Received 17 July 2015; revised 16 May 2016; accepted 25 June 2016.
Date of publication 5 October 2016; date of current version 16 May 2017.

Digital Object Identifier 10.1109/LLS.2016.2615086

A Multimetric Evaluation of Stratified Random Sampling for Classification: A Case Study

GUNJAN S. THAKUR[1], BERNIE J. DAIGLE[2], Jr., MENG QIAN[3], KELSEY R. DEAN[4], YUANYANG ZHANG[5], RUOTING YANG[6], TAEK-KYUN KIM[7], XIAOGANG WU[7], MENG LI[3], INYOUL LEE[7], LINDA R. PETZOLD[8], AND FRANCIS J. DOYLE III[1]

[1]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[2]Departments of Biological Sciences and Computer Science, The University of Memphis, Memphis, TN 38152, USA
[3]Department of Psychiatry, New York University Medical Center, New York, NY 10016, USA
[4]Department of Systems Biology, Harvard Medical School, Cambridge, MA 02138, USA
[5]Department of Computer Science, University of California at Santa Barbara, Santa Barbara, CA 93106, USA
[6]U.S. Army Center for Environmental Health Research, Fort Detrick, Frederick, MD 21702, USA
[7]Institute for Systems Biology, Seattle, WA 98109, USA
[8]Departments of Computer Science and Mechanical Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106, USA

CORRESPONDING AUTHOR: G. S. THAKUR (gthakur@g.harvard.edu)

This work was supported in part by the Institute for Collaborative Biotechnologies at the University of California at Santa Barbara through a Grant from the U.S. Army Research Office under Grant W911NF-10-2-0111 and in part by USAMRMC under Grant 09284002.

ABSTRACT Accurate classification of biological phenotypes is an essential task for medical decision making. The selection of subjects for classifier training and validation sets is a crucial step within this task. To evaluate the impact of two approaches for subject selection—randomization and clinical balancing, we applied six classification algorithms to a highly replicated publicly available breast cancer data set. Using six performance metrics, we demonstrate that clinical balancing improves both training and validation performance for all methods on average. We also observed a smaller discrepancy between training and validation performance. Furthermore, a simple analytical argument is presented which suggests that we need only two metrics from the class of metrics based on the entries of the confusion matrix. In light of our results, we recommend: 1) clinical balancing of training and validation data to improve signal-to-noise ratio and 2) the use of multiple classification algorithms and evaluation metrics for a comprehensive evaluation of the decision making process.

INDEX TERMS Classification, confusion matrix, performance metrics, random sampling, stratified sampling.

I. INTRODUCTION

ADVANCES in high-throughput technologies have enabled the genome-wide measurement of a diverse set of molecular species from tissue samples of human subjects. Data from these measurements are frequently used to perform supervised classification (SC) for diagnostic applications. SC involves learning a class-assignment rule given data from subjects with known classes (''training set''). This rule is then applied to data from an independent set of subjects whose classes have been withheld (''testing set'') to evaluate classifier performance. Broadly speaking, there are three types of classifier performance metrics [1]: 1) those based on qualitative measures of error [e.g., accuracy (ACC), true positive rate (TPR), and false negative (FN) rate]; 2) those based on probabilistic measures of error (e.g., mean absolute prediction error, logarithmic loss, and mean squared prediction error); and 3) those based on relative errors in sample ranking (e.g., area under the receiver operating characteristic (ROC) curve). Regardless of the metric chosen, many studies have emphasized the importance of using truly independent data in the testing set to accurately assess classification performance. In cases, where this is not possible, proper cross-validation (CV) procedures (see the following for details) attempt to minimize the bias in estimated performance [2].

Depending on the disease of interest, some SC studies use training sets with unequal numbers of subjects in the disease classes, often referred to as class imbalance [3], [4]. Class imbalance is known to affect the performance of many classifiers, as training and evaluation procedures often implicitly assume that training samples are relatively balanced [3]. In particular, many classifiers are biased toward the largest class, i.e., they more accurately classify new subjects from the majority class. Two approaches are commonly used in SC to address the issues caused by class imbalance: 1) biased resampling of subjects from the training data and 2) modification of the classification algorithm to explicitly account

2332-7685 © 2016 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

for training data class imbalance. Biased resampling involves randomly oversampling the minority class(es) and/or randomly undersampling the majority class(es).

