November 2007

MADALGO seminar by Kevin Chang, Max Planck Institute

Multiple pass algorithms for model selection and other clustering problems

In this talk, I will present a framework for algorithms for learning statistical models of clustering in the streaming model of computation. The streaming model is a paradigm for the design of algorithms for massive data sets, in which the algorithm may make only a small number of sequential passes over the input while using a very small amount of working memory in order to accomplish the computational task at hand.

Our algorithms will consider the following statistical model for clustering, known as a mixture of distributions. We are given k different probability distributions F_1, ..., F_k, each of which is given a weight w_i >0. A point is drawn according to the mixture by choosing a distribution F_i with probability proportional to w_i and then choosing a point according to F_i. The data are then ordered arbitrarily and placed into a data stream and the algorithm's task is to learn the density function of the mixture.

I will present a multiple pass streaming algorithm that uses P passes over the data, where P is an input parameter to the algorithm. The memory requirement of the algorithm falls significantly as a function of P, showing a strong trade-off between the number of passes and memory required. Using communication complexity, this trade-off can be proved to be nearly tight, for a slightly stronger learning problem.

I will show how this framework can be adapted to solving clustering problems in combinatorial optimization.

Host: Lars Arge, MADALGO.