A Logodds Criterion for Selection of Diagnostic Tests

Christopher J. Whitaker[1], Ludmila I. Kuncheva[1], and Peter D. Cockcroft[2]

[1] School of Informatics, University of Wales, Bangor, UK
{l.i.kuncheva,c.j.whitaker}@bangor.ac.uk
[2] Department of Clinical Veterinary Medicine, University of Cambridge, UK
pdc24@hermes.cam.ac.uk

Abstract. We propose a criterion for selection of independent binary diagnostic tests (signs). The criterion maximises the difference between the logodds for having the disease and the logodds for not having the disease. A parallel is drawn between the logodds criterion and the standard minimum error criterion. The error criterion is "progression non-monotone" which means that even for independent binary signs, the best set of two signs might not contain the single best sign. The logodds criterion is progression monotone, therefore the selection procedure consists of simply selecting the individually best features. A data set for scrapie in sheep is used as an illustration.

Keywords: feature selection, combining diagnostic tests, independent binary features, logodds criterion, veterinary medicine, diagnosis of scrapie in sheep.

1 Introduction

Diagnosis can be viewed as an example of a classification problem. The features are the diagnostic tests or the clinical signs. The classes are the possible diagnoses. Here we consider two classes, denoted by D^+ (disease present) and D^- (disease absent) and binary features (a sign can only be present or absent, and a test can only be positive or negative). We assume that the only information available to us is in the form of expert estimates of the probabilities $P(T_i^+|D^+)$ and $P(T_i^+|D^-)$, where T_i^+ stands for "test i is positive" and T_i^- stands for "test i is negative". Without loss of generality we can relabel all the signs and tests so that 'present' sign or 'positive' test indicate more strongly D^+ than disease D^-, i.e., $P(T_i^+|D^+) > P(T_i^+|D^-)$.

The common intuition is that the more *independent* signs/tests we have present/positive, the higher is the probability for D^+.

Selection of the best subset of signs or tests is an importrant topic especially in high dimensional problems. While feature selection is a well developed topic within pattern recognition [1,2], selection of classifiers to form an ensemble (diagnostic test selection) is a less developeded topic in the literature. Classifier selection belongs in the field of multiple classifier systems, usually called

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 574–582, 2004.

"overproduce and select" [4]. In our set-up the two selection tasks are equivalent, therefore feature selection techniques can be applied to selecting classifiers. Without loss of generality in the rest of the paper we will talk about test selection only.

The most common selection criterion is the minimum of the classification error. In this paper we derive a different criterion to assess the usefulness of a diagnostic test. The rationale behind this criterion is to maximize the difference between our belief that the individual has the disease (D^+) if all tests are positive and the belief that the individual does not have the disease (D^-) if all the tests are negative. The presumption is that if mixed results are obtained, further tests, probably more expensive or invasive, will be used to clarify the diagnosis. We illustrate the proposed criterion on a set of expert estimates of probabilities $P(T_i^+|D^+)$ and $P(T_i^+|D^-)$ for scrapie in sheep.

2 Criteria for Selection of Diagnostic Tests

2.1 Classification Error

No diagnostic test is perfect so some individuals with a positive test result will not have the disease while others with a negative test result will have the disease. Conventionally the accuracy of a diagnostic test is measured by two values, the sensitivity and the specificity, defined as follows

$$\text{sensitivity} = \frac{\text{number of individuals with the disease and a positive test}}{\text{number of individuals with the disease}}$$

$$\text{specificity} = \frac{\text{number of individuals without the disease and a negative test}}{\text{number of individuals without the disease}}.$$

More formally the notation often seen in the epidemiology literature is shown in the table below where α and β are probabilities

	T^+	T^-	sum
D^+	$1-\beta$	β	1
D^-	α	$1-\alpha$	1

In this table $1-\beta$ is the sensitivity of the test and $1-\alpha$ is the specificity of the test. In probabilistic terms

$$\text{sensitivity} = P(T^+|D^+) \qquad \text{and} \qquad \text{specificity} = P(T^-|D^-).$$

