Title:

Optimizing Crop Production through Data Mining Techniques in Agriculture Sector: A Comprehensive Analysis

Submitted By

Student 1: Soumil Mukhopadhyay, Roll Number:16010122257

Student 2: Mohanish Podhade, Roll Number: 16010122229

Student 3: Shrusti Vora, Roll Number: 16010122233

Student 4: Akshat Yadav, Roll Number:16010122221

Guide

Professor:Dr. Kaustubh Kulkarni

Department: COMPS

Abstract

In the agriculture sector, where decisions are pivotal and influenced by a multitude of factors, accurate yield estimation is essential for effective planning. This paper delves into the application of data mining techniques, particularly PAM, CLARA, and DBSCAN, to analyze agricultural data and optimize crop production. By leveraging large datasets encompassing crop, soil, and climatic information, these techniques aim to enhance production and fortify agriculture against climatic variability. Key parameters such as temperature, rainfall, and soil type are scrutinized to identify optimal conditions for maximizing crop yield. Additionally, multiple linear regression is employed for yield forecasting, contributing to informed decision-making in farming practices. The study also explores the utility of clustering methods in categorizing districts based on crop productivity, facilitating tailored interventions and policy formulation. Through external quality metrics, the efficacy of clustering methods is evaluated, with DBSCAN emerging as the superior approach. The paper concludes by highlighting the potential for further research in soil analysis and the integration of additional factors to bolster crop production in diverse climatic settings.

Keywords: Big Data, PAM, CLARA, DBSCAN

Contents

Nomenclature

PAM	Partition around medoids
CLARA	clustering large applications
DBSCAN	density-based spatial clustering of applications with noise

1. Introduction

In the dynamic landscape of agriculture, where decisions profoundly impact productivity and sustainability, the accurate estimation of crop yield holds paramount importance. Various factors, including soil characteristics, climate conditions, and agricultural practices, intricately shape the agricultural outcome. Consequently, leveraging advanced analytical techniques becomes imperative to harness the wealth of data available in this domain. This paper explores the application of data mining methodologies, specifically focusing on PAM, CLARA, and DBSCAN algorithms, to dissect agricultural data and optimize crop production. By scrutinizing historical data and employing predictive modeling, the study aims to equip stakeholders with actionable insights to enhance agricultural productivity.

1.1 Background

Agriculture serves as the cornerstone of the Indian economy, contributing significantly to GDP and employment. With India ranking second globally in farm output, the sector's resilience and adaptability are critical for sustained growth and food security. However, agriculture's vulnerability to climatic fluctuations and economic factors underscores the need for robust planning and risk management strategies. Traditional methods of crop estimation often fall short in capturing the complexity of agricultural systems, necessitating the adoption of data-driven approaches. By harnessing the power of big data and advanced analytics, stakeholders can gain deeper insights into crop dynamics, optimize resource allocation, and mitigate risks effectively.

1.2 Motivation

The burgeoning availability of agricultural data, coupled with advancements in data mining techniques, presents a unique opportunity to revolutionize farming practices. By leveraging algorithms such as PAM, CLARA, and DBSCAN, stakeholders can unearth hidden patterns within vast datasets, enabling informed decision-making and resource optimization. Moreover, the integration of multiple linear regression for yield forecasting further enhances the predictive capabilities of the analysis. The potential to cluster districts based on crop productivity opens avenues for tailored interventions and policy formulation, thereby fostering sustainable agricultural practices. This paper seeks to bridge the gap between data analytics and agriculture, offering actionable insights to enhance productivity and resilience in the sector.

2. Literature survey:

Clustering is widely acknowledged as an essential unsupervised classification method in various fields [4]. Numerous clustering algorithms have been developed to address different requirements [4–6]. These algorithms can be broadly categorized into Partitioning clustering, Hierarchical clustering, Density-based methods, Grid-based methods, and Model-based clustering methods.

Partitioning clustering algorithms like K-means, K-medoids PAM, CLARA, and CLARANS categorize objects into k predetermined clusters, refining the clustering quality iteratively. On the other hand, Hierarchical clustering organizes objects into tree-structured clusters, allowing lower-level clusters to be nested within higher-level ones [7]. Density-based methods determine clusters based on the density of points within a given neighborhood, ensuring that each point's neighborhood contains a minimum number of points [8].

In the realm of agricultural forecasting, researchers have explored various methodologies worldwide. For instance, Ramesh and Vishnu Vardhan analyzed agriculture data spanning from 1965 to 2009 in the East Godavari district of Andhra Pradesh, India. They employed K-means clustering to categorize rainfall data into four clusters and utilized Multiple Linear Regression (MLR) to model the relationship between rainfall and variables like year, sowing area, and production [9].

Similarly, studies have been conducted in Bangladesh to understand the impact of climate, particularly temperature and rainfall, on rice production. Majumdar et al. conducted regression analysis using climate data sourced from the Bangladesh Agricultural Research Council (BARC) over the past two decades. They divided the data into three-month intervals and analyzed attributes such as rainfall, temperature, sunlight, wind speed, humidity, and cloud coverage to identify regions with similar weather attributes [10].

