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

Kong, J.

# CHAPTER 7

# Conclusions

As class-imbalance problems attract attention from academic and industrial fields, many approaches have been proposed to improve the imbalanced classification theoretically and practically. In this thesis, we mainly conducted research on *Learning Class-Imbalanced Problems from the Perspective of Data Intrinsic Characteristics*. In the following, Section 7.1 first summarizes the main contributions of the thesis as the answers to the research questions in the Introduction chapter. Then, the strengths and weaknesses of the research work (Chapters 3-6) are also discussed following the chapter order. Finally, the outlook on future research is provided in Section 7.2.

## 7.1 Summary

**Chapter 1** introduced the scientific background and the motivation of this thesis. It showed the existence of class imbalance problems in real-world applications and emphasized the importance of learning from imbalanced data. Then, the outline of the thesis and relevant publications were given.

**Chapter 2** provided the necessary literature review. It started with the visualisation of a binary class imbalance problem. After that, the methods and the performance metrics for both binary and multi-class class imbalance scenarios were presented. Furthermore, the studies on the data complexity in the imbalanced learning domain were introduced in detail. Finally, the imbalanced benchmark datasets and the real-world imbalanced application work were reviewed.

**Chapter 3** investigated the effectiveness of several oversampling techniques, where the new ones (RACOG, wRACOG and RWO-sampling) take into account the minority class distribution, while the “classic" ones (SMOTE, ADASYN and MWMOTE) do not. These oversampling techniques were experimented with 19 benchmark datasets and our real-world inspired vehicle dataset. The experimental results first answered **research question 1**: In most cases, oversampling approaches considering the minority class distribution perform better. Different data complexity measures were taken into account with the original aim to answer **research question 2**. According to our experimental results, no apparent relationship between data complexity measures and the choice of resampling techniques can be derived. One noteworthy finding is that the F1v value strongly correlates with the potential best AUC value (after resampling).

Although “new" oversampling approaches showed effectiveness over “classical" ones in most cases, one main practical limitation must be taken into account. Due to the fact that “new" oversampling techniques consider the minority class distribution, implementing these techniques often requires more time compared to “classical" ones. When facing large datasets, huge time costs are inevitable. Therefore, the trade-off between performance improvement and time consumption must be considered while using “new" oversampling techniques.

**Chapter 4** introduced our work on hyperparameter optimisation on class-imbalance problems. Both hyperparameters in resampling techniques and classification algorithms were optimised in our experiments. Further exploration of how data complexity affects the classification improvement yielded via hyperparameter optimisation answered our **research question 3**. Applying hyperparameter optimisation for both classification algorithms and resampling approaches can significantly improve the performance of imbalanced datasets with low class overlap. However, oversampling techniques and hyperparameter optimisation do not improve performance for imbalanced datasets with high class overlap.

Despite the fact that hyperparameter optimisation improves the classification performance significantly for imbalanced datasets with low class overlap, the optimisation process always involves hundreds to thousands of iterations. The additional time consumption is significant. Moreover, different resampling techniques contain different hyperparameters. Therefore, one needs to have

an in-depth understanding of the resampling techniques in order to set the hyperparameters that need to be optimised.

**Chapter 5** conducted research on improving imbalanced classification via adding additional attributes. We proposed introducing the outlier score and four types of samples as two additional attributes of the original imbalanced datasets. We compared the classification performance of our proposed method and the resampling techniques in the literature and concluded that adding additional attributes in most cases produces significantly better or competitive classification performance. This naturally leads to the answer of **research question 4**: we can take advantage of anomaly detection techniques to improve the imbalanced classification.

We must consider the following points when using our proposed method to improve the imbalanced classification. Firstly, we have shown that the proposed attribute “type" highly correlates with the class labels under certain circumstances. Hence, we recommend choosing feature-insensitive classification algorithms when implementing our proposed method. Furthermore, considering the fact that anomaly detection problems are imbalanced problems with extreme imbalance ratios, it is recommended to add the outlier score as an additional attribute when the imbalance ratio is relatively high (no less than 5).

