Controlling True Positive Rate in ROC Analysis

T Eftestøl

Faculty of Science and Technology, University of Stavanger, Stavanger, Norway

Abstract

ROC analysis is a widely used method for evaluating the performance of classifiers. In analysis involving scarce data sets leave-one-out resampling techniques might be appropriate. This introduces a problem in terms of computing average ROC curves necessary to determine variance in the true positive and negative rates. A method to determine decision regions for a specified true positive rate is presented. The method is based on estimating the class specific probability density functions for the two classes. The functions are discretised. Dividing these yields a function where values above or below a specific threshold value corresponds to deciding class one or two respectively. It is shown how a gradual lowering of the threshold value corresponds to an increase in the true positive rate, and how a true positive rate can be specified and the corresponding threshold determined. An example with simulated data is used to demonstrate the method.

1. Introduction

ROC analysis is widely used to evaluate the performance of diagnostic markers[1]. The method is quite straightforward in use when only a single marker is evaluated[1]. However, when several markers are combined in a multi-dimensional feature vector, decision regions can be determined by use of Bayes decision theory[2]. The true positive and negative rates can be controlled by use of loss functions to set the size of the decision areas. Another issue to consider is the problem of using resampling, repeatedly determining decision regions for specific true positive and negative rates[1, 3]. With data material being scarce, the correspondence between the decision regions in the resampling iterations will be poor. A method is proposed for accurately controlling the true positive rate which can be used for problems involving small data sets and use of resampling.

2. Methods

The method estimates the probability density functions (PDF) for the two data sets

$$D_k = \{\mathbf{x}_{k_1}, \ldots, \mathbf{x}_{k_{M_k}}\}, k = 1, 2, \tag{1}$$

of feature vectors

$$\mathbf{x} = (x_1\ x_2\ \ldots\ x_d)^t, \tag{2}$$

to be discriminated and represents values of these on an evenly distributed grid of coordinates,

$$X = \{\mathbf{x}_n, \ldots, \mathbf{x}_N\}, \tag{3}$$

representing the feature space. A two class simulated data set and a grid representation of the feature space is shown in shown in figure 1.

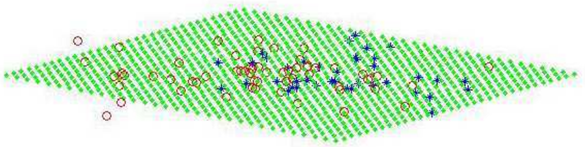

Figure 1: A two class data set ($*/\circ = \omega_1/\omega_2$) superimposed on feature space (X discretisation represented by $\cdot$).

As will be shown, the PDFs will be used to control the size of the decision regions, and thus the number of grid points in the decision area for ω_1 will determine the true positive rate resolution.

The PDFs for the two classes,

$$p(\mathbf{x}|\omega_k), k = 1, 2, \tag{4}$$

are estimated. Maximum likelihoood (ML) estimation is used under the assumption of Gaussian distribution estimated and represented on the grid and normalised (sums to 1)[2]. The estimated PDFs for the two datasets are shown in figure 2 where the computed values of $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$ on the grid are shown as blue and red dots respectively. The prior probabilities are estimated as $P(\omega_1)$ and $P(\omega_2)$. According to Bayes decision theory, choosing

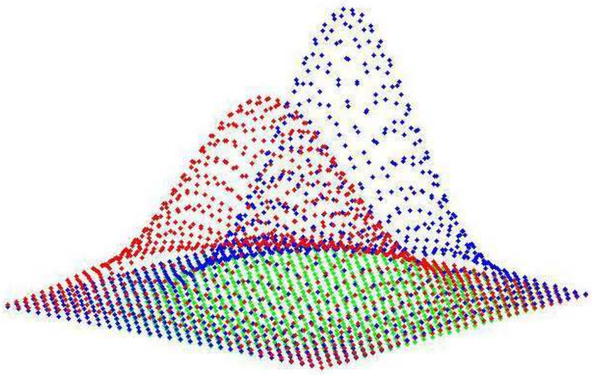

Figure 2: Discretised PDF functions. Values of $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$ are represented as $\cdot$ and $\cdot$ respectively.

the class with highest $g_k = P(\omega_k)p(\mathbf{x}_k|\omega_k), k = 1, 2$ minimises the error rate[2]. Alternatively, one might express this: Select class ω_1 if

$$G(\mathbf{x}) = \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}. \qquad (5)$$

where λ_{ij} express the loss of wrongly classifying $\mathbf{x}$ as ω_i when the trye class is ω_2. Otherwise class ω_2 is selected.By setting $\lambda_{ii} = 0, i = 1, 2$ and $\lambda_{12} = 1$ (5) is reduced to

$$G = \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{1}{\lambda_{21}} \frac{P(\omega_2)}{P(\omega_1)}. \qquad (6)$$

