Scientific Research Publishing

Intelligent Information Management, 2018, 10, 43-48
http://www.scirp.org/journal/iim
ISSN Online: 2160-5920
ISSN Print: 2160-5912

An Experiment of *K*-Means Initialization Strategies on Handwritten Digits Dataset

Boyang Li

High School Affiliated to Xi'an Jiaotong University, Xi'an, China
Email: xajdfz@hotmail.com

How to cite this paper: Li, B.Y. (2018) An Experiment of *K*-Means Initialization Strategies on Handwritten Digits Dataset. *Intelligent Information Management*, **10**, 43-48.
https://doi.org/10.4236/iim.2018.102003

Received: October 20, 2017
Accepted: February 25, 2018
Published: February 28, 2018

Abstract

Clustering is an important unsupervised classification method which divides data into different groups based some similarity metrics. *K*-means becomes an increasing method for clustering and is widely used in different application. Centroid initialization strategy is the key step in *K*-means clustering. In general, *K*-means has three efficient initialization strategies to improve its performance *i.e.*, Random, *K*-means++ and PCA-based *K*-means. In this paper, we design an experiment to evaluate these three strategies on UCI ML hand-written digits dataset. The experiment result shows that the three *K*-means initialization strategies find out almost identical cluster centroids, and they have almost the same results of clustering, but the PCA-based *K*-means strategy significantly improves running time, and is faster than the other two strategies.

Keywords

K-means, Clustering Performance Evaluation, Machine Learning, Principal Component Analysis

1. Introduction

Machine Learning, in general, is a power tool to predict the properties of unknown data based on a set of training data with or without labels. Generally, there are two types of learning methods: one is unsupervised learning and the other is supervised learning. In supervised learning, training data has explicit labels (called labeled data). However, in some cases, it is difficult to obtain the labeled training data. Unsupervised learning is the best choice to classify similar patterns into the same group without labeled data. As the results of clustering, data in the same group has higher similarity metric to each other than to those in other groups, and each group have a centroid to represent the group. In the predicting phase, the unknown data will be assigned to the group which has the minim distance between predicted data and group centroid. *K*-means algorithm performs good comparing with other clustering algorithms and it has good robustness [1].

As shown in Equation (1), in *K*-means clustering, the number of group *K* is predetermined. By initialing *K* centroids, distance metric can be calculated. For instance, the Euclidean distance between the point and centroids are calculated as shown in (2). Then, it changes the group centroids and repeats the above steps. The algorithm tries to minimize sum-squared-error criterion (SSE) of total distance metric greedily, in such a way that *K*-means finds out the group centroids and predicts the unknown data to the nearest group centroids.

$$E = \sum_{K=1}^{K} \sum_{x \in C_k} d^2 \left(x, m_k \right) \tag{1}$$

$$d^2 \left(x, m_k \right) = \sum_{n=1}^{N} \left(x_n - m_{kn} \right)^2 \tag{2}$$

Studies have shown that: the performance of *K*-means is strongly depending on the initialization strategies of centroid locations [2]. By now, Random Partition method, *K*-means++ and PCA-based *K*-means are the top three efficient strategies for the initialization. In this paper, experimental results show the different performances by comparing with aforementioned three initialization strategies for *K*-means both in the running time and the quality of the results.

2. Classical *K*-Means Initialization Strategies

In this section, we introduce the three dominate *K*-means initialization strategies. We can see that the three strategies have different influence on the results of clustering.

2.1. Random Partition Method

Random Partition initialization method [3] chooses *K* initial centroids randomly from the data set, and *K* is the estimated number of clusters. Then it calculates the distance between each point and the initial centroids, and the average distance can be computed. Then we adjust the clustering centroids and re-compute the distance. These two steps run iteratively until finding out the minim average distance, which means these *K* centroids are clustering centers.

However, Celebi *et al.* [4] found that random partition initialization method cannot guarantee the global optimum of the clustering algorithm, and the clustering results have an excessive dependence on the initial centroids [5] [6].

2.2. *K*-Means++

K-means++ [7] is an algorithm for selecting the initial cluster centroids more reasonably to avoid the poor clustering result found by the standard *k*-means algorithm. This algorithm chooses the first initial cluster centroid uniformly from the data points. In terms of iterative steps, the centroid should be updated based on two factors: one is the squared distance, and the other is the probability proportion draws from the point which is closed to existing cluster centroids.

K-means++ initial strategy dose not only speed up convergence, but also provides a better solution compared with random *K*-means solution.

2.3. *K*-Means Clustering via Principal Component Analysis (PCA-Based *K*-Means)

For the aforementioned methods, it is in the raw or original high dimensional space where the task of searching for better clustering has been performed. Recent work [8] analyzes theoretically the relationship between *K*-means clustering and principal component analysis(PCA), and draws conclusion that the smaller PCA subspace is not only contain the global solution to *K*-means clustering lies in, but identical to the cluster centroid subspace. These conclusions enlighten an effective and efficient way to find out clustering center in PCA subspace not in the original space.

3. Experiment and Results

In this experiment we compare above three initialization strategies for *K*-means in terms of runtime and quality of the results on.

3.1. Dataset

Some digits samples in UCI ML hand-written digits datasets are shown as **Figure 1**. [9] National Institute of Standards and Technology (NIST) provides normalized bitmaps of handwritten digits. We can use these samples to do the preprocessing. The steps of preprocessing are as follows:

Bitmaps of handwritten digits which derive from 43 people are divided into two parts: 30 samples for the training set and 13 samples for the test set.

Every digit is 32 × 32 bitmap, and then it is separated into 4 × 4 non-overlapping blocks. Each block records the number of one pixel.

An input matrix of 8 × 8 for each digit is generated and each matrix element is an integer in the range 0.16.

Thus, this dataset has 1797 8 × 8 images and every image is vectorized 64 feature vector with ground true labels.

Since the dataset provides basic facts, we can apply different cluster quality metrics to evaluate the goodness of fit of the cluster labels to the basic facts. It has influence in the initialization strategies of *K*-means.

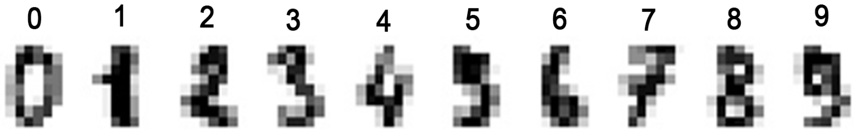

Figure 1. UCI ML hand-written digits image examples.

3.2. Clustering Performance Evaluation

3.2.1. Inertia

Inertia or within-cluster sum of squares distance is a key measure to evaluate the internally coherent of clustering. The sum of squared distance is calculated between each point and its nearest centroid.

3.2.2. Homogeneity (Shorthand as Homo)

In fact, the result of clustering should satisfy homogeneity. It means that each point only belongs to a cluster. This rule should be also independent of labels. The range of score should be standardized between 0.0 and 1.0.

3.2.3. Completeness (Shorthand as Compl)

Completeness measure how well the *K*-means algorithm assigns all the data points with a given label to the same group. Meanwhile, the score should be standardized from 0.0 to 1.0.

3.2.4. V-Measure (Shorthand as V-Meas)

Specifically, V-measure measures the harmonic criteria whether it has satisfied the homogeneity and completeness. In addition, the score is from 0.0 to 1.0.

3.2.5. Silhouette Coefficient (Shorthand as Silhouette)

The Silhouette Coefficient for a sample is defined as:

$$\text{silhouette} = \frac{a-b}{\max(a,b)}$$

where a is the mean of intra-cluster distance, b indicates the nearest-cluster distance. Moreover, the range of the parameter is $-1 \sim 1$. Specifically, 1 is the best result and -1 is the worst result. The higher the score of Silhouette Coefficient is, the more suitable the model satisfies the defined clusters.

3.3. Results

In this experiment, we compare the performance of three the classical initialization strategies based on the above-mentioned criteria. A PC with Intel® Core™ i7-6700 CPU @ 3.40 GHz × 8 is used to run this experiment.

In order to show the clustering results in 2D coordinates, we use PCA to reduce the dataset dimension to 2D, and transform the feature vector with length 64 to the 2D subspace. The reduced dataset is plotted as dot marker, and the clustering centroids are put on the figure with different markers as be showed in Figure 2.

Form Figure 2, it shows that three *K*-means initialization strategies find out almost identical cluster centroids. In addition, they have the similar accuracy of clustering.

Table 1 gives the result of Clustering performance evaluation.

As shown in Table 1, silhouette coefficient (about 0.15) shows these three clustering algorithms separated test dataset points into 10 cluster successfully

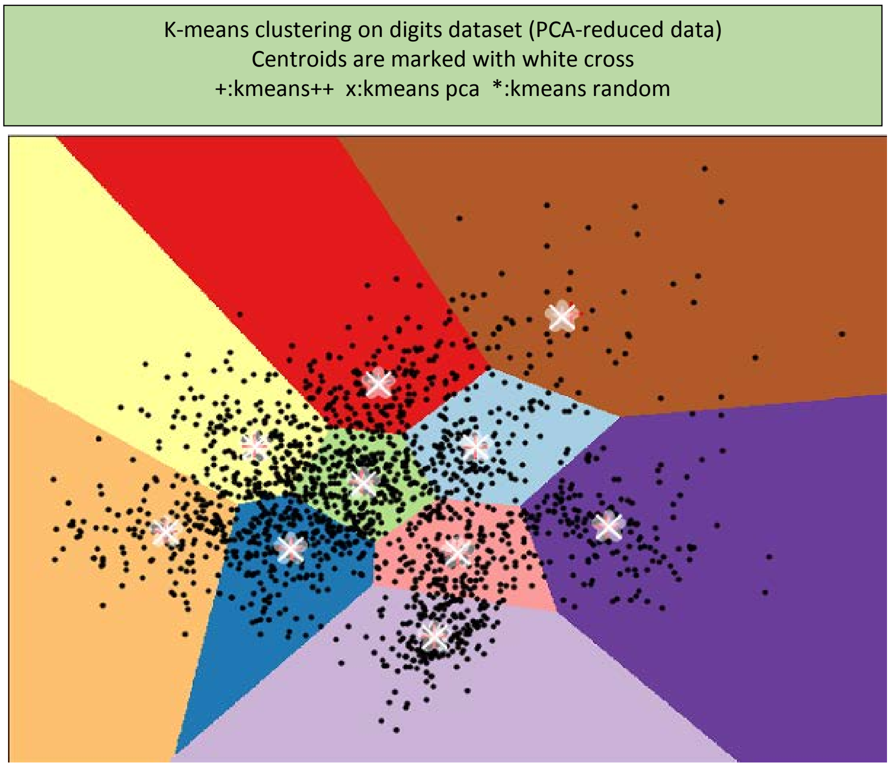

Figure 2. *K*-means clustering centroids.

Table 1. Evaluation results on three classical initialization strategies.

	Init time	Inertia	Homo	Compl	V-meas	Silhouette
K-means++	0.24 s	69432	0.602	0.650	0.625	0.146
Random	0.17 s	69694	0.669	0.710	0.689	0.147
PCA-based	0.02 s	71820	0.673	0.715	0.693	0.150

despite the separation distance is small. Homo and compl indicator are all in the range of 0.0 and 1.0 with near values, which means the results can be receivable. The values of V-means (0.625, 0.689 and 0.693) state that the accuracy of homo and compl is successfully calculated. These four evaluation indicators confirm that the three classical clustering algorithms have acceptable clustering results on test dataset.

One noticeable thing is the running time. From the evaluation results in Table 1, PCA-based *K*-means strategy significantly improves running time (about faster ten times than other two). It performs better than other strategies.

4. Conclusion

In this work, we design an experiment to evaluate the performance of three classical *K*-mean initialization strategies on UCI ML hand-written digits dataset: Random, *K*-means++ and PCA-based *K*-means. The experiment results show that the three *K*-means initialization strategies find out almost identical cluster centroids, and they have the similar accuracy of clustering. However, PCA-based K-means strategy significantly improves running time. Moreover, PCA-based K-means strategy has a better performance than other strategies, thus it is more effective for clustering. In further studies, more machine learning models like neural networks can be investigated and compared with models used in this paper.

References

[1] Pfitzner, D., Leibbrandt, R. and Powers, D. (2009) Characterization and Evaluation of Similarity Measures for Pairs of Clusterings. *Knowledge and Information Systems*, **19**, 361-394. https://doi.org/10.1007/s10115-008-0150-6

[2] Kodinariya, T.M. (2014) Survey on Exiting Method for Selecting Initial Centroids in K-Means Clustering. *International Journal of Engineering Development and Research*, **2**, 2865-2868.

[3] Hamerly, G. and Elkan, C. (2002) Alternatives to the K-Means Algorithm that Find Better Clusterings. *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (*CIKM*), McLean, VA, 4-9 November 2002, 600-607. https://doi.org/10.1145/584792.584890

[4] Celebi, M.E., Kingravi, H.A. and Vela, P.A. (2013) A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm. *Expert Systems with Applications*, **40**, 200-210, arXiv:1209.1960 . https://doi.org/10.1016/j.eswa.2012.07.021

[5] Bradley, P.S. and Fayyad, U.M. (1998) Refining Initial Points for K-Means Clustering. *Proceedings of the Fifteenth International Conference on Machine Learning*, Madison, WI, 24-27 July 1998, 91-99.

[6] Vattani, A. (2011) K-Means Requires Exponentially Many Iterations Even in the Plane. *Discrete and Computational Geometry*, **45**, 596-616. https://doi.org/10.1007/s00454-011-9340-1

[7] Arthur, D. and Vassilvitskii, S. (2007) K-Means++: The Advantages of Careful Seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, Philadelphia, PA, 1027-1035.

[8] Ding, C. and He, X.F. (2004) K-Means Clustering via Principal Component Analysis. *Proceedings of the Twenty-First International Conference on Machine Learning*, Banff, Alberta, 4-8 July 2004, 29. https://doi.org/10.1145/1015330.1015408

[9] UCI Machine Learning Repository: Hand-Written Digits Datasets. http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits