Published in final edited form as:

Am Stat. 2014 May ; 68(2): 125–126. doi:10.1080/00031305.2014.882867.

Vardeman, S. B. and Morris, M. D. (2013), “Majority Voting by Independent Classifiers can Increase Error Rates,” *The American Statistician*, 67, 94–96: Comment by Baker, Xu, Hu, and Huang and Reply

Stuart G. Baker,

National Cancer Institute

Jian-Lun Xu,

National Cancer Institute

Ping Hu, and

National Cancer Institute

Peng Huang

Johns Hopkins Medical Institution, In the Public Domain

Vardeman and Morris (VM) found a counterexample to the assertion that a majority voting classifier always performs better than its independent component classifiers. VM's counterexample applies to independent classifiers, but biostatisticians are often more interested in conditionally independent classifiers. In biomedical studies, where class is disease status, classifiers are inherently dependent simply because positivity of any reasonable classifier depends on the presence or absence of disease. Conditional independence of classifiers, given disease status, could arise if the classifiers are detecting different biological phenomenon, such as tissue abnormalities versus protein markers.

To explore how majority voting affects classification performance with conditionally independent classifiers, we investigated many examples (Figure 1). Much as we expected, we found that it generally works quite well. However, we also found that conditional independence is not a sufficient condition to ensure that majority voting always leads to better classification performance than the individual classifiers.

As with VM, we considered two classes and component classifiers with identical classification performances. To measure classification performance we used receiver operating characteristic (ROC) curves. ROC curves play a central role in the evaluation of diagnostic and screening tests (Baker 2003; Pepe 2003). In accordance with a decision theory view of ROC curves (Baker, Van Calster, and Steyerberg 2012), we restricted our investigation to ROC curves that are concave, namely with monotonically decreasing slopes from left to right. For a given cutpoint x of a score, let $\text{fpr}(x)$ and $\text{tpr}(x)$ denote the false positive and true positive rates of the component classifier. The ROC curve for the

component classifier plots $\text{tpr}(x)$ versus $\text{fpr}(x)$. At a given cutpoint, the true positive rate for the majority voting classifier is the probability of three or exactly two true positives among the component classifiers, namely $\text{tpr}M(x) = \text{tpr}(x)^3 + 3 \text{tpr}(x)^2 \{1 - \text{tpr}(x)\}$. Similarly the false positive rate for the majority voting classifier is $\text{fpr}M(x) = \text{fpr}(x)^3 + 3 \text{fpr}(x)^2 \{1 - \text{fpr}(x)\}$. The ROC curve for the majority voting classifier plots $\text{tpr}M(x)$ versus $\text{fpr}M(x)$. We considered the following six cases.

Binormal

In this standard formulation of ROC curves (Pepe 2003), the score that determines classification follows a normal distribution with a different mean for each class. Here we set $\text{fpr}(x) = 1 - \Phi(x; 0, 0.25)$ and $\text{tpr}(x) = 1 - \Phi(x; 0.12, 0.25)$, where $\Phi(x; \mu, \sigma)$ denotes the cumulative normal distribution function corresponding to mean μ , and standard deviation σ .

Mixture: VM example

Based on the online supplement of VM, the score that determines classification follows a mixture of normal distributions and gives rise to $\text{fpr}(x) = 1 - \{(12/33) \Phi(x; 0, 0.25) + (21/33) \Phi(x; 1, 0.25)\}$ and $\text{tpr}(x) = 1 - \{(4/736) \Phi(x; 0, 0.25) + (732/736) \Phi(x; 1, 0.25)\}$.

Square root

The true positive rate is $\text{tpr} = \text{fpr}^{1/2}$.

