Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 11, November -2017

Clustering Analysis for Appropriate Crop Prediction using Hierarchical, Fuzzy C-Means, K-Means and Model based Techniques

Dr Madhavi Gudavalli[1], Vidyasree P[2], S Viswanadha Raju[3]

[1]*Assistant Professor, Department of CSE, UCEN JNTUK, Narasaropet, Andhra Pradesh, India*
[2]*Assistant Professor, Department of CSE, Stanley Engineering College and Research Scholar of JNTUH, Hyderabad, Telangana, India,*
[3]*Professor, Department of CSE, JNTUH CEJ, Hyderabad, Telangana, India*

Abstract - *Data mining is a specific field of computer and information science with substantial point of view of knowledge discovery from expansive database or dataset. Different types of techniques are accessible under data mining and clustering or the unsupervised learning specifically. Clustering is a division of information into identical groups each comparative gathering is known as a cluster. Object in a group are comparative or near each other. Various methodologies are designed to implement the clustering technique very effectively. This paper represents a study on different clustering techniques that are incorporated on the seed data sets to enhance the clustering approach based on the various parameters like area, perimeter, compactness, length and width of the kernel, asymmetric coefficient and length of the kernel groove.*

Keywords - *Cluster, Fuzzy C-Means, Hierarchical, K-Means, and Model Based Clustering.*

I. INTRODUCTION

With the progress of present day systems used for scientific information gathering, tremendous measure of data are getting amassed at various databases. The data banks are growing so rapidly that it is has wound up hard to evacuate noteworthy information by using regular database methodology. Clustering is standout technique among the most extensively used techniques for exploratory data examination. Clustering is the route toward hoarding the information into classes or groups, with the goal that things inside a gathering have higher similar characteristic stood out from the things that have a place with different groups. Classification and clustering techniques plays a crucial role in the agriculture field. Agriculture field is considered to be the backbone of India and also the oldest profession adopted by the man kind. Many issues like pests, unbalanced climate conditions and other agricultural conditions were affecting the agricultural output. Utilizing the recent technologies [9] like computers and highly designed equipment in the agriculture, creates the pleasant scenario, upgrading the crop production and also comes forward in aide to one of the oldest inventions of human race. Now a day's mining the agricultural database is considered as the major concern. Worried over low rural development rate and low levels of homestead wage, the Union agrarian service is working towards an arrangement to present cluster based cultivating in the nation. Different tools also have been emerged for evaluating, justifying, upgrading and modifying the existing agricultural expert systems thus making them more useful in their intended purposes. The cluster based approach is used for framing a merged cultivable holding committed to particular nourishment grains, vegetables, foods grown from the ground cultivation crops. This helps in enhancing the agricultural management, predicting and suggesting solutions for its problems. In the process of grouping the similar objects into one class or cluster will have the higher degree of homogeneity attribute in comparison to one another but shows the high dissimilarity ratio with the objects in the different clusters. It is also comparable with the classes in the object oriented programming paradigm. The variation between the class and the cluster is that the object in the class have the identical properties where as the objects in the clusters are almost similar with the other nearby objects. There are various clustering techniques evolved and those are organized into the following categories as partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, methods for high-dimensional data and constraint-based clustering. This paper limits the discussion to hierarchical clustering, fuzzy c means clustering, K-Means clustering and model based clustering only because as these are the extensively used data mining methods in the area of agriculture and allied science.

KA Abdul Nazeer et.al. [1] Presented an updated k-means that incorporates arranging the data set and allocating arranged data set into "k" number of sets which, brought about better beginning centroids thusly upgrading the accuracy of this calculation. The calculation focalizes speedier stood out from customary calculation of K-Means. The fundamental disadvantage of this calculation is the estimation of k (number of looked for groups) still ought to be given as a information. Kunhui Lin et.al. [2] Showed an overhauled k-means clustering paper that upgraded the beginning concentrations in light of data dimensional density which, attest that these basic centers have the best difference between gatherings. This calculation is executed on the Hadoop stage (MapReduce programming model). This procedure made a difference in upgrading the unfaltering quality of the K-implies grouping. Akanksha Kapoor and Abhishek Singhal [3] did a comparative study of k-Means, K-Means ++ and fuzzy C Means algorithms. Chu-Hsing Lin.et.al. [4] Have done performance evolution by integrating Hadoop HDFS and MapReduce with R language and visualize results on Google

Maps. Omar M. Saad.et.al. [5]. Proposed automatic arrival earth quake detection using Fuzzy possibilistic C Means algorithm. Vikas A.et.al [6] demonstrated a new technique which aids in shadow removal scheme for side scan solar images using Fuzzy C Means algorithm. Ye Xuanzuo.et.al. [7] Developed an algorithm for automatic recognition of cluster centers based on local density clustering. Soumi Ghosh et.al. [8] Demonstrated a comparative report between K-Means and Fuzzy C Means (FCM) in view of the quantity of tests and K. The test comes about demonstrate that the K-means calculation is far superior to anything FCM as it requires greater investment in performing fluffy measure estimations which, brings about increment in now is the ideal time many-sided quality and henceforth impacts the outcome. Subsequently, no uncertainty FCM delivers as great outcomes as created by KM close comes about however the time unpredictability is relatively still high. A few research algorithms are been proposed for the enhancement of the agriculture production. This paper deals in this particular aspect by considering the k-Means, Fuzzy C means, Hierarchal and model based algorithms. The rest of the paper organized as follows Section 2 illustrates the proposed methodology, Section 3 demonstrates the experiment and results finally, and Section 4 concludes the paper.

