IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Special Issue-8, Apr 2019 E-ISSN: 2347-2693

Illustration of K Mean Clustering Algorithm for Analysing Laptop Utilization Dataset

V. Lakshmi Praba[1*], M.A. Saira Banu[2]

[1, 2]Department of Computer Science Rani Anna Government College for Women, Tirunelveli, India

DOI: https://doi.org/10.26438/ijcse/v7si8.6165 | Available online at: www.ijcseonline.org

Abstract— Laptops finds wide application in different fields by different users. School and College students are provided with Laptops freely distributed by the Government. These laptops are used by the students for various purposes like academic, programming, writing, editing documents, etc. To carry out the task of analysing the Laptop utilization characteristics, data has been collected from college students by supplying questionnaires. This paper examines student's perceptions related to the usage of laptop by analyzing its utilization characteristics using Simple K-Means clustering algorithm.

Keywords— Simple K-Mean Clustering, Centroids, Clusters, WEKA tool

I. INTRODUCTION

Clustering is splitting of a large dataset into clusters or groups. Every cluster or group must contain one data item and every data item must be in one cluster. Clustering is an unsupervised technique that is applicable on large datasets with a large number of attributes. It is a data modelling technique that gives a concise view of data.

A. Simple k-mean clusterin

Simple K-means clustering is an unsupervised learning algorithm that is used to cluster instances based on characteristics. It adopts the procedure of classifying a given data set into a number of clusters '*k*' which is fixed initially. The clusters are then positioned as points called centroids and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process starts over using the new adjustments until a desired result is reached. The WEKA tool helps to find out number of objects in each cluster. This leads to a statistical approach that finds points of interest for observation and further investigation.

B. Laptop utilization

This paper uses laptop utilization by college students as data set. The information has been collected from 477 students who are using laptops. From this data set, the utilization of laptop characteristics is determined by applying Simple K-Means algorithm.

II. LITERATURE SURVEY

Many authors have contributed many research articles is the area of Data mining especially in clustering techniques.

Amir Ahamad and Lipika Dey [1] proposed a modified description of cluster center to the numeric data by providing a better characterization of clusters. A new cost function and distance measure based on co-occurrence of values is resented. The measures also take into account the significance of an attribute towards clustering process.

Saroj and Kavita [2] this paper provides a comprehensive review of simple k mean clustering and modified k mean clustering techniques. In this paper simple k mean clustering has been described by using the WEKA tool with medical data set. On the other hand Modified k mean clustering has been described based on normalization and indexing approach using .NET which takes less time with minimum no. of sum of squared errors to execute the cluster.

Richa Agrawal and Jitendra Agrawal [3] In this paper, various clustering algorithm are analyzed and compared by using WEKA tool to find out which algorithm will be more comfortable for the users for performing clustering. It performed analysis with four clustering algorithms k-mean, LVQ, SOM and COBWEB. In all four algorithm result is generated on the basis of similar objects and time to create that clusters. Best algorithm found is k-mean clustering. It is taking less time than other clustering algorithm to find similar clusters through WEKA tool for air pollution dataset.

Bharat Chaudhari and Manan Parikh [4] This paper analyze the three major clustering algorithms and compare the performance of these clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm. Performance of three techniques are presented and compared using a clustering tool WEKA.

Harjot Kaur and Er. Prince Verma [5] in this paper, the focus is on data mining's algorithm i.e., clustering, a method of grouping data items in a cluster and to review its various algorithms. The paper starts by basic definition of clustering and listed all its possible algorithms for clustering. The paper concludes that portioning algorithms are efficient algorithms as compared to other algorithms in clustering technique.

Akanksha Mahajan and ER. Neena Madan [6] in this paper, analysis is carried out with Hierarchical and k means algorithm. The hierarchical clustering provides good result as compared to k means clustering for better analysis on data.

G. Shiyamala Gowri. et. al [7] Academical Data Mining is a multi-disciplinary research area that examines artificial intelligence, statistical modelling and data mining with the data mining with data generated from an Academical institution. It utilizes computational ways to deal with explicate Academical information keeping in mind the end goal to examine Academical inquiries.

