



Spring 2008 CIS Colloquium Series

Memory-Constrained Data Mining

Slobodan Vucetic

March 26, 2008

Abstract

Data mining concentrates on the development of computationally efficient and accurate tools for knowledge discovery from very large data sets. Many successful data mining algorithms scale linearly or even better with data size and can learn with high accuracy from millions of examples on a regular PC. Interestingly, linear or sub-linear scaling is often not enough in a today's world of ever increasing data collections in sciences, medicine, engineering, and business at one end and increasing use of microcontrollers with highly limited memory and processing power at another end. A critical question in this new world is how to optimize learning from data that exceeds the available computational resources by orders of magnitude.

This talk will demonstrate that highly accurate learning could be achieved using seemingly inferior computational devices. To accomplish this, a computer is treated as a reservoir that sequentially observes a large data stream and, at any given moment, maintains a data summary and a prediction model that describes the data. After each observed stream example, the reservoir content is updated such that the accuracy is increased while the memory constraint is kept satisfied. The proposed approach prefers the examples on which the current predictor is the most uncertain. In addition, a data summary is maintained instead of the individual examples.

In this talk we will present an online support vector machine algorithm that requires constant memory and achieves sublinear runtime. The proposed algorithm was tested on a number of large data sets (some of them containing more than 1 million examples) from various domains, where it was assumed that the memory can maintain information about only 100 examples. In most cases, the achieved accuracy was near the upper bound achievable using unlimited resources. On the other hand, the accuracy was substantially larger than when randomly selected 100 examples were maintained in the reservoir. This result provides strong support for the hypothesis that innovative data mining methods can enable high-quality learning from large data sets using resource-constrained computing devices.

Location: 4th Floor Conference Room (Wachman 447)

Time: 3-4pm, Wednesday, March 26, 2008

Refreshments will be served!