Unveiling Soil Diversity: Leveraging Enhanced K-Means for Classification

R.Sudha Abirami[1], M.S. IrfanAhmed[2]

Department of Computer Science & Research Centre, Alagappa University, India

Department of Science & Humanities, Anna University, India

Abstract

Agriculture is the main source for the people of India. Crop yield prediction is very popular among farmers which mainly contributes to the suitable selection of crops for sowing. For good crop yield, farmers should be known of the suitable soil type for a particular crop. The possible way to improve productivity is that chooses a right crop for the rightland type. This paper introduces an innovative method to select suitable features from a data set for crop prediction. The most important factors of the clustering process of the k-means algorithm are the selection of the initial clustering center point and the distance measurement between the data objects. The traditional k-means algorithm uses Euclidean distance to measure the distance between sample points, thus it suffers from low differentiation of attributes between data objects and is provided the optimal results. This paper proposes an improved k-means algorithm based on Manhattan distance and the initialization method of farthest first. The traditional Euclidean distance method is replaced by the Manhattan distance method for measuring the distance between sample points. The experimental results show that the improved k-means algorithm based on Manhattan distance and farthest first initialization method proposed in this paper. It produces a better clustering effect and the convergence of the algorithm is also well improved.

Keywords: Manhattan distance, kmeans algorithm, clustering algorithms

I. INTRODUCTION

Clustering is the process of forming groups which have similar data objects based on their same properties. All the data objects present in same cluster should have same properties and is similar objects [1]. Data mining algorithmic clustering methods, dimension reduction methods, parallel clustering methods, and MapReduce-based clustering methods are the different types of clustering methods [2]. In the meantime, partitioned clustering methods is also a one type of data mining algorithmic clustering that integrates different algorithms like K-means, K-modes, K-medoids, PAM, CLARA, CLARANS, and FCM. One of the commonly used algorithms for clustering implementation is the K-means clustering algorithm [4], which is very common due to the best performance for datasets [5]. In the traditional K-means algorithm, K points are firstly selected as initial centroids, where each centroid represents a cluster. Many initialization methods are available in data mining for implementing clustering concept. Then all data objects of the dataset are assigned to the centroids with the minimum distances. After the distribution of all data items, centroids are calculated again until no more objects change their cluster [6]. Mostly, Euclidean distance is applied for clustering algorithms. But the distribution may take extreme because it needs to recalculate the distance mathematical equation during each iteration. As a result, numerous mathematical metrics are put out to enhance the distance computation.

II. LITERATURE REVIEW

Numerous research techniques have been used to assess the effectiveness and performance of the K-means clustering algorithm, enhancing its running time, accuracy, and cluster quality [1, 3]. An enhanced version of the conventional K-means algorithm was used by Kaur et al. [4] to compress images more effectively and with a shorter running time.

Loohach et al implemented the K-means clustering algorithm with Euclidean distance and Manhattan distance measures and evaluated the result by using the number of iterations. Their findings indicated that applying various distance

measurements could have an impact on the number of iterations [6].

Sajana et al. [8] focused on a keen study of different clustering algorithms, highlighting the characteristics of data mining techniques and an overview of various distance methods applying in clustering methods.

Rathore et al [7] introduced a new technique to implement a K-means clustering algorithm instead of traditional K-means. A bi-part method was used to divide the dataset into distinct clusters after first improving the quality of the clusters by eliminating outlier elements. Sample data points were grouped into the closest clusters based on the distances between the remaining sample points and the initial cluster centres. The experimental results were compared with the traditional K-means algorithm and showed better accuracy by removing the de-efficiency.

III. EXISTING MODEL

Traditional K-Means Algorithm

The core idea of the k-means algorithm is: After inputting the k number of data objects, randomly select k sample points in the sample point set as the initial clustering centroid. The sample points are sorted into the nearest clusters based on the determined distances between the remaining sample data points and the original cluster centres. In the formed new clusters, new cluster centroids are identified and the data objects are grouped and categorized again until the clustering results no longer change. In the actual clustering process, after many iterations, due to several factors, the termination conditions may not be met. When number of iterations is maximum, the calculation will be stopped.

