EasyChair Preprint

№ 793

A Revisit to Clustering Techniques with its Application in Agriculture Sector

Vyoma Srivastava, Dr. K. K. Aggarwal and
Abhay Kumar Srivastava

February 25, 2019

A revisit to Clustering Techniques with its Application in Agriculture Sector

Review Article

Vyoma Srivastava
Dept. of Computer Science
Jaipur National University
Jaipur, India
vyoma.phdscholar@jnujaipur.ac.in

Dr.K.K Aggarwal
Dept. of Computer Science
Jaipur National University
Jaipur, India

Dr. Abhay Kumar Srivastava
Department in Business School
Amity Business School
Noida, India

ABSTRACT

Inspite of the big challenge of missing data for data mining algorithms, Data mining has made a great progress in recent years. Data Mining is a process where its techniques are used to extract some useful knowledge from a large data base. It is very helpful since it helps human race discover knowledge out of data and presenting it in a form that is easily understood. Data mining uses various methods and techniques which allows analysis very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details.[1]

KEYWORDS:
Data mining, clustering, classification of clustering, Mixed Attributes, Algorithms

QUICK GLANCE COMPARISON OF CLUSTERING TECHNIQUES[2]:

Name	Algorithm	Key-Idea	Type of Data	Advantages	Disadvantage
Partitional	K-means	Mean Centroid	Numerical	Simple	Sensitive to outliers
				Most Popular	Centroids not meaningful in most problems
	PAM	Mediod -centriod		robust to outliers	Cluster should be pre-determined
	CLARA			Applicable for large data set	Sensitive to outliers
	CLARANS			Handles outliers effectively	High Cost
Density Based	DBSCAN	Fixed size	Numerical	Resistant to noise	Cannot handle varying densities
				Can handle clusters of various shapes and sizes	
	OPTICS	Variable size		Good for data set with large amount of noise	Needs large no.of parameters
				Faster in computation	
	DENCLUE			Solid mathematical foundation	Needs large no.of parameters
	RDBC			More effective in discovering varied shape clusters	Cost Varying
				Handles noise effectively	
Hierarchical agglomerative	CURE	Partition Samples	Numerical	Robust to outliers	Ignores information about inter-connectivity of objects
				Appropriate for handling large dataset	
	BIRCH	Multidimensional	Numerical	suitable for large databases	Handles only numeric data
				scales linearly	sensitive to data records
	ROCK	Notion of Links	Categorical	Robust	space complexity depends on initialization of local heaps
				Appropriate for large dataset	
	S-Link	Closest pair of points	-	it does not need to specify no.of clusters	Termination condition needs to be satisfied
					Sensitive to outliers
	Ave-Link	Centriod of clusters	-	It considers all members in cluster rather than single point	It produces clusters with same variance
	Com-Link	Farthest pair of points	-	Not strongly affected by outliers	It has problem with convex shape clusters
Grid	STING	Multiple Grids	Numerical	Allows parallelization and multiresolution	Does not define appropriate level of granularity
	WaveClusters		Numerical	High-quality clusters	Cost Varying
				Successful outlier handling	
	CLIQUE	Density based grids		Dimensionality reduction	Prone to high dimensional clusters
				Scalability	
				Insensitive to noise	

Data mining tasks can be classified into two categories:

- Predictive
- Descriptive

Predictive data mining is used to predict the direct values based on patterns determined from known results while Descriptive data mining establish and elaborate the general properties of the data in the database. Prediction uses few fields or variables of the database to predict the future values of other pertinent variables. Predictive data mining approach is more commonly used across the board. Extrapolations done by using Predictive data mining techniques can

help in agriculture like predicting future crops, effects of specific fertilizer or pesticide, Weather forecasting, revenue prediction etc. etc.[3]

In this paper we humbly present a review of some Clustering technique on the basis of the algorithms. The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Clustering is a significant task in data analysis and data mining applications.[4] It is the task of arrangement a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Data mining can be done by passing through various phases. Mining can be done by using supervised and unsupervised learning. The clustering is unsupervised learning. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. Density based algorithms find the cluster according to the regions which grow with high density. It is the one-scan algorithms. Grid Density based algorithm uses the multiresolution grid data structure and use dense grids to form clusters. Its main distinctiveness is the fastest processing time. In this survey paper, an analysis of clustering and its different techniques in data mining is done.

