WHIOCE PUBLISHING PTE. LTD.
PROVIDING
FIRST-CLASS SCIENTIFIC INFORMATION
FOR TOP SCHOLARS

2025 Volume 3, Issue 6
Contemporary Education Frontiers
ISSN (Online): 3029-1860 ISSN (Print): 3029-1879

Statistical Learning in Imbalanced Data Classification

Zhe Yan
University of California, Santa Barbara, CA 93106, USA

Abstract: In statistical learning, the classification of imbalanced data is a headache-inducing problem. Why do you say so? Due to the uneven distribution of classes, some classes have a small number of students and insufficient representativeness. Coupled with the limitations of the standard assessment indicators themselves, this poses a particularly significant challenge to statistical learning. There are several methods to try, such as data-level methods, for instance, resampling, which involves re-extracting samples, and feature selection, which involves picking out key features. Algorithm-level methods, such as cost-sensitive learning, which sets different costs for different situations, and ensemble learning, which combines multiple algorithms for use; In addition, there are hybrid methods that combine data-level and algorithm-level approaches. These methods can all deal with the problem of imbalanced data classification to a certain extent. When it comes to actual operation, there are still some matters that need to be carefully considered. You need to first figure out exactly what the data is like, what kind of evaluation metrics are appropriate, model validation cannot be taken lightly, and the model must be understandable and easy to explain.

Keywords: Imbalanced data classification; Data-level methods; Algorithm-level methods; Practical considerations

Online publication: July 26, 2025

1. Introduction

Imbalanced data classification is truly a headache-inducing problem in the field of statistical learning. It exists in many aspects such as medical diagnosis, fraud detection, and rare event prediction. In these scenarios, the quantity of different categories in the dataset varies particularly greatly. Compared with most categories, there are one or two or several categories whose sample sizes are pitifully small, as rare as giant pandas.

This unbalanced situation has brought quite a few special troubles to statistical learning algorithms. It needs to be made clear. When these algorithms were first designed, most of them assumed that the quantities of each category were approximately the same. However, the reality is far from this assumption.

If we ignore the problem of this kind of imbalance in categories, the consequences will be serious. Take medical diagnosis as an example. If a model is trained with imbalanced data, it may be particularly good at identifying common diseases. But when it comes to rare diseases, it acts like a complete novice who knows nothing and is completely undetectable. This can easily lead to misdiagnosis and might even endanger the patient's life. This is no joke. For instance, in fraud detection, if a model, due to category imbalance, ignores a few types of fraudulent transactions, the resulting economic losses can be huge, just like a leaking bucket, with money flowing out in a rush.

2. Challenges Posed by Ibalanced Data

2.1. Uneven Class Distribution

The biggest and most prominent problem we encounter in the classification of imbalanced data is the class imbalance itself. What is class imbalance? That is to say, the category distribution in the problem is skewed and unbalanced. Some category samples are extremely numerous, while others are pitifully few.

Many traditional statistical learning algorithms are designed and exist with the aim of achieving the highest possible overall accuracy. When training these algorithms with balanced datasets, it is quite fair. Here, a balanced dataset means that the number of samples in each category is approximately the same. However, when it comes to class imbalance, these algorithms become “biased”, particularly favoriting the majority of classes, concentrating solely on improving the accuracy of the majority of classes and leaving the minority classes aside[1].

The reason for this bias is that accuracy is calculated. Accuracy is the ratio of the number of correctly predicted instances to the total number of instances in the dataset. Since the number of instances in the majority class is much larger, the contribution to the correct number of predictions is greater when the algorithm predicts an instance of the majority class. In other words, there is a high cost to predicting an instance of the majority class when it is not actually that class. Because of this, the algorithm will tend to perform poorly on the minority class — exhibiting a high false negative rate where actual instances are one class but are predicted to be instances of the majority class.

2.2. Limited Representation of Minority Class

Another important limitation is the lack of instances for minority class. Since the number of instances for minority class is small, the model has the problem of learning the minority class. Every instance in the dataset provides information about the class it belongs to. Therefore, if the number of instances for minority class is small, the model has limited information about what the instances for minority class look like.

This causes a problem for the model with respect to generalization. Generalization is the ability of a model to perform well on data that it has never seen before. If the data set is imbalanced and the model has a problem for generalization, it may perform well on the training data. When the model is faced with a new, unseen minority class instance in the real world, it will always predict incorrectly[2].

