MATTER: International Journal of Science and Technology
ISSN 2454-5880

Chang et al., 2018

Volume 4 Issue 3, pp. 11-24

Date of Publication: 15th November, 2018

DOI-https://dx.doi.org/10.20319/mijst.2018.43.1124

This paper can be cited as: Chang, M., Dalpatadu, R. J., Phanord, D., & Singh, A. K. (2018). A Bootstrap Approach for Improving Logistic Regression Performance in Imbalanced Data Sets. MATTER: International Journal of Science and Technology, 4(3), 11-24.

A BOOTSTRAP APPROACH FOR IMPROVING LOGISTIC REGRESSION PERFORMANCE IN IMBALANCED DATA SETS

Michael Chang
Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, United States of America (USA)
changm13@unlv.nevada.edu

Rohan J. Dalpatadu
Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, United States of America (USA)
dalpatad@unlv.nevada.edu

Dieudonne Phanord
Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, United States of America (USA)
phanordd@unlv.nevada.edu

Ashok K. Singh
William F. Harrah College of Hotel Administration, University of Nevada Las Vegas, Las Vegas, United States of America (USA)
aksingh@unlv.nevada.edu

Abstract

In an imbalanced dataset with binary response, the percentages of successes and failures are not approximately equal. In many real world situations, majority of the observations are "normal" (i.e., success) with a much smaller fraction of failures. The overall probability of correct classification for extremely imbalanced data sets can be very high but the probability of

correctly predicting the minority class can be very low. Consider a fictitious example of a dataset with 1,000,000 observations out of which 999,000 are successes and 1,000 failures. A rule that classifies all observations as successes will have very high accuracy of prediction (99.9%) but the probability of correctly predicting a failure will be 0. In many situations, the cost associated with incorrect prediction of a failure is high, and it is therefore important to improve the prediction accuracy of failures as well. Literature suggests that over-sampling of the minority class with replacement does not necessarily predict the minority class with higher accuracy. In this article, we propose a simple over-sampling method which bootstraps a subset of the minority class, and illustrate the bootstrap over-sampling method with several examples. In each of these examples, an improvement in prediction accuracy is seen.

Keywords

Binary Response, Prediction, SMOTE, Under-sampling, Over-sampling, Confusion Matrix, Accuracy, Precision, Recall, F1-measure

1. Introduction

The study of rare events is quite common in many disciplines and the use of conventional logistic regression in such cases has been questioned by many researchers. (King & Zeng, 2001) proposed a modification which involved using logistic regression with permutational distributions of the sufficient statistics for statistical inferences; they suggest alternative sampling schemes that involve sampling all available events (e.g., wars) and a fraction of nonevents (e.g., peace). This idea of under-sampling non-events to obtain a more balanced sample for logistic regression has been investigated (Chawla, Bowyer, Hall, and Kegelmeyer 2002); it was shown that a combination of under-sampling the majority class and over-sampling the minority class yields better results that over-sampling alone.

A question that is related to the study of rare events is: how many occurrences of a rare event are needed to obtain a reasonable logistic regression model. The problem of determining the number of events per predictor has been investigated using Monte Carlo simulation (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996), Concato & Feinstein, 1997); these studies confirm a rule of thumb that requires 10-20 events per predictor. (Vittinghoff & McCulloch, 2007) conducted a large simulation study and found a range of conditions in which confidence interval coverage and bias were acceptable even with less than 10 predictors per event, and

concluded that this thumb rule was too conservative. It has since been pointed out (Allison, 2012) that it is not really the rarity of the event but the small number of occurrences of the event that causes problems in estimation.

The method of under-sampling and over-sampling is used in credit scoring for prediction of binary response (Crone and Finlay 2012; García and Sánchez 2012; Namvar, Siami, Rabhi, and Naderpour 2018).

In the present article, we propose a method that involves using bootstrap (Efron and Tibshirani, 1986, 1991) on a subset of minority class cases in order to balance the data set in order to improve the predictive performance of the logistic regression model. The method is illustrated with several examples.

