

UNSUPERVISED K-MEANS CLUSTERING ALGORITHM

Badalova M.

moha89@mail.ru

Samarkand branch of Tashkent University of Information Technologies

Abstract *The k-means algorithm is generally the most known and used clustering method. There are various extensions of k-means to be proposed in the literature. Although it is an unsupervised learning to clustering in pattern recognition and machine learning, the k-means algorithm and its extensions are always influenced by initializations with a necessary number of clusters a priori. That is, the k-means algorithm is not exactly an unsupervised clustering method. In this paper, we construct an unsupervised learning schema for the k-means algorithm so that it is free of initializations without parameter selection and can also simultaneously find an optimal number of clusters. That is, we propose a novel unsupervised k-means (U- k-means) clustering algorithm with automatically finding an optimal number of clusters without giving any initialization and parameter selection. The computational complexity of the proposed U-k-means clustering algorithm is also analyzed. Comparisons between the proposed U-k-means and other existing methods are made. Experimental results and comparisons actually demonstrate these good aspects of the proposed U-k-means clustering algorithm.*

Index Terms *Clustering, K-means, number of clusters, initializations, unsupervised learning schema, Unsupervised k-means (U-k-means).*

1. INTRODUCTION

Clustering is a useful tool in data science. It is a method for finding cluster structure in a data set that is characterized by the greatest similarity within the same cluster and the greatest dissimilarity between different clusters. Hierarchical clustering was the earliest clustering method used by biologists and social scientists, whereas cluster analysis became a branch of statistical multivariate analysis. It is also an unsupervised learning approach to machine learning. From statistical viewpoint, clustering methods are generally divided as probability model-based approaches and nonparametric approaches. The probability model-based approaches follow that the data points are from a mixture probability model so that a mixture likelihood approach to clustering is used. In model-based approaches, the expectation and maximization (EM) algorithm is the most used. For nonparametric approaches, clustering methods are mostly based on an objective



function of similarity or dissimilarity measures, and these can be divided into hierarchical and partitional methods where partitional methods are the most used.

2. RELATED WORKS

In this section, we review several works that are closely related with ours. The k-means is one of the most popular unsupervised learning algorithms that solve the well-known clustering problem. Let $\mathbf{X} = \{x_1, \dots, x_n\}$ be a data set in

a d -dimensional Euclidean space \mathbb{R}^d . Let $A = \{a_1, \dots, a_c\}$

be the c cluster centers. Let $z = [z_{ik}]_{n \times c}$, where z_{ik} is a binary variable (i.e. $z_{ik} \in \{0, 1\}$) indicating if the data point x_i belongs to k -th cluster, $k = 1, \dots, c$. The

k-means

$$z_{ik} = \begin{cases} 1 & \text{if } \|x_i - a_k\| = \min_{1 \leq k \leq c} \|x_i - a_k\| \\ 0 & \text{otherwise.} \end{cases}$$

The k-means algorithm is iterated through necessary conditions for minimizing the k-means objective function $J(z, A)$ with updating equations for cluster centers and memberships, respectively, as

$$a_k = \frac{\sum_{i=1}^n z_{ik} x_{ij}}{\sum_{i=1}^n z_{ik}} \text{ and}$$

$$z_{ik} = \begin{cases} 1 & \text{if } \|x_i - a_k\|^2 = \min_{1 \leq k \leq c} \|x_i - a_k\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

where $\|x_i - a_k\|$ is the Euclidean distance between the data point x_i and the cluster center a_k . There exists a difficult problem in k-means, i.e., it needs to give a number of clusters a priori. However, the number of clusters is generally unknown in real applications. Another problem is that the k-means algorithm is always affected by initializations.

There are several clustering validity indices available for estimating the number c of clusters. Clustering validity indices can be grouped into two major categories: external and internal. External indices are used to evaluate clustering results by comparing cluster memberships assigned by a clustering algorithm with the previously known knowledge such as externally supplied class label. However, internal indices are used to evaluate the goodness of cluster structure by focusing on the intrinsic information of the data itself so that we consider only internal indices. In the paper, these most widely used internal indices, such as original Dunn's index (DNo), Davies-Bouldin index (DB), Silhouette Width (SW), Calinski and Harabasz



index (CH) , Gap statistics, generalized Dunn's index (DNg), and modified Dunn's index (DNs) are chosen for finding the number of clusters and then compared with our proposed U-k-means clustering algorithm.