Jiawei Han. et. al [8] presents data mining concepts and techniques. Different clustering techniques including K-Means clustering is well explained in this book.

III. METHODOLOGY

In this paper based on the literature survey and wide usage, K-Means clustering algorithm has been chosen for clustering the laptop utilization dataset collected from college students.

The utilization purpose has been broadly classified into three areas namely – Academic, Social and Entertainment with many attributes under each classification. For analysis, three clusters are considered namely - Effective Utilization, Medium Utilization and Poor Utilization based on the attribute characteristics under each classification.

A. Simple k-mean algorithm

Simple K-means clustering algorithm is first proposed by Macqueen in 1967 which was uncomplicated, non-supervised learning clustering algorithm. Simple K-mean is a portioning clustering algorithm. This technique is used to classify given data objects into different ***k*** clusters through the iterative method, which tends to converge to a local minimum. So the outcomes of generated clusters are dense and independent of each other. Simple K-Means is the most important flat clustering algorithm. Its objective is to minimize the average squared Euclidean distance of documents from their cluster centers where a cluster center is defined as the mean or centroid μ of the documents in a cluster ω: centroid in Eqn (1),

$$\overrightarrow{\mu(\omega)} = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x} \qquad (1)$$

B. Algorithm Description

Input: [D, k] // D→ data set of laptop utilization;
k→ no of clusters; m →mean value of each cluster
Output: k clusters
Process:
Step 1: Choose k clusters from data set.
Step 2: Assign seed value to data set.
Step 3: Calculate the Euclidean distance between each data point and cluster centroid.
Step 4: Assign each instance from data set to one of the k clusters based on Euclidian distance from the centroid.
Step 5: Repeat steps 3 and 4 until all instances are clustered.

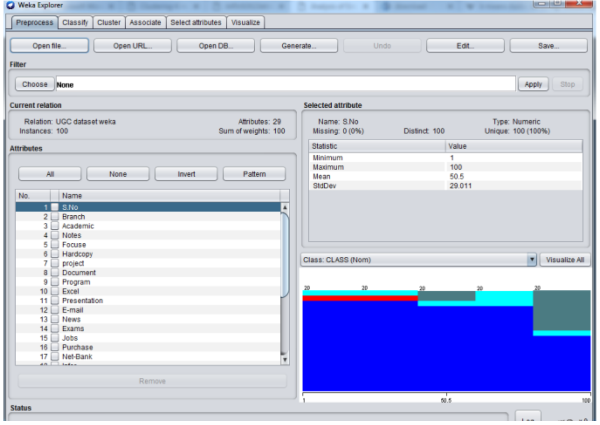
C. Dataset description

In this study, the dataset consists of 30 attributes and 477 records/instances that are used Laptop Utilization by college students. The dataset detail is as given Table 1.

Table 1: Dataset for Laptop Utilization

Dataset Name	Number of Attributes	Number of Records/Instances
Laptop Utilization	30	477

The attributes are based on data types. The data set is based on the numeric and nominal data type. The input data are fed into the Excel by the nominal or numerical values only. It is shown in Fig. 1.



Filter: None

Relation: UGC dataset weka
Instances: 100
Attributes: 29
Sum of weights: 100

Name: S.No
Missing: 0 (0%)
Distinct: 100
Type: Numeric
Unique: 100 (100%)

Statistic	Value
Minimum	1
Maximum	100
Mean	50.5
StdDev	29.011

No.	Name
1	S.No
2	Branch
3	Academic
4	Notes
5	Focuse
6	Hardcopy
7	project
8	Document
9	Program
10	Excel
11	Presentation
12	E-mail
13	News
14	Exams
15	Jobs
16	Purchase
17	Net-Bank

Class: CLASS (Nom)

Fig. 1

D. Dataset interpretation

Each attribute has been categorized as STRONGLY AGREE, AGREE, DISAGREE and STRONGLY DISAGREE which holds the numeric values as 3, 2, 1 and 0 respectively.

E. Attribute description

The selected attributes and its categories are as shown in Table 2.

