Incremental Estimation of Gaussian Mixture Models for Online Data Stream Clustering

Mingzhou Song
Department of Computer Science
Queens College and Graduate Center

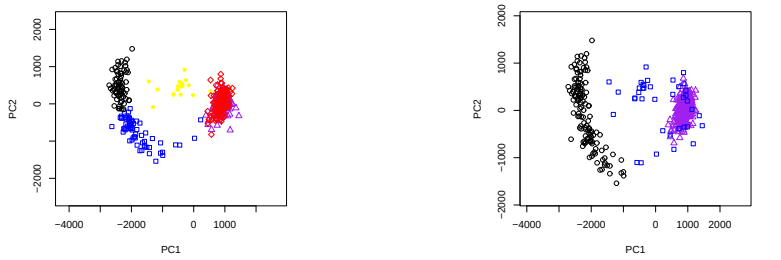
Hongbin Wang
Department of Computer Science
Graduate Center

Abstract

We present a probability density based data stream clustering approach which requires only the newly arrived data, not the entire historical data, to be saved in memory. This approach incrementally updates the density estimate taking only the newly arrived data and the previously estimated density. The idea roots on a theorem of density updating and it works naturally with Gaussian mixture models. We implement it through the expectation maximization algorithm and a cluster merging strategy by multivariate statistical tests for equality of covariance and mean. Our approach is much more practical in clustering voluminous *online* data streams than the standard EM algorithm. We demonstrate the performance of our algorithm on a simulated Gaussian mixture data stream and a real neural spike train data stream.

We consider *recent data* as all the data available in the memory from the data stream. We define *historical data* as the data observed in the data stream so far. We call unprocessed recent data *newly arrived data*. If the entire historical data were available in memory, Gaussian mixture model would have been effectively estimated using the Expectation Maximization (EM) algorithm. For voluminous data streams, however, the EM algorithm is not efficient. For a data stream without complete historical records, the EM or any of its known variations is not applicable. We argue in this paper that we can adapt probability density based clustering algorithms to solve data stream clustering problems much more efficiently than applying the EM on the entire historical data. We use the standard EM algorithm but only on newly arrived data. Our incremental Gaussian mixture model estimation algorithm merges Gaussian components that are statistically equivalent. The equivalence of two Gaussian components are determined using the W statistic for equality of covariance and Hotelling's T^2 statistic for equality of mean. The sufficient statistics of mean and covariance for the multivariate normal distribution make it possible to perform the tests and merging without resort to historical data.

A spike train is an extracellular action potential signal recorded with a probe, implanted in an animal subject under certain experimental condition. The candidate spikes were detected according to a low threshold in a spike window. Figure 1(a) shows the clusters obtained by our algorithm. Figure 1(b) shows the solution by the standard EM algorithm applied on the entire data of the same spike train. From this example, although for complex cluster shapes it differed with standard EM using all data, our algorithm produced a reasonable solution overall.



(a) Result by our incremental algorithm

(b) Result by standard EM with all data

Figure 1: Clustering on real neural spike train data stream