Majumdar *et al. J Big Data (2017) 4:20*
DOI 10.1186/s40537-017-0077-4

Journal of Big Data

RESEARCH Open Access

Analysis of agriculture data using data mining techniques: application of big data

Jharna Majumdar*, Sneha Naraseeyappa and Shilpa Ankalaki

*Correspondence:
jharna.majumdar@gmail.com
Department of M.Tech CSE,
NMIT, Bangalore 560064,
India

Abstract

In agriculture sector where farmers and agribusinesses have to make innumerable decisions every day and intricate complexities involves the various factors influencing them. An essential issue for agricultural planning intention is the accurate yield estimation for the numerous crops involved in the planning. Data mining techniques are necessary approach for accomplishing practical and effective solutions for this problem. Agriculture has been an obvious target for big data. Environmental conditions, variability in soil, input levels, combinations and commodity prices have made it all the more relevant for farmers to use information and get help to make critical farming decisions. This paper focuses on the analysis of the agriculture data and finding optimal parameters to maximize the crop production using data mining techniques like PAM, CLARA, DBSCAN and Multiple Linear Regression. Mining the large amount of existing crop, soil and climatic data, and analysing new, non-experimental data optimizes the production and makes agriculture more resilient to climatic change.

Keywords: Big Data, PAM, CLARA and DBSCAN

Background

Today, India ranks second worldwide in the farm output. Agriculture is demographically the broadest economic sector and plays a significant role in the overall socio-economic fabric of India. Agriculture is a unique business crop production which is dependent on many climate and economy factors. Some of the factors on which agriculture is dependent are soil, climate, cultivation, irrigation, fertilizers, temperature, rainfall, harvesting, pesticide weeds and other factors. Historical crop yield information is also important for supply chain operation of companies engaged in industries. These industries use agricultural products as raw material, livestock, food, animal feed, chemical, poultry, fertilizer, pesticides, seed and paper. An accurate estimate of crop production and risk helps these companies in planning supply chain decision like production scheduling. Business such as seed, fertilizer, agrochemical and agricultural machinery industries plan production and marketing activities based on crop production estimates [1, 2]. There are 2 factors which are helpful for the farmers and the government in decision making namely:

a. It helps farmers in providing the historical crop yield record with a forecast reducing the risk management.

b. It helps the government in making crop insurance policies and policies for supply chain operation.

Data mining technique plays a vital role in the analysis of data. Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database system. Unsupervised (clustering) and supervised (classifications) are two different types of learning methods in the data mining. Clustering is the process of examining a collection of "data points," and grouping the data points into "clusters" according to some distance measure. The goal is that data points in the same cluster have a small distance from one another, while data points in different clusters are at a large distance from one another. Cluster analysis divides data into well-formed groups. Well-formed clusters should capture the "natural" structure of the data [3]. This paper focuses on PAM, CLARA and DBSCAN clustering methods. These methods are used to categorize the different districts of Karnataka which are having similar crop production.

Literature survey

Clustering is considered as an unsupervised classification process [4]. A large number of clustering algorithms have been developed for different purposes [4–6]. Clustering techniques can be categorised into Partitioning clustering, Hierarchical clustering, Density-based methods, Grid-based methods and Model based clustering methods.

Partitioning clustering algorithms, such as K-means, K-medoids PAM, CLARA and CLARANS assign objects into k (predefined cluster number) clusters, and iteratively reallocate objects to improve the quality of clustering results. Hierarchical clustering algorithms assign objects in tree structured clusters, i.e., a cluster can have data point's representatives of low level clusters [7]. The idea of Density-based clustering methods is that for each point of a cluster the neighbourhood of a given unit distance contains at least a minimum number of points, i.e. the density in the neighbourhood should reach some threshold. The idea of the density-based clustering algorithm is that, for each point of a cluster, the neighbourhood of a given unit distance has to contain at least a minimum number of points [8].

There are different forecasting methodologies developed and evaluated by the researchers all over the world in the field of agriculture. Some of such studies are: Researchers like Ramesh and Vishnu Vardhan are analysed the agriculture data for the years 1965–2009 in the district East Godavari of Andhra Pradesh, India. Rain fall data is clustered into 4 clusters by adopting the K means clustering method. Multiple linear regression (MLR) is the method used to model the linear relationship between a dependent variable and one or more independent variables. The dependent variable is rainfall and independent variables are year, area of sowing, production. Purpose of this work is to find suitable data models that achieve high accuracy and a high generality in terms of yield prediction capabilities [9].

Bangladesh offers several varieties of rice which has different cropping season [10]. For this a prior study of climate (effect on temperature and rainfall) in Bangladesh and its effect on agricultural production of rice has been done. Then this study was being taken into regression analysis with temperature and rainfall. Temperature puts an adverse

consequence on the crop production. The data has been taken from the "Bangladesh Agricultural Research Council (BARC)" for past 20 years with 7 attributes: "rainfall", "max and min temperature", "sunlight", "speed of wind", "humidity" and "cloud-coverage". In Pre-processing, the whole dataset was divided in 3 month duration phases (March to June, July to October, November to February). For this duration, the average for every attribute has been taken and associated with it. This pre-processing has been done for each kind of rice variety. In clustering, the different pre-processed table has been analysed to find the sharable group of region based on similar weather attribute.

