UNIVERSITY OF SOUTH FLORIDA

Defense of a Master's Thesis

Feature Selection via Random Subsets of Uncorrelated Features
by
Long Dang
For the MSCS degree in Computer Science

The role of feature selection is crucial in many applications. In biology, gene expression micro array data sets have been used extensively in many areas of research. These data sets typically suffer from an important problem: the ratio between the number of features over the number of examples is very high. This problem mainly affects prediction accuracy because it is best to collect more labeled examples than features. A correlation based random subspace ensemble feature selector (CCC_RSM) was proposed to handle this problem. In this approach, first it determines the most relevant prediction features. Next, it groups these features based on their correlation to each other. Then, a feature is randomly chosen from each correlated group so that the selected features form a feature subset. The CCC_RSM algorithm repeats the previous step a pre-defined number of times. The proposed algorithm's performance is evaluated by combining either multiple decision trees or Support Vector Machines. This study focuses on finding alternative feature selectors in the first step so that the CCC_RSM algorithm can obtain good, or even better classification performance. We used four micro array gene expression data sets in the experiments. Based on the original algorithm, we used the Gini Index in place of Relief-F. Consequently, the alternative method gave the highest F-measure score for the Leukemia (1.00), Breast (0.98), Colon (0.91) and CNS (0.81) data sets.

Examining Committee

Lawrence Hall, Ph.D., Major Professor
Dmitry B. Goldgof, Ph.D.
Yu Sun, Ph.D.

Friday, September 18, 2020
9.30 am
Online (Microsoft Teams)
Please email for more information
longdang@usf.edu

THE PUBLIC IS INVITED

Robert Bishop, Ph.D.
Dean, College of Engineering

Dwayne Smith, Ph.D.
Dean, Office of Graduate Studies