CLUSTERING METHODOLOGIES

HIERARCHICAL CLUSTERING / CONNECTIVITY BASED CLUSTERING:

Popularly known as HCA or Hierarchical Clustering Analysis, is a method which builds a hierarchy of clusters and is either Agglomerative or Divisive. Where as Divisive Hierarchical Clustering uses the Top-Down approach, the Agglomerative Hierarchical Clustering uses Bottom-Up approach, 0020 where, each data point is put into a cluster and pairs of clusters keep getting merged as we move up the Hierarchy. In HAC (Hierarchical Agglomerative Clustering), usage of Distance matrix is very important, using which the Distance between the elements is determined and the elements closest to each other are generally merged into the cluster. The merger continues from bottom till the top based on updated distance matrix with each merger. DIANA is the basic principle behind Divisive Clustering[10]. ‘DI’visive

‘ANA’lysis clustering Algorithm starts with whole dataset as a single cluster and splits the cluster based on maximum average dissimilarity, while moving bottom from the top.

CENTROID BASED CLUSTERING / K-MEANS CLUSTERING:

Clusters are represented by central vector, which need not be a dataset in itself. When the numbers of clusters is fixed to k, k-means clustering gives an optimized solution. One of the most known ways of K-Means Clustering is Lloyd’s Algorithm[11], where a local optimum is found and is run multiple times with different random clusters. Variation of k-means include these optimizations as choosing the best of the multiple runs. It also restricts the central vector to members of data sets (k-medoids), selecting medians (k-medians clustering) or a fuzzy cluster assignment (fuzzy c-means) or in choosing initial centers less randomly (k-means++). All the algorithms demand k i.e. number of clusters in advance and of approximate similar size since they assign an object to nearest centroid. This may lead to cutting of borders which is not a big issue since this algorithm optimizes cluster center and not borders. It may be seen as a model-based clustering, and Lloyd’s algorithm as variation of Expectation-Maximization algorithm. It is closest to nearest neighbour classification conceptually, which is also used a lot in machine learning. This method partitions the dataset into Voronoi diagrams.

DISTRIBUTION BASED CLUSTERING / GAUSSIAN MIXTURE MODEL CLUSTERING:

This method uses Expectation-Maximization Algorithm for Gaussian Mixture model. Dataset is modelled with fixed number of Gaussian distributions, to avoid overfitting. Gaussian distributions used are the ones which are initialized randomly and parameters of whose are optimized post multiple iterations to better fit the dataset. All this converges to a local optimum, which may be different with every run. To obtain a hard clustering, objects are assigned to their respective Gaussian distribution. Though the same isn’t needed for soft clustering. This clustering produces complex results, which depict dependence and correlation between the attributes. Gaussian distribution is one of the most common continuous probability distributions. Expectation-Maximization Algorithm[12] is a method involving multiple iterations to find maximum likelihood estimates or parameters in statistical models.

DENSITY BASED CLUSTERING:

Clusters, in this method, are defined as areas of higher density, as compared to the other members of the dataset[13]. DBSCAN[14] happens to be most known density based clustering

method. It follows the logic of "density reachability", which is based on distance threshold being used as the connection methodology. DBSCAN is a non-complex method which needs linear number of range queries on the database. Study of DBSCAN has few important inclusions too like OPTICS, R-Tree Index and Single Linkage Clustering. A detailed research and study on these methods will be carried out subsequently. Mean-Shift is another method of Density Based Clustering in which using the Kernel Density Estimation, the object is moved to the nearest densest area, eventually, reaching to a local maxima. These local maxima can be the representative of the database.

CLUSTERING APPROACH

Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group. Clustering can be considered the most important unsupervised learning technique. Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind. It deals with finding a structure in a collection of un labeled data. Clustering is the process of organizing objects into groups whose members are similar in some way.[8]

Cluster analysis has been widely used in many applications such as business intelligence image pattern recognition web search biology and security. In business intelligence clustering can be used to organize a large number of customers into groups where customers within a group share similar characteristics. This facilitates the development of business strategies for enhanced customer relationship management. In image recognition clustering can be used to discover cluster or subclasses in handwritten character recognition system. Suppose we have a data set of handwritten digits where each digit is labeled as either 1,2,3, and so on. Note that there can be a large variance in the way in which people write the same digit. Take the number 2, for example, some people may write it with a small circle at the left bottom part, while some other may not. We can use clustering to determine sub classes for each of which represents a variation on the way in which 2 can be written. Using multiple models based on the subclasses can improve overall recognition accuracy.[9]

