IMPROVEMENT OF INITIAL CENTROIDS IN K MEANS CLUSTERING ALGORITHM

Avni Godara[1], Varun Sharma[2]
[1] Research scholar, Computer Science Engineering Department, Amity University, Rajasthan, India
[2] Assistant Professor, Computer Science Engineering Department, Amity University, Rajasthan, India

ABSTRACT

Clustering is an approach of data mining which classifies data according to correspondences and alterations and finds result frequently. Clustering has a method with name partition based clustering which tries to find out similarities and differences between data points with help of calculating distances between data points and relocate them again and again until no relocation is needed , and find accurate results. K means is also a part of partition based clustering, which also use distance measure to find similarities and differences between neighbourhood data points with Euclidean distance measure. The k means clustering algorithm generally consider as iteration based algorithm which used in daily life. The k means clustering algorithm works with spherical type clusters. In k means clustering randomly choose centroids, apply Euclidean distance measure, and form cluster of data points with minimum distance in respect to centroid. But the k means clustering has encountered a dilemma in preliminary centroid selection for the specified dataset. K means is insightful for preliminary centroids because result can vary according these preliminary centroids. The prim's algorithm makes available an enhancement in k means of giving packed in the dataset and truthful data points as preliminary centroids. This enhanced k means afford better experimental results and better accuracy then classical k means. This enhanced k means provides less iterations comparatively with classical k means clustering. The prim's algorithm tries to compact a cluster and gives centroids that has maximum probability of become center of that compact cluster and provide less re-allocation of centroids.

Keyword : - *Clustering, k means, Prim's algorithm, centroids.*

---

1. INTRODUCTION

In data mining, there are so many algorithms for find some non petty patterns in different types of databases, clustering is one of them. Clustering is a compilation of data points or objects that can be categorized by comparability and contradictories. Clustering is used on a daily basis to classify data objects like a case in point a male object belongs to a human class or cluster and a tree object belongs to plant class or cluster of some concepts. So, clustering is process define data points from a given dataset that data points belong to which class or cluster according to comparability and contradictories between them. Clustering improves a knowledge Discovery process again and again until finding results based on the behavior and impression of data points in giving datasets so this is why called unsupervised learning. Clustering brings into play a data matrix and a dissimilarity matrix for categorize data points in a cluster. Clustering uses special types of data sets like spatial datasets, related to biology datasets and there can be any type of dataset for finding non petty results but with unusual sizes and different data variables. Clustering works with binary variables, nominal variables and ordinal types of variables. There is a partition clustering one of the most bendable types of clustering. In partition based clustering, there are n data points in the dataset and segregated into k groups. The number of groups can be slighter and equivalent to data points.

According to given k clusters need to split data set into k sections from this, it will start preliminary partitioning and again repeat this procedure to categorize data points belong to which group or cluster. The end result can get if last two iterations have unchanged group matrix. So, Partitioning-based clustering methods are based on recurring repositioning of data points between clusters. Here cluster is a group who has comparable types of data points and the middle point of the cluster is called the centroid.

During iteration, data points have in minimum distance with respect to the centroid in the cluster are replaced. So, partitioning based clustering is a center based clustering. Partitioning based clustering follows rules first is each

cluster keeps in check, at least one data point, and second each data point is in the right place to exactly one cluster, but in a fuzzy partitioning, a point can belong to more than one group. The partitioned based clustering alienated in two most popular methods are k means partitioning and k medoids partitioning. The k means algorithm provides a high performance with isolated and efficient clusters. The k means algorithm one of the most utilizable algorithms in clustering based on arbitrarily selects k clusters and sum of squared error (SSQ).
The next section will be going to discuss about classical k means algorithm.

2. THE CLASSICAL K MEANS ALGORITHM

2.1 About classical k means

The initiative of k means gave by Hugo Steinhaus in 1957 but firstly used by James MacQueen in 1967.The conventional k means algorithm launched by Stuart Lloyd in 1957 when he was conducting experiments on pulse code modulation. K means is a very straightforward and momentary algorithm consider as error minimization algorithm in clustering. K means algorithm point out fix k clusters which have intra cluster comparability very sky-scraping and comparability with another cluster is very stumpy.
K means a very trouble-free way to find out prototypes which works with slighter finite datasets. After fixing of number of k cluster indiscriminately choose centroid for each cluster. In k means necessitate to presuppose all centroid's location very watchfully for example there are two poles apart clusters have centroids with minimum distance between them that can be have an effect on the consequences by getting very unusual consequence from the original ones . So a high-quality choice is that choose centroid's location very carefully with far away distance to another centroid's location. After this k means attempts to find out each data-point be in the right place to which cluster by finding distance from centroid with Euclidean distance.
After computing this bare minimum distance with respect to centroid then preliminary partitioning is done. From this minimum distance values are positioned into distance matrix and according distance matrix formulate a group matrix. In k means algorithm key point is that it re-calculates new centroids according to resemblances and resemblances compute by mean value of data points in cluster, reiterate all over process again until there is no alterations happen in group matrix. When there is no alteration means any movement in centroids are not obligatory. Typically, the square error criteria defined as

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - m_i|^2$$

Fig -2: Classical k means square error criteria.

Here, E is total square error of data points, x is given data point and mi is mean value of cluster $c_{i..}$There are k clusters with dataset be full of n data points. Now k means algorithm put into operation as:

(1) Randomly choose k data points as preliminary cluster centroids
(2) repeat
  (a))(re)assign each data point to cluster to which the data point is the majority comparable, according the mean value of data points in cluster.
  (b) Keep posted the cluster means and determines mean value of data points for each cluster.
(3) until there is no alteration.

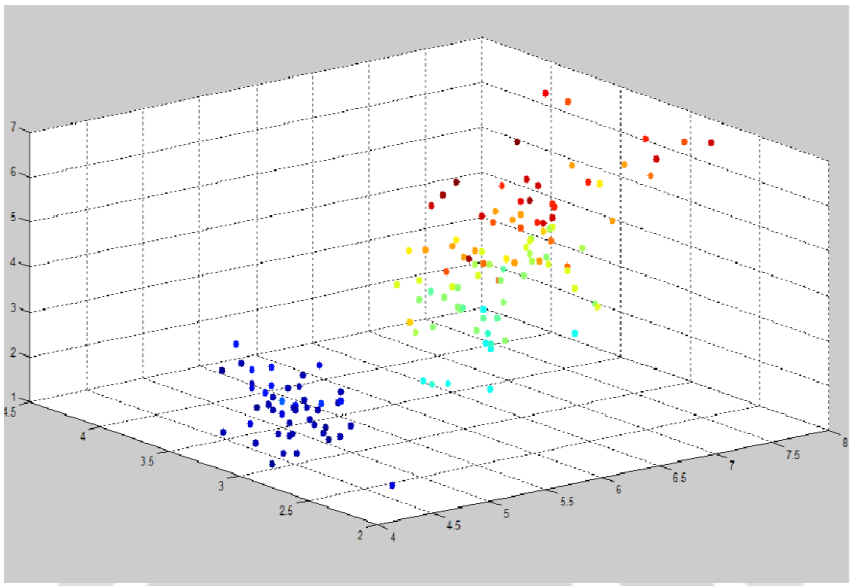

Fig -2: Name of the figure Classical k means for three randomly selected centroids

K means gives computationally more rapidly outcomes with compact clusters. It commences with k clusters over and over again updates it, so from this k means able to diminish error function. But there is a predicament with k means that it is thin-skinned with the preliminary points and outliers. Next section going to discuss enhancement of k means in this paper.

3. PREVIOUS WORK

As everybody is familiar with, that in 1957 Lloyd [1] provided k means problem, used as classical clustering. K means sat in motion with a number of data points are given and necessitate choosing k points erratically among them. At each step, it goes for data points in to cluster and then compute midpoint of each cluster. This process keeps revising itself until it is constant. But this algorithm has dilemma with preliminary selection and capriciously, bad illustration survives like worst case running time is exponential. K means having so much difficulty in it, after all this is used frequently in practice.

Ostrovsky, Rabani, Schulman and Swamy [2] planned in k means that data points should be stretched clusterable and the value of k is opted for according to that cluster instead of randomly. They distinct notion of σ separability, where the key in to k-means is supposed to be σ -separable if dipping the number of services from k to k - 1 would enlarge the rate of the best possible solution by a factor 1/ σ . They considered an algorithm with approximation ratio $1 + O((\sigma^2)$.

Tian, Jinlan, Lin Zhu, Suqin Zhang, and Lu Liu [3] gave method that offer very good result with healthier effects and with a reduction of repetitive time than classical k means algorithm and attempt to construct cluster more well-organized by use of parallel k means. This parallel k means provides complexity something like 1/m than the original k means and according to this it boost cluster analysis efficiency with time and space complexity. But result for diminutive dataset is not especially prominent.

Nazeer, KA Abdul, and M. P. Sebastian [4] presented a better version of k means ,They provide the data points and value of k necessary for work out initial centroids in cluster automatically. They applied algorithm that make sure whole method of clustering has complexity of $O(n^2)$ time and also preserve accurateness of clusters. The restriction of this projected algorithm is value of k and when size of dataset amplifies the rate up performs in better condition.

Pakhira, Malay K [5] aim to work with vacant clusters formed in k means when clusters hang on initial centroids. The problem arises in the k-means algorithm for stationary execution, this need a number of time execution for resolve the trouble but this is considered insignificant. After execute so many k means need to handle inappropriate data or outliers or noisy data: So here k means need smoothening and pre processing on the dataset to filling misplaced value, get rid of noisy data, data cleaning and selection before it gets start.

Baolin Yi, Haiquan Qiao, Fan Yang [7]for find initial center point in algorithm density based approach used with Gaussian function preserve global steadiness of clustering. In this attempt to go for most density point as primary point from the given dataset and once more use for find after that initial center in cluster, cross out the primary data

point and its adjoining data point and reiterate this method awaiting k points are there in set M. This provides good consequence than classical k means with boundary data points, density based occurrence values are appropriately defined.

Lan Huang, Shixian Du, Yu Zhang, Yaolong Ju, Zhuo Li [8] try to find out initial midpoint in dataset with kruskal's algorithm by separating the data points in k clusters. In this employ undirected graph with weighted edges by kruskal's algorithm change this into k connected sub graph , which include iterating times of algorithm and get k cluster midpoint , objective function. The pre-eminence of this algorithm, better k means (since use the alike Euclidean distance and similarity roles), is to endow with estimate of the primary centroid using a minimum spanning tree algorithm.

4. PROPOSED APPROACH

K means need a enhancement with preliminary selection of centroids in a cluster. In minimum spanning tree, two data-points or nodes connected with an edge. So at least one data point with another data point shows relationship has locality possessions. Hence, first step of improvement is selection of highly significant data points present in a cluster. According to selection we can envision how a cluster is come across. Next step is like classical k means get rid of the centroids in that cluster and provides new centroids with minimum Euclidean distance function; iterate rest of the data points to see any updates occurred in centroids.

The given input to the improved algorithm is a graph G (V, E) where E is set of edges and V is the set of vertices. Vertices demonstrate data points in graph and edges show relationship between them and the specified graph is undirected graph, so in this maximum number of clusters can be generates. Now algorithm is follows as:

1) So first step is take complete dataset and consider that dataset as a clusters.
2) Determine cost of edge added is required or not because it shows minimum distance with all neighboring data points. Minimum spanning tree algorithm applied on that cluster.
3) Now compare the cost attached of data points and get a reduced cluster.
4) Now in cluster and find degree of vertices .A node that has high degree consider as centroid or if two and more nodes have same degree than choose randomly centroid one of them.
5) Now, without going over the data points, and check the similarity and distance function from each centroid in the clusters to all its data points.

6) Each data point that is chosen has option (with either the distance or similarity functions):

(a)to substitute the existing centroid and, will make it both the similarity and distance function.

Algorithm will going to repeat itself again and again for find potential centroid until data points consider as boundary data points in that cluster.

If there is no change in distance matrix means no change in centroids then we got result.

4.1Pseudo code:

Input : Graph G(V,E) and constant k, root vertex r
Output: k cluster centers
Step 1: For each v in V[G]
Key [v] <- ∞
∏ [v] <- NULL
Step 2: key [r] = 0
Step 3: Q <- V [G]
Step 4: v <- min (Q)
Step 5: if u ϵ Q and w (v, u) < key [u]
Then : ∏ [u] = v and key [u] <- w (v, u)
Step 6: Repeat step 4 and 5 till Q is empty
Step 7: find Degree of each node.
centroidArr[][]=degree of node
Step 8: find highest degree k vertices.
For each k
Do
resultArray[]=centroidArr[][]
Step 9: End.

5. EXPERIMENTS RESULT

Here reflect on k means data points as undirected graph nodes and reduced using prim's algorithm .our ambition is compact dataset and find out easily preliminary centroids for that compact datas et. We decide on (Iris, diabetes, car) dataset from UCI repository[1].On the whole k means have complexity of O(n*k*j) where k symbolizes number of group or cluster ,n symbolizes number of data points given as input and j symbolizes number of iteration. Our functional algorithm has complexity O(E log V)+ O(n*k*j)and improved results than classical k means clustering.

Table -1: Results of proposed approach

Dataset	Number of cluster	Algorithm	Average time (ms)
Iris	3	Standard k means	0.61734
		Proposed k means	0.59347
Diabetes	3	Standard k means	0.09381
		Proposed k means	0.07653
Car	3	Standard k means	0.10554
		Proposed k means	0.08234

6. CONCLUSIONS

In daily life data used and increased and use clustering to find better results according to that data. In clustering k means is a powerful and growing algorithm used most of the application but problem of initial centroid is available. In 30 to 35 years so many paper presented to improve classical k means algorithm. To remove this problem and define data points for centroid before next iteration. The use of prim's algorithm gives better results for selection of initial centroid and choose easily data points for future iterations. Our experimental result also shows better and optimal performance for initial centroids, accuracy of result not adjusted.

7. REFERENCES

[1].Stuart Lloyd: Least Squares Quantization in PCM. Special issue on quantization, IEEE Transactions on Information Theory, (1982)

[2].Rafail Ostrovsky, Yuval Rabani, Leonard Schulman, and Chaitanya Swamy.:The Effectiveness of Lloyd-Type Methods for the k-Means Problem. FOCS, (2006)

[3].Tian, Jinlan, Lin Zhu, Suqin Zhang, and Lu Liu.:Improvement and parallelism of k-means clustering algorithm. Tsinghua Science & Technology 10, no. 3, 277-281 (2005)

[4]. Nazeer, KA Abdul, and M. P. Sebastian. :Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. Proceedings of the World Congress on Engineering, vol. 1, pp. 1-3. (2009)

Pakhira, Malay K. :A modified k-means algorithm to avoid empty clusters. International Journal of Recent Trends in Engineering 1, no. 1 (2009)

[5]. Yugal Kumar, and G. Sahoo. :A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm. International Journal of Advanced Science and Technology 62: 43-54 (2014)
[6].Baolin Yi, Haiquan Qiao, Fan Yang, Chenwei Xu.: An Improved Initialization Center Algorithm for Kmeans Clustering. IEEE (2010)
[7].Lan Huang, Shixian Du, Yu Zhang, Yaolong Ju, Zhuo Li.:K-means Initial Clustering Center Optimal Algorithm Based on Kruskal. Journal of Information & Computational Science 9: 9 (2012) 2387-2392.
[8]. Muhammad Husnain Zafar and Muhammad Ilyas . :A Clustering Based Study of Classification Algorithms. International Journal of Database Theory and Application Vol.8, No.1 (2015), pp.11-22
[9].Merz C. and Murphy P.. UCI Repository of Machine Learning Databases. Available Online from ftp://ftp.ics.uci.edu/pub/machine-leaming-databases, (1999)
[10]. Fang Yuan, Zeng-Hui Meng , Hong-Xia Zhangzs, Chun-Ru Dong :A New Algorithm To Get The Initial Centroids. Proceedings of the Third International Conference on Machine Laming and Cybernetics, Shanghai, 26-29 August (2004)
[11]. Md. Sohrab Mahmud, Md. Mostafizer Rahman, and Md. Nasim Aktar: Improvement of K-means Clustering algorithm with better initial centroids based on weighted average 7th International Conference on Electrical and Computer Engineering 20-22 December, (2012)