2. Literature Review

Data mining is the process of finding useful and actionable relationships within data sets that are long and wide, with the goal of predicting outcomes of interest, and is now commonly used in a very wide range of disciplines (Fayyad 2001; Keleş 2017). Machine learning methods are used in healthcare (Singh, 2018). Syaifudin and Puspitasari (2017) used Naïve Bayes method of classification and also the Natural Language Toolkit (NLTK) for natural language processing on data collected from Twitter in their research on Public Policy. Catanghal Jr, Palaoag and Malicdem (2017) used data mining on twitter feeds for assessing needs of a disaster hit community. Cho and Kim (2015) develop a machine learning model for evaluating video games using opinion data provided in Korean by the users. Wei and Dunbrack (2013) investigate the role of balancing training and test sets for binary classifiers in Bioinformatics. A survey of resampling techniques for improving classification performance in unbalanced datasets is available in the literature (More, 2016; Dhurjad and Banait, 2014).

3. Selective Bootstrap

The proposed method consists of first fitting a logistic regression model to the full data, predicting the binary response for each observation, and determining all observations for which the minority class was predicted correctly. This subset of the minority class is then over-sampled to obtain a balanced data set. The method is briefly described below:

Step 1: Fit a logistic regression to the full data set, and use the fitted model to predict the binary response *Y*; let $\hat{Y}$ denote the predicted response, and

$$I_{i,j} = \left\{ k \,|\, (Y_k = i) \text{ and } \left(\hat{Y} = j\right);\ i = 0,1 \text{ and } j = 0,1 \right\} \quad (1)$$

The set of indices $I_{0,0}$ corresponds to all observations for which the minority class $(Y = 0)$ is correctly predicted.

Step 2: The full data set is split into a 75% training set (TRAIN0) and a 25% test set (TEST0), and the observations in the set $I_{0,0}$ are oversampled using bootstrap to obtain a balanced data set *X*; this balanced data set *X* was next split into a 75% training set (TRAIN1) and a 25% test set (TEST1). Fit a logistic regression to the training set TRAIN1, and evaluate the logistic regression classifier on both the training set TRAIN1 and the test set TEST1 using the performance measures described below.

2.1 Performance Measures for Prediction

A large number of performance measures for multi-level classifiers exist in machine learning literature (Sokolova & LaPalme, 2009). Accuracy, precision, recall and the geometric mean F1 of precision and recall are commonly used (Guillet & Hamilton, 2007; James, Witten, Hastie, & Tibshirani, 2013). In order to compute these measures, we first need to calculate the confusion matrix. In the case of predicting a response with *K* levels, the Confusion Matrix will be a *K* x *K* matrix as shown in Table 1.

Table 1: The Confusion Matrix

PREDICTED RESPONSE	TRUE RESPONSE				
	1	2	…	*K*-1	*K*
1	$N_{1,1}$	$N_{1,2}$	…	$N_{1,K-1}$	$N_{1,K}$
2	$N_{2,1}$	$N_{2,2}$	…	$N_{2,K-1}$	$N_{2,K}$
…	…	…	…	…	…
K-1	$N_{K-1,1}$	$N_{K-1,2}$	…	$N_{K-1,K-1}$	$N_{K-1,K}$
K	$N_{K,1}$	$N_{K,2}$	…	$N_{K,K-1}$	$N_{K,K}$

where $N_{i,j} =$ number of times true response of j gets predicted as i $(i, j = 1, 2, ..., K)$.

The performance measures Accuracy, Precision, Recall and $F1$ are calculated for each category j $(j = 1, 2, ..., K)$, from the following formulas (Guillet & Hamilton, 2007):

$$\text{Accuracy} = \frac{\sum_{i=1}^{K} N_{i,i}}{\sum_{i=1}^{K}\sum_{j=1}^{K} N_{i,j}} \quad (2)$$

$$\text{Precision}_i = \frac{N_{i,i}}{\sum_{j=1}^{K} N_{i,j}} \quad (3)$$

$$\text{Recall}_j = \frac{N_{j,j}}{\sum_{i=1}^{K} N_{i,j}} \quad (4)$$

$$F1_i = \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (5)$$

For binary response problems, the Confusion Matrix reduces to a 2x2 matrix shown in Table 2.

