Analysis of Competing Classifiers using Components of Variance of ROC Accuracy Measures

Marcus A. Maloof[†]
Sergey V. Beiden[‡]
Robert F. Wagner[‡]

[†]Department of Computer Science
Georgetown University
Washington, DC 20057

[‡]Center for Devices & Radiological Health
Food & Drug Administration
Rockville, MD 20857

CS-02-01

January 2002

Abstract

Fukunaga and Hayes' theory of the effect of sample size states that for a single classifier, variance comes predominantly from the finite test sample. In this paper, we present results from an empirical study that support and extend this theory. To investigate, we conducted a Monte Carlo simulation using Bayesian and non-Bayesian classifiers for detection tasks derived from Gaussian distributions. To compare the performance of the methods, we used ROC analysis and a nonparametric technique that conducts bootstrap experiments to estimate components of variance. Results support the assertion that variance comes predominantly from the test sample for the case of a single classifier; however, for the case of two competing classifiers, they suggest that variance comes predominantly from the finite training set, not the test set. Our estimates of the variance scale with respect to sample size in accordance with the theory, but because we examined components of variance, rather than total variance, we provide a much more detailed analysis.

1 Introduction

Fukunaga and Hayes (1989) showed that as the number of test cases, N_c, increases, the variance of the error scales proportionally to $1/N_c$. As the number of training samples, N_t, increases, the variance of the error scales proportionally to $1/N_t^2$, provided that the classifier is Bayesian. Otherwise, it scales proportionally to $1/N_t$. When both N_t and N_c increase, they show that the variance of the error scales proportionally to $1/N_c$, and based on this result, conclude that the variance of a single classifier's error comes mainly from the finite test sample.

In this paper, we present results from a simulation study supporting these scaling laws and the assertion that variance comes predominantly from the test sample. However, we show for the case of two competing classifiers that variance comes predominantly from the training set, not from the test sample, an assertion that held except for problems with small sample sizes and low signal-to-noise ratios.

To investigate, we used ROC analysis to evaluate the performance of Bayesian and non-Bayesian classification methods on problems derived from two nine-dimensional normal distributions. We varied the difficulty of the detection task and sample size, and analyzed results with a nonparametric method (Beiden, Wagner, & Campbell, 2000) that uses bootstrap experiments to estimate the components of variance of a general, linear model of ROC accuracy measures (Dorfman, Berbaum, & Metz, 1992).

In the next section, we describe briefly ROC analysis, and in the section that follows, detail our experimental method. We discuss the nonparametric method for estimating components of variance in Section 4, and our empirical results appear in Section 5. We conclude with a discussion of the outcomes and implications.

2 Preliminaries

When evaluating classifiers using receiver operating characteristic (ROC) analysis (Swets & Pickett, 1982), one measures performance at various decision thresholds, thereby producing a set of true-positive and false-positive rates. The points for a given classifier form an ROC curve, and it is often convenient to characterize this curve using a single measure of performance. One such measure is the area under the curve, and one method of obtaining this metric is to use the trapezoid rule to compute the approximate area, $\hat{A}$. It is appropriate when each curve dominates some other. However, researchers have proposed analyses for when curves cross (Provost & Fawcett, 1997), for when a portion of the curve is of interest (McClish, 1989; Woods, Kegelmeyer, & Bowyer, 1997), and for tasks involving three or more decisions (Mossman, 1999; Hand & Till, 2001).

When evaluating two or more classification algorithms, one experimental design is to use multiple training sets and a single test set. One applies each algorithm to each training set and computes performance metrics—in our case, $\hat{A}$—by using the resulting classifiers for the examples in the test set. For this design, the most general linear model for $\hat{A}$ is

$$\hat{A}_{ijkn} = \mu_i + t_j + c_k + (tc)_{jk} + (at)_{ij} + (ac)_{ik} + (atc)_{ijk} + z_{ijkn}, \tag{1}$$

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, $(tc)_{jk}$ is the interaction term for the jth training set and the kth test case, $(at)_{ij}$ is the interaction term for the ith algorithm and the jth training set, $(ac)_{ik}$ is the interaction term for the ith algorithm and the kth test case, $(atc)_{ijk}$ is the 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 on the nth trial (Dorfman et al., 1992; Sahai & Ageel, 2000). In this model, algorithms are fixed effects, and the training sets and test cases are random effects.

With the exception of μ_i, each term is an independent, zero-mean random variable, which is not necessarily normal. The seven components of variance of these random variables are σ_t^2, σ_c^2, σ_{tc}^2, σ_{at}^2, σ_{ac}^2, σ_{atc}^2, and σ_z^2. We will restrict our discussion to the case of a single experimental trial (i.e., $n = 1$), so σ_{atc}^2 and σ_z^2 will be inseparable. With this understanding, we will proceed by using σ_{atc}^2, leaving six components.

$$\begin{array}{ll}
\text{trial}_0 & \left\{\begin{array}{ll}
\text{train}_0 & \left\{\begin{array}{cc}
\text{algorithm}_0 & \text{algorithm}_1 \\ \hline
\text{rating}_0^+ & \text{rating}_0^+ \\
\vdots & \vdots \\
\text{rating}_{N_c}^+ & \text{rating}_{N_c}^+ \\
\text{rating}_0^- & \text{rating}_0^- \\
\vdots & \vdots \\
\text{rating}_{N_c}^- & \text{rating}_{N_c}^-
\end{array}\right. \\
\vdots & \vdots \\
\text{train}_{10} & \vdots
\end{array}\right. \\
\text{trial}_1 & \vdots \\
\vdots & \vdots \\
\text{trial}_{300} & \vdots
\end{array}$$

Figure 1: A depiction of our experimental design.

3 Experimental Design

To investigate the effects of sample size, we conducted a Monte Carlo simulation using problems derived from two nine-dimensional normal distributions and three learning algorithms: naive Bayes (John & Langley, 1995), quadratic discriminant (Johnson & Wichern, 1982), and k-nearest neighbors, for $k = 9$ and 19 (Duda, Hart, & Stork, 2000). With these, we selected three sample sizes ($N_t = 100$, 200, and 400) and varied the difficulty of the detection task, as measured by d'.

Discriminability, d', which is equivalent to the Mahalanobis distance, is a measure of the separation between two normal 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 $d' = 1.0$ and 1.66.

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

We applied all pairs of algorithms to each of the training sets and used the resulting classifiers to *rate* each case in the test set. Instead of producing a 0-1 decision for cases, we modified the performance element of each algorithm to produce a numeric rating. For naive Bayes, we used the posterior probability of the negative class given the instance. For quadratic discriminant, we used the discriminant function. For k-NN, we used the number of votes for the negative class.

We repeated this process 300 times for each N_t, d', and pair of algorithms. Figure 1 depicts our design, but to summarize, we conducted 300 Monte Carlo trials. For each trial, we generated ten training sets and one test set. For each training set, we applied two learning algorithms, each of which generated ratings for the positive and negative cases in the test set.