$T = \frac{1}{\lambda_{21}} \frac{P(\omega_2)}{P(\omega_1)}$ can be considered a threshold, and the true positive and negative rates can be changed by adjusting the value of T which is controlled solely by λ_{21}. $G(\mathbf{x})$ is represented on the grid as shown in figure 3 where the computed values of $G(\mathbf{x})$ on the grid is shown as yellow dots. A threshold value is shown as a linear plane in cyan. The computed values of $G(\mathbf{x})$ are arranged in descending order $l = 1, 2, \ldots, L$ in a vector T'. Each element in this vector, $T'(i)$ represents a coordinate in feature space corresponding to a coordinate in the discretised feature space X, $\mathbf{x}_i$. The values in the $p(\mathbf{x}|\omega_1)$ representation is arranged correspondingly in a vector p_1 so that $p_1(i)$ is the PDF value $p(\mathbf{x}_i|\omega_1)$ corresponding to the ith largest value of T'. The values in the $p(\mathbf{x}|\omega_2)$ representation is also arranged in this way in the vector p_2. The accumulated sum of p_1 is computed and named TP. Thus, $TP(i)$, is the true positive rate corresponding to the threshold value $T'(i)$. Figure 4 illustrates this function. For a specific desired true positive value, TP is searched, and the first occurrence at or above this value $TP(i)$ is found. As the index values $1, \ldots, i$ corresponds to the i largest values in T and on the surface of $G(\mathbf{x}$, these coordnates define the decision region, R_1 for ω_1. The values of TP lower than 0.1 on

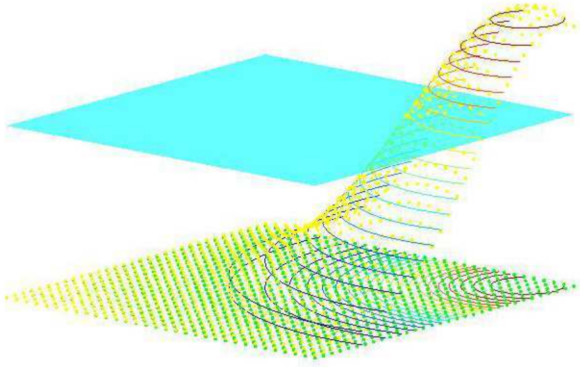

Figure 3: Discretised decision function $G(\mathbf{x})$ (values represented as $\cdot$. A possile threshold is shown as a cyan plane while contour lines on $G(\mathbf{x})$ and feature space representation X illustrates possible threshold values and decision region borders.

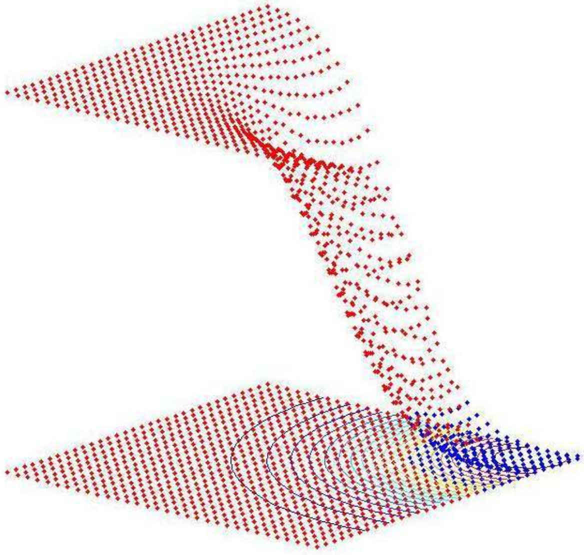

Figure 4: A representation of the true positive rate. As the threshold value is lowered to $T'(i)$, the corresponding value $TP(i)$ gives the true positive rate. On the displayed function surface, $TP(i)$ corresponds to the largest value shown in blue. The blue dots correspond to $TP(1), \ldots, TP(i)$. The corresponding coordinates in X belong to the decision region for ω_1, R_1 ($\cdot$). The remaining part belongs to R_2 ($\cdot$).

the TP function in figure 4 and the coordinates in feature space in R_1 are both indicated in blue, while the values and coordinates in the adjacent area, R_2 is shown in red.

