NJ: NUTA

# Comparative Analysis of K-Means and Enhanced K-Means Algorithms for Clustering

**Manoj Pokharel[1], Jagdish Bhatta[1*], Nawaraj Paudel[1]**
[1]Central Department of Computer Science and IT, Tribhuvan University
*Corresponding Author Email: jagdish@cdcsit.edu.np

DOI: 10.3126/nutaj.v8i1-2.44044

## Abstract

*Clustering in data mining is a way of organizing a set of objects in such a way that the objects in same bunch are more comparable and relevant to each other than to those objects in other bunches. In the modern information retrieval system, clustering algorithms are better if they result high quality clusters in efficient time. This study includes analysis of clustering algorithms k-means and enhanced k-means algorithm over the wholesale customers and wine data sets respectively. In this research, the enhanced k-means algorithm is found to be 5% faster for wholesale customers dataset for 4 clusters and 49%, 38% faster when the clusters size is increased to 8 and 13 respectively. The wholesale customers dataset when classified with 18 clusters the speedup was seen to be 29%. Similarly, in the case of wine dataset, the speed up is seen to be 10%, 30%, 49%, and 41% for 3, 8, 13 and 18 clusters respectively. Both of the algorithms are found very similar in terms of the clustering accuracy.*

**Keywords**: *Clustering, Data Mining, Elbow Method, Enhanced K-means, K-means*

## Introduction

Clustering is a craftsmanship of isolating the data items into a number of comparative bunches. The data items within the same bunches are more comparable to other data items within the same bunch instead of the items in other bunches (Ogunleye, 2021). The aim of clustering is to isolate those bunches with comparative characteristics and relegate them into clusters. The data elements inside each bunch ought to show higher degree of likeness whereas the similitude among elements of the comparative clusters ought to be as little as conceivable (Kaushik, 2016). Clustering is an unsupervised learning method. It is a process to find hidden features and patterns in set of examples. K-means algorithm is the most familiar clustering algorithm widely used in most of data science and machine learning applications. At the beginning, the choice of centroids is set randomly to some k, which are used to produce an initial randomized set of clusters. Afterwards, each centroid is set to the arithmetic mean of the cluster it defines. The strategy is rehashed until the values of centroids become stable. The ultimate centroids create last clustering of the data. (Soder, 2008). The enhanced k-means algorithm is an improvement over the conventional k-means algorithm with the use of efficient data structures.

The enhanced k-means algorithm employs the used of red-black trees and min heap over the basic data structures in traditional k-means algorithm (Yuan, 2021). Initially it starts as in traditional k-means algorithm and computes the Euclidean distance and assigns nearest cluster based on these values. Thereafter, the red-black tree is used to insert the labels of objects as keys. The min-heap corresponds to each key with corresponding value. In next iteration, if an object moves from one cluster to another, then only the distance between the clusters and object are computed and also the necessary changes in the data structures are done. The stopping condition is similar for both of the algorithms. The enhanced k-means algorithm is quite optimistic on reducing the number of unnecessary computations for achievement of better time complexity (Kumar, Puran, & Dhar, 2011).

During each iteration of the k-means algorithm, the computation of distance between each data object and each cluster is done. Based on this, if 100000 objects need to be partitioned into 100 clusters then the algorithm will have to calculate distance for 100000 × 100 × 100 $=10^9$ times. A meticulous analysis of working mechanism of the k-means algorithm derives that there is no ought to over compute the distances between each data object and each cluster repeatedly. In an iteration, only one data object can be moved from one cluster to another cluster so that other k-2 clusters being unchanged. This leads to no need of calculation of distances between these k-2 clusters and data objects in the next iteration. However, the traditional k-means algorithm still makes these unnecessary calculations.

In this context, the main objective of this study is implementation and analysis of the traditional K-Means algorithm and enhanced k-means algorithm by using various data sets and analyze their efficiency. The contribution of this work includes building an efficient framework for selection of clustering algorithms.

