International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-4 Number-3 Issue-16 September-2014

Evaluation of Modified K-Means Clustering Algorithm in Crop Prediction

Utkarsha P. Narkhede[1], K.P.Adhiya[2]

Abstract

An Agricultural sector is in need for well-organized system to predict and improve the crop over the world. The complexity of predicting the best crops is high due to unavailability of proper knowledge discovery in crop knowledgebase which affects the quality of prediction. In data mining, clustering is a crucial step in mining useful information. The clustering techniques such as k-Means, Expectation Maximization, Hierarchical Micro Clustering, Constrained k-Means, SWK k-Means, k-Means++, improved rough k-Means which make this task complicated due to problems like random selection of initial cluster center and decision of number of clusters. This works demonstrates an evaluation of modified k-Means clustering algorithm in crop prediction. The results and evaluation show comparison of modified k-Means over k-Means and k-Means++ clustering algorithm and modified k-Means has achieved the maximum number of high quality clusters, correct prediction of crop and maximum accuracy count.

Keywords

Clustering, Modified k-Means, Evaluation, Crop prediction.

1. Introduction

Data mining is a ground-breaking technology, developing with database and artificial intelligence. It is a processing overture of action of extracting trustworthy, novel, useful and understandable patterns from database. At current, data mining has been in business management, production control, electronic commerce, market analysis and scientific research and many other fields to explore a wide range of applications. Data mining with a view of its socking business vision are now becoming a data library and information strategy-making in the field of agriculture research [1],[2]. The research of agriculture field in data mining such as discovering wine fermentation and predicting yield using sensor data [3], a classification system for sorting mushrooms by grade [4], rainfall forecasting for crop growth [5] and highest humidity prediction [6] are developed. In order to do such data analysis in a clustering plays an important role for finding data information and pattern recognition in data mining.

Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in agreement to one another but are very dissimilar to objects in option clusters. A cluster of data objects can be treated collectively during the time that one group and so may be considered as a classic of data compression. Unlike classification, clustering is an effective means for partitioning the set of data into groups based on data similarity and then ascribe labels to the relatively small number of groups. Clustering is an unsupervised learning as it does not rely on predefined classes and class labelled training examples. For this reason, clustering is a form of learning by observation, rather than learning by examples. As shown in Figure 1, the three clusters are formed containing data points based on center position. The cluster center is shown by + signs. The quality of clusters depends on how dense it is. So, cluster having more number of points is cluster of good quality [2][7].

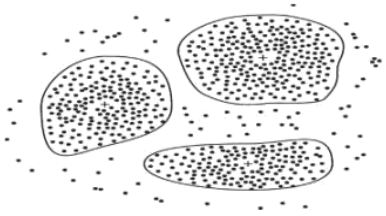

Figure 1: Cluster Analysis

The paper is organized as; initially motivation of crop prediction is described in Section II. In Section III, the clustering techniques are studied comparatively. The most efficient clustering technique leading to accurate clustering of crop records is found out. The proposed

Manuscript received August 24, 2014.
Utkarsha P.Narkhede, Department of Computer Engineering, SSBT'S College of Engineering & Technology Bambhori, Jalgaon, Maharashtra, India.

K.P.Adhiya, Department of Computer Engineering, SSBT'S College of Engineering & Technology Bambhori, Jalgaon, Maharashtra, India.

system which includes the crop knowledgebase, feature selection, three clustering approaches such as modified k-Means, traditional k-Means and k-Means++, sample testing and prediction and pattern visualization are described in Section IV. In Section V, the results and discussion describes evaluation of clustering algorithm by using parameters such as number of clusters, quality in terms of high and low and accuracy count.

2. Motivation

A crop prediction is a widespread problem that occurs. During the rising season, a farmer had curiosity in knowing how much yield he is about to expect. In the earlier period, this yield prediction become a matter of fact relied on farmer's long-term experience for specific yield, crops and climatic conditions. Farmer directly goes for yield prediction rather than concerning on crop prediction with the existing system. Unless the correct crop is predicted how the yield will be better and additionally with existing systems pesticides, environmental and meteorological parameter related to crop is not considered.

Promoting and soothing the agricultural production at a more rapidly pace is one of the essential situation for agricultural improvement. Any crop's production show the way either by interest of domain or enhancement in yield or both. In India, the prospect of widening the district under any crop does not exist except by re-establishing to increase cropping strength or crop replacement. So, variations in crop productivity continue to trouble the area and generate rigorous distress. So, there is need to attempt good technique for crop prediction in order to overcome existing problem [2][8].

