A Framework for Alternative Clustering in Stream Data

Jingyuan Zhang
OSCAR Team, SSDUT, DLUT
April. 20, 2011

Abstract

In this report we will discuss our framework for alternative clusterings in stream data, which applies ADCO [1] to generate two dissimilar results with some semi-supervised information during the clustering process. We will refer to it as AltStream-*k*means and separate it into online and offline components as the stream clustering approach CluStream [2] does. As the classical stream clustering algorithm, CluStream is easy implemented and has good performance. Based on this technique, our alternative stream clustering framework would also be simple and feasible.

This two-phased AltStream-*k*means framework will be described as follows. Firstly, we will discuss how the two dissimilar groups of micro-clusters are maintained with the clustering similarity measure ADCO throughout the stream generation process. Since the time complexity of computing the ADCO value is much higher during the process of collecting micro-clusters, we will secondly introduce a strategy proposed in [3] to efficiently reduce the computational complexity of ADCO. And thirdly we will discuss how the dissimilar micro-clusterings may be used by an offline macro-clustering component to uncover two alternatives. Some semi-supervised information will be obtained during this process to guide for two dissimilar alternative results. It is totally different with the offline macro-clustering creation in CluStream.

Finally, we will present some experimental evaluations of our AltStream-*k*means with two comparative algorithms. Limited by our experiment progress, only a part of the results have been obtained now. We will introduce the detail of our experimental framework here.

References

[1] E. Bae, J. Bailey and G. Dong. A clustering comparison measure using density profiles and its application to the discovery of alternate clusterings. Data Mining and Knowledge Discovery, 2010.

[2] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. Proc. of VLDB 2003.

[3] Mills-Tettey, G. A., A. T. Stentz, and M. B. Dias. The dynamic hungarian algorithm for the assignment problem with changing costs. Technical Report in Robotics Institute of Carnegie Mellon University, 2007.