TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 3800-3805

Usage of Clustering Algorithms in Modern Agriculture Sector and Related Areas – a Primer

[1]Pinaka Pani. R, [2]M. Jayakameswaraiah

[1]Assistant Professor, School of Computer Science and Applications, REVA University, Bangalore
[1]rppani.mca@gmail.com
[2]Assistant Professor, School of Computer Science and Applications, REVA University, Bangalore
[2]drjayakameswar@gmail.com

Article Info
Volume 83
Page Number: 3800-3805
Publication Issue:
May-June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract
Agricultural crop depends on numerous factors like earth geography, climate, and economy. In recent years, various agricultural support systems for greenhouse are planned and enforced. In fashionable Agriculture wherever farmer and business have to be compelled to create call every day and complexness involves numerous issue influencing them. Data mining algorithms and their techniques are necessary to approach for obtaining an effective result. Agriculture development is an important topic comes under big data. Our research work specializes in maximizing the assessment of the yield of cereal crops using the variety of data processing techniques like PAM, CLARA, and DBSCAN.

Keywords: *Data mining, Clustering, PAM, CLARA and DBSCAN*

1. Introduction

Agriculture plays a crucial role in day to day life. In fashionable agriculture farmers and business have to be compelled to create choices each day and complexness involves the assorted factors influencing them. Data processing in agriculture is incredibly new and up to date topic. It analyses the relationship between the plant's growth, maximizing crops, climate, wetness, and temperature. Data mining clustering algorithms are very useful in modern agriculture sector. It analysis the data in data mining methods and discovers the patterns in big information sets. It involves methods at the intersection of computing science, data mining, statistics, and info system. Cluster and Classification square measure 2 differing types of learning strategies within the data processing. A cluster is grouping the data sets into groups in step by step process. The classification process or prior knowledge will not be there while grouping the data sets.

There are many types of clustering algorithms existing such as Hierarchical and Partitional algorithms in clustering. The aim is that data within the identical cluster have tiny space from each other, whereas knowledge points in numerous cluster square measure at an oversized distance from each other. Cluster analysis divides knowledge into well-shaped teams. This research work focuses on various clustering algorithms like PAM, CLARA and DBSCAN clump strategies. These strategies measure accustomed categorize the exclusive districts of Karnataka that measure having comparable crop production [1]. According to Leonard dramatist and Peter J. Rousseeuw, PAM for "**Partition Around Medoids**". The formula is meant to search out a sequence of information referred to as medoids that square measure centrally settled in clusters. Objects that square measure tentatively outlined as medoids square measure placed into a group S of designated objects. The use of the formula is to attenuate the typical difference of objects to their highest designated object. Equivalently, we are able to minimize the total of the dissimilarities between an object and their highest designated object. The formula has two conditions.

CLARA stands for (Clustering giant Application) depends on the sampling approach to handling a giant set of information. Rather than finding medoids for the complete set of information. CLARA attracts a little set from knowledge set Associate in Nursing

May-June 2020
ISSN: 0193-4120 Page No. 3800-3805

applies the PAM formula to come up with an optimal set of methods for the sample. Density-based Spatial clustering of applications with noise (DBSCAN) is a data clustering formula planned by Martin compound, Hans-Peter Kriegel, Jörg power tool and Xiaowei Xu in 1996. DBSCAN is one among the foremost common clump algorithms and additionally most cited in scientific literature.

2. Methods

The work carried in this paper is to analyze the agricultural information by applying various clustering algorithms. The information for analysis was taken from the various parts of Karnataka.
Input knowledge set contains the parameters namely: year, District, a crop like wheat, rice, cotton etc. It depends on soil condition, varying temperature and minimum precipitation required.

In this research paper a modified method of DBSCAN methodology is utilized to cluster the information supported districts that area unit having similar conditions such as temperature, rainfall, and soil kind. These algorithms area unit accustomed cluster the information supported the districts that area unit providing most crop productivity (In planned work wheat crop is taken under consideration as an example). Supported these analyses we tend to face live obtaining the optimum parameters to produce the utmost crop production. Multiple statistical methods methodologies are utilized to forecast the year crop production.

Partition Around Medoids (PAM)

PAM is a clustering algorithm or K-medoids. PAM uses a greedy seek which might not find the greatest solution, but it's miles quicker than exhaustive seek. It divides the data into groups. It results with a bunch of objects remarked as medoids that unit of measurement at the centre of data items. With the medoids, nearest information points are computed and created it as clusters. The formula has two stages as shown in the fig(1).

