

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES **RECOMMENDATION IN AGRONOMICAL DATA USING DATA MINING** **TECHNIQUES**

Urvashi*¹ & Dr. Kanwal Garg²

*¹Research Scholar, Department of Computer Science and Applications Kurukshetra, Haryana, India

²Assistant Professor, Department of Computer Science and Applications Kurukshetra, Haryana, India

ABSTRACT

Agriculture is the backbone of Indian economy. In India various crops are grown such as cotton, wheat, rice, mustard, tea, coffee. The land of agronomical people is distributed into four categories such as small land holding farmer, moderate land holding farmer, high land holding farmer and landless farmers. Around 70% farmers are belongs to the small land holding farmer category. This reflects that majority of farmers having less area of growing. Farmer directly depends on the productivity of the crop. So, this paper is focused to perform Feature Selection and yield prediction of the crop by implementing Clustering algorithm.

Keywords: Agriculture, Yield prediction, Data Mining, K-Means, Improved K-Means, K-medoid, Clustering, Matlab, Feature Selection.

I. INTRODUCTION

Motivation

Agriculture has been a wellspring of wage for a critical level of India's peoples for a considerable length of time. An expansive mass of irrigable land, geologically shifting climatic conditions and a sizeable population having a place with the work hands guaranteed that India turned into a farming nation. All inclusive, farming still remains an overwhelming piece of the economy and the statistic. Today, around 60-70% of the nation's population relies upon agronomy.

Clustering is the unsupervised learning method.

In unsupervised learning there is not known input data. And learn by experimentation. It aimed to collect all the data points that are similar to each other and put this collected data points into cluster. The remaining data points that are outlier from the cluster are put into different cluster. There are various cluster algorithms such as: Partitioning methods, Fuzzy Clustering, Hierarchical Clustering, and Density-based, model-based are the types of Clustering.

In agriculture field, data mining is a relatively pleasing research field. It is applied to formulate the solution of agriculture problems. In this paper, a model has been designed for effective decision making done by the farmer and as well as agency side. This proposed paper, performs feature selection and recommendation using clustering algorithm that is applied to Climatic data set of Haryana of all 12 months and agriculture dataset of different districts of Haryana which are having similar crop outcome. In this, algorithms are implemented for the selection of optimal algorithm and Yield prediction.

In this Section1 contain about agronomy using data mining, Section 2 focused to described all related work to the paper work, Section 3 present the data set and methodology used, Section 4 display the result and experimental part and at last conclusion is provided by Section 5.

[Urvashi, 5(8): August 2018]

ISSN 2348 - 8034

DOI- 10.5281/zenodo.1406085

Impact Factor- 5.070

II. RELATED WORK

Various approaches were implemented by the researcher over the world to predict the outcomes of the crop. Monali Paul et al. [18] described about the analysis of soil behavior and crop yield were forecasted with the help of data mining. In this KNN and Naïve Bayes classification algorithm were implemented on the soil data set. They focused to calculate the accuracy which helped them to calculate the yield. In the experiment the soil had been divided into three categories that were low, high and moderate and best results were given by Naïve Bayes. In paper “Data Mining Techniques to Predict the Accuracy of Soil Fertility” S. Hari Ganesh [19] performed a comparative analysis between classification techniques. The soil data set were used to calculate the accuracy using naïve Bayes and Jrip. The results showed that the error rate in naïve bayes were 0.321 and in Jrip were 0.0423. This concluded that Jrip were worked better than naïve bayes in their paper.

Feature Extraction used to extract the useful features. Pooja G. Mate et al. [20] proposed a paper to extract the feature using remote sensed image for estimating the productivity of agriculture. In this feature selection and feature optimization were done. The feature was selected from agricultural dataset to calculate the crop production. The SVM was used as classification algorithm. Firstly input was applied that was remote sensed image, next step were numerous collection of feature, next feature selection and extraction, 4th step was classification of feature and selection of optimal features and at last forecasted the yield. Neetu Chahal [17] gave a study, on Agricultural Image processing using classification method. The first step was image acquisition which extracts the leaf image, next Processing to resize the image, next were image segmentation that was feature extraction and at last Classification was performed. The classification techniques were SVM, neural network and K-means clustering. At the end the diseases were classified using various classification methods.

