Hybrid K Mean Clustering Algorithm for Crop Production Analysis in Agriculture

Vandana B, S Sathish Kumar

Abstract*: The proposed research work aims to perform the cluster analysis in the field of Precision Agriculture. The k-means technique is implemented to cluster the agriculture data. Selecting K value plays a major role in k-mean algorithm. Different techniques are used to identify the number of cluster value (k-value). Identification of suitable initial centroid has an important role in k-means algorithm. In general it will be selected randomly. In the proposed work to get the stability in the result Hybrid K-Mean clustering is used to identify the initial centroids. Since initial cluster centers are well defined Hybrid K-Means acts as a stable clustering technique.

Keywords: Cluster analysis, K-Means, Precision Agriculture.

I. INTRODUCTION

Agriculture acts as a major work force in India. Satisfying the food security of increasing population is considered as an important task in recent years. Information and Communication technology tools are used in the field of agriculture to satisfy the food demand of people. The activities related to agriculture are closely associated with various factors like environmental and economic influences. Penetration of Information technology tools in the field of farming is aiding in data collection process. The key challenge is extracting the useful information from this large data which helps to the farmer community. Data analytical techniques are useful in the field of precision farming to retrieve the information from the field data.

Cluster analysis is a process of segregating data into clusters based on their related characteristics. Precision agriculture aims to manage small portion of the field and to optimize the production. Clustering can be used in agriculture to extract the data about soil moisture level and climatic conditions related to crops like temperature, humidity and rain fall statistics.

The Section II outlines the various cluster techniques and its applications. In section III, Different techniques to discover the optimal value of clusters were discussed. Section IV describes different metrics used for distance measure. In section V clustering methods were discussed in detail. Section VI describes Conclusions of the proposed work.

II. ROLE OF CLUSTER ANALYSIS

Work proposed by Qiang Fu et al. [1] describes about clustering techniques. Clustering is one of the scientific analyses used to describe the datasets based on the similarities among its data points. Partition based approaches are most widely used clustering techniques. Due to increased digitization in all domains huge amount of data is available which has to be analyzed properly to extract the information. Clustering algorithms are extensively used in the field of precision agriculture for different issues. Delineation of management zones using clustering supports variable rate fertilizer application. Fuzzy clustering technique is used to ensure the number of management zones based on variable rate fertilizer applications. It ensures increased crop yield with high quality. Research done by Dieisson pivot et al. [2] Shows the usage of sensors and information and communication technology tools in the field of smart farming which aids in increased productivity. But, Minimal education level of farmers act as a limiting factor in adopting smart farming technology. Severe climate changes and migration of people toward urban areas are affecting agriculture industry majorly. Smart farming techniques can be used to address these issues. Cluster analysis is carried out to find out the implications of smart farming and climate change. Information and communication technology tools, smart phones have changed the traditional agriculture practices. Smart phones can be used effectively to monitor the field parameters [3]. Research carried out by Jirapond Muangprathub et al. [4] shows the usage of Spontaneous irrigation system for agriculture field using sensor network. Smart phones and web application is used to manage the information efficiently. Data mining techniques are used to analyze the information for soil moisture level, temperature and humidity requirement of crops. System aims to improve the agriculture productivity through efficient data analysis. Work carried out by Soumi Ghosh et al. [5] describes how clustering techniques can be used to group the objects in precision agriculture. Usage of Fuzzy concept in k-means method is the major concept in FC-means technique. FC-means Clustering techniques are derived on fuzzy behavior it uses the method which is natural for forming the cluster-means belongs to the type of partitioning based clustering algorithm. In this technique number of final cluster (k) has to be declared in advance.

From the results of time complexity K-means algorithm performance is better than FC-means algorithm. Work done by B. Ramesh, K. Nandini [6] describes how cross-clustering techniques can be used to improve the clustering efficiency. A clustering algorithm which combines best aspects of multiple clustering algorithms will be better for practical usage. It can be used effectively for identification of outliers and to find out cluster members. This method can be used in medical as well as other domains.

Revised Manuscript Received on December 12, 2019.

