Streaming Consciousness on Streaming Clustering

Jim Keller

Electrical Engineering and Computer Science

University of Missouri

kellerj@missouri.edu

Abstract

As one who has been involved in research and applications of clustering for many years, I've come to view the clustering enterprise through three basic questions.

1. Do you believe there are any clusters in your data?

2. If so, can you come up with a technique to find the natural grouping of your data?

3. Are the clusters you found good groupings of the data?

These questions have fueled many advances to both feature vector analytics and relational data analytics. Question1 probably draws the least attention since us clustering folk want to get about our business. However for example, some nice visualization techniques have been advanced to assist with this assertion. A side benefit of not skipping this aspect of the problem is that the methods to provide an idea of whether the data has natural clusters also give hints about the big question of how many clusters to search for. There are hundreds, perhaps thousands, of answers to question 2, and always room for more. Question 3 looks at the issue of cluster validity, usually optimizing the number of clusters to provide compact and well separated groups of data.

With the explosion of ubiquitous continuous sensing (something Lotfi Zadeh predicted as one of the pillars of Recognition Technology in the late 1990s), on-line streaming clustering is attracting more and more attention. I was drawn into this world mainly due to our desire to continuously monitor the activities, and health conditions, of older adults in a large interdisciplinary eldercare research group. Roughly, the requirements are that the streaming clustering algorithm recognize and adapt clusters as the data evolves, that anomalies are detected, and that new clusters are automatically formed as incoming data dictate. Several groups are building algorithms to perform on-line clustering. But, how do those requirements conform to the long-held trust in the three questions of clustering? The purpose of this talk is to examine (my thoughts on) these questions as they relate to streaming clustering. I chose to call it "streaming consciousness" to highlight that this is not a completely defined answer, but more a flow of thoughts about this overall area.