POINTS OF SIGNIFICANCE: STRATEGIES FOR HANDLING THE CLASS IMBALANCE PROBLEM

**Fadel M. Megahed, Ying-Ju (Tessa) Chen, Aly Megahed, Yuya Ong, ....**

We previously discussed how logistic regression[1](https://paperpile.com/c/ZxOU7T/DmdM) and machine learning methods[2](https://paperpile.com/c/ZxOU7T/qJxF) can be used for predicting the class of an observation given a training dataset[3](https://paperpile.com/c/ZxOU7T/wTzc). In this article, we will discuss how these classification methods can be trained in the presence of class imbalance, when at least one of the classes within the response is rare (e.g., 30-day mortality predictions post a bypass surgery). While we recently discussed testing for rare conditions[4](https://paperpile.com/c/ZxOU7T/qpdi), the goal here is quite different since we are trying to predict the class of the response variable given a set of potential predictors (e.g., how to best predict the 30-day mortality rate given a set of patient’s medical history, biometrics, and/or socioeconomic characteristics). In this article, we focus on binary classification.

In most statistical/machine learning software, the default settings utilize *accuracy*, which is defined as the number of correctly predicted classes divided by the number of observations in the training dataset, to train binary classifiers. In this article, we show the impact of class imbalance on logistic regression[1](https://paperpile.com/c/ZxOU7T/DmdM) and decision trees[5](https://paperpile.com/c/ZxOU7T/MyEm), which are two commonly used statistical/machine learning techniques.

Per our discussion in[1](https://paperpile.com/c/ZxOU7T/DmdM), logistic regression requires numerical optimization to find the optimal estimates of the regression parameters through a maximum likelihood estimation (MLE) procedure. When the outcomes are highly imbalanced, the MLE of the logistic regression model suffers from small-sample bias[6](https://paperpile.com/c/ZxOU7T/5G8x), where the logistic regression model would underestimate the probability of observing the rare event. The proportion of observations in the rare/minority class influences the degree of bias. For example, let us consider a hypothetical study of bypass surgery predictions where the patients’ survival rate is 95%. In such a case, we can obtain a naive model (with no predictors) that is 95% accurate by completely ignoring the 5% cases where the patients did not survive (i.e., predicting all patients to survive irrespective of their medical history and demographics). Hence, training/optimizing a logistic regression model using accuracy is not suitable when the outcomes are imbalanced.

Similarly, a decision tree is built by partitioning the predictor variables to increase the class separation at each split[5](https://paperpile.com/c/ZxOU7T/MyEm). When the dataset is imbalanced and accuracy is used as the optimization metric, the decision tree would put more weights on the observations in the majority class and the influence of the minority class is underestimated. Hence, the use of accuracy to train the decision tree is not suitable when the outcomes are highly imbalanced.

In practice, there are four commonly used strategies to handle the class imbalance problem: (a) collecting more data on the rare/minority class; (b) using subsampling methods to transform the imbalanced training set into a (more) balanced dataset; (c) utilizing a suitable performance metric, e.g., the area under the precision recall curve (AUPRC); and (d) combining the use of subsampling with the choice of an appropriate metric. If one can include additional observations containing the minority class, this should be done prior to modeling. However, in many medical applications, this strategy may be time prohibitive and hence, will not be discussed further.

Subsampling, a.k.a. resampling, can be divided into under-sampling, over-sampling and hybrid techniques (Fig. 1). Under-sampling randomly removes observations from the majority class in the original training set. Over-sampling includes additional observations from the minority class through random sampling with replacement.

Subsampling methods present some limitations. Under-sampling can result in the loss of valuable information or cause some categorical predictors to be invariant. On the other hand, over-sampling can result in overfitting which can be assessed by comparing the predictive performance of the model based on the training and holdout datasets. Hybrid methods can mitigate the limitations by synthesizing new observations for the minority class and in some instances under-sampling the majority class.

Another technique for handling the class imbalance problem is to select an appropriate performance metric. In a previous article[3](https://paperpile.com/c/ZxOU7T/wTzc), we showed how the area under the precision curve (AUPRC) is more suitable than accuracy and the area under the receiving operating characteristic curve (AUROC) for class imbalance problems.

|  |
| --- |
| **Figure 1: The application of subsampling methods to obtain a balanced dataset, with an illustrative example in the bottom left of each panel and its corresponding application to the “new-thyroid2” dataset.**    **(a)** The original dataset containing 35 positive outcomes and 180 negative outcomes. **(b)** The under-sampled data, where 35 observations from the majority/negative class were randomly selected without replacement, resulting in a total of 70 observations (35 positive and 35 negative) to be used in model training. **(c)** The over-sampled data, where 145 additional observations were obtained using a sampling with replacement strategy on the minority/positive class. **(d)** The use of a hybrid method to generate synthetic examples to be added to the minority class. |

The subsampling and performance metric selection strategies do not have to be mutually exclusive since they can be combined together. Consider a dataset where the imbalance ratio,, in its binary response is larger than 100. In this case, a modeler may choose to subsample to partially balance the data, i.e. the imbalance ratio remains larger than one, in order to not lose so much information (under-sampling) or create too many repetitive/synthetic observations (over-sampling). To account for the remaining imbalance in the dataset, the modeler can now use the AUPRC to train their statistical/machine learning model(s).

