INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 1 - V8I1-1439)
Available online at: https://www.ijariit.com

Clustering Algorithms and Classification Method for the Analysis of the Crop Yielding Dataset

Vani Yelamali
vani.yelamali@kletech.ac.in
KLE Technological University, Hubli, Karnataka

ABSTRACT

Prediction of the crop is a dominant element of agriculture and even for farmers[1]***. Currently, India is in the second position in the world for agricultural produce. The main economic sector of India is agriculture, which plays an important role in the growth of the economy[1]. Prediction of a crop is a challenging issue, so comprehensivevarieties of crop predictionmethods are used. In this paper the essential data is collected from districts of Karnataka and Tamilnaduwith many parameters like year, district, area,temperature, rainfall, crop, and yield in tons for the year2005 to 2016. For crop prediction, methods like K-Means and the J48 algorithm are applied in the existing system for clustering and classification. In the proposed system for clustering, the PAM (PartitionaroundMedoid) is applied and for classification, J48 is applied. A dataset is collected and clustered as stated by the attribute of the PAM algorithm by using the Euclidean formula, calculatingthe distance between the points. In this work tools like eclipse mars.2 and the programming language, Java is applied. The result acquired through an existing algorithm such as K-Means &J48 in comparison with a proposed algorithm like PAM and J48 gives better results in terms of accuracy and time complexity.***

Keywords: *Crop Prediction, Clustering, Classification, K-Means, PAM, j48.*

1. INTRODUCTION

India is the second-largest producer of crops[2]. Agriculture is one of the major sources of income and least paid occupations in India, so for farmer's cropsYieldprediction is essential[3]. Every farmer tries to know the yield he gets for the crop[3]. Inpast, the prediction of crop yield was calculated by analyzing the previous experience of the farmer. The agricultural yields mainly rely on weather conditions and Planning of harvest operations. The yield of the Cropkeeps on changing every year dynamically based on the Parameters like temperature, humidity, rainfall, wind speed, etc.[4][1]. Climate change plays a major challenge in agriculture which can be overcome by the prediction ofyield. Now a day's Data Mining plays an important rolein agriculture. Data Mining is a proportionately young and multidisciplinary field of artificial intelligence. Data Mining is the process that ascertains patterns in a large dataset. It makes use of methods at the intersection of artificial intelligence, machine learning, statistics, and database system. Further, the data mining process is used to retrieve information from the data set and transform the information into a logical structure.

This paper focuses on data mining techniques such as clustering and classification performed on the crop yielding dataset. K-Means and PAM is performed on the crop yielding dataset.

The number of clusters is passed to form the clusters.

- The clusters are formed based on the instanceswhich belong to the same class
- The J48 classifier is applied to the dataset toanalyze the attribute value which affects the most
- The J48 classifier increases the accuracy rate ofthe procedure of data mining**.**

2. LITERATURE REVIEW

K-Means clustering

By KardiTeknomo, Ph.D., K-Means is the strategy that includes all values of the group. Firstly, recognize the K bunch focus; contingent upon the k Cluster focuses make a group. Presently the mean estimation of the group is called K-Means.In k-means calculation, since it is viewed as that there is a "k" number of groups; consider that there are "k" number of bunch means (group focuses), where the groupmeans is the normal of the considerable number of
information focuses falling under every group. The final consequence of the k-implies bunchingcalculation is that every information point in the information set is assembled into "k" groups around the "k" group means.

Partition Algorithms– A Study and Emergence of Mining Projected Clusters in High-Dimensional Dataset[5]

According to K. Naveen Kumar, the PAM is also called a K-Medoidalgorithm the clusters arerepresented by medoids. This algorithm tries to find the objects that are centrally located in clusters calledmedoids. The K-medoid is the same as the K-means clustering algorithm and both algorithms use partitional

clustering. In K-medoid clusters are formed based on the data set of n objects with k no ofclusters. When compared to the K-means algorithm K-medoid algorithm chooses the data points as the center of the cluster.

Improved J48 Classification Algorithm for thePrediction of Diabetes

According to Gaganjot Kaur, Amit ChhabraInstances can be sorted down the tree from the root to the leaf node by using decision tree classification. In the tree, the test is specified on each node of the attribute ofthe instances and the branches of the tree correspond to the values of the attributes. The classification of instance is done starting from the root node, then testing the other specified attributes of the node, then going down the attribute specified in the example.The process is repeated. The pruning of the tree is performed on the data to remove the anomalies whichcause noise and outliers.