Soil characteristics are studied and analysed using data mining techniques. As an example, the k-means clustering is used for clustering soils in combination with GPS-based technologies [11]. Authors like Alberto Gonzalez-Sanchez, Juan Frausto-Solis and Waldo Ojeda-Bustamante have done extensive study on predictive ability of machine learning techniques such as multiple linear regression, regression trees, artificial neural network, support vector regression and k-nearest neighbour for crop yield production [12]. Wheat yield prediction using machine learning and advanced sensing techniques has done by Pantazi, DimitriosMoshou, Thomas Alexandridis and Abdul Mounem-Mouazen [13]. The aim of their work is to predict within field variation in wheat yield, based on on-line multi-layer soil data, and satellite imagery crop growth characteristics. Supervised self-organizing maps capable of handling existent information from different soil and crop sensors by utilizing an unsupervised learning algorithm were used. The software tool 'Crop Advisor' has been developed by S. Veenadhari, B. Misra and CD Singh [14] is an user friendly web page for predicting the influence of climatic parameters on the crop yields. C4.5 algorithm is used to find out the most influencing climatic parameter on the crop yields of selected crops in selected districts of Madhya Pradesh.

Methods

The objective of proposed work is to analyse the agriculture data using data mining techniques. In proposed work, agriculture data has been collected from following sources:

Dataset in agricultural sector [https://data.gov.in/, http://raitamitra.kar.nic.in/statistics],

Crop wise agriculture data [html://CROPWISE_NORMAL_AREA],

Agriculture data of different districts [http://14.139.94.101/fertimeter/Distkar.aspx], http://raitamitra.kar.nic.in/ENG/statistics.asp],

Agriculture data based on weather, temperature, and relative humidity [http://dmc.kar.nic.in/trg.pdf].

Input dataset consist of 6 year data with following parameters namely: year, State-Karnataka (28 districts), District, crop (cotton, groundnut, jowar, rice and wheat.), season (kharif, rabi, summer), area (in hectares), production (in tonnes), average temperature (°C), average rainfall (mm), soil, PH value, soil type, major fertilizers, nitrogen (kg/Ha), phosphorus (Kg/Ha),Potassium(Kg/Ha), minimum rainfall required, minimum temperature required.

In proposed work, modified approach of DBSCAN method is used to cluster the data based on districts which are having similar temperature, rain fall and soil type. PAM and CLARA are used to cluster the data based on the districts which are producing maximum crop production (In proposed work wheat crop is considered as example). Based

on these analyses we are obtaining the optimal parameters to produce the maximum crop production. Multiple linear regression method is used to forecast the annual crop yield.

Modified approach of DBSCAN

DBSCAN is a base algorithm for density based clustering containing large amount of data which has noise and outliers. DBSCAN has two parameters namely Eps and MinPts. However, traditional DBSCAN cannot produce optimal Eps value [15]. Determination of the optimal Eps value automatically is the one of the most necessary modification for the DBSCAN. Figure 1 briefs the modified approach of the DBSCAN method.

Modified DBSCAN proposes the method to find the minimum points and Epsilon (radius value) automatically. KNN plot is used to find out the epsilon value where input to the KNN plot (K value) is user defined. To avoid the user define K value as input to the KNN plot, Batchelor Wilkins clustering algorithm is applied to the database and obtain the K value along with its respective cluster centres. This K value is given as input to the KNN Plot.

Determination of Eps and Minpts

The Epsilon (Eps) value can be found by drawing a "K-distance graph" for entire data-points in dataset for a given 'K', obtained by the Batchelor Wilkins Algorithm [16]. Initially, the distance of a point to every 'K' of its nearest-neighbours is calculated. KNN plot is plotted by taking the sorted values of average distance values. When the graph is plotted, a knee point is determined in order to find the optimal Eps value [15].

Partition around medoids (PAM)

It is a partitioning based algorithm. It breaks the input data into number of groups. It finds a set of objects called medoids that are centrally located. With the medoids, nearest data points can be calculated and made it as clusters. The algorithm has two phases:

