The Art of Balancing

Zhen Zhang, Justin Croft, Kendell Churchwell

C Spire, Department of Marketing, 1018 Highland Colony Parkway, Ridgeland, Mississippi, 39157, USA

ABSTRACT

Data mining technology has been widely used in strategic marketing to uncover actionable information for a wide spectrum of critical marketing decisions. A common problem in many data mining applications is that data is often skewed, and skewed data often leads to degenerated algorithms that assign most or all cases to the most common outcome. For the modeling projects that have extremely skewed targets, such as churn prediction or fraud detection, data balancing techniques applied prior to modeling process are crucial steps to ensure a useful model. As modeling cases are domain and algorithm sensitive, there is no one-size-fit-all solution for the right balancing strategy. In this paper we present empirical guidelines on balancing strategies for extremely skewed data with binary outcome. Best practices are suggested pertaining to decision trees, logistic regression algorithm and neural network models.

Keywords: data mining, skewed, target, class imbalance, binary classifier, model, algorithms, majority class, minority class, overall accuracy, target accuracy, SPSS Modeler.

1. INTRODUCTION

Data mining is a process of extracting hidden information in data and translated it into actionable knowledge, patterns, or insights for business decisions. One of the challenges data miners often face is skewed data. Skewed data, also called class imbalance, refer to the situation where classification categories are not approximately equal. The class that has the greater percentage is called the majority class, whereas the class that has the smaller percentage is called the minority class. In many real-world situations the minority classes, such as churn, fraud, or network failure are often the modeling target for business applications (Weiss 2004; Longadge et al. 2013).[1,2] However, directly modeling imbalanced data can be problematic. This is because most commonly used classification algorithms assume a relatively balanced class distribution. Consequently, these algorithms attempt to build models with the goal of maximizing over all accuracy, which results in degenerated classifiers that assign most or all cases to the most common outcome (Sun et al. 2007; Seiffert et al. 2008).[3,4] The issue with class imbalance become more pronounced with the applications of machine learning algorithms, therefore the data mining community has paid considerable attention to improving performance of modeling imbalanced data (Batista et al. 2004; Chawla, 2005).[5,6] The commonly applied data-level approach can be generally categorized into Under Sampling, Over Sampling and Advanced Sampling. Under Sampling, or Fully Reducing, balances data by keeping all cases in the minority class while randomly eliminate the majority class; Over Sampling, or Fully Boosting, replicates the minority class to balance the majority cases. For severely imbalanced data, Fully Reducing technique can result in a weak model because the feature dimensions of the remaining small portion of the majority class often cannot adequately represent that of the original data. Fully Boosting, on the other hand, often results in an over-fitted classifier. To counter these problems, Advanced Sampling techniques are developed to intelligently reduce majority class by removing redundant cases or cases borderline with minority class, or intelligently add minority cases, such as SMOTE technique.[1]

While Advanced Sampling strategies are reported to improve classifier quality at various degrees, they are often complex to apply. In our practice of modeling skewed binary target, we have observed that sampling 10% or slightly greater of majority class is often sufficiently representative of its original population. When data is severely skewed, there is still a need to replicate majority class to balance the 10% majority class sampled, but the replication is on a much reduced scale, therefore the over-fitting problem inherent to Fully Boosting method is mitigated. We call this the "10% Rule" sampling strategy. Compared to other Advanced Sampling strategies, the "10% Rule" strategy is simple & straight forward to apply.

In this study we've conducted statistical tests to estimate & compare the performance of three sampling strategies: Fully Reducing, Fully Boosting, and the "10% Rule", in combination with different modeling algorithms. Table 1 demonstrates the three sampling strategies using an example of 50,000 records in the training set, and the target size is 2%.

Table 1. Balancing Strategies Example

Sampling Strategy	Majority Class	Minority Class (target)	Majority Class Sampled (%)	Minority Class Replicated
Fully Reducing	1,000	1,000	2%	1 x
Fully Boosting	49,000	49,000	100%	49 x
"10% Rule"	5,000	5,000	10%	5 x

.

2. METHODS

2.1. Sampling Method for Evaluation Datasets

We used data from C Spire business database. Potential predictive fields derive from the CRM and billing system; Modeling target is customer churn. Through random sampling with replacement, the original dataset is repeatedly split 50/50 into training and testing datasets. These training and testing pairs are then subjected to various combinations of modeling algorithms and balancing strategies.