Table 2: Binary Response Confusion Matrix

PREDICTED RESPONSE	TRUE RESPONSE	
	0	1
0	$N_{0,0}$	$N_{0,1}$
1	$N_{1,0}$	$N_{1,1}$

The accuracy in the binary response case reduces to:

$$\text{Accuracy} = \frac{N_{00} + N_{11}}{N_{00} + N_{01} + N_{10} + N_{11}} \quad (6)$$

Precision and Recall for category 0 are given by:

$$\text{Precision}_0 = \frac{N_{0,0}}{N_{0,0} + N_{0,1}} \quad (7)$$

$$\text{Recall}_0 = \frac{N_{0,0}}{N_{0,0} + N_{1,0}} \quad (8)$$

Similarly, Precision and Recall, for category 1 are given by:

$$\text{Precision}_1 = \frac{N_{1,1}}{N_{1,0} + N_{1,1}} \quad (9)$$

$$\text{Recall}_1 = \frac{N_{1,1}}{N_{0,1} + N_{1,1}} \quad (10)$$

The $F1$ measures are given by:

$$F1_i = \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i};\ i = 0,1. \quad (11)$$

3. Example

To illustrate the proposed method, the breast cancer survival data set (Bozorgi, Taghva, & Singh, 2017) is used. The pre-processed data of 338596 observations on the binary response variable (breast cancer survivability) and 19 predictors has 38381 cases (11.34%) of response 0 and 300215 cases (88.66%) of response 1, and is clearly unbalanced. Table 3.1 (from Bozorgi, Taghva, & Singh, 2017) shows a brief explanation of predictors; the predictors race, marital status, grade, and radiation are categorical, and age (at diagnosis), tumor size, csEODTumorSize, regionalNodesPositive, csEODExtension, and regionalNodesExamined are continuous.

Table 3: Explanation of Predictors

Variable	Variable Definition	Values
patientIdNumber	Patient ID	up to 8 digits
race	race identifier	01-99, white = 01, black = 02
maritalStatus	one digit code for marital status	1-9, single = 1, married = 2, etc.
behaviorCode	code for benign etc.	0-4, benign = 0, malignant potential = 1, etc.
grade	cancer grade	1-9, Grade I = 1, etc.
vitalStatusRecord	alive or not	1-4, alive = 1, dead = 4
histologicType	microscopic composition of cells	4-digit code

csExtension	extension of tumor	2-digits code
csLymphNode	involvement of lymph nodes	2-digits code
radiation	radiation type code	0-9, none = 0, Beam = 1, etc.
SEERHistoricStageA	codes for stages	0-9, in situ = 0, localized = 1
ageAtDiagnosis	First diagnosis age	00-130, actual age, missing = 999
csTumorSize	size in millimeters	000-888, no tumor = 000
regionalNodesPositive	negative vs positive nodes	00-99, number of positive nodes
regionalNodesExamined	positive and negative nodes examined	00-99, number
survivalMonths	number of months alive	000-998, number of months, for missing = 9999
COD	Cause of Death	5-digit code, breast cancer = 2600, alive = 00000
yearOfDiagnosis	This visit year	4-digits code

3.1 Step 1 (Fit logistic model to full data)

This large data set was split into a training set and a test set by randomly selecting 25% of the observations for the test set. We will refer to this training set as TRAIN0. To establish a baseline for precision, recall, and F1, we first fitted a logistic regression model to the binary response Y (breast cancer survivability). In order to address the issue of multicollinearity among predictors, generalized variance inflation factor (GVIF) values (Fox & Monette, 1992) were computed and predictors with GVIF above 5 were removed, and then statistically insignificant predictors were removed to obtain the final logistic regression model for the full data. Table 3.2 shows the final logistic model, and Table 3.3 shows the GVIF values for the explanatory variables in the model; all GVIF values are close to 1, indicating that there is no multicollinearity in the fitted model.

Table 4: Final logistic regression model for the entire data

Predictor	Estimate	SE	z-value	P-value
Intercept	4.76	0.05	105.45	0.00
race_2	-0.55	0.02	-30.62	0.00

race_Other	0.24	0.02	9.78	0.00
maritalStatus_2	0.26	0.02	13.70	0.00
maritalStatus_4	-0.01	0.02	-0.40	0.69
maritalStatus_5	-0.18	0.02	-7.93	0.00
maritalStatus_Other	0.16	0.03	4.80	0.00
grade_2	-1.09	0.03	-35.22	0.00
grade_3	-2.18	0.03	-72.82	0.00
grade_4	-1.68	0.04	-39.45	0.00
grade_9	-1.35	0.03	-42.79	0.00
radiation_1	0.39	0.01	31.51	0.00
radiation_2	1.59	0.14	11.74	0.00
radiation_5	0.28	0.10	2.85	0.00
radiation_8	-0.02	0.04	-0.47	0.64
radiation_9	-0.28	0.13	-2.09	0.04
ageAtDiagnosis	-0.01	0.00	-21.96	0.00
csEODTumorSize	0.00	0.00	-52.06	0.00
regionalNodesPositive	-0.01	0.00	-68.36	0.00
csEODExtension	-0.01	0.00	-79.78	0.00
regionalNodesExamined	-0.02	0.00	-40.30	0.00