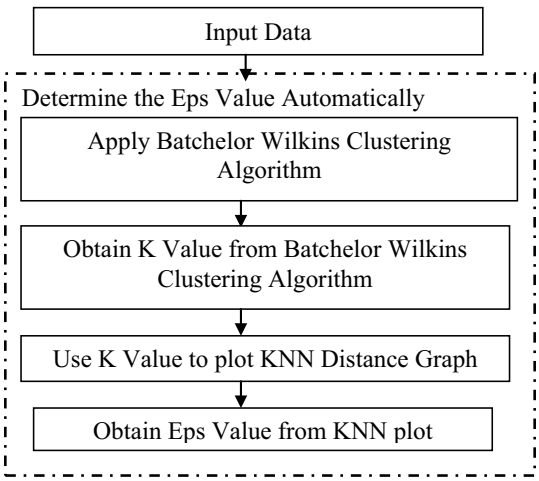


Fig. 1 Determine the Eps value automatically

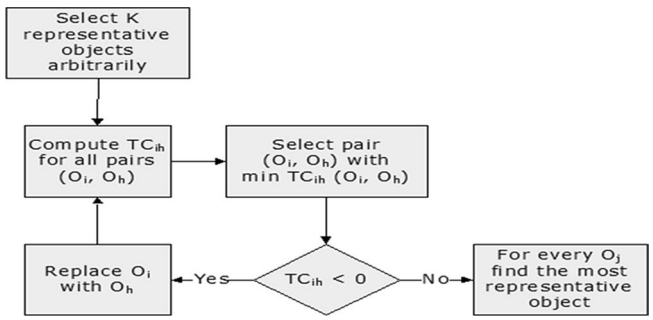


Fig. 2 PAM Algorithm steps

1. BUILD phase, a collection of k objects are selected for an initial set S.

- Arbitrarily choose k objects as the initial medoids.
- Until no change, do.

  – (Re) assign each object to the cluster with the nearest medoid.
  – Improve the quality of the k-medoids (randomly select a non medoid object, O random, compute the total cost of swapping a medoid with O random).

2. SWAP phase, one tries to improve the quality of the clustering by exchanging selected objects with unselected objects. Choose the minimum swapping cost.

Example: For each medoid m1, for each non-medoid data point d; Swap m1 and d, recompute the cost (sum of distances of points to their medoid), if total cost of the configuration increased in the previous step, undo the swap Fig. 2 depicts the steps involved the PAM algorithms.

CLARA (clustering large applications)

It is designed by Kaufman and Rousseeuw to handle large datasets, CLARA (clustering large applications) relies on sampling [17, 18]. Instead of finding representative objects for the entire data set, CLARA draws a sample of the data set, applies PAM on the sample, and finds the medoids of the sample. To come up with better approximations, CLARA draws multiple samples and gives the best clustering as the output. Here, for accuracy, the quality of the clustering is measured based on the average dissimilarity of all objects in the entire data set. Figure 3 briefs about the steps involved in the CLARA Algorithm.

Multiple linear regression to forecast the crop yield

Multiple linear regression is a variant of "linear regression" analysis. This model is built to establish the relationship that exists between one dependent variable and two or more independent variables [19].For a given dataset where $x_1 \ldots x_k$ are independent variables and Y is a dependent variable, the multiple linear regression fits the dataset to the model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

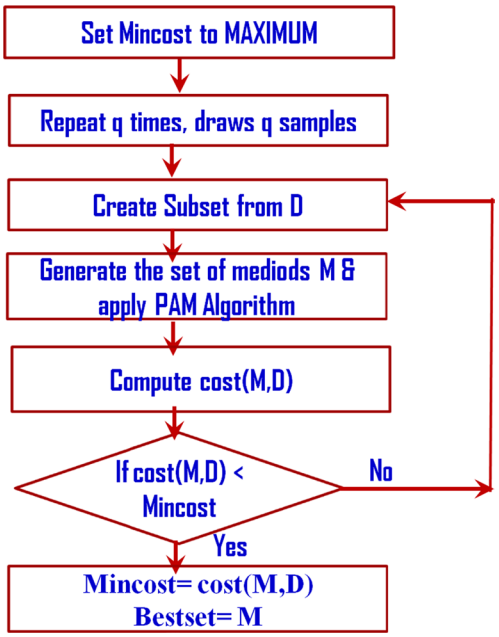


Fig. 3 Steps involved in CLARA algorithm

where β_0 is the y-intercept and $\beta_1, \beta_2, \ldots, \beta_k$ parameters are called the partial coefficients. In matrix form

$$\mathrm{Y} = \mathrm{XB} + \mathrm{E}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} E = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

Before applying the multiple linear regression to forecast the crop yield, it's necessary to know the significant attributes from the database. All the attributes used in the database will not be significant or changing the value of these attributes will not affect anything on the dependent variables. Such attributes can be neglected. P value test is performed on the database to find the significant attributes and multiple linear regression is applied only on the significant values to forecast the crop yield.

Evaluation methods

Data mining algorithms work with different principles, being able to be influenced by different kinds of associations on data. To ensure fairer conditions in evaluation, this work finds the optimal clustering method for agriculture data analysis. Proposed work adopts the external quality metrics [3] like Purity, Homogeneity, Completeness, V Measure, Rand Index, Precision, Recall and F measure to compare the PAM, CLARA and DBSCAN clustering methods.

```
.....................................
Total number of clusters : 7
.....................................
cluster point 1 is : BIJAPUR
.....................................
cluster point 2 is : DAKSHINAKANNADA
.....................................
cluster point 3 is : KOPPAL
.....................................
cluster point 4 is : SHIMOGA
.....................................
cluster point 5 is : UTTARAKANNADA
.....................................
cluster point 6 is : BAGALKOT
.....................................
cluster point 7 is : CHIKMAGALUR
```

Fig. 4 Cluster centres obtained from the Batchelor Wilkins algorithm

Purity of the clustering is computed by assigning each cluster to the class which is most frequent in the cluster. Homogeneity represents the each cluster contains only members of a single class. Completeness represents the all members of a given class are assigned to the same cluster. V-measure is computed as the harmonic mean of distinct homogeneity and completeness scores. Rand Index measures the percentage of decisions that are correct. Precision is calculated as the fraction of pairs correctly put in the same cluster. Recall represents the fraction of actual pairs that were identified. F measure indicates the harmonic mean of precision and recall. Higher quality metrics value represents the better cluster quality.