II. METHODOLOGY

In this paper, a study on clustering approach for agriculture to analyze the seed dataset using K-Means, Fuzzy C-Means is done by passing different parameters like area, perimeter, compactness, length and width of the kernel, asymmetric coefficient and length of the kernel groove utilizing R programming.

Step1: Data acquisition

Wheat seeds are vital in cultivating and different varieties of wheat seeds are giving us distinctive yields which must made expanded every year for take care of demand for the general population. The seed dataset gives the names of three distinct seeds as Kama, Rosa, and Canadian. These following wheat seeds give diverse yields in cultivating. The chose data from UCI machine learning repository contains both categorical and continuous characteristic esteems. The dataset contains following properties like area, perimeter, compactness, length, width, asymmetric Coefficient, lk Groove, type of wheat seed.

Step 2: Preprocessing

The selected dataset contains no missing values in the table. Identifying the missing values can be done either by eliminating the record or by replacing the missing values by calculating the mean. In this paper utilizing the R programming aids in finding the missing values with the help of na.omit () function which returns the object with list wise deletion of missing values. After pre-processing the dataset was visualized as in Figure 1.

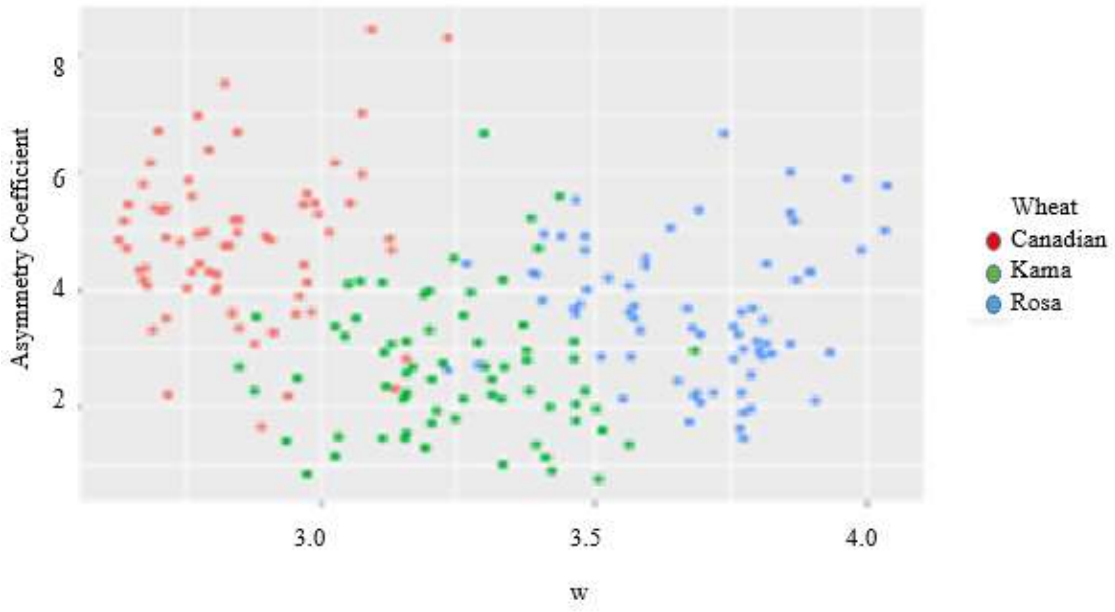

Figure 1: Visualization of pre-processed seed dataset

Step 3: Applying different algorithms on Seed data by using R-programming

The procedure used is shown in Figure 2

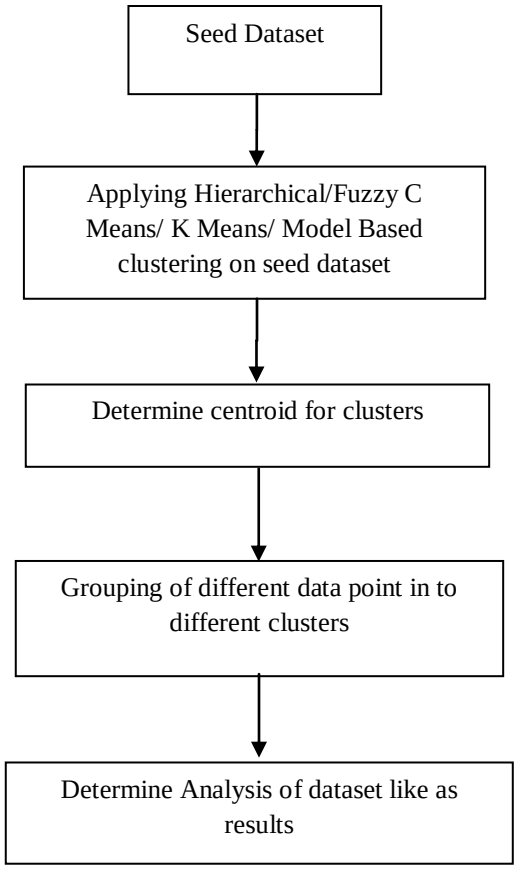

Figure 2: Analyzing different clustering algorithms on seed dataset

(i) Applying K-Means algorithm

K-means is a measurable, unsupervised, non-deterministic, iterative technique for gathering the diverse articles into clusters. The k-means calculation has ended up being to a great degree intense in making clusters in various handy applications in creating regions, for instance, in Bioinformatics, advertise division, PC vision, geostatistics, stargazing and cultivation. It is least complex unsupervised learning calculations known for its speed, ease, and convenience.

(ii) Applying Fuzzy C Means algorithm

Fuzzy C-Means is withal named as soft clustering. The FCM calculation endorses a data point to have a place with every one of the groups with the participation in the middle of 0 and 1. If the data point is more proximate to the cluster focus its enrollment will be more towards a specific group. At the end of each emphasis both the inside and enrolment should be refreshed appropriately. FCM alongside Genetic calculation can be adjusted to cause a recommender framework predicated on homogeneous characteristic measures.

(iii) Applying Hierarchal algorithm

Hierarchical clustering allocate objects into tree like structures, , where cluster can have a data point or representation of low level cluster. Various leveled clustering might be separated into two classes: agglomerative and divisive.

Comparative analysis of Fuzzy C Means and hierarchal algorithms with K-Means algorithm:

K-Means and Fuzzy-C Means Clustering are fundamentally the same as in approaches. The primary distinction is that, in Fuzzy-C Means bunching, each point has a weighting related with a specific group, so a point doesn't sit "in a group" as much as has a powerless or solid relationship to the group, which is controlled by the converse separation to the focal point of the cluster.