For a given pair of algorithms (e.g., naive Bayes and quadratic discriminant) and the results from each of the 300 Monte Carlo trials, we conducted a family of *bootstrap experiments* to estimate the components of variance of the linear model in Equation 1. We then averaged the components of variance over the 300 applications of this procedure, which we describe in the next section. We also computed each component's standard deviation over these 300 trials.

4 Components of Variance of ROC Measures

Beiden et al. (2000) use a series of bootstrap experiments to estimate the components of variance of Equation 1 for a context similar to ours. Originally designed to assess the performance of multiple human "readers" across multiple imaging modalities on cancer screening tasks, there is an isomorphism between this task and those investigated in our community. Taking advantage of this isomorphism, we used this method to assess the performance of pairs of classifiers.

One application of the procedure detailed in the previous section yielded for each N_t and d', ten sets of positive and negative case ratings for each of the two algorithms. Using these, we conducted six bootstrap experiments, each of which estimated an *observed variance* of $\hat{A}$.

As mentioned previously, one method of calculating $\hat{A}$ is the trapezoid rule. However, researchers have shown this measure to be equal to the Mann-Whitney two-sample statistic (DeLong, DeLong, & Clarke-Peterson, 1988), which is convenient for numeric case ratings such as ours. Given m ratings of negative cases, $\mathbf{r}^-$, and n ratings of positive cases, $\mathbf{r}^+$,

$$\hat{A} = \frac{1}{m\,n} \sum_{i=1}^{m} \sum_{j=1}^{n} I(r_i^-, r_j^+),$$

where

$$I(r^-, r^+) = \begin{cases} 1 & \text{if } r^- > r^+; \\ \frac{1}{2} & \text{if } r^- = r^+; \\ 0 & \text{if } r^- < r^+. \end{cases}$$

Given the ten sets of positive and negative case ratings for two algorithms, we conducted one bootstrap experiment by assuming that algorithms (a) and training sets (t) were fixed effects and test cases (c) was a random effect. In the following equations, we use the vertical bar ($|$) to distinguish between fixed and random effects. Letters that precede the vertical bar are fixed, letters that follow are random. Therefore, the observed variance of the performance metric $\hat{A}$ for this bootstrap experiment is the sum of the components of variance that involve the random effect c:

$$\operatorname{var}(\hat{A}_{at|c}) = \sigma_c^2 + \sigma_{tc}^2 + \sigma_{ac}^2 + \sigma_{atc}^2. \tag{2}$$

Estimating $\operatorname{var}(\hat{A}_{at|c})$ entailed selecting an algorithm and a training set. Using these, we drew randomly with replacement from the corresponding set of positive and negative case ratings and computed $\hat{A}$. We repeated this process 15,000 times[1] and calculated the variance of $\hat{A}$ in the usual way. We then estimated the variance of $\hat{A}$ for all other algorithms and training sets in the same manner, and so $\operatorname{var}(\hat{A}_{at|c})$ is simply the average of these bootstrapped variances over all algorithms and training episodes.

The observed variance in which algorithms are fixed effects and training sets and test cases are random effects is

$$\operatorname{var}(\hat{A}_{a|tc}) = \sigma_t^2 + \sigma_c^2 + \sigma_{tc}^2 + \sigma_{at}^2 + \sigma_{ac}^2 + \sigma_{atc}^2. \tag{3}$$

Again, the terms in this equation are simply those components of variance involving the random effects t and c. (Note that when all factors are assumed fixed, $\operatorname{var}(\hat{A}_{atc|}) = \sigma_z^2 = 0$.) As before, we estimated $\operatorname{var}(\hat{A}_{a|tc})$ by iterating over all algorithms and training episodes. However, since t is a random effect in this bootstrap experiment, we drew training episodes randomly with replacement. For an algorithm and a training episode, we again drew randomly with replacement from the case ratings to estimate $\hat{A}$, repeated 15,000 times, and calculated $\operatorname{var}(\hat{A})$. After applying this procedure for all algorithms and training episodes, we averaged these bootstrapped variances to yield $\operatorname{var}(\hat{A}_{a|tc})$.

To compare two classifiers, we computed the observed variance of differences between individual estimates of $\hat{A}$ in a similar manner: by estimating with the same and different algorithms, the same and different

[1] The procedure had converged after about 1000 iterations, but we also estimated confidence intervals for each of the bootstrapped variances, requiring the additional calculations. Presenting these confidence intervals was not necessary for this paper.

training sets, and combinations of these. There were four such observed variances of interest, one being when we compared two different algorithms, a and a', using the same training sets and test cases:

$$\mathrm{var}(\hat{A}_{a|tc} - \hat{A}_{a'|tc}) = 2(\sigma^2_{at} + \sigma^2_{ac} + \sigma^2_{atc}). \tag{4}$$

Similarly, the observed variance for an algorithm and two different training sets, t and t', is

$$\mathrm{var}(\hat{A}_{at|c} - \hat{A}_{at'|c}) = 2(\sigma^2_{tc} + \sigma^2_{atc}). \tag{5}$$

The observed variance for two different algorithms trained using the same data set is

$$\mathrm{var}(\hat{A}_{at|c} - \hat{A}_{a't|c}) = 2(\sigma^2_{ac} + \sigma^2_{atc}). \tag{6}$$

Finally, the observed variance for two different algorithms and two different training sets is

$$\mathrm{var}(\hat{A}_{at|c} - \hat{A}_{a't'|c}) = 2(\sigma^2_{tc} + \sigma^2_{ac} + \sigma^2_{atc}). \tag{7}$$

After computing these six observed variances, we expressed Equations 2–7 as the linear system

$$\mathbf{var} = \mathbf{M}\boldsymbol{\sigma}^2, \tag{8}$$

where **var** is the vector of average variances estimated in the six bootstrap experiments (i.e., the left-hand sides of Equations 2–7), $\boldsymbol{\sigma}^2$ is the vector of the six components of variance, $(\boldsymbol{\sigma}^2)^T = \left[\, \sigma^2_t \; \sigma^2_c \; \sigma^2_{tc} \; \sigma^2_{at} \; \sigma^2_{ac} \; \sigma^2_{atc} \,\right]$, and $\mathbf{M}$ contains the coefficients from Equations 2–7:

$$\mathbf{M} = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 2 & 2 & 2 \\ 0 & 0 & 2 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 2 & 0 & 2 & 2 \end{bmatrix}.$$

Solving this linear system gave estimates of the components of variance for a given Monte Carlo trial. Although there are other bootstrap experiments one could imagine (e.g., see Roe & Metz, 1997), those presented are sufficient for estimating $\boldsymbol{\sigma}^2$.