Experimental results

Modified approach of DBSCAN

Before applying DBSCAN algorithm on the dataset user needs to determine the Minpts and Eps values. The Batchelor Wilkins algorithm is applied on the dataset in order to determine the K value (Number of clusters) automatically. For the dataset used in the proposed work, K value obtained from the Batchelor Wilkins is 7 with following districts as cluster centres. Results of Batchelor Wilkins algorithm are shown in Fig. 4.

KNN plot is plotted using K value obtained from the Batchelor & Wilkins' Algorithm to determine the epsilon value and the min points for the DBSCAN.

Figure 5 depicts the result of KNN plot. The KNN plot is plotted using K value obtained from the Batchelor & Wilkins' Algorithm (i.e. here $K = 7$). Eps value is calculated by taking the slope of the line from any point and sought-after pair of points that have the greatest slope to locate the point. The slope of the line is located at the point of 0.4, a point which is the optimal value Eps [20].

DBSCAN clustering algorithm is applied on the dataset to cluster the different districts of Karnataka which are having similar rain fall, temperature and soil type using optimal Eps value.

Figure 6 depicts the different districts of Karnataka which are considered for the purpose of analysis.

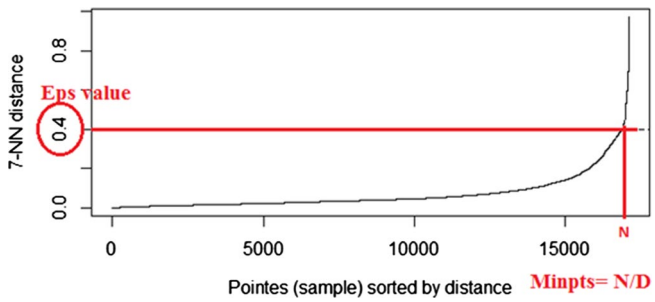


Fig. 5 KNN plot for given dataset

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			

Fig. 6 Districts of Karnataka considered for the analysis

Figures 7, 8 and 9 depicts the different districts of Karnataka which are having similar temperature range, rain fall range and soil types respectively.

PAM

To apply the PAM algorithm on the dataset, initially user need to give k (Number of clusters), where k is given as 3 in current experiment. Crop yield is categorised into LOW, MODERATE and HIGH production. Total districts are clustered into 3 clusters using PAM clustering method. Resultant clusters are shown in the Table 1.

- Study and analysis of wheat crop production in different districts of Karnataka as shown in Fig. 10.

As a result of the analysis, North Karnataka districts such as Bijapur, Dharwad, Bagalkot, Belgaum, Raichur, Bellary, Chitradurga and Davangere are the districts which have maximum wheat crop production.

CLARA

Districts in the dataset are clustered into 3 clusters using CLARA algorithm. Clusters are shown in the Fig. 11. It represents the districts which are having similar factors like

