ISSN: 2581-3404 (Online)
International Journal of Innovative Research in Technology and Management, Vol-5, Issue-5, 2021.

Inhabited Community Detection Using Improved K-Means Clustering Algorithm

Kanishka Sisodia[1], Chinmay Bhatt[2], Varsha Namdeo[3]

CSE, SRK University, Bhopal, India[1,2,3]

Kanishkasisodia@gmail.com[1], chinmay20june@gmail.com[2], varsha_namdeo@yahoo.com[3]

Abstract-: *In clustering objects those have comparative nature will lies in a similar cluster and on the off chance that they are of unmistakable nature, they will be in various cluster. However Standard K-means is prime calculations of the clustering yet it experience the ill effects of certain detriments these are as per the following 1) Performance relies upon introductory groups which are picked haphazardly in standard K-means. 2) The standard K-means calculation has time intricacy of 0(nkl) that is an excessive amount of costly. 3) The standard K-means calculation likewise experiences the dead unit issue those outcomes in clusters without any information focuses. 4) In standard K-means we do arbitrary instatement which makes them chat at nearby minima. Numerous upgrades were proposed to work on the presentation of the standard K-means calculation yet the vast majority of them address just each of them in turn. In this paper, we address introductory focus just as calculation intricacy issue in one calculation. Now a day, population growth rate increase rapidly. So the size of the population database is increased exponentially. It is very difficult to find information from this huge dataset. Both clustering and classification algorithm is used to extract data from population database. The size of family, Population Density, Birth Rate, Death Rate, number of Employed person, Unemployment, Prediction of Male person, Prediction of Female population, Prediction of Budget for the year, Prediction of members in each caste, Prediction of Rural population and Prediction of Urban population etc.*

Keywords:- K Means, Clustering, Community Detection, Machine Learning.

Introduction

Now a day, population growth rate increase rapidly. Today's size of the population database is increased exponentially. It is very difficult to find information from this huge dataset. Both clustering and classification algorithm is used to extract data from population database. To conquer the constraint of the past applied calculation referenced in the above similar investigation of different exploration papers. We can improve the time intricacy; decrease no. of Iteration to find the centroid of the groups. Naturally get the underlying centroids without entering it. We can anticipate populace of various kinds and furthermore foresee Academic execution by directing different assessments, appraisals and one more type of estimations.

Anyway scholarly execution might shift from one understudy to another as every understudy has the distinctive degree of execution. The scholarly exhibition of an understudy is typically put away in different configurations like records, archives, records and so on the accessible information would be dissected to remove helpful data. In the event that we early discover understudy with low-execution record, we can give some therapeutic to those understudies. This will help in working on the nature of the understudies. (Bidgoli et al., 2003)

K-implies is experiencing numerous weaknesses, for example, its presentation relies upon beginning bunches which are picked haphazardly in standard K-implies, the standard K-implies calculation is computationally costly, standard K-implies calculation contains the dead unit issue that outcomes in void groups, In standard K-implies we do arbitrary instatement which drives standard K-means to merge to nearby minima. It is extremely challenging to plan a nonexclusive bunching calculation. So there are numerous improvements are given by various specialist. Every one of them thinks about each of them in turn. So there requires a decent compromise between time intricacy and nature of bunch.

The Basic K-means grouping procedure is easy to execute, and we start with a portrayal of the fundamental calculation. In Basic K-mean we initially pick no of required group here it is addressed by k. As fundamental K-means is partitional calculation so it is needed to give some of a group as information. Subsequent to indicating the quantity of the group we need to give their relating starting centroids. Here introductory centroid ought to likewise be given as the info boundary. Essential K-means is the iterative technique where at each progression we attempt to decrease the intra bunch distance. For great bunching, we ought to boost the between group remove and limit the intra-group distance. In Basic K-means we initially compute the distance of each point from all bunch centroids. Dole out the highlight the bunch that has least separation from group centroid. We will stop the cycle when old group and new bunch are same. It is hypothetically demonstrated that K-means chat to bring about a limited number of step, yet some time number of steps increments to an extreme so we can adjust the ending compel.

