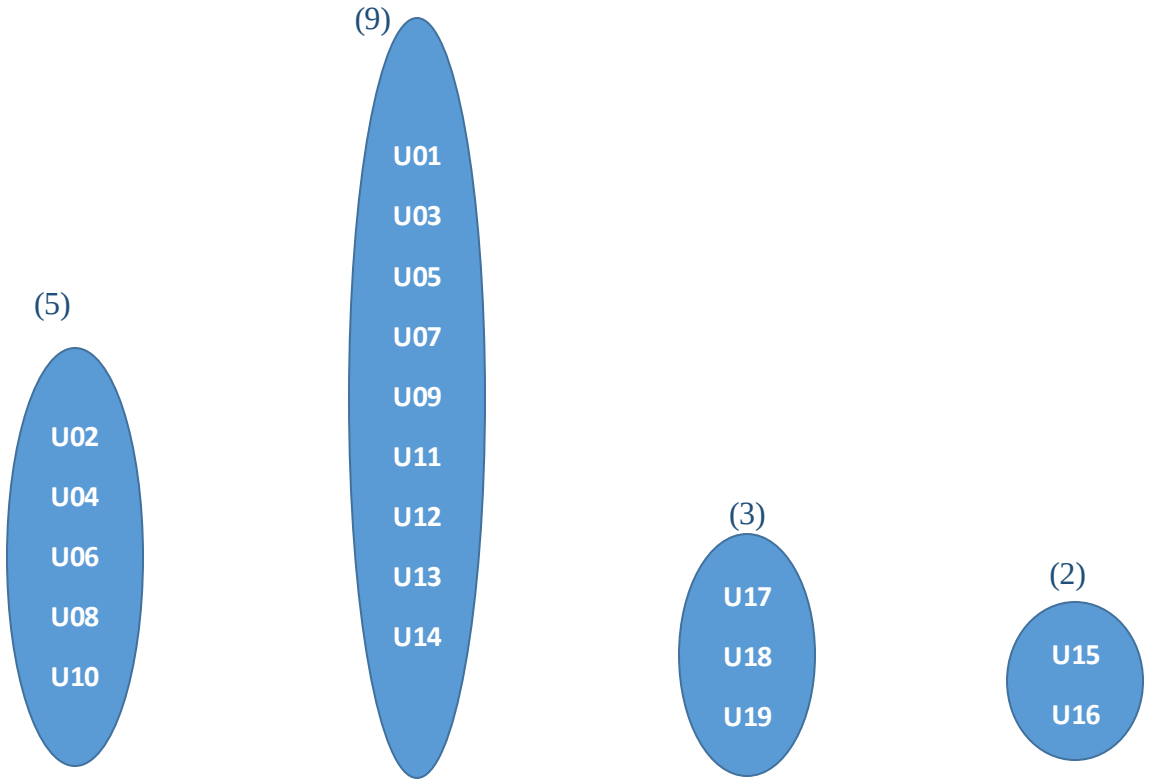
K-Means Clustering-On Program Access Data



```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 32.89596098857761
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute             Full Data         0          1          2          3
                           (19)       (11)        (4)        (1)        (3)
=============================================================================
User                        U01        U02        U01        U18        U16
Machine                Multiple   Multiple        M01        M18        M19
Totalprogexec           20.2632    23.9091         19         10         12
typeprogexec             4.6316     4.3636          3         10          6
progexecweekday         19.2105    23.9091         19          0     8.6667
progexecworkhrs         19.1579    21.8182         18         13         13
progexecaftworkhrs       1.4211     2.0909          1          0          0
progexecweekend          2.4737     2.0909          1         10     3.3333
totalexectime        15662.1053 19945.4545      11910       6720       7940
progtypename                UP         UP         LP         UP         LP



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      11 ( 58%)
1       4 ( 21%)
2       1 (  5%)
3       3 ( 16%)
```

