Novel Criteria for Estimation of Diagnostic Test Performance and its Practical Using

Leonid S. Fainzilberg
International Research and Training Center of Information Technologies and Systems
Prospect Academica Glushkova, 40, Kiev-680, Ukraine, 03680
fainzilberg@voliacable.com

Abstract:

The paper introduces a new approach to estimating a diagnostic test's utility from the point of view of risk reduction. Equations are given for the utility of a test, or separate diagnostic feature, supported by proofs. The greatest lower bound on a test's sensitivity and a test's specificity are obtained. Improvement to the traditional ROC-analysis is shown, based on the formal restriction of a ROC-plot.

1. Introduction

Diagnostic test performance is one of the basic characteristic of modern decision support systems for diagnostics in technical and medical applications. Adopting to statistical approach any diagnostic test traditionally is considered as useful if it is based on features having different conditional distributions for classes, for example different distributions for healthy persons and patients with some disease [1]. As a matter of fact till now such approach is widely used to evaluate tests diagnostic performance based on area below so-called receiver operating characteristic (ROC) plot [2-4].

However from the paper [5] it is known that the fact of distinction of feature's conditional distributions is only necessary but not sufficient condition from the point of view of decrease of the average error probability. The sufficient conditions ensuring of reduction of the average error probability was obtained in the paper [6].

But it is well known that the average probability of the error does not take into account the rates between losses of false positive and false negative mistakes. At the same time for some application including medical diagnostics such loss are not equivalent. The problem of dealing with highly imbalanced classes is an important issue also.

So the problem to build criteria ensuring the performance of a diagnostic test from the point of view of risk reduction is very important for practical using.

The purpose of this paper is to establish a connection between a test utility and an average risk reduction under possible values of sensitivity, specificity, class prevalence and errors loses. This paper quantifies the true positive and true negative rates that must be achieved to produce a reduction in expected cost over classifying everything as a single class. It establishes bounds on the these rates in terms of misclassification costs and priors.

We present novel mathematical conditions render a binary classifier (test) having expected losses less than that of a default classifier that merely uses *a prior* class distributions and then ties this conditions to ROC analysis, showing how to restrict portions of the traditionally ROC curve to regions where the test is useful.

2. Basic Definition

Let's consider a traditionally "binary" discriminating task: distinguishing between two classes, for example, distinguishing between persons in different disease states, nominally, those with "disease" (class V_1) and those "without disease" (class V_2). The results of testing based on some diagnostic test are dichotomized into two decision: $\delta = 1$ (the "positive" decision when person is assign to the class V) and $\delta = 2$ (the "negative" decision when person is assign to the class V_2).

Let we need to evaluate the performance of the novel diagnostic test directed on decision of this task using it's results obtained on persons with known state. Observed data are arrayed in 2×2 matrix including numbers of True Positive (*TP*), True Negative

(*TN*), False Positive (*FP*) and False Negative (*FN*) cases (Table 1.)

Table 1: Results of testing.

Actual class	Predicted Class	
	V_1 ($\delta = 1$)	V_2 ($\delta = 2$)
V_1	*TP*	*FN*
V_2	*FP*	*TN*

From this table a variety of known diagnostic indexes are derived including the True-Positive Fraction (*TPF*) which is commonly referred as the test's sensitivity

$$C_E = \frac{TP}{TP + FN}$$

and the False-Positive Fraction (*FPF*) which is equal to 1 minus the true negative fraction, or 1 minus the so-called test's specificity

$$C_P = \frac{TN}{TN + FP}.$$

There is a natural question: what sensitivity C_E and what specificity C_P is acceptable for practical using of this test ?

To study this problem let's suppose that we assume to use the test for screening patients from representative group in the sense that group members reflected to *a priory* class probabilities $P(V_1)$ and $P(V_2) = 1 - P(V_1)$, where $P(V_1)$ - known prevalence of the disease. In this case the expected risk (the average loss) is defined by the formula

$$R = \sum_{k=1}^{2} \sum_{j=1}^{2} L_{kj} P(V_k, \delta = j), \qquad (1)$$

where $P(V_k, \delta = j)$ denotes the probability of the random event when tested person really belong to group V_k ($k = 1,2$) but test's result is $\delta = j$ ($j = 1,2$) and L_{kj} be corresponding loss.

