SAN DIEGO STATE UNIVERSITY
Leadership Starts Here

SDSU
presents
a thesis defense for
Master of Science
degree in
Computer Science

Tuesday,
December 2, 2014

10:00am
GMCS 405

Matthew Shaw

K-Means Clustering with Automatic Determination of K Using a Multi-Objective Genetic Algorithm with Applications to Microarray Gene Expression Data

Abstract

As the role of large scale data analysis continues to expand, the task of extracting useful information becomes ever more important. Clustering, the task of grouping together data points that share similar features, provides a way to present high-dimensional data in a format that can be more easily comprehended by humans, while also allowing inferences to be drawn about previously unseen data points based on the known characteristics of other points in the same cluster. Over the years several techniques have been developed for clustering data. While effective, most of these algorithms require that the number of clusters to partition the data into be known in advance. In situations where the domain is well understood, this requirement is a minimal burden, but this becomes problematic when little is known about the data being analyzed. The work that follows investigates using a Multi-objective Genetic Algorithm to discover an optimal number of clusters (K) to partition the data into while simultaneously finding high quality clustering solutions. The genetic algorithm finds the most appropriate value of K for the data set, and the approach does not depend on the underlying data. As a proof of concept, this algorithm is applied to clustering microarray gene expression data.

Thesis Committee

Robert Edwards, Thesis Chair, Department of Computer Science
Roger Whitney, Department of Computer Science
Peter Salamon, Department of Mathematics & Statistics