World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:01, 2018

An Adaptive Oversampling Technique for Imbalanced Datasets

Authors : Shaukat Ali Shahee, Usha Ananthakumar

Abstract : A data set exhibits class imbalance problem when one class has very few examples compared to the other class, and this is also referred to as between class imbalance. The traditional classifiers fail to classify the minority class examples correctly due to its bias towards the majority class. Apart from between-class imbalance, imbalance within classes where classes are composed of a different number of sub-clusters with these sub-clusters containing different number of examples also deteriorates the performance of the classifier. Previously, many methods have been proposed for handling imbalanced dataset problem. These methods can be classified into four categories: data preprocessing, algorithmic based, cost-based methods and ensemble of classifier. Data preprocessing techniques have shown great potential as they attempt to improve data distribution rather than the classifier. Data preprocessing technique handles class imbalance either by increasing the minority class examples or by decreasing the majority class examples. Decreasing the majority class examples lead to loss of information and also when minority class has an absolute rarity, removing the majority class examples is generally not recommended. Existing methods available for handling class imbalance do not address both between-class imbalance and within-class imbalance simultaneously. In this paper, we propose a method that handles between class imbalance and within class imbalance simultaneously for binary classification problem. Removing between class imbalance and within class imbalance simultaneously eliminates the biases of the classifier towards bigger sub-clusters by minimizing the error domination of bigger sub-clusters in total error. The proposed method uses model-based clustering to find the presence of sub-clusters or sub-concepts in the dataset. The number of examples oversampled among the sub-clusters is determined based on the complexity of sub-clusters. The method also takes into consideration the scatter of the data in the feature space and also adaptively copes up with unseen test data using Lowner-John ellipsoid for increasing the accuracy of the classifier. In this study, neural network is being used as this is one such classifier where the total error is minimized and removing the between-class imbalance and within class imbalance simultaneously help the classifier in giving equal weight to all the sub-clusters irrespective of the classes. The proposed method is validated on 9 publicly available data sets and compared with three existing oversampling techniques that rely on the spatial location of minority class examples in the euclidean feature space. The experimental results show the proposed method to be statistically significantly superior to other methods in terms of various accuracy measures. Thus the proposed method can serve as a good alternative to handle various problem domains like credit scoring, customer churn prediction, financial distress, etc., that typically involve imbalanced data sets.

Keywords : classification, imbalanced dataset, Lowner-John ellipsoid, model based clustering, oversampling

Conference Title : ICADMA 2018 : International Conference on Advanced Data Mining and Applications

Conference Location : Sydney, Australia

Conference Dates : January 29-30, 2018

Open Science Index, Computer and Information Engineering Vol:12, No:01, 2018 publications.waset.org/abstracts/83833.pdf