Table 2: Laptop Utilization Dataset Attributes

ID	ATTRIBUTE	DATA TYPE

1	S.NO.	Numerical
2	Branch	Nominal
3	Academic	Numerical
4	Notes	Numerical
5	Focus	Numerical
6	Hardcopy	Numerical
7	Project	Numerical
8	Document	Numerical
9	Program	Numerical
10	Excel	Numerical
11	Presentation	Numerical
12	E-mail	Numerical
13	News	Numerical
14	Exams	Numerical
15	Jobs	Numerical
16	Purchase	Numerical
17	Net-Bank	Numerical
18	Infor.	Numerical
19	Download	Numerical
20	Soc.Net.	Numerical
21	Chatting	Numerical
22	Movies	Numerical
23	Music	Numerical
24	Video	Numerical
25	Images	Numerical
26	Games	Numerical
27	Design	Numerical
28	Academic	Numerical
29	Social	Numerical
30	Entertainment	Numerical

F. Performance metrics Sum of squared error within clusters

The most commonly used clustering strategy is based on the square-root error criterion. To minimize the square-error where square-error is the sum of the Euclidean distances or any other distances between each instance/observation and its cluster center.

The sum of squared error (SSE^2) indicates how compact a cluster is: the lower the value, the better.

$$\mathrm{SSE}\ (X, \Pi) = \sum_{i=1}^{K} \sum_{x_j \epsilon C_i} \left\| x_j - m_i \right\|^2 \qquad (2)$$

Eqn (2) is being counted as the sum of square differences between the value of the attribute of each instance and the value of the centroid of the given attribute.

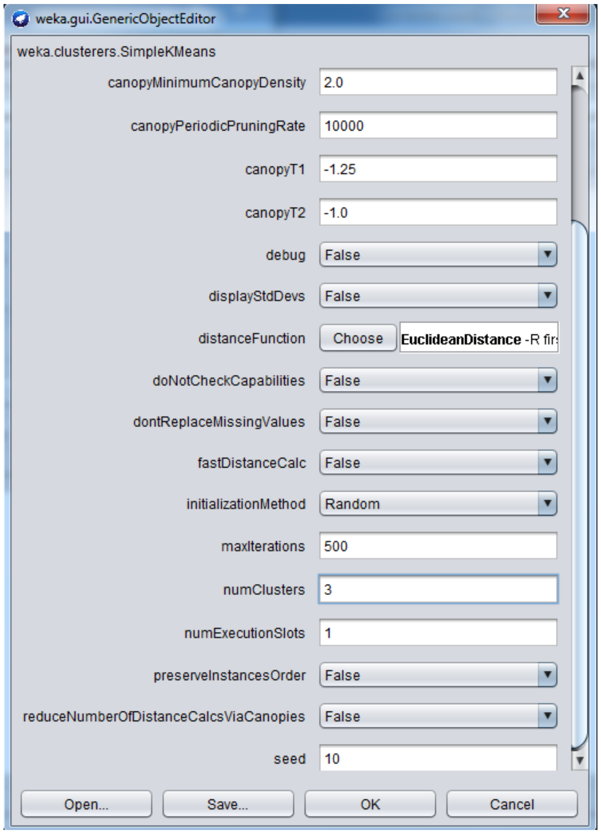
IV. EXPERIMENTAL SETUP

A. WEKA tool

WEKA is a data mining system developed at the University of Waikato in New Zealand that implements data mining algorithms. WEKA is a state-of-the-art facility for developing machine learning techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, classification, regression, clustering, association rules; it also includes a visualization tools. The new machine learning schemes can also be developed with this package. WEKA is open source software issued under the GNU General Public License. This work is carried out using WEKA tool version 3.8.3.