3. EXISTING SYSTEM

In the existing work, the J48 algorithm with the K- means clustering algorithm is performed on the crop yielding dataset. The result obtained from thesealgorithms is accurate but J48 exhibits limitations with applications, learning time, and the computation complexity of the network. K-means algorithm is used in this work to overcome these two problems. In this, the dataset used is of Karnataka from a period of the year 2005-2016. Parameters considered for crop prediction are Temperature, rainfall, crop yield, atmospheric pressure.

4. DATASETS USED

The dataset used in this experiment is in CSV file format with 667 records and 10 parameters. Withdifferent combinations of techniques, Crop predictionis done in this proposed work. For this, the data is collected from Karnataka and Tamilnadu from the period of the year 2005-2016. Particular attributes are considered for this work as crop, season, area (in hectares), production (in tones), average temperature,rainfall, and soil.

5. TECHNIQUES USED

5.1 K-Means Algorithm:

Mainly, the K-Means clustering algorithm focus to divide N observations into K clusters based on the attributes. The main important thing in K-Means is used to define the clusters with K-centroids. Each statement isin the cluster with the closest mean.

The Step by Step Process:

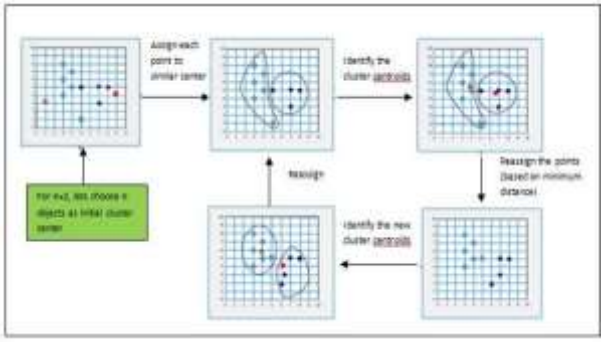

Fig-1:Step by step process of K-Means Algorithm

Step-1: Data is collected and stored in the database.
Step-2: StructuralCrop Yielding database splits clusters.
Step-3: Each object is assigned to the closest centroid.
Step-4: After the objects are assigned, recalculate thecentroid.
Step-5: Repeat step 3 and 4 until centroids move nolonger.

Where k is the number of clusters to be formed,IntnoOfClusters = k;

```
for(int j=0;j<noOfClusters;j++){
i=0;
String clusterOutput="";int count=0; for(intclusterNum:
assignments){
if(j==clusterNum){
count++;

clusterOutput+=inputLines.get(i)+"\r\n";
}
i++;
}
```

5.2 : K-MedoidAlgorithm (PAM):

```
Instances trainingSet, InstancestestingSet) throws Exception {
Evaluation evaluation =
new
Evaluation(trainingSet);
```

The functioning of the K-Medoid clustering algorithm is near to the k-Means clustering algorithm. The dissimilarities between each data itemand its corresponding medoidare reduced in the method.

Mean is more influenced by outliers or other extreme testingSet);

```
}
model.buildClassifier(trainingSet);
evaluation.evaluateModel(model,
```

Returnevaluation ;

values than a Medoid[6], as PAM is more robust thanK-Means in the presence of noise and outliers.PAM works efficiently for small datasets but does not scalewell for large datasets.

5.2.1 : PAM Algorithm:

1. To represent the cluster, use the data items from the dataset.
2. Arbitrarily choose K representative objects as medoids.
3. Calculate the total swapping cost $S(P_iM_k)$ where P_i is a non-medoid item and M_k is chosenmedoid.If $S<0$, M_k is replaced by P_i. Assign each data item to the cluster with the most similar representative item that is medoid.

5.3 :J48 Classifier:

1. In the J48 classifier, the instances which are the same belong to the same class which represents theleaf of the tree.

2. The information gain is calculated for each attributefrom the given test attributes[3].

The best attributes is selected based on the criteria selected and the branches are formed based on the selected branches

Input: Instances Output: Prediction
Input format: Weka's ARFF formatReading from ARFF file;
BufferedReaderdatafile = readDataFile("Crop Yielding quality.arff");
Instances data1 = new Instances(datafile);

The prediction is done using the classifier; Evaluation classify(Classifier model,

5.3.1 :J48 algorithm flow

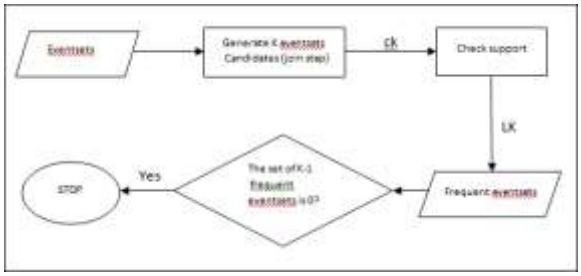

Fig -2:J48 algorithm flow

The above Fig-2 depicts the j48 algorithm flow.