III. PROPOSED WORK

The objective of a research paper is addressed step by step below such as:

A. Data Set

All the databases used in this research paper are collected from Indian Government Agricultural Records of various districts of Haryana. From the large dataset, only the limited factors are fetched out which are important for agriculture production. All the factors are considered for a period of years 2009-2016. Some selected factors are shown below such as:

Crop: The wheat and Mustard crop is selected as the crop of which prediction occurs. The crop details are collected from [Haryana Agriculture Department]

Year: From 2009-2016 data are analyzed and collected from a government of Haryana.

Area: The area used in this are calculated in terms of acres.

District: In this, 10 districts are taken such as Rohtak, Fatehabad, Sirsa, Sonipat, Kurukshetra, Karnal, Hissar, Jhajjar, Bhiwani, Mahendergarh.

Months: The data of January to December are collected from date and time.

Temperature: The min and max and mean temperature is used.

Humidity: As it affects the productivity of crops and the value represent %.

Wind Speed: The speed of wind taken in km/h format.

Visibility: It is a measure of the distance at which the object can be clearly seen, it present in terms of km.

Pressure: It is denoted by 1001 mbar.

Precipitation: It is any form of water falling from the sky, and calculated in terms of mm.

Production: The production is taken into Quantile per acre.

B. Methodology Used:

a. **Matlab:** Matlab is a multi-paradigm numerical computing environment. It was developed by Math Works. The functionalities of Matlab are Manipulation of a matrix, Function plotting, data plotting, algorithm implementation, user interface creation and interfacing with a program written in languages such as C, C++, Python.

[Urvashi, 5(8): August 2018]
ISSN 2348 - 8034
DOI- 10.5281/zenodo.1406085
Impact Factor- 5.070

b. Feature Selection: The Feature Selection reduced the number of inputs or selecting the required information. From the existing data only useful information are extracted from the available dataset.

Information Gain: The information is a formula use to select important attributes of the dataset.

Pseudocode:

- a. Calculate the Feature weight.
- b. Sort them
- c. Select a number of features needed for experimentation.
- d. Save the information gain.

c. Yield Prediction: The Yield can be predicted using Clustering algorithm.

Formula to calculate Yield:

1. Base Prediction Formula: It is equal to the average of previous year production. It uses only its own district dataset.

$\text{predProduction2} = \text{mean}(\text{production}(\text{predDis}, 1 : \text{totalYears}-1));$

$\text{error2} = \text{abs}(\text{actualProduction}-\text{predProduction2});$

$$\sum_{i=1}^n \sum_{j=1}^m \text{Production}_{i,j} / (n * m)$$

Here i is production value of previous year

j is production value of current year

n is a ending limit of i

m is ending limit of j

2. Proposed Prediction Formula: It is calculated selecting particular clustering algorithm. w1, w2 are assigned weights which are multiplied with previous year production.

$w1=0.2;$

$w2= (1-w1);$

$\text{predProduction1} = (w1 * \text{predProduction} + w2 * \text{predProduction2});$

$\text{error1} = \text{abs}(\text{actualProduction}-\text{predProduction1});$

d. Clustering Algorithms: In, proposed work, K-means, Improved K-Means, K- Mediods clustering are compared together on the basis of data collection of some districts of Haryana of wheat and mustard crop having similar temperature, Humidity, wind speed and other climatic factors of the complete year. Based on the examination the optimal parameter is achieved in order to get the high production of a crop.

K-Means: The similar items are grouped together, form a cluster. In this, the center of the cluster is the mean of measurement in the cluster. The similarity is calculated, using Euclidean distance.

$$d = \sqrt{(x2-x1)^2 + (y2-y1)^2}$$

Where d is calculated distance,

x1, x2= point x coordinate;

y1, y2= y coordinate;

Pseudocode:

- a. Randomly select k cluster center (centroids)
- b. Centroid collects all those points that are close to it.
- c. Use batch update-with the closest centroid each observation is assigned to the cluster.
- d. In each cluster, the average is computed to obtain knew centroid location.
- e. Until the maximum number of iteration is not reached repeat steps b-d.

[Urvashi, 5(8): August 2018]

ISSN 2348 - 8034

DOI- 10.5281/zenodo.1406085

Impact Factor- 5.070

Improved K-Means: In K-Means algorithm, randomly centroids are selected. It doesn't provide efficient outputs sometimes due to a random selection of centroid, whereas in the improved-means algorithm the sorting is performed when the split occurs, the average is calculated then at the end centroids are formed.