5 Results and Analysis

In this section, we present results for the comparison of naive Bayes and quadratic discriminant. The variance due to the case sample for a single algorithm is the sum of the variance components involving c: $\mathrm{var}_c(\hat{A}) = \sigma^2_c + \sigma^2_{tc} + \sigma^2_{ac} + \sigma^2_{atc}$. Similarly, the variance due to the training sample is the sum of the components involving t: $\mathrm{var}_t(\hat{A}) = \sigma^2_t + \sigma^2_{tc} + \sigma^2_{at} + \sigma^2_{atc}$. (Naturally, we can disregard terms common to both expressions for the purpose of comparison, since they contribute equally to both variances.)

Figures 2–4 show the components of variance for the comparison between naive Bayes and quadratic discriminant for $d' = 1.66$ and $N_t = 100$, 200, and 400. In these graphs, the variance due to the case sample is greater than that due to the training sample. For instance, referring to Figure 2, when $N_t = 100$, $\mathrm{var}_c(\hat{A}) = 0.002949$ and $\mathrm{var}_t(\hat{A}) = 0.001692$. In the other figures, we can also see that the variance comes predominantly from the finite test sample, as shown by Fukunaga and Hayes (1989).

However, if we look only at the components of variance that include the factor for algorithms, then we see that when comparing two algorithms, it is the variance from training set that dominates, not the variance from the test set. That is, disregarding σ^2_{atc}, which contributes equally to both variances, $\sigma^2_{at} > \sigma^2_{ac}$. This was true for all sample sizes.

Results from our experiments with $d' = 1.0$ were similar, but not as stark, as Figures 5–7 indicate. The results appearing in Figure 5 suggest that the test sample and the training sets contributed roughly the same amount of variance. However, as sample size increased, the variance from the test cases became more dominant, as shown in Figures 6 and 7.

We obtained similar results from our comparisons between naive Bayes and k-NN, for $k = 9$ and 19. See the end of this document for these plots.

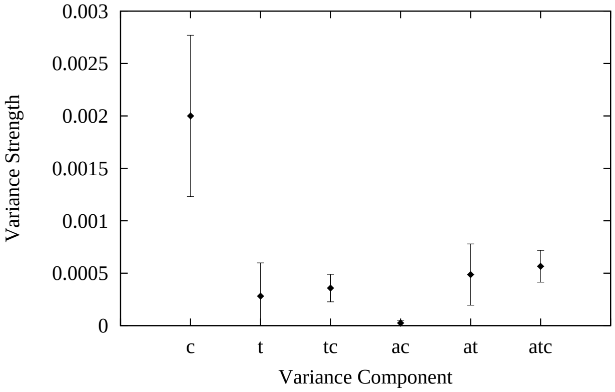

Figure 2: Naive Bayes versus quadratic discriminant, $d' = 1.66$, $N_t = 100$.

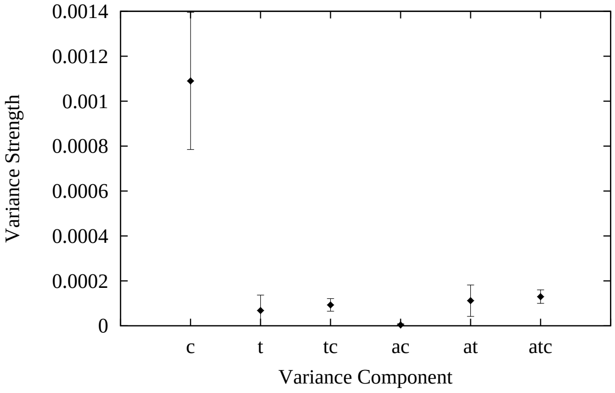

Figure 3: Naive Bayes versus quadratic discriminant, $d' = 1.66$, $N_t = 200$.

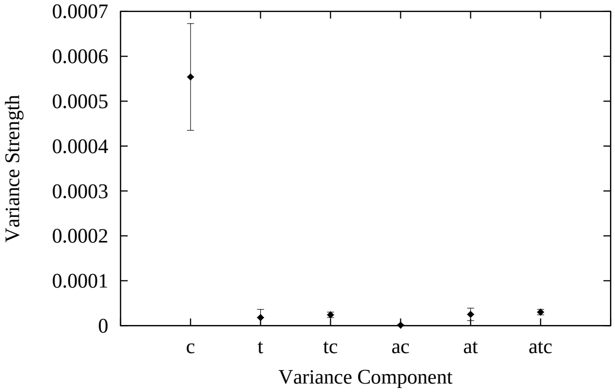

Figure 4: Naive Bayes versus quadratic discriminant, $d' = 1.66$, $N_t = 400$.

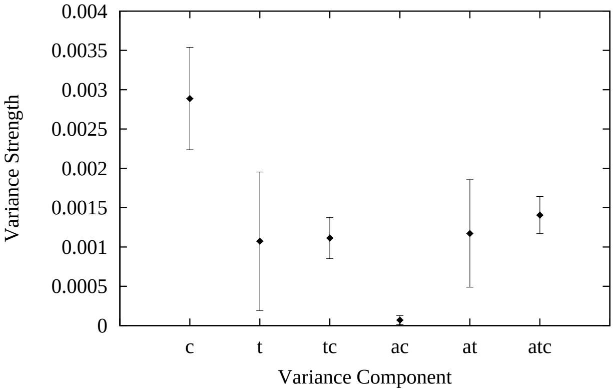

Figure 5: Naive Bayes versus quadratic discriminant, $d' = 1.0$, $N_t = 100$.

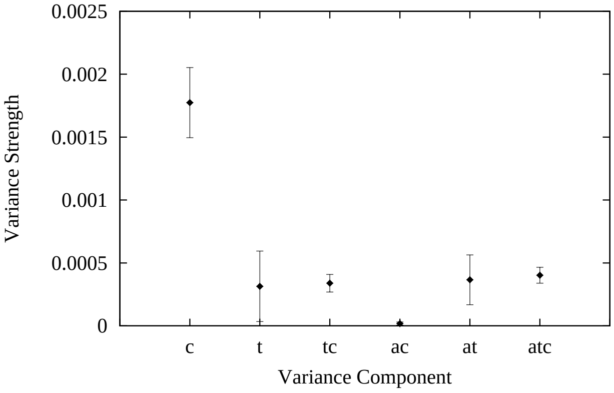

Figure 6: Naive Bayes versus quadratic discriminant, $d' = 1.0$, $N_t = 200$.

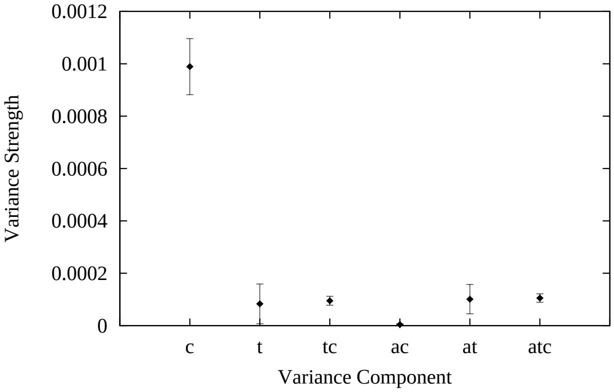

Figure 7: Naive Bayes versus quadratic discriminant, $d' = 1.0$, $N_t = 400$.

