Alternative Clustering on Stream Data

Jingyuan Zhang
OSCAR Team, SSDUT, DLUT
Nov. 30, 2010

Abstract

In recent years, a large amount of streaming data, such as network flows, sensor data and web click streams have been generated and various data stream clustering algorithms are developed. However, just as the traditional data, multiple alternate clusterings may exist in stream data and they are all reasonable in some perspective.

In this presentation, I will first use a simple example of text data streams [1] to illustrate the meaning and feasibility of finding alternative clustering on stream data. Then, I will introduce two completely unsupervised techniques [2] to find two alternative clustering simultaneously on non-stream data. These methods would be immensely useful in the stream data domain because, with vast amounts of data arriving fast and continuously, it is infeasible to require some a-priori knowledge to search for the various alternate clusterings sequentially just as the COALA does. Finally I will talk about the existed method to compare clustering dissimilarity for stream data [3] and the quality measurement in stream clustering techniques [4], both of which are the main factors needed to be considered if we want to find alternative clustering on stream data.

For future work, I need to investigate generalizing the existed unsupervised algorithms to handle alternative clustering on stream data by combining the dissimilarity and quality together. Maybe you could give me some useful suggestions.

References

[1] Y. B. Liu, J. R. Cai, J. Yin, W. A. Fu. Clustering text data streams. Journal of Computer Science and Technology, 2008.

[2] P. Jain, R. Meka, and I. S. Dhillon. Simultaneous unsupervised learning of disparate clusterings. Proc. of SDM 2008.

[3] 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.

[4] F. Cao, M. Ester, W. Qian, and A. Zhou. Density-based clustering over an evolving data stream with noise. Proc. of SDM 2006.