Innovativity in Modeling and Analytics Journal of Research
vol. 5, 2020, pp.1-8, Published online: 15 May 2020
http://imajor.info/JDA

©*iMAJOR*
ISSN 2534-9619

The Bootstrap Procedure in Machine Learning Problems: Summary of Practical Advantages

Borislava P. Vrigazova
Sofia University St. Kliment Ohridski
Faculty of Economics and Business Administration
Bulgaria
vrigazova@uni-sofia.bg

Abstract. The aim of this paper is to summarize the practical advantages of the bootstrap procedure in machine learning problems. Although the bootstrap is not a novel resampling method, its advantages in machine learning problems have not been researched in detail. We did series of research on its practical advantages and highlight a few that can boost classification problem's performance. Our findings show that the bootstrap can decrease computing time, while increase prediction accuracy. This makes it a reliable alternative to cross validation in classification problems with or without variable selection methods.

Key Words: the bootstrap, classification, ANOVA.

Subject Classification Codes: C38 ,C52, C55

1 Literature review

The most common way to identify the value of the shrinkage parameter in the lasso [1], ridge [2] and adaptive lasso [3] is cross-validation [5]. Cross-validation reduces the possibility of overfitting by choosing the value of the shrinkage parameter that reduces the number of nonzero coefficients. Cross validation is also used in Support Vector Machines [6] and in feature selection for Support Vector Machines [4]. Cross validation can be used for splitting the dataset into training and test set to validate prediction in machine learning. Cross validation, however, can result in slower classification with the increase of the size of the dataset [7].

The bootstrap, on the other hand, is widely used for calculating confidence intervals for regression estimation [8]. Some researchers show that it can also be applied to surveys [9], non-parametric prediction [11] and quantile regression [10]. Improvements of existing classification method choose cross validation as the resampling method [12]. Luo [13] used the bootstrap procedure to build an algorithm for selecting the right number of clusters in pattern categorization. Traditional applications of the bootstrap procedure do not involve being used as alternative to cross validation in classification problems.

Instead, the asymptotic properties of the bootstrap make it appropriate to be used in multiplicative error models [14]. It is also used to correct the behaviour of certain statistical tests in the presence of dependencies [15]. Zou [15] devised the block bootstrap method for

detecting the seasonal unit roots in various economic time series. The bootstrap also finds applications in other fields like renewable energy. Li [16] used the bootstrap procedure to estimate confidence intervals in thermal security region of bulk power grid.

Cross validation, however, is most often used for model selection in various fields. Kerbaa [18] used cross validation as model selection technique in sea radar clutter used for adaptive target detection. Li [17] used cross validation to reconstruct state traffic and match missing data between probe and stationary sensors. The property of model selection in cross validation is the underlying technique in this research, Lobo [19] proposed a version of cross validation that choose the best geospatial model with decreased computing time.

As academic literature suggests, the practical applications of cross validation and the bootstrap contain different purposes. Cross validation is used as model selection technique, while the bootstrap as a confidence interval estimation technique. However, the research we conducted led to the finding that in machine learning problems the bootstrap and cross validation can be used for similar purposes. Furthermore, we show that the bootstrap can be better alternative to cross validation in classification problems.

The rest of the paper is organized as follows: Section 2 summarizes our methodology, section 3 presents the results. Section 4 concludes.

2 Methodology

We have used the bootstrap as a resampling method in classification problems using two approaches. The first one involved applying the bootstrap in ANOVA-classification problems to identify the number of features that result in the best classification metrics and accuracy. The second approach focused on studying the behavior of the Support Vector Machines with bootstrap with non-standard rule for train/test split.

2.1 The bootstrap for ANOVA classification

We follow the approach in [7] to perform ANOVA-bootstrapped classification on several datasets and compare the performance of the modified classification methods (A-BOOT LR, DC, KNN) to their ANOVA cross-validated versions (ANOVA LR, DC, KNN). We compare the performance of the logistic regression, decision tree classifier and the k-nearest neighbour. In this approach, we do not apply preliminary transformation on the dataset. We fit ANOVA classification models and replace the ten-fold cross validation with the bootstrap procedure. To split the dataset into training and test set we use 70/30 rule. We explore the classification metrics, error rate, time and prediction accuracy of each model. Our experiments show that without preliminary transformations of data, the bootstrap produced similar accuracy and classification metrics to those from the tenfold cross validation. The computing time, however, decreased significantly. We also run the standard versions of the classifiers without ANOVA or any feature selection and using tenfold cross-validation (standard LR, DC, KNN)

In [20] we conducted similar research concerning the Support Vector Machines. However, we applied transformations on datasets prior to applying the bootstrap in the ANOVA-SVM. We show that not only the resampling method matters, but also the preliminary data

transformations for improving the accuracy of the model. The time for fitting the ANOVA-SVM with the bootstrap decreased compared to other modifications of the ANOVA-SVM. The improvement of accuracy and classification metrics were the result from applying various types of preliminary transformations on input data.