Given an imbalanced training data set, factors such as sample size, class separability (i.e., the degree of data separation between subjects in different classes), and intraclass stratification (i.e., the presence of biological subgroups within each class) can further affect classifier performance. To minimize the effects of the latter factor, clinical training and testing sets are often age and ethnicity matched across disease classes. This strategy can be extended to facilitate the matching of additional confounding clinical factors. In particular, many clinical trials [5], [6] balance treatment groups using variables that are known to influence the disease prognosis in a process known as stratified randomization [7]. For the development of prognostic models, clinical variables have also been used in conjunction with molecular data (e.g., in Cox regression models) to improve classifier performance [8]. Finally, clinical variables have been integrated into molecular Bayesian network results to improve performance. In this letter, we provide the first real-world demonstration (to the best of our knowledge) of the importance of clinically informed sampling for creating high-throughput training and testing data sets for SC. Specifically, we apply six different classification methods to predict metastasis using publicly available breast cancer data before and after clinically informed sampling, evaluating performance with six commonly used metrics. We also provide an analytical explanation for the minimum number of metrics based on the qualitative measures of error, needed to properly evaluate the performance of each classifier.

II. METHODS

We used a publicly available breast cancer metastasis data set [9] to quantify the effects of clinically informed sampling when creating training and testing sets for SC. The data set consists of gene expression data for 286 subjects with breast cancer that either did (107) or did not (179) develop metastasis within five years of surgery. We split the data set into two cohorts using two strategies. In the first approach, we partitioned data into training and testing data sets, each comprising of 100 subjects, by randomly undersampling subjects from the two classes (avoiding class imbalance). Each subset of data comprises of 50 subjects with metastasis and 50 subjects without metastasis. Second, we partitioned the data set into two cohorts using stratification randomization. Specifically, all 200 subjects selected by the random sampling strategy were redistributed into training and test sets, such that estrogen receptor (ER) status, tumor stage, age, progesterone receptor (PR) status, the outcome case status, and menopause status, were balanced. This procedure was performed using SAS 9.4 (SAS Institute, Cary, NC). Table 1 provides the number of subjects in different strata, defined by the clinical variables, present in the training and testing set obtained by the two partitioning approaches considered here. In order to estimate the classification performance in an unbiased manner, we applied nested CV (NCV) [10] to the training data for each classifier. NCV performs two levels of CV—internal and external—where the former is used to select features/classification parameters and the latter is used to evaluate performance. In the external (internal) CV loop, the complete (external) training data set is randomly split N_E (N_I) times into K_E (K_I) subsets. Each external (internal) training set is composed of $K_E - 1$ ($K_I - 1$) subsets and the remaining subset is used as an external (internal) test set. We computed actual performance metrics by training each classifier on the entire training data set using ordinary CV, followed by evaluating the corresponding classifiers on the entire testing data set. In the remainder of this section, we briefly describe the six classifiers used to assess the effects of clinically informed data sampling.

TABLE 1. **Number of subjects corresponding to different clinical variable in partitions obtained by two strategies—block ("random") and stratified ("balanced") randomization.**

Clinical Variables		Random			Balanced		
		"train."	"test"	diff.	"train."	"test"	diff.
ER status	ER+	79	73	6	75	72	3
	ER-	21	27	6	25	28	3
Tumor stage	I	49	52	3	50	48	2
	II	48	47	1	49	48	1
	III, IV	3	1	2	1	4	3
Age	20-50	41	51	10	43	44	1
	51-85	59	49	10	57	56	1
PR status	PR+	58	54	4	58	57	1
	PR-	40	39	1	39	40	1
Relapse	"0"	50	50	0	50	50	0
	"1"	50	50	0	50	50	0
Meno-pause	"pre"	50	62	12	55	55	0
	"post"	50	38	12	45	45	0

A. NEAREST SHRUNKEN CENTROIDS

We applied the nearest shrunken centroids (NSCs) classifier as implemented in the PAMR R package [11]. NSC efficiently performs both feature selection and classification while requiring the specification of only a single parameter. For the NCV procedure, we used $N_E = 10$, $K_E = 5$, $N_I = 1$, and $K_I = 50$.