The error of a test is

$$P(\text{error}) = P(T^- \ \wedge \ D^+) + P(T^+ \ \wedge \ D^-) \tag{1}$$

For this paper we will assume that both types of error are of equal consequence. Also we will assume that we are equally unsure about whether an

individual does or does not have the disease ($P(D^+) = P(D^-) = \frac{1}{2}$). In these circumstances equation (1) leads to

$$\begin{aligned} P(\text{error}) &= P(T^- \ \wedge \ D^+) + P(T^+ \ \wedge \ D^-) \\ &= P(T^-|D^+) \times P(D^+) + P(T^+|D^-) \times P(D^-) \\ &= \frac{1}{2} \times \big(P(T^-|D^+) + P(T^+|D^-)\big) \\ &= 1 - \frac{1}{2} \times ((1 \text{ - } \beta) + (1 \text{ - } \alpha)) \end{aligned}$$

Minimising the error is a sensible criterion. However it is not without its problems. Toussaint (1971) [5] showed that the best test is not necessarily in the best pair of tests. The following hypothetical data shows this. For three tests (labelled A, B and C) we have

test	sensitivity	specificity	P(error)
A	0.90	0.90	0.100
B	0.80	0.95	0.125
C	0.70	0.99	0.155

The best individual test is A, followed by B then C. If we use two tests to make the diagnosis, e.g., A and B, and assume that the tests are independent, then the error can be calculated using the method shown in Table 1.

Table 1. Calculation of the error of the combined test (A and B) for independent A and B

Test results (X)	$P(X\|D^+)$	$P(X\|D^-)$	minimum
$TA^+ \ \wedge \ TB^+$	$0.9 \times 0.8 = 0.72$	$0.1 \times 0.05 = 0.005$	.005
$TA^+ \ \wedge \ TB^-$	$0.9 \times 0.2 = 0.18$	$0.1 \times 0.95 = 0.095$	.095
$TA^- \ \wedge \ TB^+$	$0.1 \times 0.8 = 0.08$	$0.9 \times 0.05 = 0.045$	.045
$TA^- \ \wedge \ TB^-$	$0.1 \times 0.2 = 0.02$	$0.9 \times 0.95 = 0.855$	.020

When calculating the value for P(error) we have to remember that the probabilities need to be weighted by the probabilities of having or not having the disease. As these are both assumed to be $\frac{1}{2}$ here then we can sum the minimum probabilities and multiply by $\frac{1}{2}$ to get .0825. Similarly we find P(error) is .0695 and .05975 for the pairs A & C and B & C respectively. Thus the best pair is B & C, which does not include the best individual test. P(error) using all three tests is .0495. The errors associated for each possible combination of the three tests are shown in Figure 1.

It is findings like these that make feature selection a difficult problem. We shall refer to this phenomenon by calling the error criterion 'progression non-monotone'. We note that the error criterion is monotone with respect to nested sets, i.e. for any independent tests A and B

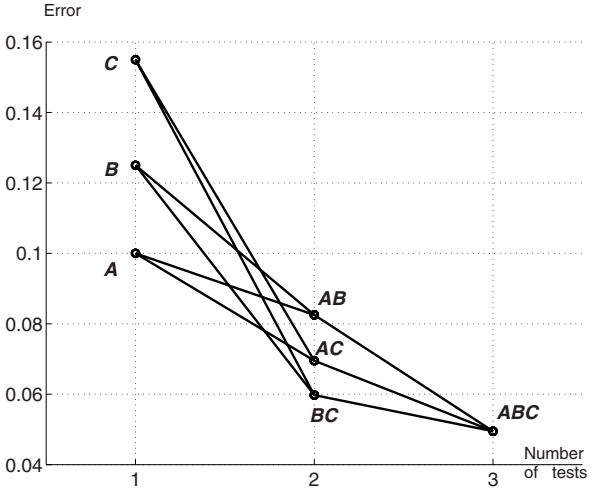

Fig. 1. The error for each combination of tests

$$P(\text{error for A} \wedge \text{B}) \leq P(\text{error for A}) \quad \& \quad P(\text{error for A} \wedge \text{B}) \leq P(\text{error for B})$$

The fact that the error criterion is progression non monotone, even for independent features, means that there is no simple straightforward procedure for selecting the optimal feature subset. Can we find another criterion which is both intuitive for diagnostic purposes, and is progression monotone?