2.2 The bootstrap in Support Vector Machines

In this study [21] we tested the performance of the bootstrap procedure on raw data without variable selection procedure. We explored how the bootstrap procedure affects the behaviour of the Support Vector Machines (SVM). We fitted SVMs with the bootstrap procedure as resampling technique and split the dataset into training and test set using 30/70 rule. We compared its performance to SVMs fitted with leave-one-out cross validation, random train/test split and tenfold cross validation. In all experiments, we fix C=1. A key finding from our research is that the bootstrap can optimize the performance of the SVMs by decreasing computing time and improving accuracy. The improvement of accuracy in this case was related to the smaller number of features chosen as a result of bootstrapped ANOVA procedure.

Despite this, our experiments are sensitive to the value of the parameter C in the Support Vector Machines. In all experiments we used ten iterations of the bootstrap procedure. Thus, we called the procedure the tenfold bootstrap. We show in [7] that ten iterations are enough to optimize the performance of the classification model, which is the biggest advantage of our approaches.

3 Results

3.1 The bootstrap for ANOVA classification

Table 1 presents the results from the classic ANOVA cross validated logistic regression, k-nearest neighbor, decision tree and the support vector classifier (ANOVA LR, DC, KNN, SVM). We compare those results to the results from their ANOVA-bootstrapped versions (A-BOOT LR, DC, KNN) and their cross-validated classical versions without ANOVA (standard LR, DC, KNN). We perform experiments on the adult dataset , fraud and the cover dataset. The datasets were retrieved from [23].

As table 1 shows the standard logistic regression that uses untransformed data, cross validation and no feature selection method resulted in similar accuracy to the cross-validated and bootstrapped ANOVA models. In the case of the fraud dataset, the bootstrapped ANOVA logistic regression improved the accuracy compared to the other two models. The ANOVA-bootstrapped LR was able to retain similar accuracy or improve it compared to its standard versions. However, the computing time in the case of the bootstrapped ANOVA logistic regression was much lower than the classical ones. The cover dataset is the biggest dataset, containing 581,013 observations. The standard logistic regression ran in 1686.25 seconds, while the bootstrapped version outperformed the standard and the ANOVA cross validated version, running in 109.55 seconds.

The computational advantage of the bootstrapped ANOVA was also retained in the decision tree classifier. Table 1 shows that it resulted in similar accuracy and error rate to the

Table 1: Performance measures of the Logistic regression, k-nearest neighbor, decision tree and support vector machines

dataset	method	accuracy	error rate	time (seconds)	% of features
Adult	standard LR	79.7%	20.3%	4.05	100
	ANOVA LR	79.7%	20.3%	0.47	50
	A-BOOT LR	79.1%	20.9%	0.16	100
Fraud	standard LR	92.8%	7.2%	4.93	100
	ANOVA LR	85.0%	15.0%	0.03	100
	A-BOOT LR	93.7%	6.3%	0.02	100
Cover	standard LR	66.1%	33.9%	1686.25	100
	ANOVA LR	62.7%	37.3%	153.09	40
	A-BOOT LR	66.1%	33.9%	109.55	100
Adult	standard DC	80.9%	19.1%	3.70	100
	ANOVA DC	81.8%	18.2%	0.65	70
	A-BOOT DC	80.5%	19.5%	0.14	100
fraud	standard DC	98.2%	1.8%	0.24	100
	ANOVA DC	96.9%	3.1%	0.03	100
	A-BOOT DC	97.7%	2.3%	0.02	100
cover	standard DC	91.6%	8.4%	105.87	100
	ANOVA DC	64.0%	36.0%	4.69	50
	A-BOOT DC	90.4%	9.6%	5.48	100
adult	standard KNN	77.4%	22.6%	5.28	100
	ANOVA KNN	77.7%	22.3%	14.77	20
	A-BOOT KNN	74.3%	25.7%	0.42	100
fraud	standard KNN	97.1%	2.9%	1.91	100
	ANOVA KNN	97.0%	3.0%	0.03	100
	A-BOOT KNN	97.4%	2.6%	0.02	100
cover	standard KNN	96.3%	3.7%	68.08	100
	ANOVA KNN	63.0%	37.0%	14400.00	40
	A-BOOT KNN	95.6%	4.4%	8.76	100

Table 2: Computing time algorithms 1-4.

