IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Cluster Analysis in Precision Agriculture

[1*]Vandana B., [2]S. Sathish Kumar

[1]Visvesvaraya Technological University, Belagavi, India
[2]Dept. of Computer Science and Engineering, RNS Institute of Technology, Bengaluru, India

[*]*Corresponding Author: vandanabcse4@gmail.com, Tel.: 9148924876*

DOI: https://doi.org/10.26438/ijcse/v7i4.473477 | Available online at: www.ijcseonline.org

Accepted: 08/Apr/2019, Published: 30/Apr/2019

Abstract- In this paper, the use of clustering techniques in the field of precision agriculture has been discussed. Types of clustering techniques discussed are k means clustering, mean shift clustering; Density based spatial clustering of applications with noise (DBSCAN), Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM) and Hierarchical clustering. As clustering is a method of identifying similar groups of data in a data set, Clustering has a huge number of uses spread crosswise over different spaces. In data science clustering analysis is used to gain some valuable insights from the data by looking at what groups the data point belongs to which group when clustering algorithm is applied. **Few applications of cluster analysis in the field of agriculture are using k means, hierarchical** agglomerative clustering approach, pam clustering method and divisive clustering approach to form the clusters based on soil fertility, crop production, irrigation requirements etc.

Keywords— Precision Agriculture, k_means clustering, Density based spatial clustering of applications with noise(DBSCAN), Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM), Pam clustering, Hierarchial clustering.

I. INTRODUCTION

Clustering is a method of identifying similar groups of data in a data set. Entities in each group are nearly more like elements of that aggregate than those of alternate groups. Clustering is the errand of isolating the populace or information focuses into various groups with the end goal that information focuses in similar groups are increasingly like other information focuses in similar groups than those in different groups. In basic words, the point is to isolate clusters with comparative qualities and appoint them into groups. Clustering has a huge number of uses spread crosswise over different spaces. The absolute most prevalent uses of grouping are

- Market segmentation
- Social network analysis
- Search result grouping
- Medical imaging
- Image segmentation
- Anomaly detection

Clustering is an unsupervised machine learning approach; however would it be able to be utilized to enhance the exactness of directed machine learning calculations too by clustering the information focuses into comparable groups.

Precision agriculture focuses on usage of Information and Communication technology tools in the field of agriculture to increase the crop yield. Clustering can be used extensively in the field of precision agriculture to form the grouping of objects in which we don't have clear class labels.

The work focuses on different types of clustering methods that can be used in the field of precision agriculture. Selecting optimal number of cluster plays an important role in cluster analysis. Techniques like Elbow method and Silhouette method are used to select optimal number of clusters. The work emphasizes on formation of clusters based on K-mean cluster for crop yield data.

Rest of the paper is organized as follows, section I contains the Introduction of Clustering and its Applications, Section II contains related work of Clustering and Data mining technology, Section III contains types of clustering algorithms, Section IV contains Results and Discussions, Section V contains applications of clustering, Section VI concludes the work with the future plans of the research work.

II. RELATED WORK

Data mining and cluster analysis should be a part of agriculture because they can improve the accuracy of

decision systems. The cluster heuristic allows data to be combined into useful patterns that may lead to better decisions. Effective techniques can be carved out for solving different agricultural problems of various complexities by intelligent use of data mining and its tools such as cluster analysis.

Few applications of the cluster analysis in agriculture fields were discussed such as hierarchical agglomerative clustering approach, fuzzy clustering, hierarchical divisive clustering and Kohonen self-organizing feature maps along with an application of each of these techniques in the field of agriculture is also presented [1].

Data clustering Algorithms are essential approaches to analyze the agricultural data and also to achieve the practical and effective solutions for agricultural problems such as suitable crop for the particular soil type, crop which can produce maximum production the environments like more temperature, less rain fall, less nitrogen content in the soil etc. Variations in the environmental conditions like sudden raises in temperature, reduced / increased rainfall, variations in the market prices, etc. with all these aspects it is difficult for formers to take critical farming decisions.

Several data clustering methods are implemented on the input data in order to achieve maximum rice production. DBSCAN and CLIQUE clustering algorithm are used to obtain the optimum requirement of climatic conditions for rice, like optimum range of best temperature, high temperature, rain fall and PH value to obtain maximum production of rice crop. [2]

Author describes the importance of cluster analysis techniques in segmentation and pattern extraction. It focuses on selecting optimal clustering technique to extract the valuable information from the given dataset [3].