The DNo, DNg, and DNs are supposed to be the simplest (internal) validity index where it compares the size of clusters with the distance between clusters. The DNo, DNg, and DNs indices are computed as the ratio between the minimum distance between two clusters and the size of the largest cluster, and so we are looking for the maximum value of index values. Davies-Bouldin index (DB) measures the average similarity between each cluster and its most similar one. The DB validity index attempts to maximize these between cluster distances while minimizing the distance between the cluster centroid and the other data objects. The Silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. Thus, positive and negative large silhouette widths (SW) indicate that the corresponding object is well clustered and wrongly clustered, respectively. Any objects with the SW validity index around zero are considered not to be clearly discriminated between clusters. The Gap statistic [20] is a cluster validity measure based upon a statistical hypothesis test. The gap statistic works by comparing the change in within-cluster dispersion with that expected under an appropriate reference null distribution at each value c . The optimal number of clusters is the smallest c .

3. THE UNSUPERVISED K-MEANS CLUSTERING ALGORITHM There always exists a difficult problem in the k-means algorithm and its extensions for a long history in the literature. That is, they are affected by initializations and require a given number of clusters a priori. We mentioned that the X-means algorithm has been used for clustering without given a number of clusters a priori, but it still needs to specify a range of number of clusters based on BIC, and it is still influenced by initializations. To construct the k-means clustering algorithm with free of initializations and automatically find the number of clusters, we use the entropy concept. We borrow the idea from the EM algorithm by Yang et al. We first consider proportions α_k in which the α_k term is seen as the probability of one data point belonged to the k th class. Hence, we use $-\ln \alpha_k$ as the information in the occurrence of one data point belonged to the k th class, and so $-\sum_{k=1}^c \alpha_k \ln \alpha_k$ becomes the average of information. In fact, the term $-\sum_{k=1}^c \alpha_k \ln \alpha_k$ is the entropy over proportions α_k . When $\alpha_k = 1/c, \forall k = 1, 2, \dots, c$, we say that there is no information about α_k . At this point, we have the entropy achieve the maximum



value. Therefore, we add this term to the k-means objective function $J(z, A)$ as a penalty. We then construct a schema to estimate α_k by minimizing the entropy to get the most information for α_k . To minimize $-\sum_{k=1}^c \alpha_k \ln \alpha_k$ is equivalent to maximizing $\sum_{k=1}^c \alpha_k \ln \alpha_k$. For this reason, we use $\sum_{k=1}^c \alpha_k \ln \alpha_k$ as a penalty term for the k-means objective function $J(z, A)$. Thus, we propose a novel objective function as follows: $\beta \geq 0$ $J_{U-k-means}(z, A, \alpha) = \sum_{i=1}^n \sum_{k=1}^c z_{ik} \|x_i - a_k\|^2 - \beta n \sum_{k=1}^c \alpha_k \ln \alpha_k$ (1) In order to determine the number of clusters, we next consider another entropy term. We combine the variables membership z_{ik} and the proportion α_k . By using the basis of entropy theory, we suggest a new term in the form of $\sum_{i=1}^n \sum_{k=1}^c z_{ik} \ln \alpha_k$. Thus, we propose the unsupervised k-means (U-k-means) objective function as follows:

$$J_{U-k-means}(z, A, \alpha) = \sum_{i=1}^n \sum_{k=1}^c z_{ik} \|x_i - a_k\|^2 - \beta n \sum_{k=1}^c \alpha_k \ln \alpha_k - \gamma \sum_{i=1}^n \sum_{k=1}^c z_{ik} \ln \alpha_k$$

We know that, when β and γ in Eq. are zero, it becomes the original k-means. The Lagrangian of Eq. is

$$J_{U-k-means}(z, A, \alpha) = \sum_{i=1}^n \sum_{k=1}^c z_{ik} \|x_i - a_k\|^2 - \beta n \sum_{k=1}^c \alpha_k \ln \alpha_k - \gamma \sum_{i=1}^n \sum_{k=1}^c z_{ik} \ln \alpha_k$$

Algorithm by Yang et al. This is the robust-learning fuzzy c-means (RL-FCM) proposed by Yang and Nataliani. In Yang and Nataliani, they gave the RL-FCM objective function $P J(U, \alpha, A) = \sum_{i=1}^n \sum_{k=1}^c \mu_{ik} \|x_i - a_k\|^2 - r_1 \sum_{i=1}^n \sum_{k=1}^c \mu_{ik} \ln \alpha_k + r_2 \sum_{i=1}^n \sum_{k=1}^c \mu_{ik} \ln \mu_{ik} - r_3 n \sum_{k=1}^c \alpha_k \ln \alpha_k$ with μ_{ik} , not binary variables, but fuzzy c-memberships with $0 \leq \mu_{ik} \leq 1$ and $\sum_{k=1}^c \mu_{ik} = 1$ to indicate fuzzy memberships for the data point x_i belonging to k-th cluster. If we compare the proposed U-k-means objective function $J_{U-k-means}(z, A, \alpha)$ with the RL-FCM objective function $J(U, \alpha, A)$, we find that, except μ_{ik} and z_{ik} with different membership representations, the RL-FCM objective function $J(U, \alpha, A)$ in Yang and Nataliani gave more extra terms and parameters and so the RL-FCM algorithm is




more complicated than the proposed U-k-means algorithm with more running time. For experimental results and comparisons in the next section, we make more comparisons of the proposed U-k-means algorithm with the RL-FCM algorithm. We also analyze the computational complexity for the U-k-means algorithm. In fact, the U-k-means algorithm can be divided into three parts: (1) Compute the hard membership partition z_{ik} with $O(ncd)$; Compute the mixing proportion α_k with $O(nc)$; (3) Update the cluster center a_k with $O(n)$. The total computational complexity for the U-k-means algorithm is $O(ncd)$, where n is the number of data points, c is the number of clusters, and d is the dimension of data points. Compared with the RL-FCM algorithm, the RL-FCM has the total computational complexity with $O(nc^2d)$.

REFERENCES

- [1] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.
- [2] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. New York, NY, USA: Wiley, 1990.
- [3] G. J. McLachlan and K. E. Basford, Mixture Models: Inference and Applications to Clustering. New York, NY, USA: Marcel Dekker, 1988.
- [4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," J. Roy. Stat. Soc., Ser. B, Methodol., vol. 39, no. 1, pp. 1–38, 1977.
- [5] J. Yu, C. Chaomurilige, and M.-S. Yang, "On convergence and parameter selection of the EM and DA-EM algorithms for Gaussian mixtures," Pattern Recognit., vol. 77, pp. 188–203, May 2018.
- [6] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, Jun. 2010.
- [7] M.-S. Yang, S.-J. Chang-Chien, and Y. Nataliani, "A fully-unsupervised possibilistic C-Means clustering algorithm," IEEE Access, vol. 6, pp. 78308–78320, 2018.
- [8] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp. Math. Statist. Probab., vol. 1, 1967, pp. 281–297.
- [9] M. Alhawarat and M. Hegazi, "Revisiting K-Means and topic modeling, a comparison study to cluster arabic documents," IEEE Access, vol. 6, pp. 42740–42749, 2018.





[10] Y. Meng, J. Liang, F. Cao, and Y. He, “A new distance with derivative information for functional k-means clustering algorithm,” *Inf. Sci.*, vols. 463–464, pp. 166–185, Oct. 2018.

[11] Z. Lv, T. Liu, C. Shi, J. A. Benediktsson, and H. Du, “Novel land cover change detection method based on k-Means clustering and adaptive majority voting using bitemporal remote sensing images,” *IEEE Access*, vol. 7, pp. 34425–34437, 2019. [12] J. Zhu, Z. Jiang, G. D. Evangelidis, C. Zhang, S. Pang, and Z. Li, “Efficient registration of multi-view point sets by K-means clustering,” *Inf. Sci.*, vol. 488, pp. 205–218, Jul. 2019.