Another problem is the lack of diversity for minority class samples. If the instances for minority class are very similar to each other, the model will learn the patterns for minority class instances that are very similar to each other. Therefore, the model will not be able to generalize to instances for minority class that are in the real world, but different from the available minority class instances. For example, if the application is fraud, the instances for minority class may include only a few types of fraud and these types of fraud may be very similar to each other.

2.3. Evaluation Metric Limitations

In statistical learning, the standard evaluation metrics we commonly use, such as accuracy, precision and recall rate, often fail to comprehensively and accurately assess the performance of the model when encountering data imbalance. As mentioned earlier, if the proportion of the majority type in the dataset is particularly large, then the accuracy metric can easily lead people astray. Just think about it. If there were a model that directly predicted all examples to belong to the majority class no matter what situation it encountered, the accuracy of the calculation would seem quite high. But in reality, it simply couldn’t complete the crucial task of correctly identifying the minority class. Isn’t that a scam?

Accuracy and recall rate are not omnipotent either; they also have their own shortcomings. Accuracy refers to the proportion of positive cases that are truly predicted correctly among all the cases predicted as positive. In the case of data imbalance, if a model is particularly conservative and only classifies examples into the minority class when it is absolutely certain, then its accuracy for the minority class may seem quite high. However, this conservative approach also has problems; it can lead to a lower recall rate. What is the recall rate? It refers to the proportion of positive cases that are truly predicted correctly among all the actual positive cases. If a model has a high accuracy for minority classes but a low recall rate, it

will miss many examples that actually belong to minority classes. Such a model is simply not suitable for applications where detecting minority classes is particularly crucial.

In order to break through the limitations of the previous methods, we usually adopt some other evaluation indicators. Take this F1 score for example. It is the harmonized average of the two indicators, accuracy and recall rate. Why calculate it this way? By considering both accuracy and recall rate together, it is possible to have a more comprehensive and balanced understanding of how the model performs. For instance, it's like watching someone work. You can't just look at how fast they do it, but also how well they do it and if there are any omissions. F1 scores take both aspects into account[4].

And the area under the receiver's working characteristic curve, this indicator is used to measure the model's ability to distinguish between two types of data under different judgment criteria. Just like setting different passing lines in an exam, the model's ability to distinguish between two types of data under each passing line varies. This area can comprehensively reflect the model's ability in this aspect. In addition, the precise recall curve pays particular attention to the model performance related to a few categories. This indicator is particularly useful when the number of categories varies greatly, that is, when the category imbalance is very serious. It is like a "magnifying glass", which can see the situations of a few categories more clearly.

These selectable evaluation metrics enable us to assess the model's performance in imbalanced data scenarios more meticulously and accurately. With these indicators, people in this field can more intelligently and reliably determine whether a certain model is suitable for a specific application.

3. Statistical Learning Techniques for Imbalanced Data

3.1. Data-Level Methods

3.1.1. Resampling Techniques

Resampling technology mainly adjusts the training data to make the distribution of different types of data more balanced. Specifically, oversampling means increasing the number of minority class data, while undersampling means reducing the number of majority class data.

There is a particularly common oversampling technique called synthetic minority oversampling, and its English abbreviation is SMOTE. How does this SMOTE operate? It will perform an "insertion" operation among the existing minority class data, thereby generating some new synthetic data for the minority classes. The advantage of doing this is that it can increase the representativeness of the minority classes in the data. Moreover, it does not simply copy the existing minority class data. If it is directly copied, it is very likely to cause the model to "memorize mechanically", that is, overfitting occurs. However, SMOTE also has its preconditions. It assumes that the feature space of the data is continuous and that the data of the minority classes are distributed relatively reasonably, which can make this "insertion" operation meaningful. If the actual situation does not meet these assumptions, the synthetic data generated by SMOTE may not be very reliable and cannot accurately represent minority groups.

3.1.2. Feature Selection and Extraction

Feature Selection and Extraction. Feature Selection and Extraction methods can be also be very helpful in case of imbalanced data. By selecting proper features for classification and eliminating unimportant features for reducing dimensions of data and directing model more precisely to the useful features, these methods can be applied. Feature extraction methods such as principal component analysis (PCA) or linear discriminant analysis (LDA) can spread the original feature space to a feature space with lower dimensionality, which may capture the separation between classes more effectively. In case of imbalanced data, selecting proper features for classification, which are more discriminative for minority class, can help the model to learn more effective and useful information and increase the performance[6].