Author describes about importance of data mining and its applications. It focuses on predictive analysis and cluster mining in the field of IT security . IT Security is emerging as a major sector in IT Industries. Data mining techniques can be associated with security techniques to improve the level of security in IT industry [4].

III. TYPES OF CLUSTERING

Different types of clustering can be selected based on the application requirement and type of datasets. Major clustering techniques used in the field of precision agriculture are:

A. K-Means Clustering

K-Means is presumably the most well-known clustering algorithm. It initially selects various classes/groups to utilize and arbitrarily introduce their individual center points. The centre point is vectors of indistinguishable length from every datum point vector. Each information point is arranged by processing the separation between that point and each group center, and after that ordering the point to be in the group whose center is nearest to it. Based on these ordered centers, the group center is recomputed by taking the mean of the considerable number of vectors in the group. These means are repeated for a set number of iterations or until the point that the group center doesn't change much between cycles. likewise pick to arbitrarily introduce the group center a couple of times, and after that select the run that seems as though it gave the best outcomes. K-Means has the preferred standpoint that it's entirely quick, as all we're truly doing is processing the separations among centers and group centers. Then again, K-Means has several disservices. Initially, it is required to choose what number of groups/classes there are. K-implies additionally begins with an arbitrary decision of group focuses and it might yield diverse bunching results on various keeps running of the calculation. Along these lines, the outcomes may not be repeatable and need consistency.

B. Mean Shift Clustering

Mean shift clustering is a sliding-window-based calculation that endeavours to discover dense areas of data points. It is a centroid-based calculation implying that the objective is to find the center points of each groups/class, which works by updating candidates for center points to be the mean of the points within the sliding-window. These applicant windows are then sifted in a post-preparing stage to dispense with close copies, framing the last arrangement of center points and their comparing groups.

Mean shift considers highlight space as an empirical probability density function. Mean shift considers them as sampled from the underlying probability density function, if the input is a set of points. using Mean Shift, we can also identify clusters associated with the given mode for each data point; Mean shift associates its dataset's probability density function with its nearby peak. For each data point, Mean shift computes the mean of the data point when it defines a window around it and then it repeats the algorithm till it converges after it shifts the center of the window to the mean the window shifts to a more denser region of the dataset, after each iteration.

Mean Shift, at the high level can be specified as follows:

- Around each data point, fix a window.
- Within the window, compute the mean of data.
- After shifting the window to the mean, repeat till convergence.

C. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a density based clustered algorithm similar to mean-shift, but it has a pair of advantages over it. DBSCAN is one of the clustering methods of Density based clustering methods. The working principle of DBSCAN algorithm is depends on the user defined parameters those are radius (Eps) and minimum number of points to be present within the threshold radius (MinPts). Determining the optimal EPs value is challenging task and difficult for the user who has no prior knowledge about the database.

D. Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMMs) is more flexible than K-Means. We assume that the data points are Gaussian distributed with GMMs; which is less prohibitive presumption than saying they are round by utilizing the mean. That way, two parameters are used to describe the clusters shape: the mean and the standard deviation. GMM model accommodates mixed membership. GMM allows for mixed membership of points to clusters is another implication of its covariance structure. In k means, a point belongs to one and only one cluster, whereas in GMM a point belongs to each cluster to a different degree. The degree is based on the probability of the point being generated from each cluster's (multivariate) normal distribution, with cluster center as the distribution's mean and cluster covariance as its covariance. Mixed membership may be more appropriate (e.g. news articles can belong to multiple topic clusters) or not (e.g. organisms can belong to only one species), depending on the task.

E. Hierarchical Clustering

Hierarchical clustering is a cluster analysis method which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two categories.

1) Agglomerative:.

1) Agglomerative Hierarchical Clustering: This is a bottom up approach: here, each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. This algorithm works on the basis of the nearest distance measure of all the pairwise distance between the data point by grouping the data one by one. Again distance between the data point is recalculated but when the groups has been formed, which distance to consider. For this there are many available methods. Some of them are.

- Single-nearest distance or single linkage.
- Complete-farthest distance or complete linkage.
- Average-average distance or average linkage.
- Centroid distance.
- Ward's method - sum of squared Euclidean distance is minimized.

a) Algorithm:

Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering is defined as follows:

- Start with M clusters, and a single sample indicates one cluster.
- Find the closest (most similar) pair of clusters and merge them into a single cluster, so that it has one cluster less.
- Compute distances (similarities) between the new cluster and each of the old clusters.
- Repeat steps until all items are clustered into a single cluster of size M

2) Divisive Hierarchical Clustering: This is a top down approach: here, all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. The single cluster splits into 2 or more clusters. Splitting continues till the number of clusters becomes equal to the number of samples or as specified by the user, whichever is less.

b) Algorithm:

- Initially start with a single cluster encompassing all elements.
- Select one, the largest cluster or the cluster with highest diameter;
- Find the element f in one that has the lowest average similarity to the other elements in one;
- f is the first element added to the splinter group while the other elements in one remain in the original group;
- Find the element g in the original group that has highest average similarity with the splinter group;
- If the average similarity of g with the splinter group is higher than its average similarity with the original group then assign g to the splinter group and go to previous step; otherwise do nothing.
- Repeat until each element belongs to its own cluster.