B. Simple k-means setup in WEKA tool

Simple K-Mean method is chosen with 'Use training set' cluster mode. Euclidean distance and Random initialization method is set. The number of desired clusters is set as 3. The maximum number of iteration is set to 500 and initial seed value set to 10 instances. K-Means clustering parameter setting used for analysis is shown in Fig. 2



weka.gui.GenericObjectEditor

weka.clusterers.SimpleKMeans

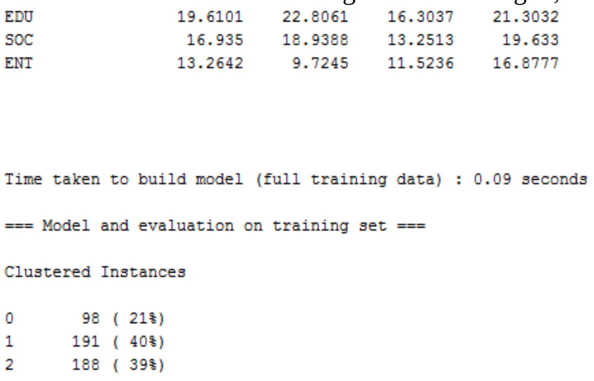
canopyMinimumCanopyDensity 2.0
canopyPeriodicPruningRate 10000
canopyT1 -1.25
canopyT2 -1.0
debug False
displayStdDevs False
distanceFunction Choose EuclideanDistance -R firs
doNotCheckCapabilities False
dontReplaceMissingValues False
fastDistanceCalc False
initializationMethod Random
maxIterations 500
numClusters 3
numExecutionSlots 1
preserveInstancesOrder False
reduceNumberOfDistanceCalcsViaCanopies False
seed 10

Fig. 2 Parameter settings for clustering

V. RESULT ANALYSIS

The dataset values are set as 3, 2, 1 and 0 for STRONGLY AGREE, AGREE, DISAGREE and STRONGLY DISAGREE respectively. The utilization purpose has been broadly classified into three areas namely – Academic (9 attributes), Social (9 attributes) and Entertainment (7 attributes) under each classification. A total of 25 attributes with a maximum value of 75 (all STRONGLY AGREE) is possible. For analysis, three clusters are considered namely - Effective Utilization, Medium Utilization and Poor Utilization based on the attribute characteristics under each classification.

The obtained result for clustering is shown in Fig. 3,



```
EDU                 19.6101    22.8061    16.3037    21.3032
SOC                  16.935    18.9388    13.2513     19.633
ENT                 13.2642     9.7245    11.5236    16.8777



Time taken to build model (full training data) : 0.09 seconds

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

Clustered Instances

0       98 ( 21%)
1      191 ( 40%)
2      188 ( 39%)
```

Fig. 3 Cluster Result

The utilization of laptop has been classified into 3 categories (3 clusters) namely, Effective Utilization (Cluster 2), Medium Utilization (Cluster 0) and Poor Utilization (Cluster 1).

With 477 instances it has been observed that utilization for entertainment is lesser with the value of 13.26 as centroid and 16.93 for Social utilization and with the high rating of 19.61 for Academic utilization.

Similarly from the obtained results it is observed that in all the clusters the value for Entertainment is lesser. Comparing with Social, Academic values always shoots higher.

Effective Utilization cluster values are compared with all data and the following characteristics are observed.

Fig. 4 shows the bar chart depicting Effective Utilization.

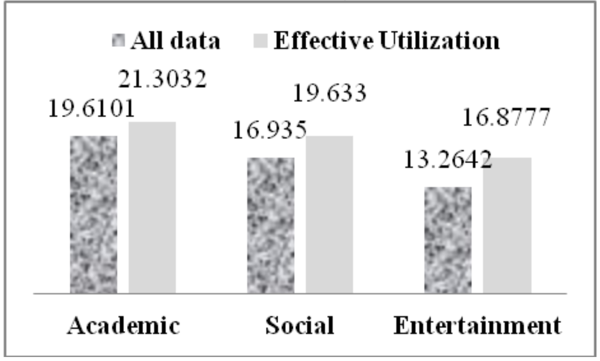


Fig. 4 All data vs. Effective Utilization

The result of Effective Utilization cluster indicates that it has an average of 8.63% of improvement when compared with the entire dataset value for Academic utilization purpose. Similarly, the Percentage of improvement for Social utilization and Entertainment utilization is 15.93 and 27.24 respectively.