B. SUPPORT VECTOR MACHINE

Using NCV with $N_E = 20$ and $K_E = 2$, the internal training data was processed according to the following three steps. Firstly, differential analysis using the R LIMMA package [12], we performed a moderated t-test to obtain informative genes (p-value < 0.01; intensity $> =$ average value for each gene). Secondly, feature selection applying the MATLAB ReliefF algorithm [13] with $N_I = 1$ and $K_I = 10$, we ranked all genes in order of informativeness. For ReliefF parameters, we set number of nearest neighbors $= 20$, sigma $= 5$, and default values for all others. For each internal training and testing data set, we selected the minimum set of genes from the top-ranked 200 that achieved the highest classification performance values [area under the curve (AUC) and ACC] for subsequent analysis. Lastly, performance evaluation—we implemented the support vector machine (SVM) model [14] in MATLAB to evaluate the top-selected features. For SVM parameters, we chose the nu-SVC classification type and a linear kernel, using default values for all others. This linear SVM diagnostic model with the top-selected features from step 2) was trained on the internal training data and tested on the internal testing data to evaluate classification performance. We fixed the soft margin constant (marginal parameter) at 1. Upon the completion of the internal step of NCV, the optimized biomarker panel was determined by selecting genes with the highest frequency of appearance in all sets of top-selected features. A classifier based on this panel was then trained using the samples in the external training set and evaluated on samples from the external testing set.

C. RANDOM FOREST

The "training data set" was first divided into 500 pairs of external training and validation sets. Then, within the external training sets, we further divided it into ten pairs of internal training and validation sets (the nested loop). We ranked the genes by the ascending moderated t-test p-values using the LIMMA R package on the internal training set, then trained random forest classifiers using CMA R package [15] and validated on top 10, 20,..., 100 genes, and finally, reported a local optimal feature length with maximum AUC using the ROCR R package. We then ranked the genes in the whole external training set, picked the top genes using the average of the ten local optimal feature lengths, and trained classifiers and validated on the corresponding external validation set. Multiple metrics, including AUC and error rate, were reported for each of the 500 external loops, and average scores were achieved as predicted performance.

D. GAUSSIAN PROCESSES CLASSIFICATION

Gaussian processes classification [16] treats the sample values in the training data set as samples from a multivariate Gaussian distribution, defining a Gaussian process operating on these values. More specifically, we define the second-order statistics of the Gaussian process as a kernel function of the data values and learn the process parameters by maximizing the posterior distribution of the data. When applying Gaussian processes to classification problems, we use a logistic function to convert the unbounded Gaussian output to a probability value within [0, 1]. In order to train the classifier, we use a Laplace approximation implemented in the R KERNLAB package [17]. For NCV, used $N_E = 10$, $K_E = 10$, $N_I = 1$, and $K_I = 10$.

E. COMBINER-BASED LINEAR DISCRIMINANT ANALYSIS

Yang *et al.* [18] developed COMBINER, a robust pathway based biomarker discovery tool. This algorithm takes multiple cohorts of high-throughput data as input and produces a panel of candidate biomarkers as output. The COMBINER algorithm performs two steps—inference (identify "driver genes" from a given pathway, using data from cohort 1) and recursive feature elimination (identify candidate biomarker pathways, using data from cohort 2). We extended the above-mentioned algorithm to also obtain a decision boundary for diagnostics. The breast cancer data set is divided into three subsets (cohorts). As before, inference is done on the first subset (cohort 1) and candidate biomarker pathways are identified on the second subset (cohort 2). Restricting ourselves only to candidate biomarker pathways, various classification performance metrics are estimated on the third subset by employing LDA and CV. The above-mentioned threefold partition is repeated ten times and the mean performance metrics are computed. Next, LDA coefficients on the subspace spanned by the most frequent biomarker pathways are computed for every threefold partitioning. The final decision boundary is obtained by taking the median of the above coefficients. The actual performance metrics are computed using the boundary obtained earlier on the testing data.