The pseudo-code of the Traditional k-means algorithm:

Input: data set, k value

Output: divided into k clusters

1. select k points from the sample Euclidean from sample point xi to each cluster center
2. **repeat**
3. **for** j=1, 2, ... , m
4. compute the Euclidean distance from sample point xi to each cluster center
5. determine the cluster class mark of xi according to the closest distance
6. group the sample points into corresponding clusters
7. **end for**
8. compute new cluster center point.
9. **until** the cluster is allocated and result remains unchanged

IV. INITIALIZATION METHODS

As *k*-means clustering aims to converge on an optimal set of cluster centroids and cluster objects based on distance from these centroids via successive iterations, the fewer iterations of the *k*-means clustering algorithms will be required for grouping. There are several initialization strategies:

i) Random Partition Method
ii) k means++
iii) Canopy clustering algorithm
iv) Furthest first initialization.

i) Random Partition Method

Using this procedure, a random cluster ID is assigned at random to each data point. After that, the cluster points are created using their cluster IDs, and the starting points are obtained by calculating the average (per cluster ID). It is well known that the Random Partition method can determine the yield beginning points that are closest to the data objects' mean value.

ii) k Means++

First, we select a random point from the dataset. Next, we choose the point that is most likely to be located far from the first point. The squared distance between a point and the first centre is used to compute the sampling data object from a probability distribution.

A probability distribution proportional to each point's squared distance from its closest centre generates the

remaining points. A data object is therefore more likely to be sampled if it is far from its closest cluster.

iii) Canopy clustering algorithm

The **Canopy clustering algorithm** is pre-clustering and unsupervised clustering algorithm. It is mainly used in grouping the objects before the cluster process is done in K-means or Hierarchical clustering algorithm. Its goal is to expedite clustering processes on big data sets, where it would not be feasible to use a different technique directly because of the size of the data set.

1. Start with the set of data points to be clustered.
2. Remove a point from the set, beginning a new 'canopy' containing this point.
3. For each point left in the set, assign it to the new canopy if its distance to the first point of the canopy is less than the loose distance.
4. If the distance of the point is additionally less than the tight distance, remove it from the original set.
5. Repeat from step 2 until there is no more data points present in the set to cluster.
6. A more costly but precise approach can be used to sub-clustered these comparatively inexpensively clustered canopies.

iv) Furthest first initialization

By calculating the Manhattan distance—which is the total of the absolute differences between the two vectors point to all other potential points—it aims to improve K-means findings. The point that is the furthest away from the first centre in terms of Euclidean distance is selected as the second initial centre.

V. DISTANCE MEAUSRES

The K-Means approach computes the distance between each dataset point and each initialization centroid. Points are allocated to the centroid with the shortest distance based on the values discovered. Therefore, the clustering algorithm relies heavily on this distance measure.

i) Euclidean distance:

The Euclidean distance formula is to find the distance between two points on a plane. The formula for Euclidean Distance is the distance between two points (x1,y1) and (x2,y2) is $\mathbf{d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}}$.

ii) Manhattan distance:

Manhattan distance is calculated with the sum of the absolute differences between the two vectors. The formula for the Manhattan Distance between two points (X1, Y1) and (X2, Y2) is given by |X1 – X2| + |Y1 – Y2|

VI. PROPOSED MODEL

i) Flow Diagram

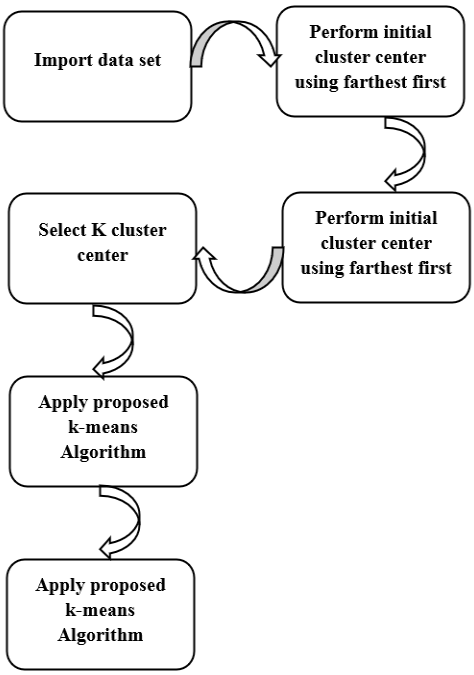

Figure 1 Flow Diagram

ii) Enhanced K-Means algorithm

Step 1: For a given data set, randomly select k data sample points as the initial cluster center using farthest first method

Step 2: Use each sample point's attribute value to create the evidence body of each sample.

