POINTS OF SIGNIFICANCE: THE CLASS IMBALANCE PROBLEM

*leadfirst*

We previously discussed how statistical/machine learning methods (e.g., logistic regression1 and decision trees2) can be used for predicting the class of an observation given a training dataset3. Unfortunately, when one of the classes within the response is rare, the classifier can underestimate the probability of observing a rare event – the greater the imbalance, the greater this small-sample bias4. To illustrate how class imbalance affects classifiers and how it can be mitigated, we will focus on binary classification with decision trees.

One of the key metrics (and often the default) in evaluating classifiers is accuracy – the fraction of correctly predicted classes. However, when one class is rare, accuracy can be misleading (because it will vary across classes) and is unsuitable for training classifiers (because it would assign too much weight on the majority class and insufficient weight on the minority class). For example, suppose we want to predict the outcomes of bypass surgery, which has an average survival rate of 95%. We can immediately create a trivial classifier that predicts that every patient survives. Our accuracy across classes would be excellent (95%), but the model is useless because its accuracy within the minority class is 0%.

Class imbalance is typically expressed using the imbalance ratio, *IR* = *M*/*m* where M and m are the number of observations in the majority and minority class, respectively. There are two practical strategies to handle the class imbalance problem: (a) applying subsampling methods to transform the imbalanced training set into a (more) balanced dataset, and/or (b) utilizing a suitable performance metric. It is also possible to collect more data on the rare class but this can be time prohibitive and we will not discuss it here.

To illustrate the data imbalance problem and the impact of subsampling techniques on model performance, let us return to our example in which we measure levels of two enzymes A and B in normal and diseased cells5. Recall that normal cells had enzyme A and B levels distributed as *N*(1, 0.22) and Exp(1). In diseased cells these levels were distributed as *N*(1.2, 0.22) and Exp(1/0.29). For this simulated example, we fixed the total number of observations in the training set to 1,000 observations, and examined six different IR values. The first case corresponds to IR = 1, a balanced simulated scenario with 500 normal and 500 diseased cases. Furthermore, we examined five other imbalanced training sets of IR ≃ 5, 10, 25, 100, and 333 corresponding to 167, 91, 38, 10 and 3 diseased cases respectively.

Using a reduced of 110 total observations and IR = 10 data set as a visual example (**Fig 1a**), we can balance it by resampling in two ways. We can under-sample (**Fig. 1b**), where we randomly remove observations from the majority class in the original training set. This can result in the loss of valuable information, e.g., some infrequent levels of a categorical predictor may be removed from the training set. Alternatively, we can over-sampling (**Fig. 1c**) by including additional observations from the minority class. This can be achieved with random sampling with replacement, which can result in overfitting, where the model performs better on training data than on test data. Finally, we can over-sample by generating synthetic observations in the minority class (**Fig. 1d**). One scheme to do this is Synthetic Minority Over-sampling TEchnique, SMOTE, which uses each minority class observation to generate synthetic observations along the line segments joining its *k* minority class nearest neighbors6. The number of *k-* nearest neighbors depends on the amount of over-sampling required6.

Diagram

Description automatically generated

**(two column) Fig. 1 | The use of subsampling methods to obtain a balanced dataset. a.** The original dataset contained 10 diseased cells and 100 normal cells. **b.** The under-sampled data, where 90 observations from the majority/normal class were randomly selected without replacement, resulting in 20 observations (10 diseased and 10 normal) to be used in model training. **c.** The over-sampled data, where 90 additional observations were obtained using a sampling with replacement strategy on the minority/diseased class. **d.** The generation of synthetic examples that were added to the minority class.

Under- and over-sampling are not necessarily mutually exclusive. For extremely unbalance cases (e.g. IR > 100), a hybrid/combined sampling approach can be very effective. In this scheme, first the majority class is randomly under-sampled until a pre-specified IR is attained. Second, an over-sampling technique (e.g., SMOTE) is used to further balance the dataset by increasing the number of observations in the minority class. If a robust performance metric is used achieving full balance is not required. For example, the area under the ROC curve (AUROC)3,5 often outperforms overall accuracy as a training metric with relatively small IRs (e.g., IR<10). To explore the extent to which these resampling schemes can mitigate the negative effects of class imbalance on the training of a classifier, we’ll track the performance of a decision tree classifier trained on 1,000 cells and tested on 10,000 cells. Two main observations can be made based on the testing results (**Fig 2.**). First, when no subsampling is applied (NULL), the performance of our classifier deteriorates quickly with increasing IR. For IR ≥ 10, the classifier training process completely ignores the minority class and the classifier naively labels all cells as normal. This behavior corresponds to zero sensitivity (how well we predict diseased cells) and the AUROC = 0.5 as the specificity (how well we predict normal cells) is equal to one. On the other hand, when under-sampling was used to balance the training data, the model’s performance improved for IR ≤ 100. At IR = 333 the classifier performed naively but this time labeling all cells as diseased.