***** Correspondence Author

Vandana B*, Research Scholar, Department of Computer Science and Engineering, RNSIT, Visvesvaraya Technological University, Belagavi, India. Email: vandanabcse4@gmail.com.

Dr. S Sathish Kumar, Professor, Department of ISE, RNSIT, VTU, Bengaluru, India. Email: sathish_tri@yahoo.com.

Various clustering techniques can be used to cluster the elements depending on application requirements. Uses of cluster analysis were discussed in detail.[7] Research carried out by Pradeep Nagendra Hegde et al. [8] describes that how efficiency of clustering technique can be analyzed using the metrics like sum of squared error and cluster distance. K-Means and FC Means are analyzed for the quality metrics. K-Means gives best output for the agriculture data. The work done by Beste Eren [9] shows how big datasets are analyzed using k-means algorithm efficiently. K-means is used to analyze mobile datasets. It can be used to develop models for the larger population. In the work proposed by Keshav Sanse and Meena Sharma [10] different clustering methods are compared using cluster validity metrics. K-means is the simple clustering technique but K-value has to be specified in advance. Hierarchical clustering technique forms clusters by splitting or combining data objects. Density based algorithms are used to develop arbitrary shaped clusters Grid based techniques are used to cluster the objects of spatial data. Work done by Xiaoli Cui et al. [11] shows that k-means technique can be used in big data architecture. Cluster validation techniques are used to assess the quality of clustering methods on different data sets. The research done by Tanvir Habib Sardar et al. [12] describes different types of partition based clustering techniques which can be used in Hadoop distributed platform to extract the knowledge from different datasets. Research carried out by Ramzi A. Haraty [13] shows how K-Means method is applied to discover the patterns in health care data. Sensor generated images, audio, text contents are difficult to analyses using traditional approach. K-Means technique is used to analyses such content effectively. An enhancement is made to traditional k-Means using greedy approach to produce the preliminary centroids. Work done by E. Ezhilan et al.[14] proposes Optimized k-means approach to address the initial cluster selection problem.

III. PROCEDURE TO DETERMINE THE OPTIMAL CLUSTERS VALUE

The optimal K value selection is often unclear. Finding out optimal value of clusters is essential in k-means algorithm. In the proposed work optimal value of clusters is identified by three different approaches. Silhouette Method, Elbow Method, and Gap Statistic Method are used to select the K-value as described below.

A. Elbow Method

A Graph is used to denote the proportion of variance given by the clusters versus the total number of clusters. The first cluster will have a lot of variance, at some point the marginal gain will reduce, a sharp angle will be formed in the graph. It appears like an elbow. After that elbow point including more clusters will not contribute major Value. That point is considered as the desired K. Partition based method like K-means ensures that clusters within cluster sum of square is reduced. Within cluster indicates closeness of clustering. It should be minimal as much as possible. It minimizes the intra cluster variation. In elbow technique within cluster sum of square is plotted against total number of clusters. Optimal clusters is identified in such a way that adding one more cluster will not improve the within sum of square. Fig. 1 shows the identification of optimal number of clusters based on elbow method. Three clusters are considered as an optimal value of clusters for the selected agriculture data using this technique.

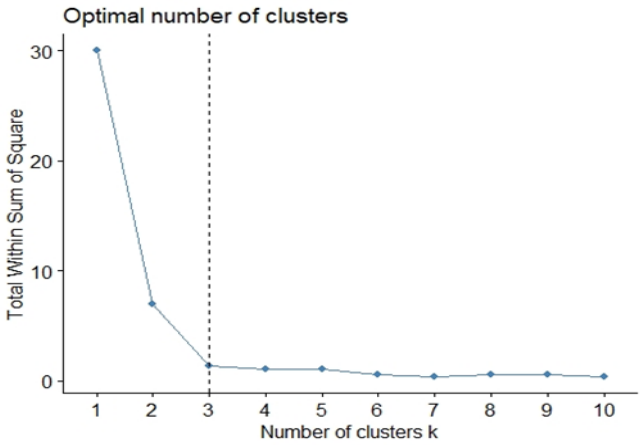

Fig. 1. **Elbow technique to select the number of clusters K.**

B. Silhouette Method

It defines how fine each object is associated within its cluster. A high value of silhouette width value denotes a fine clustering. The k value that maximizes the average silhouette value for a range of possible values of k is considered as an optimal k value. Fig. 2 shows the selection of optimal clusters based on Silhouette method. 3 clusters are considered as optimal clusters for the selected production data.