INTRODUCTION

Database of agriculture is increasing day by day. Though, the application of various data mining techniques in the field of agriculture is still relatively new. In this paper we present a brief review of a various of Data Mining techniques that have been applied in the agricultural field. Though limited, data mining techniques like Fuzzy C-Means, Support Vector Machine (SVM), K nearest Neighbour, K-Means, Naïve Bayes Classifier, Neural Networks (NN) and bi-Clustering. How appropriate the data mining technique is, is determined, to some extent, by the problem which is being worked upon or the types of different agricultural data. Few surveys summarize the application of data mining techniques and predictive modelling application in the agriculture field.[5]

Agriculture is the backbone of our country's economy. As Mahatma Gandhi said, "India lives in villages and agriculture is the soul of Indian economy"[6]

A country's population survives on the food it grows and that makes a country independent or dependent. With India being an Agricultural economy, India is one of the largest producers of various agricultural products and hence can be called as the "*thali of the world*", Cuba can be called as Sugar Bowl of the world only based on it's sugar production.

Procurement due to dip in MSP in State

The Gujarat & Rajasthan government conducted groundnut procurement in coordination with the National Agricultural Cooperative Marketing Federation of India Limited (NAFED). NAFED didn't procure groundnut and other agriculture crops because of the inadequate infrastructure with the State agencies. NAFED's concerns came previous years citing of irregularities and malpractices at groundnut warehouses.

MSP driven down

Huge quantities of groundnuts arrive at the markets across the State. This pulls down the prices below the minimum support prices (MSP), necessitating procurement measures.

According to Gujarat State itself, estimates, kharif groundnut crop for the year 2018-19 is likely to be close to 27 lakh tonnes, lesser than the 32 lakh tonnes recorded last year. The market prices have already dipped below MSP of ₹4,890 per quintal thereby prompting the State to announce procurement under the price support scheme (PSS).

The Gujarat government announced a bonus of ₹110 per quintal over-and-above MSP, making the effective procurement price ₹5,000 per quintal. The market prices thus came in the range of ₹4,360-4,500 per quintal.

As per the latest available data, so far 8700 tonnes of groundnut worth ₹43 crore has been procured from over 3700 farmers in the State [7].

Ensuring food availability and stabilizing food prices through large scale interventions in the food grain sector has a long tradition in India. Currently, India's food program encompasses public procurement, storage and distribution of wheat, coarse grains and rice.

India promotes co-operative marketing of agricultural produce to benefit the farmers, by undertaking procurement of agricultural commodities for internal trade as well as for exports. National Agricultural Cooperative Marketing Federation of India (NAFED) is one of the central Nodal Agencies for procurement of 16 notified agricultural commodities and is having godown capacity of 50,200 MT, Cold storage capacity of 9565 MTs, Onion storage capacity of 4400 MTs and Empty Container Yard of 20,008 Sq. Mt. Quality checking at the time of procurement is done for physical parameters such as moisture, shriveled grains, broken grains, admixture etc., in accordance with the Fair Average Quality norms.

Data mining process results in discovering new patterns in large data sets. It is the process of analyzing data from different perspectives and summarizing it into useful information without any restriction to the type of data that can be analyzed. The goal of the data mining process is to extract knowledge from an existing data set and transform it for advanced use.

The data availability can be as either a text file or a web server log or a data warehouse or a relational database. Effective Analysis of data in needs understanding of appropriate techniques of data mining. The intention of this paper is to use different data mining techniques in perspective of agriculture domain with respect to Quality assessment of Groundnuts so as to develop a model for application of the model to Quality decision of procurement of any of the 16 notified agricultural commodities like Oilseeds, Pulses and Cotton in accordance with Fair Average Quality Norms.

The paper will use various Data Mining Techniques including Classification, Clustering and Regression to device and develop a model which can be used for accessing Quality and making apt decisions based on the Quality norms specified for 16 agricultural commodities. Research paper is developing suitable data models to achieve high precision and high generality with respect to data points like Maximum limits of tolerance of Foreign Matter (like dust, dirt, stones, lumps of earth, chaff, stem/straw or any other impurity), Damaged Pods, Shrivelled & Immature pods, Pods of other variety, Shelling (kernels/pods) and Moisture Content.