2.2 Modeling Algorithms and Software

Modeling algorithms chosen for this research are popular techniques for binary classifiers: Logistic Regression, Neural Network and Decision Tree. The Chi-square Automatic Interaction Detection (CHAID) algorithm is used to represent Decision Tree. Software used for this work is IBM SPSS Modeler.

2.3 Balancing Strategies

SPSS Modeler offers two default options for balancing training data set: Fully Reducing or Fully Boosting. These two approaches are compared to the proposed "10% Rule" strategy. The"10% Rule" strategy aims at keeping a sufficiently representative sample for the majority class while minimizing the number of replicates required for minority class.

Since the modeling target is approximately 1.25%, the "10% Rule" translates as randomly selecting 10% of the active customers in training set while making approximately 8 replicates of all churn cases.

2.4 Statistical Tests

ANOVAs are applied to compare both overall accuracy and target accuracy. Tests for homogeneity of variance are conducted to determine the proper Post-Hoc multiple comparison test. Dunnett's T3 test is chosen when the homogeneity of variance test is rejected, otherwise Bonferroni test is applied.

3. RESULTS

3.1. General Trend

The overall accuracy increases as the balancing strategy goes from Fully Reducing to "10% Rule" to Fully Boosting. On the other hand, the target accuracy decreases as the balancing strategy moves in the same direction. Similar general trend is observed across all three modeling algorithms tested. Figures 1 and 2 show this trend.

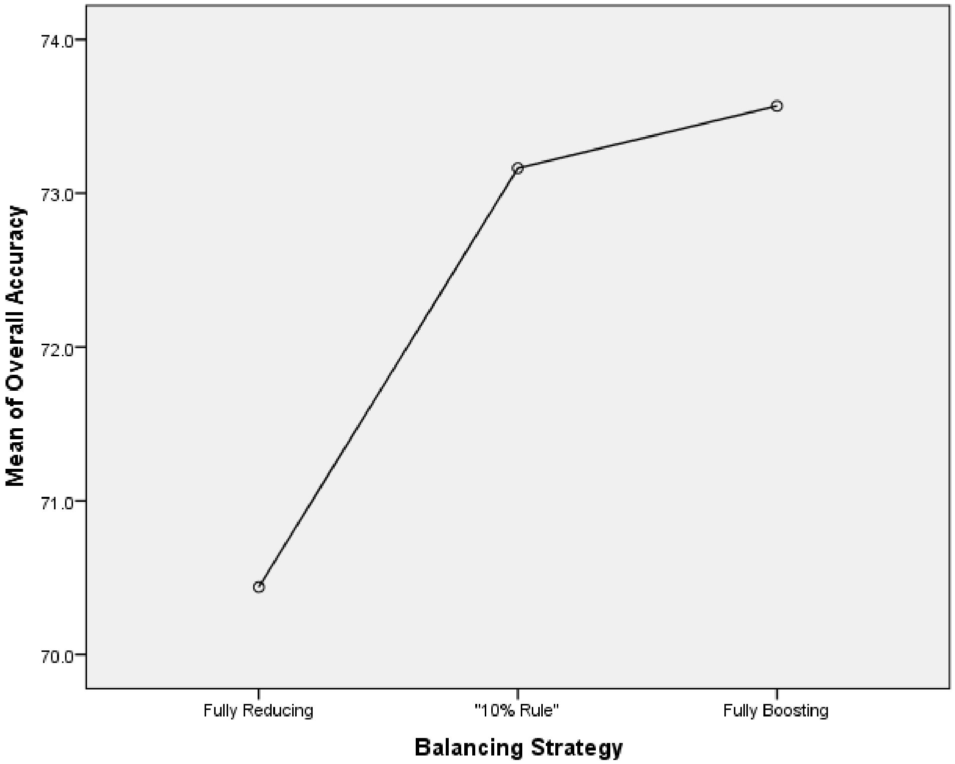

Figure 1. Comparison of Overall Accuracy by Balancing Strategy

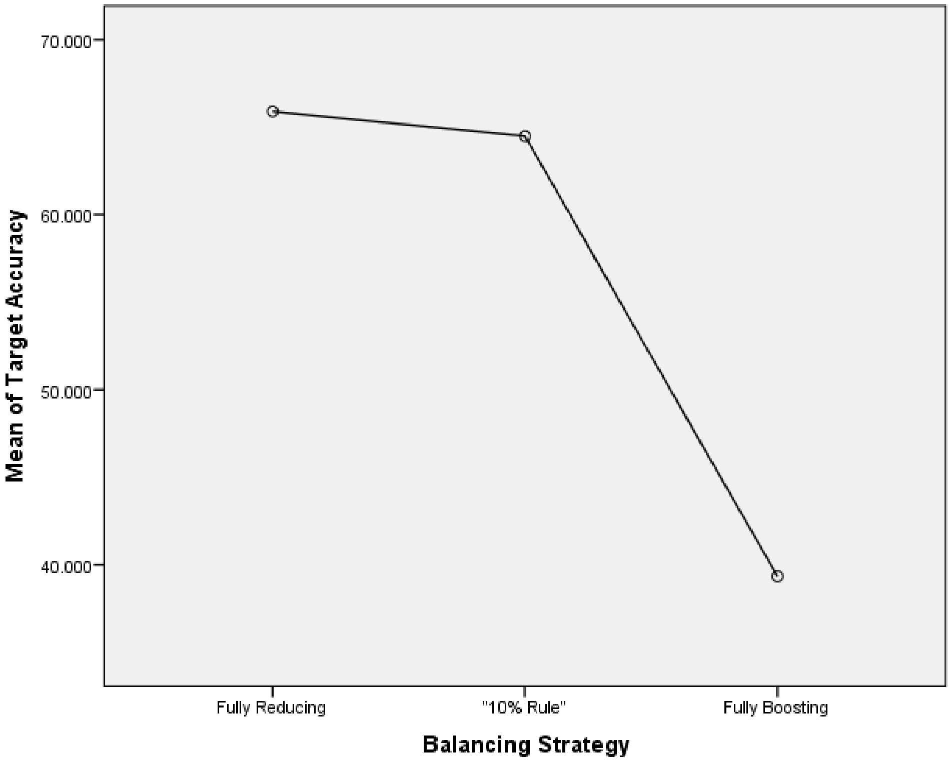

Figure 2. Comparison of Target Accuracy by Balancing Strategy

3.2. Neural Net

ANOVA and Post-Hoc tests indicate that there is significant difference in both overall accuracy and target accuracy between the different sampling strategies. Table 2 shows the mean and 95% of confidence interval of overall and target accuracies.

3.3. Decision Tree

With CHAID, the “10% Rule” and Fully Reducing strategies performs similarly. Fully Boosting strategy produces classifier with significantly higher overall accuracy and significantly lower target accuracy, compared to the other two strategies. Table 3 shows the descriptive results.

3.4. Logistic Regression

Logistic regression algorithm demonstrates the robustness to sampling methods. Though there is significant difference between sampling methods, the scale of difference is smaller compared to the other two algorithms. Table 4 shows the mean and 95% of confidence intervals for overall accuracy and target accuracy.

Table 2. Neural Net: Comparison of Overall Accuracy and Target Accuracy between Sampling Strategies

				95% Confidence Interval	
		N	Mean	Lower Bound	Upper Bound
Overall Accuracy	Fully Reducing	50	67.3	66.4	68.3
	"10% Rule"	50	72.8	72.3	73.3
	Fully Boosting	50	84.1	83.7	84.4
Target Accuracy	Fully Reducing	50	65.9	64.8	67
	"10% Rule"	50	64.5	63.5	65.4
	Fully Boosting	50	39.4	38.5	40.1

Table 3. Decision Tree: Comparison of Overall Accuracy and Target Accuracy between Sampling Strategies

				95% Confidence Interval	
		N	Mean	Lower Bound	Upper Bound
Overall Accuracy	Fully Reducing	50	68.1	66.8	69.4
	"10% Rule"	50	69.1	68.2	69.9
	Fully Boosting	50	71.1	70.4	71.7
Target Accuracy	Fully Reducing	50	65.1	63.7	66.5
	"10% Rule"	50	63.3	62.2	64.5
	Fully Boosting	50	57.7	56.5	58.8

Table 4. Neural Net: Comparison of Overall Accuracy and Target Accuracy between Sampling Strategies

