A Technique For Improving Classification Accuracy Of Highly Imbalanced And Sparse Datasets

Sindhura Bonthu, Graduate Student, Computer Science
Faculty Mentor: Dr. Qiuming Zhu, Computer Science

Over recent years, Big data has gained significance in various fields like medicine, e-commerce, banking etc., due to the useful information that could be derived from the collected datasets. Incorporating machine learning in big data analytics helps to build models useful for understanding the implications of the data and predicting future trends. The data collected from different sources have different characteristics that could have either positive or negative effects on the accuracy and efficiency of the machine learning process. For example, one of the negative characteristics is the data sparseness. Data is said to be sparse when we have multiple classes within the dataset, but the majority of the data is biased to belong in one direction of the classes. In such cases, the model resulting from a machine learning process could be a one that misclassifies the data and makes wrong predictions. In order to handle the imbalanced data properly, this research will make use of resampling techniques such as over-sampling and under-sampling. These resampling strategies will be combined with different machine learning algorithms to compare the performance metrics of data models with different neural network architectures.