IV. RESULTS AND DISCUSSIONS

Selecting an optimal number of clusters is an important stage in partition based clustering algorithm like k-means or K medoids. Crop yield data is considered for applying clustering algorithm. R Tool is used to carry out the cluster analysis.

Elbow method is used to find out the optimal number of cluster. Data sets are collected from Directorate of Economics and Statistics, Karnataka, India. Fig.1. describes the optimal number of cluster as four for the selected dataset.

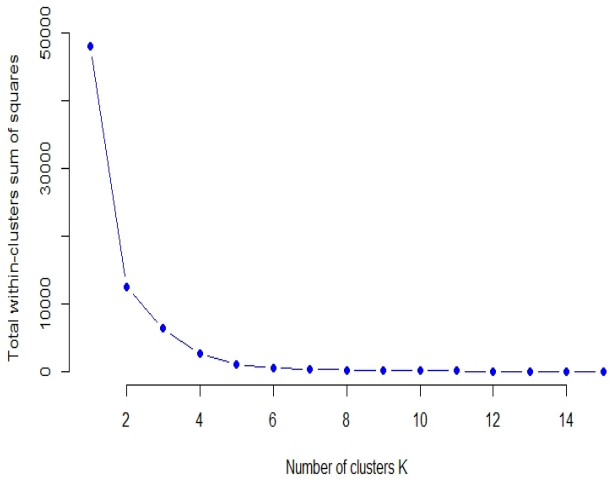

Fig.1. Optimal Number of Cluster selection using Elbow technique.

Crop yield data is clustered by applying K-Means algorithm. Four clusters are formed based on the result of elbow approach. Fig. 2. Describes the formation of different clusters based on the crop yield data.

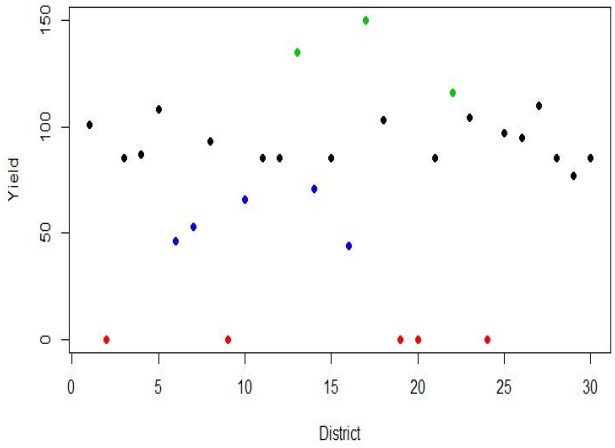

Fig.2. Four clusters formed using crop data.

Results describes that cluster analysis can be used effectively in the field of precision agriculture. Regions can be clearly grouped based on the crop yield parameter.

V. APPLICATIONS OF CLUSTERING

Applications of cluster analysis in the field of agriculture are hierarchical agglomerative clustering approach, divisive clustering, K-means clustering, Pam clustering etc. which can be used to group the regions based on different types of crop parameters.

Precision agriculture is mainly concerned with the integration of various Information and Communication Technologies in the field of agriculture. This technology is getting embedded into various agricultural equipments, due to integration, agricultural equipments are useful to farmers.

One of the tasks is the delineation of management zones; the delineation of management zones has been used as a method of subdividing fields into parts by taking different properties for a long time.

The initial step of spatially dividing the information data points may be accomplished by overlaying a network. Because of the anomalies in the field shape and the holes just as openings in natural data density, running a k-means algorithm on the directions (co-ordinates) of the data points in the data set gives a more adaptable answer for the underlying tessellation. An upper bound for the parameter k is given by the span of the subsequent littlest zone while zones beneath a limit that is being given by the precision of used farming equipment which cannot be managed. A lower bound for the k parameter is set by the granularity of the last administration zones and by the measure of heterogeneity on the field.