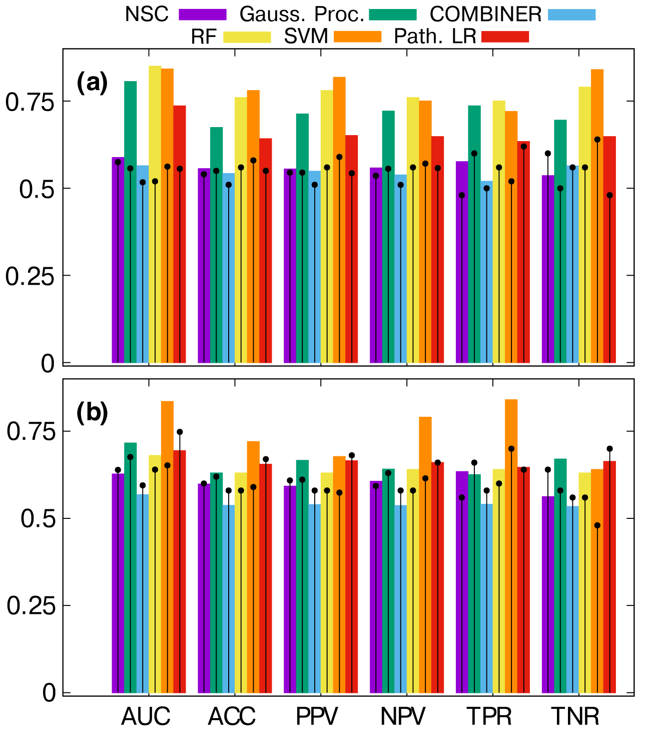

FIGURE 1. Breast cancer data set with partition. (a) Random block partitioning and (b) stratified random sampling based on clinical variables or balanced partition. Predicted performance metrics for six different SC algorithm shown as bars, and actual performance metrics on the test data set shown by stem plots overlapping their corresponding bars. Metrics used: AUC = area under the curve, ACC = accuracy, PPV = positive predictive value, NPV= negative predictive value, TPR = true positive rate, and TNR = true negative rate. Classifier used: NSC = nearest shrunken centroid, Gauss. Proc = Gaussian process regression, COMBINER = COre module biomarker identification with network exploration, RF = random forest, SVM = Support vector machine, and Path. LR = Pathifier-based logistic regression.

F. PATHIFIER-BASED LOGISTIC REGRESSION

As a second method for incorporating known gene network information, we used the pathifier method of Drier *et al.* [19] to project all expression data (training and testing) onto

KEGG pathway curves to compute "pathway deregulation scores." These scores indicate each sample's distance along the pathway principal curve, providing a dimensional reduction to 1320 features for each sample. All sample class labels (training and testing) were randomized during pathifier projection to ensure that no over-fitting occurs during this dimensional reduction stage. Next, we computed pathway scores and used them to train and test a logistic regression classifier using the CMA Bioconductor package. For NCV, we used $N_E = 100$, $K_E = 5$, $N_I = 1$, and $K_I = 3$, performing both feature selection and tuning of the L2 penalization parameter within each CV step.

III. RESULTS AND DISCUSSION

We employed six commonly used metrics to assess the performance of six different classifiers in this letter (see Fig. 1). The two main trends we observed in our results were: 1) the disagreement in the values of various predicted (bar plots) and actual (stem plots) classifier performance metrics is, on average, less pronounced in the case of the clinically balanced data set when compared with the unbalanced data set and 2) on average, the classification performance is improved when using the balanced data set. Using the Wilcoxon signed rank test, we compared the differences in the average performance for a given metric value, over all algorithms on the testing data. We obtained p-value of 0.0313 for AUC, ACC, negative predictive value (NPV), and TPR, 0.0625 for positive predictive value (PPV), and of 0.4375 for true negative rate (TNR).