Fuzzy C means will tend to run slower than K means, since it's really accomplishing more work. Each point is assessed with each bunch, and more operations are engaged with every assessment. K-Means simply needs to do a separation count, though fluffy c implies necessities to do a full reverse separation weighting.

The proposed work represents K-means clustering algorithm performance and accuracy in distributed environment. The analysis on k-mean and hierarchical algorithm by applying validation measures like area, width, compactness and other parameters. The experimental results shows that kmean algorithm performs better as compared to hierarchical algorithm and takes less time for execution. But on the other hand, hierarchical algorithm provides good quality of results corresponding to k-mean. By applying hierarchal algorithm quality can be achieved whereas k-means helps to faster the clustering process.

(iv) Model based clustering

Model based methodologies expect an assortment of information models and apply greatest probability estimation and Bayes criteria to distinguish most likely model and number of clusters.. In particular, the Mclust( ) work in the mclust packages chooses the ideal model as indicated by Bayesian Information Criterion (BIC) for Expectation-Maximization (EM) introduced by hierarchical clustering for parameterized Gaussian mixture models. One picks the model and number of clusters with the biggest BIC. Finally, find of the centroid of the each cluster so that it helps to make different clusters with different points which is easy to categorize and analysis the wheat data.

Implementing the different clustering algorithms based on R-Programming makes the clustering process very easy on seed dataset to enhance the wheat production.

III. EXPERIMENT AND RESULTS

Cluster analysis divides the data into groups that are meaningful, useful or both. It is also used as a starting point for other purposes of data summarization. The results are tested on UCI machine learning data repository for Seed data analysis. The dataset was analyzed with different clustering algorithms with the help of R data mining tool. Every algorithm has its uniqueness and antithetical behaviour. Each algorithm is analyzed by passing different parameters like area, perimeter, compactness, length and width of the kernel, asymmetric coefficient and length of the kernel groove utilizing R programming. Wheat seeds are vital in cultivating and different varieties of wheat seeds are giving us distinctive yields which must made expanded every year for take care of demand for the general population. Collecting the wheat dataset with the number of instances: 210, number of attributes: 8, missing values: nil, number of continues attributes: 3, number of classes: 3 as depicted in the Figure 3.

	A	B	C	D	E	F	G	H
1	area	perimeter	compactn	length	width	asymmetryCoefficient	lkgroove	wheat
2	15.26	14.84	0.871	5.763	3.312	2.221	5.22	Kama
3	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	Kama
4	14.29	14.09	0.905	5.291	3.337	2.699	4.825	Kama
5	17.63	15.98	0.8673	6.191	3.561	4.076	6.06	Rosa
6	16.84	15.67	0.8623	5.998	3.484	4.675	5.877	Rosa
7	17.26	15.73	0.8763	5.978	3.594	4.539	5.791	Rosa
8	13.07	13.92	0.848	5.472	2.994	5.304	5.395	Canadian
9	13.32	13.94	0.8613	5.541	3.073	7.035	5.44	Canadian
10	13.34	13.95	0.862	5.389	3.074	5.995	5.307	Canadian

Figure 3: The Seed dataset is analyzed based on different parameters

Pre-processing is done on the seed dataset in order to remove the missing values are depicted in the Figure 4.