The second step of redundantly converging of two zones has two constraints: first, zones which are to be blended must be comparable in their qualities; second, they should be immediate neighbors in land space (spatial limitation). As the outcome of both the conditions, it would be guaranteed that the subsequent zones will rather be homogeneous, as indicated by the principal condition and contiguous, agreeing to the second condition.

Unusual rainfall fluctuations that differs from year to year and from one region to other region is a major reason for decreased crop yield. Crops require various amount of water in various growing season. Accordingly, it is hard to know the variations of rainfall in a region. The identification of pattern of rainfall becomes an essential task for regional and local planners as well as managers. It is very much essential to classify meteorological data such as rainfall and humidity

features to identify the patterns of moisture level which influences the crop yield.

Cluster based approach is mainly focused in agriculture as well as allied sectors. In this approach commonly referred to as cluster farming, real profit is generated only by merging several small farms to a mother farm. Data mining and cluster analysis improves the accuracy of decision systems. The cluster heuristic allows data to be combined into useful patterns that may lead to better decisions.

VI. CONCLUSION AND FUTURE SCOPE

Clustering techniques are used in various applications. The paper gives an overview about different types of clustering techniques. Clustering and its application in agricultural field are discussed. In present scenario the application of cluster analysis has already gained momentum, still there are lot of areas where a great deal of efforts is still required where clustering techniques can be used to form production domains, market regions etc., which will be implemented in future.

REFERENCES

[1] Mamta Tiwari, Dr.Bharat Misra,"Application of Cluster Analysis in Agriculture", International Journal of Computer Applications Volume 36– No.4, (0975 – 8887), December 2011.

[2] K.Ranjini, Dr.N.Rajalingam,"Performance Analysis of Hierarchical Clustering Algorithm", Int. J. Advanced Networking and Applications Volume: 03, Issue: 01, Pages: 1006-1011, 2011.

[3] R.S. Walse, G.D. Kurundkar, P. U. Bhalchandra, "A Review: Design and Development of Novel Techniques for Clustering and Classification of Data", International Journal of Scientific Research in Computer Science and Engineering, Vol.06, Issue.01, pp.19-22, 2018.

[4] Shilpa Mahajan, "Convergence of IT and Data Mining with other technologies", International Journal of Scientific Research in Computer Science and Engineering, Vol.01, Issue.04, pp.31-37, 2013.

[5] Ms. Shilpa Ankalaki, Jharna Majumdar, "Applications of Clustering Algorithms for Analysis of Agriculture Data", International Journal of Engineering & Technology, 7, 3, 638-643, 2018.

[6] Dileep Kumar Yadav," A Comparative Analysis of Clustering Algorithm for Agriculture Data", International Journal of Current Research Vol. 7, Issue, 07, pp.18361-18364, July, 2015.

[7] Jiawei Han, Micheline Kamber," Data Mining Concept and Techniques ", 2nd Ed. - Morgan Kaufmann Publishers.

[8] JharnaMajumdar, Sneha Naraseeyappa and Shilpa Ankalaki," Analysis of agriculture data using data mining techniques: application of big data", Majumdar et al. J Big Data, 4:20 DOI 10.1186/s40537-017-0077-4, 2017.

[9] Vandana B, S. Sathish Kumar, " Big Data Analysis through R for Weather Monitoring", Global Journal of Engineering Science and Researches, ICRTCET-2018, pp 99-106, 2019

[10] Rahmah N, Sitanggang S. Determination of optimal epsilon (Eps) value on DBSCAN algorithm to clustering data on peatland hotspots in Sumatra. IOP conference series: earth and environmental. Science. 2016.

Authors Profile

Vandana B is a Research Scholar in Visvesvaraya Technological University, Belagavi, Karnataka, India. She is pursuing research at R N S Institute of Technology, Bengaluru and She is working as an Assistant Professor in the department of Computer Science and Engineering, RajaRajeswari College of Engineering, Bengaluru. She received her B.E degree from K. S. Institute of Technology, Visvesvaraya Technological University, Bengaluru, Karnataka and M. Tech degree from Jawaharlal Nehru National College of Engineering, Shimogga, Karnataka, India.

S Sathish Kumar is an Associate Professor in R N S Institute of Technology, Bengaluru, Karnataka, India. He received his B.E degree from Madurai Kamaraj University, Tamilnadu, India, M.E degree from Regional Engineering College, Tiruchirappalli, Bharathidasan University, Tamilnadu, India. and Doctor of Philosophy Degree from Dr. M. G. R University, Chennai, Tamilnadu, India. His field of Interest includes Data Mining, Big Data, Bio Medical Engineering and Cloud Computing.