2.2 Logodds Criterion

From a Bayesian viewpoint the probability that an individual has the disease after we have observed a positive test result ($P(D^+|T^+)$) is a measure of our belief that the individual actually has the disease. Probabilities (P) are often more conveniently expressed as odds $\frac{P}{1-P}$, which can then be combined after taking their logarithms. Here the odds of having the disease is

$$\text{ODDS}(D^+ : D^-|T^+) = \frac{P(D^+|T^+)}{1 - P(D^+|T^+)} = \frac{P(D^+|T^+)}{P(D^-|T^+)}$$

Intuitively we would like this value to be high as it means that if we get a positive test result then we increase our belief that the individual has the disease. Similarly we would like our belief in the individual having the disease to be decreased when we get a negative test result. The odds of having the disease given a negative test result is $\text{ODDS}(D^+ : D^-|T^-) = \frac{P(D^+|T^-)}{1-P(D^+|T^-)} = \frac{P(D^+|T^-)}{P(D^-|T^-)}$. We would like this to be as small as possible.

We define the *diagnostic value* of a particular test T to be

$$v(T) = \log(\text{ODDS}(D^+ : D^-|T^+)) - \log(\text{ODDS}(D^+ : D^-|T^-)).$$

This criterion maximises the difference in our belief of being diseased between a positive test result being obtained and a negative test result being obtained. From the definition of Bayesian probability we have

$$P(D^+|T^+) = \frac{P(T^+|D^+) \times P(D^+)}{P(T^+|D^+) \times P(D^+) + P(T^+|D^-) \times P(D^-)} \tag{2}$$

$$= \frac{(1-\beta) \times \delta}{(1-\beta) \times \delta + \alpha \times (1-\delta)} \tag{3}$$

where $\delta = P(D^+)$ and α and β are as defined above.

Similarly

$$P(D^-|T^+) = \frac{\alpha \times (1-\delta)}{\alpha \times (1-\delta) + (1-\beta) \times \delta}$$

This leads to

$$\text{ODDS}(D^+ : D^-|T^+) = \frac{\delta}{1-\delta} \times \frac{1-\beta}{\alpha}$$

Taking the logarithm of this leads to our measure of belief in being diseased when a positive test result is obtained.

$$\log(\text{ODDS}(D^+ : D^-|T^+)) = \log \frac{\delta}{1-\delta} + \log \frac{1-\beta}{\alpha}$$

Similarly our measure of belief in being diseased when a negative test result is obtained is

$$\log(\text{ODDS}(D^+ : D^-|T^-)) = \log \frac{\delta}{1-\delta} + \log \frac{\beta}{1-\alpha}$$

Taking the difference of these two logodds terms leads to our logodds criterion

$$v(T) = \log \frac{1-\beta}{\alpha} - \log \frac{\beta}{1-\alpha} \tag{4}$$

$$= \log \frac{\text{sensitivity}}{1-\text{sensitivity}} + \log \frac{\text{specificity}}{1-\text{specificity}} \tag{5}$$

Notice that δ is not involved in the criterion. This means that the prior probability we have for an individual being diseased plays no part in deciding which test is the best. This has important consequences when we come to look at the situation when we combine the results from more than one test.

Suppose that T is a combined test consisting of $T_1, T_2, \ldots, T_n$. A positive combined test will be equivalent to all individual tests giving positive results, i.e., $T^+ = T_1^+ \cap T_2^+ \cap \ldots \cap T_n^+$. A negative combined test will be equivalent to all individual tests giving positive results, i.e., $T^- = T_1^- \cap T_2^- \cap \ldots \cap T_n^-$. Using the assumption of independence, we can show that

$$v(T) = \sum_{i=1}^{n} v(T_i).$$

for any values of the prior probability δ. Let $\mathcal{S}$ denote the set of all available diagnostic tests. Our proposed criterion is

$$\max_{T \subseteq \mathcal{S}} v(T) = \max_{T \subseteq \mathcal{S}} \sum_{T_i \in T} v(T_i).$$

For the example data shown above the logodds criterion values can be calculated as shown in Table 2. Using this criterion the best test is C, followed by A then B. The real advantage of this criterion can now be seen. After we have got the results of the best test (C here) then our prior probability of the disease has changed. But this plays no part in deciding which test to use in conjunction with it. So the second best test to use in combination with test C is test A. The result of combining all three tests and showing all possible results is given in Figure 2.