The paper is indexed as: a Chapter on capturing the details of the data points. A Chapter which discusses various Data Mining Techniques to be used. A Chapter covers the application of the Techniques to the data.

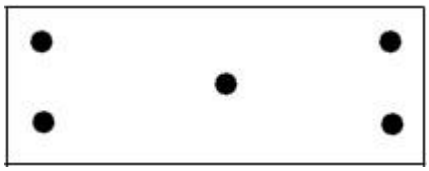
Data Collection

Consignment received at warehouses are considered in lots of maximum 500 Tons or such part thereof as constitutes a single consignment.

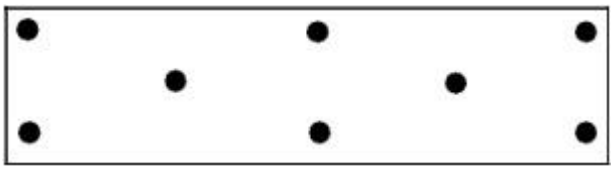
Increments of samples are taken from different parts of a bag (e.g. bottom, middle and top). In case of number of bags in a consignment being upto 10, sample needs to be collected from each bag, while for consignments having 10-100 bags, any 10 bags can be selected at random. In case of consignments with more than 100 bags, approx. square root of total number bags are to be sampled.

Consignments received via lorries, following sampling methodology is followed:

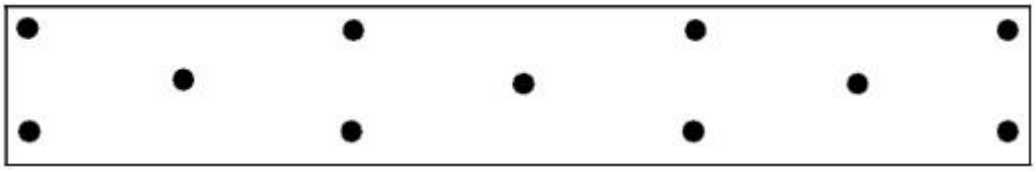
Upto 15 Ton: 5 sampling points



From 15 Ton to 30 Ton: 8 sampling points



From 30 Ton to 500 Ton: 11 sampling points



Survey Reporting

Trained surveyors carry various tools with the to do the sampling, testing and reporting. Based on their tests, they submit a daily EOD report covering following parameters:

1. Surveyor Name
2. Date of Survey
3. Warehouse Type – Public or Private
4. Warehouse Name
5. Vehicle ID
6. Procurement agency
7. Procurement scheme
8. Weight of Bags in Metric Ton
9. Percentage of Foreign matter
10. Percentage of Damaged
11. Percentage of Immature
12. Percentage of Pods of other variety
13. Percentage of Shrivelled
14. Percentage of Kernels
15. Variety of Groundnut
16. Percentage of Moisture
17. Number of Bags Accepted
18. Number of Bags Rejected

SCOPE OF PAPER

India promotes co-operative marketing of agricultural produce to benefit the farmers, by undertaking procurement of agricultural commodities for internal trade as well as for exports. National Agricultural Cooperative Marketing Federation of India (NAFED) is one of the central Nodal Agencies for procurement of 16 notified agricultural commodities and is having godown capacity of 50,200 MT, Cold storage capacity of 9565 MTs, Onion storage capacity of 4400 MTs and Empty Container Yard of 20,008 Sq. Mt. Quality checking at the time of procurement is done for physical parameters such as moisture, shriveled grains, broken grains, admixture etc., in accordance with the Fair Average Quality norms.

Scope of data includes Third Party Quality Assessment (drawing and analysis) of samples of Groundnuts from the different storage facilities.

Data mining process results in discovering new patterns in large data sets. It is the process of analyzing data from different perspectives and summarizing it into useful information without any restriction to the type of data that can be analyzed. The goal of the data mining process is to extract knowledge from an existing data set and transform it for advanced use.

The data availability can be as either a text file or a web server log or a data warehouse or a relational database. Effective Analysis of data in needs understanding of appropriate techniques of data mining. The intention of this paper is to use different data mining techniques in perspective of agriculture domain w.r.t. Quality assessment of Groundnuts so as to develop a model for application of the model to Quality decision of procurement of any of the 16 notified agricultural commodities like Oilseeds, Pulses and Cotton in accordance with Fair Average Quality Norms.