3.2. Algorithm-Level Methods

3.2.1. Cost-Sensitive Learning

Cost-sensitive learning incorporates the different costs that majority classes and minority classes have to pay when they are misclassified. Just think about it. In reality, misclassifying a few categories may lead to much more serious consequences than misclassifying the majority. Therefore, cost-sensitive learning sets a higher "cost" for the misclassification of minority classes. As a result, the algorithm will pay more attention to minority classes and strive to make more accurate predictions.

Moreover, depending on different learning algorithms, there are various methods to achieve cost-sensitive learning. Take the decision tree algorithm for example. When deciding which feature to use to segment data at each node, the original segmentation standard is like an "old-fashioned" one, only considering some fixed factors. Now that cost-sensitive learning is available, this criterion can be adjusted to take into account the cost of misclassification as well. In this way, when the algorithm divides the data, it will be more inclined to favor a few classes.

3.2.2. Ensemble Learning

Ensemble is an approach that uses multiple base learners. In case of imbalanced data classification, ensemble methods can use the diversity of base learners to learn the imbalanced data classification better.

Bagging constructs multiple bootstrap samples from the training data and trains a separate base learner on each bootstrap sample. The predictions of base-learners are combined by voting or averaging. Since each base-learner is trained on different bootstrap samples, it is expected that they will be diverse from each other and therefore learn different parts of the data which may help in learning the minority class better.

Boosting trains the base learners sequentially, where the base learner t is trained to correct the misclassifications of base learner t−1. This process is repeated such that the model gradually learns the imbalanced data classification better by giving more weight to the misclassified instances.

AdaBoost is one of the well-known boosting algorithms that has been applied successfully on many imbalanced data classification problems.

3.3. Hybrid Methods

Hybrid methods combine data-level and algorithm-level approaches to achieve better performance in imbalanced data classification. For instance, a hybrid method may first apply resampling to balance the class distribution and then use a cost-sensitive learning algorithm to train the model. By leveraging the strengths of both approaches, hybrid methods can often achieve superior performance compared to using either approach alone[7].

4. Practical Considerations and Best Practices

4.1. Data Understanding and Preprocessing

Thoroughly understanding the data and the problem at hand is crucial. This includes analyzing the class distribution, identifying potential outliers or noise, and understanding the business or application context. Such understanding can guide the choice of appropriate preprocessing techniques. For example, if the data contains a significant amount of noise in the minority class, more sophisticated noise reduction techniques may be required before applying resampling or other classification methods.

4.2. Evaluation Metric Selection

The choice of evaluation metric depends on the application's goals. Recall that we've already noted that standard metrics such as accuracy are not appropriate for imbalanced data, and alternative metrics such as the F1-score, AUC-ROC, and precision-recall curves may be more meaningful. Sometimes the cost of misclassifying the minority class may be much more than misclassifying the majority class. In these instances, an evaluation metric that accounts for the different

misclassification costs may give a more accurate evaluation of a model's performance.

4.3. Model Validation and Testing

You should also test the model on independent test sets to make sure. We can also use cross validation methods like 5-fold cross validation to validate the model. In 5-fold cross validation, we partition the data into 5 subsets and the model is run 5 times, with each of the 5 subsets used as the test set once and the remaining subsets used as the training set. The result is an average estimate of the model performance that is more reliable than a single train/test split.

4.4. Interpretability and Explainability

As discussed in the imbalanced data classification post, in some practical applications it might be important to be able to explain to the user why the model made a specific prediction. This is particularly important when the prediction might have significant consequences if the user does not agree with it. In such cases, being able to provide some form of explanation and/or sanity checks is extremely helpful.

Feature importance: if using a model supporting this feature, extracting the most important features and explaining to the user why the prediction was made based on those features. Decision tree visualization: similar to the previous point, but specifically for decision trees. Model agnostic explanations: extracting the SHAP values of the input features for each prediction and visualizing them.