The threshold value for these decision regions are found as $T(i)$ and shown as a linear plane in cyan color in figure 3.

Once these functions are computed, a desired level of true positives can be found from a search in the TP vector, and then the corresponding threshold value, $T(i)$ is found. From this threshold value, λ_{21} can be computed if that is desirable. The true negative (TN) value can be computed by summing the values of $p(\mathbf{x_n}|\omega_2)$ in R_2 which corresponds to $TN = \sum_{l=i+1}^{L} p_2(i)$.

As the number of grid points can be freely chosen, the problem of maintaining the same true positive value throughout resampling can be handled.

In cases where the data material is scarce, resampling by leave-one-out can be used. If we consider the complete data set as the union of D_1 and D_2 as

$$D = \{D_1, D_2\} = \{\mathbf{x}_{1_1}, \ldots, \mathbf{x}_{1_{M_1}}, \mathbf{x}_{2_1}, \ldots, \mathbf{x}_{2_{M_2}}\}, \quad (7)$$

we train $M = M_1 + M_2$ classifiers from the datasets

$$D_j = \{\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \mathbf{x}_M\}, i = 1, \ldots, M, \quad (8)$$

each time leaving out $\mathbf{x}_j$ for testing. For D_j a classifier is trained and the thresholds for the desired true positive rates $TP_d(m), m = 1, \ldots, Q$ is determined as described above. For each threshold value, $TP_d(j)$, the classification of $\mathbf{x}_j$ is determined as $C_k(m, j) = k, k = \{1, 0\}$ where 1 denotes a correct classification and 0 a false classification. Thus, the matrix $\mathbf{C_k}, \mathbf{k = 1, 2}$ represent all the classifications for all left out samples at all threshold values for the two data sets $D_k, k = 1, 2$.

$$\mathbf{C_k} = \begin{pmatrix} C_k(1,1) & \cdots & C_k(1,j) & \cdots & C_k(1,M_k) \\ \vdots & \vdots & & \vdots & \\ C_k(m,1) & \cdots & C_k(m,j) & \cdots & C_k(j,M_k) \\ \vdots & \vdots & & \vdots & \\ C_k(Q,1) & \cdots & C_k(Q,j) & \cdots & C_k(Q,M_k) \end{pmatrix} \quad (9)$$

where row number m gives all the classifications for the treshold trained to $TP(m)$ and column number j gives all classifications for a specific $\mathbf{x}_j$. The mean true positive value $TP(m)$ for the tested feature vectors can be computed as

$$TP_t(m) = \frac{\sum_{j=1}^{M_1} C_1(m, j)}{M_1}. \quad (10)$$

The corresponding mean true negative value $TN(m)$ can be computed as

$$TN_t(m) = \frac{\sum_{j=1}^{M_2} C_2(m, j)}{M_2}. \quad (11)$$

Confidence limits can be computed according to the method proposed for binomial data[4, 5].

3. Examples

Data were simulated for two two-class classification problems. Gaussian distributions were assumed with mean values $\mu_1 = (1\ 1)^t$, $\mu_2 = (-1\ -1)^t$ and covariance being equal to the the identity matrix for problem number 1. The parameters for the distributions were the same for problem 2 except for $\mu_1 = (2\ 2)^t$, $\mu_2 = (-2\ -2)^t$. The prior probabilites for both problems were set to $P(\omega_1) = 0.4$ and $P(\omega_1) = 0.6$. 500 feature vectors were generated for each problem and the classifiers were designed for the full data set and reduced data sets at 250, 100, 50 and 25 samples.