Furthermore, data mining techniques have been applied to analyze soil characteristics. For example, k-means clustering combined with GPS-based technologies has been employed to cluster soils [11]. Additionally, machine learning techniques such as Multiple Linear Regression, Regression Trees, Artificial Neural Networks, Support Vector Regression, and k-Nearest Neighbor have been explored for crop yield prediction [12]. Pantazi, Dimitrios Moshou, Tomas Alexandridis, and Abdul Mounem-Mouazen focused on wheat yield prediction using machine learning and advanced sensing techniques based on soil data and satellite imagery [13].

Moreover, software tools like 'Crop Advisor' have been developed to predict crop yields based on climatic parameters [14]. These tools utilize algorithms like C4.5 to identify the most influential climatic parameters on crop yields in specific districts.

This literature review underscores the diverse applications of data mining techniques in agriculture, ranging from clustering for soil analysis to predictive modeling for crop yield optimization.

3.Conclusions and Scope for further work:

3.1 Conclusions:

This study demonstrates the effectiveness of data mining techniques like PAM, CLARA, and DBSCAN in optimizing crop production. Integration of multiple linear regression enhances predictive capabilities. Clustering districts based on productivity prioritizes interventions effectively. DBSCAN emerges as a superior clustering approach.

3.2Scope for Further Work:

1. Integrate additional data sources like satellite imagery for more accurate predictions.
2. Refine clustering algorithms for improved accuracy and scalability.
3. Incorporate spatial analysis to understand spatial dependencies in agricultural data.
4. Employ longitudinal analysis to identify temporal trends in crop productivity.
5. Explore the influence of socioeconomic factors on agricultural outcomes for more equitable policies.

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.

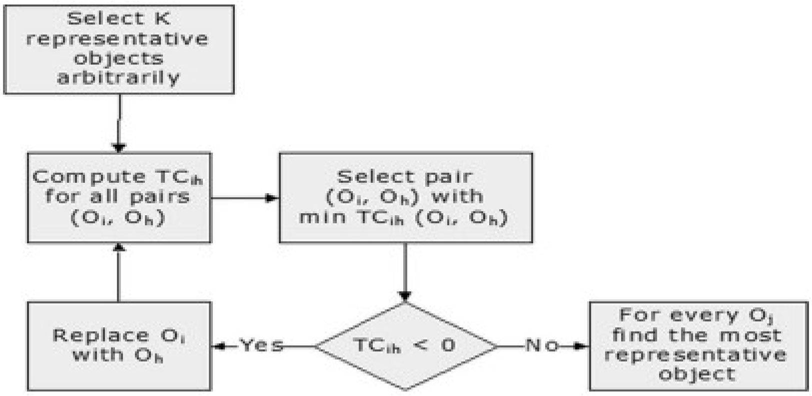
4. Jain A, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv. 1999;31(3):264–323.
5. Jain AK, Dubes RC. Algorithms for clustering data. New Jersey: Prentice Hall; 1988.
6. Berkhin P. A survey of clustering data mining technique. In: Kogan J, Nicholas C, Teboulle M, editors. Grouping multidimensional 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 classifcation using morphological soil profle 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

