International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Initial centroid selection for K- means clustering algorithm using the statistical method

N Sujatha [1,*], Latha Narayanan Valli [2], A Prema [1], SK Rathiha [3] and V Raja [3]

[1] *Department of Computer Science, Sri Meenakshi Government Arts College for Women, Madurai, Tamil Nadu, India.*
[2] *Standard Chartered Global Business Services Sdn Bhd, Kuala Lumpur, Malaysia.*
[3] *Department of Physics, Government Arts College, Melur, Tamil Nadu, India.*

International Journal of Science and Research Archive, 2022, 07(02), 474–478

Publication history: Received on 07 November 2022; revised on 20 December 2022; accepted on 22 December 2022

Article DOI: https://doi.org/10.30574/ijsra.2022.7.2.0309

Abstract

An iterative process that converges to one of the many local minima is used in practical clustering methods. K-means clustering is one of the most well-liked clustering methods. It is well known that these iterative methods are very susceptible to the initial beginning circumstances. In order to improve K-means clustering's performance, this research suggests a novel method for choosing initial centroids. The suggested approach is evaluated with online access records, and the results demonstrate that better initial starting points and post-processing cluster refinement result in better solutions.

Keywords: Web data clustering; Web Usage Mining; K-means; Initial Centroids; Web access logs; Genetic Algorithm;

1. Introduction

Clustering techniques have become very popular in a number of areas, such as engineering, medicine, biology, and data mining [1,2]. A good survey on clustering algorithms can be found in [3]. The k-means algorithm [4] is one of the most widely used clustering algorithms. The algorithm partitions the data points (objects) into C groups (clusters), so as to minimize the sum of the (squared) distances between the data points and the center (mean) of the clusters.

To apply the k-means algorithm, do the following:

- Choose C data points to initialize the clusters
- For each data point, find the nearest cluster center that is closest and assign that data point to the corresponding cluster
- Update the cluster centers in each cluster using the mean of the data points which are assigned to that cluster
- Repeat steps 2 and 3 until there are no more changes in the values of the means

In spite of its simplicity, the k-means algorithm involves a very large number of nearest-neighbor queries. The high time complexity of the k-means algorithm makes it impractical for use in the case of having a large number of points in the data set. Reducing a large number of nearest neighbor queries in the algorithm can accelerate it. In addition, the number of distance calculations increases exponentially with the increase in the dimensionality of the data [5-7].

Many algorithms have been proposed to accelerate the k-means. In [5,6], the use of kd- trees[8] is suggested to accelerate the k-means. However, backtracking is required, a case in which the computation complexity is increased [7]. Kd-trees are not efficient for higher dimensions. Furthermore, it is not guaranteed that an exact match of the nearest

* Corresponding author: N Sujatha

neighbor can be found unless some extra search is done as discussed in [9]. Elkan [10] suggests the use of triangle inequality to accelerate the k-means. In [11], it is suggested to use RTrees. Nevertheless, R-Trees may not be appropriate for higher dimensional problems. In [12-14], the Partial Distance (PD) algorithm has been proposed. The algorithm allows early termination of the distance calculation by introducing a premature exit condition in the search process.

In this study, we propose a new algorithm to accelerate the k-means by choosing the initial centroids based on statistical modes. The paper is organized as follows: the following section presents the general k-means algorithm. Section 3 presents our proposed initial refinement procedure. Section 4 presents the results and the work is concluded in section 5.

2. Standard K-Means Algorithm

One of the most popular clustering techniques is the k-means clustering algorithm. Starting from a random partitioning, the algorithm repeatedly (i) computes the current cluster centers (i.e. the average vector of each cluster in data space) and reassigns each data item to the cluster whose centre is closest to it. It terminates when no more reassignments take place. By this means, the intra- cluster variance, that is, the sum of squares of the differences between data items and their associated cluster centers is locally minimized. k - Means' strength is its runtime, which is linear in the number of data elements, and its ease of implementation. However, the algorithm tends to get stuck in suboptimal solutions (dependent on the initial partitioning and the data ordering) and it works well only for spherically shaped clusters. It requires the number of clusters to be provided or to be determined (semi-) automatically. In our experiments, we run k-means using the correct cluster number _

1. Choose a number of clusters k
2. Initialize cluster centers 1,... k
    a. Could pick k data points and set cluster centers to these points
    b. Or could randomly assign points to clusters and take means of clusters
3. For each data point, compute the cluster center it is closest to (using some distance measure) and assign the data point to this cluster.
4. Re-compute cluster centers (mean of data points in a cluster)
5. Stop when there are no new re-assignments.

3. Initial refinement

The initial cluster centers are normally chosen either sequentially or randomly as given in the standard algorithm. The quality of the final clusters is based on these initial seeds. It may lead to a local minimum; this is one of the disadvantages in k- means clustering. To avoid this, in our proposed method, we are selecting the modes of the data vector as initial cluster centers. Based on the number of clusters, the modes are selected one after another. Initially, the first mode value is selected as the center for the first cluster and the next highest frequently occurred value is (next mode value) assigned as the center for the next cluster. With this modification, the k-means algorithm is tested with web usage data received from the Graphic, Visualization, & Usability Center's (GVU) 8th WWW User Survey (http://www.cc.gatech.edu/gvu/user_surveys/surve y-1998-10/). This web usage dataset contains a total of 10104 instances with 72 attributes. GVU runs the Surveys as a public service and as such, all results are available online. And the data set is clustered based on the users' occupation type such as computer, management, professional, education, and others.

4. Results

For clustering, two measures of cluster "goodness" or quality are used. One type of measure that allows us to compare different sets of clusters without reference to external knowledge is called an internal quality measure. As mentioned in the previous section, we will use a measure of "overall similarity" based on the pairwise similarity of data items in a cluster. The other type of measure lets us evaluate how well the clustering is working by comparing the groups produced by the clustering techniques to known classes. This type of measure is called an external quality measure. One external measure is entropy [15], which provides a measure of "goodness" for un-nested clusters or for the clusters at one level of hierarchical clustering. Another external measure is the F-measure, which, as we use it here, is more oriented toward measuring the effectiveness of hierarchical clustering. The F measure has a long history but was recently extended to data item hierarchies in.

There are many different quality measures and the performance and relative ranking of different clustering algorithms can vary substantially depending on which measure is used. However, if one clustering algorithm performs better than other clustering algorithms on many of these measures, then we can have some confidence that it is truly the best clustering algorithm for the situation being evaluated. As we shall see in the results sections, the bisecting k-means algorithm has the best performance for the three quality measures that we are about to describe.

4.1. Entropy

We use entropy as a measure of the quality of the clusters (with the caveat that the best entropy is obtained when each cluster contains exactly one data point). Let CS be a clustering solution. For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the "probability" that a member of cluster j belongs to class i. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula

$$E_j = -\sum_i P_{ij} \ log(P_{ij} \ )$$

Where the sum is taken over all classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster.

$$E_{cs} = \sum_{j=1}^{m} \frac{n_j * E_j}{n}$$

Where n_j is the size of cluster j, m is the number of clusters, and n is the total number of data points.

4.2. F measure

The second external quality measure is the F measure[16] , a measure that combines the precision and recall of ideas from information retrieval. We treat each cluster as if it were the result of a query and each class as if it were the desired set of data items for a query. We then calculate the recall and precision of that cluster for each given class. More specifically, for cluster j and class i

$$\text{Recall (i, j)} = n_{ij} / n_i$$

$$\text{Precision(i, j)} = n_{ij} / n_j$$

Where n_{ij} is the number of members of class i in cluster j, n_j is the number of members of cluster j and ni is the number of members of class i. The F measure of cluster j and class i is then given by

$$\text{F(i, j)} = (2 * \text{Recall(i, j)} * \text{Precision(i, j)}) / ((\text{Precision(i, j)} + \text{Recall(i, j)})$$

For an entire hierarchical clustering the F measure of any class is the maximum value it attains at any node in the tree and an overall value for the F measure is computed by taking the weighted average of all values for the F measure as given by the following:

$$F = \sum_i \frac{n_i}{n} \max\{F(i,j)\}$$

Table1. Presents the results, showing that our proposed method outperforms than the standard method.

Table 1 Clustering results with initial refinement

Methods	**Quality Measures**		**Time Complexity (in secs)**
	E	**F**	
Refined K-Means Standard	0.1730	0.9614	0.516
K-means	0.2373	0.9599	0.953

5. Conclusion

The k-means algorithm is one of the most widely used clustering algorithms. A practical approach to clustering uses an iterative procedure that converges to one of the numerous local minima. These iterative techniques are known to be especially sensitive to initial starting conditions. In this paper, we have proposed a novel method to improve the cluster quality from the k-means algorithm by choosing the initial cluster centers based on statistical mode-based calculation to allow the iterative algorithm to converge to a "better" local minimum. The proposed algorithm is tested with the web usage data and shows that refined initial starting points of clusters lead to improved solutions. Experimental results show that the proposed algorithm gave better results than the conventional algorithm when applied to real data sets. At present, we are applying a Genetic Algorithm (GA) to improve the cluster quality as a post-refinement process.

Compliance with ethical standards

Acknowledgments

The team of authors would like to thank all those who have helped to complete this work.

Disclosure of conflict of interest

The authors have no conflicts of interest to declare

References

[1] Lv T., Huang S., Zhang X., and Wang Z, "Combining Multiple Clustering Methods Based on Core Group", Proceedings of the Second International Conference on Semantics, Knowledge and Grid (SKG'06), pp. 29-29, 2006.

[2] Nock R., and Nielsen F., "On Weighting Clustering," IEEE Transactions and Pattern Analysis and Machine Intelligence, vol 28, no. 8, pp. 1223-1235, 2006.

[3] Xu R., and Wunsch D., "Survey of clustering algorithms", IEEE Trans. Neural Networks, vol. 16, no. 3, pp. 645-678, 2005.

[4] MacQueen J., Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. and Prob, pp. 281-97, 1967.

[5] Kanungo T., Mount D.M., Netanyahu N., Piatko C., Silverman R., and Wu A.Y., "An efficient k-means clustering algorithm: Analysis and implementation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892, 2002.

[6] Pelleg D., and Moore A., "Accelerating exact k-means algorithm with geometric reasoning," Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, pp. 727-734, 1999.

[7] Sproull R., "Refinements to Nearest-Neighbor Searching in K-Dimensional Trees," Algorithmica, vol. 6, pp. 579-589, 1991.

[8] Bentley J., "Multidimensional Binary Search Trees Used for Associative Searching," Commun. ACM, vol. 8, no. 9, pp. 509-517, 1975.

[9] Friedman J., Bentley J., and Finkel R., "An Algorithm for Finding Best Matches in Logarithmic Expected Time," ACM Trans. Math. Soft. Vol. 3, no. 2, pp. 209-226, 1977.

[10] Elkan, C., "Using the Triangle Inequality to Accelerate k-Means," Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), pp. 609-616, 2003.

[11] Yang, J.; Wang, J. Tag clustering algorithm lmmsk: Improved k-means algorithm based on latent semantic analysis. J. Syst. Electron. 2017, 28, 374–384.

[12] Proietti, G. and Faloutsos C., "Analysis of Range Queries and Self-spatial Join Queries on Real Region Datasets Stored using an R- tree," IEEE Transactions on Knowledge and Data Engineering, vol. 5, no. 12, pp. 751-762, 2000.

[13] Md. Zubair 1,□, MD.Asif Iqbal 1,b, Avijeet Shil 1,c, Enamul Haque 2,d , Mohammed Moshiul Hoque 1,e, and Iqbal H. Sarker 1 , An Efficient K-means Clustering Algorithm for Analysing COVID-19. Dec.2020

[14] Bei C., and Gray, R., "An Improvement of the Minimum Distortion Encoding Algorithm for Vector Quantization," IEEE Transactions on Communications, vol. 33, no. 10, pp. 1132- 1133, 1985.

[15] Shannon CE., "A mathematical theory of communication, Bell System Technical Journal", vol. 27, pp. 379-423 and 623-656, July and October, 1948.

[16] Larsen B., and Aone C. "Fast and Effective Text Mining Using Linear-time Document Clustering," KDD-99, San Diego, California, 1999.

[17] Ahmed, M.; Barkat Ullah, A.S.S.M. Infrequent pattern mining in smart healthcare environment using data summarization. J. Supercomput. 2018, 74, 5041–5059