3. Literature Review

To understand the advancement of clustering techniques, it is essential to briefly examine their history. In data mining, clustering plays an important role for finding data information and pattern recognition. Hierarchical micro clustering algorithm, constrained k-Means algorithm, SWK k-Means algorithm, expectation maximization algorithm, improved rough k-Means, k-Means++ and Beehive algorithm are clustering techniques.

In 2003,Hwanjo Yu et al. [9] presented clustering based support vector machine (CB-SVM) designed for handling very large data sets. Basically SVM is data classification method whose training complexity highly depends on size of data. So it is not worked for large dataset. In order to work authors have designed CB-SVM for handling large dataset. CB-SVM is a hierarchical micro clustering algorithm that scan entire data set only once to provide an SVM with high quality of sample that carry statistical summaries of the data. Hierarchical micro-clustering algorithm has randomly selected the number of cluster value and initial cluster center. It shows good quality of cluster for large dataset but it is expensive to update and store the cluster, also splitting and merging the data degrades performance.

In 2005, Kiri Wagstaff et al. [10] developed HARVIST (Heterogeneous Agricultural Research Via Interactive, Scalable Technology) graphical interface that allows user to interactively run automatic classification and clustering algorithm. They have used constrained k- means clustering algorithm for pixel clustering which merge the concept of constraint-based and partitioning methods. It shows good quality of clusters for huge datasets and also give better performance than hierarchical clustering, but it has drawbacks such as local optima problem, fix number of cluster, difficult to get initial value of cluster center, sensitive to noise.

In 2007, A Majid Awan et al. [11] has developed a software system for predicting Oil-Palm Yield from climate and plantation data. They used unsupervised partitioning of data for finding spatio-temporal patterns using kernel method. By using only k-means partitioning method it is burdensome to deal with abstract data so authors have incorporated kernel method. It shows good quality of clusters for huge datasets and also gives better performance than hierarchical clustering, but it has drawback of decision of number of cluster value.

In 2011, Sun Kim et al. [12] proposed model for theme-based clustering algorithm that capture probabilistic for text documents. Probabilistic clustering comes under model-based clustering methods in which data are generated by mixture of probability distributions. Given text, subject terms are extracted and used for clustering document in a probabilistic framework. An EM algorithm is used for learning the proposed model in order to ensure annals are assigned to correct themes. An EM is good in handling with real world dataset but it randomly

selects k value and becomes sensitive to noise and also highly complex in nature.

In 2012, DUAN weing-ying et al. [13], proposed improved k means clustering algorithm with weighted based on density. They proposed a solution to search initial central points and combine it with a distance measure with weight. An Improved k-Means clustering algorithm requires additional parameter such as density, threshold and number of cluster. It also reduces impact of noise data.

In 2007, David Arthur et al. in [14] proposed K-Means++ clustering algorithm by using randomized seeding technique. They proposed an optimal clustering by giving solution of speed and accuracy improvement over k-means algorithm. It is also good in handling large dataset but has drawback of number of cluster value and decision of initial center.

In 2013, M. Gunasundari et al. in [1], suggested crop yield prediction model which is used to predict crop yield from historical crop data set. A relational cluster Bee Hive algorithm is proposed for extracting yield patterns across multiple data sets. The outcome helps in identification of and investigates areas of unusually high or low yield. The Beehive algorithm is good in handling large dataset, initial cluster center and efficient in finding optimal solutions. But it has drawback of having number of tuneable parameters and k value.

Table 1: Comparison Chart

Algorithm	I/P Parameter	K value	Initial Centroids	Dataset	Shape	Noise
Hierarchical Micro Clustering	Branching factor, Diameter threshold	Sensitive	Randomly chosen	Large Dataset	Spherical	Yes
Constrained k-Means	Must- link, Cannot-Link, No.of cluster	Sensitive	Randomly chosen	Small Dataset	Spherical	No
SWK k-Means	Kernel Matrix, No. of Cluster, wt, penalty term	Sensitive	-	Large Dataset	Spherical	Yes
k-Means++	Number of Cluster, Cluster centers	Sensitive	Randomly Chosen	Huge Dataset	Spherical	Yes
EM	Number of Cluster	Sensitive	-	Real world Dataset	Spherical	No
Improved Rough k-Means	Density Threshold, Number of Cluster	Sensitive	-	Huge Dataset	Spherical	Yes
Beehive	Number of scout bee, site selected, qualified site, number of bees selected for best sites, size of patch	Sensitive	Randomly Chosen	Large Dataset	Spherical	Yes

The comparison table 1 shows that hierarchical micro clustering, constrained k-Means algorithm, k-Means++ and beehive are very sensitive in decision of number of cluster value and decision of initial centroid are always chosen randomly, EM, SWK k-Means and improved rough k-means algorithm have taken the calculated value of initial centroid but they are also sensitive in decision of number of cluster value. Excluding Beehive algorithm, all algorithms formed their cluster in spherical shape whereas beehive formed hexagonal shaped cluster. Except EM algorithm and constrained k-means algorithm, all algorithms are good enough to deal with noise.

4. Proposed System

With reference to literature work, authors have noticed the initial problems for clustering in this paper. The proposed solution introduces a better way for clustering by doing enhancement in partitions. In order to understand the working of architecture of crop prediction, it is essential to study the required blocks for architecture. The architecture of crop prediction includes the crop knowledge-base, feature selection, three clustering approaches such as k-Means, k-Means++, proposed modified k-Means algorithm, pattern visualization and sample testing and prediction.

4.1 Architecture of Crop Prediction

Architecture is a system that unifies its components or elements into a coherent and functional whole. The architecture of crop prediction is shown in Figure 2 and the block description is as follows.

Crop knowledge base: The crop knowledge base consists of farm knowledge such as crop types, soil types, soil-ph value, crop disease and pesticides [15], seasonal parameter such as kharif, rabi and summer crops. The knowledge-base also consists of zones as well as district information, environmental parameter such as maximum and minimum temperature value and average rainfall [16].

Feature Selection: The feature selection module is responsible for selection of attribute from crop knowledge-base for partitioning. It selects one record at a time from 396 x 10 records and performs calculation for partitioning. The three important steps for partitioning are as follows:

$$\text{Let } X = \begin{bmatrix} x_{1,a1} & x_{1,a2} & \dots & x_{1,an} \\ x_{2,a1} & x_{2,a2} & \dots & x_{2,an} \\ . & . & & . \\ x_{n,a1} & x_{n,a2} & \dots & x_{n,an} \end{bmatrix}$$

Where $X = \{x_1, x_2, \dots, x_n\}$ be the `n' objects.

$A = \{a_1, a_2, \dots, a_n\}$ be the `a' variables.
$Z = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$ be the minimum value of variables from X.

1. Compute the difference `β' of each records with minimum value of X.

$$\beta = (x_1 - \alpha_1) + (x_1 - \alpha_2) + , \dots . , (x_1 - \alpha_n) \dots . \quad (3.1)$$

2. Compute the summation in matrix M_{nx1} of each difference.

$$M_{nx1} = \sum_{i=0}^{n} \beta \quad \dots . (3.2)$$

3. Sort the matrix M_{nx1} and partition it according to number of iteration.

Instead of taking initial cluster center randomly, it is calculated based on partition data.

Center Value = round(size(dataclust)/2) …. (3.3)

Where, dataclust is size of partition data.

Instead of taking input as number of cluster value after partitioning, it is calculated first based on number of iteration value.

Suppose $W = \{w_1, w_2, \dots, w_n\}$ be calculated centroids. Apply both partition values as well as center values to k-Means function. Assign center value 'W' to the position and compute distance $d(x_i, w_j)$ for all W centers.

$$d_{i,j} = \sum_{i=1}^{a} \sqrt{|(x_{ij} - w_j)^2|} \quad \dots . (3.4)$$

Assign x_i to the cluster with minimum distance and for each w_j center value move the position of w_j to the mean of points in cluster. Then k-means function provides an output in form of idx and c value. Where, idx is cluster indices of each points and c is w x a matrix of centroid value. Compute the mean of idx value at every number of iteration and consider the highest mean value among all to perform better clustering. The index of highest mean is determined as number of cluster.

Clustering Approaches: The three clustering approaches is used such as modified k-Means, k-Means++ and traditional k-Means. The determined value of number of clusters and initial cluster centers is provided to modified k-Means clustering algorithm. Because of the number of clusters (k value) is required at starting for traditional k-Means and k-Means ++, the same calculated value of number of clusters is provided and initial cluster centers are uniformly chosen. All three approaches performed clustering and provide output in the form cluster number and centroid matrix.

Sample Testing and Prediction: There is need to provide input parameters such as zone, district, and selection of seasons, soil type, maximum temperature, minimum temperature and average rainfall for sample testing. Based on the output values of each clustering, the test data calculates the distance measure with clustering output and selects minimum distance as a predicted value. Then, the predicted cluster value is founded in output cluster number (idx) and as per the priority the very first output value of predicted cluster is selected. Then, the similar number of records of output value is founded in expected value and accuracy in terms of its count value is calculated. The accuracy count is shown by pie chart.

Pattern Visualization: Pattern visualization is done in order to observe the scenario of clustering. It has shown by Silhouette plot [17] and pie chart. Silhouette refers to a method of interpretation and validation of clusters of data. The technique provides a brief graphical representation of how well each object lies within its cluster. It was first described by Peter J. Rousseeuw in 1990. This plot shows number of clusters on X axis and dataset values on Y axis.

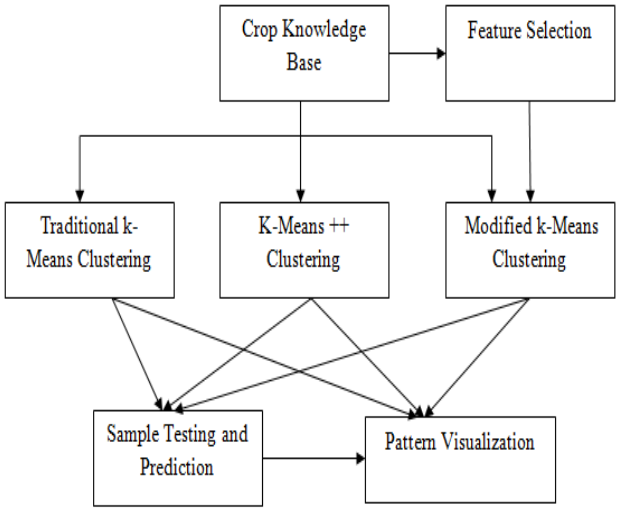

Figure 2: Architecture of Crop Prediction

4.2 Algorithms

The modified k-Means clustering algorithms is partitioning algorithm which differ in determination of number of cluster and selection of initial cluster centroid. The processing of number of cluster and initial cluster centroid by modified k-Means algorithms are as follows.

The modified k-Means algorithm as shown in algorithm 1 is proposed to determine initial cluster centers and number of cluster. The algorithm begins with finding the minimum value of attribute of X dataset. The minimum value is then subtracted with each record and added those subtracted value to get one representative value of each record. The complete representative values are sorted for partitioning. The partitioning is done on the basis of number of iterations. Each value of partitioned data is stored to calculate partitioned centroids one by one. The X dataset, iteration number and centroids value are given input to k-Means function. The steps of k-Means function are as follows.

1. The euclidian distance is computed of each point from X with centroids value and minimum distance value is assigned to the cluster.
2. The center value is move to the position by computing mean of all points in cluster. This is repeated until centroid value gets same.
3. The output value is acquired in terms of cluster index and centroid value. For cluster representation, the silhouette function is used for plotting those points.

Then, the mean value of cluster index is computed to get compact value. This is continued until number of iteration is covered. At end maximum mean is computed of all mean and its index value is determined as number of cluster. According to index value the sorted representative value is again partitioned and same procedure is repeated up to the k-Means function to get refined output values.

Algorithm 1: Proposed Modified k-Means
Require:X(dataset contain 'n'number of points), **Z**(set of minimum value of variable from X),**Sz**(size of X), **T**(number of iteration).
Output: Idx(Number of cluster) and C(Centroid)
1: **for all** i such that i Є **X do**
2: Compute the difference with Z….Equation 3.1
3: Compute the summation of the subtracted value... Equation 3.2.
4: **end for**
5: Sort the computed summation.
6: **for all** ii such that ii Є **T do**
7: Partition sorted values as per T
8: Store partitioned data in data-clust(clust)
9: **for all** j such that j Є **ii do**
10: Compute the size of data-clust(clust)
11: Compute the initial centroid value…Equation 3.3
12: Store the centroid value in W.
13: **end for**
14: Apply X, initial number of partition and clustered-data to k-Means function.
15: Assign initial centroids from W.
16: **for all** x_i Є X do
17: Compute the Euclidean distance $d(x_{i,j},w_i)$ for each W….Equation 3.4
18: Assign x_i to the cluster with minimum distance and for each w_i Є W center value move the position of w_i to the mean of a points in cluster.
19: Repeat step(17) and step(18) until the value of centroids is same.
20: **end for**
21: Store the cluster index and centroid value in idx and c respectively.
22: Assign the idx value and X matrix to Silhouette function for plotting of data points.
23: Compute the mean of idx value
24: **end for**
25: Compute the maximum mean of all mean values of idx.
26: Partition the sorted value as per the maximum mean.
27: Repeat Step(8) to Step(22) until all index values is covered.

The sample testing, as shown in algorithm 2, is performed with the help of test data. The Euclidian distance is calculated between test data and centroids which is output value of clustering algorithm. The minimum distance of euclidian distance is considered as predicted distance and its cluster number is considered as predicted value. The predicted cluster number is found in output value of modified k-Means clustering algorithm, and stored it with all the records that belong to same predicted cluster. Then test data is found in all the records that belong to predicted cluster and its first output value is stored as predicted output value. The accuracy count is computed by counting predicted output value in all output values of all records that belong to predicted cluster.

Algorithm 2: Sample Testing
Require: B(Test data contains I number of points), **Idx (**number of clusters**)** and **C**(Centroids)
Output: Predicted Crop and Accuracy Count
1: **for all** i such that i Є **B do**
2: Compute Euclidean distance $d(B_{i,j}, c_i)$ for each C
3: **end for**
4: Get the minimum distance measured(B,C) and its cluster number is predicted value.
5: Find predicted cluster number in Idx and store it in 'aa'. 'aa' contains all records that belong to predicted cluster.
6: Find the test data 'B' in 'aa' and store its output value as predicted output value.
7: Compute the accuracy count by counting predicted output value in out(aa). 'out(aa)' contains all output values of all records that belong to predicted cluster.

5. Results and Discussion

The clustering algorithm such as proposed modified k-Means, traditional k-Means and k-Means++ are evaluated on crop dataset by using parameters such as number of clusters, quality in terms of high and low and accuracy count. The simulation environment used for evaluation of modified k-Means clustering algorithm for crop prediction is Matlab R2012a. For the experiments the dataset [15],[16] is provided by department of agriculture Maharashtra state to build the crop knowledge base. In order to perform clustering the number of iteration is important parameter to set. Based on number of iteration, partition is performed. The number of cluster and cluster centroid is determined and given as an input to k-Means function of MatlabR2012a. The k-Means gives the `Idx' and `C' value as an output. The representation of cluster is done by Silhouette plot of MatlabR2012a. For testing, the input parameter needs to be set are zone, districts, types of season, soil types, soil ph, temperature and rainfall value. The distance measure between these input value and obtained centroid value are computed and minimum distance is considered as predicted. For accuracy the count of predicted value is computed to get similar number of records.

The experimental results are shown in silhouette representation to determine the difference between cluster efficiently. The Silhouette plot gives the range of [0-1] such that representation of data points is plotted on X axis and number of clusters on Y axis. The range [0-1] shows the quality of cluster. The cluster towards ONE value is of best quality of cluster than ZERO [18]. Silhouette representations based on number of iteration are as follows.

Figure 3 shows the k-Means output when number of iteration is FIVE. The number of cluster formed on Y axis is THREE. As the X axis shows the quality of cluster. k-Means formed no cluster such that it reached towards ONE value and formed ONE cluster that reached towards -0.2 means cluster is not properly grouped. Based on this, k-Means gives 30 similar records in terms of accuracy.

