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

Kong, J.

CHAPTER 1

Introduction

1.1 Background

Machine learning is a broad multidisciplinary research area which is based on many different branches of mathematics and science, including computer science, statistics, cognitive psychology, engineering, and optimisation theory (Soofi and Awan, 2017). Machine learning algorithms aim to learn from the data and build the models in order to make predictions on unseen samples. These algorithms have numerous applications, for instance, medical diagnosis (Acharya, Chowriappa, Fujita, Bhat, S. Dua, Koh, Eugene, Kongmebhol, and K. Ng, 2016), product recommendations (Misra, Wan, and McAuley, 2018), defect detection (Haddad, S. Yang, Karam, Ye, Patel, and Braun, 2018), video surveillance (Radtke, Granger, Sabourin, and Gorodnichy, 2014), computer vision (X. Zhang, Zhuang, W. Wang, and Pedrycz, 2018) and self-driving cars (Carranza-García, Lara-Benítez, García-Gutiérrez, and Riquelme, 2021).

Supervised and unsupervised learning are the two main sub-categories of machine learning algorithms. Supervised learning algorithms are trained with a given set of samples with input attributes and outputs, and the goal is to find the relationship between the input attributes and the output responses. On the other hand, no output responses are given when training the unsupervised learning algorithms, and the goal is to discover the structure of the data itself, for example clustering the data points.

Classification and regression problems are the two essential branches of supervised learning. Classification is an important research area in the field of machine learning and data mining. Classification predictive modelling refers to the task of estimating the mapping function from input attributes to discrete class labels. In other words, training a classification model is to predict the class labels of samples with given input attributes. For example, spam email filtering can be regarded as a two-class classification task, where the two class labels are *spam* and *not spam*. Regression predictive modelling refers to the task of estimating the mapping function from input attributes to continuous outputs, i.e. a regression predictive model returns a quantity. For example, predicting the house price based on the size, location, age, and other aspects of the house is a regression task, where these aspects of the house are the input attributes, and the house price is the continuous output.

Several machine learning algorithms have been proposed in the literature to handle classification tasks, for instance, logistic regression, decision tree, random forest, support vector machine, and k-Nearest Neighbour (KNN). Most of the proposed algorithms are designed under two main assumptions:

- The classes are equally distributed;
- The cost of classification errors, so-called misclassification costs, are equal.

However, both assumptions do not always hold in real-world applications. Many real-world classification problems suffer from significant differences in the number of samples in each class. Moreover, the classification costs in these real-world classification problems cannot be treated equally. Taking an example of cancer diagnosis, a cancer case is much less likely to occur than healthy cases, i.e. samples in the *cancer* class and *healthy* class are not equally frequent. In this problem, the rare samples (cancer) are more important, and their classification costs are higher. Failure to identify a cancer case will lead to a person's life loss. Such problems where one or more classes are underrepresented are known as *class-imbalance problems*. Standard classifiers aim to maximise the overall accuracy, i.e. the proportion of all correctly classified validation samples in percentage, and always perform poorly on such problems. For example, if a standard classifier is used to handle a cancer diagnosis task with 95 healthy cases and 5 cancer cases, even if it would classify all samples into *healthy* class and still achieve 95% accuracy overall. However, the classification accuracy on *cancer* class is 0%, which makes the classifier useless in practice.

Class-Imbalance Learning Strictly speaking, any dataset with an unequal class distribution can be considered imbalanced. However, in the imbalanced domain, a dataset is defined as imbalanced only when the samples in different classes have a significant or even extreme gap in the number of samples (Fernández, García, Galar, Prati, Krawczyk, and Herrera, 2018). In other words, one or more classes significantly outnumber the other class(es) in an imbalanced dataset. The classes with more samples are called *majority* classes, while the underrepresented classes are called *minority* classes. The skewed distribution in the dataset will make the classifiers biased toward the majority class(es) and the minority class(es) to be overlooked. However, from the application point of view, the underrepresented (minority) class is usually the class of interest and has higher misclassification costs in the problem. For example, it is more important to correctly identify the minority class samples in medical diagnosis, email filtering, and fault detection. The price of misclassifying the minority samples would be a massive loss of money in fault diagnosis, an unqualified product in anomaly detection and a person's life in medical diagnosis. In contrast, the misclassification cost on majority class samples is only a double check. Hence, it is of vital importance to study class-imbalance problems.

Class-imbalance problems have caught growing attention from both academic and industrial fields. Many techniques have been developed to alleviate the influence of *class imbalance* and can be categorized into four broad groups.

1. *Data-level* approaches rebalance the class distribution directly via resampling the data space. These approaches manipulate the data directly and are easy to implement in real-world applications as a preprocessing step.

2. *Algorithm-level* approaches adapt the classification algorithms to force the learning bias toward the minority class. However, the adjustments always require a deep understanding of the corresponding algorithms, i.e. the adjustments are algorithm-specific.

3. *Cost-sensitive* learning techniques handle the class-imbalance problems by considering the unequal misclassification costs. This technique can be combined with *data-level* and *algorithm-level* approaches.

4. *Ensemble-based* methods in imbalanced learning domain usually combine an ensemble learning algorithm and one of the approaches above.

Data Complexity in Imbalanced Learning Domain The *class imbalance* was widely considered as the determinant for performance degradation. However, researchers have observed that in some cases, good classification performances can be achieved even in the presence of significant class imbalance for problems with low complexity, such as a linearly separable problem. This suggests that the class imbalance itself cannot be considered as the main reason for the performance degradation, and naturally extends the research direction to the study of data complexity. The difficulty of a supervised classification problem can be characterized by several data complexity measures in the literature, including (i) *feature overlapping measures*; (ii) *measures of the separability classes*; and (iii) *measures of geometry, topology and density of manifolds*. These measures are used to gain insights into the performance of data-level approaches in (Weng and Poon, 2006). The relationship between data complexity measures and imbalanced classification performance is studied in (M. S. Santos, Soares, Abreu, Araujo, and J. Santos, 2018). Small disjuncts are problematic under class-imbalance situations because they are difficult to distinguish from noise (Jo and Japkowicz, 2004; López, Fernández, García, Palade, and Herrera, 2013). It is also interesting to study the samples as different types: *safe*, *borderline*, *rare* samples and *outliers* according to their local characteristics (Napierala and Stefanowski, 2016) or their distances to the decision boundary (Kubat, Matwin, et al., 1997).

Motivated by the studies on *Data Complexity in Imbalanced Learning Domain* (López, Fernández, García, Palade, and Herrera, 2013; Luengo, Fernández, García, and Herrera, 2011; Prati, Batista, and Monard, 2004; M. S. Santos, Soares, Abreu, Araujo, and J. Santos, 2018; Weng and Poon, 2006), as well as the study showing that "*Better Data*" *is Better than* "*Better Data Miners*" (Agrawal and Menzies, 2018), this thesis mainly conducts the research on *Learning Class-Imbalanced Problems from the Perspective of Data Intrinsic Characteristics*.

1.2 Research Questions

In this thesis we focus on learning class-imbalanced problems from the perspective of data intrinsic characteristics. To achieve our objective, the following research questions are considered.

RQ1: What is the difference between the effectiveness of “classical” and “new” resampling techniques?

Resampling techniques have been proven effective in handling class-imbalance problems, and many resampling techniques have been proposed in the literature. However, most empirical studies and application work still focus on “classical" resampling techniques and do not take newly developed ones into account. Distinguishing the oversampling techniques into “classical" and “new" and studying their effectiveness will provide researchers with insights on choosing appropriate techniques.

RQ2: What is the relationship between data complexity measures and the choice of oversampling techniques?

Researchers in the imbalanced learning domain not only focus on developing novel approaches but also emphasize the importance of understanding the problem at a deeper level. The more complex the data, the more difficult the classification is. *Imbalance* is not the unique factor hindering the classification. It is also of vital importance to understand how other data characteristics influence the imbalanced classification performance and investigate the relationship between data complexity measures and the choice of oversampling techniques.

RQ3: What is the relationship between the degree of class overlap and the classification improvement obtained via hyperparameter tuning?

Hyperparameter optimisation has shown great effectiveness for many machine learning algorithms. When dealing with class-imbalance problems, hyperparameters in both resampling and classification algorithms should be considered in the experiments. A minor variation of these hyperparameters might influence the performance significantly. Nevertheless, this topic has not been studied in detail in the context of learning from imbalanced data. Therefore, we explore the potential of applying hyperparameter optimisation to construct high-quality classifiers for imbalanced data automatically. From **RQ2**, we already mentioned that data complexity influenced classification performance. Our question in this part becomes, will data complexity affect the classification improvement yielded via hyperparameter tuning?

RQ4: Can we take advantage of anomaly detection techniques to improve imbalanced classification?