Table 5: *The GVIF values of predictors in the final logistic regression model based on the entire data*

	GVIF	Df	GVIF^(1/(2*Df))
Categorical_race	1.06	2.00	1.02
Categorical_maritalStatus	1.40	4.00	1.04
Categorical_grade	1.12	4.00	1.01
Categorical_radiation	1.05	5.00	1.00
ageAtDiagnosis	1.42	1.00	1.19

csEODTumorSize	1.03	1.00	1.01
regionalNodesPositive	1.18	1.00	1.09
csEODExtension	1.03	1.00	1.02
regionalNodesExamined	1.10	1.00	1.05

The Confusion Matrix for the full model using the entire data set is shown in Table 6.

Table 6: *Confusion Matrix*

PREDICTED RESPONSE	TRUE RESPONSE		
	0	1	Total
0	5160	33221	38381
1	2496	297719	300215
Total	7656	330940	338596

The Precision, Recall, and *F*1 values for category 1 are all excellent (Table 7), but Precision and *F*1 for category 0 are quite poor (Table 8).

Table 7: *Category 1 precision, recall and F1 of the final logistic regression model for the full data set*

Precision	99.17%
Recall	89.96%
*F*1	94.34%

Table 8: *Category 0 Precision, Recall, and F1 of the final logistic regression model for the full data set*

Precision	13.44%
Recall	67.40%
*F*1	22.42%

It is worth mentioning that the above results are as to be expected since 88.66% of the observations in the full data set correspond to the majority class (*Y*=1) and only 11.33% are in the minority class (*Y*=0), and therefore it is easier to predict the survival of a breast cancer patient but it is harder to predict that a patient will not survive.

3.2 Step 2: (Selective bootstrap)

The set $I_{0,0}$ of observations for which both the observed and predicted Y are 0 turned out to have 5160 observations:

$$I_{0,0} = \left\{ k \mid (Y_k = 0) \text{ and } \left(\hat{Y} = 0\right) \right\} \quad (12)$$

The training set of 75% of all observations was randomly selected from the full data; this training set has $n_0 = 28,724$ failures (0) and $n_1 = 225,223$ successes (1). The set $I_{0,0}$ was bootstrapped $n_1 - n_0 = 196,499$ times, and these observations were combined with the training set TRAIN0 to get a balanced data set *X* of 450446 observations. The balanced data set *X* was split in a training set TRAIN 1 of 75% of rows in *X*, and test set TEST1 of the remaining rows. Table 9 shows the logistic regression obtained, Table 10 shows the GVIF values of the predictors in the model, and Tables 11 and 12 display the confusion matrices obtained from this training and test sets TRAIN1 and TEST1.

Table 9: *Final logistic regression model for the balanced training data set (TRAIN1)*

Predictor	Estimate	SE	z-value	P-value
(Intercept_	6.27	0.06	104.60	0.00
race_2	-0.95	0.02	-47.82	0.00
race_Other	0.37	0.03	11.48	0.00
maritalStatus_2	0.34	0.02	14.77	0.00
maritalStatus_4	-0.04	0.03	-1.48	0.14
maritalStatus_5	-0.34	0.03	-12.49	0.00
maritalStatus_Other	0.23	0.04	5.60	0.00
grade_2	-1.30	0.04	-29.07	0.00
grade_3	-2.88	0.04	-67.66	0.00
grade_4	-2.05	0.06	-34.75	0.00
grade_9	-1.56	0.04	-34.83	0.00
radiation_1	0.63	0.02	39.29	0.00
radiation_2	2.21	0.20	11.30	0.00

radiation_5	0.52	0.13	3.99	0.00
radiation_8	0.08	0.05	1.68	0.09
radiation_9	-0.47	0.15	-3.17	0.00
ageAtDiagnosis	-0.02	0.00	-30.05	0.00
csEODTumorSize	0.00	0.00	-129.24	0.00
regionalNodesPositive	-0.02	0.00	-113.48	0.00
csEODExtension	-0.01	0.00	-151.88	0.00
regionalNodesExamined	-0.03	0.00	-95.26	0.00

Table 10: *The GVIF values of predictors in the logistic regression fitted to the balanced training set TRAIN1*