				95% Confidence Interval	
		N	Mean	Lower Bound	Upper Bound
Overall Accuracy	Fully Reducing	50	70.4	70.1	70.8
	"10% Rule"	50	73.2	73	73.3
	Fully Boosting	50	73.6	73.4	73.7
Target Accuracy	Fully Reducing	50	69.6	69.1	70.2
	"10% Rule"	50	68.6	68	69.1
	Fully Boosting	50	68.1	67.8	68.6

4. DISCUSSION AND GENERAL RECOMMENDATION

Increasing the sampling percentage of the majority class in training data set enhances overall accuracy of the testing data. This makes intuitive sense because the larger the sampling percentage, the better it represents its original population. On the other hand, as the sampling size of the majority increases, it requires more replicates of minority class to maintain the approximate 50/50 balance. As target being artificially enlarged, classifier becomes over confident with the training data, which results in an over-fitted classifier that does not perform well with testing data. For skewed data, overall accuracy alone is no longer the proper criteria for assessing the quality of a classifier. For example, when the majority and minority class ratio of the target is 99:1, even random selection would achieve a very good overall accuracy. In such cases the goal of modeling is to boost up target accuracy while maintain a good level of overall accuracy. In this study we've demonstrated that the "10% Rule" sampling strategy seems to work well to achieve this goal. We further demonstrate that modeling algorithms react differently to sampling variations.

Fully Boosting does not work well with Neural Net algorithm. The many fold copies of minority class brought in by this sampling strategy seems to be confusing to the "neurons" of the artificial neural network. The result is a dramatic drop of target accuracy. Similarly, CHAID also shows significant decrease in target accuracy for Fully Boosting, albeit the decrease is not as steep as neural net. This renders the Fully Boosting method unfavorable for both neural net and tree algorithms. For Neural Net, the "10% Rule" approach is a clear winner because it brings a jump of overall accuracy while offering the same target accuracy, compared to Fully Reducing strategy. As for CHAID, it seems safe to choose either Fully Reducing or the "10% Rule" strategies, because no significant differences are observed either between the overall accuracies or the target accuracies.

Among the three modeling algorithms studied, Logistic Regression demonstrates most robustness to sampling variations. There are significant changes among the three sampling strategies tested, but the incremental changes are small. It is also noticed that the standard deviations for the mean of model accuracies are also smaller, compared to those of the other two algorithms. This confirms from another angle the robustness of the logistic regression classifier to sampling methods. While it seems either of the three sampling strategies are likely to produce satisfactory logistic regression classifiers, the preferred choice again falls on the "10% Rule" approach, in that it generate a larger increase of overall accuracy at the cost of barely detectable decrease in target accuracy, compared to Fully Reducing.

While algorithms can be domain sensitive, the authors observed that generally speaking, Logistic Regression algorithm works better with extremely small binary target, that is, binary target that is smaller than 5%.

REFERENCES

Weiss, G.M. 2004. Mining with Rarity: A Unifying Framework. In Sigkdd Explorations, 6(1), 7-19. [1]

Longadge, R., S.S. Dongre, and L. Malik. 2013. Class Imbalance Problem in Data Mining: Review. In *International Journal of Computer Science and Network*, 2(1), 2277-5420. [2]

Sun, Y., M.S. Kamel, A.K.C. Wong, and Y. Wang. 2007. Cost-sensitive boosting for classification of imbalanced data. In *Pattern Recognition*, 40, 3358-3378. [3]

Seiffert, C., T.M. Khoshgoftaar, J.V. Hulse, and A. Napolitano. 2008. Building Useful Models from Imbalanced Data with Sampling and Boosting. In *Proceedings of the Twenty-First International FLAIRS Conference*, 306-311. [4]

Batista, G.E.A.P.A., R.C. Prati, and M.C. Monard. 2004. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. In *Sigkdd Explorations*, 6 (1), 20-29. [5]

Chawla, N.V. 2005. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Handbook*, pp 853-867. [6]

Submitting author

Zhen Zhang, Ph.D., Department of Marketing, 1018 Highland Colony Parkway, Ridgeland, MS 39157, USA. Phone (601) 540-7157
E-mail: zzhang@cspire.com

Acknowledgements

This work is supported by C Spire, the nation’s largest privately held wireless communication & technology company.