CCSU
DEPARTMENT OF MATHEMATICAL SCIENCES

COLLOQUIUM

Tuesday, May 8, 2007
12:30 – 1:30 PM
Maria Sanford, Room 101

Scalable Regularized K-Means Clustering with Probabilistic Support for High Dimensional Data

SAMIRAN GHOSH

INDIANA UNIVERSITY-PURDUE UNIVERSITY

Abstract

Over the last decade a variety of clustering algorithms have evolved. However one of the simplest (and possibly overused) partition based clustering algorithm is K-means. It can be shown that the computational complexity of K-means does not suffer from exponential growth with dimensionality rather it is linearly proportional with the number of observations and number of clusters. The crucial requirements are the knowledge of cluster number and the computation of some suitably chosen similarity measure. For this simplicity and scalability among large data sets, K-means remains an attractive alternative when compared to other competing clustering philosophies especially for high dimensional domain. However being a deterministic algorithm, traditional K-means have several drawbacks. It only offers hard decision rule, with no probabilistic interpretation. In this paper we have developed a decision theoretic framework by which traditional K-means can be given a probabilistic footstep. This will not only enable us to do a soft clustering, rather the whole optimization problem could be recasted into Bayesian modeling framework, in which the knowledge of cluster number could be treated as an unknown parameter of interest, thus removing a severe constrain of K-means algorithm. Our basic idea is to keep the simplicity and scalability of K-means, while achieving some of the desired properties of the other model based or soft clustering approaches.

For further information:

http://www.math.ccsu.edu/gotchev/colloquium/
gotchevi@ccsu.edu 860-832-2839
castanedan@ccsu.edu 860-832-2851