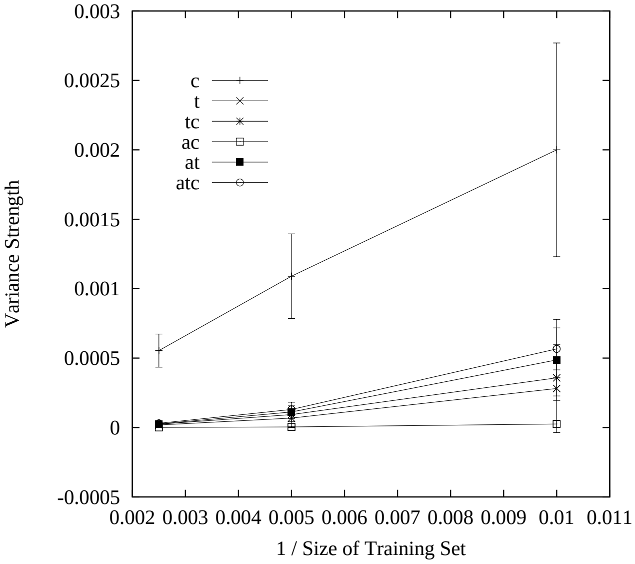

Figure 8: Naive Bayes versus quadratic discriminant, $d' = 1.66$.

We also examined how each component of variance scaled with respect to N_t and N_c. As mentioned previously, Fukunaga and Hayes (1989) showed that as the number of test cases, N_c, increases, the variance of the error scales proportionally to $1/N_c$. As the number of training samples, N_t, increases, the variance of the error scales proportionally to $1/N_t^2$, provided that the classifier is Bayesian. Otherwise, it scales proportionally to $1/N_t$. When both N_t and N_c increase, the variance of the error scales proportionally to $1/N_c$.

Our results generally support these scaling laws, and since we estimated components of variance, rather than total variance, we were able to sketch a much more detailed picture—at least empirically—than has previously existed. Figure 8 shows how the components of variance change with sample size for the comparison of naive Bayes and quadratic discriminant for $d' = 1.66$. We plotted variance strength against $1/N_t$ to show more clearly linear and nonlinear trends. We obtained similar results with other pairs of algorithms, and overall, the components of variance scaled with respect to N_c and N_t as follows:

$\sigma_c^2 \sim 1/N_c$,
$\sigma_t^2 \sim 1/N_t^2$,
$\sigma_{tc}^2 \sim 1/(N_t N_c)$,
$\sigma_{ac}^2 \sim 0$,
$\sigma_{at}^2 \sim 1/N_t^2$ when classifiers were Bayesian,
$\sigma_{at}^2 \sim 1/N_t$ when one classifier was non-Bayesian (i.e., k-NN), and
$\sigma_{atc}^2 \sim 1/(N_t N_c)$.

Relating these results to Fukunaga and Hayes' (1989) theory of the effect of sample size, since $\mathrm{var}_c(\hat{A}) = \sigma_c^2 + \sigma_{tc}^2 + \sigma_{ac}^2 + \sigma_{atc}^2$, the term σ_c^2 will dominate for large values of N_c, so $\mathrm{var}_c(\hat{A}) \sim 1/N_c$. Similarly, $\mathrm{var}_t(\hat{A}) = \sigma_t^2 + \sigma_{tc}^2 + \sigma_{at}^2 + \sigma_{atc}^2$, and if we ignore terms involving $1/(N_t N_c)$, as do Fukunaga and Hayes, then $\mathrm{var}_t(\hat{A}) \sim 1/N_t^2$. However, when the classifier is non-Bayesian, the term σ_{at}^2 goes to $1/N_t$ and will dominate for large values of N_t, so in this situation, $\mathrm{var}_t(\hat{A}) \sim 1/N_t$. Finally, when both N_t and N_c increase, our results suggest that since $\mathrm{var}(\hat{A})$ is the sum of all variance components and the term σ_t^2 will dominate, so $\mathrm{var}(\hat{A}) \sim 1/N_t$. All of this analysis is consistent with Fukunaga and Hayes.

When comparing classifiers, since $\mathrm{var}_{ac}(\hat{A}) = \sigma_{ac}^2 + \sigma_{atc}^2$, $\mathrm{var}_{ac}(\hat{A}) \sim 1/(N_t N_c)$. Since $\mathrm{var}_{at}(\hat{A}) = \sigma_{at}^2 + \sigma_{atc}^2$, $\mathrm{var}_{at}(\hat{A}) \sim 1/N_t^2 + 1/(N_t N_c)$ when classifiers are Bayesian. Otherwise, $\mathrm{var}_{at}(\hat{A}) \sim 1/N_t + 1/(N_t N_c)$. Finally, since $\mathrm{var}_a(\hat{A}) = \sigma_{at}^2 + \sigma_{ac}^2 + \sigma_{atc}^2$, $\mathrm{var}_a(\hat{A}) \sim 1/N_t^2 + 2/(N_t N_c)$ when classifiers are Bayesian. Otherwise, $\mathrm{var}_a(\hat{A}) \sim 1/N_t + 2/(N_t N_c)$.

Although our results support Fukunaga and Hayes' (1989) theory of the effect of sample size, because we estimated components of variance, we sketched a much more detailed picture empirically than has previously existed. And while our results also support Fukunaga and Hayes' assertion that for a single classifier, variance comes predominantly from the finite test sample, our study suggests that for the case of two competing classifiers, variance comes predominantly from the finite training sample, not the finite test sample.

6 Conclusion

Using ROC analysis and bootstrap experiments to estimate the components of variance of a linear model of ROC accuracy measures, we examined the performance of several pairs of classifiers on detection tasks derived from nine-dimensional Gaussian distributions. Results from a Monte Carlo simulation support Fukunaga and Hayes' (1989) theory on the effect of increasing samples, but we extended the theory using a detailed empirical analysis. Our results support the assertion that for a single classifier, variance comes predominantly from the finite test sample. However, when comparing two classifiers, our study suggests that variance comes predominantly from the training set, not the test set. Results also support Fukunaga and Hayes' scaling laws, but we presented laws for components of variance and for variance correlated with algorithms, rather than total variance for a single classifier, which provides a more detailed understanding of the effect of sample size on classifier performance.

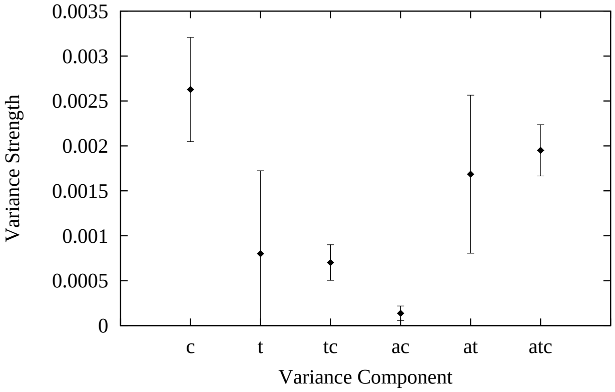

Figure 9: Naive Bayes versus k-NN, for $k = 9$, $d' = 1.0$, $N_t = 100$.