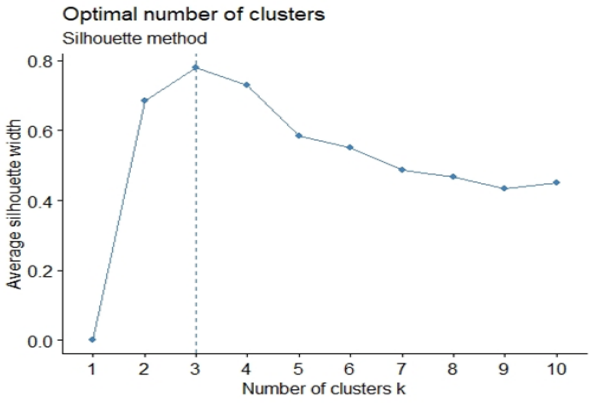

Fig. 2. **Silhouette method to select K value**

C. Gap Statistic Method

It relates within cluster deviance for k values with predictable values for null reference dispersal of data. K value which maximizes the gap statistic is considered as optimal cluster value. Fig. 3 shows the selection of k clusters based on Gap statistics method. 3 clusters are considered as an optimal value of clusters for the selected production data.

Retrieval Number: B10021292S19/2019©BEIESP
DOI: 10.35940/ijitee.B1002.1292S19

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

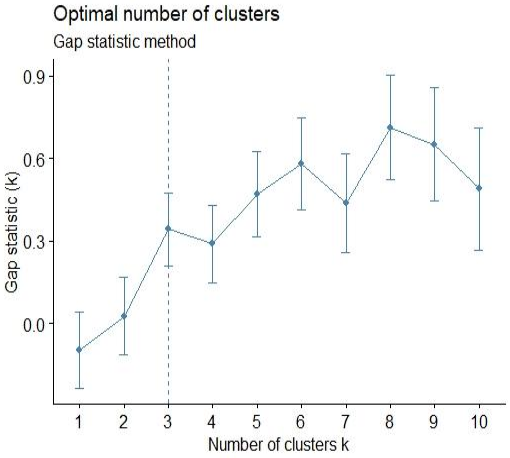

Fig. 3. **Gap Statistics technique to select the K value.**

By considering all methods three is considered as an optimal number of clusters for the selected dataset.

IV. DISTANCE METRICS BETWEEN TWO CLUSTERS

Collection of one or more records forms a cluster. Distance between the clusters can be measured using following techniques. Take cluster P, which contains the x records P1,P2…..Px and cluster Q, which contains y records Q1,Q2……Qy. The commonly used measures of distance between clusters are:

1) Minimum Distance between Clusters

The distance between the records Pi and Qj that is neighboring:

min(distance(Pi;Qj)); i = 1,2…….x; j = 1,2…… y.

2) Maximum Distance between Clusters

The distance between the pair of records Pi and Qj that is furthermost:

max(distance(Pi;Qj)); i = 1, 2…….x; j = 1, 2……. y.

3) Average Distance between the Clusters

The average distance of all likely distances between elements in one cluster and elements in the other cluster:

average(distance(Pi;Qj)); i = 1,2……x; j = 1, 2…….y.

4) Centroid Distance between the Clusters

The remoteness between the two cluster centroids is considered as follows.

distance (centroid P; centroid Q).

If clusters are in chain format then minimum distance is good option. In this method cluster members may not be close to one another but new member which has to be added to the cluster is close to one another. If clusters are in spherical format then Maximum and Average distance is considered as a better option.