5. Conclusion

The classification of imbalanced data in the field of statistical learning is truly a special and difficult challenge. Why do you say so? The main reason is that the quantity distribution of various categories in the data is uneven, and this unevenness is the major source of complexity.

In real life, it is all too common for the number of one category to far exceed that of another. Take the fraud detection system for example. The number of legitimate transactions is far greater than that of fraudulent ones. Just like a drop of water in the ocean, fraudulent transactions are pitifully few. In the medical diagnosis of rare diseases, healthy cases account for the majority in the dataset, while cases of rare diseases are extremely rare. This uneven situation has disrupted the normal learning process of traditional statistical models. It should be noted that when these traditional models are designed, they usually assume that the distribution of each category is balanced.

Another issue that makes this matter even more difficult is that the number of instances represented by a few classes is too limited. Because there are few available examples for a few classes, it is difficult for the model to capture its unique features and patterns. Each instance of a minority class is valuable information. However, if the number is too small, it means that the model may not have enough examples to effectively draw inferences by analogy. This can easily lead to overfitting, that is, the model records a limited number of class data but fails to learn the true underlying rules. So, when the model encounters new and unseen instances of small categories, it is very likely to make wrong predictions.

Moreover, standard assessment metrics also have significant limitations, which is like a major obstacle on the way forward. Metrics like accuracy, which are calculated by dividing the number of correctly predicted instances by the total number of instances, can easily deceive people in an unbalanced data scenario. If there were a model that, without any hesitation, predicted all instances as majority classes, its accuracy might be quite high. However, it simply couldn't complete the crucial task of correctly identifying minority classes. Although accuracy and recall rate can provide more information, they also have their own drawbacks. A model may predict with extreme caution to achieve high accuracy for minority classes, but this might come at the expense of recall rate, that is to say, it will miss many true instances of minority classes.

However, despite all these challenges, there are still some effective ways to develop models that perform well in situations of category imbalance. Data-level methods can play a crucial role in solving this problem. Resampling

techniques are quite useful. For instance, oversampling a few classes and undersampling the majority classes can make the distribution of classes more balanced. Oversampling methods like Synthetic Minority Oversampling (SMOTE) can generate some synthetic instances for minority classes, increasing the representativeness of minority classes without simply replicating existing data. Conversely, undersampling means reducing the number of instances of the majority class, making the dataset more balanced. Feature selection is also a very important data-level method. By identifying and selecting the most relevant features, we can enable the model to focus on the aspects of the data that are most useful for distinguishing different categories, especially for a few categories.

The hybrid approach is even more powerful. It combines the advantages of both data-level and algorithm-level methods. For instance, a hybrid approach might first use resampling to balance the distribution of categories, and then train the model with cost-sensitive learning algorithms. This combination usually works better than using any one of the methods alone, just like 1 plus 1 is greater than 2.

Nowadays, in various fields such as finance, healthcare and security, the demand for accurate and reliable classification models is increasing. So for data scientists and machine learning practitioners, mastering the skills of classifying imbalanced data is quite valuable. If we can effectively address the challenge of data imbalance, we will be able to fully exploit the potential of statistical learning and promote innovation in areas where accurate classification is particularly important for success. Whether it is detecting rare diseases, identifying fraudulent transactions, or predicting rare events, the ability to handle imbalanced data is the key to developing advanced and influential machine learning systems.

Disclosure statement

The author declares no conflict of interest.

References

[1] He H, Bai Y, Garcia E A, et al., 2008, ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE.

[2] García, Salvador, Herrera, et al., 2009, Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy. Evolutionary Computation.

[3] Fernández A, López V, Galar M, et al., 2013, Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. Knowledge-Based Systems, 42: 97-110.

[4] Yin Q Y, Zhang J S, Zhang C X, et al., 2014, A Novel Selective Ensemble Algorithm for Imbalanced Data Classification Based on Exploratory Undersampling. Mathematical Problems in Engineering.

[5] Liang G, Zhang C, 2012, A Comparative Study of Sampling Methods and Algorithms for Imbalanced Time Series Classification. Springer, Berlin, Heidelberg.

[6] Yan Y, 2018, Deep Learning Based Imbalanced Data Classification and Information Retrieval for Multimedia Big Data. ProQuest LLC.

[7] Jin Y, Wang N, Wu R, et al., 2024, Ultra-imbalanced classification guided by statistical information.