J48 algorithm is performed in two steps
i) Join step
ii) Prune
The algorithm is classified into two steps:

Step 1: Find out frequent event sets with K items in a database applying for the minimum support.
Step 2: With the help of frequent K-events find frequentevent sets with K+1 items applying the self-join rule. Repeat this process from K=1 to the point where it is unable to apply the self-join rule. For analysis, Clustering is done, by applying the farthest first algorithm, and classification is done using the J48 algorithm. With these two combinations of algorithms, better accuracy can be obtained compared to the existing combination of algorithms.

6. SYSTEM ARCHITECTURE

Fig-3 depicts the analysis of the Crop Yielding dataset. The purpose of clustering is to group the data and place them in one group. Groups can be formed based on similar or dissimilar objects.

The J48 classifier is used to predict the Crop Yielding dataset. The J48 can be used to generate the decision trees. It uses the attributes of the dataset to decide by
splitting the data into small subsets. If the subset of the instances belongs to the same class then the splitting willstop. That class will act as the leaf node in the decision tree.

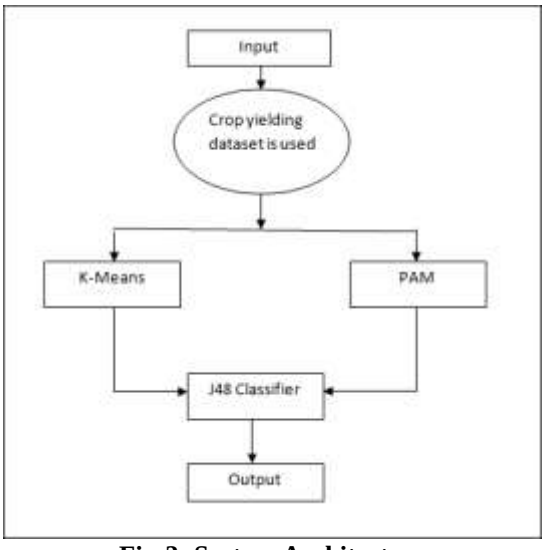

Fig-3: System Architecture

Fig-4 depicts the Dataflow diagram which shows the flowof the application

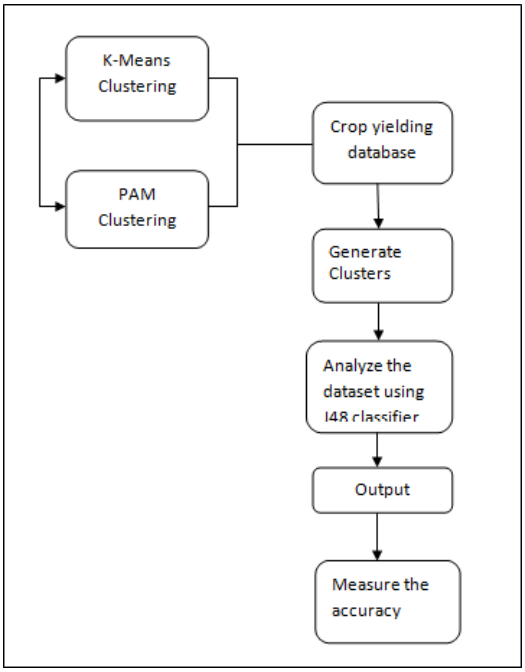

Fig-4: Data flow Diagram

7. RESULT

The clusters formed using K-Means and PAM isobtained and then the condition is predicted. The performance of the clustering algorithm is compared based on the time taken to form the cluster. The accuracyplot is obtained after comparing both the K-Means and PAM algorithms. It is found that the accuracy of PAM is more when compared to the K-Means. The J48 classifier is used to increase the accuracy. To generate the J48 classifier data mining tool called WEKA is used.