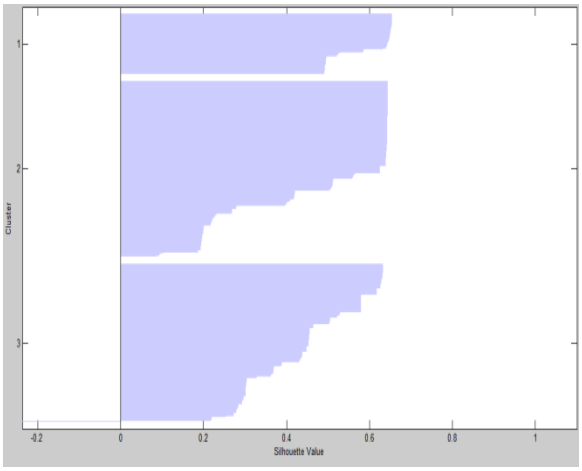

Figure 3: k-Means Output for Number of Iteration = 5

Figure 4 shows the k- Means++ output when number of iteration is FIVE. The number of clusters formed on Y axis is THREE. As the X axis shows the quality of cluster. k- Means++ formed no cluster such that it reached towards ONE and ZERO value which means clusters are grouped on average. Based on this k-

Means++ gives 34 similar records in terms of accuracy.

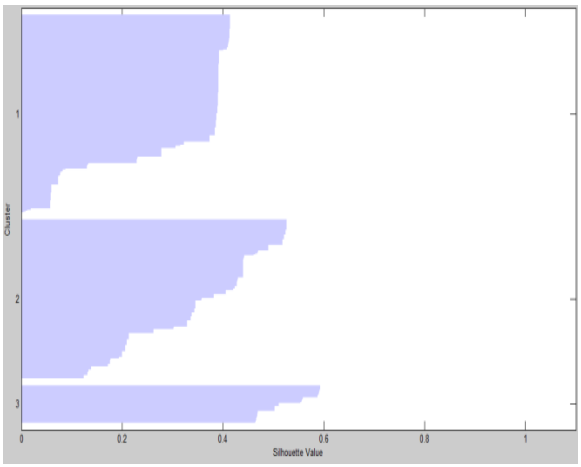

Figure 4: k-Means++ Output for Number of Iteration = 5

Figure 5 shows the proposed modified k-Means output when number of iteration is FIVE. The number of clusters formed on Y axis is THREE. As the X axis shows the quality of cluster. Proposed modified k-Means formed ONE cluster such that it reached towards ONE value with more number of points which means cluster are tightly grouped. Based on this, the proposed modified k-Means gives 36 similar records in terms of accuracy. Figure 6 shows pie chart of accuracy in terms of similar records. Out of three clustering algorithm, proposed modified k-Means shows maximum accuracy count.

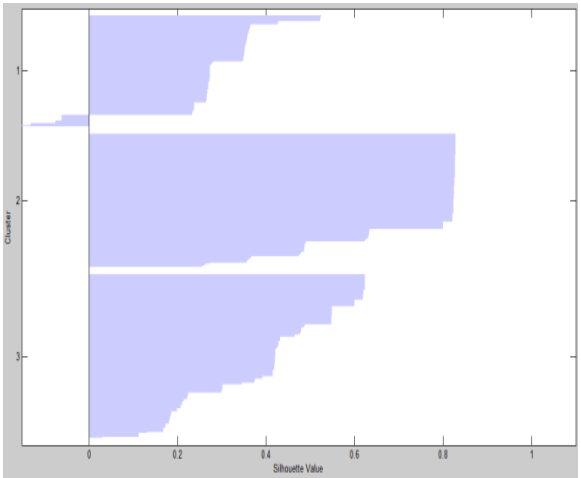

Figure 5: Modified k-Means Output for Number of Iteration = 5

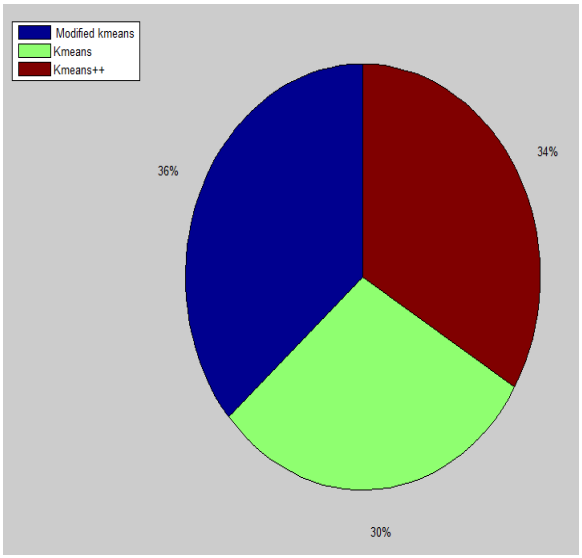

Figure 6: Accuracy for Number of Iteration = 5

The overall comparison of proposed modified k-Means, k-Means and k-Means++ clustering approaches in terms of Number of cluster, Quality, Accuracy Count are as follows. Table 2, 3 and 4 show comparison of three clustering approaches for number of iteration (NOI) = 5, 10 and 20 respectively. Out of those, proposed modified k-Means has shown the good results by achieving high quality of cluster and maximum number of similar count.