Step 3: Use Manhattan distance method to calculate the distance from each sample point to each initial cluster centroids, select the center with the smallest distance, and add the cluster center to the cluster.

Step 4: Select k cluster center points again.

Step 5: Find whether the clustering center has been changed or not, if it has changed. Repeat the iteration, if the cluster center points remain the same. Then it shows the output of the corresponding clustering result.

Data Set:

The data set used comes from the UCI data set. The name of the data set is cpdata and its attributes are

@attribute temperature numeric
@attribute humidity numeric
@attribute ph numeric
@attribute rainfall numeric
@attribute label

iii) Methodology

(1) Import the data set and enter the number of k objects in cluster.

(2) The traditional k-means method which is implemented by using Euclidean distance metrics and the improved k-means method by using Manhattan distance metrics are used for clustering, respectively.

(3) Perform clustering as many times, find the execution time and number of iterations as the final result.

(4) Compare the experimental results with enhanced k-means algorithm and traditional k-means algorithm.

iv) Result and discussion

In the cpdata dataset, the running time of enhanced k-means algorithm by using Manhattan distance is better than that of the traditional k-means algorithm by using Euclidean distance method and the number of iterations is less. Therefore, the improved-K-means algorithm that introduced in this paper has better convergence.

Initialization methods	No of iterations	Time taken
Random	16	0.27
K means++	21	0.39
Canopy	4	0.12
Farthest first	2	0.08

Table 1: Performance of enhanced k means algorithm for different initialization methods

Table 1 shows the value of the number of iterations of the algorithm and the algorithm running time for different types of initialization methods.

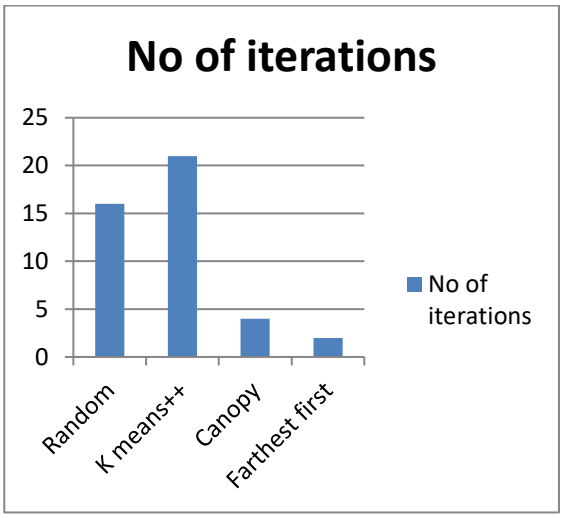

Figure 2: Number of iterations for different initialization methods.

The analysis of the results in this figure 2 shows that the Farthest first algorithm used in this enhanced k-means method by using soil data set has the lowest number of iterations and the running time is comparable with Traditional K-means clustering algorithms with enhanced k-means clustering algorithms.

	K means (Euclidian distance)	Manhattan distance
Time taken to build the model	0.09	0.05
Number of iterations	13	5

Table 2: Performance of K-Means algorithm with different distance methods

The comparison of the performance of running time and number of iterations of the K-means

algorithm by using Euclidean and Manhattan distance methods are displayed in the Table 2.