Algorithm PAM
Input:
D = {t_1, t_2, t_3, ……., t_n} // set of elements
A // adjacency matrix showing distance between elements.
k // Number of desired clusters.
Output:
K // set of clusters.
PAM Algorithm:
Arbitrarily select k medoids from D;
repeat
For each t_h not a medoid do
For each medoid t_i do
Compute square error function TC_{ih};
Find i, h where TC_{ih} is the smallest;
If $TC_{ih} < 0$ then
Replace medoid t_i with t_h;
Until $TC_{ih} >= 0$;
For each $t_i \in D$ do
Assign t_i to K_j where dis(t_i, t_j) is the smallest over all medoids;

Figure 1: The formula has two stages as shown

Example: The flow chart of an algorithm is shown in the below fig(2). It's a working steps PAM algorithm.

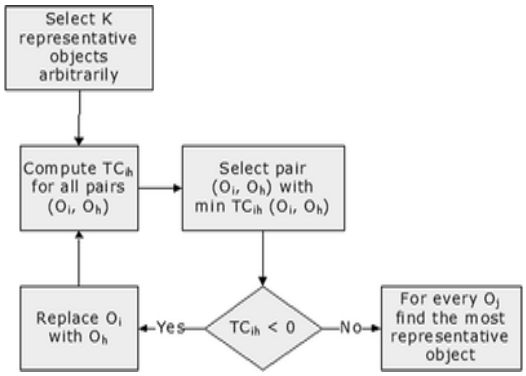


Figure 2: The flow chart of an algorithm is shown

CLARA (Clustering LARge Application)

CLARA stands for Clustering LARge Application. It was designed by dramatist and Rousseeuw to handle giant information. It attracts multiple samples of the info set and applies PAM on every sample, provides the most effective clustering as output and finds the medoids of the sample. It handles the giant dataset than PAM. Here, for accuracy, the standard of the clump is measured supported the typical difference of all objects within the entire knowledge set. The fig(4) algorithm briefs concerning the steps concerned within the CLARA.

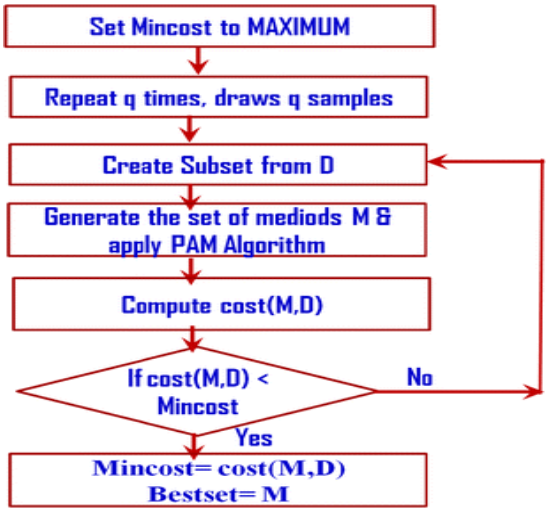


Figure 3:

DBSCAN

DBSCAN is the formula for clusters containing a large number of data and information. It has two parameters particularly Eps and Mints. But in ancient or traditional DBSCAN cannot manufacture the optimum Eps value. The most priority and necessary modification required to be done in to find the optimal Eps value repeatedly and determining the Eps value.

May-June 2020
ISSN: 0193-4120 Page No. 3800-3805

Figure 2 shown below explains the updated way of DBSCAN methodology.

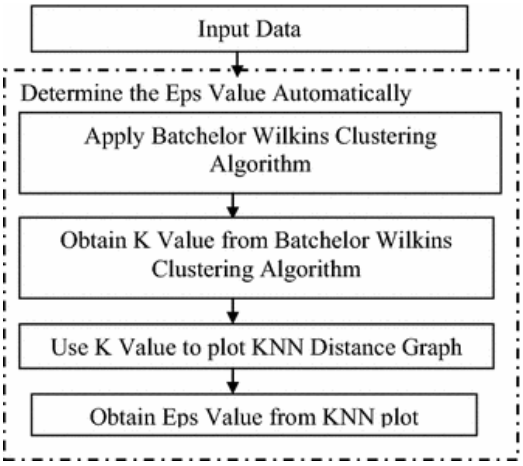


Figure 4:

