ISSN:0975-9646

A Novel Clustering Algorithm Using K Harmonic Means and Improved Time Complexity

Rangasubramanian K
Student, Department of Electronics and Communication,
Sri Sivasubramaniya Nadar College of Engineering,
Kalavakkam, Tamil Nadu, India

Abstract **- Clustering algorithm is one of the most popular unsupervised learning algorithms in machine learning. K means clustering is one of the widely used clustering methods for various applications in data mining, image processing and computer vision. Many solutions have been offered to make the k-means clustering algorithm more efficient. This paper proposes an improved k-means clustering algorithm by initializing cluster seeds and improving the time complexity of the algorithm.**

Keywords **- K-means clustering, unsupervised learning.**

I. INTRODUCTION

Data clustering is identified as a method in which a set of data is grouped such that the data in the same group show a higher similarity in certain properties compared to the data in other groups [1]. Data clustering finds use in many applications such as data mining, pattern recognition and image segmentation [2]. Clustering can be broadly classified into hierarchical and partitional methods [3].

K means clustering is a numerical, unsupervised, non-deterministic and iterative method [4]. It partitions a given set of data into k clusters [1]. Though the k means clustering algorithm is efficient, it may converge to a local minima due to randomness in its initialization and thus the results would become counterintuitive [1]. Moreover, it's efficiency is affected by cluster sizes [1].

Therefore, it would be more efficient if initial cluster centers are fixed. Naturally, the number of clusters should be given a priori [5]. There have been many attempts to fix initial cluster centers [1, 8, 10].

Efficiency of the algorithm is also measured by its time complexity [2, 4, 5, 9]. There have also been attempts to improve the efficiency of k means by reducing computational complexity [2-4, 6-9].

In this paper, we propose an improvement in the k means clustering algorithm by using efficient methods for initialization and reducing the computational complexity of the algorithm.

II. LITERATURE REVIEW

There have been many attempts in literature to improve the efficiency of the k means algorithm. Initial cluster center selection has been attempted by using the leader method, where cluster radius and distance between two clusters are used to obtain the number of clusters [3]. Initialization has also been attempted by using K harmonic means, after which subtractive clustering has been used to enhance efficiency [8]. Density optimization and distance optimization have also been used to initialize cluster centers [10]. Constraints have been included to find cluster size and assign initial cluster centers [1].

Computational complexity of k means has been reduced by ignoring the data points whose affiliations to cluster centers are known and the respective distances from the cluster centers either remain the same or are lesser than the one obtained from the previous iteration [4].The same has been attempted using a global k means algorithm [2]. K mean and k medoid algorithm have been used for initialization, after which clustering is performed based on the cluster-error criterion, where two nearest clusters are combined and the iteration is continued [5]. There has also been an attempt to improve efficiency of the k means algorithm by using top-n merging, cluster pruning and optimized updates [9].

III. PROPOSED ALGORITHM

The algorithm proposed in this paper attempts to make the clustering algorithm more efficient by initializing the initial cluster centers and reducing computational complexity. This is achieved by using K harmonic means for initialization and using an efficient method for allocation of data points to cluster centers.

Initialization by K Harmonic Means:

Consider the whole dataset $X = \{x_1, x_2, \dots, x_n\}$ which has n unlabeled examples in the d-dimensional space $R^{d\cdot}$ The following algorithm divides X into m clusters [8]:

1. Obtain m initial cluster centers c_j for KHM algorithm, where $j = 1,2,\dots..m$ and let $\widehat{H} = 0$.
2. According to the following function H(x), calculate H. $\propto$ is a parameter such that $\propto \geq 2$.

$$H(x) = \sum_{i=1}^{n} \frac{m}{\sum_{j=1}^{m} \frac{1}{\|x_i - c_j\|^{\alpha}}}$$

3. Based on the following equation, get each element t_{ij} of the matrix T, where $i = 1,2,\dots,n$ and $j = 1,2,\dots,m$.

$$t_{ij} = \frac{\|x_i - c_j\|^{-\alpha-2}}{\sum_{j=2}^{m} \|x_i - c_j\|^{-\alpha-2}}$$

4. Obtain the weight W_i of each data x_i according to the following equation.

$$W_i = \frac{\sum_{j=1}^{m} \left\| x_i - c_j \right\|^{-\alpha-2}}{\left(\sum_{j=1}^{m} \left\| x_i - c_j \right\|^{-\alpha} \right)^2}$$

5. Update each cluster center c_j using the following equation.

$$c_j = \frac{\sum_{i=1}^{n} t_{ij} W_i x_i}{\sum_{i=1}^{n} t_{ij} W_i}$$

6. If $\left| \widehat{H} - H \right| > ¢$, then let $\widehat{H} = H$and go to 2, else go to 7.
7. For each data point x_i, assign it to cluster q by the following equation and end KHM.

$$q = \arg\max_{j=1,2,\cdots,m} \left(t_{ij} \right)$$

8. Let the parameter $\lambda \cdot m$ be a positive integer.
9. After applying KHM, $\lambda \cdot m$ cluster centers are obtained. Add these $\lambda \cdot m$ cluster centers into the data set X . Thus the new data set X^* is formed. Now the number of elements in the set X^* is $n + \lambda \cdot m$.

K means clustering with reduced computational complexity:

The standard k means calculates the distance of each data point in every iterative step of the algorithm. This takes up a lot of execution time. This can be reduced by storing the distances of data points from their cluster centers so that they can be used for successive iterations. Thus, if the distance of a data point, in a particular iteration, is lesser than or equal to the one from its previous cluster center, the data point stays in its cluster and consequently, there is no need to calculate its distances from the other $k - 1$ clusters [4].

We have our dataset X^* obtained by initialization.

1. Calculate the distance between each data object x_i and all cluster centers c_j as Euclidean distance $d(x_i, c_j)$ and assign data object x_i to the nearest cluster with cluster center c_j.
2. Store the label of the cluster center c_j and the distance of the data point x_i from c_j, where c_j is the cluster center of the cluster to which x_i has been assigned, in two arrays *Clust[]* and *Dist[]* respectively.
   Set *Clust[i]* = j
   Set *Dist[i]* = $d(x_i, c_j)$, which is the Euclidean distance between x_i and c_j.
3. For each cluster j, recalculate its cluster center.
4. For each data point x_i, compute its distance to the center of the present nearest cluster. If this distance is lesser than or equal to *Dist[i]*, then the data object stays in its cluster. If not, then calculate the distance between each data object x_i and all k cluster centers c_j as Euclidean distance $d(x_i, c_j)$ and assign data object x_i to the nearest cluster with cluster center c_j.
   Set *Clust[i]* = j
   Set *Dist[i]* = $d(x_i, c_j)$, which is the Euclidean distance between x_i and c_j.
5. Repeat steps 3 and 4 until convergence criterion is met.

IV. DISCUSSIONS

The aim of this paper is to propose an efficient k means clustering algorithm. Since initialization and computational complexity of the algorithm affect its efficiency, this paper proposes a k means clustering algorithm where initialization improves the performance of the actual clustering procedure whose computational complexity has also been reduced.

K harmonic means is insensitive to initial cluster seeds. Therefore, the algorithm employs k harmonic means to arrive at cluster centers. However, since the efficiency of the k means depends on initialization, these cluster centers are used as initial seeds for the clustering algorithm.

The time complexity of the traditional k means is given by $O(nkt)$, where n is the number of data points, k is the number of clusters and t is the number of iterations. Incidentally, t is also the number of data points that move from their previous clusters. However, the proposed algorithm has its time complexity given by $O(nk)$, thus improving the speed and efficiency of the k means clustering algorithm.

V. CONCLUSION

Clustering is one of the most important unsupervised learning methods which is used in various applications. K means is one of the most popular clustering algorithms. While there have been many attempts to improve the efficiency and speed of the k means algorithm, this paper proposes an improved k means by initialization using k harmonic means and reducing the computational complexity of the algorithm. Improving clustering results further with an algorithm that functions at a higher speed can be considered as an area for future research.

REFERENCES

[1] Nuwan Ganganath, Chi-Tsun Cheng, and Chi K. Tse, Data Clustering with Cluster Size Constraints Using a Modified K-means Algorithm, International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 158-161, 2014.

[2] Juanying Xie, and Shuai Jiang, A simple and fast algorithm for global K-means clustering, Second International Workshop on Education Technology and Computer Science, pp. 36-40, 2010.

[3] K. Mahesh Kumar, and A. Rama Mohan Reddy, A Fast K-Means Clustering Using Prototypes for Initial Cluster Center Selection, IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO), 2015.

[4] Shi Na, Liu Xumin, and Guan Yong, Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm,

Third International Symposium on Intelligent Information Technology and Security Informatics, pp. 63-67, 2010.

[5] Saurabh Shah, and Manmohan Singh, Comparison of A Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid algorithm, International Conference on Communication Systems and Network Technologies, pp. 435-437, 2012.

[6] Greg Hamerly, Making k-means even faster, SDM DOI: 10.1137/1.9781611972801.12, pp. 130-140, 2010.

[7] Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii, Scalable k-means++, Journal Proceedings of the VLDB Endowment, Vol. 5 Issue 7, pp. 622-633, 2012.

[8] Lei Gu, A Novel Subtractive Clustering by Using K Harmonic Means Clustering for Initialization, 7th IEEE International Conference on Software Engineering and Service Science(ICSESS), pp. 840-843, 2016.

[9] Jianpeng Qi, Yanwei Yu*, Lihong Wang,and Jinglei Liu, K*-Means: An Effective and Efficient K-means Clustering Algorithm, IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom), pp. 242-249, 2016.

[10] Caiquan Xiong, Zhen Hua, Ke Lv, Xuan Li, An Improved K-means text clustering algorithm By Optimizing initial cluster centers, 7th International Conference on Cloud Computing and Big Data, pp. 265-268, 2016