Pseudocode:

- Sort the points, split them into clusters, calculate the average and at the end k cluster center is achieved (centroids).
- All the points that are close are collected by centroids.
- Each measurement is assigned to the cluster with the closest centroid.
- In each cluster, the average is computed to obtain new centroid location.
- Repeat steps b-d until the maximum number of iteration is not achieved.

K-Medoids:

K-means is similar to K-Medoids. Both aimed is to divide a set of observation into k clusters so that Sum of the distance between an observation and center of the observed cluster is minimized. In the K-Medoids, a center of the cluster is a member of the cluster called medoids. The importance of k-medoids is that it can be implemented where mean doesn't exist within the dataset because k medoid returns the medoids which are the actual data points. The distance calculated by k-medoids is Manhattan distance.

$$d = \sum_{i=1}^n |X_i - Y_i|$$

Pseudocode:

- Each k cluster has medoid which is achieved by using a technique defined by 'START' name-value pair argument.
- Testing of each point is done by checking if the sum of distance within cluster gets smaller using that point as the medoids. If true new medoids are assign. Each point is assigned with the value closest to medoid.

IV. RESULTS

Feature Selection: In this a useful features are selected from the environmental dataset. Out of 8 features only 6 are selected implemented information gain formula on matlab is shown in fig 1. Below such as.

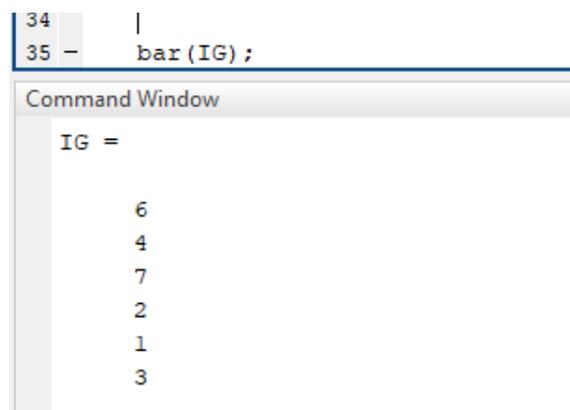


Fig1: Selected Features from Environment dataset.

In this selected features are: Pressure, Humidity, Precipitation, Month, District, and Temperature.

Important Terms:

- 1. Execution time:** The execution time taken by algorithm is represented by et.

[Urvashi, 5(8): August 2018]

ISSN 2348 - 8034

DOI- 10.5281/zenodo.1406085

Impact Factor- 5.070

2. Number of iteration: It is number of time centroid is changed to select correct center. Here it is denoted by iter.

3. Distance: It is sum of the entire intermediate node. Here it is shown by Dist.

The Feature selection is used to improve the output of clustering algorithm used without feature selection. For this 3 Clustering algorithms are used such as K-Medoids, K-Means, Improved K-Means, all of them are described below such as:

Basic K-Means: In Basic K-Means algorithm, 2 inputs are applied that are:

a. **Clustering data:** The environmental dataset.

b. **Number of cluster:** Here $k=2$ (number of clusters)

Output: It calculate 3 outputs such as et1 is execution time, iter1 is number of iteration, totalDist1 is total distance of Basic K-Means, idx1, c1.

With the use of feature selection, it computes the Execution time, Number of Iteration, Distance of the algorithm. The computed results are such as:

In this the average is calculated by running 10 times, then the computer average execution time is 0.013, average total distance is 2.7 and average number of iteration is 2.8

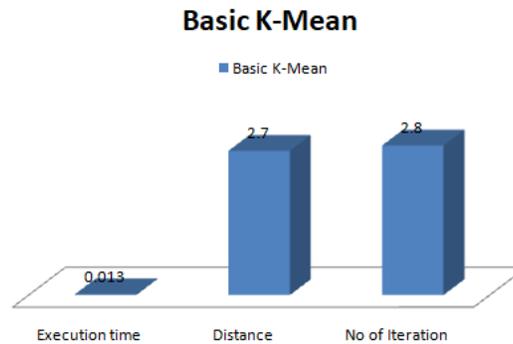


Fig 2: Execution time, number of iteration, distance

From the figure2, it is concluded that clusters are well separated from each other

