Proceedings of the Sixteenth International Conference on Pattern Recognition (ICPR-02)
Los Alamitos, CA: IEEE Press

On Machine Learning, ROC Analysis, and Statistical Tests of Significance

Marcus A. Maloof
Department of Computer Science
Georgetown University
Washington, DC 20057, USA
maloof@cs.georgetown.edu

Abstract

ROC analysis is being used with greater frequency as an evaluation methodology in machine learning and pattern recognition. Researchers have used ANOVA *to determine if the results from such analysis are statistically significant. Yet, in the medical decision making community, the prevailing method is* LABMRMC*. Although this latter method uses* ANOVA*, before doing so, it applies the Jackknife method to account for case-sample variance. To determine whether these two tests make the same decisions regarding statistical significance, we conducted a Monte Carlo simulation using several problems derived from Gaussian distributions, three machine-learning algorithms,* ROC *analysis,* ANOVA*, and* LABMRMC*. Results suggest that the decisions these tests make are not the same, even for simple problems. Furthermore, the larger issue is that since* ANOVA *does not account for case-sample variance, one cannot generalize experimental results to the population from which the data were drawn.*

1. Introduction

Receiver Operating Characteristic (ROC) analysis [16] has proven invaluable for empirical studies of machine-learning algorithms (e.g., [14, 4, 18, 9]). Researchers have typically used analysis of variance, or ANOVA [15], to determine whether results from ROC analysis are statistically significant [4, 8]. Yet, in the medical decision making community, which has a long tradition of conducting research on ROC analysis, the prevailing method in the context of multiple readers and multiple cases (MRMC) is LABMRMC [5]. Although LABMRMC conducts an analysis of variance, before doing so, it uses the Jackknife method [7] to account for the case-sample (i.e., *test-sample*) variance. Researchers use LABMRMC to evaluate human performance on detection tasks, so it is not clear whether the additional statistical machinery present in LABMRMC is necessary for relatively simplistic machine decision makers.

To investigate this issue, we conducted a Monte Carlo experiment using several problems derived from Gaussian distributions, three machine-learning algorithms, ROC analysis, ANOVA, and LABMRMC. Results suggest that these tests do not make the same decisions regarding statistical significance, an outcome that has important ramifications for researchers designing and conducting experiments with learning algorithms.

This paper makes two contributions. First, we describe an experimental design and analysis using LABMRMC that takes into account more sources of variance and may provide greater statistical power than popular designs. Second, we present empirical results suggesting that ANOVA and LABMRMC make different decisions about statistical significance.

In the next section, we detail the differences between ANOVA and LABMRMC, and lay the foundation for our empirical study, which we describe in Section 3. After presenting our experimental results, we discuss implications and then conclude with caveats and directions for future work.

2. Background

Recently, researchers have begun using ROC analysis with greater frequency to evaluate learning algorithms, which entails measuring performance at several different decision thresholds and plotting the true-positive and false-positive rates. One way to obtain a single measure of performance, A_z, is to approximate the area under the curve formed by these points using the trapezoid rule. This measure is useful for comparing the performance of classifiers when one ROC curve dominates another. Others have proposed analyses for situations in which curves intersect [11] and in which only a portion of the curve is of interest [17]. Since we evaluate classifiers at different decision thresholds, ROC analysis is a method of evaluating performance that is unconfounded by unequal but unknown costs of misclassification error, by skew in the data set, and by differences in inductive bias amongst the learning methods.

Researchers often use ten-fold cross validation to estimate performance of algorithms on a task. However, this design introduces a source of correlation since one uses examples for training in one trial and for testing in another. Thus, we are suggesting that one should partition the data into, say, eleven sets, train using each of the ten partitions, and test all classifiers using the eleventh. As we have argued elsewhere [1], this is isomorphic to the multiple-reader, multiple-case design in medical imaging.