	A	B	C	D	E	F	G	H
1	area	perimeter	compactn	length	width	asymmetryCoefficient	lkgroove	wheat
2	15.26	14.84	0.871	5.763	3.312	2.221	5.22	Kama
3	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	Kama
4	14.29	14.09	0.905	5.291	3.337	2.699	4.825	Kama
5	17.63	15.98	0.8673	6.191	3.561	4.076	6.06	Rosa
6	16.84	15.67	0.8623	5.998	3.484	4.675	5.877	Rosa
7	17.26	15.73	0.8763	5.978	3.594	4.539	5.791	Rosa
8	13.07	13.92	0.848	5.472	2.994	5.304	5.395	Canadian
9	13.32	13.94	0.8613	5.541	3.073	7.035	5.44	Canadian
10	13.34	13.95	0.862	5.389	3.074	5.995	5.307	Canadian

Figure 4: Dataset after pre-processing using replace missing values.

Seed dataset was analyzed using K-Means, Fuzzy C-Means is done by passing different parameters like area, perimeter, compactness, length and width of the kernel, asymmetric coefficient and length of the kernel groove and type of the seed through utilizing R programming. Different clusters are formed by integrating the two parameters at each time .Figure 5 shows that different varieties of wheat formed into cluster based on area and perimeter of wheat clustering.

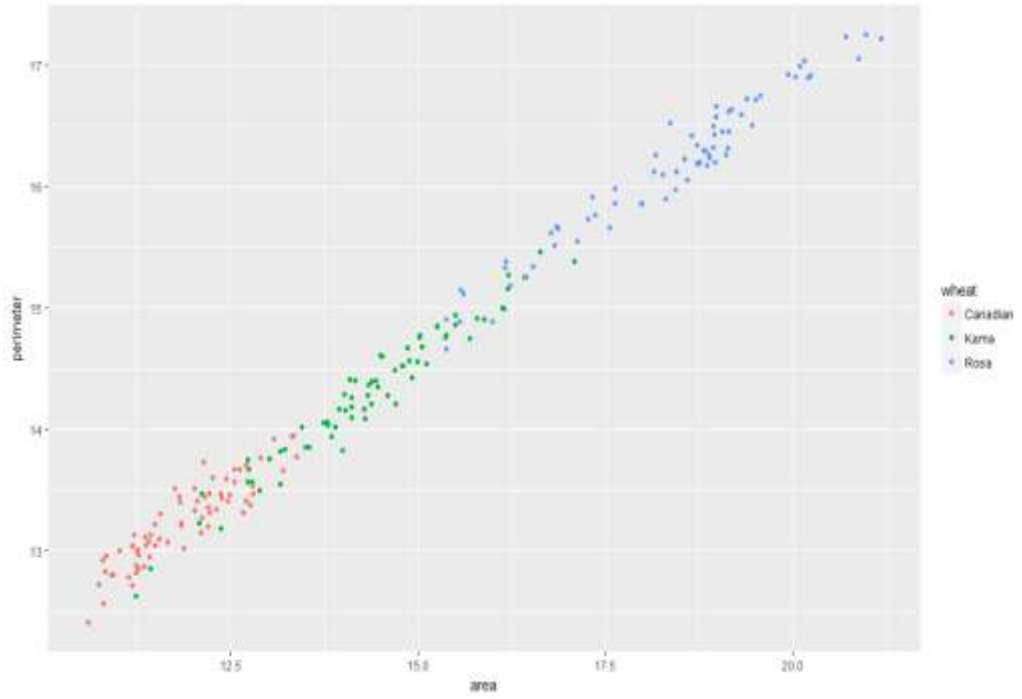

Figure 5: Area and perimeter for wheat clustering

Figure 6 and 7 illustrates the length and width of wheat clustering and compactness and area of wheat clustering.

