Announcing the Final Examination of Orestis Panagopoulos for the degree of Doctor of Philosophy

Time & Location: June 3, 2016 at 10:00 AM in Engineering II 312
Title: DATA MINING MODELS FOR TACKLING HIGH DIMENSIONAL DATASETS AND OUTLIERS

High dimensional data and the presence of outliers in data each pose a serious challenge in supervised learning.

Datasets with significantly larger number of features compared to samples arise in various areas, including business analytics and biomedical applications. Such datasets pose a serious challenge to standard statistical methods and render many existing classification techniques impractical. The generalization ability of many classification algorithms is compromised due to the so-called curse of dimensionality. A new binary classification method called constrained subspace classifier (CSC) is proposed for such high dimensional datasets. CSC improves on an earlier proposed classification method called local subspace classifier (LSC) by accounting for the relative angle between subspaces while approximating the classes with individual subspaces. CSC is formulated as an optimization problem and can be solved by an efficient alternating optimization technique. Classification performance is tested in publicly available datasets. The improvement in classification accuracy over LSC shows the importance of considering the relative angle between the subspaces while approximating the classes. Additionally, CSC appears to be a robust classifier, compared to traditional two step methods that perform feature selection and classification in two distinct steps.

Outliers can be present in real world datasets due to noise or measurement errors. The presence of outliers can affect the training phase of machine learning algorithms, leading to over-fitting which results in poor generalization ability. A new regression method called relaxed support vector regression (RSVR) is proposed for such datasets. RSVR is based on the concept of constraint relaxation which leads to increased robustness in datasets with outliers. RSVR is formulated using both linear and quadratic loss functions. Numerical experiments on benchmark datasets and computational comparisons with other popular regression methods depict the behavior of our proposed method. RSVR achieves better overall performance than support vector regression (SVR) in measures such as RMSE and adjusted coefficient of determination while being on par with other state-of-the-art regression methods such as robust regression (RR). Additionally, RSVR provides robustness for higher dimensional datasets which is a limitation of RR, the robust equivalent of ordinary least squares regression. Moreover, RSVR can be used on datasets that contain varying levels of noise.

We also present the development of novel models that deal with the problem of one-class classification in the presence of outliers as well as the execution of feature selection and dimensionality reduction in high dimensional noisy data.

Major: Industrial Engineering

Educational Career:
Bachelor's of Production Engineering and Management, BS, 2010, Technical University of Crete
Master's of Industrial and Systems Engineering, MS, 2014, University of Florida

Committee in Charge:
Petros Xanthopoulos, Chair, Industrial Engineering and Management Systems
Luis Rabelo , Industrial Engineering and Management Systems
Qipeng Phil Zheng , Industrial Engineering and Management Systems
Damian Dechev , Computer Science

Approved for distribution by Petros Xanthopoulos, Committee Chair, on May 20, 2016.

The public is welcome to attend.