The anomaly detection problem can be considered as a class-imbalance problem with an extreme imbalance in terms of class distribution. There are many techniques available in the literature to detect anomalies. Considering the similarity between the two problems, it is very interesting to study if we can improve performance for class-imbalance problems with anomaly detection ideas.

RQ5: How can the idea of four types of samples be effectively extended to multi-class imbalanced scenarios?

The idea of studying different types of samples (*safe, borderline, rare* samples and *outliers*) was first proposed and evaluated on binary imbalanced classification problems (Napierala and Stefanowski, 2016). Since the idea was proposed, it has attracted widespread attention in the field of imbalanced learning, and more than 200 papers have cited the original paper so far. Some studies then extended this idea to multi-class scenarios without considering the more complicated relationships among classes in multi-class imbalanced scenarios. Therefore, proposing improved sample type identification for multi-class imbalanced classification is worth studying.

RQ6: How applicable are the developed approaches to real-world problems?

The performances of the developed approaches are typically tested on benchmark problems. However, given that real-world problems are more complex, these approaches may fail in real-world situations. Therefore, it is important to validate the performance of the approaches on real-world applications.

1.3 Outline of the Thesis

This thesis is organised as follows.

Chapter 2 presents a gentle introduction to class imbalanced problems. It starts with the literature review on binary and multi-class class imbalance problems, including problem illustration, existing approaches and performance metrics. Moreover, it presents the data complexity measures and introduces studies on the data complexity in the imbalanced learning domain. Finally, it shows the benchmark datasets for learning from imbalanced problems and imbalanced applications in real-world scenarios.

Chapter 3 introduces an empirical investigation comparing several oversampling techniques. Apart from experimenting with imbalanced benchmark datasets, further exploration through data from a real-world inspired digital vehicle model is presented.

Chapter 4 presents our study on hyperparameter optimisation on class-imbalance problems. We consider optimising the hyperparameters in both resampling techniques and classification algorithms. Furthermore, we investigate the relationship between the degree of class overlap and the improvement yielded via hyperparameter tuning.

Chapter 5 introduces our idea of improving imbalanced classification via adding additional attributes. We propose introducing the outlier score, an important indicator to evaluate whether a sample is an outlier, as an additional attribute of the original imbalanced datasets. Apart from this, we also introduce the four types of samples (*safe, borderline, rare* samples and *outliers*) as another additional attribute.

Chapter 6 introduces our proposed improved sample type identification for multi-class imbalanced classification. We first show the drawbacks when applying the existing identification rule directly to multi-class scenarios. After that, we emphasize the importance of proposing a new identification rule for multi-class scenarios and introduce the improved type identification rule.

Chapter 7 presents the main conclusions of this thesis and the potential future research directions.

1.4 Publications

The main contributions of this thesis are based on the following publications:

- **Kong, J.**, Kowalczyk, W., Nguyen, D.A., Bäck, T. and Menzel, S., 2019, December. Hyperparameter optimisation for improving classification under class imbalance. In 2019 IEEE symposium series on computational intelligence (SSCI) (pp. 3072-3078). IEEE.

- **Kong, J.**, Rios, T., Kowalczyk, W., Menzel, S. and Bäck, T., 2020, May. On the performance of oversampling techniques for class imbalance problems. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 84-96). Springer, Cham.

- **Kong, J.**, Kowalczyk, W., Menzel, S. and Bäck, T., 2020, September. Improving imbalanced classification by anomaly detection. In International Conference on Parallel Problem Solving from Nature (pp. 512-523). Springer, Cham.

- **Kong, J.**, Kowalczyk, W., Jonker, K., Menzel, S. and Bäck, T., 2022, July. Improved Sample Type Identification for Multi-Class Imbalanced Classification with Real-World Applications. In International Conference on Data Science. (Accepted, publication in process)

Other work by the author:

- Rios, T., **Kong, J.**, van Stein, B., Bäck, T., Wollstadt, P., Sendhoff, B. and Menzel, S., 2020, December. Back to meshes: Optimal simulation-ready mesh prototypes for autoencoder-based 3D car point clouds. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 942-949). IEEE.

- Nguyen, D.A., **Kong, J.**, Wang, H., Menzel, S., Sendhoff, B., Kononova, A.V. and Bäck, T., 2021, October. Improved automated cash optimization with tree parzen estimators for class imbalance problems. In 2021 IEEE 8th international conference on data science and advanced analytics (DSAA) (pp. 1-9). IEEE.