	GVIF	Df	GVIF^(1/(2*Df))
Categorical_race	1.08	2.00	1.02
Categorical_maritalStatus	1.52	4.00	1.05
Categorical_grade	1.11	4.00	1.01
Categorical_radiation	1.06	5.00	1.01
ageAtDiagnosis	1.53	1.00	1.24
csEODTumorSize	1.05	1.00	1.02
regionalNodesPositive	1.16	1.00	1.07
csEODExtension	1.03	1.00	1.02
regionalNodesExamined	1.06	1.00	1.03

Table 11: Confusion Matrix for TRAIN1

PREDICTED RESPONSE	TRUE RESPONSE		
	0	1	Total
0	153381	15506	168887
1	5700	163248	168948
Total	159081	178754	337835

Table 12: Confusion Matrix for TEST1

PREDICTED RESPONSE	TRUE RESPONSE		
	0	1	Total
0	51101	5235	56336
1	1876	54399	56275
Total	52977	59634	112611

Table 13 shows the Precision, Recall, and *F*1 values computed using the confusion matrices of Tables 11 and 12; Table 13 clearly shows that the performance of the logistic classifier has improved using the proposed approach.

Table 13: Precision, Recall, and *F*1 measures for the TRAIN1 and TEST1

CATEGORY	PRECISION	RECALL	*F*1
TRAIN1 - 1	0.91	0.97	0.94
TRAIN1 - 0	0.96	0.91	0.94
TEST1 - 1	0.91	0.97	0.94
TEST1 - 0	0.96	0.91	0.93

References

Allison, Paul (2012, February 13). *Statistical Horizons.* Retrieved from https://statisticalhorizons.com/logistic-regression-for-rare-events

Bozorgi, Mandana, Taghva, Kazem, & Singh, Ashok (2017). Cancer Survivability with Logistic Regression. Computing Conference 2017 (18-20 July 2017) London, UK. https://ieeexplore.ieee.org/document/8252133/citations

Catanghal Jr, R. A., Palaoag, T. D. and Malicdem, A. R. (2017). Crowdsourcing approach for disaster response assessment. Matter: International Journal of Science and Technology.

Chawla, Nitesh V. (2005). Data Mining and Knowledge Discovery Handbook. Maimon, Oded, Rokach, & Lior (Eds.), *Data mining for imbalanced data: an overview*, (pp. 853-867). New York, Springer.

Chawla, N. V., Bowyer, K, Hall, L, & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

Concato J, Feinstein AR (1997). Monte Carlo methods in clinical research: applications in

multivariable analysis. Journal of Investigative Medicine, 45(6), 394-400.

Crone, S. F. and Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. International Journal of Forecasting 28 224–238.

Efron, B. and Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Volume 1, Number 1, pp. 54-75.

Efron, B. and Tibshirani, R. (1991). Statistical Data Analysis in the Computer Age. Science, Vol. 253, pp. 390-395.

Fox, John & Monette, Georges. (1992). Generalized collinearity diagnostics. Journal of the American Statistical Association, 87(417), 178-183.

Guillet, F., & Hamilton, H., J. (Eds.). (2007). Quality measures in data mining. (Vol.43). New York: Springer.

Keleş, Mümine Kaya (2017). An overview: the impact of data mining applications on various sectors. Technical Journal 11, 3(2017), 128-132.

King, Gary & Zeng, Langche. (2001). Logistic Regression in Rare Events Data. Political Analysis, 9(2), 137-163.

Namvar, A., Siami, M., Rabhi, F., Naderpour, M. (2018). Credit risk prediction in an imbalanced social lending environment. arXiv preprint arXiv:1805.00801

Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, 49 (12), 1373-1379.

Ramadhan, M. M., Sitanggang, I. S. and Anzani, L. P. (2017). Classification model for hotspot sequences as indicator for peatland fires using data mining approach. Matter: International Journal of Science and Technology, Special Issue Volume 3 Issue 2, pp. 588-597.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management, *45*(4), 427–437.

Syaifudin, Y. W. and Puspitasari, D. (2017). Twitter data mining for sentiment analysis on Peoples feedback against government public policy. Matter: International Journal of Science and Technology, Special Issue Volume 3 Issue 1, pp. 110 – 122.

Vittinghoff, Eric & McCulloch, Charles E. (2007). Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression, American Journal of Epidemiology, 165(6), 710–718.

https://doi.org/10.1093/aje/kwk052

Wei Q, Dunbrack RL Jr (2013) The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. PLoS ONE 8(7): e67863.