II. Related Work

A brief overview of literature on K-means and student performance is given in this section. The aim of the review is to come up with set of research a lot of research has been done in recent years in the area of K-means and student performance.

K-Medoids Clustering algorithm is good if your data contain the outliers in the data .This algorithm is more efficient if our data set contain the noise in the data. In k-means algorithm we are trying to reduce the intra cluster distance in the data set while in k-medoids we trying to reduce the absolute error in the data. It is also an iterative algorithm as k-means is it terminate if each representative object is actually the medoid of the cluster. Here we have a random point xi and m is the representative of the cluster and we swap the xi with the cluster representative m. These may moves closer to new representative point or it may happen that it will move closer to other representative point. There is cost associated with each of the swapping process and it is calculated on the basis of criteria for k-medoids. For each swapping operation we calculate the swapping function and we add them all to find the overall swapping cost. (Kaufman and Rousseeuw, 2009).

Intelligent k-means algorithm use a principle according to that the farther a point is from the centroid the more interesting it becomes. In this algorithm we fallow the principle of principal component analysis and we search for the point that are too far from the centroid, because according to principal component analysis farther point corresponds to maximum data scatter. In this algorithm we form the anomalous pattern clusters that are formed from these points. (Buttrey, 2006).

K-means++ is a method to initialize the number of cluster k which is given as an input to the k-means algorithm. Since choosing the right value for k in prior is difficult, this algorithm provides a method to find the value for k before proceeding to cluster the data. (Arthur and Vassilvitskii, 2007).

In this paper author proposed a variant of K-means that is more efficient for multispectral image segmentation. A multispectral image is an image that captures specific wavelength. The author uses both spectral as well as the spatial property of the image. For this author uses image encryption Rauf et al. (2012).

Since the universities desire to improve their educational quality, the usage of data mining in higher education to help the universities, instructors, and students to improve their performance has become more and more attractive to both university managers and researchers. For example, to help improving the student performance, some research questions should be explored such as how the students learn? How quickly or slowly the students adapt with the new problems? Is it possible to infer the knowledge requirements to solve the problems directly from student performance data? (Thai-Nghe et al., 2010).

K-Means calculation experiences the numerous issues talked about in this segment we will address them. One of the primary impediments is the high time intricacy. In any distance based bunching approaches like k-implies, all information focuses or existing group are examined in a similar way independent of their distance to settle on a choice over bunching. This examining of all information focuses or the bunches doesn't have straight time adaptability and it falls flat for enormous datasets. The k-implies is innately sluggish on the grounds that k-implies bunching calculation takes O(nk) for single cycle. In this way calculation becomes difficult to be utilized for huge datasets as it would take a few cycles. Distance computations are one reason which makes the calculation slow. Deciding the quantity of groups ahead of time has been a test in k-implies. Expanding the worth of k diminishes mistake bringing about bunching. The blunder is the amount of the squared Euclidean good ways from information focuses to the bunch habitats of the allotments to which information focuses have a place. In outrageous case there is plausible of zero blunder in case grouping is performed by thinking about every information point as its own bunch. Generally worth of k is picked by for the most part suppositions, earlier information. The k-means calculation is successful in creating groups for some useful applications. In any case, the Computational intricacy of the first k-implies calculation is exceptionally high, particularly for huge informational collections. Besides, this calculation brings about various sorts of groups relying upon the arbitrary decision of introductory centroids. A few endeavors were made by specialists for working on the exhibition of the k-implies bunching calculation. The significant disadvantage of this calculation is that it produces various bunches for various arrangements of upsides of the underlying centroids. Nature of the last bunches intensely relies upon the determination of the underlying centroids. The k-implies calculation is computationally costly and requires time relative to the result of the quantity of information things, number of groups and the quantity of cycles.

III. Proposed Work

Our work is divided into two parts in first part we try to predict the performance of the student based on their marks in examination. We cluster student based on their marks. Find out the students who have less mark. Help these students so that we can achieve the better result in the near future. In the second part we are trying to reduce the time complexity as well as improve the performance at the same time. Data mining can be applied to educational databases to identify undesirable student behavior which was previously unknown. We can construct coursework, plan and schedule classes, and model students, predict their performance and provide recommendations for students, using data mining techniques. But originally k means algorithm suffer from many shortcoming like, it depends upon initial centroid, no of cluster etc.