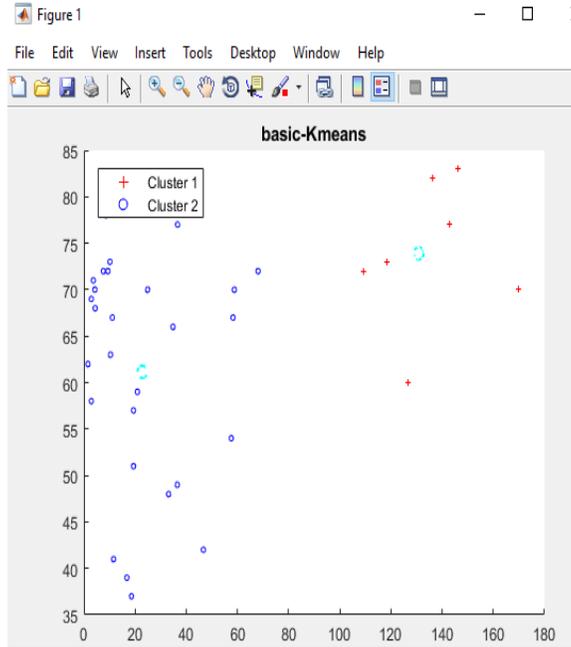


Fig 3: Shows Cluster1 and Cluster 2 using basic k-means

Improved K-Means: In Improved K-Means algorithm, 2 inputs are applied that are:

- a. **Clustering data:** The environmental dataset.
- b. **Number of cluster:** Here k=2(number of clusters)

Output: It calculate 3 outputs such as et2 is execution time, iter2 is number of iteration, totalDist2 is total distance of Improved K-Means.

Using feature selection, it computes the Execution time, Number of Iteration, Distance of the algorithm. The computed results are such as:

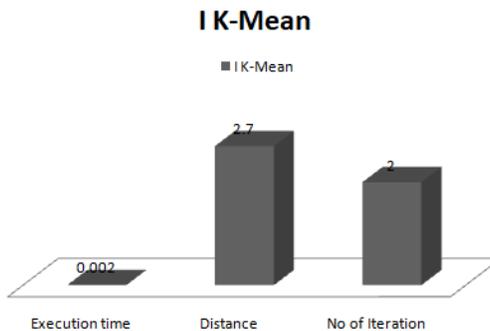


Fig 4: Execution time, number of iteration, distance computed by improved K-means.

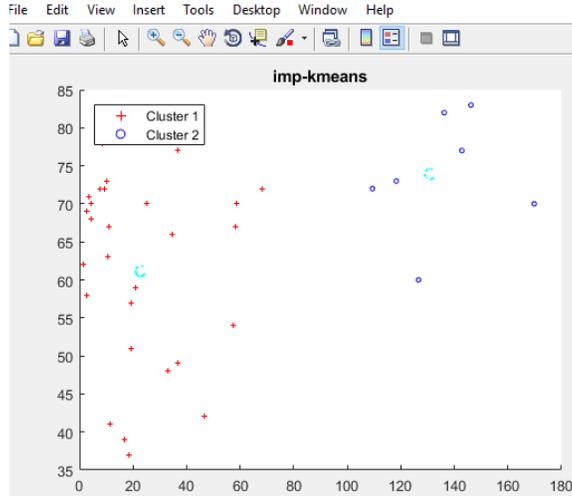


Fig 5: Shows Cluster1 and Cluster 2 using improved k-means.

K-Medoids: In K-Medoids algorithm, 3 inputs are applied that are:

- a. **Clustering data:** The environmental dataset filtered after feature selection.
- b. **Number of cluster:** Here k=2(number of clusters)
- c. **City Blocks:**

Output: It calculate 3 outputs such as et3 is execution time, iter3 is number of iteration, totalDist3 is total distance of Improved K-Medoids.

With the use of feature selection, it computes the Execution time, Number of Iteration, Distance of the K-Medoids algorithm. The computed results are such as:

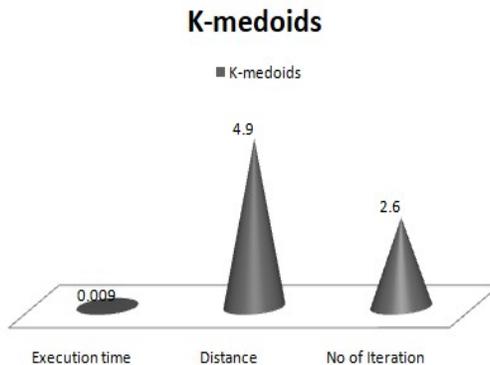


Fig 6: Execution time, number of iteration, distance of K-Medoid

In this the plus sign represent cluster 1, circle represent cluster2. From fig 6, it is clearly viewed that both the cluster are far a parted from each other

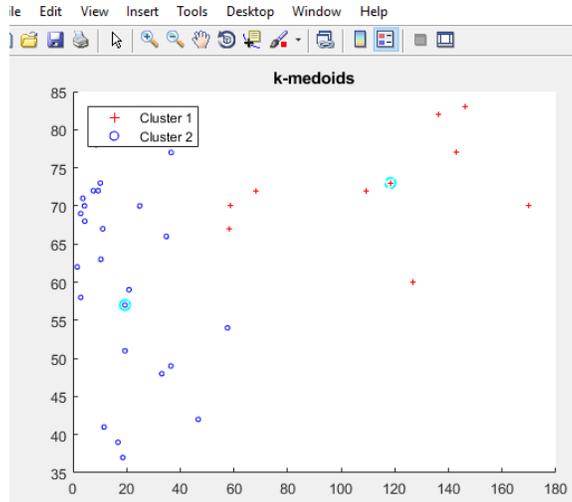


Fig 7: Shows Cluster1 and Cluster 2 using K-medoids

Experimental and Evaluation: After feature extraction the results are better as compare to the basic k-means, Improved K-means, K-medoids clustering algorithms implemented without feature selection. When Feature selection is not used that time the Improved K-Means works better then K-Medoids and basic k-means but having very high number of iteration. The high number of iteration doesn't make it optimal algorithm but this can be improved using feature extraction. This can be visually shown below such as:

Fig8: Execution time, number of iteration, distance of all three clustering algorithm.

Clustering algorithm	Et	Iter	Dist
Basic K-Means	0.013	2.8	2.7082
Improved K-Means	0.002	2	2.7082
K-Medoids	0.009	2.7	4.935

It is shown by Fig9 that the algorithms without feature extraction takes more time to execute, more number of iteration to find centroid, more distance is taken as compare to when feature extraction is applied.

Fig9: Comparison between with and Without Feature selection using Basic K-means

K-means	Without FS	With FS
Et	0.051	0.013
Iter	5.1	2.8
Dist	2.8045	2.7082

From Fig10, it is concluded that improved k-means algorithm execute faster, having less number of iteration to reach at correct centroid and distance is also less.

Fig10: Comparison between with and Without Feature selection using Improved k-means.

Improved K-mean	Without FS	With FS
Et	0.009	0.002
Iter	5.6	2
Dist	2.8045	2.7082

Fig11 shows that after feature extraction the execution time is reduced, distance is also reduced. But number of iterations is increased.

Fig11: Comparison between with and Without Feature selection using K-Medoids.

K-Medoids	Without FS	With FS
et	0.020	0.009
iter	2	2.7
Dist	5.31283	4.935

Yield Prediction: The yield is predicted using Improved K-Means, K-medoids.

Important Terms:

1. **AP:** It stands for Actual Production.
2. **PP1:** Production using Clustering algorithms.
3. **E1:** Difference between AP and PP1
4. **PP2:** Production using Average calculation.
5. **E2:** Amount of error due to wrong predicting the PP2.
6. **District:** Included Haryana district

Improved K-Means:

Input: The 2 inputs are applied to Improved K-Means algorithms as:

1. **Data:** It is combination of the environment and production data. In this normalization is used to remove unwanted noise.
2. **K:** The k is number of Cluster here k=2.

Output: The output includes:

1. **District:** The 10 districts of Haryana are used.
2. **Actual output:** The output actually gained from the crop. It is used to calculate the error between predicted and actual.
3. **Proposed Production:** The output obtained using its district plus all similar districts to it.
4. **Error in Proposed Production:** The difference between actual and predicted using its own district + similar kind of district (Clustering Concept).
5. **Base Production:** The output obtains in its own district.
6. **Error in Base Production:** The amounts of error come in between actual production and B.P

Actual	C.B.P	C.B.Err	B.P	B.Err
18.51	18.486	0.024	18.714	0.204
19.01	18.921	0.089	19.336	0.326
17.6	17.681	0.081	17.564	0.036
18.1	18.046	0.054	18.086	0.014
17.52	17.616	0.096	17.471	0.049
18.5	19.052	0.0582	19.566	1.066
19.2	18.837	0.363	19.214	0.014
18.2	18.267	0.067	18.4	0.2
18.21	18.267	0.056	18.4	0.19
19.8	19.087	0.713	19.571	0.229
		2.1		2.3

Fig12: Output using Improved K-means.

The above line in fig13, show the amount of difference in base yield prediction. The below line show the error using proposed yield prediction method. As, it is clearly seen that the base yield prediction method error rates is higher as compared to proposed yield prediction error rate.

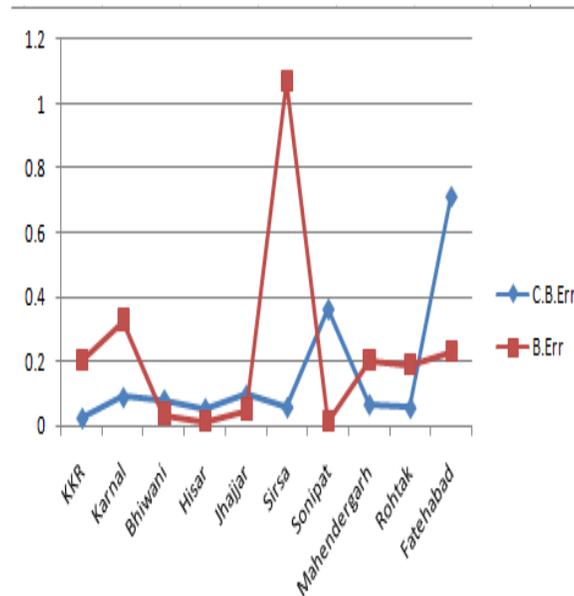


Fig13: comparison of erroneous using Improved K-means.

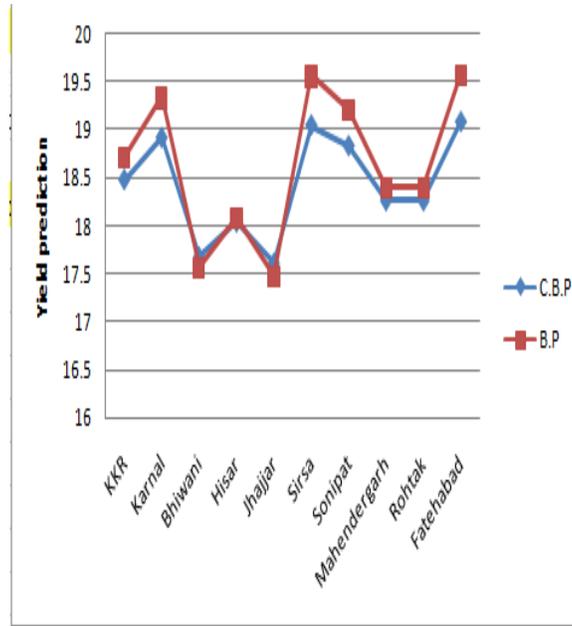


Fig14: comparison of base production, proposed production using Improved K-means.

K-Medoids:

Input: The inputs are applied to K-Medoids algorithms as:

1. **Data:** It is combined dataset of production and environment factors.
2. **K:** k is considered as 2.
3. **Cityblock:** The city name.

Clusters Formation:

Fig 15: Districts of Haryana considered for analysis.

District	Symbol
KKR	+
Karnal	+
Bhiwani	0
Hissar	0
Jhajjar	+
Sirsa	0
Sonipat	+
Mahendergarh	+
Rohtak	+
Fatehabad	0

The fig15 show that the district having + sign fall in cluster1 and with 0 symbols falls in cluster2.

Output: The output includes:

1. **District Name:** Name of all used Haryana districts.

[Urvashi, 5(8): August 2018]

ISSN 2348 - 8034

DOI- 10.5281/zenodo.1406085

Impact Factor- 5.070

2. **Actual output:** The obtained Output after cutting the crop.

3. **Proposed Production:** The outcome using K-medoid clustering algorithm plus output of base prediction

4. **Error in Proposed Production1:** The amount of error during forecasting the yield using C.B.Err

5. **Base Production:** The outcome achieved in its own district.

