Announces the Ph.D. Dissertation Defense of

Robert K.L. Kennedy

for the degree of Doctor of Philosophy (Ph.D.)

"Novel Techniques for Handling Imbalanced Data with Unsupervised Methods"

December 2nd, 2024, 10:30 a.m.
In-person Room EE-405

DEPARTMENT:
Electrical Engineering and Computer Science

ADVISOR:
Taghi M. Khoshgoftaar, Ph.D.

Ph.D. SUPERVISORY COMMITTEE:
Taghi M. Khoshgoftaar, Ph.D., Chair
Imadeldin Mahgoub, Ph.D.
Mehrdad Nojoumian, Ph.D.
DingDing Wang, Ph.D.

ABSTRACT OF DISSERTATION

In the modern data landscape, vast amounts of unlabeled data are continuously generated, necessitating development of robust unsupervised techniques for handling unlabeled data. This is the case for fraud detection and healthcare sectors analyses, where data is often significantly imbalanced. This dissertation focuses on novel techniques for handling imbalanced data, with specific emphasis on a novel unsupervised class labeling technique for unlabeled fraud detection datasets and unlabeled cognitive datasets. Traditional supervised machine learning relies on labeled data, which is often expensive and difficult to create, particularly in domains requiring expert input. Additionally, such datasets suffer from challenges associated with class imbalance, where one class has significantly fewer examples than another, complicating model training and significantly reducing performance. The primary objectives of this dissertation include developing a novel unsupervised cleaning method and an innovative unsupervised class labeling method. We validate and evaluate our methods across various datasets, which include two Medicare fraud detection datasets, a credit card fraud detection dataset, and three datasets used for detecting cognitive decline.

Our unique approach involves using an unsupervised autoencoder to learn from the datasets' features and synthesize labels. Primarily targeting imbalanced datasets, but still effective for balanced datasets, our method calculates an error metric for each instance. This metric is used to distinguish between fraudulent and legitimate cases, allowing us to assign a binary class label. To further improve label generation, we integrate an unsupervised feature selection method that ranks and identifies the most important features without using class labels. This approach aims to enhance the quality of the synthesized class labels, simplify models, and reduce computational costs, which is well suited in large highly imbalanced datasets such as the Medicare fraud and credit card fraud detection datasets. Our novel techniques only use the datasets'

Born in Florida, USA