Exponential Probability Clustering

Hou Yuxi, Cheol Hoon Park
School of Electrical Engineering and Computer Science
KAIST
Email: hyyyx@kaist.ac.kr

Abstract

K-means is a popular one in clustering algorithms, and it minimizes the mutual euclidean distance among the sample points. But K-means has some demerits, such as depending on initial condition, unsupervised learning and local optimum. However mahalanobis distancecan deal this case well. In this paper, the author proposed a new clustering algorithm, named exponential probability clustering, which applied Mahalanobis distance into K-means clustering. This new clustering does possess not only the probability interpretation, but also clustering merits. Finally, the simulation results also demonstrate its good performance compared to K-means algorithm.

I. Introduction

1.1 K-means Clustering

K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.[1]The main idea is to define k centroids, one for each cluster. The next step is to associate each point to the nearest centroid. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. Then a loop has been generated, the k centroids change their location step by step until no more changes are done. K-means clustering update centers as following [2]:

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \quad (1)$$

1.2 Mahalanobis distance

Mahalanobis distance is a distance measure introduced by P. C. Mahalanobis in 1936. It is based on correlations between variables by which different patterns can be identified and analyzed. It is a useful way of determining similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant, i.e. not dependent on the scale of measurements [3]. Formally, the Mahalanobis distance from a group of values with mean $\mu = (\mu_1 \quad \mu_2 \quad \cdots \quad \mu_p)$ and covariance matrix $\Sigma$for a multivariate vector $x = (x_1 \quad x_2 \quad \cdots \quad x_p)^T$is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (2)$$

II. Exponential probability clustering

Like K-means clustering algorithm, the objective function is given by:

$$maximize\ J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}|\Sigma_k|^{-1/2}\exp\{-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\} \quad (3)$$

Which represents the sum of the squares of the Mahalanobis distance of each data point to its assigned vector μ_k. Our goal is to find values for the $\{r_{nk}\}$ and the $\{\mu_k\}$ so as to minimize J.Where

$$r_{nk} = \begin{cases} 1 & if\ k = argmin_j \frac{1}{2}(x_n-\mu_j)^T\Sigma_j^{-1}(x_n-\mu_j) \\ 0 & otherwise \end{cases} \quad (4)$$

Now considering the optimization of the μ_k with the r_{nk} held fixed, and the objective function could be minimized by setting its derivative with respect to μ_k to zero giving

$$\frac{\partial J}{\partial \mu_k} = 0 \Rightarrow \sum_{n=1}^{N} -r_{nk}|\Sigma_k|^{-1/2}\exp\{-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\}\cdot\Sigma_k^{-1}(x_n-\mu_k) = 0$$

(5) Which we can easily solve for μ_k to give

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} r_{nk}\cdot\exp\{-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\}\cdot x_n}{\sum_{n=1}^{N} r_{nk}\cdot\exp\{-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\}} \quad (6)$$

Where $\Sigma_k^{new} = \sum_{n=1}^{N} r_{nk}\cdot(x_n-\mu_k^{new})(x_n-\mu_k^{new})^T$ (7)

So, the iteration algorithm can be described as: Step1: setting initial condition for parameter μ_k and Σ_k Step2: calculating values for series $\{r_{nk}\}$ by equation (4) Step3: updating μ_k and Σ_k by equation (6) and (7) Step4: repeating step2 and 3 until the objective function value doesn't increase any more;

Ⅲ. Experimental Results and Analysis

A two-dimensional sample data is generated as: one part is centered at point (0,0) with covariance matrix [3 0;0 1] while the other part centered at (5,5) with covariance[1 0;0 2], shown as Figure1. Both K-means clustering and exponential probability clustering are applied into this data. The red marked star is from the former and red circle "0"by the later.As shown as figure1, the center estimated by K-means is: [-0.65 -0.01] and [4.97 4.45], compared to [-0.08 0.04] and [4.97 4.84] by exponential probability clustering. Obviously, our clustering estimate centers more correctly. Figure2 shows us the relationship of objective function value increasing as iteration going.

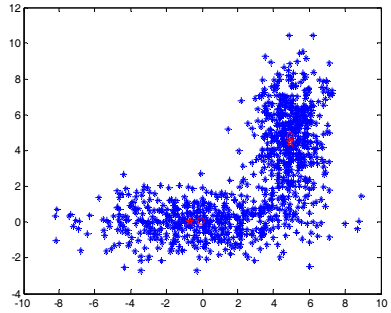

Figure1 Test data and center estimation results

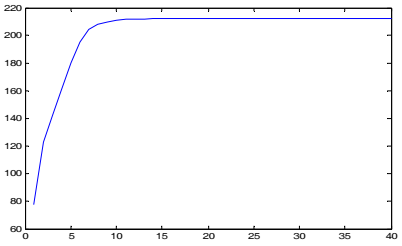

Figure2 Objective function value increasing as iteration

Ⅳ. Conclusion

Exponential probability clustering is connected to probability interpretation more than K-means, and compared to euclidean distance, mahalanobis distance represents data relationship better. Because of this, our proposed clustering has better performance in previous simulation. When every cluster data has covariance as p*I, p real number, I unit matrix, both two clustering algorithm has same results. Accordingly, from this sense K-means cluster is only one special case of exponential probability clustering.

Reference

[1] Tapas Kanungo, An Efficient k-Means Clustering Algorithm: Analysis and Implement- ation, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002.

[2] Christopher M.Bishop, Pattern Recognition and Machine Learning, 2006 Springer Science + Business Media, LLC, 2006.

[3] Matthias W¨olfel and Hazim Kemal Ekenel, Feature Weighted Mahalanobis Distance: Improved Robustness for Gaussian Classifiers.