Incremental Generalized Eigenvalue Classification on Data Streams

Mario R. Guarracino, Salvatore Cuciniello, Davide Feminiano

High Performance Computing and Networking Institute
National Research Council, Italy

Abstract

As applications on massive data sets are emerging with an increasing frequency, we are facing the problem of analyzing the data as soon as they are produced. This is true in many fields of science and engineering: in high energy physics, experiments have been done to transfer data at a sustained rate of 150 gigabits per second. In Y2007, that speed will enable the delivery to users of data continuously produced by the LHC particle accelerator located at CERN. Other examples can be found in network traffic analysis, telecommunications data mining, discrimination of data from sensors that monitor pollution and biological hazards, video and audio surveillance. In all cases, computational procedures have to deal with a large amount of data that are delivered in form of data streams. Traditional data mining techniques assume that the dataset is static and, to increment knowledge, random samples are extracted from the dataset. In this study, we use Incremental Regularized Generalized Eigenvalue Classification (I-ReGEC), a supervised learning algorithm, to continuously train a classification model from a data stream. The advantage of this technique is that the classification model can be update incrementally. The algorithm online decides which are the points that contain new information and updates the available classification model. We show through numerical experiments, on a synthetic dataset, the method performance, highlighting its behavior with respect to the number of incremental training set, the accuracy classification and the throughput of the data stream.