Determines the Eps price automatically. Updated DBSCAN shows the strategy of searching the minimum values and area value

DBSCAN the modified proposes the strategy to seek out the lowest amount factors and Epsilon (radius price) consistently. KNN plot is employed to seek out the epsilon value where an input to the KNN plot (K fee) is mentioned. To avoid the user-defined K price as entering the KNN plot, the Batchelor Wilkins clustering method is carried out to the database and gain the K fee along with its respective clusters. Value of K is given initially as input to the KNN Plot. The value of Epsilon (Eps) can be calculated by drawing a "K-distance graph" for complete facts-points in the course of a dataset for a given 'K', obtained through the Batchelor Wilkins Algorithm. Initially, the gap of a factor to each 'K' of its nearest-neighbors is computed. KNN plot is deliberate by using taking the looked after values of common distance values.

Evaluation Method

The techniques of mining Algorithm works with different methods. By considering the association on various parameter it is very flexible to apply to the evaluation method. For agriculture data mining this technique will be very useful in adopting the different parameter like soil analysis, temperature and rainfall conditions. By using this technique we get homogenetic of each cluster that contains solely a member of one class [3].

Rand index of this technique measures the position of pic selection which squares measures are correct. By applying this techniques method we get accurate results when we are considering the fraction of pairs properly placed within the cluster of similar behavior. The higher values of these techniques give the best cluster quality.

PAM

we considered the data set which provides the number of cluster K, where K is given as 3 in the present experiment. The yield of the crops is categorized into high, low and moderate production values. To apply the PAM clustering method we also consider the districts of Karnataka into three clusters using this method. The result of the same is shown in the table(1). The analysis of the production of the wheat crop in Karnataka state as shown in figure(3).

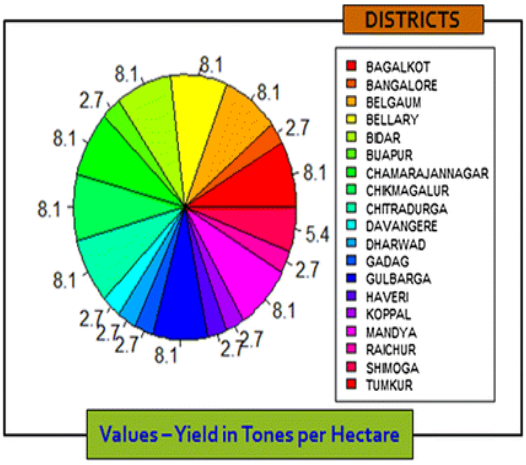


Figure 5:

Yield in tonnes per hectare of various districts Table 1

Low crop Producing areas	**High crop Producing areas**	**High crop Producing areas**
Bellary, Raichur, Mandya, Gadag, Gulbarga	Haveri, Bidar, Bijapur Tumkur, Belgaum, Koppal and Chamarajanagar	Bangalore, Shimoga, Chikman Galur

A result of the PAM Algorithm. As observed the north side of Karnataka produces maximum yield..

CLARA

This is another eminent data mining technique in clustering. For this experiment, we considered a similar parameter like rainfall, soil analysis, temperature area, and production. The result of this algorithm is given in table 2.

May-June 2020
ISSN: 0193-4120 Page No. 3800-3805

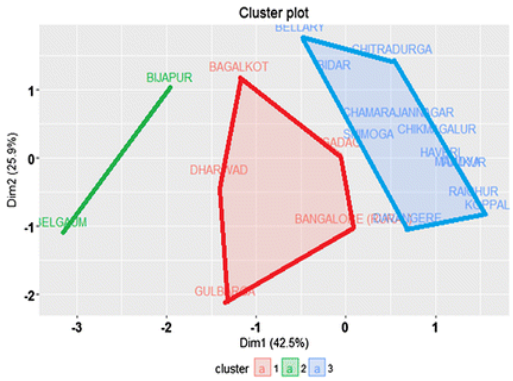


Figure 6:

A results of CLARA algorithm using R Language

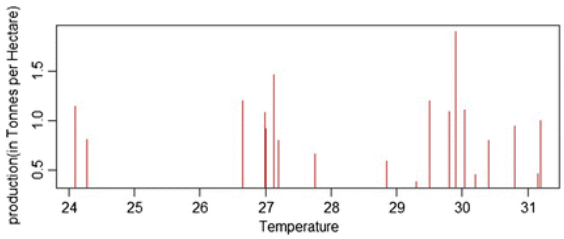


Figure 7:

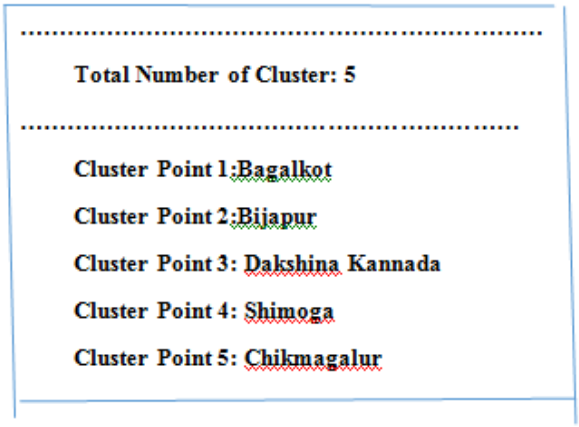
Plot the graph Temperature vs Production A result of CLARA Algorithm Table 2

Large yield low Rainfall. Temp Range(24 -26)	Medium Area, Production and High Rainfall Temp range (27-29)	Less Area, Production and Moderate Rainfall (29-30)
Bijapur and Belgaum	Gadag, Gulbarga, Dharwad ,Bangalore and Bagalkote	Bidar, Koppal , Davengere, Tumkur Mandya, Bellary, Shimoga.

DBSCAN

Updated DBSCAN shows the strategy of searching the minimum values and area value by itself. The value can be found by sketching a K- distance graph for an entire data point during a data set for an obtained value K. It is found by Batchelor and Wilkin's formula continuously.

The distance value of all K is the nearest values is calculated.



KNN graph is plotted using K value and the min points for the DBSCAN is shown in fig[7].

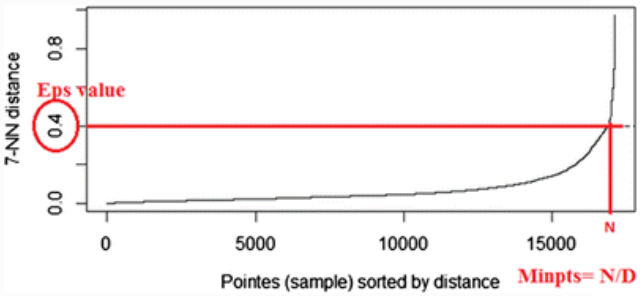


Figure 8:

DISTRICT	Symbol	District	Symbol
BAGALKOT		GULBARGA	
BANGALORE (RURAL)		HASSAN	
BANGALORE (URBAN)		HAVERI	
BELGAUM		KODAGU(COORG)	
BELLARY		KOLAR	
BIDAR		KOPPAL	
BIJAPUR		MANDYA	
CHAMARAJANNAGAR		MYSORE	
CHIKMAGALUR		RAICHUR	
CHITRADURGA		SHIMOGA	
DAKSHINAKANNADA		TUMKUR	
DAVANGERE		UDUPI	
DHARWAD		UTTARAKANNADA	
GADAG			

Figure 9:

DBSCAN

We considered the various parts of Karnataka state which are identical in rainfall, soil and heat measurement considering base EPs values for executing DBSCAN clustering algorithm.

We have selected the different parts of Karnataka for the purpose of analyzing the various parameters like temperature, rainfall range, and soil type are considered from fig[8].

The fig(9) and fig(10) gives the regions of Karnataka which are in identical parameter values.

May-June 2020
ISSN: 0193-4120 Page No. 3800-3805

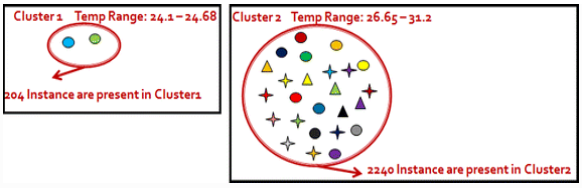


Figure 10:

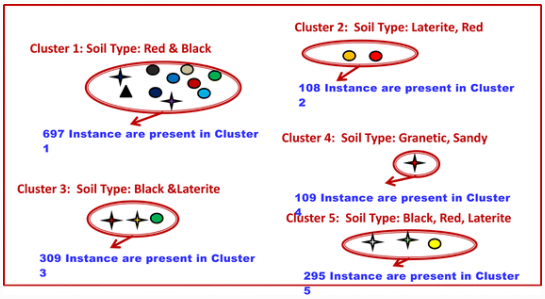


Figure 11:

Comparison of clustering methods

In this, we are comparing the clustering algorithms. The purity, oneness, completeness , Rand Index etc are being compared.