			Time for fitting SVM (s)			
Dataset	n	p	Algorithm 1	Algorithm 4	Algorithm 2	Algorithm 3
glass	175	9	0.02	0.00	0.26	0.02
leaf	286	7	0.04	0.00	0.46	0.02
wells	3020	4	0.56	0.32	358.14	0.98
fraud	3255	4	3.01	0.59	726.29	2.12
abalone	4177	8	2.51	0.64	807.09	1.70
ed	5785	5	9.15	1.86	4623.18	6.43
monica	6367	11	5.21	0.93	2628.48	3.50
food	23971	5	763.03	30.33	>28800	337.05
adult	45222	13	2434.4	528.61	>28800	1757.79

standard versions of the decision tree classifier. However, the computing time was reduced greatly. The classic versions fitted the decision tree classifier in 3.70 and 0.65 seconds (adult data), while the bootstrapped version in 0.14 seconds. Interesting case is the case of the cover dataset, where the bootstrap managed to increase accuracy to 90.4% compared to the cross-validated ANOVA version, which resulted in 64% accuracy. The boostrapped ANOVA decision tree classifier resulted in accuracy close to the standard decision tree – 91.6% but at a reduced speed. Although the cover dataset should be investigated for overfitting and which model suits best the data, we are more interested in the algorithm's performance rather than the special characters of a particular dataset.

The computational advantage of the bootstrapped ANOVA k-nearest neighbor is visible in table 1. As the size of the dataset increases, the computational advantage of the bootstrap became more visible. The cross-validated KNN in the cover dataset ran in 14400 seconds, wile the bootstrapped version ran in 8.76 seconds. The computational advantage of the bootstrap in classification problems is an important finding as it can significantly shorten the time for prediction despite the increasing size of the dataset. This advantage is present on non-transformed data and your experiments show that it is preserved on transformed data as well.

3.2 The bootstrap in Support Vector Machines

We also explored the bootstrap's performance in Support Vector Machines [21]. We compared the boostrapped SVM (algorithm 4) with ten-fold cross-validated SVMs (algorithm 1), leave-one-out cross validation (algorithm 2) and repeated random train/test split (algorithm 4). Table 2 compares the computing time from the four algorithms.

As table 2 shows algorithm 4 was the fastest one. The leave-one-out cross validation and the random train/test split were the slowest. The bootstrapped SVMs reduced computing time around 4 times in the adult dataset compared to the ten-fold cross-validated version. The tenfold cross-validation is a faster method than the leave-one-out cross validation and the random train/test split. Despite this, the increase of the size of the dataset can slow

Table 3: Accuracy of algorithms 1 vs 4

Dataset	n	p	average accuracy algorithm 1	average accuracy algorithm 4
glass	175	9	0.7	0.62
leaf	286	7	0.58	0.69
wells	3020	4	0.54	0.54
fraud	3255	4	0.65	0.65
abalone	4177	8	0.53	0.53
ed	5785	5	0.87	0.87
monica	6367	11	0.88	0.87
food	23971	5	0.86	0.86
adult	45222	13	0.75	0.75

prediction down. The bootstrap avoids this advantage of the cross validation without loss of accuracy as table 3 shows.

As table 3 shows that the bootstrapped SVMs resulted in accuracy similar to that of the cross-validated SVMs. The bootstrapped version, however, was much faster. The same number of features and larger size of dataset did not slow the bootstrap down. This finding is in line with the results from the other classification methods. The bootstrap can accelerate prediction time in classification models without loss of accuracy.

In [20] and [22] we show that bootstrap can increase the accuracy on some datasets. Our experiments show that when the bootstrapped SVMs is applied with ANOVA variable selection, the accuracy and classification metrics can increase. This finding is particularly important in cancer datasets as it can improve the accuracy of detecting malignant and benignant cancer. All these findings show that the bootstrap can be used not only for finding confidence intervals but also as a reliable alternative of other resampling method.