## Materials and Methods

In this study two datasets namely 'Wholesale customers Data Set (Margarida, 2014) and 'Wine Data Set' (Aeberhard, 1991), available in the UCI: Repository of Machine Learning are used to evaluate the performance of the k-means algorithm and the enhanced k-means algorithm. The wholesale customers dataset has 440 instances with 8 attributes while the wine dataset has 178 instances with 13 attributes. Among the two data sets the 'Whole customers Data Set' dataset contained a categorical attribute 'Region' with the value 'Libson', 'Oporto', 'Other'. Hence to perform clustering the attribute sex was enumerated as Libson: 1 and Oporto: 2 and Other: 3. Both the data sets contained no any missing values. Similarly, the same dataset contained a categorical attribute 'Channel' with the values 'Horeca' and 'Retail' and they were also enumerated as 'Horeca: 1' and 'Retail: 2'.

## K-Means Clustering Algorithm

The K-means clustering algorithm partitions 'n' objects into 'k' clusters. Each object in 'n' belongs to the cluster with the nearest mean value. This strategy has capacity to produces precisely 'k' distinctive clusters with most prominent conceivable distinction. Since the best number of clusters 'k' is not known in advance, it must be computed from the data. The main objective of K-means clustering approach is to minimize total intra-cluster variance or the squared error function (Sayad, n.d.). The squared error function (Sayad, n.d.) is given by the equation below

$$\sum_{j=1}\sum_{i=1}\left\|x_i^{(j)} - c_j\right.$$

Where, J is the squared error function (objective function) to be minimized, k is the number of clusters, n is the number of cases, $x_i$ is the $i^{th}$ case, $c_j$ is the centroid for cluster j and $\left\|x_i^{(j)} - c_j\right\|$ is the distance function.

### Enhanced K-Means Clustering Algorithm

Since the k-means algorithm includes unnecessary computation of distances between data objects, it leads to extra computational costs. These overheads are reduced in enhanced k-means (Yuan, 2021). For this, the idea is to use the min-heap and red-black trees to compute the distance of only those objects that move from one cluster to another. Since, the traditional k-means algorithm computes the distance of every object in each iteration, the enhanced algorithm saves considerable amount of time (Kumar, Puran, & Dhar, 2011).

### Selection of k using Elbow method

The optimal value of k, for both of the dataset is obtained using the elbow method. The optimal k-value guides the clustering model towards better accuracy. The elbow method plots the value of cost function produced by different values of k and the elbow of the plotted curve is selected as a number of clusters to use. In general, the elbow method determines the within-cluster sum of square value for various k values and helps determine the most appropriate k value (M A Syakur, 2018).

### Implementation

The algorithms are implemented in the python programming language. Pandas is used to import dataset, obtain and work with data frames, obtain statistical description of dataset. Seaborn is used for data visualization. Sklearn library is used to implement elbow method and clustering. Matplotlib is used for plotting graphs.

## Results

### Optimal Selection of k

Upon the implementation of the elbow method on both the data sets as stated in previous chapter following plots are obtained.

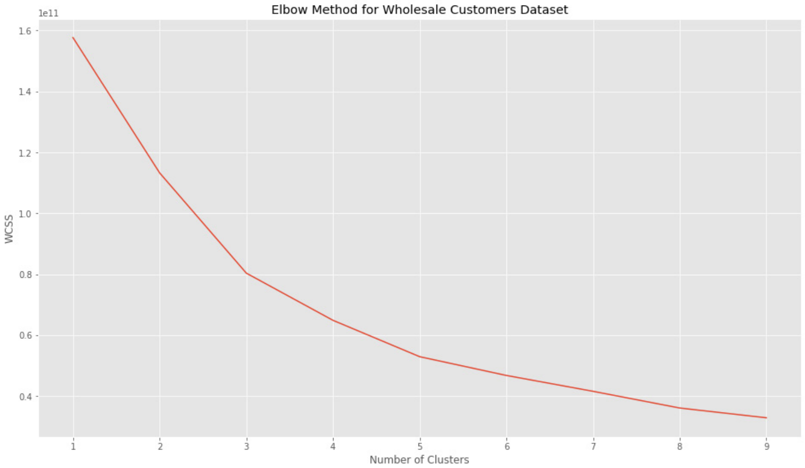

**Figure 1: Elbow plot for wholesale customers dataset**

From the plot in figure 1, it is observed that the optimal number of clusters for the wholesale customers dataset is four. The keyword optimal in the wholesale customers dataset references the value four. Similarly, the elbow plot for the wine dataset is obtained as shown below.