Figure 7

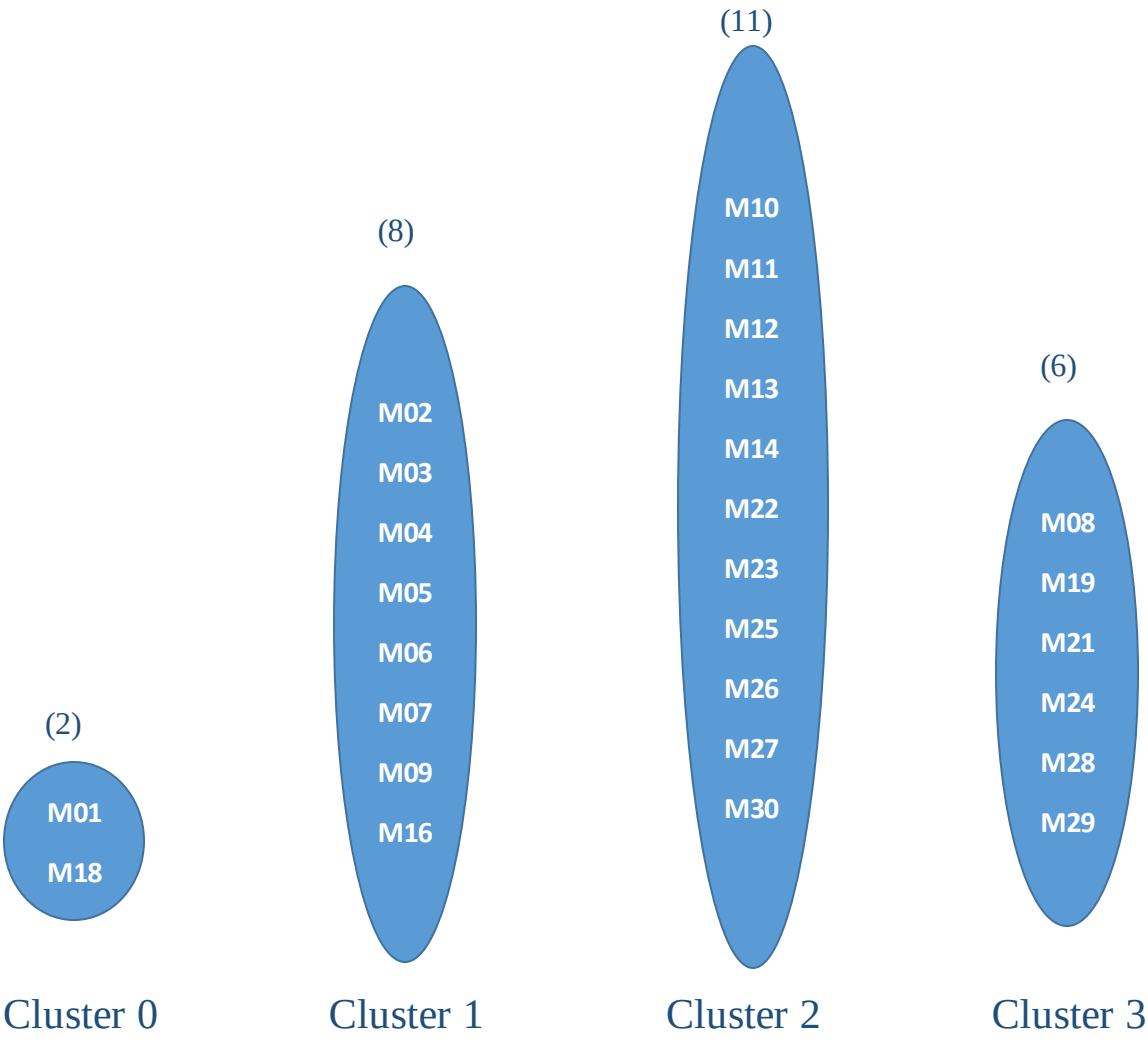
K-Means Clustering-On Email Usage Data



```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 43.20666666666666
Missing values globally replaced with mean/mode

Cluster centroids:
                                     Cluster#
Attribute                Full Data          0          1          2          3
                               (19)        (5)        (9)        (3)        (2)
=============================================================================================
User                            U01        U02        U01        U17        U15
Machine                    Multiple        M02   Multiple        M19   Multiple
Emailprogram                    E1         E1         E1         E4         E1
SentorReceived                Sent   Received       Sent   Received   Received
TotalNo.ofemails           11.2632         12         11         11         11
EmailMaxemail        jones@pqr.com smith@abc.org xyz@sai.org bob@xyz.com bob@xyz.com
Totsizeoemails/rKb       5062.0526       5065       5061       5061       5061
No.ofAttachments                10         10         10         10         10
Emailworkhrs               7.7368          9          8          6          6
Emailaftworkhrs            2.2632          1          2          4          4
Emailweekday              10.3158         12         11          5         11
Emailweekend               0.9474          0          0          6          0
s/rt/fmom@icare.com             No        Yes         No         No         No



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       5 ( 26%)
1       9 ( 47%)
2       3 ( 16%)
3       2 ( 11%)
```