Once conducting an experiment in this manner and producing a set of areas for two or more learning methods, we might use a test, like Mixed, Two-way ANOVA [15], to determine whether results are statistically significant. In this case, the linear model for the performance metric A_z is

$$A_z^{ijn} = \mu_i + t_j + (at)_{ij} + z_{ijn},$$

where μ_i is the overall average effect of the ith algorithm, t_j is the effect of the jth training set, $(at)_{ij}$ is an interaction term for the ith algorithm and the jth training set, and z_{ijn} is the random effect due to experimental error of the nth repeated trial. Here, we will assume that $n = 1$, so $z_{ijn} = 0$. In this model, the algorithms are fixed effects and the training sets are random effects, but there is no term to account for the variance due to the cases in the test set. Consequently, statistically significant results generalize only to the population of training sets, not to new examples drawn from the population. Note that generalizing to new test sets is not the same as generalizing to new examples.

To generalize to new examples, the analysis must account for the *case-sample variance*, a factor present in Swets and Pickett's model [16]. One possible way to estimate this variance is to split the data set into subsamples, conduct experiments using each subsample, much like one does traditionally for machine-learning experiments, and use Mixed, Three-way ANOVA [15]. Problems include the ad hoc nature by which one splits the data, which leads to unreliable estimates [7], the inability to obtain a maximum-likelihood estimate because of a small sample [5], and as discussed previously, correlation from using an example for training and for testing, which violates the assumption of ANOVA that random effects are uncorrelated [15]. Ideally, we want to draw new random samples from the population for the test sets, but we cannot because of, say, the possibility and expense of collecting such data.

The most general mathematical model for this experimental design is [3]:

$$\begin{aligned} A_z^{ijkn} &= \mu_i + t_j + c_k + (at)_{ij} + (ac)_{ik} \\ &\quad + (tc)_{jk} + (atc)_{ijk} + z_{ijkn}, \end{aligned}$$

where μ_i is the overall average effect of the ith algorithm, t_j is the effect of the jth training set, c_k is the effect of the kth test case (if it were available), $(at)_{ij}$ is an interaction term for the ith algorithm and the jth training set, $(ac)_{ik}$ is an interaction term for the ith algorithm and the kth test case, $(tc)_{jk}$ is an interaction term for the jth training set and the kth test case, $(atc)_{ijk}$ is an interaction term for the ith algorithm, the jth training set, and the kth test case, and z_{ijkn} is the random effect due to experimental error of the nth random trial. Since we are assuming $n = 1$, $z_{ijkn} = 0$.

LABMRMC [5] uses the Jackknife method [7] on case ratings to estimate pseudo-values for areas under a set of ROC curves, thereby accounting for case-sample variance; it then applies ANOVA to estimate significance. Roe and Metz [13] validated this method through computer simulation, and showed it to be conservative in its decisions for small samples. Beiden, Wagner, and Campbell [3] assume this same linear model, but conduct a family of bootstrap experiments to estimate nonparametrically the model's components of variance.

Of concern is whether ANOVA and LABMRMC make the same decisions about statistical significance of experiments with machine-learning algorithms. To address this issue, in the next section, we describe Monte Carlo experiments designed to compare these two statistical tests.

3. Description of Experiments

To investigate whether ANOVA and LABMRMC make the same decisions regarding statistical significance, introduced in the previous section, we conducted a Monte Carlo experiment using problems derived from two nine-dimensional normal distributions and using three learning algorithms: naive Bayes, nearest neighbor, and k-nearest neighbor (k-NN), for $k = 9$. With these, we selected four sample sizes ($n = 100$, 200, 300, 400) and varied the difficulty of the detection task, as measured by d'.

Discriminability, or d', which is equivalent to the Mahalanobis distance, is a measure of the separation between two Gaussian distributions:

$$(d')^2 = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1).$$

In our experiments, with $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\mu}_1 = \mathbf{1}$, we set $\sigma_{ij} = 0$, for $i \neq j$, and set σ_{ii} such that $d' = 1.0, 1.33,$ and 1.66.

For each d' and n, we generated ten training sets by randomly drawing n examples from $N(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0^2)$ for the negative class and n examples from $N(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$ for the positive class. Similarly, we generated a single test set by drawing randomly $\frac{1}{4}n$ samples from each distribution.

For each of the ten training sets, we applied implementations of naive Bayes, nearest neighbor, and k-NN, for $k = 9$, and used the resulting classifiers to predict the class of the cases in the test set. In the first experimental condition, we evaluated the decisions of the algorithms at different thresholds, thereby producing a set of true-positive and false-positive rates. Using the trapezoid rule, we computed the

Table 1. The number of times ANOVA (A) and LABMRMC (L) determined that results were statistically significant at $p < 0.05$.

