Building accurate classifiers for predicting group membership is made difficult when data is skewed or imbalanced which is typical of real world data sets. The classifier has the tendency to be biased towards the over represented group as a result. This imbalance is considered a class imbalance problem which will induce bias into the classifier particularly when the imbalance is high.

Class imbalance data usually suffers from data intrinsic properties beyond that of imbalance alone. The problem is intensified with larger levels of imbalance most commonly found in observational studies. Extreme cases of class imbalance are commonly found in many domains including fraud detection, mammography of cancer and post term births. These rare events are usually the most costly or have the highest level of risk associated with them and are therefore of most interest.

To combat class imbalance the machine learning community has relied upon embedded, data processing and ensemble learning approaches. Exploratory research has linked several factors that perpetuate the issue of misclassification in class imbalanced data. However, there remains a lack of understanding between the relationship of the learner and imbalanced data among the competing approaches. The current landscape of data processing approaches have appeal due to the ability to divide the problem space which allows for simpler models. However, most of these approaches have little theoretical bases although in some cases there is empirical evidence supporting the improvement.

The main goals of this research is to introduce newly proposed a priori based sampling methods that improve concept learning within class imbalanced data. The results in this work highlight these techniques performance within publicly available data sets from different domains containing various levels of imbalance. In this research the theoretical and empirical reasons are explored and discussed.

Major: Modeling and Simulation

Educational Career:
Bachelor's of Computer Information Systems, BS, 2012, DeVry
Master's of Computer Science, MS, 2013, University of West Florida
Master's of Modeling and Simulation, MS, 2015, University of Central Florida

Committee in Charge:
Petros Xanthopoulos, Chair, Industrial Engineering & Management Systems (IEMS)
R. Paul Wiegand, Institute for Simulation and Training (IST)
J. Peter Kincaid, Institute for Simulation and Training (IST)
Waldemar Karwowski, Industrial Engineering & Management Systems (IEMS)

Approved for distribution by Petros Xanthopoulos, Committee Chair, on October 6, 2015.

The public is welcome to attend.