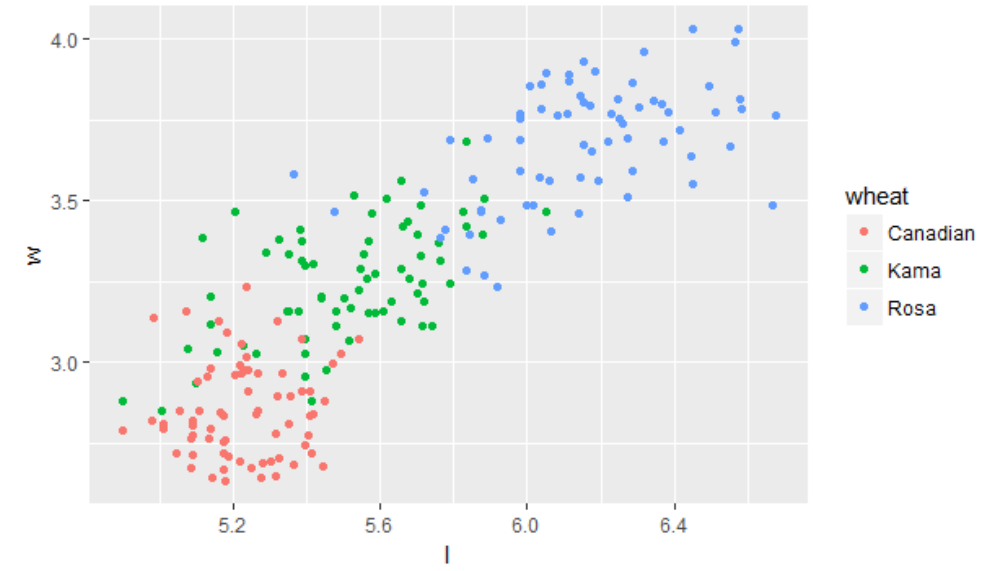

Figure 6: Length, width for wheat clustering

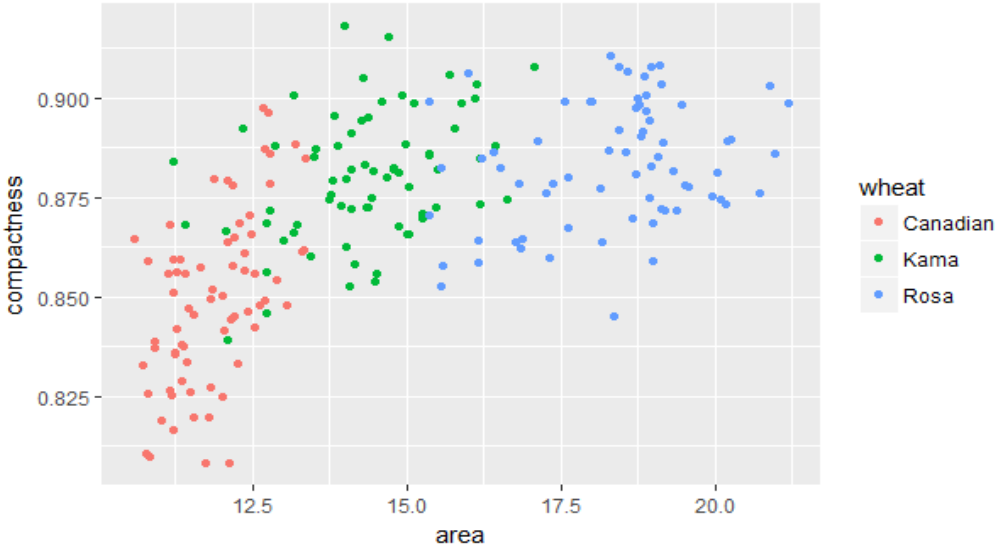

Figure 7: Compactness, area for wheat clustering

Figure 8 and 9 depicts the asymmetric coefficient and area and also lkgroove and area for wheat clustering.

