DOI: http://doi.org/10.4038/sljastats.v21i3.8062

Sri Lankan Journal of Applied Statistics, Vol (21-3)

Application of K-Means and Fuzzy K-Means to Rice Dataset in Sierra Leone

Bangura R. M.[1*], Johnson S. D.[2], Mbulayi O[3]

[1]Biometric Unit, Sierra Leone Agricultural Research Institute, Freetown, Sierra Leone,
[2]Agronomy Department, Rokupr Agricultural Research Centre, Kambia district, Sierra Leone,
[3]Mathematics and Computer Science, University of Kinshasa, Democratic Republic of Congo,
[*]Corresponding Author: mannahray@yahoo.com

Received: 26th October 2020 / Revised: 10th December 2020 / Published: 31st December 2020

ABSTRACT

As k-means and fuzzy k-means are regarded as unsupervised dimensional reduction learning techniques, we present an application of this technique from the Agronomic data collected in 2015 to demonstrate the efficiency of fuzzy k means over k means of eight different types of rice varieties in Sierra Leone. Also, we identified different rice varieties as outliers from the silhouette clusters (segment).

Keywords: Clusters, K-means, Outliers

1. Introduction

Clustering of k-means forms part of the main topics in machine learning. Machine learning is widely used in physical or natural sciences, as it helps to get an intuition about the structure and pattern of the data. Clustering identifies similar or different subgroups in a given dataset (Hartigan and Wong, 1976). The homogeneity identifies similar clusters according to their data points.

K-means is a clustering of n-observation into partitions. The k-means method uses a prototype (centroids) to represents clusters by optimizing the squared error function (Bradley, et al, 1998). It is considered an iterative algorithm because the data, is partitioned into clusters (subgroups), thereby making the data points as similar (homogeneous) as possible (Bradley & Fayyad, 1998).

Fuzzy k-means, on the other hand, is regarded as a soft (flexible) method than k-means because each point can belong to two centroids with different quality (Bradley & Fayyad, 1998). Fuzzy K-means is more statistically formalized and discovers soft clusters, where a particular point can belong to more than one cluster with a certain probability

In this paper, we present an application of clustering analysis to Agronomy, with eight varieties. The paper is divided into five parts. In the next section, a summary of the dataset is presented in the methodology, followed by results and discussion. We give our conclusion in section four.

2. Methodology

This data was collected in the year 2015. From this dataset, we only considered eight rice varieties which are; Nerica 1, Nerica 3, Nerica 6, Rok 3, Rok 16, Rok 17 and Pa Gbonko. Also, there are nine independent variables, which are; panicle, tillers, plant height, number of filled grains, 50% days to flowering, days to maturity, and grain yield. We consider one independent variable, days to maturity. We chose one independent variable because, the number of observation points is 384, which is sufficient for analysis. Statistical Analysis Systems (SAS), ARiS and XLSTAT were used to generate tables and graphics.

3. Results and Discussion

Here we present a cluster analysis of eight different types of rice varieties in Sierra Leone. We aim at segmenting these varieties into subgroups to demonstrate similarities among them. From table 1, it shows that all the clusters are closer to '1'and also a mean width of 0.706 implies a good choice as it is going towards '1'.

Table 1: Average silhouette width of each cluster

Clusters	Value
Cluster 1	0.623
Cluster 2	0.720
Cluster 3	0.674
Cluster 4	0.742
Cluster 5	-0.072
Mean width	0.706

Table 2 shows a summary of cluster for days of maturity. It is seen that, cluster 5 has the lowest size, while cluster 4 with the highest. It could also be seen that the minimum and maximum distances to the centroid is zero.

Table 2: Cluster summary

Cluster Type	Size	Average distance to centroid
Cl.1	33	0.000
Cl.2	105	0.000
Cl.3	91	0.000
Cl.4	153	0.000
Cl.5	2	0.003

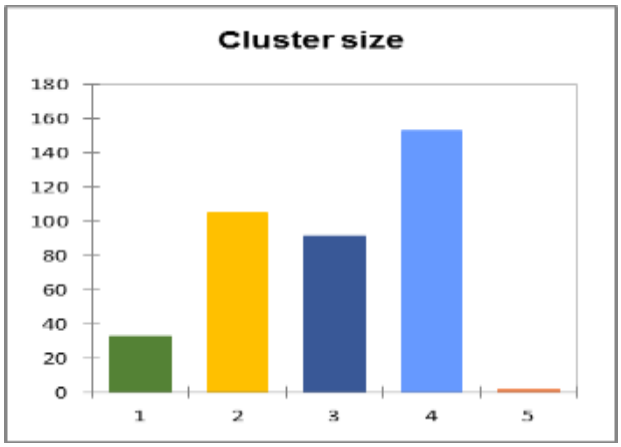

Figure 1: Cluster size for days to maturity

Table 3 shows the respective centers for both days to maturity clusters. It could be seen that days to maturity have better means are they are very much closer to '1'. Also, table 4 shows a summary of statistics for days to maturity.