To provide some data-driven guidance on strategy choice, we examined 58 imbalanced binary classification datasets from the KEEL repository[7](https://paperpile.com/c/ZxOU7T/QgQV). Our criteria for selecting these datasets were: the minority class being the “positive” category, having at least 25 observations for the minority class, and the number of predictors provided in the metatable to match those observed in the raw data. We provide an overview for these 58 datasets in Table 1, where we have grouped the datasets by their imbalance ratio. In our numerical experiments, we considered 2 classifiers (logistic regression and the CART implementation for decision trees), 4 subsampling strategies (none, under, over, and SMOTE), and 3 performance metrics (accuracy, AUROC, and AUPRC) which were used to train the classifiers.

**Table 1: A summary of the 58 imbalanced datasets for binary classification examined in our analysis. The Interquartile Ranges (IQR) are in parentheses.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Imbalance Ratio (IR)**  **Category** | **# Datasets** | **Median IR** | **Median # observations** | **Median # predictors** | **Median # numeric predictors** | **Median # categorical predictors** |
|  | 13 | 2.78 (1.0) | 683 (632.0) | 9.0 (10.0) | 9.0 (10.0) | 0 (0) |
|  | 15 | 8.79 (2.9) | 514 (707.5) | 8.0 (1.5) | 8.0 (1.5) | 0 (0) |
|  | 15 | 22.10 (13.1) | 1484 (1262.5) | 7.0 (2.5) | 7.0 (4.0) | 0 (3) |
|  | 15 | 58.28 (32.7) | 1916 (801.5) | 8.0 (1.5) | 8.0 (2.5) | 1 (2) |

The impact of the three general strategies for handling imbalanced datasets (subsampling, metric, and both) on four representative KEEL datasets (one from each imbalance ratio category) is depicted in Fig. 2. Based on the sensitivity values (how well we detect the true positives), there are three important observations to be made. First, the no subsampling strategy (NULL) generally results in lower sensitivity values when compared to the under-sampling and SMOTE strategies, irrespective of the metric used to optimize the dataset and the dataset’s imbalance ratio. Second, the larger the imbalance ratio, the no subsampling approach converges to a naive model, where the sensitivity values tend to 0 (i.e., we do not detect any positive outcomes). Third, the use of AUPRC and AUROC does improve the sensitivity results for the no subsampling case (especially for the *new-thyroid2* and *yeast-1\_vs\_7* datasets); however, the impact of metric is not substantial when a subsampling approach is utilized. This is not surprising since we used subsampling to achieve a balanced dataset (IR = 1). Note that our three observations may not generalize to other datasets.

|  |
| --- |
| **Figure 2: 5-fold cross validation results from the decision tree model applied to four randomly selected datasets from each imbalance ratio category. The columns capture the choice of metric to train the decision tree and the rows capture our four subsampling strategies (with no subsampling in the top row).**    The animated figure is changed every 10 seconds following the imbalance ratio category. |

We performed an ANOVA on the sensitivity results to capture how the imbalance ratio category, subsampling and method and their pairwise interactions impacted our obtained results for all 58 datasets (Fig. 3). We observed that the no subsampling strategy yielded the lowest sensitivity values, with the average sensitivity values deteriorating rapidly when the IR > 10. On the other hand, under-sampling resulted in higher sensitivity results, on average, when compared to over-sampling and had significantly better results at the higher and lower ends of the imbalance ratio categories. For the examined datasets, logistic regression outperformed the decision tree model especially when the IR > 10.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Figure 3: Result of Hierarchical Regression Analysis**   |  |  |  | | --- | --- | --- | | (a) Type II Sum of Squares and Anova F-tests | (b) Imbalance Raito and Subsampling as Predictors | (c) Imbalance Raito and Method as Predictors | |  | | | |

If we were to provide a general strategy based on our numeric experiments, we would recommend the use of down-sampling as an initial approach to training the machine learning model(s) of choice. If the results are acceptable, there may be no need for utilizing more advanced and computational subsampling methods. If the modelers chose to utilize their subsampling methodology to achieve a balanced dataset, we suspect that the optimization metric will not be impactful. However, if the IR is not equal to 1, the choice of optimization metric can be important (see Fig. 2).

**Supplementary Materials**

Our code can be accessed at <https://github.com/fmegahed/subsampling/>.

**References**

1. [Lever, J., Krzywinski, M. & Altman, N. Logistic regression. *Nat. Methods* **13**, 541–542 (2016).](http://paperpile.com/b/ZxOU7T/DmdM)

2. [Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: supervised methods. *Nat. Methods* **15**, 5–6 (2018).](http://paperpile.com/b/ZxOU7T/qJxF)

3. [Lever, J., Krzywinski, M. & Altman, N. Classification evaluation. *Nat. Methods* **13**, 603–604 (2016).](http://paperpile.com/b/ZxOU7T/wTzc)

4. [Altman, N. & Krzywinski, M. Testing for rare conditions. *Nat. Methods* **18**, 224–225 (2021).](http://paperpile.com/b/ZxOU7T/qpdi)

5. [Krzywinski, M. & Altman, N. Classification and regression trees. *Nat. Methods* **14**, 757–758 (2017).](http://paperpile.com/b/ZxOU7T/MyEm)

6. [King, G. & Zeng, L. Logistic Regression in Rare Events Data. *Polit. Anal.* **9**, 137–163 (2001).](http://paperpile.com/b/ZxOU7T/5G8x)

7. [Alcalá-Fdez, J. *et al.* Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *J. Mult.-Valued Logic Soft Comput.* **17**, (2011).](http://paperpile.com/b/ZxOU7T/QgQV)