V. CLUSTERING METHODS

The Research work focuses on the usage of clustering techniques to form the clusters based on crop data. In this approach similar data elements are grouped in a cluster and different data is separated by identifying patterns within the data. The main goal is to minimize the objective function i.e. the total distance between all elements from their cluster centers has to be minimized.

A. K-means

K-means generates k clusters by partitioning the data. In the initial step it chooses the cluster members or centers arbitrarily then distributes the remaining elements among the clusters and modifies the new value of cluster center. The method is to categorize given data by using K clusters. K centroids have to be recognized and placed apart from each other at different locations. Each object is allocated to the nearest centroid. Group of elements allocated to a cluster forms a cluster. Based on new objects cluster centroid will be updated to the new value. This process repeats until there is no modification in the cluster centroid. Random initialization of centroids results in different results in different execution. Selecting an initial centroid properly is an important step in k-means technique. Steps involved are:

1. Clusters k is selected as initial clusters.
2. In each stage element is allocated to the nearest centroid cluster.
3. Recalculate the centroids of modified clusters. Repeat Step 2.
4. Stop when further records movement between clusters increases cluster scattering.

B. Hierarchical Clustering Algorithm

Hierarchical clustering technique is represented by a tree like structure called a dendrogram. It displays cluster and sub cluster relationship. Two basic approaches in hierarchical clustering are:

1) Hierarchical Divisive Approach:

Starts with one cluster, in each step cluster are divided till the formation of singleton cluster. It also known as top down approach.

Hierarchical Divisive Clustering Algorithm is described as follows:

1. Start with whole sample as one cluster.
2. Split into two sub clusters.
3. Repeat the process until each data is in its own singleton cluster.

2) Hierarchical Agglomerative Clustering

Each point is considered as an individual cluster at each stage closest pair of clusters is merged with each other. It follows bottom up approach.

Hierarchical Agglomerative Clustering Algorithm is described as follows:

1. Start with *n* clusters as each element is considered as one cluster.
2. The two nearest elements are merged into one cluster.
3. At every stage, the two clusters having the smallest distance are merged.

The distances measures minimum distance, maximum distance, average distance, and centroid distance can be implemented in the hierarchical method as mentioned below:

- Clustering using Single Linkage

The Minimum distance value is used as a distance measure in single linkage clustering technique. This technique cluster together at early stage elements those are distant from each other because of a chain of intermediate elements in the same cluster.

- Clustering using Complete Linkage

The maximum distance value is used as a distance measure in complete linkage clustering technique. This technique produce clusters at the early stages with elements that are within a narrow range of distances from each other. The elements in such clusters have spherical shapes.

- Clustering using Average Linkage

The average distance value is used as a distance measure in average linkage clustering technique. It is based on all possible pairs of elements in the clusters.

- Centroid Linkage Clustering Technique

In Centroid linkage clustering centroid distance is used as a distance measure. Clusters are represented by the mean values for each variable. In average linkage method, pairwise distance is calculated. Average of all such distances is calculated. In centroid distance clustering the distance between group means is calculated.

- Ward's Method of Clustering Technique

Ward's method uses agglomerative approach. It combines elements and groups together incrementally to create larger clusters. Ward's method considers the loss of information that happens when records are clustered together.

C.Hybrid K-Means Algorithm

Hybrid K-Means Algorithm includes the advantages of hierarchical method and K-means algorithm. High accuracy of hierarchical method is combined with fast convergence of K-Mean technique. Cluster analysis is carried out to group the valuable information available in the clusters. In K-means random observations are selected as initial clusters. K means results are sensitive to these random centers. Results may not be stable each time when k-means is computed. To address this problem and hybrid algorithm is used by combining hierarchical clustering and k means clustering. Proposed Hybrid K-means technique is used to determine early centroids for K-means method.