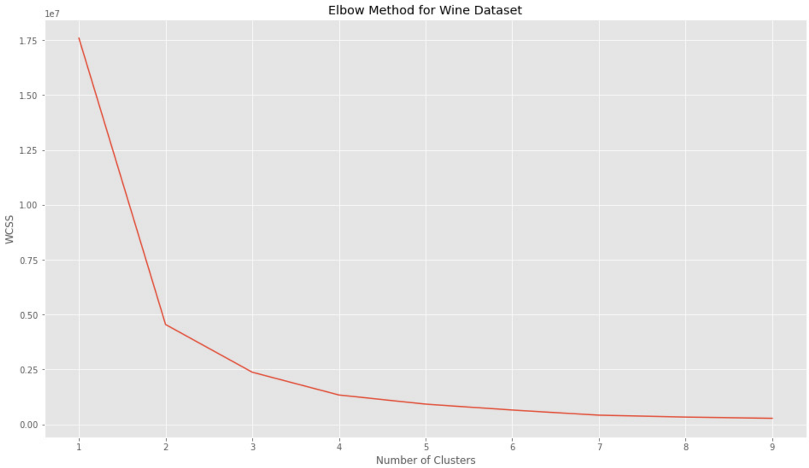

**Figure 2: Elbow plot for wine dataset**

### Time Performance Analysis

After the optimal values are obtained for each data set, both the k-means and enhanced k-means algorithm are run with varying k values for the comparative analysis. The time stamps required for obtaining each of the clusters, for varying value of k are obtained for both wholesale customers and wine dataset. The keyword 'Optimal' refers to value 4 for both the wholesale customers and wine dataset. The numerical estimates are tabulated as follows.

**Table 1: Clustering Time in Seconds**

| K-value | K = Optimal | | K = 8 | | K = 13 | | K = 18 | |
|---|---|---|---|---|---|---|---|---|
| Data set<br>Algorithm | Wholesale customers | Wine | Wholesale customers | Wine | Wholesale customers | Wine | Wholesale customers | Wine |
| K-Means Time | 1.40 | 1.68 | 2.21 | 2.09 | 2.89 | 2.90 | 3.56 | 2.36 |
| Enhanced K-Means Time | 1.25 | 1.21 | 1.82 | 1.52 | 2.2 | 1.86 | 2.73 | 2.03 |

From the table the results imply that the enhanced algorithm has little improvement over datasets with fewer number of features, however there is noticeable improvement in clustering time taken to classify datasets with large number of features as well as large number of instances.

The above results are interpreted by the following figures.

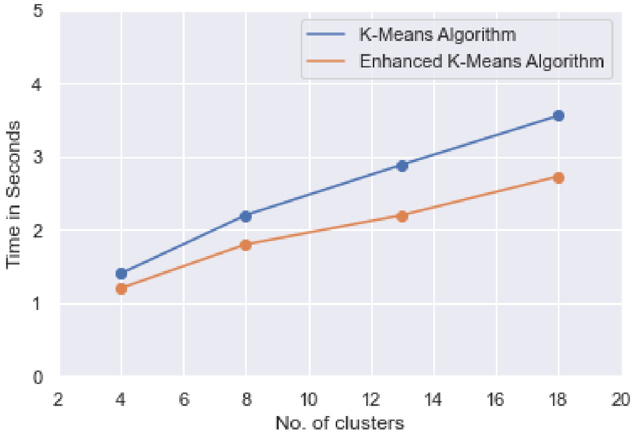

**Figure 3: Time taken by k-means and enhanced k-means for clustering wholesale customers dataset**

From figure 3 it can be clearly depicted that the improved k-means algorithm is far better than the traditional k-means algorithm. The time complexity seems to be far lower in case of the improved algorithm.

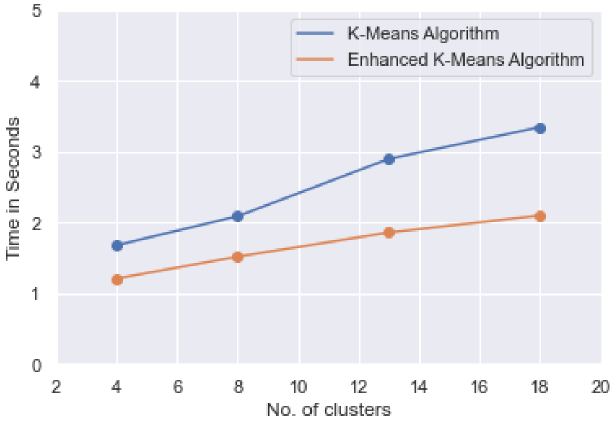

**Figure 4: Time taken by k-means and enhanced k-means for clustering wine dataset**

Similarly, in figure 4, the improved algorithm seems better for clustering purpose. The time differences between the algorithms provides clearer image of the performance of algorithms.

### Cluster Size Analysis

The below clustering graph is obtained as a result of clustering wholesale customers dataset with four centroids with aid of the elbow method by k-means algorithm.

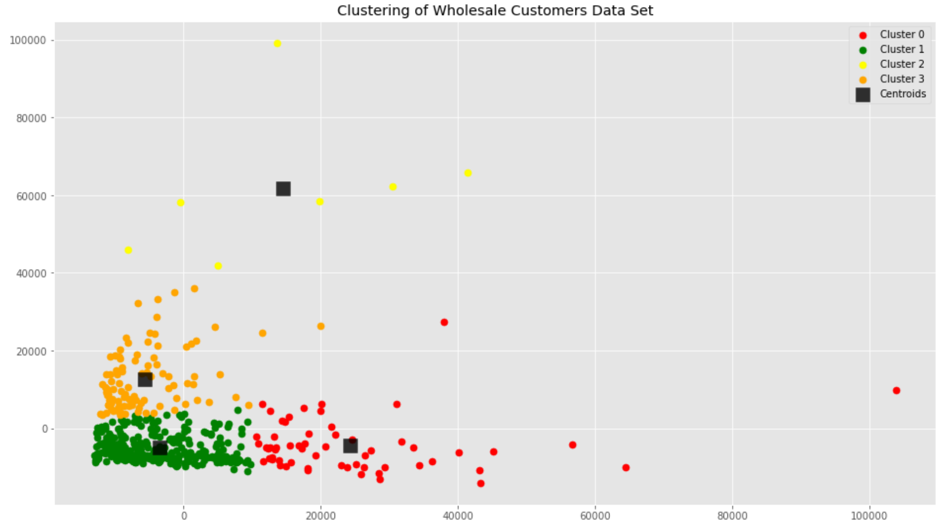

**Figure 5: Clustering of Wholesale customers dataset with k = 4 with elbow method.**

The further results are visualized as follows:

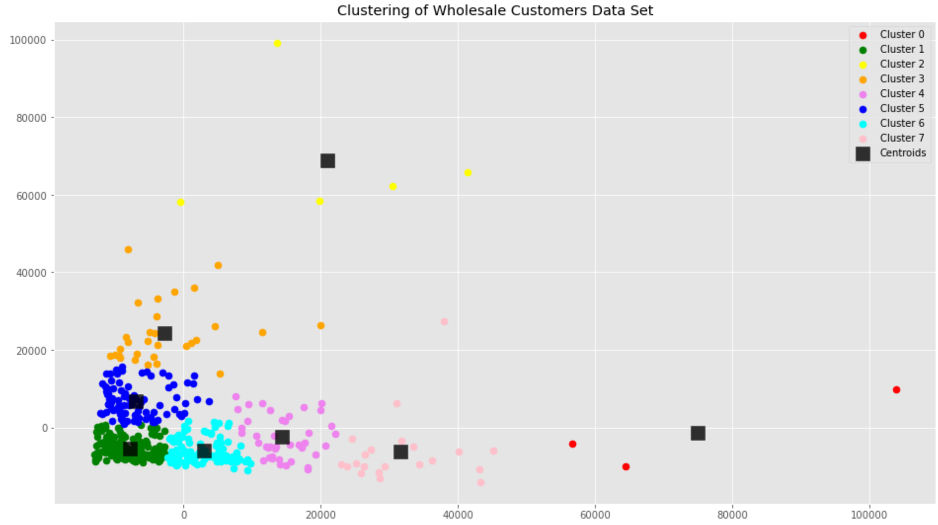

**Figure 6: Clustering of Wholesale customers dataset with k = 8 and elbow method.**

Insight on the clusters of the wine dataset can be visualized through the following figures. The elbow method was implemented in all the cases and then the clustering algorithm was applied.