	$d' = 1.0$		$d' = 1.33$		$d' = 1.66$	
n	A	L	A	L	A	L
100	897	686	965	754	981	780
200	980	954	997	956	998	950
300	995	951	1000	958	999	944
400	999	935	999	935	999	941

Table 2. The number of times ANOVA and LABMRMC agreed (=) or disagreed ($\neq$) on whether results were statistically significant at $p < 0.05$.

	$d' = 1.0$		$d' = 1.33$		$d' = 1.66$	
n	$=$	$\neq$	$=$	$\neq$	$=$	$\neq$
100	617	383	724	275	765	235
200	934	66	953	47	948	52
300	947	52	958	42	944	55
400	935	64	935	64	941	58

area under the ROC curve implied by each set of points, a process resulting in ten areas. For the three algorithms and their ten performance measures, we used ANOVA to determine if the differences among the means were statistically significant at $p < 0.05$.

In the second experimental condition, we applied the algorithms to each of the training sets and used the resulting classifiers to *rate* each case in the test set. LABMRMC requires such ratings, so instead of producing a 0-1 decision for cases, we modified each algorithm to produce a numeric rating. For naive Bayes, we used the posterior probability of the negative class given the instance. For nearest neighbor, we used the difference of the Euclidean distances from the query to the nearest neighbors of the negative and positive classes. For k-NN, we used the number of votes for the negative class. For the three algorithms and their ten sets of ratings, we used LABMRMC to determine if the differences among the means were statistically significant, also at $p < 0.05$.

We repeated this process 1000 times, tabulating the number of times ANOVA and LABMRMC rejected the null hypothesis, and whether the two tests agreed or disagreed about the statistical significance of the outcome, as we will discuss in the next section.

4. Experimental Results and Analysis

Table 1 shows the number of times ANOVA and LABMRMC determined that results were statistically significant at $p < 0.05$. If comparing two identical classifiers and rejecting the null hypothesis at a level of $p < 0.05$, then we would expect a test in the limit to reject the null hypothesis five percent of the time. When comparing different algorithms, especially for small samples, we expect a test to reject the null hypothesis more frequently.

Yet, as we can see, ANOVA appears to be overly optimistic, especially for the higher values of d' and for samples sizes greater than 100. LABMRMC was certainly more conservative than ANOVA, but it was probably not pessimistic, given that LABMRMC has been validated [13].

What we cannot infer from Table 1 is the number of times ANOVA and LABMRMC agreed or disagreed when rejecting the null hypothesis, and Table 2 reports this information. Ideally, these two tests would agree and disagree perfectly, but since ANOVA appears to be overly optimistic, this can not be the case. At the very least, we would hope that whenever LABMRMC fails to reject the null hypothesis that ANOVA also fails to reject it. This also is not the case.

By comparing the results in Table 1 and Table 2, we can see that for $n = 100$, there will be cases for which LABMRMC rejects the null hypothesis and ANOVA fails to reject it, and vice versa. It is not until ANOVA fails to reject the null hypothesis for almost all Monte Carlo trials that the disagreements between the tests consist of those cases for which LABMRMC rejects the null hypothesis and ANOVA fails to reject, an outcome that has several implications, which we discuss further in the next section.

5. Discussion of Results

One immediate implication of this study is that the use of ANOVA for the experimental design we considered may lead to overly optimistic conclusions when determining the significance of means of areas under ROC curves. Researchers in the medical decision making community have investigated this issue extensively, but this work appears to be largely unknown in the communities that evaluate machine-learning algorithms.

Variance contributes to error. If ANOVA does not take into account a strong source of variance, then it follows that it will be overly optimistic when rejecting the null hypothesis. Our study suggests that case-sample variance is such a source in machine-learning experiments. LABMRMC accounts for this variance using the Jackknife method.

There may be other experimental designs and other analyses of variance that properly estimate or account for case-sample variance. Indeed, machine-learning researchers are often in the comfortable position of having too many

cases—a luxury our colleagues in the medical community rarely enjoy—and for these occasions, traditional experimental designs and analyses may adequately take into account case-sample variance.