Hierarchical technique builds clusters hierarchy. This arranges elements in the form of a tree structure. Elements are linked by very short branches if they are having high similarity with one another and long branches if they are having low similarity. As a first step in the hierarchical approach distance matrix between the elements of a cluster is calculated. Two nearest elements are joined by a node referred as a pseudonode. Two elements are removed from the list which has to be processed. These elements are represented by a pseudonode which indicates a new branch. Total number of clusters k is identified using silhouette method. Entire tree is divided into k clusters. Center of each cluster is identified. These centers are considered as initial centroids for k-means technique. It is described as mentioned below:

1. Define A={ai | i=1,……q} as each data of X where X={xi | i=1,…..r} is attribute of n-dimensional vector
2. Define total clusters as k
3. Set m as a total iterations
4. Define p=1 as an initial counter
5. Compute k means algorithm
6. Note the centroids as Cp= {cpj | j=1……k}
7. Update counter p=p+1
8. Step 5 is repeated while p<m.
9. Consider C={Cp | p=1……m} as a Data set with total number of cluster k
10. Compute Hierarchical algorithm
11. Note the centroids as A={ap| p=1,…k}
12. Consider A as initial cluster centroids.
13. Compute k-means clustering.

VI. RESULTS AND DISCUSSIONS

To analyze the efficiency of the Hybrid k-means clustering technique two datasets are used in the proposed work. US arrest Dataset and Crop Production Dataset. K means and Hybrid K means technique are computed on selected dataset for the experimental purpose. The following parameters are considered to analyze the cluster efficiency.

A. Strength

Strength is defined as between cluster sum of square divided by total sum of square. It is a measure of total variance in the data set. K-means maximize the between group dispersion and minimizes the within group dispersion. By assigning elements into k clusters rather than total number of samples there is a reduction in sum of squares.

B. Between Cluster Sum of Square

It defines between cluster sum of squares. It indicates squared average distance between all centroids. Euclidean distance from a given cluster centroid to all other centroids are calculated. This process is repeated for all other clusters. All values are added together which is referred as Between cluster sum of square. It measures the variance between all clusters. A Large value represents clusters are well separated. A small value indicates clusters are very close to one another.

C. Total within Cluster Sum of Square

It defines total sum of squares within the cluster. It measures goodness of the cluster in terms of compactness. It should be small as much as possible. Two different datasets are considered to compare the clustering techniques.

1) US Arrests Dataset

This dataset describes the arrest statistics per 100,000 residents for various types of crimes in 50 States of US in 1973. It also describes population living in urban areas. [15] k means and Hybrid k means technique are computed on US Arrest Dataset. Strength of the cluster, between cluster sum of square and total within cluster sum of square is considered as parameters to compare two algorithms. The results are represented in the Table 1.

Retrieval Number: B10021292S19/2019©BEIESP
DOI: 10.35940/ijitee.B1002.1292S19

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table 1: Comparison between K means and Hybrid K means For Us Arrest Dataset.

Algorithm	Strength	Between SS	Total within SS
K means	64.5%	126.3	69.6
Hybrid k means	71.2%	139	56

2) Crop Production Data

This data set contains description about wheat production statistics in the state of Karnataka from 2009 to 2016.

The data is collected from Directorate of Economics and Statistics, Karnataka. Data contains wheat production statics in various districts of Karnataka. Both kmeans and hybrid kmeans are computed for Crop production data. The results are described as shown in the Table 2.

Table 2: Comparison between k means and Hybrid k means for Crop Production Dataset.

Algorithm	Strength	Between SS	Total within SS
K means	96.5 %	28.93	1.06
Hybrid k means	97.3 %	29.18	0.81

as described in the Table 1 and Table 2 strength and between cluster sum of square of Hybrid kmeans is more when compared to kmeans. Intra cluster variation is minimal when compared to kmeans algorithm. It shows that Hybrid kmeans algorithm performs well when compared to kmeans and it can be used efficiently in agriculture dataset.

VII. CONCLUSION

It clearly shows that clustering can be used effectively to extract the knowledge in precision agriculture field. Traditional k-mean method is combined with Hierarchical algorithm for the centroid selection. Elbow technique, Silhouette technique and Gap statistics methods are used to select the optimal value of clusters.

In future it can also be used for management zone delineation where it is possible to apply the variable rate of fertilizers based on soil nutrients.