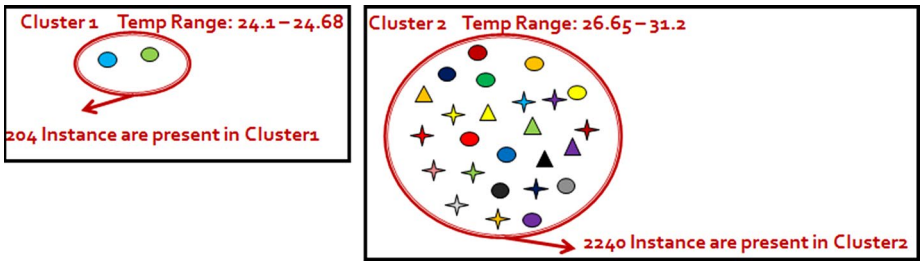


Fig. 7 Districts which are having similar temperature range

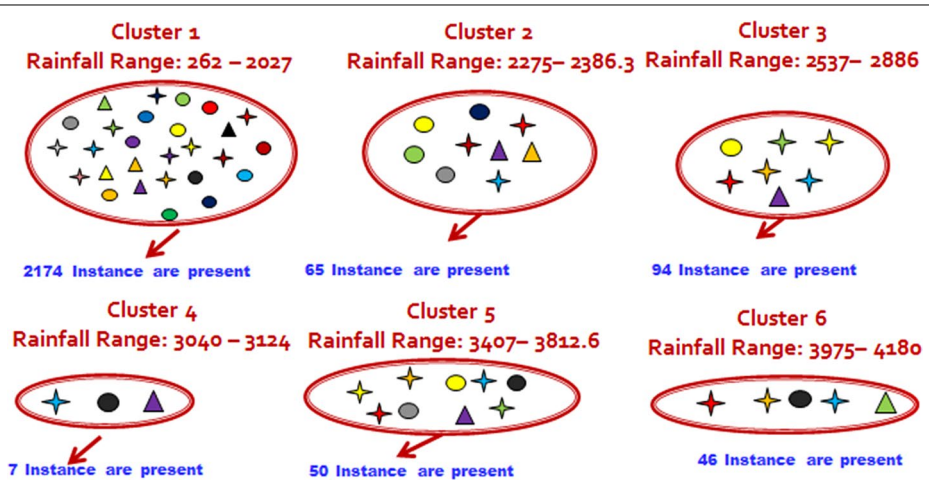


Fig. 8 Districts which are having similar rain fall during 6 year duration

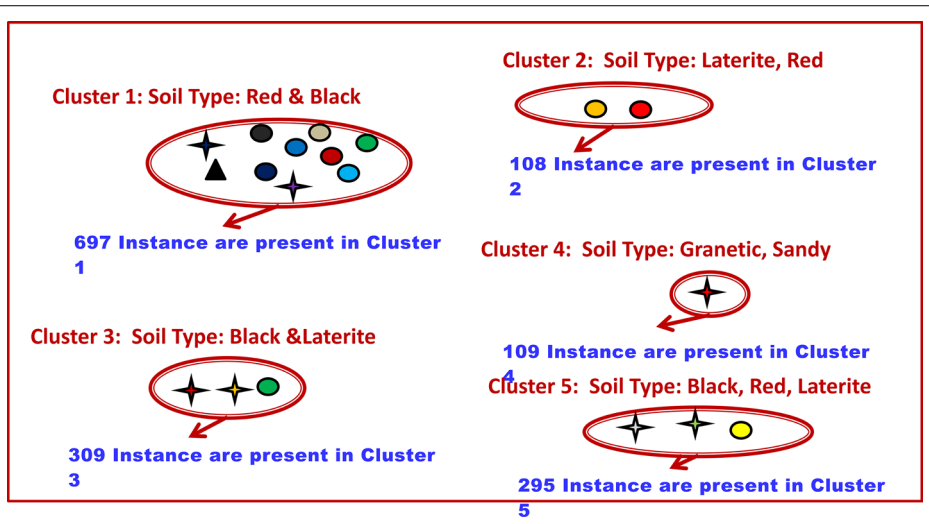


Fig. 9 Districts which are having similar soil type

Table 1 Results of PAM algorithm

Low-moderate production	High production	Moderate-high production
Mandya, Raichur, Gadag, Gulbarga, Bellary	Koppal, Dharwad, Haveri, Bijapur, Bidar, Chamarajannagar, Belgaum, Tumkur	Davangere, Shimoga, Chikmagalur, Bangalore

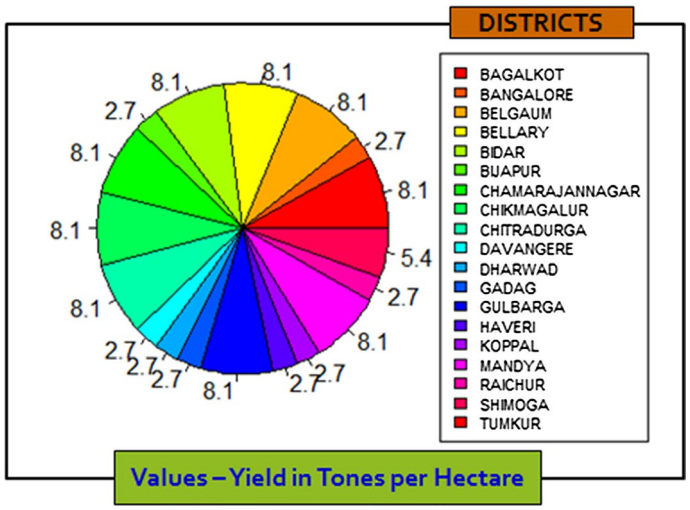


Fig. 10 Production of yield in tonnes per hectare of different districts

area, production, rainfall and temperature. Result of the CLARA algorithm is shown in the Table 2.

- Study and analysis of temperature and wheat crop production in different districts of Karnataka as shown in Fig. 12. From the Fig. 12, we can analyze that the optimal temperature for Wheat crop production is 29.9 °C.

