FAU COLLEGE OF ENGINEERING AND COMPUTER SCIENCE
FLORIDA ATLANTIC UNIVERSITY

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’ features for the labeling process, allowing for this research to adhere to the unsupervised paradigm. We detail empirical results and statistical analyses that demonstrate substantial improvements in label accuracy. These findings show the efficacy of our iterative cleaning method and class label synthesis method on unlabeled data, including the incorporation of unsupervised feature selection.

BIOGRAPHICAL SKETCH

Born in Florida, USA
B.S., Florida Atlantic University, Boca Raton, Florida 2014
M.S., Florida Atlantic University, Boca Raton, Florida 2018
Ph.D., Florida Atlantic University, Boca Raton, Florida 2024

CONCERNING PERIOD OF PREPARATION
& QUALIFYING EXAMINATION

Time in Preparation: 2019- 2024

Qualifying Examination Passed: Spring 2019

Published Papers:

Kennedy, Robert KL, Taghi M Khoshgoftaar, Flavio Villanustre, and Timothy Humphrey. "A parallel and distributed stochastic gradient descent implementation using commodity clusters." *Journal of Big Data* 6.1 (2019): 16.

Kennedy, Robert KL, and Taghi M. Khoshgoftaar. "Accelerated deep learning on hpcc systems." *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2020.

Kennedy, Robert KL, and Taghi M. Khoshgoftaar. "An Examination of Neural Networks on Cluster Computers." *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2021.

Kennedy, Robert KL, Justin M. Johnson, and Taghi M. Khoshgoftaar. "The effects of class label noise on highly-imbalanced big data." *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2021.

Kennedy, Robert KL, Zahra Salekshahrezaee, and Taghi M. Khoshgoftaar. "A novel approach for unsupervised learning of highly-imbalanced data." *2022 IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2022.

Kennedy, Robert KL, Zahra Salekshahrezaee, Flavio Villanustre, and Taghi M Khoshgoftaar. "Iterative cleaning and learning of big highly-imbalanced fraud data using unsupervised learning." *Journal of Big Data* 10.1 (2023): 106.

Johnson, Justin M., Robert KL Kennedy, and Taghi M. Khoshgoftaar. "Learning from Highly Imbalanced Big Data with Label Noise." *International Journal on Artificial Intelligence Tools* 32.5 (2023).

Kennedy, Robert KL, Zahra Salekshahrezaee, and Taghi M. Khoshgoftaar. "Unsupervised anomaly detection of class imbalanced cognition data using an iterative cleaning method." *2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2023.

Kennedy, Robert KL, and Taghi M. Khoshgoftaar. "A Novel Approach to Synthesize Class Labels in Highly Imbalanced Large Data." *2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2023.

Kennedy, Robert KL, Flavio Villanustre, Taghi M Khoshgoftaar, and Zahra Salekshahrezaee. "Synthesizing class labels for highly imbalanced credit card fraud detection data." *Journal of Big Data* 11.1 (2024): 38.

Kennedy, Robert KL, and Taghi M. Khoshgoftaar. "Synthesizing Class Labels for Balanced and Highly Imbalanced Cognition Data." *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2024.

Kennedy, Robert KL, and Taghi M. Khoshgoftaar. "Impact of Class Imbalance on Unsupervised Label Generation for Medicare Fraud Detection." *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2024.

Kennedy, Robert KL, Flavio Villanustre, and Taghi M Khoshgoftaar. " Unsupervised Feature Selection and Class Labeling for Credit Card Fraud." *Journal of Big Data* (under review 2024).