The paper will use various Data Mining Clustering Techniques and others to device and develop a model which can be used for accessing Quality and making apt decisions based on the Quality norms specified for 16 agricultural commodities. Research paper is developing suitable data models to achieve high precision and high generality w.r.t. to data points like Maximum limits of tolerance of Foreign Matter (like dust, dirt, stones, lumps of earth, chaff, stem/straw or any other impurity), Damaged Pods, Shrivelled & Immature pods, Pods of other variety, Shelling (kernels/pods) and Moisture Content.

The dataset involved has been sourced from National Agricultural Cooperative Marketing Federation of India National Agricultural Cooperative Marketing Federation of India, NAFED's survey report namely, "*Third Party Quality Assessment of Pulses and Oil Seeds during Procurement for National Agricultural Cooperative Marketing Federation of India, NAFED*" in 2018. The said survey was conducted by Quality Council of India (QCI). The dataset, based on this survey report, is of 268,000+ cells and has been kept focused on Procurement of Groundnuts across 2 states of Rajasthan and Gujarat.

SUMMARY

In a country like India, the economy is a lot influenced by Agricultural sector. The success or failure of Agricultural sector is dependent on the rainfall, climate of all seasons. Though, there is a lot of use of technology in the field of Agriculture, usage of Data Mining Techniques in the Agricultural sector of India is still minimal. When it specially comes of usage of humungous data to draw predictive models, very few researches have been done till now. Some of good application of Data Mining Techniques have been seen in studies done by Saeed Soltani and Reza Modarres. Still, the field of Data Mining can be termed as relatively unexplored, especially in Indian context.

Today, where Data Mining has gained momentum across various Industries, Agriculture field must also work on it diligently. We believe that Data Mining techniques like Agglomerative Clustering, DBSCAN, EM Algorithms, K-Means will bring in an advancement that Agricultural sector has long been waiting for.

Devising of models to help procuring the right quality of seeds, would go a long way in shortening the process time between Growing in fields to the shelves of the retail stores. Such models will also help a lot, in a way, to cut down the costs involved with warehousing especially on storage costs, losses because of rotting etc.

REFERENCES

[1] Amandeep Kaur & Navneet Kaur: Clustering Techniques

[2] Shraddha K.Popat et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 805-812

[3] **Developing innovative applications in agriculture using data mining- Sally Jo Cunningham and Geoffrey Holmes**

[4] **TanujWala: International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), May 2015**

[5] **Dharmendra patel: International Journal of Computer Applications." A Brief survey of Data Mining Techniques Applied to Agricultural Data"**

[6] **http://www.klientsolutech.com/agriculture-in-india/**

[7] **https://www.thehindubusinessline.com/markets/commodities/gujarat-nafed-to-procure-groundnut-jointly/article25540886.ece**

[8] **Margaret H. Dunham, "Data Mining Introductory and Advanced Topics", Dorling Kindersley Pvt. Ltd. India, Sixth Edition,2013.**

[9] **Tayel, Salma, et al. "Rule-based Complaint Detection using RapidMiner", Conference: RCOMM 2013, At Porto, Portugal, Volume: 141- 149,2014**

[10] **Everitt, Brian (2011). *Cluster analysis*. Chichester, West Sussex, U.K**

[11] **Lloyd, S. (1982). "Least squares quantization in PCM".**

[12] **Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm"**

[13] **Kriegel, Hans-Peter; Kröger, Peer; Sander, Jörg; Zimek, Arthur (2011). "Density-based Clustering".**

[14] **Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise".**

ABOUT AUTHOR

Vyoma Bisaria is an engineering professional with an extensive analytics and IT experience of over 6 years. With the rich experience of premier Organizations like Quality Council of India, NEISBUD and many others, Vyoma has gathered a very practical approach and experience towards the application of Data Mining Techniques in the real world, thus helping her research from the perspective of a very thorough researcher. Vyoma has completed her Bachelor of Technology from Vishveshwarya Institute of Technology, Gautam Buddha Technical University, India and has done her Master of Engineering from Indraprastha Engineering College, Mahamaya Technical University, India. She excelled in both BTech and MTech in 1st Division.