6. **Error in Base Production 2:** The difference between Actual and Base production. It is represented by B.Err.

Predicted Production: The fig14 show the actual production of 2016. C.B.P is computed using clustering based prediction method. The C.B.Err is the amount of difference occurs in between actual production and predicted production using clustering algorithm. The B.P is prediction attained using base concept. The B.Err is the amount of error occurs during prediction.

District	Actual	C.B.P	C.B.Err	B.P	B.Err
KKR	18.51	18.405	0.105	18.714	0.204
Karnal	19.01	18.84	0.17	19.336	0.326
Bhiwani	17.6	17.804	0.204	17.564	0.036
Hisar	18.1	18.169	0.069	18.086	0.014
Jhajjar	17.52	17.535	0.015	17.471	0.049
Sirsa	18.5	19.205	0.705	19.566	1.066
Sonapat	19.2	18.755	0.455	19.214	0.014
Mahender	18.2	18.185	0.015	18.4	0.2
Rohtak	18.21	18.185	0.025	18.4	0.19
Fatehabad	19.8	19.204	0.591	19.571	0.229
			2.3		2.3

Fig16: Production according to districts

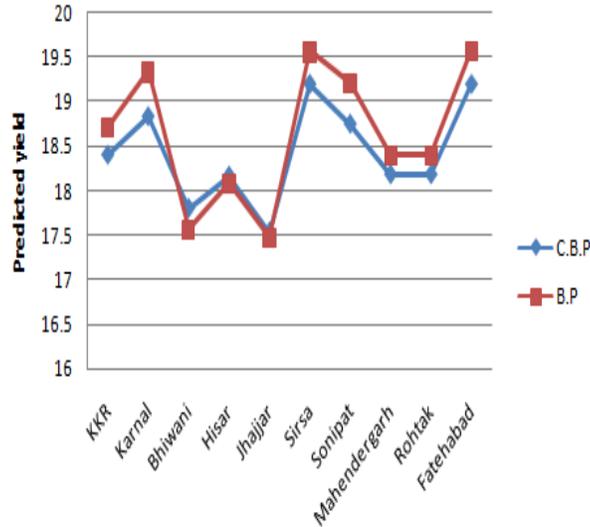


Fig17: Production using base and proposed method.

From Fig 18, it is clearly seen that the error rate in proposed method is less than the base method.

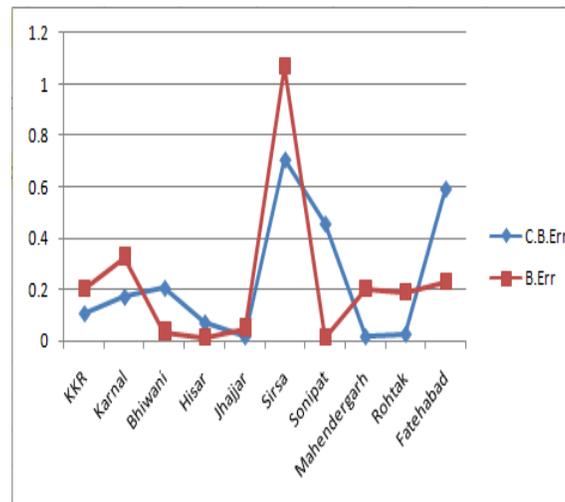


Fig18: Error rate in between base and proposed method implemented by K-Medoids

V. CONCLUSION

From the paper, it is concluded that the without feature extraction the improved works better than k-means, k-medoids but having very high number of iteration. The iteration number is reduced after using feature selection. This makes Improved K-Means as optimal Algorithm. Beside it, K-Medoids also work better than Basic K-Means in both feature selection or without feature selection. So, overall Improved K-Means and K-medoids are efficient algorithm on agricultural data. To predict the yield both algorithms are implemented. The result provided by K-Medoids forecast production near to actual production and improved K-Means lacks. At the end of the, it is paper consolidate

[Urvashi, 5(8): August 2018]

ISSN 2348 - 8034

DOI- 10.5281/zenodo.1406085

Impact Factor- 5.070

that K-Medoids and Improved K-Means both are optimal algorithms. One gives close output related to yield, other gives better results in feature selection. In future the work can be carried out on larger data set.