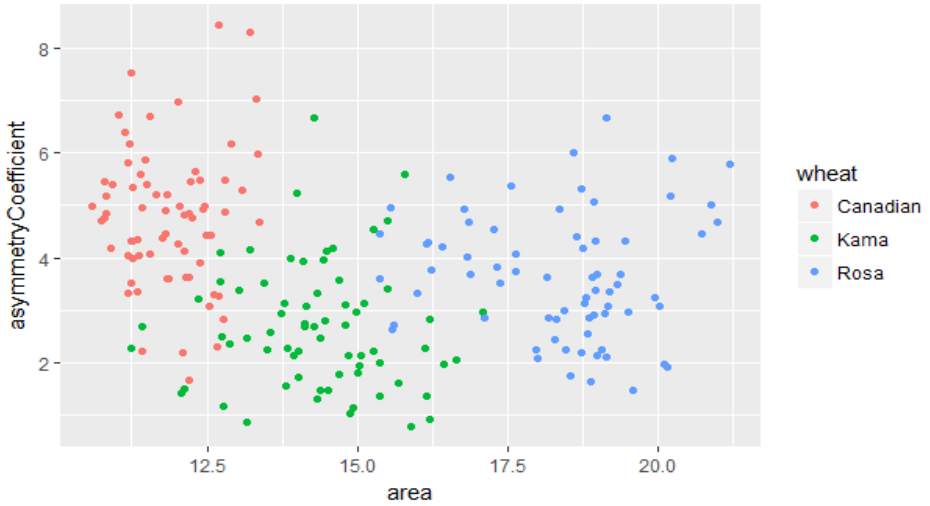

Figure 8: AsymentryCoeifficient,areafor wheat clustering

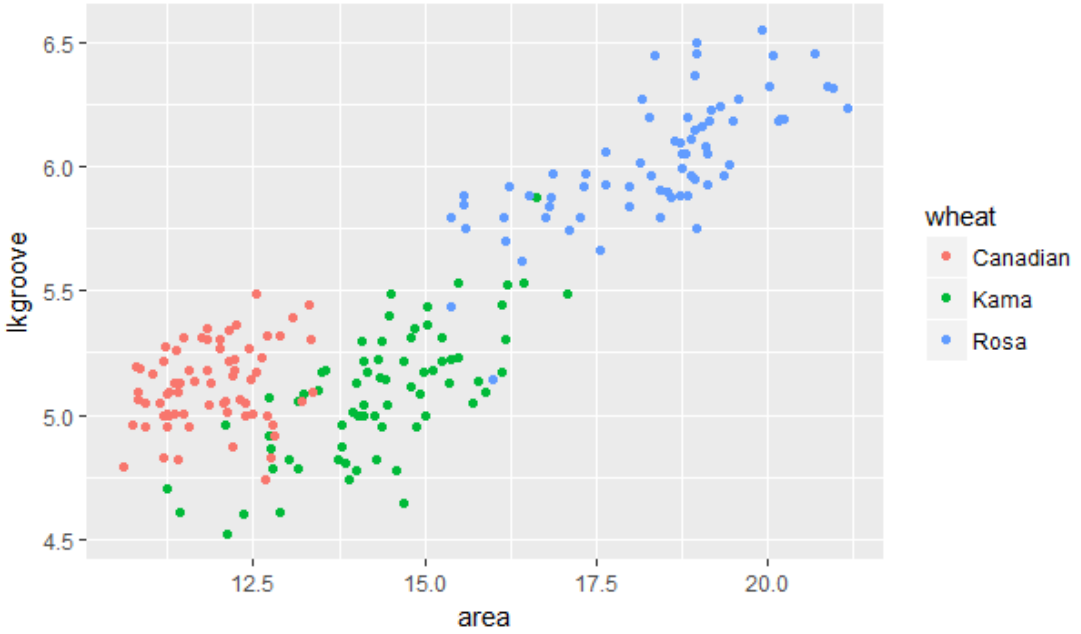

Figure 9: Area,lkgroove for wheat clustering

Various algorithms are implemented by passing the above parameters to forms clusters as follows. Based on the above parameters hierarchal algorithm is implemented by forming into 3 clusters as shown in Figure 10.

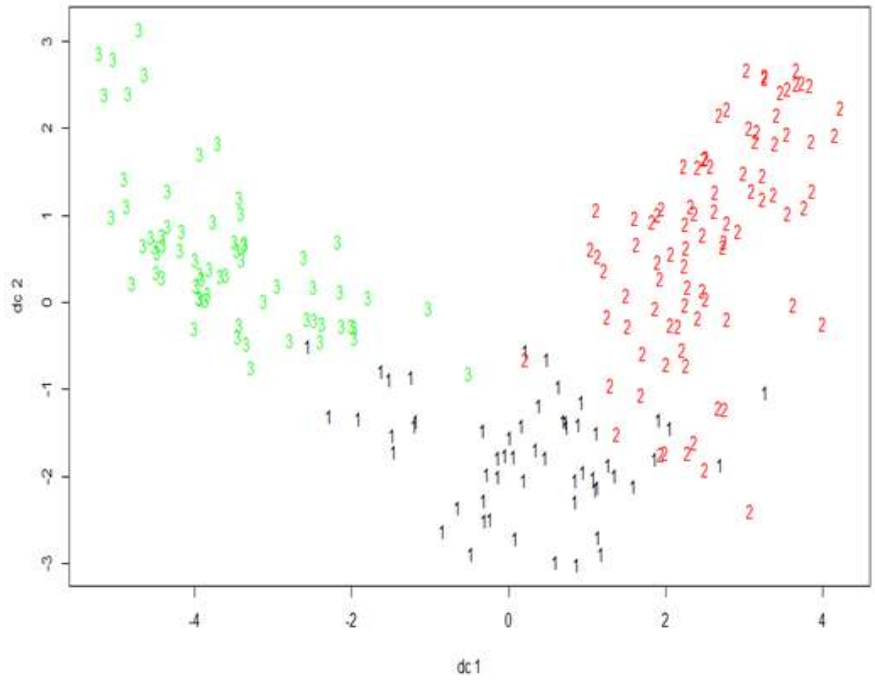

Figure 10: Hierarchical clustering

Fuzzy c means algorithm is implemented on the seed dataset and form 2 clusters whereas k-mean algorithm is implemented as depicted in Figure 11.

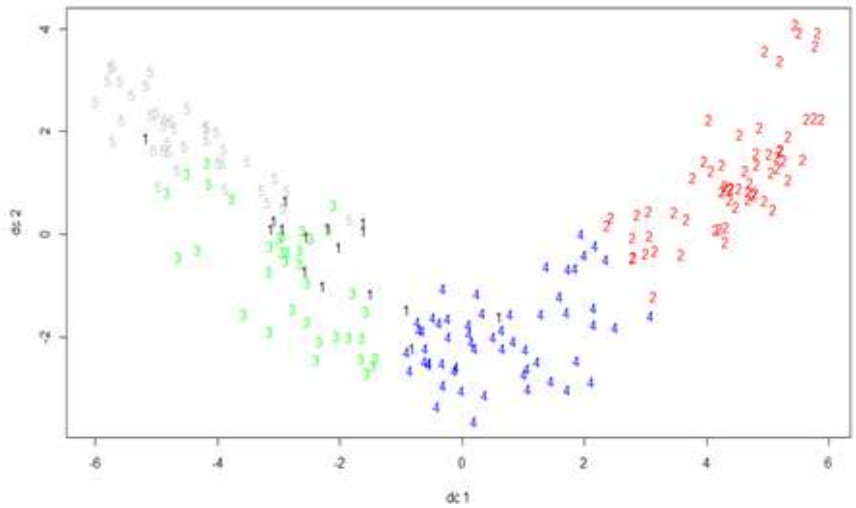

Figure 11: K-Means clustering

Model based clustering is implemented based on BIC, classification; uncertainty and density are shown in the Figures 12, 13, 14 and 15.