Definition 1. The diagnostic test will be named as useful if the average risk R accepted on its decisions $\delta = 1,2$ is strongly less then *a priory* risk R_0, based on *a prior* disribution of classes, i.e.

$$R < R_0. \qquad (2)$$

Our goal is to define formal conditions guaranteeing performance of a strict inequality (2).

3. Formal conditions for estimation of diagnostic test utility

We have proved the following theorems being based on theories of statistical decisions.

Theorem 1. Any test is useful in the sense of (2) if and only if

a) $C_E > \theta(1 - C_P)$ *when* $\theta \geq 1$; (3)

b) $C_E > 1 - \theta + \theta(1 - C_P)$ *when* $\theta < 1$; (4)

where θ is defined by the formula

$$\theta = \frac{1 - P(V_1)}{\omega P(V_1)} \qquad (5)$$

in which the dimensionless value

$$\omega = \frac{L_{12} - L_{11}}{L_{21} - L_{22}}, \qquad (6)$$

determing the loss ratio where L_{11}, L_{22} define losses of correct decisions and L_{12}, L_{21} define losses connected to false negative and false positive mistakes. Naturally $L_{12} > L_{11}$, $L_{21} > L_{22}$ and so $\omega > 0$. It is supposed also that $P(V_1) \neq 0$.

As numerical examples frequently are more convincing than formal reasonings we shall consider the expected results of testing shown on figure 1.

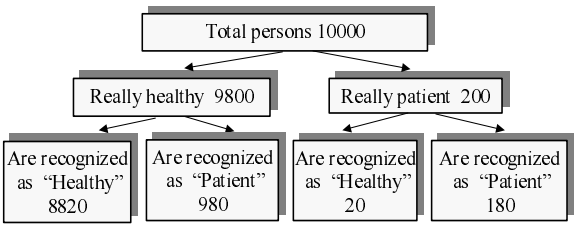

Figure 1: Expected results of testing.

Let's assume that we use some test having the sensitivity $S_E = 90\%$ and the specificity $C_P = 90\%$ for revealing patients from the large group (10000 persons) and known prevalence of disease $P(V_1) = 2\%$. We shall believe also that nothing costs of correct decisions, i.e. $L_{11} = L_{22} = 0$ and losses connected to mistakes are $L_{12} = 5$ and $L_{21} = 1$.

As prevalence is 2% we may assume that only 200 persons from 10000 testing persons are really sick. So before testing expected losses is

$$R_0 = (L_{12} \cdot 200)/10000 = (5 \cdot 200)/10000 = 0{,}1.$$

But after testing it may happened that 980 really healthy persons will be falsely recognized as "Patient" and 20 really sick persons will not recognized as "Patient". So after testing expected losses is bigger than before testing:

$$R = (1 \cdot 980 + 5 \cdot 20)/10000 = 0{,}108.$$

Hence the test is useless. It is easily verified that the formal condition of Theorem 1 is not valid.

The following theorems are direct consequences of the Theorem 1.

Theorem 2. Let $\theta \geq 1$. Then for any possible sensitivity $0 \leq C_E \leq 1$ the test is obviously useless if the specificity of the test satisfies to the condition

$$C_P \leq \frac{1 - P(V_1)(1+\omega)}{1 - P(V_1)}. \qquad (7)$$

Theorem 3. Let $\theta < 1$. Then for any possible specificity $0 \leq C \leq 1$ the test is obviously useless if the sensitivity of the test satisfies to the condition

$$C_E \leq 1 - \frac{1 - P(V_1)}{\omega P(V_1)}. \qquad (8)$$

The very important result follows from Theorems 2 and 3: there are the low boundary of test's specificity and also the low boundary a test's sensitivity are defined as right-hand sides of (7) and (8).

As seen from figure 2, the specificity of the useful test must be more than 89 % when $P(V_1) < 0{,}02$ and $\omega = 5$ and more than 97 % when $\omega = 1$.

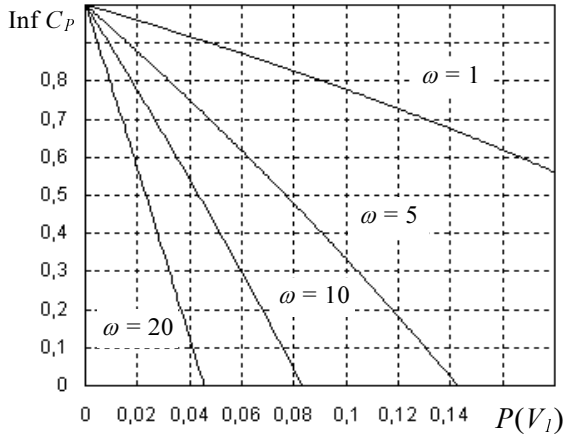

Figure 2: Low boundary of useful test specificity.

4. Comparison to the traditional ROC analysis

It is known that traditional ROC analysis provides a concise description of trade-off available between sensitivity C_E and specificity C_P - the two related but distinct aspects of diagnostic performance. An empirical ROC plot consists of the sequence of discrete points in co-ordinates C_E and $1 - C_P$ obtained by varying the diagnostic cut-off value (threshold). If distributions of classes are different then

$$C_E > 1 - C_P, \qquad (9)$$

and so the ROC plot is placed over diagonal of the ROC-space.

Of course this is true but for practice it is also important to know: whether the given test will allow to reduce the *a priory* risk or not?

From Theorem 1 follows that in general the inequality (9) isn't sufficient to guarantee that the test is really useful. That is why we propose to reinforce the traditional ROC analysis by limiting a permissible portion of the ROC curve.

Figure 3 illustrate suggested notion to the case $\theta \geq 1$. According to condition (3) in this case the boundary line begins from the point with coordinates (0,0) to the point with coordinates ($1/\theta$, 1).

Line OA correspond to the case when the prevalence of disease is equal 15 % and $\omega = 1$. We see that in this case the test is useless because the ROC-curve is placed bellow line OA. But if we suppose that the cost of false negative error is four times over than false positive one and nothing costs of correct decisions, i.e. $\omega = 4$ then corresponding line (OB) crosses the ROC curve. So we may use the test having the sensitivity $C_E = 62{,}5\%$ and the specificity $C_P = 80\%$ because corresponding point C belongs to the permissible part of the ROC-curve.

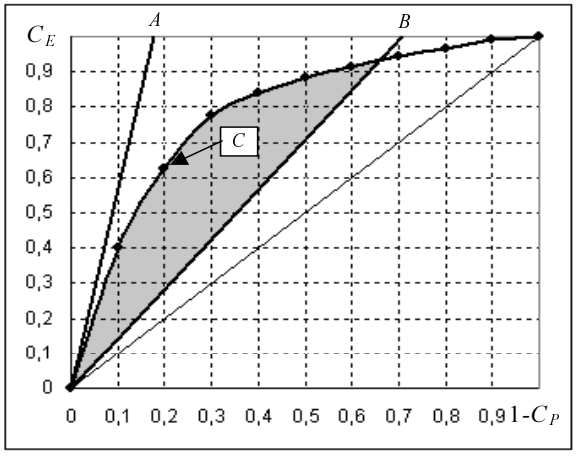

Figure 3: The limited ROC-plot for the case $\theta \geq 1$

Using the condition (4) we may limit the ROC-plot for the case $\theta < 1$ also. But in this case the limiting line begins from the point having coordinates $(1,1)$ to a point having coordinates $(0, 1-\theta)$ and so it limits the ROC-curve from the other side (see Figure 4).

Finally, we can use conditions (3) and (4) to decide the inverse task: what allowable range of loss ratio can be chosen for useful test when the prevalence $P(V_1)$, the sensitivity C_E and the specificity C_P are known. This range may be obtained by the formula:

$$\frac{1-P(V_1)}{P(V_1)}\frac{1-C_P}{C_E} \le \omega \le \frac{1-P(V_1)}{P(V_1)}\frac{C_P}{1-C_E}. \quad (10)$$

