Identifying clusters structure of rare events using random forest clustering

ABSTRACT

Given highly imbalanced data, most learning algorithms faced the challenge to accurately predict rare events, while such cases were the ones that carry importance and useful knowledge. In a binary class label dataset, the rare events are the ones in the minority class. This study used a stroke dataset with a binary class label and the class imbalance ratio was 54:1. In addition to that, the dataset contained missing values and mixed data types. To identify the intrinsic structures in the minority class (the stroke group), Random Forest Clustering was used to produce the proximity matrix and fed to Partition around Medoid (PAM) clustering method to identify the optimal number of clusters. The proximity plot seems to show there could be cluster tendency and k=2 was identified to be the best as compared to k=3 to k=5. Based on the internal cluster validation, however, the silhouette coefficient width was small (0.1) indicating much of the data objects were within the other boundary of the other class. We have suggested a further investigation plan in this paper for the next action.