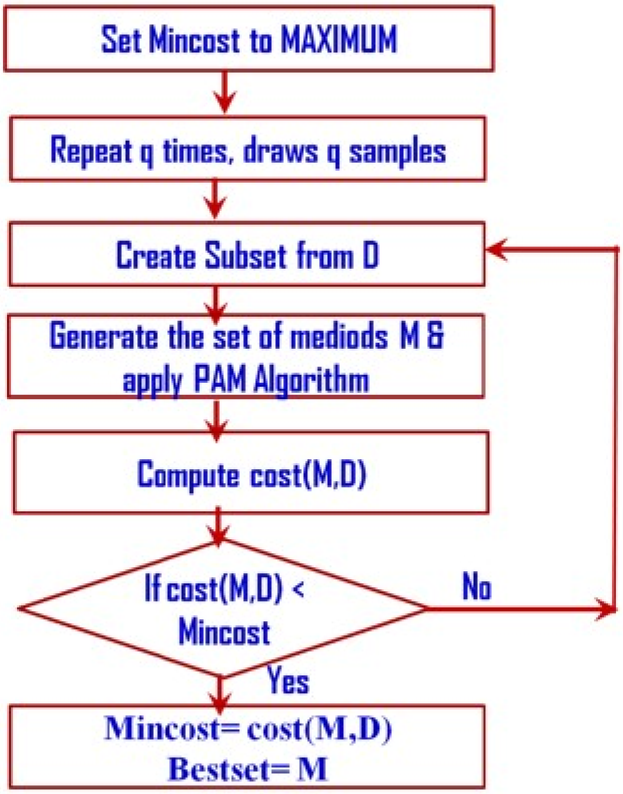
4. List of Figures

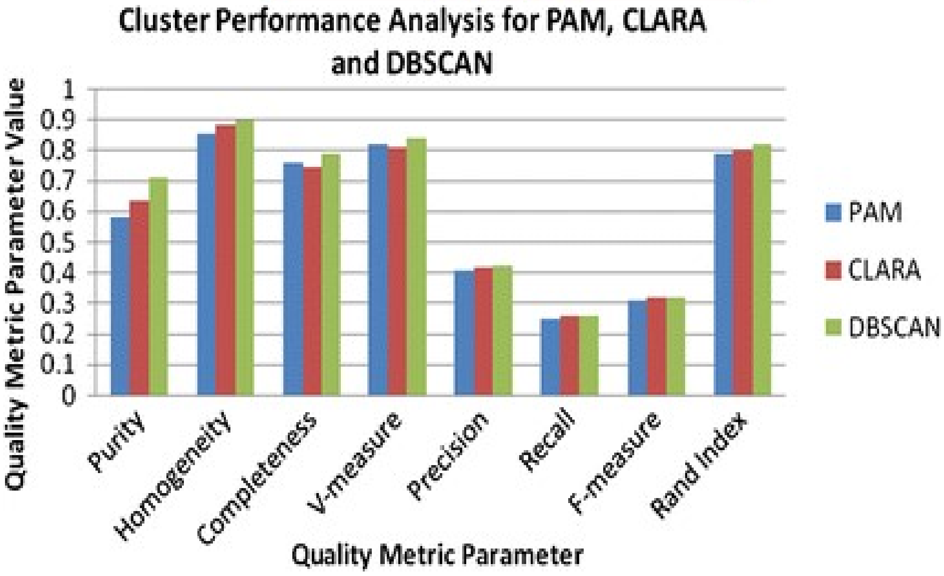
4.1Vector representation of document

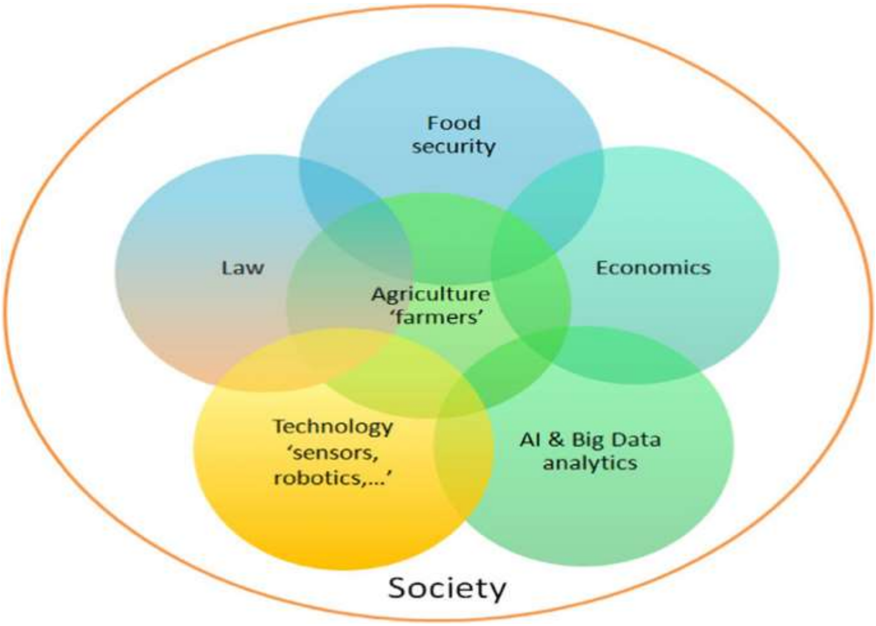
PAM:



CLARA:


Set Mincost to MAXIMUM
Repeat q times, draws q samples
Create Subset from D
Generate the set of mediods M &
apply PAM Algorithm
Compute cost(M,D)
If cost(M,D) <
Mincost
No
Yes
Mincost= cost(M,D)
Bestset= M


Cluster Performance Analysis for PAM, CLARA
and DBSCAN
Quality Metric Parameter Value
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Purity
Homogeneity
Completeness
V-measure
Precision
Recall
F-measure
Rand Index
Quality Metric Parameter
PAM
CLARA
DBSCAN



4.2 Citation:

TY - JOUR
AU - Majumdar, Jharna
AU - Naraseeyappa, Sneha
AU - Ankalaki, Shilpa
PY - 2017
DA - 2017/07/05
TI - Analysis of agriculture data using data mining techniques: application of big data
JO - Journal of Big Data
SP - 20
VL - 4
IS - 1
AB - 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.
SN - 2196-1115
UR - https://doi.org/10.1186/s40537-017-0077-4

DO - 10.1186/s40537-017-0077-4
ID - Majumdar2017
ER -

Plagiarism index: