View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Queensland eSpace

Sample Size Estimation using the Receiver Operating Characteristic Curve

Andrew P. Bradley and I.D. Longstaff
Cooperative Research Centre for Sensor Signal and Information Processing (CSSIP)
School of Information Technology and Electrical Engineering
The University of Queensland, St Lucia, QLD 4072, Australia

Abstract

In this paper we describe two related approaches to estimating the sample sizes required to statistically compare the performance of two classifiers: acceptable failure rates (AFR) and the area under the receiver operating characteristic (ROC) curve (AUC). In particular, we consider rare event detection problems, where the prior class probabilities are highly skewed, and measure performance at a specific operating point and for the whole ROC curve. It is shown that the use of AUC as a performance measure is preferable to AFR as it requires a smaller data set to demonstrate superiority of one classifier over another.

1. Introduction

Estimating the required sample size to adequately train and then test pattern recognition systems is of great practical importance. However, this issue is not trivial for a particular problem *a priori* because of the complex inter-relationship of a number of domain specific unknowns. For example, either prior knowledge or some form of exploratory data analysis must be used to estimate the:

- Intrinsic dimensionality of the problem, i.e., the number of features required to classify the data [1];
- Complexity of the problem, i.e., the type of decision boundary (classification scheme or number of free parameters) appropriate for this data [2];
- Expected performance, i.e., the level of performance we can hope to achieve on this data.

More commonly perhaps, the sample size problem is turned on its head, when one is presented with a pre-specified number of examples from which to design and evaluate a pattern recognition system. We are then faced with an optimization problem where we attempt to minimize generalization error, ε, (or some other measure of performance) by varying the number of features and the type, number, or complexity of classifier. This is often a computationally expensive exercise and so it may not be feasible to exhaustively test all of possible models in the solution space [2]. In this case, it is important to use a performance measure that is maximally sensitive to the differing performance of the classifiers under test.

In this paper we investigate the sample size estimation problem from a practical perspective, specifically where the goal is to detect rare events from a large underlying population. In particular, we attempt to estimate the number of samples required to train and test a classifier so that we can demonstrate one of two things:

1. That the generalization performance has not been obtained by a purely random labeling of the examples in the test set (Hypothesis: H_1); and
2. That the generalization performance obtained using one method (say, feature set, classifier, or parameter setting) is superior to another (Hypothesis: H_2).

Both of these require us to formulate a null hypothesis and then to specify a level of significance (confidence limit) with which we can either reject or not reject that hypothesis [3]. We have chosen a statistical significance of less than 0.05, i.e., 1 in 20 or better.

The paper is organized as follows: first, we describe typical underlying statistics for a universal screening application and then utilize an acceptable failure rates (AFR) analysis to estimate the required sample size. Next, we propose the use of the Wilcoxon test to estimate sample size by highlighting its direct link to the area under the receiver operating characteristic (ROC) curve (AUC) [4]. We then discuss the differences between the two methods in terms of which method requires the smaller number of samples to perform a statistically significant comparison.

2. Universal Screening

In this paper we use the term "universal screening" to cover the broad range of medical tests that are applied *universally* to a target demographic in order to detect an underlying disease or abnormality. The goal of universal screening programs, such as those for breast cancer, cervical cancer, and neonatal hearing impairment, is to classify examples (patients) into one of two classes:

1. Negative, that is the test is within normal limits; or

0-7695-2128-2/04 $20.00 (C) 2004 IEEE

2. Positive, that is the test is abnormal and therefore requires further review.

Cleary the vast majority of cases will be normal and so we can expect to observe extremely skewed prior class probabilities when screening the target population, e.g., $\pi_n = 0.95$ and $\pi_p = 0.05$. Note, these priors do not represent the underlying probabilities of the condition being screened for, but reflect the expected referral rates of a typical screening test. That is, positives will consist of cases showing the condition being screened for, a pre-cursor, or an abnormality of unknown significance.

Universal screening tests (either manual or automated) are never 100% accurate. That is, they have a certain sensitivity, *Se* (proportion of positive examples that produce a positive test result: the true positive rate, TPR) and specificity, *Sp* (proportion of negative examples that produce a negative test result: the true negative rate, TNR). Typical values for automated neonatal hearing screening are *Se* = 98% and *Sp* = 90% [5], and *Se* = 51%, *Sp* = 98% for the manual Pap smear screening [6]. Clearly for a specific screening test, there is a trade-off between *Se* and *Sp* and hence the cost-effectiveness of the screening program.

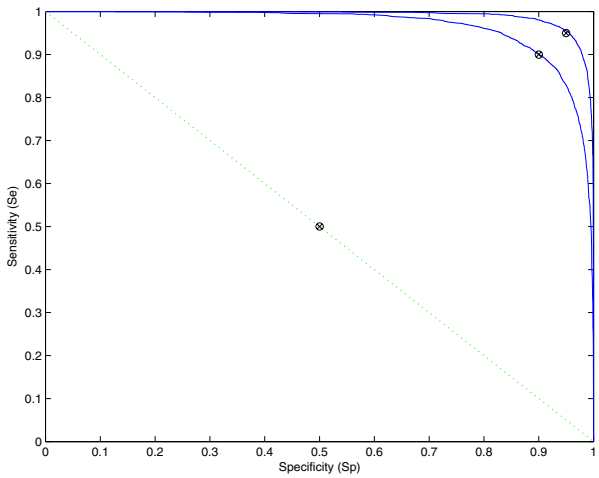

Figure 1: ROC curve and Bayes operating point of three classifiers to be compared.

In this paper, we illustrate the sample size estimation problem using performance criteria calculated from data with a known Normal distribution. That is, for Normal class conditional probabilities we can directly relate the Bayes operating point, with known error probabilities, to a specific ROC curve, and therefore to AUC. The three levels of performance we wish to compare are:

1. A random classifier with *Se* = 50%, *Sp* = 50%, and AUC = 50%;
2. The target classifier with *Se* = 90%, *Sp* = 90%, and AUC = 96%;
3. The improved classifier with *Se* = 95%, *Sp* = 95%, and AUC = 99%;

The ROC curves and specified Bayes operating points for these three cases are shown in Figure 1. Here, the operating points (*Se* = *Sp*) of the target classifiers have been chosen to have a performance suitable for practical application in a universal screening program and AUC is calculated via Monte Carlo simulation.

It should also be noted that, as with any statistical test, there is also an implicit assumption that the samples used to train and test a pattern recognition system are representative of the general population. If this assumption is true, then performance estimated on this data should be indicative of performance in general.

3. Sample Size Estimation Methods

Clearly, in any universal screening application there will be highly skewed prior class probabilities. In such cases it is not efficient to train a pattern recognition system using a random sample of the underlying population [7]. Therefore, it is common practice to enrich the training sample with additional positive cases, so that the priors are approximately equal. In the following we calculate sample sizes based on:

1. The naturally occurring prior probabilities: $\pi_n = 0.95$, $\pi_p = 0.05$; and
2. Equal priors of $\pi_n = 0.5$, $\pi_p = 0.5$, so that number of negative and positive samples is equal ($C_n = C_p$).

However, it should be noted that if we use ε as our performance measure, we will have to adjust results obtained with equal priors to estimate performance on the underlying population [7]. Alternatively, a performance measure such as AUC, which is independent of class priors, can be used [4].

In the following, we have not specifically taken advantage of the fact that when comparing classifiers, performance is nearly always estimated on the same subset of data. That is, the experimental design is often paired, or blocked, and is therefore more sensitive [3]. However, a paired comparison may not always be possible for universal screening, especially if it involves performing two separate tests on the patient. Therefore, the sample sizes estimated below, which do not assume paired data, will be conservative relative to a paired comparison. In addition, it is assumed in the following that some form of cross-validation (say, 10-fold cross-validation or leave-one-out) will be used for training and testing the system [7]. In this way, each data point is used as a test point for the classifier (trained on a sub-set of the remaining data) only once. Again, the sample sizes

estimated may be conservative if some form of stratified cross-validation or boot-strapping is used.

3.1. Acceptable Failure Rates (AFR)

In this section we will determine the sample size necessary to show that a proposed classifier has better performance than another, baseline classifier, in terms of the false negative rate (FNR). Here, FNR has been chosen because positive samples are naturally the most infrequent events and so determine the lower bound on the sample size. In addition, false negatives often have the greatest misclassification cost in a universal screening program and so they are of critical importance.

To test the first hypothesis, H_1, we need to compare the random classifier (FNR = 50%) to the target classifier (FNR = 10%). The (rare) events, which have a probability of 0.5 and 0.05 respectively, will generate a Binomial expectation with a mean, $\mu = pC_p$ and a variance, $\sigma^2 = (1\text{-}p)pC_p$. For our purposes, there is a reasonably large number of samples (say, > 10), and so we can assume the distribution to be Normal. Therefore, we use the large sample z test,

$$z = \frac{x_p - \mu}{\sigma}.$$

For H_1 we wish to test the (null) hypothesis that a FNR of 10% was generated by a purely random labeling of the positive samples. At a statistical significance of less than 0.05, this relates to $z > 1.645$. Therefore, $C_p = 5$ giving a sample size $N = 10$ assuming equal class priors, or $N = 100$ ($C_p = 5$, $C_n = 95$) assuming a prior of $\pi_p = 0.05$.

For H_2 we wish to test the hypothesis that an improvement in performance of reducing the FNR from 10% to 5% was obtained by-chance. In this case, $C_p = 98$ giving a sample size $N = 196$ assuming equal class priors, or $N = 1960$ ($C_p = 98$, $C_n = 1862$) assuming natural priors. Therefore, a sample size greater than or equal to this will allow us to determine that a classifier with a FNR of 10% will deliver an estimate of less than 5% with a by-chance probability of 0.05.

3.2. The Area under the ROC Curve (AUC)

A classifier effectively performs a (perhaps non-linear) projection from the multi-dimensional feature space into a one-dimensional classification score. The classifier is trained to associate low values of the classification score (normally 0 or –1) with negative samples and high values of the classification score (normally +1) with positive samples. When a sample of unknown class is presented to the classifier the classification score has to be thresholded to determine the actual class to which the example belongs. By varying the decision threshold (cut point) it is possible to graph the variation in the probability of a false positive against the probability of a true positive. This type of graph is often referred to as a receiver operating characteristic (ROC) curve [8]. The area under the ROC curve (AUC) is a useful overall performance measure of classifier performance, e.g., an AUC of 1 indicates perfect classification, whilst an AUC of 0.5 indicates a classifier that randomly assigns samples to classes. It is known that the AUC represents the probability that a randomly chosen positive example is correctly rated (ranked) with a larger classification score than a randomly selected negative example, i.e., $P(x_p > x_n)$ [4]. Moreover, this probability of correct ranking is the same quantity estimated by the non-parametric Wilcoxon statistic, W (also often referred to as the U, or Mann-Whitney test) [3]. Therefore, the Wilcoxon test can be used to test the (null) hypothesis that two samples come from identical populations *without* having to make any assumptions about the shape of those distributions. In addition, we can also calculate the standard error of W,

$$\sigma_W = \sqrt{\frac{W(1-W) + (C_p - 1)(Q_1 - W^2) + (C_n - 1)(Q_2 - W^2)}{C_p C_n}}.$$

Where, W is the value of the Wilcoxon test and

$$Q_1 = \frac{W}{(2-W)} \qquad Q_2 = \frac{2W^2}{(1+W)}$$

If C_n and C_p are both reasonably large (say, > 8) the *sampling* distribution of W can be approximated closely with a Normal distribution [3]. Therefore, a large sample z test can again be performed. However, this time we have to determine the value of C_n and C_p that reduce σ_W to a level where the difference in W (mean AUC) becomes statistically significant.

In order to perform equivalent tests to those performed in Section 3.1, hypothesis, H_1, must test whether a classifier with an AUC of 0.96 was a by-chance result from a random-choice classifier, with an AUC of 0.5. Therefore, to obtain a probability of less than 0.05 ($z > 1.645$) of accepting the null hypothesis with an AUC = 0.96, we require $C_p = 3$ giving a sample size $N = 6$ assuming equal class priors, or $N = 40$ ($C_p = 2$, $C_n = 38$) assuming a prior of $\pi_p = 0.05$.

For H_2, we must test a hypothesis that is equivalent to the decrease in FNR from 10% to 5% (see Section 3.1). Therefore, H_2 corresponds to a null hypothesis than an AUC = 0.99 was a by-chance result from a classifier with an AUC = 0.96. The alternate hypothesis is that an improved classifier with AUC = 0.99 is significantly

better than the target classifier with an AUC of 0.96. For this difference (of 0.03) to be statistically significant ($z > 1.645$) we require $C_p = 71$ giving a sample size $N = 142$ assuming equal class priors, or $N = 1320$ ($C_p = 66$, $C_n = 1254$) with natural priors. These results are summarized in Table 1.

Table 1: Estimated sample sizes using acceptable failures rates (AFR) and the Wilcoxon test (AUC) for equal and natural class priors.

Method	AFR $\pi_p = 0.5$	AFR $\pi_p = 0.05$	AUC $\pi_p = 0.5$	AUC $\pi_p = 0.05$
H_1 sample size	10	100	6	40
H_2 sample size	196	1960	142	1320

4. Discussion

Clearly, to test hypothesis H_1: that the performance of the classifier is equivalent to a random classifier; we do not a large amount of data using either AFR or AUC. This is not surprising as it would be very unlikely to attain the level of performance at the target operating point purely by-chance. However, in both cases (equal and natural class priors) AUC requires fewer samples to reject this hypothesis.

In order to test hypothesis H_2: that, at the specified operating point, the FNR rate has decreased from 10% to 5% (AUC increased from 0.96 to 0.99); we require significantly more samples of data. In both cases (equal and natural class priors), AUC requires less samples than AFR to reject this hypothesis. This clearly illustrates that AUC is a more sensitive measure of performance than either *Se* or *Sp* and confirms previous experimental results [8]. In addition, it is natural to expect AUC to be more sensitive as it is obtained by effectively averaging a number of operating points (equal to the number of data points) and so will have a lower variance than the performance measured at any one operating point.

Figure 2 shows the z values for a constant 0.05 improvement in AUC and sensitivity for various levels of performance between 0.5 and 0.95. In this case, the sample size is fixed ($N = 200$) and the class priors are equal. It can be seen that as performance increases the 5% improvement in performance becomes significant when AUC = 0.85, but the 5% change in sensitivity does not become significant until $S_e = 0.9$. Therefore, with a sample size of 200 a change in AUC from 0.85 to 0.9 is statistically significantly, while a change in S_e from 0.85 to 0.9 is not. However, in this case there is no guarantee that a sensitivity of 0.85 lies on a ROC curve with an AUC of 0.85 and so these results should be taken as purely illustrative.

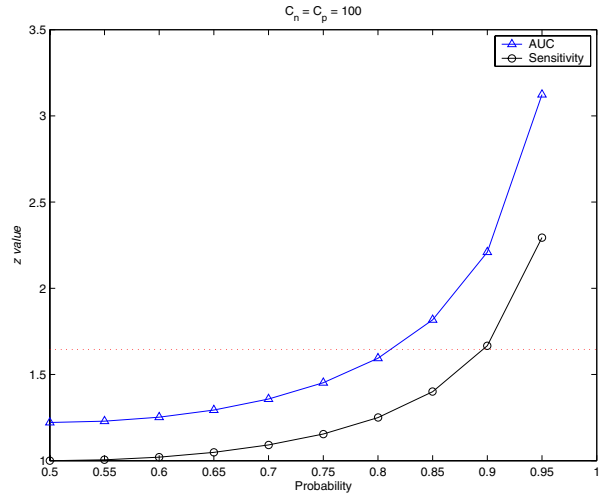

Figure 2: AUC and sensitivity z values, for a 5% improvement in AUC and S_e, $N = 200$ ($C_n = C_p$).

5. Conclusions

Enriching the data by altering the priors is an effective way of reducing the required sample size to train and test a classifier. While moderate amounts of data are required to test if a classifier is superior to a random classifier, significantly more data is required to test if two similar classifiers are significantly different. Use of AUC as a performance measure is preferable to AFR as it is independent of class priors and requires less data to demonstrate superiority of one classifier over another.

6. References

[1] Verveer, P.J. and Duin, R.P.W., "An Evaluation of Intrinsic Dimensionality Estimators," *IEEE Trans Pattern Analysis and Machine Intelligence*, 17 (1), 1995, pp. 81-86.

[2] Jain, A.K., Duin, R.P.W., Mao, J., "Statistical Pattern Recognition: A Review," *IEEE Trans Pattern Analysis and Machine Intelligence*, 22 (1), 2000, pp. 4-37.

[3] Walpole, R.E., *Introduction to Statistics*, 3rd Ed, Macmillan, New York, 1982.

[4] Hanley, J.A. and McNeil, B.J., "The meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, 143, 1982, pp.29-36.

[5] Davis A, *et al*, "A critical review of the role of neonatal hearing screening in the detection of congenital hearing impairment," *Health Technology Assessment*, 1 (10), 1997.

[6] McCrory, D.C., *et al*, "Evaluation of cervical cytology," Evidence report/technical assessment Number 5, AHCPR Pub No. 99-010, Rockville, MD, 1999.

[7] Ripley, B.D., *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996.

[8] Bradley, A.P., "The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms," *Pattern Recognition*, 30 (7), 1997, pp. 1145-1159.