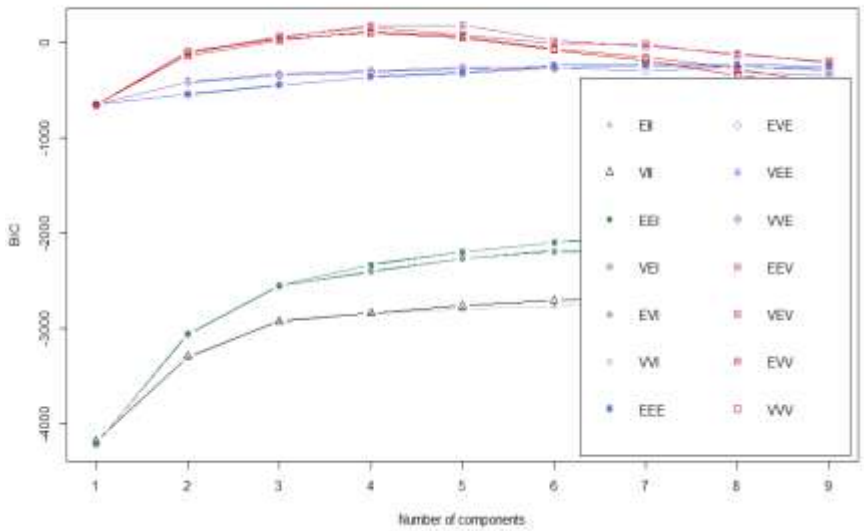

Figure 12: Model based clustering based on BIC

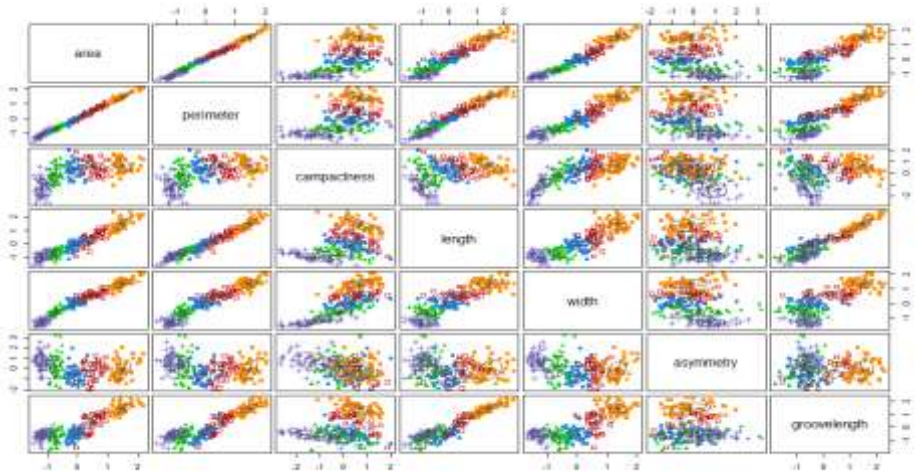

Figure 13: Model based clustering based on classification

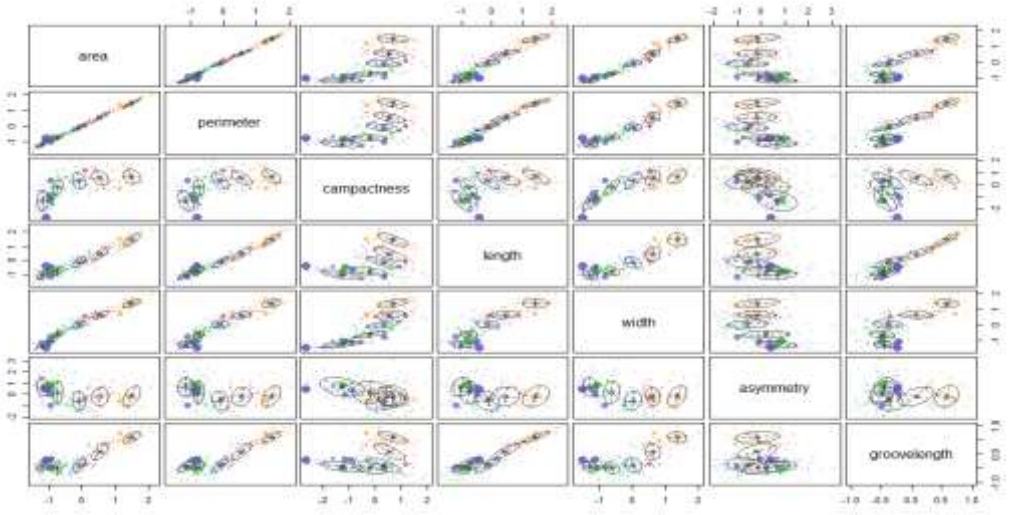

Figure 14: Model based clustering based on Uncertainty

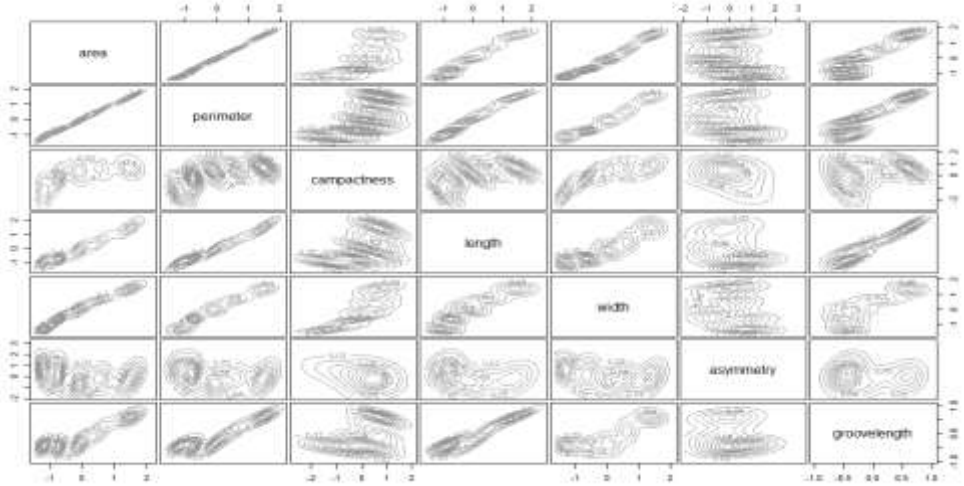

Figure 15: Model based clustering based on density

The above results conclude that using the R-Programming implementation of different algorithms is quiet easy compared to other data mining tools.

IV. CONCLUSION

After a conscientious and detailed study of various clustering techniques, a nascent approach has been proposed in which various parameters are passed into K-Means, Fuzzy C-Means, hierarchical and model based clustering algorithms. This cluster based approach using R-Programming is used for framing a merged cultivable holding committed to particular nourishment grains, vegetables, foods grown from the ground cultivation crops. This helps in enhancing the agricultural management, predicting appropriate crops and suggesting the solutions for agricultural problems. It also concludes that k-mean algorithm is good for large dataset and hierarchical is good for small datasets. The future work makes a comparison among these algorithms on the premise of normalization, by considering normalized and un-normalized data that will result in diverse outcomes.

REFERENCES

[1] K A Abdul Nazeer, S D Madhu Kumar, "Enhancing the k-means clustering algorithm by using a O(n logn) heuristic method for finding better initial centroids" in Second International Conference on Emerging Applications of Information Technology, 2011: pp.261-264.

[2] K.Lin, X.Li, J. Chen, Z. Zhang, "A K-means Clustering with Optimized Initial Center Based on Hadoop Platform" in The 9th International Conference on Computer Science & Education, 2014:pp. 263-266.