Constant odds ratio

In the constant odd ratio (OR) formulation $\text{tpr} (1 - \text{fpr}) / \{(1 - \text{tpr}) \text{fpr}\} = \text{OR}$, which gives $\text{tpr} = \text{fpr} \text{OR} / \{1 + \text{fpr} (\text{OR} - 1)\}$. In this formulation an odds ratio of 3, which is large by epidemiological standards, implies poor classification performance (Pepe et al. 2004). Here we set $\text{OR} = 10$.

Three-segment A

The three-segment ROC curve successively connects points $(0,0)$, (a_1, b_1) , (a_2, b_2) , and $(1,1)$ with line segments using the formula $\text{tpr} = (b_1/a_1) \text{fpr}$ if $\text{fpr} \leq a_1$, $\{(b_2 - b_1)/(a_2 - a_1)\} (\text{fpr} - a_1) + b_1$ if $a_1 < \text{fpr} \leq a_2$, and $\{(1 - b_2)/(1 - a_2)\} (\text{fpr} - a_2) + b_2$, if $\text{fpr} > a_2$. Here $(a_1, b_1) = (0.3, 0.5)$ and $(a_2, b_2) = (0.5, 0.7)$.

Three-segment B

For this three-segment ROC curve, $(a_1, b_1) = (0.1, 0.5)$ and $(a_2, b_2) = (0.75, 0.99)$.

Two-segment VM

The two-segment ROC curve successively connects points $(0,0)$, (a, b) , and $(1,1)$ with line segments using the formula $\text{tpr} = (b/a) \text{fpr}$ if $\text{fpr} \leq a$ and $\{(1 - b)/(1 - a)\} (\text{fpr} - a) + b$ if $\text{fpr} > a$. Let f_c denote the binary outcomes of component classifiers $c = 1, 2, 3$ in Table 1 of VM, from which we computed $a = \text{pr}(f_1 = 1 | Y = 0) = \text{pr}(f_2 = 1 | Y = 0) = \text{pr}(f_3 = 1 | Y = 0)$, and $b = \text{pr}(f_1 = 1 | Y = 1) = \text{pr}(f_2 = 1 | Y = 1) = \text{pr}(f_3 = 1 | Y = 1)$. For classifier 1, $\text{pr}(f_1 = 1 | Y = y) =$

$\Sigma_{j=(0,1)} \Sigma_{k=(0,1)} \text{pr}(f_1=1, f_2=j, f_3=k, y) / \Sigma_{i=(0,1)} \Sigma_{j=(0,1)} \Sigma_{k=(0,1)} \text{pr}(f_1=i, f_2=j, f_3=k, y)$ and similarly for the other classifiers. This procedure gave $(a, b) = (0.636, 0.995)$.

Two-segment A

In this two-segment ROC curve, $(a, b) = (0.1, 0.50)$.

Two-segment B

In this two-segment ROC curve, $(a, b) = (0.75, 0.99)$.

In three of these six cases, the ROC curve for the majority voting classifier was superior to the ROC curves of the conditionally independent component classifiers over the entire range of false positive rates. However, to our surprise, in three cases (*Two-Segment B*, *Two-Segment VM*, *Three-Segment B*) the ROC curve for the conditionally independent classifier was slightly greater (by at most 0.02, 0.01, 0.003, respectively) than the ROC curve for the majority voting classifier over a small range of false positives ([0.647, 0.718], [0.585, 0.613], [0.670, 0.688], respectively). To better understand how these results arise, we derived a simple set of sufficient conditions for part of these ranges for the two-segment cases (Appendix); a key quantity is the ratio b/a for $a > 1/2$.

