Journal of Advances and Scholarly Researches in Allied Education
Vol. 17, Issue No. 1, April-2020, ISSN 2230-7540

Study on Clustering Method Based on K-Means Algorithm

Yogeesh N.*

Assistant Professor of Mathematics, Government First Grade College, Tumkur, Karnataka, India

Abstract – In this paper we join the biggest least distance algorithm and the conventional K-Means algorithm to propose a further developed K-Means clustering algorithm. This further developed algorithm can make up the weaknesses for the conventional K-Means algorithm to decide the underlying point of convergence. The further developed K-Means algorithm adequately tackled two detriments of the conventional algorithm, the first is more noteworthy reliance to decision the underlying point of convergence, and another is not difficult to be caught in neighborhood minimum[1][2].

Keywords – Clustering, K-Means Algorithm

----------X----------

1. INTRODUCTION

Bunch investigation depends on different kinds of items' disparities and utilizations distance capacities' guidelines to make model order [3]. If the grouping is truly make a distinction is rest with the appropriation type of example character vectors. In the event that the commitments of specks of vectors is bunched and test dabs in a similar gathering are focused and test dabs in various gatherings are far off, it will be not difficult to utilize distance capacities to order the spots, which will quite far make insights in a similar gathering be comparable and measurements in various gathering be unique. The eigenvector of the entire example design assemblage can be treated as dabs which convey in highlight space. The distance work between specks might go about as the proportion of closeness of examples. As indicated by the nearness of dabs' distance, the action can be utilized to characterize designs. In this paper we join the biggest least distance algorithm and the conventional K-Means algorithm to propose a further developed K-Means clustering algorithm. This further developed algorithm can make up the weaknesses for the conventional K-Means algorithm to decide the underlying point of convergence. The further developed K-Means algorithm viably addressed the disservice that the conventional K-Means algorithm relies a lot upon the choice of introductory central focuses.

2. K-MEANS ALGORITHM

K-Means algorithm dependent on separating [4] [5] is a kind of bunch algorithm, and it is proposed by J.B.MacQueen. This algorithm which is solo is generally utilized in information mining and example acknowledgment. Targeting limiting bunch execution list, square-mistake and blunder measure are establishments of this algorithm. To seek the optimalizing result, this algorithm attempts to discover K divisions to fulfill a specific rule. First and foremost, pick a few specks to address the underlying group central points(usually, we pick the principal K example spots of pay to address the underlying bunch point of convergence); also, assemble the excess example dabs to their central focuses as per the standard of least distance, then, at that point, we will get the underlying order, and if the characterization if outlandish, we will adjust it(calculate each group central focuses once more), emphasize drearily till we get a sensible arrangement. K-Means algorithm dependent on partitioning is a kind of bunch algorithm, and enjoys benefits of quickness, effectiveness and celerity. Be that as it may, this algorithm relies very much upon starting specks and the distinction in picking beginning examples which consistently prompts various results. In addition, this algorithm dependent on track work consistently utilizes slope strategy to get extremum. The heading of search in slope strategy is consistently along the bearing where energy diminishes, which will prompts the way that when the underlying group point of convergence isn't legitimate, and afterward the entire algorithm will effortlessly sink into nearby least point.

3. RELATED CONCEPTION

Euclidean distance (short for distance)

Assume that X and Z are two examples of example vectors

$X = (x_1, x_2, \cdots, x_n)^T$ $Z = (z_1, z_2, \cdots, z_n)^T$ and we define the distance between X and Z as.

$$D = \|X - Z\| = [\sum_{i=1}^{n} (x_i - z_i)^2]^{\frac{1}{2}} \quad (1)$$

Simple to know that the more modest D is, the more comparative are X and Z (D is the distance of X and Z in ndimensional space)

Cluster Criterion Function

The sample pattern congregation is $\{X\} = \{X_1, X_2, \cdots, X_N\}$, and we classify it to C classes, they are $S_1, S_2, \cdots, S_c$. M_j and S_j are mean vectors. So:

$$M_j = \frac{1}{N_j} \sum_{X \in S_j} X, \quad N_j = |S_j| \quad (2)$$

Also, N j and S j are the quantity of tests. Then, at that point, we characterize group model capacity as:

$$J = \sum_{j=1}^{c} \sum_{X \in Sj} \|X - M_j\|^2 \quad (3)$$

J addresses the quadratic amount of mistake of a wide range of classes of tests and their mean worth. We can likewise consider it the amount of distances of tests and their mean worth. Along these lines, we should make an honest effort to get the base worth of [6].