[3] Kapoor, A., & Singhal, A. (2017, February). A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms. In Computational Intelligence & Communication Technology (CICT), 2017 3rd International Conference on (pp. 1-6). IEEE.

[4] Lin, C. H., Liu, J. C., & Peng, T. C. (2017, May). Performance evaluation of cluster algorithms for Big Data analysis on cloud. In Applied System Innovation (ICASI), 2017 International Conference on (pp. 1434-1437). IEEE.

[5] Saad, O. M., Shalaby, A., & Sayed, M. S. (2017, June). Automatic arrival time detection for earthquakes based on fuzzy possibilistic C-Means clustering algorithm. In Recent Advances in Space Technologies (RAST), 2017 8th International Conference on (pp. 143-148). IEEE.

[6] Vikas, A., Sreenivasan, A., & Sudha, K. L. (2017, February). Fuzzy C-means clustering and Criminisi algorithm based shadow removal scheme for side scan sonar images. In Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on (pp. 411-414), IEEE.

[7] Xuanzuo, Y., Dinghao, L., & Xiongxiong, H. (2017, May). An algorithm for automatic recognition of cluster centers based on local density clustering. In Control And Decision Conference (CCDC), 2017 29th Chinese (pp. 1347-1351). IEEE.

[8] S. Ghosh, S.K. Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms" in International Journal of Advanced Computer Science and Applications, Vol. 4, 2013:pp. 35-39.

[9] ViswanadhaRaju, S., Vidyasree, P., Gudavalli, M., & Sekhar, B. C. (2016, November). Minimum Cost Fused Feature Representation and Reconstruction with Autoencoder in Bimodal Recognition System. In Proceedings of the International Conference on Big Data and Advanced Wireless Technologies (p. 17). ACM.