Fig. 11 Results of CLARA in R language

Table 2 Results CLARA algorithm

Large area, production and moderate rainfall, temperature (24–26)	Moderate area, production and high rainfall, temperature (27–29)	Low area, production moderate rainfall, temperature (29–30)
Bijapur, Belgaum	Gadag, Gulbarga, Dharwad, Bangalore, Bagalkote	Koppal, Davangere, Shimoga, Haveri, Chikmagalur, Bidar, Chamarajannagar, Tumkur, Mandya, Raichur, Bellary

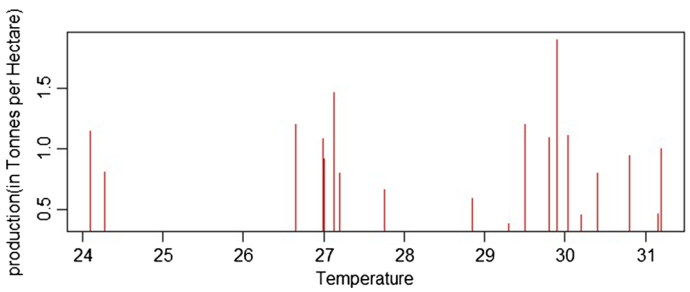


Fig. 12 Plot temperature vs. production

Multiple linear regression

Before applying the multiple linear regression, the "p value test" is performed on the dataset to determine the significant attributes. Table 3 depicts the significant values. An independent variable which has a "p value" of less than 0.05, specifies that the "null-hypothesis" can be rejected means it will have effect on regression analysis. So these independent values can be added to the model. Whereas if the p value is more than common alpha level i.e. 0.05, the variable will said to be not significant to the model.

Table 4 shows the multiple linear regression equation for different crop yield. For example, for Wheat crop, if all the independent variables are zero, the yield becomes 112. 1 unit increase in temperature level reduces the yield by 4.14e−02 units, 1 unit increase in rainfall will increase yield by 1.34e−04 units, 1 unit increase in pH will increase the yield by 0.079153 units, 1 unit increase in Nitrogen reduces the yield by 1.31e−03 units, 1 unit increase in potassium level decreases the yield by 0.00167 units and 1 unit increase in water requirement decreases the yield by 0.28125 unit.

Table 3 P value test: significant attributes

	Cotton	Groundnut	Jowar	Rice	Wheat
Temperature	*0.547536*	3.41E−07	3.86E−07	0.003139	0.001137
Rainfall	*0.784625*	1.86E−06	*0.653187*	*0.105878*	0.018042
Ph	0.011752	2.55E−05	0.029733	5.08E−07	0.01834
Nitrogen	5.85E−05	*0.071873*	*0.349257*	0.000841	8.6E−06
Phosphorus	*0.071843*	0.043345	*0.464847*	0.025816	*0.209524*
Potassium	2.82E−07	*0.643528*	0.050831	1.43E−05	0.021422
Water	4.95E−05	4.92E−49	1.2E−102	1.22E−26	NA

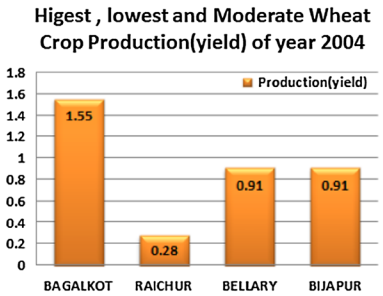
The italicized cells are representing the insignificant independent attributes for each crop as the values are more than 0.05. Regression Equation is formed using the independent variables

Table 4 Multiple linear regression equation for different crop yield

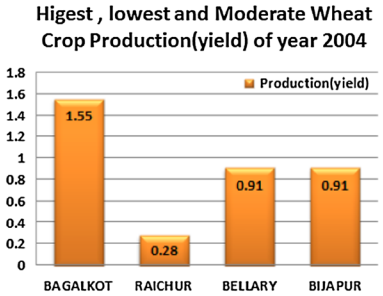
Crop	Yield forecast equation
Cotton	Yield = (7.149372) + (−0.14468)pH + (−0.00131) Nitrogen + (−0.00405) Potassium + (−0.00405) Water Required
Groundnut	Yield = (2.79115) + (0.029217) Temperature + (5.78e−05) Rainfall + (−0.05681) pH + (−0.00127) Phosphorus + (−0.00492) Water Required
Jowar	Yield = (−1.62694) + (−5.35e−02) Temperature + (0.051512) pH + (−0.00113) Potassium + (0.01685436) Water Required
Rice	Yield = (−0.18503) + (0.041593) Temperature + (0.172042) pH + (−8.27e−04) Nitrogen + (−4.28e−03) Phosphorus + (−0.00264) Potassium + (9.15e−04) Water Required
Wheat	Yield = (112) + (−4.14e−02) Temperature + (1.34e−04) Rainfall + (0.079153) pH + (−1.31e−03) Nitrogen + (−0.00167) Potassium + (−0.28125) Water Required

For 1 unit increase in pH, the crops like Jowar, Rice, and Wheat yield will increase but Groundnut and Cotton yield will decrease.

