Data Stream Mining

Tutorial by Michael Hahsler

Abstract:

Typical statistical and data mining methods (e.g., clustering, regression, classification and frequent pattern mining) work with “static” data sets, meaning that the complete data set is available as a whole to perform all necessary computations. Well known methods like k-means clustering, linear regression, decision tree induction and the APRIORI algorithm to find frequent itemsets scan the complete data set repeatedly to produce their results. However, in recent years more and more applications need to work with data which are not static, but are the result of a continuous data generating process which is likely to evolve over time. Some examples are web click-stream data, computer network monitoring data, telecommunication connection data, readings from sensor nets and stock quotes. These types of data are called data streams and dealing with data streams has become an increasingly important area of research. This tutorial will introduce data streams and methods used to manage and mine such data streams. Methods like sampling, moving windows, and data stream clustering and classification will be introduced and demonstrated using the R package stream.