Table 2: Comparison Table [NOI = 5]

Algorithms	No. of Cluster	Quality		Count of similar number of records
		Low	High	
k-Means	3	1	-	30
k-Means++	3	-	-	34
Modified k-Means	3	1	-	36

Table 3: Comparison Table [NOI = 10]

Algorithms	No. of Cluster Form	Quality		Count of similar number of records
		Low	High	
k-Means	8	1	1	31
k-Means++	9	1	1	33
Modified k-Means	9	1	5	36

Table 4: Comparison Table [NOI = 20]

Algorithms	No. of Cluster	Quality		Count of similar number of records
		Low	High	
k-Means	12	2	5	29
k-Means++	20	5	10	30
Modified k-Means	20	4	8	40

Based on the quality of cluster, the crops are predicted. Table 5 shows the sample results in which proposed modified k-Means has predicted crop correctly as compared to k-Means and k-Means++. Table 6 shows attribute details required for crop prediction.

Table 5: Sample Results

Z	D	Kh	Rb	S	ST	SP	MxT	MiT	AvR	Mk	k++	k
0.1	0.01	1	0	0	0.1	7.3	0.346	0.192	0.075	**Rice**	**Rice**	**Rice**
0.1	0.01	1	1	1	0.1	7.3	0.346	0.192	0.075	**Maize**	Rice	**Maize**
0.2	0.14	1	0	1	0.3	5.75	0.28	0.14	0.16	**Groundnut**	Rice	**Groundnut**
0.2	0.15	1	0	0	0.2	5.75	0.29	0.13	0.45	**Rice**	**Rice**	**Rice**
0.2	0.15	1	1	0	0.2	5.75	0.29	0.13	0.45	**Jowar**	Rice	**Jowar**
0.3	0.15	0	1	0	0.3	6.5	0.28	0.14	0.16	**Potatoes**	Rice	**Potatoes**
0.3	0.15	1	1	1	0.3	6.5	0.28	0.14	0.16	**Chillies**	Rice	**Sugarcane**
0.4	0.14	1	0	1	0.4	7.8	0.4	0.05	0.095	Sunflower	Rice	Maize
0.5	0.21	1	0	0	0.5	6	0.35	0.23	0.3150	**Ragi**	Rice	Rice
0.5	0.21	1	0	0	0.5	6	0.35	0.23	0.3150	**Coconut**	Rice	Sugarcane

Table 6: Attribute Details

Abbreviation	Attribute - Name
Z	Zone
D	District
Kh	Kharif
Rb	Rabi
S	Summer
ST	Soil Types
SP	Soil Ph
MxT	Maximum Temperature
MiT	Minimum Temperature
AvR	Average Rainfall
Mk	Modified k-Means
k++	k-Means++
k	k-Means

6. Conclusion and Future Work

Clustering is a data mining algorithm and plays significant role for extracting knowledge and update of information. Clustering technique applied in crop dataset has resulted in novel approach which has significance success in predicting crop. However, the main drawback of existing clustering algorithms like random initialization of cluster centers and uniform provision of number of clusters as an input are pointed out. The drawbacks are overcome by proposing modified k-Means clustering algorithm which used the formulated value to initialize cluster centers and to determine number of clusters. This work demonstrates about modified k- Means clustering in crop prediction by increasing quality and accuracy count. The modified k-Means clustering algorithm is evaluated by comparing k-Means and k-Means++ algorithms and achieved the maximum number of high quality clusters, correct prediction of crop and maximum accuracy count. Data mining plays a crucial role in Agriculture sector for better prediction of crop. The proposed work is done on crop dataset belong to Maharashtra State. Our future work includes to consider geographical area using world geographic information system for global crop prediction.

Acknowledgment

The authors feel a deep sense of gratitude to Prof. Dr. G. K. Patnaik HOD of Computer Science and Engineering Department for his motivation and support during this work. The authors are also thankful to the Principal, SSBT's College of Engineering and Technology Bambhori, Jalgaon for being a constant source of inspiration.

