A Report

on

Applying K-means Clustering Algorithm on HCI data

By:

Manasi Kattel
CE 3rd year

Contents

Introduction

HCI dataset is a data collected from a moodle system of the students of Department of Computer Science and Engineering, Kathmandu University of the course Human Computer Interaction (HCI). The data initially consisted of the attributes : Time, User full name, Event context, Component, Event name, Description, Origin and IP address.
Time denotes the particular date and time at which a certain event was performed by a user. User full name includes the full name of the user who performed a certain event. Event context denotes the context of the event whether it is course related or related to other contexts like research paper or mini project. Event name is basically the name of the event occurred and event description is a brief description of that event. IP address includes the IP addresses of the devices from which moodle system was accessed.

Time	User full n	Event con	Compone	Event nam	Descriptio	Origin	IP address
3/02/18, 1	Sushil shre	Course: C(	Logs	Log report	The user v	web	
3/02/18, 1	Sushil shre	Course: C(	Logs	Log report	The user v	web	27.34.49.92
3/02/18, 1	Sushil shre	Course: C(	Activity re	Activity re	The user v	web	27.34.49.92
3/02/18, 1	Sushil shre	Course: C(	System	Course vie	The user v	web	27.34.49.92
3/02/18, 1	Sushil shre	Course: C(	System	Course vie	The user v	web	27.34.49.92
3/02/18, 1	Sushil shre	Course: C(	System	Course vie	The user v	web	27.34.49.92
3/02/18, 1	Sushil shre	Course: C(	System	Course se	The user v	web	27.34.49.92
3/02/18, 1	Sushil shre	Course: C(	System	Course vie	The user v	web	27.34.49.92
3/02/18, 1	Sushil shre	Course: C(	System	Course vie	The user v	web	27.34.49.92
3/02/18, 1	Sushil shre	Course: C(	System	Course vie	The user v	web	27.34.49.92
3/02/18, 1	Sushil shre	Course: C(	System	Course vie	The user v	web	27.34.49.92

Fig: Snippet of initial HCI data

Since K-means clustering algorithm best works with numerical data, a new dataset was created that includes name of the user and the number of times he accessed the moodle system. Since Sushil Shrestha is the course instructor, his name was removed from the list of users.

Name	Count
'Manish R	120
'Shristi Sha	107
'Pratit Raj	87
'Dipesh Kh	86
'Monika S	81
'Sushmi Sh	69
'Shashwot	68
'Rabindra	58
'Rubina Ph	49
'Shaileaf T	47
'Sakshi Sh	45
'Suprim La	45
'Suman Dh	43
'Ashin Pot	42
'Sadikshya	41
'Anushara	39

Fig: Updated data

Result

In Weka tool, the K-means clustering algorithm is now applied to the data considering the count of times each user has accessed the website. The results of the clustering are shown in figure below :-

```
Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1
                (26.0)     (18.0)      (8.0)
=============================================
Count           47.3077    30.7778       84.5



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

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

Clustered Instances

0      18 ( 69%)
1       8 ( 31%)
```

Fig: Result of K-means clustering on HCI data

From the above figure, it can be inferred that the algorithm has clustered the data into two clusters: one with centroid value 30.7778 and other with centroid value 84.5. This means that users who have accessed the system around 30 times are in the cluster 0 and the users who have accessed the system around 84 times are in cluster 1. Hence, cluster 1 is a cluster of active users whereas cluster 0 is a cluster of less active users.

Visualization of clusters

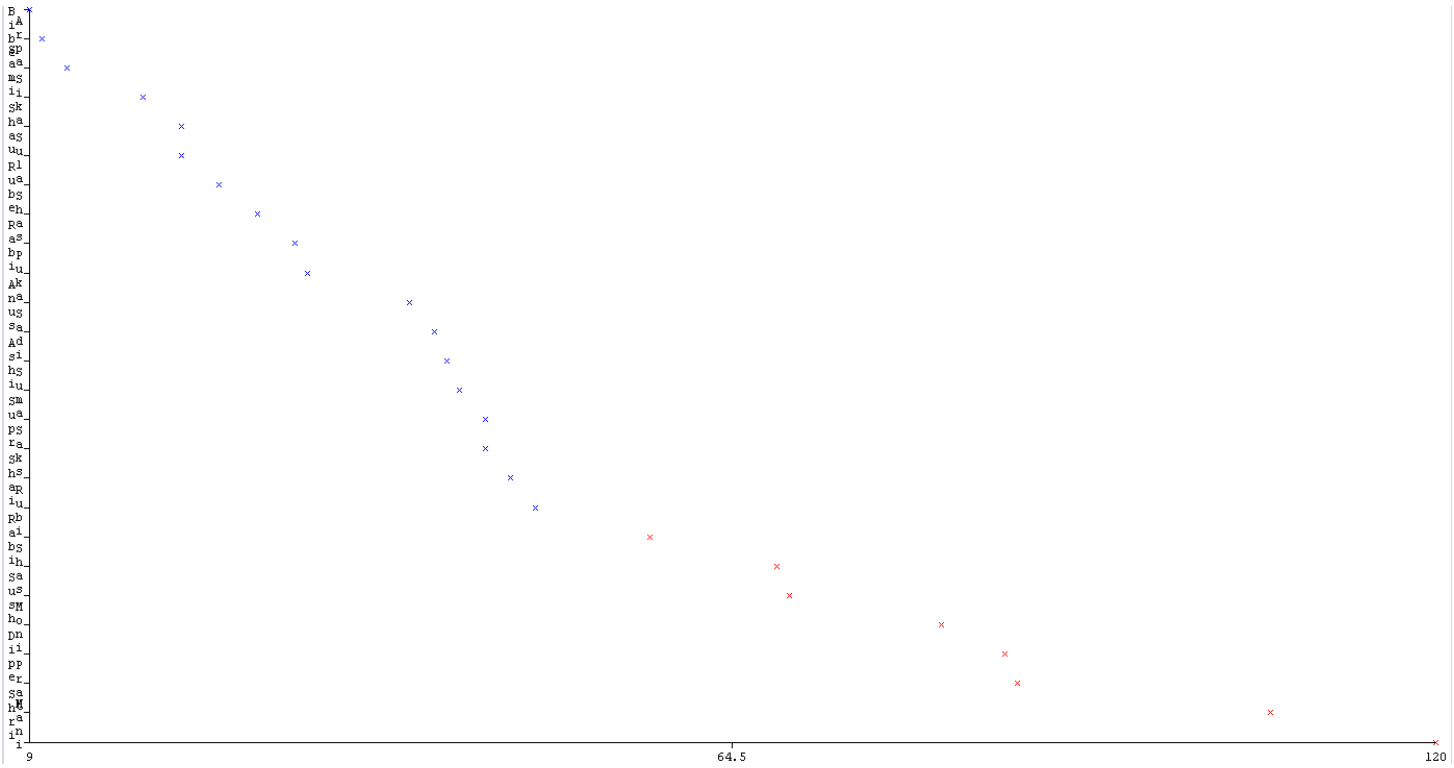

Fig: Visualization of cluster assignments

Above figure shows visualization of cluster assignments. It is a plot of count in X-axis and Names in Y-axis. The blue points represent cluster 0 and red points represent cluster 1. Since the names are long, the names aren't clearly visible in the Y-axis. However, on Weka, we can click on the desired instance to know the information regarding that instance.

Conclusion

Thus, from HCI data clustered instances, it can be concluded that there are less number of active users and more number of inactive users. Out of 26 students, only 8 of them are active whereas 18 of them are less active.