Existing K-Means clustering Algorithm

K-Means clustering is a dividing clustering strategy which moves objects by moving starting with one bunch then onto the next beginning from an underlying parceling. The point of the group investigation is to segment n a perception into K bunches in which every perception has a place with the bunch with the closest mean. It is one of the easiest solo learning algorithms that take care of the notable clustering issue. The K-Means algorithm is a developmental algorithm that acquires its name from its strategy for activity. The algorithm bunches perceptions into K gatherings, where K is given as an info boundary. It then, at that point, relegates every perception to bunch dependent on the perception's nearness to the mean of the group. The bunch's mean is then recomputed and the cycle starts once more. Here's the way the algorithm works:

Step 1: The algorithm self-assertively chooses K focuses as the underlying group places ("means")

Step 2: Each point in the dataset is alloted to the shut group, in view of the Euclidean distance between each point and each bunch place.

Step 3: Each bunch place is recomputed as the normal of the focuses in that group.

Step 4: Steps 2 and 3 repeats until the clusters converge.

Intermingling might be characterized diversely relying on the execution, yet it typically means that either no perceptions change groups when stages 2 and 3 are rehashed or that the progressions don't make a material contrast in the meaning of the bunches.

K-Means is a notable strategy in unaided learning and vector quantization. The K-Means clustering is defined by limiting a conventional target work, meansquared-blunder contortion.

$$minimum\ MSE(P) = \sum_{i=1}^{N} \|x_i - C_{p(i)}\|^2$$

Where

N is the number of data samples;

K is the number of clusters;

d is the dimension of data vector;

$X = \{x_1, x_2, \ldots . x_N\}$ is a set of N data samples;

$P = \{p(i) | i = 1, \ldots N\}$ is class label of X;

$C = \{c_j | j = 1, \ldots k\}$ are k cluster centroids.

Because of its straightforwardness for execution, the ordinary K-Means can be applied to a given clustering algorithm as a post handling stage to work on the last arrangement. Notwithstanding, the principle challenge of the customary K-Means is that its arrangement execution profoundly depends on the chose starting allotment. All in all, with the majority of randomized beginning segments, the traditional K-Means algorithm unites to a locally ideal arrangement. A drawn out adaptation of K-Means, the K-Median clustering, serves an answer for defeat this restriction.

The K-Median algorithm look through each group centroid from information tests with the end goal that the centroid limits the summation of the good ways from all information focuses in the bunch to it. In any case, practically speaking, there are no effective arrangements known to a large portion of the figured K-Median issues that are NP-Hard. A further developed method is to form the K-Means clustering as a kernel machine in a

profoundly dimensional component space. The kernel machine tackles K-clustering issue in a profoundly dimensional Hilbert space rather than its feedback space. The enhancement of k-clustering issues in d dimensional space has ended up being NP-hard in K, be that as it may, for one-dimensional element space, a plan dependent on powerful programming can fill in as a device to drive an internationally insignificant arrangement. Subsequently, a heuristic methodology is utilized to assess the underlying part of K-Means clustering issues. Propelled by Wu's work on shading quantization, this can be tackled by powerful programming in the key part subspace. Specifically, a nonlinear bend can be chosen as the main course, for example a kernel head part.

4. IMPROVED K-MEANS ALGORITHM

To determine the initial cluster focal point

It can't be more fundamental to depict the determination of the underlying group point of convergence, in any case, normally; the choice of point of convergence is very stochastic, which prompts the way that the result of bunch result is likewise very stochastic. While in reasonable applications, we not just need the underlying central focuses to be decentralized yet additionally need them to be more delegate. The biggest least distance algorithm depends on test in the field of example acknowledgment. Its fundamental idea is to pick the example wherein pairwise distances are farther separated however much as could reasonably be expected to be bunch point of convergence. Accordingly, we cannot just decide the best starting group point of convergence mentally yet in addition increment the productivity of isolating beginning information gathering. Additionally, it has no likelihood that the underlying group central focuses will be excessively nearby, which might occur while utilizing K-Means algorithm. In this paper, at the primary spot, we utilize the biggest least distance algorithm to decide K introductory bunch central focuses, and afterward we join it with the conventional K-Means algorithm, finally, achieve the grouping of example assembly.

The further developed K-Means algorithm is clearly better compared to conventional one in perspectives, for example, the accuracy of bunch, the speed of group, strength, etc. In this paper, we take on Euclidean distance as the basis, on the grounds that the computation of Euclidean distance in both hyperspace and two-dimensional space are comparative. Thus, we will utilize two-dimensional space as an illustration to break down the further developed K-Means algorithm in this paper.

Algorithm Description

Given N samples of pattern $\{x_1, x_2, \cdots, x_N\}$ which are waiting for classifying, they are. They need to be classified to K clusters.

1. Choose any one among $\{x_1, x_2, \cdots, x_N\}$ to act as the role of first cluster focal point 1 z , for example, we choose $z_1 = x_1$

2. Choose another point which is as much as possible far apart to z_1 to be the focal point of the second cluster and calculate the distance between each sample and z_1 :

$$\|x_i - z_1\|, \quad i = 1,2,\cdots,N$$

If:

$$\|x_j - z_1\| = \max\{\|x_i - z_1\|, \quad i = 1,2,\cdots N\}, \quad j = 1,2,\cdots N \tag{4}$$

Then choose x_j to be the focal point of the second cluster, and $z_2 = x_j$

3. Calculate the distance between each sample among $\{x_1, x_2, \cdots, x_N\}$ and $\{z_1, z_2\}$ one by one.

$$d_{i1} = \|x_i - z_1\|, \quad i = 1,2,\cdots N \tag{5}$$

$$d_{i2} = \|x_i - z_2\|, \quad i = 1,2,\cdots N \tag{6}$$

Choose the minimum of the outcomes:

$$\min(d_{i1}, d_{i2}), \quad i = 1,2,\cdots N$$

Gather the minimums of all samples of pattern and $\{z_1, z_2\}$. Choose the maximum among the minimums to be the third cluster focal point z_3

If:

$$\min(d_{j1}, d_{j2}) =$$

$$\max\{\min(d_{i1}, d_{i2}), \quad i = 1,2,\cdots N\}, \quad j = 1,2,\cdots N \tag{7}$$

Then:

$$z_3 = x_j. \tag{8}$$

4. Suppose that we have got r $(r < k)$ cluster focal points $\{z_i, \quad i = 1,2,\cdots r\}$, now we need to determine the r+1th cluster focal point, namely if:

$$\min(d_{j1}, d_{j2}, \cdots, d_{jr}) = \max\{\min(d_{i1}, d_{i2}, \cdots, d_{ir}), \quad i = 1,2,\cdots N\} \quad j = 1,2,\cdots N$$

Then

$$z_{r+1} = x_j .$$

5. Repeat, till r+1=k

6. Now we have chosen K initial cluster focal point $\{z_1(1), z_2(1), \cdots, z_k(1)\}$. The numbers in parenthesis are serial numbers used in iterative operations to seek cluster points.

7. According to the rule of minimizing distance, allocate $\{x_1, x_2, \cdots, x_N\}$ to one of the K clusters, namely, if:

$$\|x - z_j(t)\| = \min\{\|x - z_i(t)\|, \ i = 1,2,\cdots,K\}, \ j = 1,2,\cdots,K \qquad (9)$$

Then

$$x \in s_j(t) .$$

The image t in the recipe is the chronic number of iterative activities, j s represents the j th bunch, and the group point of convergence is z_j.

8. Calculate the new vector values of each cluster focal point:

$$z_j(t+1) \ , \ j = 1,2,\cdots,K .$$

Calculate the mean vectors of samples of each cluster:

$$z_j(t+1) = \frac{1}{N_j} \sum_{x \in s_j(t)} x , \quad j = 1,2,\cdots,K \qquad (10)$$

The symbol N_j in the recipe above represents the quantity of tests of the j th bunch s Calculate the mean vectors of tests of the K groups separately. Making mean vectors be new groups can limit bunch rule work J_j.

$$J_j = \sum_{x \in s_j(t)} \|x - z_j(t+1)\|^2 , \quad j = 1,2,\cdots,K \qquad (11)$$

9. If: $z_j(t+1) \neq z_j(t)$, $j = 1,2,\cdots,K$, then turn back to 7v, classify samples of pattern one by one again, and repeat iterative operations. If $z_j(t+1) = z_j(t)$, $j = 1,2,\cdots,K$, then the convergence of the algorithm is finished

5. THE ANALYSIS OF EXPERIMENT

Targeting testing the effectiveness of the further developed K-Means algorithm, we use emulation information assembly introduced in table 1. The information gathering is made out of 20 arbitrary information and is grouped to five classes as indicated by the level of bunch. We can see that the contrasts between each class are very self-evident. The trial takes benefits of Visual C++ 6.0 advancement climate [7].

Table 1

pattern	abscissa	ordinate	pattern	abscissa	ordinate
X1	1	1	X11	1.69	0.93
X2	1.5	1.5	X12	0.3	1.1
X3	1.5	1.1	X13	7	7.4
X4	81	80	X14	6.9	6.9
X5	7.3	8	X15	22.2	20.5
X6	35.7	33.4	X16	23	21
X7	8	7.3	X17	80.6	73.2
X8	21.2	20	X18	36.7	38.55
X9	81	73	X19	34.76	33.6
X10	6.9	7.6	X20	81	73.6

Table 2

	Standard K-Means	Improved K-Means
iterations	9	6
Cluster Criterion Function J	657.603	58.3263
First class	X1,X12	X1,X2,X3,X11,X12
Second class	X5,X7,X10, X13,X14	X5,X7,X10,X13, X14
Third class	X2,X3,X11	X8,X15,X16
Forth class	X4,X9,X17, X20	X4,X9,X17,X20
Fifth class	X6,X8,X15, X16, X18,X19	X6,X18,X19

The results of two kinds of algorithms are depicted by Table 2. We might arrive at the resolution that the result of standard K-Means isn't unreasonably acceptable, cause its underlying bunch central focuses are excessively arbitrary, which will cause temperamental group result. While the further developed K-Means gives somewhat wonderful result.

CONCLUSION

As indicated by the scholastic examination and aftereffect of test over, the further developed K-Means not just keeps the high productivity of standard K-Means yet additionally raises the speed of assembly viably by working on the method of choosing starting group point of convergence. The further developed K-Means is clearly better compared to standard K-Means in both group accuracy and solidness. Particularly, the benefits will be more clear when gotten to the heart of the matter of bunch issues which have huge scope and totally arbitrary appropriated information.

REFERENCES

[1] Usama M. Fayyad Cory A. Reina Paul S. Bradley (1998). Initialization of Iterative Refinement clustering algorithms[C].Proc.4th International Conf. On Knowledge Discovery & Data Mining.

[2] DUDA R O, HART P E. (1973). Pattern classification and scene analysis[M].New York:John Wiley & Sons,1973.

[3] Bian ZhaoQi and Zhang Xuegong (2000). Pattern Recognition Beijing Tsinghua University Press.

[4] JAIN A K, DUBES R C. (1988). Algorithms for clustering data[M].New Jersey:Prentice-Hall.

[5] Zhang Yufang (2003) etc. A kind of improved K-means algorithm [J]. Computer Application,p31?33, (8).

[6] SELIM S Z, ISMAIL M A.K-means type algorithms: a generalized convergence theorem and characterization of local optimality [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, p81?87, 1984, 6(1).

[7] Sun Xin and Yu Anping VC++ in depth detailed introduction [M] Beijing: Electronic Industry Press, 2006.

[8] Ravindra, R.; Rathod, R.D.G. (2017). Design of electricity tariff plans using gap statistic for K-means clustering based on consumers monthly electricity consumption data. Int. J. Energ. Sect. Manag., 2, pp. 295–310.

[9] Han, L.; Wang, Q.; Jiang, Z.; Hao, Z. (2010). Improved K-means initial clustering center selection algorithm. Comput. Eng. Appl., 46, pp. 150–152.

[10] UCI. UCI Machine learning repository. Available online: http://archive.ics.uci.edu/ml/ (accessed on 30 March 2019).

[11] Tibshirani, R.; Walther, G.; Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. J. R. Statist. Soc. Ser. B (Statist. Methodol.), 63, pp. 411–423.

[12] Xiao, Y.; Yu, J. (2007). Gap statistic and K-means algorithm. J. Comput. Res. Dev., 44, pp. 176–180.

[13] Kaufmn, I.; Rousseeuw, P.J. (1990). Finding Groups in Data an Introduction to Cluster Analysis; New York John Wiley & Sons: Hoboken, NY, USA.

Corresponding Author

Yogeesh N.*

Assistant Professor of Mathematics, Government First Grade College, Tumkur, Karnataka, India

yogeesh.r@gmail.com