SPANNING TREES AS TOOLS FOR DATA ANALYSIS

By

Adam Petrie

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: Decision Sciences and Engineering Systems

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Thomas R. Willemain, Thesis Adviser

Mark J. Embrechts, Member

John E. Mitchell, Member

Malik Magdon-Ismail, Member

Rensselaer Polytechnic Institute
Troy, New York

April 2007
(For Graduation May 2007)

ABSTRACT

We develop new methods to describe high-dimensional data based on the graph-theoretic concept of a spanning tree. The minimum spanning tree (MST), longest spanning tree (LST), and short Hamiltonian path (the “snake”) are graphs that can be constructed on a dataset using Euclidean distances as edge lengths. These graphs require computational time that grows at worst linearly in the number of dimensions and quadratically with the number of datapoints. The node degree and edge length distributions of these trees serve as easy-to-visualize 1-dimensional summaries of multidimensional data, and statistics derived from these distributions can characterize key properties of the data.

Non-asymptotic theoretical properties of spanning trees are extremely difficult to derive analytically, so Monte Carlo simulations are necessary to examine the average edge length, node degree, and topological properties of trees with finite numbers of nodes. Building MSTs and LSTs on random data with up to a million points, we report the results of extensive simulations regarding the mean, standard deviation, and normality of node degree frequencies; the maximum node degree; the distribution of the longest edge length; and the mean, standard deviation, and normality of the total length of the tree, and compare with asymptotic results in the literature. Further, we analyze the correlation structure and run length distribution of snakes on uniform data.

We also present five new tree-based methods for the analysis of high-dimensional datasets. The maximum node degree in the LST is used for outlier detection. The standard deviation of the node degree distribution and a linear combination of the first three node degree frequencies are used to identify the implicit dimension of a dataset. The topology of the MST is used to find the center of the tree and a corresponding robust estimate of location. The topology of the MST is also used in the analytical derivation of the multisample generalization of the Friedman and Rafsky (1979) statistic for the multivariate two-sample test. Finally, run length statistics based on the sequence of segment lengths in the snake provide a test for multivariate uniformity almost regardless of the domain or shape of the data cloud.