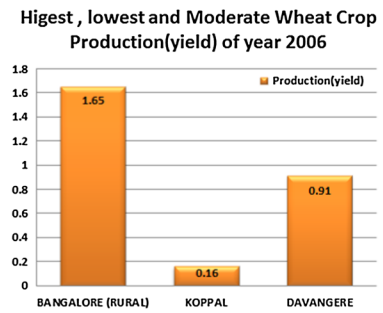
Results for optimal temperature and rainfall for wheat—Table 5



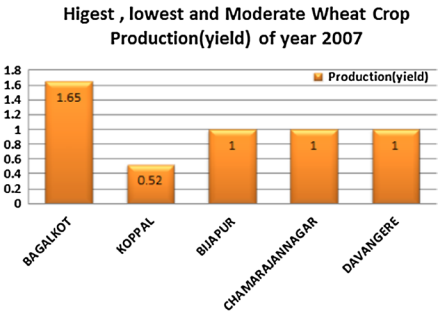
I .Yield plot-2004



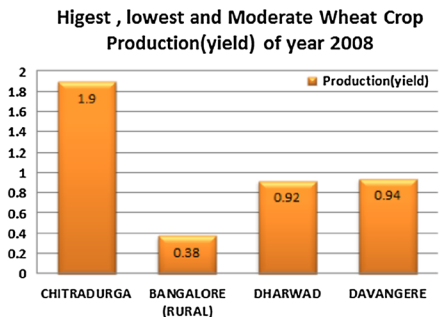
II. Yield plot-2005



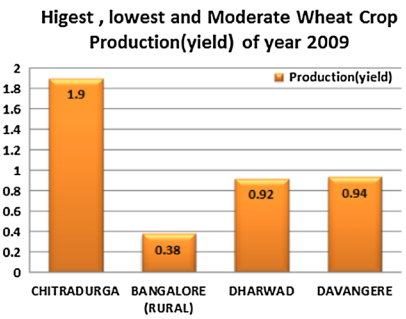
III. Yield plot-2006



IV. Yield plot-2007



V. Yield plot-2008



VI. Yield plot-2009

Table 5 shows the optimal parameters to achieve the higher wheat production.

Table 5 Optimal parameters to achieve higher production

Optimal parameters to achieve higher production	
Optimal temp	25.4–29.9
Worst temp	30.2–31.15
Rainfall	548–580

Comparison of clustering methods

As mentioned earlier, clustering comparison has done using four performance quality metrics. Table 6 shows the comparison of PAM, CLARA and DBSCAN methods for clustering the districts which are having similar crop productivity.

Table 6 and Fig. 13 depicts the comparison of PAM, CLARA and DBSCAN clustering methods. Higher quality metric values indicates better clustering quality. Analysis of the quality metrics parameters for different clustering methods is shown in the Fig. 13. From Fig. 13, DBSCAN has higher value for most of the quality metrics parameter. DBSCAN gives the better clustering quality than PAM and CLARA, CLARA gives the better clustering quality than the PAM.

Discussion

The crops are usually selected by its economic importance. However, the agricultural planning process requires a yield estimation of several crops. In this sense, five crops were selected for this work using the data availability as the key measure. Thus, a crop

Table 6 Comparison of clustering methods

Number of clusters k = 3			
	PAM	**CLARA**	**DBSCAN**
Purity	0.578947	0.631578	0.708512
Homogeneity	0.853526	0.879624	0.895275
Completeness	0.758264	0.782356	0.786854
V-measure	0.814447	0.805181	0.83757
Precision	0.40369	0.415365	0.42152
Recall	0.24856	0.25634	0.25655
F-measure	0.307677	0.317028	0.318966
Rand index	0.785364	0.796352	0.814561

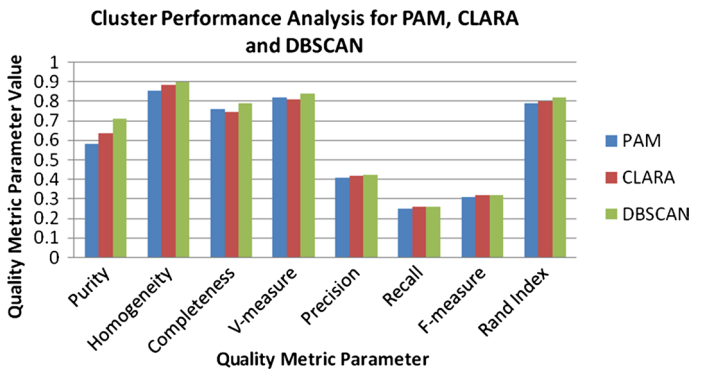


Fig. 13 Comparison of clustering methods

was selected when enough data samples appeared in the range of 6 years under analysis. In presents works, research is commonly limited to the 5 crops those are cotton, wheat, ground nut, jowar and rice. Example wheat crop analysis is discussed in this paper.

The present work covers the PAM, CLARA, Modified DBSCAN clustering methods and multiple linear regression method. PAM and CLARA are the traditional clustering methods where as DBSCAN method is modified by introducing the Batchelor Wilkins clustering method to determine the 'k' value and KNN method to determine the minimum points and radius value automatically. Using these methods crop data set is analysed and determined the optimal parameters for the wheat crop production. Multiple linear regression is used to find the significant attributes and form the equation for the yield prediction.

Some works measure the quality of the clustering methods using internal quality metrics [21], some other uses the external quality metrics. However, in these works, research is limited to the external quality metrics which are combination of several metrics those are [22]: set matching metrics, metrics based on counting pairs and metrics based on Entropy. The quality metrics were ranked, from the best to the worst, according to purity, homogeneity, completeness, v measure, precision, recall and rand index results, in the following order: DBSCAN, CLARA and PAM.

Conclusion

Various data mining techniques are implemented on the input data to assess the best performance yielding method. The present work used data mining techniques PAM, CLARA and DBSCAN to obtain the optimal climate requirement of wheat like optimal range of best temperature, worst temperature and rain fall to achieve higher production of wheat crop. Clustering methods are compared using quality metrics. According to the analyses of clustering quality metrics, DBSCAN gives the better clustering quality than PAM and CLARA, CLARA gives the better clustering quality than the PAM. The proposed work can also be extended to analyse the soil and other factors for the crop and to increase the crop production under the different climatic conditions.

Authors' contributions

JM, Dean R&D, Prof & HOD of Dept of M.Tech CSE at NMIT, has 40 years of experience in India and abroad has guided and given extensive help to develop the data mining algorithms. SN, Assistant Professor of Dept of M.Tech CSE at NMIT has developed the PAM and CLARA algorithms with the help of Dr. Jharna Majumdar. SA Assistant Professor of Dept of M.Tech CSE at NMIT has developed Modified approach of DBSCAN, Multiple Linear Regression and quality metrics for cluster comparison with the guidance and help of Dr. Jharna Majumdar. All authors together analysed the crop data set to determine the optimal parameters to maximise the crop yield. All authors read and approved the final manuscript.

Acknowledgements

The authors express their sincere gratitude to Prof N.R Shetty, Advisor and Dr H.C Nagaraj, Principal, Nitte Meenakshi Institute of Technology for giving constant encouragement and support to carry out research at NMIT.

The authors extend their thanks to Vision Group on Science and Technology (VGST), Government of Karnataka to acknowledge our research and providing financial support to setup the infrastructure required to carry out the research.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Research Department of Computer science, Nitte Meenakshi Institute of Technology.

Publisher's Note

Received: 25 February 2017 Accepted: 31 May 2017
Published online: 05 July 2017

References

1. Veenadhari S, Misra B, Singh CD. Data mining techniques for predicting crop productivity—A review article. In: IJCST. 2011; 2(1).
2. Gleaso CP. Large area yield estimation/forecasting using plant process models.paper presentation at the winter meeting American society of agricultural engineers palmer house, Chicago, Illinois. 1982; 14–17
3. Majumdar J, Ankalaki S. Comparison of clustering algorithms using quality metrics with invariant features extracted from plant leaves. In: Paper presented at international conference on computational science and engineering. 2016.
4. Jain A, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv. 1999;31(3):264–323.
5. Jain AK, Dubes RC. Algorithms for clustering data. New Jersey: Prentice Hall; 1988.
6. Berkhin P. A survey of clustering data mining technique. In: Kogan J, Nicholas C, Teboulle M, editors. Grouping multi-dimensional data. Berlin: Springer; 2006. p. 25–72.
7. Han J, Kamber M. Data mining: concepts and techniques. Massachusetts: Morgan Kaufmann Publishers; 2001.
8. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Paper presented at International conference on knowledge discovery and data mining. 1996
9. Ramesh D, Vishnu Vardhan B. Data mining techniques and applications to agricultural yield data. In: International journal of advanced research in computer and communication engineering. 2013; 2(9).
10. MotiurRahman M, Haq N, Rahman RM. Application of data mining tools for rice yield prediction on clustered regions of Bangladesh. IEEE. 2014;2014:8–13.
11. Verheyen K, Adrianens M, Hermy S Deckers. High resolution continuous soil classification using morphological soil profile descriptions. Geoderma. 2001;101:31–48.
12. Gonzalez-Sanchez Alberto, Frausto-Solis Juan, Ojeda-Bustamante W. Predictive ability of machine learning methods for massive crop yield prediction. Span J Agric Res. 2014;12(2):313–28.
13. Pantazi XE, Moshou D, Alexandridis T, Mouazen AM. Wheat yield prediction using machine learning and advanced sensing techniques. Comput Electron Agric. 2016;121:57–65.
14. Veenadhari S, Misra B, Singh D. Machine learning approach for forecasting crop yield based on climatic parameters. In: Paper presented at international conference on computer communication and informatics (ICCCI-2014), Coimbatore. 2014.
15. Rahmah N, Sitanggang IS. 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;31:012012.
16. Forbes G. The automatic detection of patterns in people's movements. Dissertation, University of Cape Town. 2002.
17. Ng RT, Han J. CLARANS: A Method for Clustering Objects for Spatial Data Mining. In: IEEE Transactions on Knowledge and Data Engineering. 2002; 14(5).
18. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Wiley. 1990. doi:10.1002/9780470316801.
19. Multiple linear regression-http://www.originlab.com/doc/Origin-Help/Multi-Regression-Algorithm. Accessed 3 July 2017.
20. Elbatta MNT. An improvement for DBSCAN algorithm for best results in varied densities. Dissertation, Gaza (PS): Islamic University of Gaza. 2012
21. Kirkl O, De La Iglesia B. Experimental evaluation of cluster quality measures. 2013. 978-1-4799-1568-2/13. IEEE.
22. Meila M (2003) Comparing clustering. In: Proceedings of COLT 2003.