In this part we will propose an improved adaptation of standard K-implies, this calculation functions admirably when size of the informational index is enormous. Our improved K-implies proposes to redress two prime downside of the standard K-implies. First is to pick starting centroid and the subsequent one is to decrease the time intricacy of a standard K-implies calculation. In the first place, we address the underlying centroid issue, in the event that we haphazardly pick the underlying centroid, this prompts the diverse intra bunch distance. To resolve this issue

one way is to run the calculation more than once and this prompts an alternate intra bunch distance, as each time we pick distinctive introductory centroid. From this diverse intra group distance we pick least one and comparing to that intra bunch distance select beginning centroid, however running the calculation number of time is exceptionally tedious. There are some different improvements, for example, take an example of information from the entire informational index and bunch them utilizing various leveled grouping and centroid acquired in this strategy can be utilized as starting centroid for our calculation. This arrangement additionally experiences the disadvantage that it functions admirably provided that the size of an example is little and various bunches are little in number.

To resolve the main issue we first check is information contain any bad worth in its trait assuming indeed, play out the standardization on the information for standardization find biggest negative worth from the informational index in that relating quality then, at that point, increase the value of all worth of that property. This makes entire informational index to positive. This standardization needed as we are working out the separation from the beginning. In the event that there is no regrettable worth in the relating property, no compelling reason to play out the standardization. In the subsequent stage, we sort the entire informational collection and separation the entire informational index into k equivalent parts. Here k is the quantity of required bunches. Then, at that point, find the worth at the center file of each set, these qualities are treated as the underlying centroid. In the second piece of the calculation, we need to decrease the time intricacy of the calculation. In essential K means calculation in the event that any information point goes from one bunch to other, we work out the distance of all information point from all the centroid. In case there are n number of information point and k is the quantity of groups and l is the no of cycle then fundamental K-implies calculation have a period intricacy of O (NKL). To lessen this we can utilize past advance outcomes, for this we track the closest bunch distance for every information point. In the following stage if the new centroid distance is not exactly or equivalent to the past distance then the point stay in a similar bunch, so there is no compelling reason to work out its separation from other group centroid. This will work on the grounds that in the K-implies bunch are round.

Algorithm: Enhanced K-means Algorithm

Inputs:

K-number of cluster

Data points

Output:

Cluster with data points

1: for i ←1, N do

2: find minimum of Oi

3: end for

4: if Oi < 0 then

5: for i ←1, N do

6: Oi = Oi - min

7: end for

8: end if

9: sort(Oi)

10: stepsize = N/K

11: for i←1, K, C = stepsize/2 do

12: Ci = Oc

13: C = C + stepsize

14: end for

15: repeat

16: for i←1, N do

17: for j ← 1, K do

18: rj ← dist(Cj , Oi)

19: end for

20: Parent[i] = min(row;K; i)

21: Object[i] = indexof(Parent[i]; row;K)

ISSN: 2581-3404 (Online)
International Journal of Innovative Research in Technology and Management, Vol-5, Issue-5, 2021.

22: add to Group

23: end for

24: for j =1← 1 , K do

25: OCj = Cj

26: Cj = mean(Cj)

27: end for

28: if !equal(OC,C,K) then

29: for i = 1 ← 1, K do

30: clear Gi

31: end for

32: end if

33: until !equal(OC,C,K)

The proposed thought originates from the way that the standard k-means calculation finds round formed group, whose middle is the gravity focus of focuses in that group, this inside moves as new indicates are included or, then again expelled from it Fahim et al. (2006). In this proposed algorithm we need to calculate distance for only those point that become father to centre if a point become closer then there is no need to calculate distance for these points. If points are no moving then use this distance function.

Algorithm: New Distance Function

1: for l = 1 ← 1, N do

2: for m = 1 ← 1, K do

3: rm = dist(Cm, Ol)

4: if rm < Parent[l] then

5: break

6: end if

7: end for

8: add to group

9: end for

IV. Result Analysis

Analysis of Enhanced K-Means Algorithm

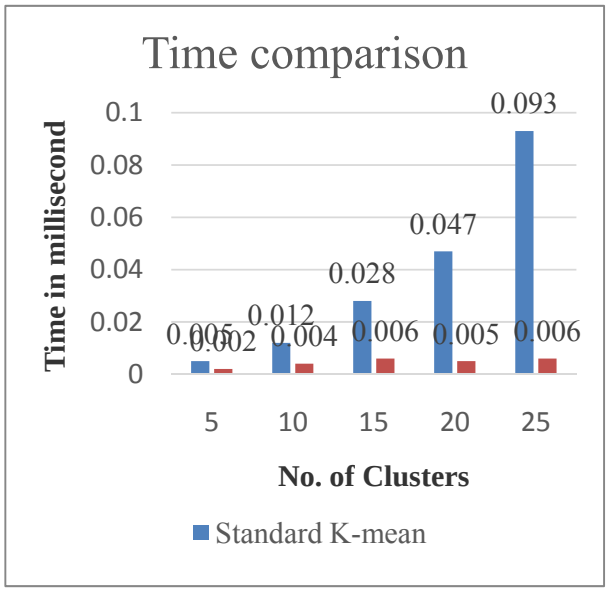

Figure (a): Time taken in millisecond

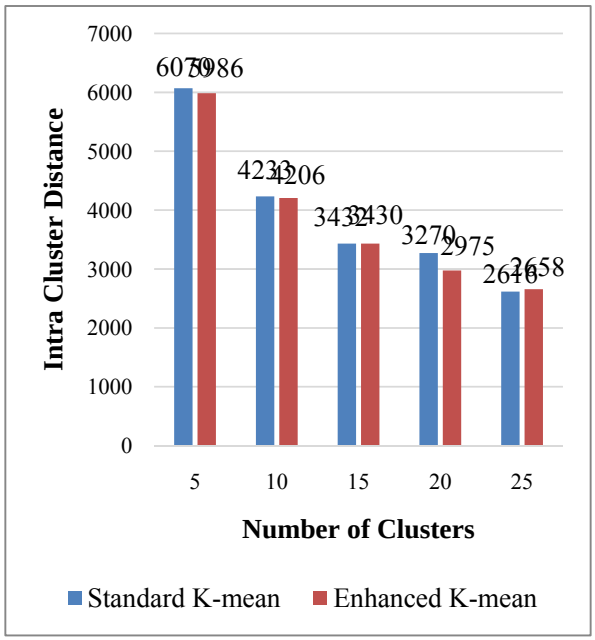

Figure (b): intra cluster distance

Figure 1: Relative Performance of Standard K-means and Enhanced K means

We use Census of India inhabited population data set for our algorithm performance evaluation. Our data set contain 6632 rows and each row contains one attribute. We run two algorithms standard K-means and Enhanced K-Means for the same data set with a different number of clusters. Numbers of clusters are 5, 10, 15, 20, 25 and calculate time taken to run the respective algorithm and what is their intra cluster distance. Our aim is to reduce the intra cluster distance as well as reduce the time taken to run the enhanced algorithm. Standard K-means algorithm has time complexity of O(nkl) Yedla et al. (2010), where n is a number of data points, k is the number of clusters and l is the number of iteration. Our proposed algorithm is of two phase and both phases are independent. In the first phase, we calculate the initial centroid for clusters. As in this process, we find min value that takes time O(n) and then sorts the data points and for sorting we are using heap sort. Heap sort has O(nlogn) complexity in best, average as well worst case and then partition them in k equal part this process will take time of O(k). So total complexity of this phase is O(nlogn) +O(k) + O(n), that can be written as O(nlogn). For second phase of over algorithm two cases are arrived in first, if point remain in the same cluster then it will take complexity of O(1) and if point do not remain in the same cluster then it will take O(k). So on average we can say that half point are moving from one cluster to other. So in average case complexity will be O(nk/2). As K-means converses to local minima so we can say that number of iteration also reduces and it is because no of points moves from one cluster to other reduce. We can consider no of iteration as ål i=1 1=i that is O(logl). So total complexity of second phase is O(nk log l). So total complexity of the complete algorithm is O(nk log l) + O(nlogn) that is approximately equal to O(nklogl). Standard K-means complexity is O(nkl). Our algorithm is computationally better than standard K-means. My experiment result are given in figure a and figure 1b.

Table 1: Comparative Result of Standard k-mean and Enhanced K-mean

S. No.	Criteria	Standard K-mean	Enhanced K-mean
1	Time (in ms)	0.093	0.006
2	Number of Iteration	69	3
3	Intra cluster Distance	2616	2658
4	Number of cluster	25	25

Table 1: shows that Enhanced K-means is better than Standard K-mean in terms of time and iterations.

Table 2: Predicted inhabited population

Population in each village	Number of villages	Results
27	42	42 villages contain less population
62	49	Average.
122	52	Average.
267	73	Highest number of villages which have around 267 inhabited population.
3246	11	Lesser number of villages has more population.

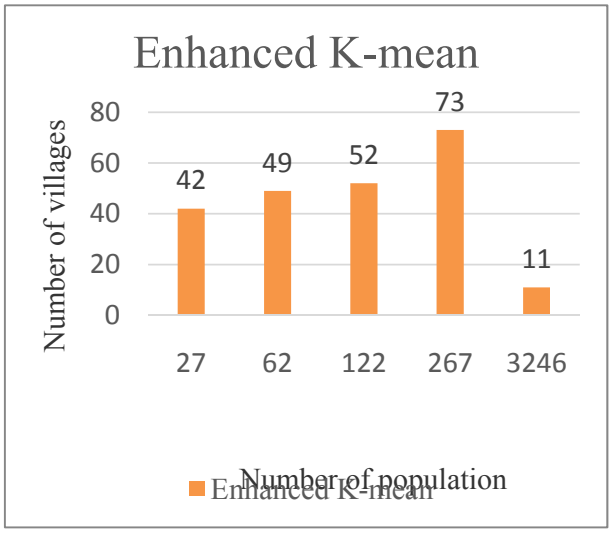

Figure 2: Predicted Inhabited population using Enhanced K-mean.

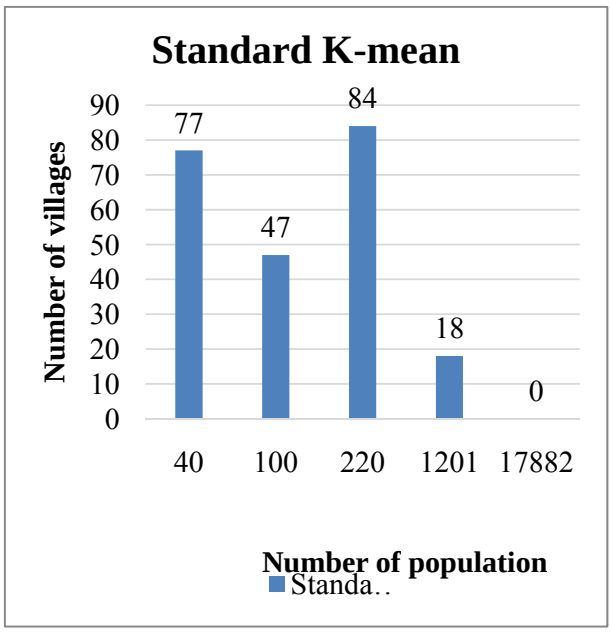

Figure 3: Predicted Population using Standard K-mean

In figure 3 when average population is 17882 then standard K-means return 0 number of village. It means that this is the situation of dead unit problem i.e. the cluster which has no data points. This situation is also the drawback of the standard k-mean which is overcome by proposed methodology i.e. enhanced K-mean.

V. Conclusion

Standard K-means is one of the most popular clustering algorithms but it suffers from some shortcoming. In this paper we address three of this shortcoming that are initialization of initial center problem, high time complexity and dead unit problem. Proposed algorithm reduces time complexity from O(nkl) to O(nklogl) and to get rid of random initial center initialization proposes an algorithm that initial center in the best way in O(nlogn). Here initialization algorithm does not have any constraints such as threshold value. Even after proposed enhancement, we require giving a number of a cluster as input to the algorithm we can automate the number of the cluster in future. Our Experiment consider only single attribute data set, we can extend this approach to multi-attribute data. While predicting inhabited population we are considering only one attribute. We can include more number of the attributes.

REFERENCES:

[1] Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications, volume 27. ACM.

[2] Ahmed, A. B. E. D. and Elaraby, I. S. (2014). Data mining: A prediction for student's performance using classification method. World Journal of Computer Application and Technology, 2(2):43–47.

[3] Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pages 1027–1035. Society for Industrial and Applied Mathematics.

[4] Bezdek, J. C. (1981). Cluster validity. In Pattern Recognition with Fuzzy Objective Function Algorithms, pages 95–154. Springer.

[5] Bidgoli, B. M., Kashy, D., Kortemeyer, G., and Punch, W. F. (2003). Predicting student performance: An application of data mining methods with the educational web-based system lon-capa. In Proceedings of ASEE/IEEE frontiers in education conference.

[6] Birch, Z. T. (1996). an efficient data clustering method for very large databases/t. zhang, r. ramakrishnan, m. livny. In Proceedings of the 1996 ACM SIGMOD international conference on Management of data (SIGMOD'96).-New York: ACM, pages 103–114.

[7] Buttrey, S. E. (2006). Clustering for data mining: A data recovery approach. Campos, M. M., Milenova, B. L., and McCracken, M. A. (2009). Enhanced k-means clustering. US Patent 7,590,642.

[8] Everitt, B., Landau, S., Leese, M., and Stahl, D. (1974). Cluster analysis. –john wiley & sons. Ltd., New York, page 330.

[9] Fahim, A., Salem, A., Torkey, F. A., and Ramadan, M. (2006). An efficient enhanced kmeans clustering algorithm. Journal of Zhejiang University-Science A, 7(10):1626– 1633.

[10] Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3):283–304.

[11] Jain, A. K. and Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall, Inc. Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3):264–323.

[12] Kaufman, L. and Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis, volume 344. John Wiley & Sons.

[13] Mitchell, T. M. (2006). The discipline of machine learning, volume 3. Carnegie Mellon University, School of Computer Science, Machine Learning Department.

[14] Nazeer, K. A. and Sebastian, M. (2009). Improving the accuracy and efficiency of the k-means clustering algorithm. In Proceedings of the World Congress on Engineering, volume 1, pages 1–3.

[15] Pelleg, D., Moore, A. W., et al. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In ICML, volume 1, pages 727–734.

[16] Rauf, A., Sheeba, S. M., Khusro, S., and Javed, H. (2012). Enhanced k-mean clustering algorithm to reduce number of iterations and time complexity. Middle-East Journal of Scientific Research, 12(7):959–963.

[17] Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. ii—recent progress. IBM Journal of research and development, 11(6):601–617.

[18] Tan, P.-N., Steinbach, M., and Kumar, V. (2013). Data mining cluster analysis: basic concepts and algorithms. Introduction to data mining.

[19] Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., and Schmidt-Thieme, L. (2010). Recommender system for predicting student performance. Procedia Computer Science, 1(2):2811–2819.

[20] Theiler, J. P. and Gisler, G. (1997). Contiguity-enhanced k-means clustering algorithm for unsupervised multispectral image segmentation. In Optical Science, Engineering and Instrumentation'97, pages 108–118. International Society for Optics and Photonics.

[21] Wagstaff, K., Cardie, C., Rogers, S., Schr¨odl, S., et al. (2001). Constrained k-means clustering with background knowledge. In ICML, volume 1, pages 577–584.

[22] Yedla, M., Pathakota, S. R., and Srinivasa, T. (2010). Enhancing k-means clustering algorithm with improved initial center. International Journal of computer science and information technologies, 1(2):121–125.

[23] Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: an efficient data clustering method for very large databases. In ACM Sigmod Record, volume 25, pages 103–114. ACM.