Figure 8

K-Means Clustering-On Machine Usage Data



```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 45.253128464456445
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute          Full Data          0          1          2          3
                        (27)        (2)        (8)       (11)        (6)
=============================================================================
Machine                  M01        M01        M02        M10        M08
shared                    No         No         No        Yes        Yes
totalusedtime       7053.3704     8063.5   11092.625   3700.2727  7478.3333
frequser             Multiple        U01        U02   Multiple   Multiple
weekdayuse                Yes        Yes        Yes        Yes        Yes
weekenduse                Yes        Yes        Yes         No        Yes
keybdchartyped    161563.3333   167407.5     246995   83860.4545 188161.6667
CPUuseproc        158094.4444     163247     241844   83620.3636 181246.6667



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       2 (  7%)
1       8 ( 30%)
2      11 ( 41%)
3       6 ( 22%)
```

Figure 9

IMPLEMENTATION OF DATA MINING SECURITY TECHNIQUES:

One Of the Key points that has be considered while working on data is the user privacy. Since the given data does not contain any private information about the user or any Social Security information, there is no major threat of data security of the user. Confidentiality and integrity of user data has been all through the project work.

However the login, printer ,machine, file access details of the user are all required by the organization's management to keep track of the employees, and as this information is not as sensitive as credit card or other private information data mining security is not of much concern with respect to the given data in our personal opinion.

The data is not been modified or tampered, the integrity of the data has been maintained. Since the name of the user also is hidden while studying the data, so privacy preserving data mining techniques are considered.

CONCLUSIONS

After careful analysis of data using different data mining techniques we have come to the following conclusions about the user behavior based on the Login Access pattern, CPU & Keyboard Usage pattern, File Access, Printer Access, Program Access, email usage and machine usage.

About Login Access Pattern

1) From the type 1 data we found out that all the Users have logged in to the machines on 22 days in the given month counting both weekdays and weekends, except for Users U18 and U19 who logged in for only 8 days in the given month.

2) There are few users who have logged in at a particular time interval and logged out at a particular time interval like Users U01,U03,U05,U07,U11 while the remaining users logged into the machines at multiple times.

About CPU & Keyboard Usage pattern

1) From the type 1 user data it is very apparent that all the users have typed 246995 characters and Total CPU Use, 241844 except for the Users U15, U18 and U19.

2) Only two Users have considerable CPU Usage and keyboard Usage during weekends (U18, U19) while the remaining users have zero usage.

About Machine Usage pattern

1) From the type 1 User data we observed that there are a total of 27 machines among which 15 are used exclusively by users and 12 are shared between multiple users.

2) Machines M08 and M19 are used for maximum time and Machines M18 and M19 are the only two machines that are used during weekends in the given month.

About Program access

1) From the Type 2 User data we observed that the UP type of programs are executed maximum number of times and have a larger contribution over the total execution time when compared to the LP Programs.

2) From the additional attributes we derived, it is apparent that most number of programs are executed during working hours and on weekdays.

About Printer access

1) From the type 2 user data we observed that Printer PR 2 is used by maximum number of Users.

2) From the additional attributes we derived we found out that Use U13 has printed the maximum number of pages for the given month.

About File access

1) From the type 2 user data we observed that Users U06, U08 and U10have updated the files maximum number of times.

2) User U13 has opened files maximum number of times and Users U18 and U19 are the only two users who have accessed the files during weekends.

About email access

1) From the type 3 user data, we have observed that the Users U02, U04, U06, U08, U10 have sent or received personal mails.

i.e. mails to/from mom@icare.com

2) Most number of emails are received only from smith and Bob.

REFERENCES:

http://faculty.wiu.edu/C-Amaravadi/is524/res/dm_c_ov.pdf

www.cs.bu.edu/fac/gkollios/ada05/.../lect19-05.ppt

http://abbottanalytics.blogspot.com/2014/05/why-overfitting-is-more-dangerous-than.html

http://charuaggarwal.net/outlierbook.pdf

https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

http://ijcttjournal.org/Volume4/issue-2/IJCTT-V4I2P129.pdf