Table 3: Centers for days to maturity

Cltrs.	CL. 1	CL.2	CL.3	CL.4	CL.5
Days to maturity	0.999	1	1	1	0.983

Table 4: Summary statistics (Items)

Var	Obs	Ob	ObWO.	Min	Max	Mean	Std.
Days to maturity	384	0	384	12	145	117.9	14.9

Var-Variable, Obs-Observation, Ob-Observation with missing data, ObWO- Observation without missing data, Min-Minimum, Max-Maximum and Std.-Standard deviation

Wilks' lambda statistic was used to test the differences between means of identified groups (clusters) of subjects on a combination of dependent variables (Nicola; 2000). In table 5, 24 iterations were used for days to maturity and the Wilks' lambda is given as 0.018, which is significant at 0.05.

Table 5: Summary of clusters

No.	Iterat.	criteria	B-c	wcv	wLt	MW
5	24	0.006	0.001	0	0.018	0.706

N^0: No. of clusters, WCV: Wth class variance, wLt: Wilks' Lambda test, Mw: Mean width and Bc: Mean width

A silhouette is a method of validating the consistency of different clusters. It also shows how well one cluster matched as compared to other clusters. In this instance, we consider eight different varieties taking from a huge dataset from two locations in Sierra Leone (Rokupr and Bo). It is also used to measure the degree of separation between clusters. From figure 2, we have five clusters. Cluster 1, 2, 3 and 4 seems to have high values but with different width. Cluster five (5) has few clusters and one outlier. Also, Nerica 3 is an outlier in clusters 1 and 3. ROK16 and Nerica1 are also outliers in cluster 2 and cluster 5 respectively. This implies that the clusters are appropriate because most of the varieties have high values within their clusters.

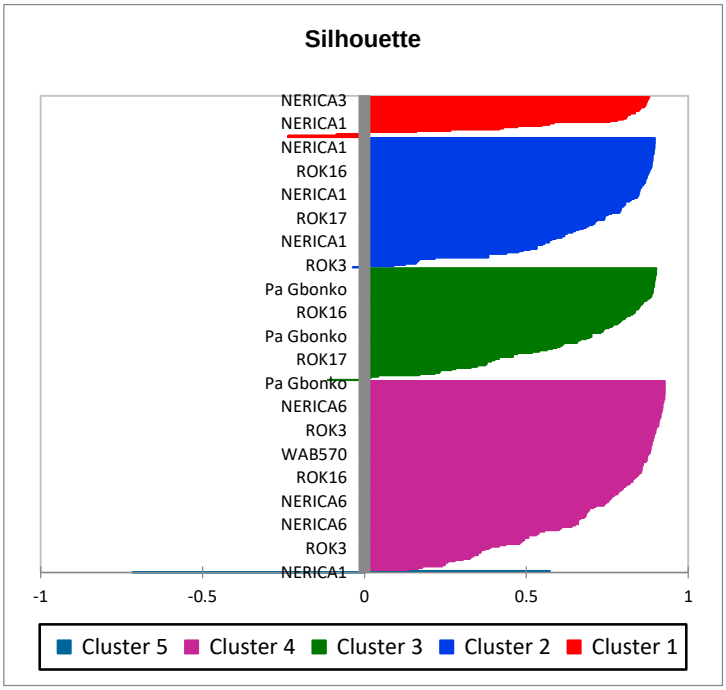

Figure 2: Silhouette for days to maturity

4. Conclusion

We presented both k-means and fuzzy k-means clustering for the variable, days to maturity. We had demonstrated that fuzzy k-means is more accurate than k-means because of their flexibility. From the silhouette, we deduce that, Nerica 3, is an outlier in clusters 1 and 3, while ROK 16 and Nerica 1 are also outliers in cluster 2 and cluster 5 respectively. With 24 iterations for days to maturity, the Wilks' lambda showed a highly significant difference at $p < 0.05$.

Acknowledgements

We appreciate the Sierra Leone Agricultural Research Institute and Rokupr Agricultural Research Center for providing data for this work.

References

1. Amorim, R. C. and Mirkin, B., (2012). Metric, Feature Weighting and Anomalous Cluster Initialization in k-Means Clustering. *Pattern Recognition. 45 (3):* 1061–1075.

2. Bradley, P., Paul, S., Fayyad, U. and Usama, M. (1998). Refining Initial Points for k-Means Clustering. *Proceedings of the Fifteenth International Conference on Machine Learning*.

3. Ding, C. and Xiaofeng, H. (2004). K-means clustering via Principal Component Analysis. *Proceedings of International Conference on Machine Learning (ICML 2004): 225–232.*

4. Ding, C. and He, X. (2004). K-means clustering via principal component analysis. In: ICML, p. 29. Fisher, W. D. (1958). "On Grouping for Maximum Homogeneity". In: *Journal of the American Statistical Association 284*, pp. 789–798.

5. Ding, C. and He, X. (2004). Linearized cluster assignment via spectral ordering. *Proceedings of the International Conference on Machine Learning.*

6. Drake and Jonathan (2012). *Accelerated k-means with adaptive distance bounds*. The 5th NIPS Workshop on Optimization for Machine Learning, OPT2012

7. Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1*, 183–187.