Chart

Description automatically generated

**(two column) Fig. 2 | The holdout predictive performance of the CART decision tree model for classifying normal and diseased cells with no subsampling and under-sampling applied to training datasets having IR of 1, 5, 10, 25, 100, and 333.** The three panels capture **a.** specificity, **b.** sensitivity, and **c.** AUROC predictive performances versus the imbalance ratio in the training dataset with no resampling (under-sampling was used to obtain *IR* = 1 in all cases). The x-axis, Imbalance Ratio, spacing is scaled using a logarithmic transformation to facilitate the visualization of the highly skewed IR values.

Under-sampling isn’t necessarily the best approach, however. In **Fig. 3**, we repeat the assessment for over-sampling and SMOTE. Both under- and SMOTE sampling generally improve the sensitivity, and AUROC for *IR* ≤ 100 when compared to over-sampling. All subsampling strategies resulted in a naive model for *IR* = 333, with no diseased cell predictions for under-sampling, and no normal cell predictions for the over-sampling and SMOTE cases. This is not surprising since: (a) 6 total (3 diseased and 3 normal cases) are not sufficient for training a representative decision tree model that captures normal and diseased cells not observed in the training set; (b) 3 diseased cases are not sufficient to generate synthetic examples, with the default *k-*nearest neighbor value equal to 5, and hence, the algorithm’s implementation converged to an over-sampling method; and (c) the over-sampling based on 3 diseased observations resulted in an over-fit model where we were able to detect diseased cases in training but not in the test/holdout application.

Chart, line chart

Description automatically generated

**(two column) Fig. 3 | The holdout predictive performance of the CART decision tree model for classifying normal and diseased cells with under-sampling, over-sampling and SMOTE applied to training datasets having IR of 1, 5, 10, 25, 100, and 333.** The three panels capture **(a)** specificity, **(b)** sensitivity, and **(c)** AUROC predictive performances versus the imbalance ratio in the original training dataset with no resampling (resampling was used to obtain IR=1 in all cases). The x-axis spacing is scaled using a logarithmic transformation to facilitate the visualization of the highly skewed IR values.

**Fig. 4** shows the ROC and PR curves for the no subsampling (NULL) and under-sampling classifier from **Fig. 3** when *IR*=10. When under-sampling is applied, the model achieves both greater AUROC and AUPRC. In addition, the model recalls about 70% of diseased cells in the holdout data at the threshold *f0=* 0.63 while misclassifying only about 28% of the normal cells and the PR curve shows the classifier achieves a precision of 71.2% at a power of 69.7%. When no subsampling is applied, the ROC curve is on a 45-degree line which indicates the classifier is no better than random guessing.

Graphical user interface

Description automatically generated with medium confidence

**(two column) Fig. 4 | The ROC and PR curves for the decision tree classifier of normal and diseased states, corresponding to the no- (NULL) and under-sampling scenarios.** The curves were generated based on a balanced holdout set of 10,000 observations (5,000 for each cell outcome), and the models were trained based on the IR=10 Scenario depicted in Fig. 1.

Subsampling strategies present one potential solution to the class imbalance problem, which can result in overfit models. If we were to provide a general strategy for handling the class imbalance problem, we would recommend the use of under-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, the optimization metric will likely not be impactful and hence we did not examine it in our simulated example. When the imbalance ratio is high, we suspect that subsampling alone may not result in “good” predictive performance as shown in our simulated diseased cell example. In such cases, it may be best to combine the use of subsampling with the choice of a suitable optimization metric (e.g., AUPRC based on our previous discussion in[3](https://paperpile.com/c/iCic9d/s8TOJ)).

**Replication and Supplementary Materials**

We host a Markdown containing R code and results for: (a) our simulated example; (b) R code that can be used for the readers’ own datasets; and (c) supplementary experiments on the use of both subsampling and optimization metrics to handle the class imbalance for 58 popular datasets. It can be accessed at <https://ying-ju.github.io/subsampling.github.io/>.

**Fadel M. Megahed1, Ying-Ju Chen2, Aly Megahed3, Yuya Ong3, Naomi Altman4 & Martin Krzywinski5\***

1Farmer School of Business, Miami University, Oxford, OH, USA. 2Department of Mathematics, University of Dayton, Dayton, OH, USA. 3IBM Research - Almaden, San Jose, CA, USA. 4Department of Statistics, The Pennsylvania State University, State College, PA, USA. 5Canada’s Michael Smith Genome Sciences Centre, Vancouver, British Columbia, Canada,

\**e-mail:* [*martink@bcgsc.ca*](mailto:martink@bcgsc.ca)

**References**

1. Lever, J., Krzywinski, M. & Altman, N. Logistic regression. *Nat. Methods* **13**, 541–542 (2016).

2. Krzywinski, M. & Altman, N. Classification and regression trees. *Nat. Methods* **14**, 757–758 (2017).

3. Lever, J., Krzywinski, M. & Altman, N. Classification evaluation. *Nat. Methods* **13**, 603–604 (2016).

4. King, G. & Zeng, L. Logistic regression in rare events data. *Polit. Anal.* **9**, 137–163 (2001).

5. Naomi Altman, A. M. K. Graphical Assessment of Tests and Classifiers. *Nat. Methods* (2021).

6. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).