Various studies have attempted to investigate the underlining relationship between different classification metrics using techniques, including correlation analysis [1] and factor analysis [20]. Ferri *et al.* [1] suggest that the AUC metric correlates well with others in the same class, i.e., those based on relative errors in sample ranking. However, the optimal number of representative metrics from other classes is not well established. In this letter, we have used metrics belonging to the sample ranking (AUC) and the qualitative error measure (ACC, PPV, NPV, TPR, and TNR) classes. Metrics belonging to the latter class represent different functions of the confusion matrix, which contains entries counting the numbers of TPs, false positives (FPs), TNs, and FNs. If we let A_1 and A_2 be the actual numbers of subjects in Class I and Class II (given the labels of the test data set) and P_1 and P_2 be the predicted numbers of subjects in the two classes (given the classifier predictions), we see that a total of four variables (TP, FP, TN, and FN) and four equations ($\mathrm{TP}+\mathrm{FN} = A_1$, $\mathrm{FP}+\mathrm{TN} = A_2$, $\mathrm{TP}+\mathrm{FP} = P_1$, and $\mathrm{FN}+\mathrm{TN} = P_2$) are operating in this system. This well-posed problem has only two degrees of freedom, suggesting that only two independent performance metrics from the qualitative error measures class are required.

IV. CONCLUSION

In this letter, we quantitatively evaluated the impact of stratified random partitioning (based on clinical variables) of the data into training and test data sets. In comparison with randomized partitioning, the former approach improved the overall performance of the six classifiers considered in this letter. It also reduced the discrepancy between predicted and actual performance, thus improving confidence in training set-based evaluations of classifier performance. Finally, using a simple analytical argument, we showed that two performance metrics from the qualitative error measures class are sufficient to capture all of the available performance information.

REFERENCES

[1] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognit. Lett.*, vol. 30, no. 1, pp. 27–38, 2009.

[2] A.-L. Boulesteix, "Over-optimism in bioinformatics research," *Bioinformatics*, vol. 26, no. 3, pp. 437–439, 2010.

[3] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit.*, vol. 23, no. 4, pp. 687–719, 2009.

[4] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

[5] M. Zelen, "The randomization and stratification of patients to clinical trials," *J. Chronic Diseases*, vol. 27, no. 7, pp. 365–375, 1974.

[6] W. N. Kernan, C. M. Viscoli, R. W. Makuch, L. M. Brass, and R. I. Horwitz, "Stratified randomization for clinical trials," *J. Clin. Epidemiol.*, vol. 52, no. 1, pp. 19–26, 1999.

[7] K. Suresh, "An overview of randomization techniques: An unbiased assessment of outcome in clinical research," *J. Human Reproduct. Sci.*, vol. 4, no. 1, pp. 8–11, 2011.

[8] J. S. Parker *et al.*, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *J. Clin. Oncol.*, vol. 27, no. 8, pp. 1160–1167, 2009.

[9] Y. Wang *et al.*, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[10] S. Kothari, J. H. Phan, A. N. Young, and M. D. Wang, "Histological image classification using biologically interpretable shape-based features," *BMC Med. Imag.*, vol. 13, no. 1, p. 9, 2013.

[11] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 10, pp. 6567–6572, May 2002.

[12] M. E. Ritchie *et al.*, "*limma* powers differential expression analyses for RNA-sequencing and microarray studies," *Nucl. Acids Res.*, vol. 43, no. 7, p. e47, Apr. 2015.

[13] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Appl. Intell.*, vol. 7, no. 1, pp. 39–55, 1997.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[15] M. Slawski, M. Daumer, and A.-L. Boulesteix, "CMA—A comprehensive bioconductor package for supervised classification with high dimensional data," *BMC Bioinformat.*, vol. 9, no. 1, p. 439, 2008.

[16] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced Lectures on Machine Learning*. New York, NY, USA: Springer, 2004, pp. 63–71.

[17] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab—An S4 package for kernel methods in R," *J. Statist. Softw.*, vol. 11, no. 9, pp. 1–20, 2004. [Online]. Available: http://www.jstatsoft.org/v11/i09/

[18] R. Yang, B. J. Daigle, Jr., L. R. Petzold, and F. J. Doyle, III, "Core module biomarker identification with network exploration for breast cancer metastasis," *BMC Bioinform.*, vol. 13, no. 1, p. 12, 2012.

[19] Y. Drier, M. Sheffer, and E. Domany, "Pathway-based personalized analysis of cancer," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 16, pp. 6388–6393, 2013.

[20] N. Seliya, T. M. Khoshgoftaar, and J. Van Hulse, "A study on the relationships of classifier performance metrics," in *Proc. 21st Int. Conf. Tools Artif. Intell. (ICTAI)*, 2009, pp. 59–66.