Table 2. Calculation of the logodds criterion

test	sensitivity $1-\beta$	specificity $1-\alpha$	$\log \frac{1-\beta}{\alpha}$	$\log \frac{\beta}{1-\alpha}$	logodds criterion
A	0.90	0.90	2.197	-2.197	4.394
B	0.80	0.95	2.773	-1.558	4.331
C	0.70	0.99	4.248	-1.194	5.442

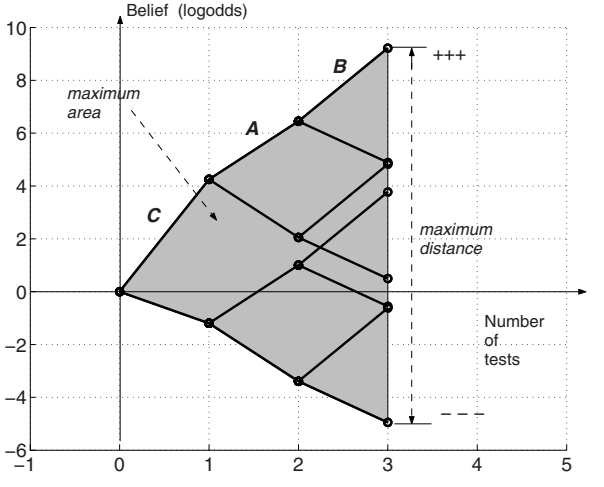

Fig. 2. The logodd criterion for all possible results of 3 tests

Visually the logodds criterion finds the test with the largest difference between the logodds for a positive and a negative test result.

2.3 Differences between the Logodds and the Error Criteria

The following example shows that the logodds criterion is different from the error criterion. Consider again the three tests A, B and C. Let $\alpha_1 = 0.1$ and

$\beta_1 = 0.1$ be the respective errors for test A. Figure 3 depicts the regions for α_2 and β_2 for another test. If the point specified by the pair (α_2, β_2) is inside one of the shaded regions (e.g. Test C), then the second test is worse than A on the error criterion but better than A on the logodds criterion. However if the point is not inside the shaded region (e.g. Test B) then the test is worse than A on both criteria.

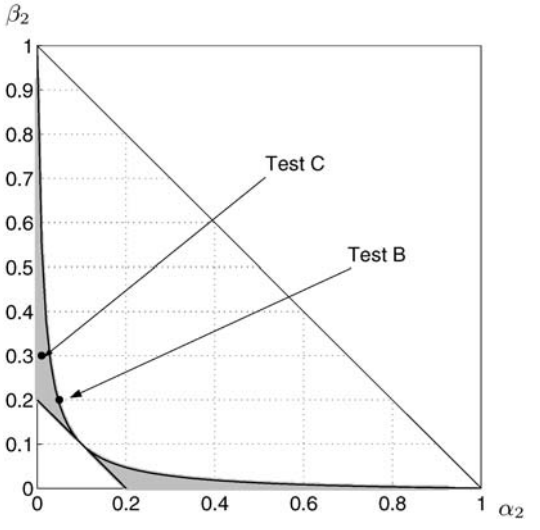

Fig. 3. Regions for α_2 and β_2 where the (individual) error criterion and logodd criterion disagree

An important difference between the two criteria is the computational effort required for estimating the error for a set of tests $T = \{T_1, \ldots, T_n\} \subseteq \mathcal{S}$. For the error criterion, the error is the sum of 2^n terms, one for each possible combination of test results. These are calculated from the expert estimates of the probabilities. Since the error criterion is progression non monotone, special procedures have to be applied to navigate through the subsets T. On the other hand, the logodds criterion is progression monotone. It is a simple sum whereby we can maximise each component so as to get the a total maximum. Thus the set of the individually best n features is the best set of n features according to the logodds criterion.

3 Results for Scrapie Data

Scrapie is a notifiable disease of sheep. We consider 285 clinical signs that may be observed in either scrapie or in 62 differential diagnoses [6]. The available data is in the form of probabilities $P(T_i^+|D^+)$ and $P(T_i^+|D^-)$, where $i = 1, \ldots, 285$, D^+ is scrapie, and D^- is any of the other diagnoses. It is not practical to observe

all 285 clinical signs on sheep suspected for scrapie. Kuncheva et al. (2003) [3] report an analysis which uses sequential feature selection (SFS) to choose 15 signs based on the classification error criterion. Table 3 shows the 15 signs selected by the error criterion through the SFS procedure and the 15 signs selected by the logodds criterion. The numbers in front of the signs show the *individual rank* according to the error criterion. The sign with rank 1, Hyperaesthesia, is the most important discriminatory sign according to the error criterion.