If one ROC curve is higher than another ROC curve over the entire range of false positives, the classifier for the former is clearly superior to that for the latter. How does one compare classifiers when the ROC curves cross? In the three cases discussed, the crossover was so slight as to make little difference in classification performance. VM noted “the importance of precision of language when saying what mathematical results mean in practice.” In this spirit, we mention a precise way to compare classifiers with crossing ROC curves based on an anticipated cost/benefit ratio for a practical application. Typically one class is positive for an event, such as disease, and one class is negative for the event. The optimal point on the ROC curve, which corresponds to the largest net benefit, occurs where the slope of the ROC curve equals $\{(\text{cost of a false positive})/(\text{benefit of a true positive})\} \times \{(\text{probability of no event})/(\text{probability of event})\}$ (Metz 1978; Baker and Kramer 2007). The decision-maker chooses the classifier whose optimal point on the ROC curve (which also determines the optimal cutpoint on the classifier) has the largest net benefit. When considering a range of cost-benefit ratios, analysts can more easily compare classifiers via decision curves (Vickers 2008) or relative utility curves (Baker 2009; Baker, Van Calster, and Steyerberg 2012).

Appendix

For the two-segment ROC curve with point (a, b) , we derive a set of five simple conditions that are sufficient for the component classifier to perform better than the majority voting classifier. We begin with Condition 1, $\text{fpr} < a$, and Condition 2, $\text{fpr} > a$. Conditions 1 and 2 imply $\text{tpr} = (a/b) \text{fpr}$ and $\text{tpr} = (b/a) \text{fpr}$, where tpr is the true positive rate of the component classifier corresponding to false positive rate fpr . Condition 3, $\text{fpr} > (3/2)/(1 + b/a)$, implies the superiority of the component classifier at fpr , namely $\text{tpr} > \text{tpr}_M$. We next need to specify the acceptable values of a and b for these Conditions to hold. Combining Conditions 1 and 3 gives $b > 3/2 - a$; with the additional requirement that $b > a$

for a proper ROC curve, we obtain Condition 4, $b = \text{Max}(a, 3/2-a)$. Combining $b = 1$ with Condition 4 implies Condition 5, $a = 1/2$. Consider *Two-Segment B*. Because $a = 0.75$ and $b = 0.99$, Conditions 4 and 5 hold. Under Condition 2, $\text{fpr}M = 0.75$, which implies $\text{fpr} = 0.674$, thereby satisfying Condition 1. Under Condition 3, $\text{fpr} > 0.647$. Consider *Two-Segment VM*. Because $a = 0.636$ and $b = 0.995$, Conditions 4 and 5 hold. Under Condition 2, $\text{fpr}M = 0.636$, which implies $\text{fpr} = 0.592$, thereby satisfying Condition 1. Under Condition 3, $\text{fpr} > 0.585$.

References

- Baker SG. The Central Role of Receiver Operating Characteristic (ROC) Curves in Evaluating Tests for the Early Detection of Cancer. *Journal of the National Cancer Institute*. 2003; 95:511–515. [PubMed: 12671018]
- Baker SG. Putting Risk Prediction in Perspective: Relative Utility Curves. *Journal of the National Cancer Institute*. 2009; 101:1538–1542. [PubMed: 19843888]
- Baker SG, Kramer BS. Peirce, Youden, and Receiver Operating Characteristic Curves. *The American Statistician*. 2007; 61:343–346.
- Baker SG, Van Calster B, Steyerberg EW. Evaluating a New Marker for Risk Prediction Using the Test Tradeoff: An Update. *International Journal of Biostatistics*. 2012; 8:5.
- Metz CE. Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine*. 1978; VIII:283–298. [PubMed: 112681]
- Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Cambridge, MA: Oxford University Press; 2003.
- Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker. *American Journal of Epidemiology*. 2004; 159:882–890. [PubMed: 15105181]
- Vardeman SB, Morris MD. Majority Voting by Independent Classifiers Can Increase Error Rates. *The American Statistician*. 2013; 67:94–96.
- Vickers AJ. Decision Analysis for the Evaluation of Diagnostic Tests, Prediction Models and Molecular Markers. *The American Statistician*. 2008; 62:314–320. [PubMed: 19132141]

Figure 1.

Comparison of ROC curves for majority voting classifier and conditionally independent component classifiers. The 45-degree line is included for reference.