REFERENCES

1. Trimi Neha Tete, Sushma Kamlu, "Detection of Plants Diseases using Threshold, k-Means Cluster and Ann Algorithm", 2017 I2CT.
2. Dimo Dimov, Fabina Low, Mirzahayot, Ibrakhimov, Sarah-Schonbrodt-Stitt, Christopher Conrad, "Feature extraction and machine learning for the classification of active cropland in the Aral Sea Basin", 2017 IGARSS.
3. R.Pratheba, A. Sivasangari, D. Saraswady, "Performance analysis of pest detection for agricultural field using clustering techniques", 2014 ICCPCT.
4. Onur Yuzugullu, Esra Erten, Irena Hajnsek, "Rice Growth Monitoring by Means of X-Band Co-Polar SAR: Feature clustering and BBCH Scale", 2015 IEEE Geoscience and Remote Sensing Letters.
5. D. Ashok Kumar, N.Kannathasan, "A Study and Characterization of Chemical Properties of Soil Surface Data using K-Means algorithm", 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering.
6. Edmund W.Schuster, Sumeet Kumar, Sanjay E. Sarma, Jeffrey L. Willers, George A. Milliken, 2011 8th International Conference and Expo on Emerging Technologies for a Smarter World.
7. Renata Ribeiro do Valle Goncalves, Jurandir Zullo, Luciana Alvim Santos Romani, Bruno Ferraz do Amaral, Elaine Parros Machado Sousa, "Agricultural Monitoring using Clustering Techniques on Satellite Image Time Series of Low Spatial Resolution", 2017 9 International Workshop on the Analysis of Multitemporal Remote Sensing Image.
8. Sweta Singh, Sanya Ambegaokar, Kiran Singh Champawat, Animesh Gupta, Shirish Sharma, "Time Series Analysis of Clustering High Dimensional Data in Precision Agriculture", 2015 International Conference on Innovations in Information, Embedded and Communications Systems.
9. Jeromia J, K.V. Anusuya, "Energy Efficient Cluster Formation Algorithm and Sink Relocation Algorithm for Precision Agriculture", 2016 IC-GET.
10. Hilal Ahmad, Kalyanaraman Rajagopal, Ashiq Hussain Shah, Arif Hussain Bhat, Kalyanaraman Venugopal, "Study of Bio-Fabrication of Iron nanoparticles and their Fungicidal Properties against Phytopathogens of Apple Orchards", 2017 IET Nanobiotechnology.
11. Utpal Kumar Paul, Sudipta Chattopadhyay, "An Energy Saving Routing Scheme for WSN based Crop Field Monitoring System", 2016 IEEE.
12. Sonam Maurya, Vinod Kumar Jain, "Threshold Sensitive Region-Based Hybrid Routing Protocol for Precision Agriculture", 2016 IEEE Wireless Communications and Networking Conference.
13. "High Spatial Resolution Remote Sensing Images", 2015 IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.
14. Sudhir Gupta, Vinay Pandit, K. S. Rajan, "Remote Sensing Based Season Calendar for Indian districts using MODIS data", 2009 IEEE.
15. Kavi Kumar Khedo, Mohammad Riyad Hosseney, Mohammad Ziyad Toonah, "PotatoSense: A Wireless Sensor Network System for Precision Agriculture", 2014 IST-Africa Conference.
16. Xin Feng, Xiaoxi Hu, Yang Liu, "Radar Signal Sorting Algorithm of K-Means Clustering based on Data Field", 2017 ICCS.
17. Neetu Chahal, Anuradha, "A Study on Agricultural Image Processing along with Classification Model", 2015 IEEE International Advanced Computing Conference.
18. Monali Paul, Santosh K. Vishwakarma and Ashok Verma, "Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach", IEEE.
19. Dr. S. Hari Ganesh, Mrs. Jayasudha, "Data Mining Technique to Predict Accuracy of the Soil Fertility", International Journal of Computer Science and Mobile Computing.
20. Pooja G. Mate, Kavita R. Singh, Anand Khobragade, "Feature Extraction Algorithm for Estimation of Agriculture Acreage from Remote Sensing Images", 2016 IEEE.

RESEARCHERID



THOMSON REUTERS

[Urvashi, 5(8): August 2018]
ISSN 2348 - 8034
DOI- 10.5281/zenodo.1406085
Impact Factor- 5.070