References

[1] M. Ananthara, T. Arunkumar, and R. Hemavathy, Cry: An improved crop yield pre-diction model

using bee hive clustering approach for agricultural data sets," in Pattern Recognition, Informatics and Medical Engineering (PRIME), 2013 IEEE International Conference , pp. 473-478.

[2] U.P. Narkhede and K.P.Adhiya, " A Study of Clustering Techniques for Crop Prediction - A Survey", American International Journal of Research in Science, Technology, Engineering & Mathematics, vol 1, Issue 5, ISSN no: 2328-3491, pp. 45-48 , 2014.

[3] A. Mucherino and G. Rub, "Recent developments in data mining and agriculture", Proceedings of russ 2011 International Conference on advances in data mining: Leipzig, Germany: IBaI Publishing, pp. 1-9, September, 2011.

[4] S. J. Cunningham and G. Holmes, "Developing innovative applications in agriculture using data mining", Proceeding of international conference on SEARCC', 1999.

[5] M. Kannan, S.Prabhakaran, and P.Ramachandran, "Rainfall forecasting using data mining technique," International Journal of Engineering and Technology, vol. 2, no. 0975-4024, pp. 397-401, 2010.

[6] Kiruthika.V.G, Arutchudar.V, and S. Kumar.P, "Highest humidity prediction using data mining techniques," International Journal of Applied Engineering Research, vol. 9, no. 16, pp. 3259{3264, 2014.

[7] J. Han, M. Kamber, and J. Pei, Data Mining, Second Edition: Concepts and Techniques, ser. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2006. [Online]. Available: http://books.google.co.in/books?id=AfL0t-YzOrEC.

[8] R. A.A. and K. R.V., Review - role of data mining in agriculture," International Journal of Computer Science and Information Technologies(IJCSIT), 2013,vol. 4(2), no. 0975-9646, pp. 270-272.

[9] J. Y. Hwanjo Yu and J. Han, in Classifying Large Data Sets Using SVMs with Hierarchical Clusters", Proceedings of Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 306-315, New York, NY, USA, 2003.

[10] D. M. Kiri L. Wagstaff and and S. R. Sain, "Harvist: A system for agricultural and weather studies using advanced statistical methods," 2005. [Online]. Available: https://www.agriskmanagementforum.org.

[11] A.Awan and M.Md.Sap," A framework for predicting oil-palm yield from climate data", International Journal of Information and Mathematical Sciences 3(2), 111-118 ,2012.

[12] S. Kim and Wilbur, " An em clustering algorithm which produces a dual representation," Proceedings of 10th IEEE International Conference on Machine Learning and Applications and Workshops (ICMLA), vol. 2, pp. 90-95,2011.

[13] Duan, Weng-ying, Tao-rong Qiu, Long-zhen Duan, Qing Liu, and Hai-quan Huan. "An improved Rough K-means algorithm with weighted distance measure." In Granular Computing (GrC), 2012 IEEE International Conference on, pp. 97-101. IEEE, 2012.

[14] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in SODA '07 Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithm, pp. 1027-1035, 2007-2013.

[15] D. More, Ed., Mahatma Phule Krishi Vidyapeeth, Rahuri (An Agriculture University), pp 38-129, 2013.[Online] Available: http://www.mpkv.mah.nic.in/.

[16] "Department of agriculture, maharashtra state," Accessed on 12-Feb-2014. [Online]. Available: http://www.mahaagri.gov.in/.

[17] W. L. Martinez and A. R. Martinez, Exploratory Data Analysis with MATLAB, Chapman & Hall/CRC Press LLC,UK, pp 147 -150, 2005.

[18] "Documentation center," Accessed on 10-April-2014. [Online].Available:http://www.mathworks.in/help/stats/k-means clustering.html#brah7f1-1.

Utkarsha P. Narkhede, Research Scholar, received Bachelor's Degree in computer science from North Maharashtra University in 2012 and pursuing Master's Degree in Shram Sadhana Bombay Trust's College of Engineering and Technology, Bambhori, Jalgaon, India.

Krishnakant P. Adhiya is working as Associate Professor in Computer Engineering Department at Shram Sadhana Bombay Trust's College of Engineering and Technology, Bambhori, Jalgaon, India. He has teaching experience of 23 yrs. He has completed Bachelor's Degree in 1990 from Govt. College of Engineering Amravati, India and obtained Masters Degree in 1996 from M.N.R.E.C., Alahabad, India.