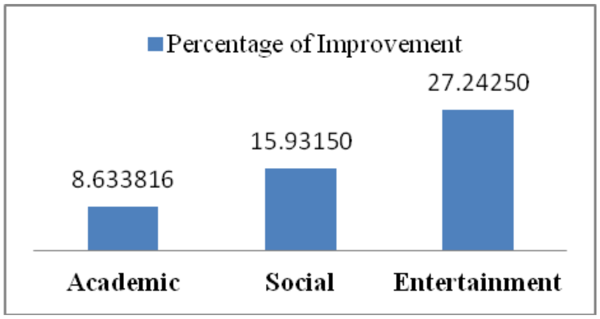


Fig. 5 Percentage of Improvement

From the Fig. 5 it is evident that for all the three utilization categories, there is a considerable Percentage of Improvement under the Effective utilization cluster category.

Fig. 6 shows the bar chart depicting Medium Utilization.

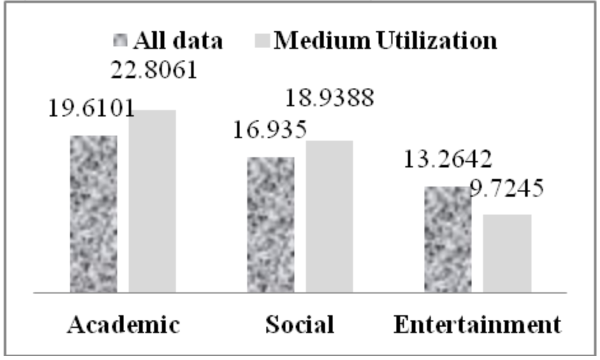


Fig. 6 All data vs. Medium Utilization

The result of Medium Utilization cluster indicates that it has the Percentage of improvement of 16.29 and 11.83 for Academic utilization and Social utilization respectively. The Percentage of value has been pulled down to -26.68 in the case of utilization for Entertainment.

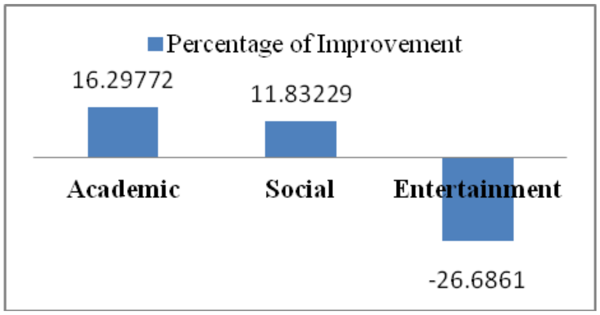


Fig. 7 Percentage of Improvement

From the Fig. 7 it is evident that for all the three utilization categories, there is a considerable Percentage of Improvement under the Medium utilization cluster category.

Fig. 8 shows the bar chart depicting Poor Utilization.

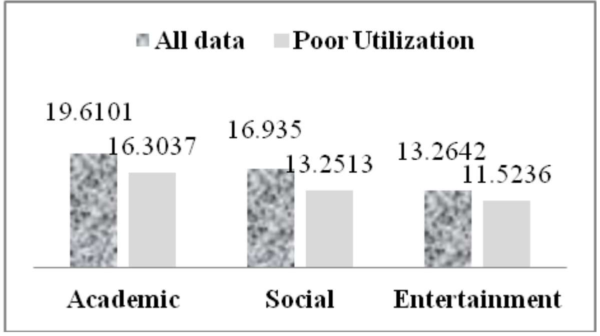


Fig. 8 All data vs. Poor Utilization

The result of Poor Utilization clearly indicates that for all the three categories of Academic utilization and Entertainment utilization there is no percentage of improvement. It is shown in Fig 9.

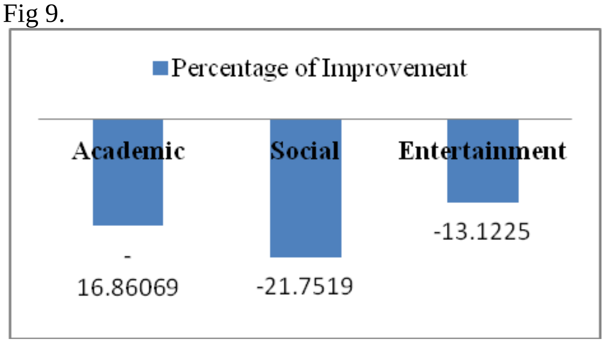


Fig. 9 Percentage of Improvement

Altogether, while considering the entire dataset 39% of students are effectively utilizing the laptop, 21% with Medium utilization and 40% with Poor utilization (Fig. 10).

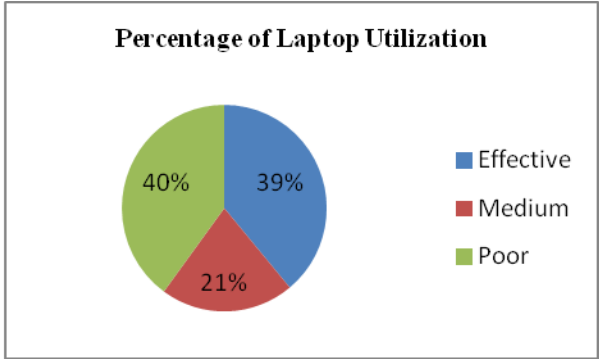


Fig. 10 Overall Percentage of Laptop Utilization

VI. CONCLUSION

A detailed study has been carried out with the collected dataset and the results are observed. The result clearly indicates that laptop usage has a definite impact on educating the students for their betterment. Observed result indicates 39% of the students use the laptop effectively and 21% with medium utilization covering to a maximum of 60% better utilization.

In this work, K means clustering algorithm is considered taking three attributes into consideration for analysis. In future more attributes like combination of these basic attributes can be considered and analysed.

ACKNOWLEDGEMENT

This work is carried out as part of the project funded by UGC Minor Research Project.

REFERENCES

[1] Amir Ahamad, Lipika Dey, “A K-Mean clustering algorithm for mixed numerical and categorical data”, Data & Knowledge Engineering, pp 503-527, Vol.63, Iss. 2, November 2007.

[2] Saroj, Kavitha, “Study on Simple k Mean and Modified k Mean Clustering Technique”, IJCSET, (279-281) Vol. 6- No.7, July 2016.

[3] Richa Agarwal, Jitendra Agarwal, “Analysis of Clustering Algorithm of WEKA Tool on Air Pollution Dataset”, International Journal of Computer Applications, (0975-8887) Vol. 168- No.13, June 2017.

[4] Bharat Chaudhari, Manan Parikh, “A Comparative Study of Clustering algorithms Using WEKA tools”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), (2319-4847) Vol. 1- No.2, October 2012.

[5] Harjot Kaur, Er. Prince Verma, “Comparative WEKA Analysis of Clustering Algorithm’s”, I.J. Information Technology and Computer Science, (56-67) No.8, August 2017.

[6] Akanksha Mahajan, Er. Neena Madan, “Survey of K means Clustering and Hierarchical Clustering for Road Accident Analysis”, International Reearch Journal of Engineering and Technology, (2395-0056) Vol. 4- No.6, June 2017.

[7] G. Shiyamala Gowri, Ramasamy Thulasiram and Mahindra Amit Baburao, “Academical Data Mining Application for Estimating Students Performance in WEKA Environment”, ICSET, Vol. 14, 2017.

[8] Jiawei Han, Micheline Kamber, Jian Pei, “Data Mining Concepts and Techniques”, 3rd Edition.

Authors Profile

Dr.V.Lakshmi Praba is currently working as Assistant Professor in the Research Department of Computer Science of Rani Anna Government College for Women, Tirunelveli. She has more than 15 years ofresearch experience. She has published more than 25 papers in International Journals. She has authored 2

books. She also has received grant from UGC-MRP. Her areas of interest include Image Processing, Data
Mining and Network Security.

Ms.M.A.Saira Banu is an M.Phil Scholar studying in Research Department of Computer Science of Rani Anna Government College for Women, Tirunelveli. She did her PG in the same Institution and passed with Distinction. Her areas of Interest include Data mining and Machine Learning.