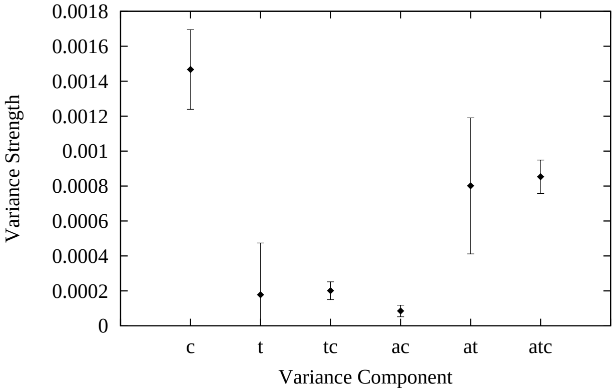

Figure 10: Naive Bayes versus k-NN, for $k = 9$, $d' = 1.0$, $N_t = 200$.

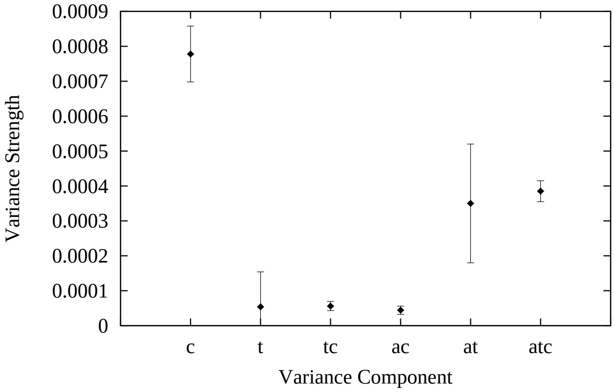

Figure 11: Naive Bayes versus k-NN, for $k = 9$, $d' = 1.0$, $N_t = 400$.

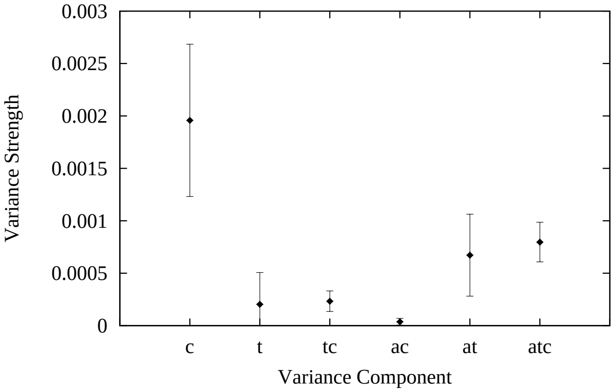

Figure 12: Naive Bayes versus k-NN, for $k = 9$, $d' = 1.66$, $N_t = 100$.

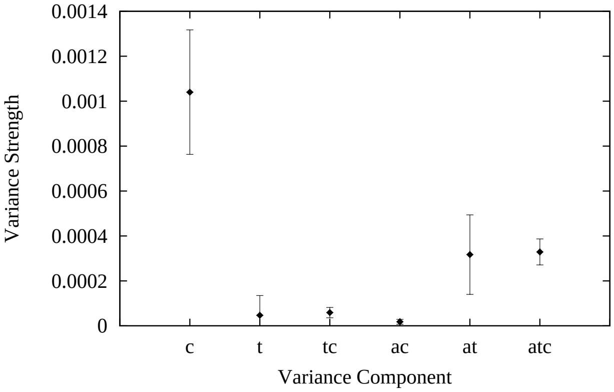

Figure 13: Naive Bayes versus k-NN, for $k = 9$, $d' = 1.66$, $N_t = 200$.

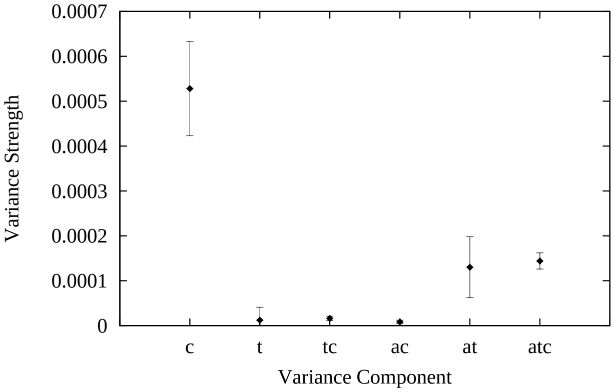

Figure 14: Naive Bayes versus k-NN, for $k = 9$, $d' = 1.66$, $N_t = 400$.

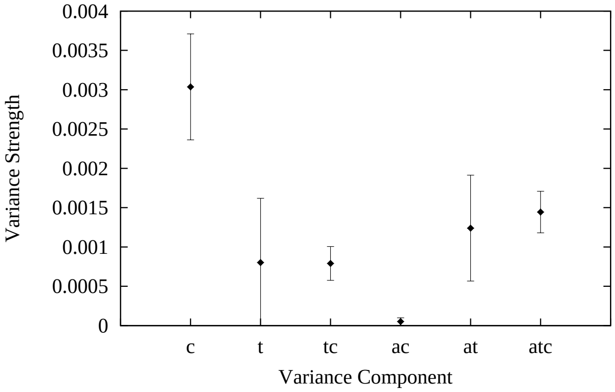

Figure 15: Naive Bayes versus k-NN, for $k = 19$, $d' = 1.0$, $N_t = 100$.

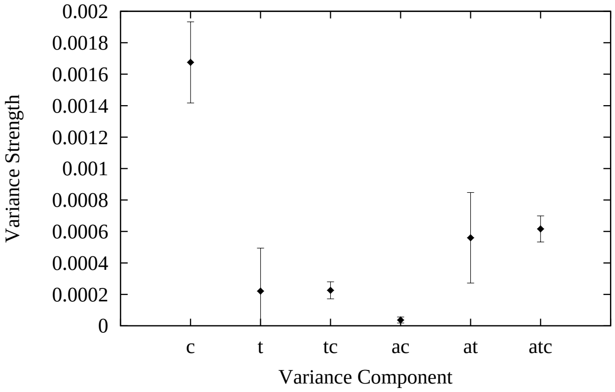

Figure 16: Naive Bayes versus k-NN, for $k = 19$, $d' = 1.0$, $N_t = 200$.

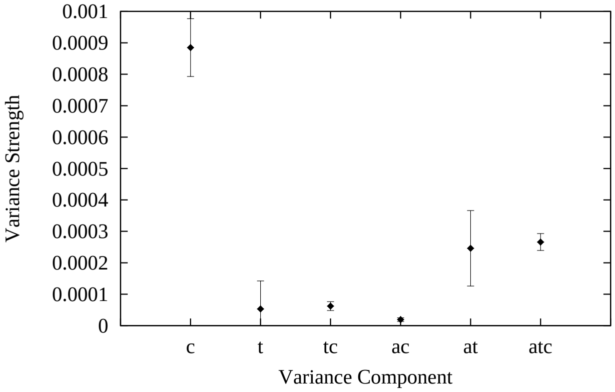

Figure 17: Naive Bayes versus k-NN, for $k = 19$, $d' = 1.0$, $N_t = 400$.

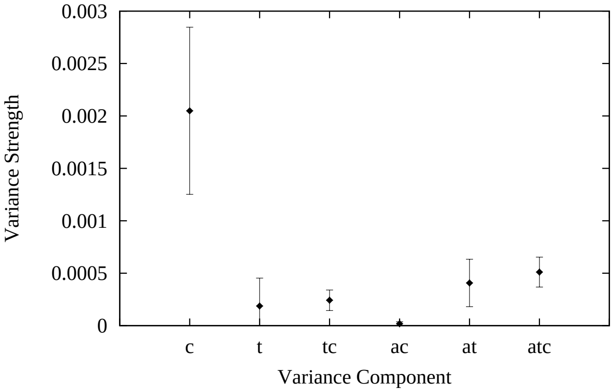

Figure 18: Naive Bayes versus k-NN, for $k = 19$, $d' = 1.66$, $N_t = 100$.

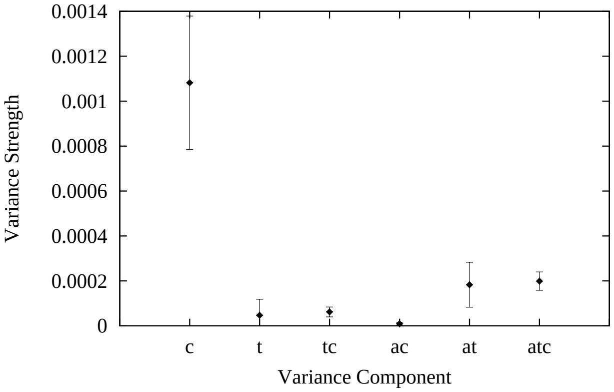

Figure 19: Naive Bayes versus k-NN, for $k = 19$, $d' = 1.66$, $N_t = 200$.

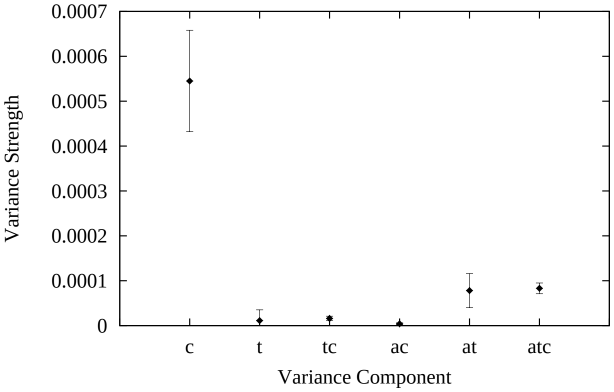

Figure 20: Naive Bayes versus k-NN, for $k = 19$, $d' = 1.66$, $N_t = 400$.

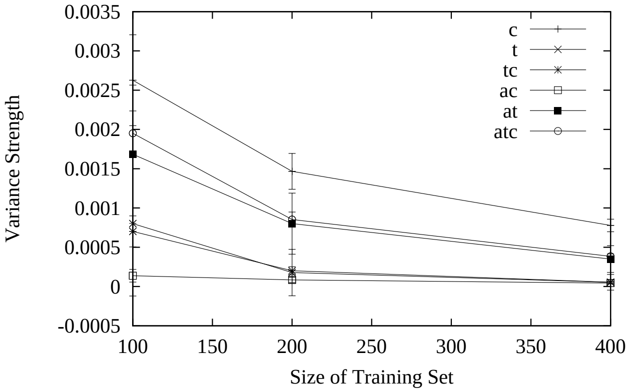

Figure 21: Naive Bayes versus k-NN, for $k = 9$, $d' = 1.0$.

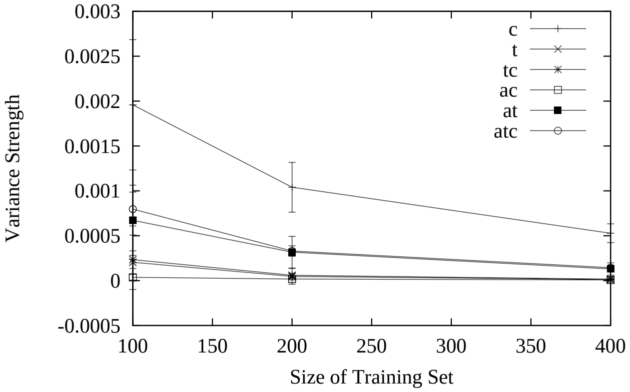

Figure 22: Naive Bayes versus k-NN, for $k = 9$, $d' = 1.66$.

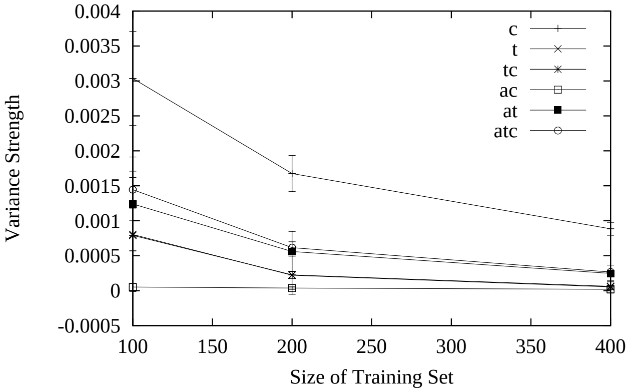

Figure 23: Naive Bayes versus k-NN, for $k = 19$, $d' = 1.0$.

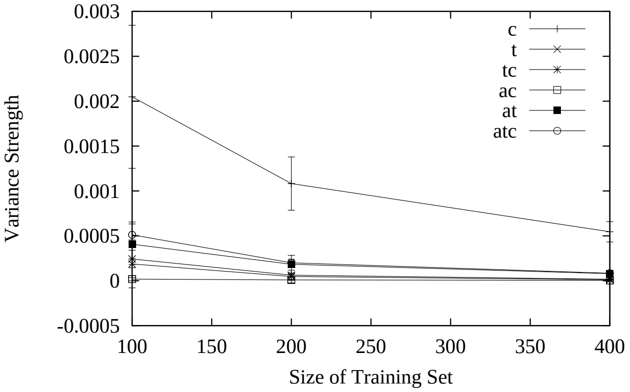

Figure 24: Naive Bayes versus k-NN, for $k = 19$, $d' = 1.66$.

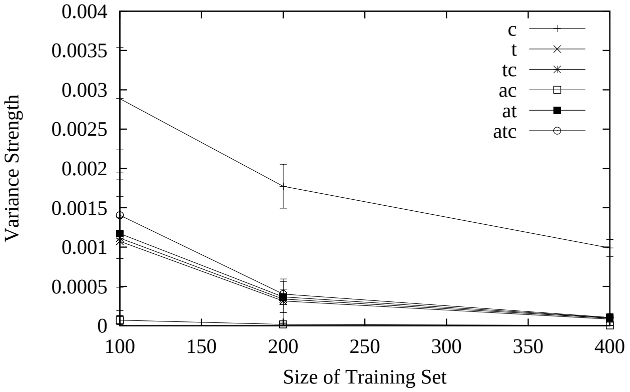

Figure 25: Naive Bayes versus quadratic discriminant, $d' = 1.0$.

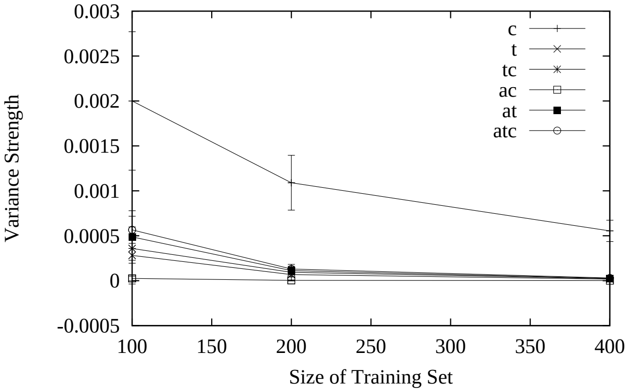

Figure 26: Naive Bayes versus quadratic discriminant, $d' = 1.66$.

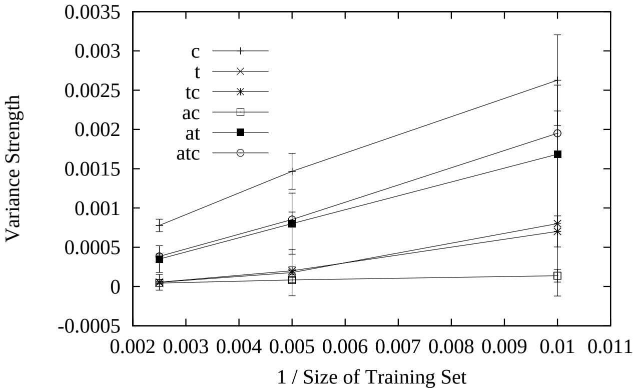

Figure 27: Naive Bayes versus k-NN, for $k = 9$, $d' = 1.0$.

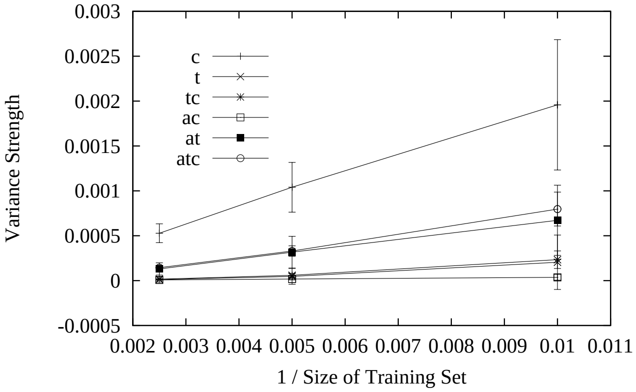

Figure 28: Naive Bayes versus k-NN, for $k = 9$, $d' = 1.66$.

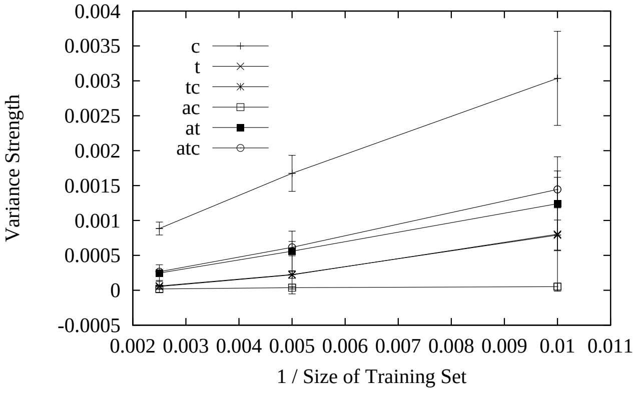

Figure 29: Naive Bayes versus k-NN, for $k = 19$, $d' = 1.0$.

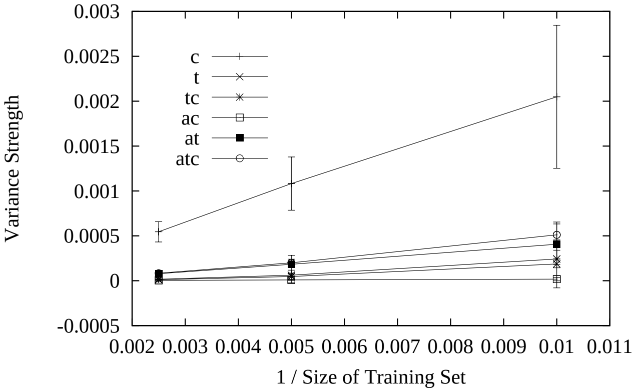

Figure 30: Naive Bayes versus k-NN, for $k = 19$, $d' = 1.66$.

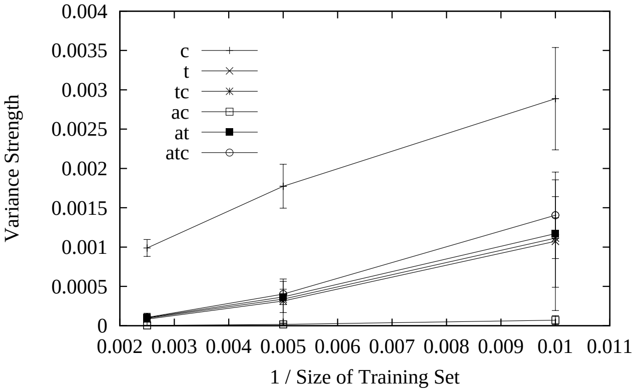

Figure 31: Naive Bayes versus quadratic discriminant, $d' = 1.0$.

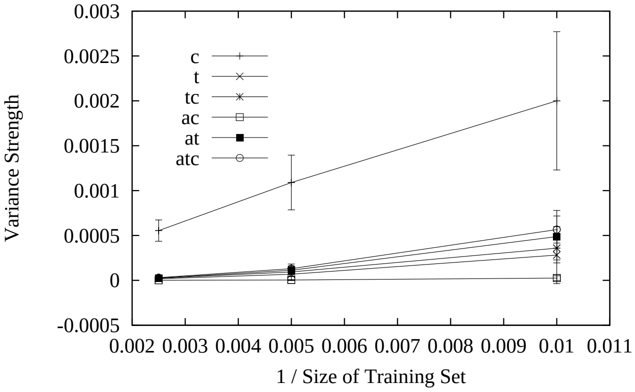

Figure 32: Naive Bayes versus quadratic discriminant, $d' = 1.66$.