Table 3. The 15 diagnostic signs (tests) selected through the error criterion (via SFS) and the logodds criterion. Given in brackets are $(P(T^+|D^+)/P(T^+|D^-))$. 'R' is the *individual* rank of the sign by the error criterion

R Sign	R Sign
1 Hyperaesthesia (.87/.06)	2 Weight loss (1/.25)
2 Weight loss (1/.25)	1 Hyperaesthesia (.87/.06)
3 Pruritus (.8/.07)	3 Pruritus (.8/.07)
21 *Increased respiratory rate (0/.19)	4 Abnormal behaviour (.7/.06)
4 Abnormal behaviour (.7/.06)	5 Underweight (.9/.25)
5 Underweight (.9/.25)	21 *Increased respiratory rate (0/.19)
9 Tremor (.63/.09)	22 *Sudden death (0/.18)
22 *Sudden death (0/.18)	18 Abortion or weak newborns (.3/.02)
6 Dysmetria (.67.1)	6 Dysmetria (.67.1)
7 Ataxia (.77/.2)	9 Tremor(.6/.09)
8 Grinding teeth (.67/.11)	10 Trembling (.6/.09)
10 Trembling (.6/.09)	8 Grinding teeth (.7/.11)
11 Alopecia (.57/.08)	23 *Tachycardia (0/0.13)
12 Seizures or syncope (.52/.1)	25 *Reluctant to move (0/.13)
13 Rumen hypomotility (.47/.1)	11 Alopecia (.6/.08)

Notes:
1. The signs marked with a * had to be relabeled so as to ensure that $P(T_i^+|D^+) > P(T_i^+|D^-)$.
2. For calculation purposes, all 0's were reassigned to 0.01 meaning "very rare" and all 1's were reassigned to 0.99 meaning "almost sure".

The table shows that the logodds criterion has selected 12 of the 15 features that were chosen by the error criterion. Two of the three different features chosen by the logodds criterion have sensitivities of 0. If the expert estimates are incorrect in the extremes (0 or 1), then the logodds criterion might give undue weight to such features.

4 Conclusions

The difference between the error and logodds criteria can be viewed in the following way. A pharmaceutical company creates a diagnostic test and measures its worth by means of the number of misdiagnoses that ensue. This is a perfectly

sensible for the company. However an individual is only interested in the results of the test on themselves. The best test for the individual is the one that leads to the most information about whether the individual actually has the disease. The logodds criterion attempts to measure this. In this sense the criterion is intuitive for diagnostic purposes.

The logodds criterion is progression monotone. This means that we can determine at the outset the ordering of the worth of the tests using this criterion. The advantage is that as the prior belief plays no part in the criterion then this ordering is the one that applies every time we want to include the results of another test. The computational efficiency savings follow from the progression monotonicity.

References

1. D. W. Aha and R. L. Bankert. A comparative evaluation of sequential feature selection algorithms. In *Proc. 5th International Workshop on AI and Statistics*, pages 1–7, Ft Lauderdale, FL, 1995.
2. A. K. Jain and D. Zongker. Feature selection: evaluation, application and small sample performance. *IEEE Transactions on PAMI*, 19(2):153–158, 1997.
3. L. I. Kuncheva, P. D. Cockcroft, C. J. Whitaker, and Z. S. Hoare. Pre-selection of independent binary features in differential diagnosis. submitted.
4. F. Roli and J. Kittler, editors. *Proc. 2nd International Workshop on Multiple Classifier Systems (MCS 2001), Lecture Notes in Computer Science LNCS* **2096**. Springer-Verlag, Cambridge, UK, 2001.
5. G. T. Toussaint. Note on optimal selection of independent binary-valued features for pattern recognition. *IEEE Transactions on Information Theory*, 17:618, 1971.
6. M. E. White. Diagnosis, information management, teaching and record coding using the consultant database. *Canadian Veterinary Journal*, 29:271–273, 1988.