Table-1: Time taken by each algorithm to form clusters

Algorithms Clusters	K-Means (in seconds)	PAM (in seconds)
2	31.41	13.95
3	25.11	12.04
4	17.47	10.53
5	12.89	8.65

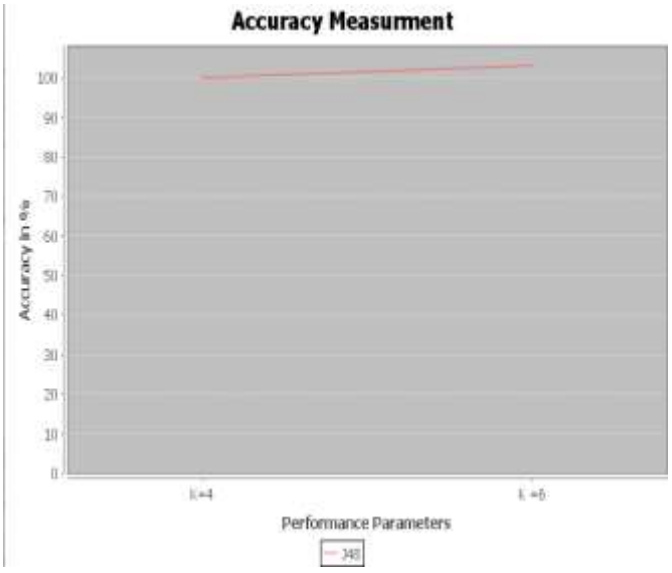

Fig-5: Classification Accuracy

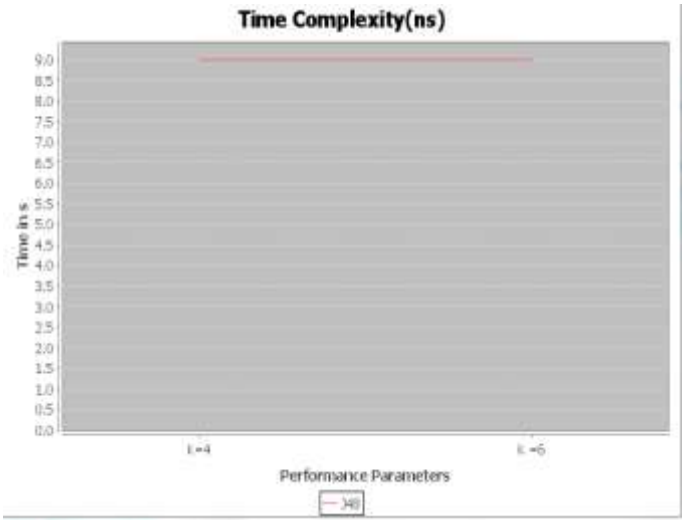

Fig-6. Time Complexity

Around 98% of classification accuracy is achieved forthe crop yielding dataset in the proposed technique which is shown in the above Fig-5. Similarly proposed algorithm performance time complexity is 8.5s which is shown in Fig-6.

8. CONCLUSION AND FUTURE ENHANCEMENT

A new algorithm PAM which is similar to K-means is used. The PAM algorithm performs well when compared with the K-means algorithm. The Crop Yielding dataset isused and with the help of the J48 classifier, the condition is predicted. The performance of the clustering algorithm is compared based on the time taken to form the cluster. Firstly clustering is done based on the features of the K- medoid algorithm then by using the J48 algorithm, classification has been made. These clustering algorithms are used only onthe static databases where the climate keeps on changing frequently.

Future: In the future new clustering algorithms can beused and the Crop Yielding can be predicted for the dynamic dataset. Compare with the different algorithms and then calculate the accuracy of the algorithm.

9. REFERENCES

[1] R. Medar, "Crop Yield Prediction using MachineLearning Techniques," pp. 1–5, 2019.

[2] S. S. Kale and P. S. Patil, "A Machine Learning Approach to Predict Crop Yield and Success Rate," *2019 IEEE Pune Sect. Int. Conf. PuneCon 2019*, pp. 1–5, 2019, doi: 10.1109/PuneCon46936.2019.9105741.

[3] V. Bhuyar, "Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District," no. April2014, 2016.

[4] R. Kumar, M. P. Singh, P. Kumar, and J. P. Singh, "Crop Selection Method to maximize cropyield rate using machine learning technique," *2015 Int. Conf. Smart Technol. Manag. Comput. Commun. Control. Energy Mater. ICSTM 2015 - Proc.*, no. May, pp. 138–145, 2015, doi: 10.1109/ICSTM.2015.7225403.

[5] S. Sahay, S. Khetarpal, and T. Pradhan, "Hybrid data mining algorithm in cloud computing using MapReduce framework," *Proc. 2016 Int. Conf. Adv. Commun. Control Comput. Technol. ICACCCT 2016*, no. 978, pp. 507–511, 2017, doi:10.1109/ICACCCT.2016.7831691.

[6] B. Pardeshi and D. Toshniwal, "Hierarchical Clustering of Projected Data," vol. m, pp. 551–559, 2011.

[7] www.semanticsscholar.org

