Learning class-imbalanced problems from the perspective of data intrinsic characteristics

Kong, J.

Summary

The class-imbalance problem is a challenging classification task and is frequently encountered in real-world applications. Various techniques have been developed to improve the imbalanced classification performance theoretically and practically. Apart from developing new approaches, researchers also address the importance of understanding the data itself, which will provide more insight into what actually hinders the imbalanced classification performance.

This thesis conducted research on Learning Class-Imbalanced Problem from the perspective of Data Intrinsic Characteristics. The empirical investigation comparing several data-level algorithms shows that oversampling approaches considering the minority class distribution can provide better imbalanced classification performance in most cases. Although data complexity measures cannot provide any guidance on the choice of resampling techniques, we find the potential best AUC value can be predicted by the F1v measure (the Directional-vector Maximum Fisher's Discriminant Ratio). Both conclusions are also verified on a real-world inspired vehicle mesh dataset from Honda Research Institute

Hyperparameter optimisation has shown great effectiveness for many machine learning classification algorithms. For example, the maximum depth of the tree and the minimum number of samples required to split an internal node are critical for tuning the Decision Tree to achieve the best performance. Therefore, we emphasize the importance of hyperparameter tuning for data-level approaches.

The anomaly detection problem is a class-imbalance problem with an extreme imbalanced ratio. Techniques for anomaly detection problems can be applied to class-imbalanced problems with fine adjustment. In this thesis, we propose to introduce the Local Outlier Score, which is an important indicator to evaluate whether a sample is an outlier, as an additional attribute of the original imbalanced

dataset. This proposal is more than borrowing the knowledge from anomaly detection research field but also provides researchers with another possibility to acquire more insight from the data rather than undersampling/oversampling.

In the final part of the thesis, an improved sample type identification is proposed for dealing with multi-class imbalanced classification and applied on a real-world surface defects dataset from TATA Steel. Meanwhile, we address the importance of understanding the different data intrinsic characteristics for binary and multi-class scenarios.