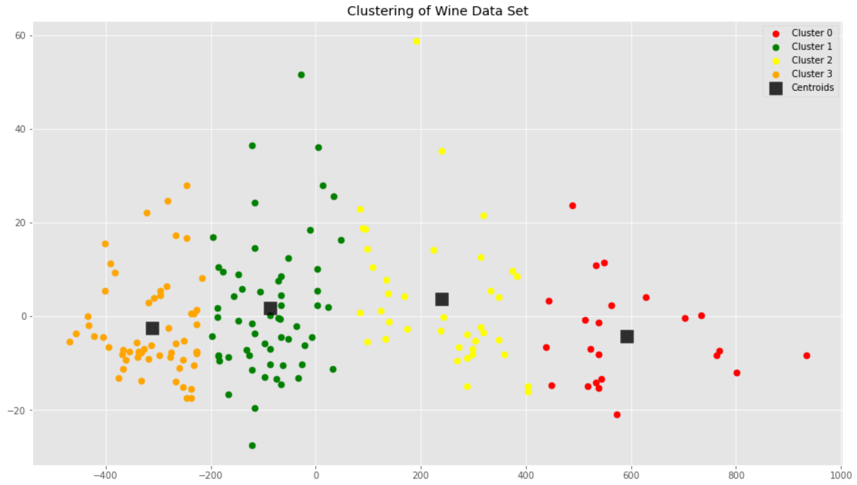

**Figure 7: Clusters in wine dataset with k =4**

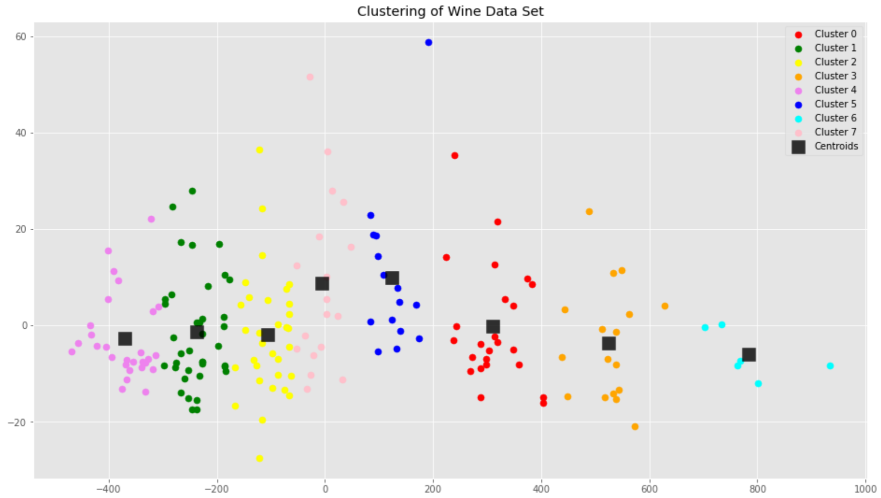

**Figure 8: Clusters in wine dataset with k = 8**

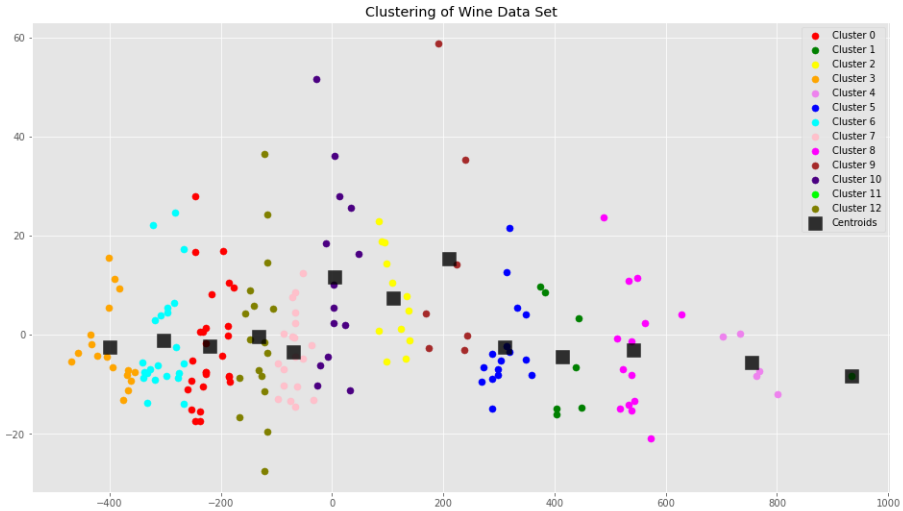

**Figure 9: Clusters in wine dataset with k = 13**

## Discussions

Based on above findings, it can be concluded that the enhanced k-means algorithm shows significant improvement in terms of the time complexity. Both the algorithms when employed with elbow method show that the initial choices for centroids are far better than ones performed without the aid of elbow method. In case of the wholesale customers dataset the largest speedup difference is found to be 49% for 13 clusters and for wine dataset it was found to be 29% for 18 clusters. There seems no any discrepancy in between the clusters produced by both the approaches.

## Conclusion

The two algorithms, traditional k-means and improved k-means are used for the comparative study for clustering purposes. The datasets Wholesale customers and Wine with varying attributes and instances are chosen for study. The experimental results illustrate that the improved k-means algorithm outperforms than the traditional algorithm in terms of the time complexity. Both the algorithms in most cases have produced similar results in terms of the clustering accuracy and hence can be considered equally accurate. This speedup is the direct advantage of using the additional data structures to store the intermediate results and dynamically relocate them whenever required. However, these additional data structures pose a severe drawback; the improved algorithm is lagging behind in terms of space complexity.

## References

Aeberhard, S. (1991, July 01). *UCI Machine Learning Repository*. (National Science Foundation) Retrieved June 1, 2021, from https://archive.ics.uci.edu/ml/datasets/wine

Kaushik, S. (2016, November 5). *Analytics Vidhya*. Retrieved June 1, 2021, from https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

Kumar, R., Puran, R., & Dhar, J. (2011). Enhanced K-Means Clustering Algorithm Using Red Black Tree and Min-Heap. *International Journal of Innovation Management and Technology, II*(1), 49-54.

Kumar, S. (2017, October 21). *Github*. Retrieved June 15, 2021, from https://github.com/siddheshk/Faster-Kmeans/blob/master/Code/enhancedKmeans.py

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S. & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, 1-6.

Margarida, G. (2014, March 13). *UCI Machine Learning Repository*. (National Science Foundation) Retrieved June 1, 2021, from https://archive.ics.uci.edu/ml/datasets/wholesale+customers

Ogunleye, J. O. (2021). *The Concept of Data Mining.* IntechOpen Book Series.

Sayad, S. (n.d.). *K-Means Clustering*. Retrieved 2 5, 2021, from https://www.saedsayad.com/clustering_kmeans.htm

Soder, O. (2008, May 29). *Phonetic Sciences*. Retrieved June 01, 2021, from https://www.fon.hum.uva.nl/praat/manual/k-means_clustering_1__How_does_k-means_clustering_work_.html

Yuan, R. (2021). An improved K-means Clustering Algorithm for Global Earthquake Catalogs and Earthquake Magnitude Prediction. *Journal of Seismology*, 1005–1020.