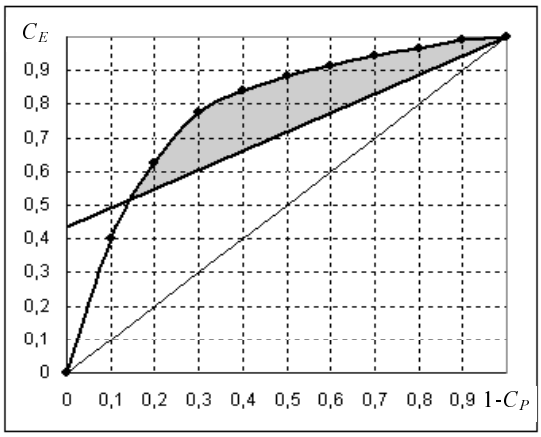

Figure 4: The limited ROC-plot for the case $\theta < 1$

5. Practical application

Obtained results have helped us to prove a diagnostic performance of novel diagnostic test for predict organic heart's diseases, first of all, initial stage of ischemia (CAD). The test is based on measurement of some original parameter β describing the form of person's electrocardiogram in the phase space [7].

To study the diagnostic performance of this test we used the clinical material from four German clinics: Essen University Hospital, Katholical Hospital "Phillpusstift" (Essen), Heart and Diabetes Center of North Rhein-Weasfalia (Bad-Oeynhausen) and German Heart Center (Berlin). The clinical material included 441 ECG records of verified CAD patients and 387 ECG records of healthy volunteers included in the control group.

It is interesting that all records including records of really CAD-patients had not any abnormalities of traditional for cardiology ECG's parameters.

At the same time, average values of parameter β were different for healthy persons and CAD-patients. It allowed us to construct the simple test rule

$$\begin{aligned} &\textbf{CAD}, \quad if \ \beta > \beta_0, \\ &\textbf{Healthy} \quad if \ \beta \le \beta_0, \end{aligned} \quad (11)$$

where $\beta_0 = const$ - threshold value. This test have sensitivity $S_E = 81\%$ and specificity $S_P = 78\%$.

Because CAD-prevalence is about 6 % so from (10) follows that the offered test is useful for the practice in a wide range of loss ratio:

$$4{,}2 \le \omega \le 64{,}1.$$

6. Conclusion

The diagnostic test or the separate diagnostic feature are useful in the sense of a risk reduction (Determination 1) if and only if for given *a priory* probabylity $P(V_1)$ and given loss ratio ω the test's sensitivity C_E and the test's specificity C_P satisfy to conditions of the Theorem 1. We estimate the low boundary of a useful test sensitivity and the low boundary of a useful test specificity (Theorem 2 and 3) which can be used on practice. Obtained condition allows also improving the tradition ROC analysis by formal limiting a permissible portion of the ROC-plot (Figure 3 and Figure 4).

Based on obtained result we have proved a performance of the novel diagnostic feature of ECG in the phase space for diagnostic initial stage of a heart organic disease.

References

[1] M. Ben-Bassat. Irrelevant Features in Pattern Recognition. In *IEEE Trans. Comput,* vol. C-27, No. 8, 749-766, 1978.

[2] C.E.Metz. Fundamental ROC analysis. In *Progress in Medical Physics and Psychophysics Handbook of Medical Imaging*, vol. 1, SPIE Press, Bellingham, WA, 754-769, 2000.

[3] T.Fawcett. Using Rule Sets to Maximize ROC Performance. In *Proceedings of the IEEE International Conference on Data Mining (ICDM-2001).* Los Alamitos, CA, IEEE Computer Society. 131-138, 2001.

[4] C.Ferri, J.Hernández-Orallo., M.A.Salido. Volume Under the ROC Surface for Multiclass Problems. Exact Computation and Evaluation of Approximations. In *Technical Report DSIC.* Univ. Politèc. València, 2003

[5] L.S. Fainzilberg. Interconnection between feature properties and probability of error in statistical recognition of two classes. In *Proc. of the 12th Int. Conf. on Pattern Recognition (ICPR'94, Jerusalem), IEEE Comp. Soc.Press. LA, California*, vol. 2, 544-546, 1994.

[6] L.S. Fainzilberg. Why relevant features may be use less in statistical recognition of two classes. In *Proc. of the 13th Int. Conf. on Pattern Recognition (ICPR'96, Viena), IEEE Comp. Soc.Press. LA, California,* vol. 2, 730-734, 1996.

[7] L.S. Fainzilberg. Nowa metoda interpretacji zapisu EKG w balaniach skriningowych oraz w opiece domowej. In *Zdrowie publiczne (Public health),* vol. 115, No. 4, 458-464, 2005.