ACKNOWLEDGMENT

The authors extend their gratitude to the Management, Head of the Department and Principal of RRCE, RNSIT, Bengaluru and Visvesvaraya Technological University, Belagavi, for the immense support they have provided.

REFERENCES

1. Qiang Fu_, Zilong Wang, Qiuxiang Jiang "Delineating soil nutrient management zones based on fuzzy clustering optimized by PSO" Mathematical and Computer Modelling 51 (2010) 1299_1305.
2. Dieisson Pivoto at el. "Scientific development of smart farming technologies and their application in Brazil" INFORMATION PROCESSING IN AGRICULTURE 5 (2018) 21–32.
3. Vandana B, S Sathish Kumar, " A smart phone based information sharing system in precision agriculture" AIP Conference Proceedings 2039,020002(2018);doi:10:1063/1.5078961.
4. Jirapond Muangprathub at. el. " IoT and agriculture data analysis for smart farm" Computers and Electronics in Agriculture, 156 (2019) 467–474.
5. Soumi Ghosh, Sanjay Kumar Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms" ((IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013, 35-39.
6. B. Ramesh, K. Nandhini, "Clustering Algorithms – A Literature Review",International Journal of Computer Sciences and Engineering, Volume-5,Issue-10 E-ISSN: 2347-2693
7. Vandana B, S Sathish Kumar, "Cluster Analysis in Precision Agriculture",International Journal of Computer Science and Engineering, Volume-7,Issue-4, April 2019, 473-477.
8. Pradeep Nagendra Hegde et. al. "Performance Analysis of Data Mining Techniques for Real Time Applications" International Journal of Current Engineering and Scientific Research (IJCESR) ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-4, ISSUE-3, 2017, 66-70
9. Beste Eren, "K-Means Algorithm Application on Big Data", Proceedings of the World Congress on Engineering and Computer Science 2015 Vol II, WCECS 2015, October 21-23, 2015, San Francisco, USA.
10. Keshav Sanse, Meena Sharma, "Clustering methods for Big data analysis" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 3, March 2015, ISSN: 2278 – 1323, 642-648.
11. Xiaoli Cui et. al. "Optimized big data K-means clustering using MapReduce", Journal of Supercomputer 70:1249-1259, DOI
12. 10.1007/s11227-014-1225-7, Springer.
13. Tanvir Habib Sardar, Zahid Ansari, "Partition based clusteringof large datasets using MapReduce framework: An analysis of recent themes and directions", Future Computing and Informatics Journal 3 (2018) 247-261, Science Direct.
14. Ramzi A. Haraty, "An Enhanced K-Means Clustering Algorithm for Pattern Discovery in Healthcare Data", International Journal of Distributed Sensor Networks Volume 2015, Article ID 615740, http://dx.doi.org/10.1155/615740.
15. E. Ezhilan et/. al. "Implementation of Optimized K-Means Clustering on Hadoop Platform", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 2, ISSN: 2277128X.
16. McNeil, D.R , Interactive Data Analysis.New York, Wiley. (1977).

AUTHORS PROFILE

Ms. Vandana B has done B.E degree from KSIT, Bengaluru, and M.Tech degree from JNNCE, Shimogga. She is a Research Scholar in RNSIT, VTU, Belagavi, Karnataka, India. She has published more than 20 papers in various journals and conferences. Her research interest includes Data analytics, Data Mining, Network, Big Data and Cloud Computing. She is a Life Member of ISTE, Member of IAENG and CSTA. Currently she is working as an Assistant professor in the department of CSE, RRCE Bengaluru.

Dr. S Sathish Kumar has done B.E degree from Madurai Kamaraj University, Tamilnadu, M.E degree from Regional Engineering College, Tiruchirappalli, Bharathidasan University, Tamilnadu, and Doctor of Philosophy Degree from Dr. M. G. R. University, Chennai, India. He has published research papers in various journals and International conferences. His research field covers Data Analytics, Bio Informatics, Communication Networks and Cloud Computing. He has served as a BOE/BOS member of various colleges. Currently he is working as a Professor in the department of ISE, R N S I T, Bengaluru.