4 Concluding Remarks

In this paper we summarized the practical advantages of the bootstrap as a resampling method for classification problems. These advantages include shortened computing time and improvement of accuracy that persist with the increase of the dataset. The review of our analysis shows that the bootstrap can be used as alternative to cross validation in classification problems, which widen the practical applications of the bootstrap.

References

[1] R. Tibshirani, *Regression shrinkage and selection via the lasso: a retrospective*, J. R. Statist.Soc.B, 73(3), 1996, 267–268.

[2] A. Hoerl and R., Kennard, *Ridge regression. Applications to nonorthogonal problems*, Technometrics, 12(1), 1970, 69–82.

[3] H. Zou, *The adaptive lasso and its Oracle properties*, Journal of the American Statistical Association, 101(746), 2006, 1418–1429.

[4] J. Weston and S. Mukherjee and O. Chapelle and M. Pontil and T. Poggio and V. Vapnik, *Feature selection for SVMs*, Advances in neural information processing systems, 2001, 668–674.

[5] T. Wong, *Parametric methods for comparing the performance of two classification algorithms evaluated by k-fold cross validation on multiple data sets*, Pattern Recognition, 2017, 65, 97–107.

[6] C. Cortes and V. Vapnik, *Support-vector networks*, Machine Learning, 1995, 20(3), 273–297.

[7] B. Vrigazova and I. Ivanov, *The bootstrap procedure in classification problems*, International Journal of Data Mining, Modelling and Management, 2020, 12(4), 273–297, in press.

[8] F. Hu and J. Kalbfleisch, *Estimating equations and the bootstrap*, Lecture Notes Monograph Series, Institute of Mathematical Statistics, 1997, 32, 405—416.

[9] J. Shao, *Impact of the bootstrap on sample surveys*, Statistical Science, Institute of Mathematical Statistics, 2003, 18, 191—198.

[10] X. Feng and X. He and J. Hu, *Wild bootstrap for quantile regression*, Biometrika, 2011, 98, 995—999.

[11] T. Fushiki and F. Komaki and K. Aihara, *Nonparametric bootstrap prediction*, International Statistical Institute (ISI) and Bernoulli Society for Mathematical Statistics and Probability., 2005, 11, 9293—307.

[12] P. Bugata and P. Drotar, *Weighted nearest neighbors feature selection*, Knowledge-Based Systems, 2019, 163, 749–761.

[13] X. Luo and X. Zhu and E. lim, *A parametric bootstrap algorithm for cluster number determination of load pattern categorization*, Energy, 2019, 180, 50–60.

[14] I. Perera and M. Silvapulle, *Bootstrap based probability forecasting in multiplicative error models*, Journal of Econometric, 2020, 180, in press.

[15] N. Zou and D. Politis, *Bootstrap seasonal unit root test under periodic variation*, Econometrics and Statistics, 2020, in press.

[16] X. Li and J. Jiang and G. Liu and L. Bai and H. Cui and F. Li, *Bootstrap-based confidence interval estimation for thermal security region of bulk power grid*, International Journal of Electrical Power & Energy System, 115, 2020.

[17] J. Li and K. Perrine and L. Wu and C. Walton, *Cross-validating traffic speed measurements from probe and stationary sensors through state reconstruction*, International Journal of Transportation Science and Technology, 8(3), 2019, 290–303

[18] T. Kerbaa and A. Mezache and H. Oudira, *Model Selection of Sea Clutter Using Cross Validation Method*, Procedia Computer Science, 158, 2019, 394-400

[19] V. Lobo and T. Fonseca and F. Moura, *Bayesian cross-validation of geostatistical models*, Spatial Statistics, 35, 2020.

[20] B. Vrigazova and I. Ivanov, *Optimization of the ANOVA Procedure for Support Vector Machines*, International Journal of Recent Technology and Engineering, 8(4), 2019, 5160-5165 (paper 869, ISSN: 2277-3878), https://www.ijrte.org/wp-content/uploads/papers/v8i4/D7375118419.pdf

[21] B. Vrigazova and I. Ivanov, *Tenfold bootstrap procedure for support vector machines*, Computer Science 21(2) 2020: 241–257. https://journals.agh.edu.pl/csci/article/view/3634

[22] B. Vrigazova, *Detection of Malignant and Benign Breast Cancer Using the ANOVA-BOOTSTRAP-SVM*, Journal of Data and Information Science, 5(2), 2020, 62–75. http://manu47.magtech.com.cn/Jwk3$_{-}$jdis/EN/10.2478/jdis-2020-0012

[23] www.kaggle.com