However, without properly accounting for the case-sample variance, thereby taking into account only the variance due to the random subsampling of the data set, then we can generalize results, if significant, only to the population of random subsamples of the data set. We can make no inferences about results we might obtain if we had new examples drawn from the population [10], which is precisely the inference we want to make.

6. Conclusion

We have reported the outcome of a Monte Carlo simulation designed to compare ANOVA and LABMRMC for determining statistical significance of results from machine-learning experiments. Since ANOVA does not estimate case-sample variance, it tended to be overly optimistic, especially for large samples. LABMRMC, on the other hand, accounts for case-sample variance using Jackknife, requires a simple experimental design, and lets us generalize to the population from which test cases were drawn.

For the future, we hope to report results for non-normal distributions. We have preliminary results for distributions formed with mixtures of Gaussians that vary in the degree of overlap between the distributions corresponding to the negative and positive classes. For these and similar problems, we also have results for other learning methods, such as k-NN for other values of k, quadratic discriminant, and C4.5 [12].

Related to this study is our work on a general framework for understanding the uncertainty of accuracy measures of competing classifiers [1]. For the MRMC paradigm, we have conducted a Monte Carlo simulation similar to that reported in this paper, but used bootstrap experiments to estimate the components of variance of the general linear model [3]. Results support Fukunaga's theory of the effect of design samples [6] and contradict conventional wisdom about the relative contribution of sources of variance to error. We hope to report the details of this study soon [2].

Acknowledgements. This research was conducted in the Department of Computer Science at Georgetown University. The author thanks Robert Wagner and Sergey Beiden for their helpful comments on earlier drafts of this paper.

References

[1] S. Beiden, M. Maloof, and R. Wagner. Analysis of competing classifiers in terms of components of variance of ROC summary accuracy measures. In *Proceedings of the SPIE International Symposium on Medical Imaging: Image Processing*, volume 4684, 2002.

[2] S. Beiden, M. Maloof, and R. Wagner. Analysis of competing classifiers in terms of components of variance of ROC summary accuracy measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, under review.

[3] S. Beiden, R. Wagner, and G. Campbell. Components-of-variance models and multiple-bootstrap experiments: An alternative method for random-effects ROC analysis. *Academic Radiology*, 7:341–349, 2000.

[4] A. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[5] D. Dorfman, K. Berbaum, and C. Metz. Receiver Operating Characteristic rating analysis: Generalization to the population of readers and patients with the Jackknife method. *Investigative Radiology*, 27:723–731, 1992.

[6] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, Boston, MA, 1990.

[7] D. Hinkley. Jackknife methods. In S. Kotz, N. Johnson, and C. Read, editors, *Encyclopedia of statistical sciences*, volume 4, pages 280–287. John Wiley & Sons, New York, NY, 1983.

[8] M. Maloof, P. Langley, T. Binford, and R. Nevatia. Generalizing over aspect and location for rooftop detection. In *Proceedings of the Fourth IEEE Workshop on Applications of Computer Vision*, pages 194–199, IEEE Press, Los Alamitos, CA, 1998.

[9] M. Maloof, P. Langley, T. Binford, R. Nevatia, and S. Sage. Improved rooftop detection in aerial images with machine learning. *Machine Learning*, to appear.

[10] C. Metz. Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigative Radiology*, 24:234–245, 1989.

[11] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 445–453, Morgan Kaufmann, San Francisco, CA, 1998.

[12] J. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Francisco, CA, 1993.

[13] C. Roe and C. Metz. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: Validation with computer simulation. *Academic Radiology*, 4:298–303, 1997.

[14] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 203–208, IEEE Press, Los Alamitos, CA, 1996.

[15] H. Sahai and M. Ageel. *The analysis of variance: Fixed, random, and mixed models*. Birkhäuser, Boston, MA, 2000.

[16] J. Swets and R. Pickett. *Evaluation of diagnostic systems: Methods from signal detection theory*. Academic Press, New York, NY, 1982.

[17] M. Thompson and W. Zucchini. On the statistical analysis of ROC curves. *Statistics in Medicine*, 18:452–462, 1986.

[18] K. Woods and K. Bowyer. Generating ROC curves for artificial neural networks. *IEEE Transactions on Medical Imaging*, 16(3):329–337, 1997.