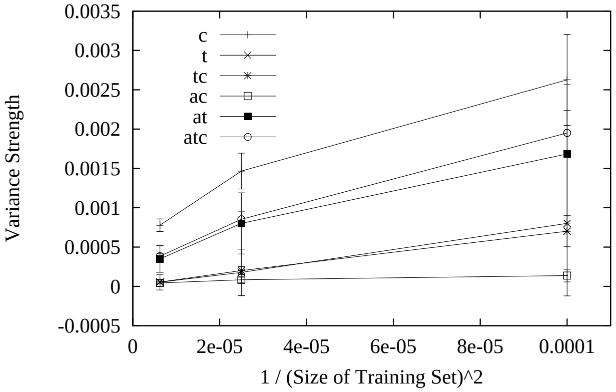

Figure 33: Naive Bayes versus k-NN, for $k = 9$, $d' = 1.0$.

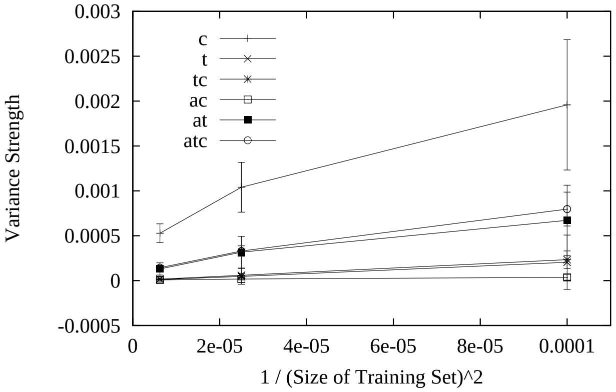

Figure 34: Naive Bayes versus k-NN, for $k = 9$, $d' = 1.66$.

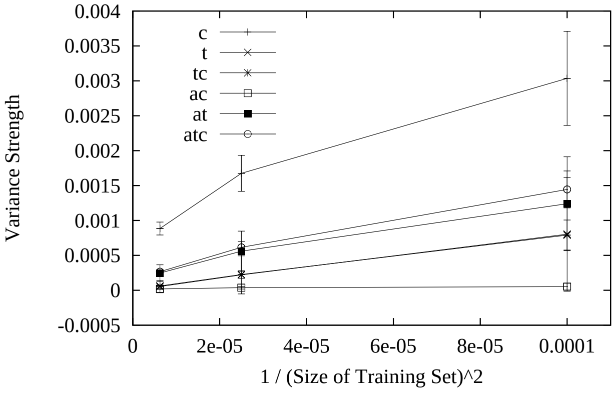

Figure 35: Naive Bayes versus k-NN, for $k = 19$, $d' = 1.0$.

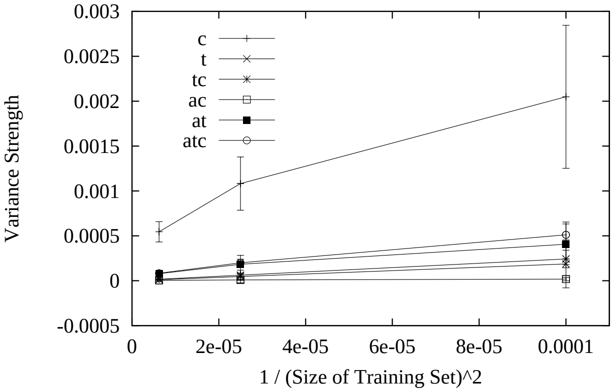

Figure 36: Naive Bayes versus k-NN, for $k = 19$, $d' = 1.66$.

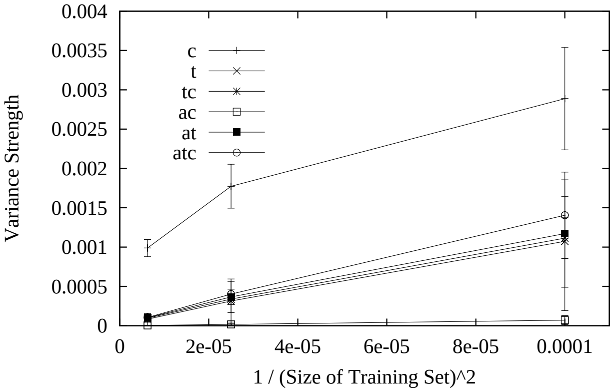

Figure 37: Naive Bayes versus quadratic discriminant, $d' = 1.0$.

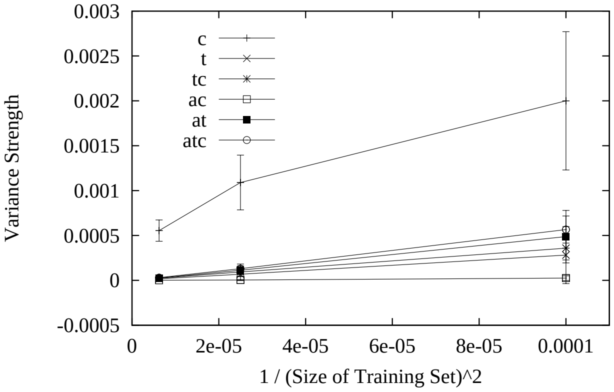

Figure 38: Naive Bayes versus quadratic discriminant, $d' = 1.66$.

Acknowledgements

This research was conducted in the Department of Computer Science at Georgetown University and in the Center for Devices & Radiological Health of the Food and Drug Administration.

References

Beiden, S., Wagner, R., & Campbell, G. (2000). Components-of-variance models and multiple-bootstrap experiments: An alternative method for random-effects Receiver Operating Characteristic analysis. *Academic Radiology*, *7*, 341–349.

DeLong, E., DeLong, D., & Clarke-Peterson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.

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

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification* (2nd ed.). New York, NY: John Wiley & Sons.

Fukunaga, K., & Hayes, R. (1989). Estimation of classifier performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(10), 1087–1101.

Hand, D., & Till, R. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, *45*, 171–186.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 338–345). San Francisco, CA: Morgan Kaufmann.

Johnson, R., & Wichern, D. (1982). *Applied multivariate statistical analysis.* Englewood Cliffs, NJ: Prentice-Hall.

McClish, D. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making*, *9*, 190–195.

Mossman, D. (1999). Three-way ROCs. *Medical Decision Making*, *19*, 78–89.

Provost, F., & Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 43–48). Menlo Park, CA: AAAI Press.

Roe, C., & Metz, C. (1997). Variance-component modeling in the analysis of receiver operating characteristic index estimates. *Academic Radiology*, *4*, 587–600.

Sahai, H., & Ageel, M. (2000). *The analysis of variance: Fixed, random, and mixed models.* Boston, MA: Birkhäuser.

Swets, J., & Pickett, R. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory.* New York, NY: Academic Press.

Woods, K., Kegelmeyer, W., & Bowyer, K. (1997). Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(4), 405–410.