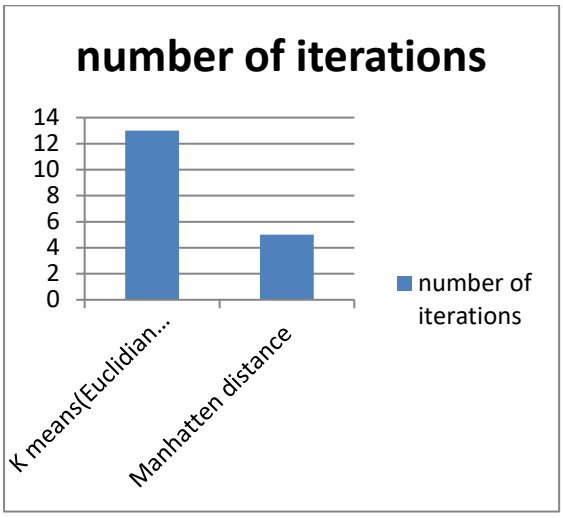

Figure 3 Number of iterations for different distance metrics

It is observed that Manhattan distance methods take the less time and a smaller number of iterations than the Euclidean distance method. The enhanced k-means algorithm using Manhattan distance metrics performs good results when compared to traditional k-means algorithm which is implemented by using Euclidean distance methods. The Figure 3 is shown different distance metrics.

VII. CONCLUSION

The enhanced k-means Algorithms takes pH, humidity, rainfall and temperature as input and predicts the crop based on particular soil. The rules forecast the yield of maize, wheat, and rice depending on the characteristics of the soil. The distance measure between sample points is performed using the Manhattan distance instead of the Euclidean distance. Finally, the k-means algorithm using Manhattan distance method is applied to cluster the soil. Through experimental comparison, the improved k-means algorithm based on Manhattan distance that proposed in this paper has good clustering effect and convergence. The initial clustering centers are still selected using farthest first method instead of randomly, so it can be further optimized.

VIII. REFERENCES

[1] G. Tzortzis and A. Likas, "The min-max k-means clustering algorithm," Pattern Recognition, vol. 47, no. 7, pp. 2505–2516, 2014.

[2] M. Emre Celebi a, Hassan A. Kingravi, Patricio A. Vela, comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Systems with Applications, vol. 40, pp. 200–210, 2013.

[3] Sofianita Mutalib, S-N-Fadhlun Jamian, Shuzlina Abdul Rahman, Azlinah Hj Mohamed, Soil classification: An application of self organising map and k-means, 10th International Conference on Intelligent Systems Design and Applications, ISDA, DOI:5687224, 2010.

[4] Kazheen Ismael Taher, Adnan Mohsin Abdulazeez and Dilovan Asaad Zebari, Data Mining Classification Algorithms for Analyzing Soil Data, Asian Journal of Research in Computer Science 8(2): 17-28, 2021; Article no.AJRCOS.68035 ISSN: 2581-8260, 2021.

[5] S.Manimekalai, K.Nandhini, Survey on Classification Techniques for Soil Data Prediction to Better Yielding of Crops, International Journal of Computer Sciences and Engineering, Volume-6, Issue-1 E-ISSN: 2347-2693, 2018.

[6] N. Saranya, A. Mythili, Classification of Soil and Crop Suggestion using Machine Learning Techniques, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181,Vol. 9 Issue 02, 2020.

[7] Jharna Majumdar, Sneha Naraseeyappa & Shilpa Ankalaki, Analysis of agriculture data using data mining techniques: application of big data, Journal of Big Data volume 4, Article number: 20, 2017.

[8] R. Sudha Abirami, M. S. Irfan Ahmed, "A Comparison of Data Mining Classification Algorithms using Soil Dataset", Journal of Emerging Technologies and Innovative Research (JETIR), Volume 9, Issue 5, ISSN-2349-5162, May 2022.

[9] Ramesh Vamanan, K. Ramar, Classification of Agricultural Land Soils: A Data Mining Approach,

Agricultural Journal 6(3):82-86, DOI:10.3923/aj.2011.82.86, 2011.

[10] V. Rajeswari and K. Arunesh, Analysing Soil Data using Data Mining Classification Techniques, Indian Journal of Science and Technology, DOI: 10.17485/ijst/2016/v9i19/93873, Volume: 9, Issue: 19, Pages: 1-4, 2016.