**Chapter 6** presented our improved sample type identification for multi-class imbalanced classification. We showed the drawbacks of the existing identification rule in multi-class scenarios, (i) a higher percentage of unsafe samples in minority classes and (ii) the false identification of outliers. The proposed rule answered the **research question 5**, we can improve the sample identification by adjusting $k$ according to the imbalance ratio and considering neighborhood information of the neighbors. The proposed approach was tested on a challenging real-world problem, the steel surface defects detection task. The experimental results answered **research question 6**, showing the industrial applicability of our method.

We used two performance metrics to evaluate the experimental results, *MinAcc* for assessing the performance of minority class(es) and *MAUC* for measuring the overall performance of all classes. According to our experimental setup, the proposed identification rule significantly better classifies minority class(es) while producing competitive overall classification performance. Hence, one main

limitation of the proposed method is that a significant better performance cannot be guaranteed if the given task only focuses on the overall performance.

## 7.2 Future Work

This thesis mainly conducted the research on *Learning Class-Imbalanced Problems from the Perspective of Data Intrinsic Characteristics*. Despite the achievements presented here that have revealed interesting insights, learning the data intrinsic characteristics in imbalanced datasets and how to efficiently use these characteristics to obtain guidance on choosing the imbalanced techniques still need to be completed. Furthermore, much work is yet to be done to apply the class imbalance techniques to handle complex real-world scenarios. Several possible future research directions for extending the work in this thesis are discussed as follows.

**Software Tool for Learning from Class Imbalance Datasets** As we have shown in this thesis, the class imbalance problem has been studied extensively from different aspects, including data interpolation, algorithm adjusting, cost-sensitive learning, data complexity etc. Given a class imbalance problem, one has to try several techniques and choose the best one for the specific situation. However, these techniques are available in different languages, *Python*, *R* and *C*, which makes it challenging for researchers to implement and compare. Therefore, software with the following functions would greatly contribute to the community.

- Main class-imbalance techniques, e.g. various resampling techniques, different algorithm-level approaches, cost-sensitive learning approaches and ensemble learning methods.
- Efficient hyperparameter optimisation algorithms to choose the optimal combination of hyperparameters.
- Data complexity analysis to provide some algorithm selection insights.
- Several benchmark examples to help beginners understand the functions of the software.

**Comparative Study of Anomaly Detection and Class Imbalance Problem** The anomaly detection problem can be considered a class imbalance problem with an extreme imbalance ratio. In this thesis, we proposed to add the *Local Outlier Score* as an additional attribute to gain more information for the original imbalanced dataset. In future work, other anomaly detection techniques, such as the clustering-based local outlier score (CBLOF) (Z. He, Xu, and Deng, 2003) and histogram-based outlier score (HBOS) (Goldstein and Dengel, 2012) could be included in the analysis. It is also interesting to explore other potential attributes to be added.

**Data Complexity in Real-Time Processing** In this thesis we mainly focused on stationary imbalanced datasets, whereas in many applications, such as fault diagnosis and bank commercial monitoring systems (H. M. Nguyen, Cooper, and Kamei, 2011), the data is constantly arriving and real-time analysis must be given. This scenario refers to the topic of *Online Class Imbalance Learning from Imbalanced Data Streams* (Fernández, García, Herrera, and Chawla, 2018; M. Last, 2002), which combines the difficulties of data stream mining and class imbalance problems (Fernández, García, Galar, Prati, Krawczyk, and Herrera, 2018; S. Wang, Minku, and Yao, 2014). In this type of problem, the new learning instances arrive in a time-based manner and the class distribution is dynamic. The imbalance ratio may evolve over time, making the relationship dynamic so that the algorithms with fixed imbalance ratio assumptions are not valid anymore. For example, when the imbalanced problem evolves into a balanced problem, it will lead to the failure of the previous imbalanced algorithm. When the majority class evolves into the minority one (or vice versa), the algorithm may even bring more imbalance bias to the problem. Thus, analysing the data complexity dynamically and adjusting the approaches accordingly would significantly benefit applications.