The University of Chicago
Department of Statistics

MASTER'S THESIS PRESENTATION

LI AN CHEN

Department of Statistics
The University of Chicago

An Empirical Study on Variants of k-Means Clustering

WEDNESDAY, February 16, 2011, at 11:00 AM
110 Eckhart Hall, 5734 S. University Avenue

ABSTRACT

Cluster analysis is a widely applied statistical data mining technique that is applied in various areas, within which k-means is one of the most mature methods. This paper will talk about its relations with EM and its underlying limitations; an improved k-means algorithm, x-means, will then be discussed. It is a highly scalable method compared with k-means and efficient in choosing number of clusters. The analysis will also incorporate an algorithm, ODIN (Outlier Detecting using Indegree Numbers) before executing x-means so as to eliminate random noises. The combined method will be used to analyze two sets of school data, comprising of numerical attributes, from which we will derive meaningful clusters and hence a valid structure of the data; comparison between the two datasets and the outcome using differing sets of attributes will be discussed and a proposal as for how the result can be used will be presented.

Information about building access for persons with disabilities may be obtained in advance by calling Sandra Romero at 773.702-0541 or by email (sandra@galton.uchicago.edu).