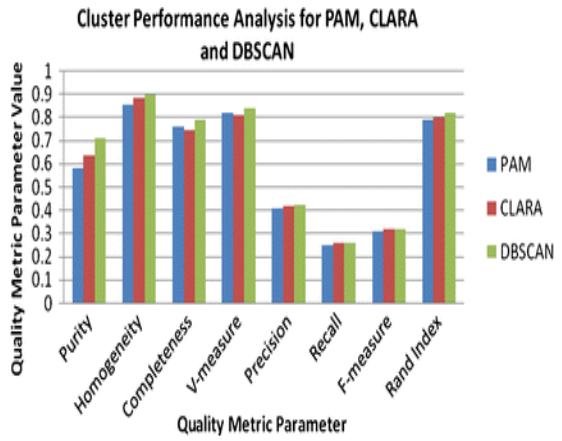


Figure 12:

In figure(11) is an example to show the comparison of the algorithm. From the example, we can see that the performance of the DBSCAN algorithm shows good results than PAM and CLARA.

3. Discussion

The crops square measure hand-picked supported its economic importance. However, the agricultural arising with methodology wants a yield estimation of the many crops. Thus, a crop was hand-picked once enough knowledge samples appeared within the vary of five to six years below analysis. It presents works, analysis on wheat crop is analyzed and mentioned on the varied issue of this paper. The present work paper deals with PAM, DBSCAN, and CLARA, and changed DBSCAN clump strategies. PAM and CLARA square measure the standard clump strategies wherever as DBSCAN methodology is changed clump methodology by introducing the Batchelor Wilkins that verify the 'k' price and KNN methodology to see the minimum points and radius price mechanically. Mistreatment these strategies crop knowledge set is analyzed and determined the best parameters for the wheat crop production. In these works, an analysis is restricted to the external quality metrics that square measure combination of many metrics those square measure set the same metrics, metrics supported investigation pairs and metrics supported Entropy.

The standard metrics were hierarchic, from the simplest to the worst, in step with various parameter DBSCAN, CLARA, and PAM.

4. Conclusion

In this paper, numerous data mining clustering Algorithms are used on the dataset to assess the most effective yielding methodology. This paper uses data mining clustering algorithms ie, PAM, CLARA, and DBSCAN are used to get a higher or better result with optimum climate demand of the crops on an finest range of best or worst temperature, and rainfall to attain more productivity of a crop. Clustering methods are in comparison to the usage of quality metrics. According to the analyses of clustering pleasant metrics, DBSCAN is used to provide a higher result than CLARA and PAM, comparably CLARA is better and more efficient than PAM.

This research work can be explored further in alkalizing the nutrients of soil condition and various other attributes to increase productivity.

References

[1] Veenadhari, S,Singh CD, Misra B, Data Mining Techniques For Predicting Crop Productivity—A review article, IJCST, 2017;

[2] CP Gleason, Large Area Yield Estimation/Forecasting Using Plant Process Models, presented at American society of agricultural engineers palmer house, Chicago, Illinois- 2018;

[3] Jain A, Flynn PJ, Murty,MN. Data Clustering: a review. ACM - 2018;31(3):264–323.

[4] Dubes RC , Jain AK, algorithms for clustering data. New jersey: Prentice Hall; 2014.

[5] Kogan J, Teboulle M, Nicholas C, A survey of clustering data mining technique, Grouping multidimensional data. Berlin: Springer; 2016. p. 25–75.

[6] Kamber M, Han J,. Data Mining: Concepts And Techniques. Morgan Kaufmann Publishers; 2011.

[7] Ester M, Sander J, Kriegel HP, Xu X. A Density-Based Algorithm For Discovering

Clusters In Large Spatial Databases With Noise. Presented at International conference on knowledge discovery and data mining. 2017.

[8] Vishnu Vardhan B, Ramesh D, Data Mining Techniques And Applications To Agricultural Yield Data. International Journal Of Advanced Research In Computer And Communication Engineering. 2018; 2(9).