The results for problem 1 is shown in figure 5. The first

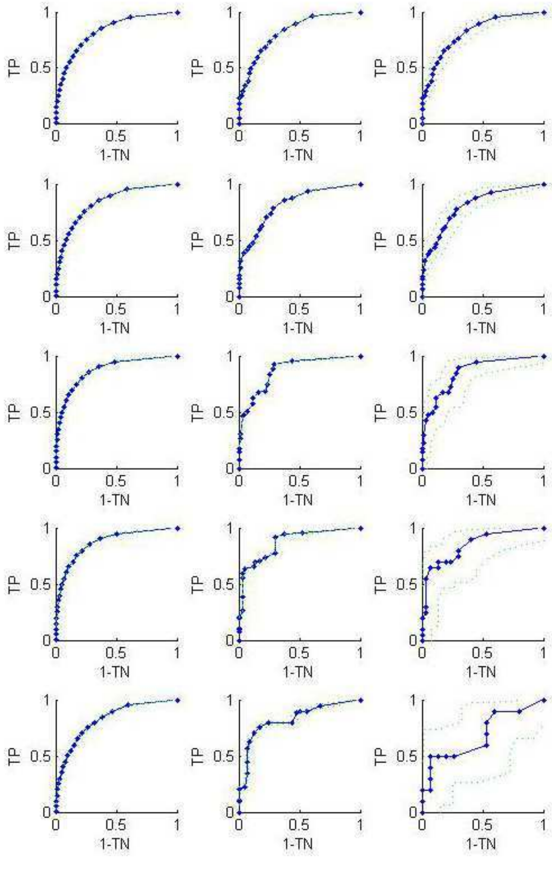

Figure 5: ROC curves for problem 1. The first column shows the ROC curves for the model based TP and TN values. The second column shows the ROC curves for the data used in training. The third column shows the ROC curves for the resampled data. The number of samples decrease from the top: 500, 250, 100, 50, 25 samples.

column shows the ROC curves for the model based TP and TN values. The second column shows the ROC curves for the data used in training. The third column shows the ROC curves for the resampled data. The green curves show the confidence limits in the third column and the standard deviations from the mean values in the two other columns. From top to bottom, the number of samples decrease.

The corresponding ROC curves for problem 2 is shown in figure 6.

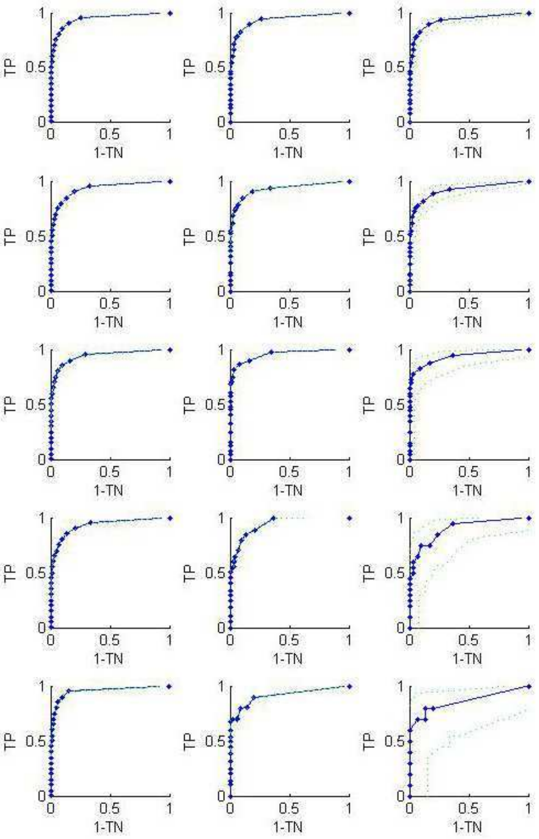

Figure 6: ROC curves for problem 2. The first column shows the ROC curves for the model based TP and TN values. The second column shows the ROC curves for the data used in training. The third column shows the ROC curves for the resampled data. The number of samples decrease from the top: 500, 250, 100, 50, 25 samples.

4. Discussion and conclusions

As can be seen from the ROC curves of the test data, the uncertainty increases as the number of samples decreases. It is also interesting to note the discrepancy between the ROC curve for the test data compared to the ROC curves for the model and the training data. This might indicate that the the number of samples is too low. In addition this happens for problem 1 where the two classes are less separated.

One might speculate that good correspondence between the model, training data and resampled test data ROC curves indicates good generality in the classifier. These relationships might be investigated further in further development of this work. Further investigations should be made to compare the present method to established methods in ROC analysis.

A possible limitation of the method that should be investigated is the effect of limiting the feature space.

A method for controlling the true positive rates for multidimensional feature vectors has been described and demonstrated on artificial data.

References

[1] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters 2006;27:861–874.

[2] Duda RO, Hart PE, Stork DG. Pattern Classification. Second edition. New York: John Wiley and Sons, Inc., 2001.

[3] Macskassy S, Provost F. Confidence bands for ROC curves: Methods and an empirical study. In Proceedings First Workshop on ROC Analysis in AI (ROCAI-04). 2004; .

[4] Agresti A, Coull B. Approximate is better than ”exact” for interval estimation of binomial proportions. The American Statistician 1998;52(2):119–126.

[5] Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Statistical Science 2001;16(2):101–133.

Address for correspondence:

Trygve Eftestøl
University of Stavanger, N-4036 Stavanger, Norway
tel./fax: ++47-5183-2035/1750
trygve.eftestol@uis.no