An Optimal Reject Rule for Binary Classifiers

Francesco Tortorella

Dipartimento di Automazione, Elettromagnetismo,
Ingegneria dell'Informazione e Matematica Industriale
Università degli Studi di Cassino
Cassino (FR), Italy
tortorella@unicas.it

Abstract. Binary classifiers are used in many complex classification problems in which the classification result could have serious consequences. Thus, they should ensure a very high reliability to avoid erroneous decisions. Unfortunately, this is rarely the case in real situations where the cost for a wrong classification could be so high that it should be convenient to reject the sample which gives raise to an unreliable result. However, as far as we know, a reject option specifically devised for binary classifiers has not been yet proposed. This paper presents an optimal reject rule for binary classifiers, based on the *Receiver Operating Characteristic* curve. The rule is optimal since it maximizes a classification utility function, defined on the basis of classification and error costs peculiar for the application at hand. Experiments performed with a data set publicly available confirmed the effectiveness of the proposed reject rule.

1 Introduction

Many complex classification problems involve binary decisions, since they require to choose between two possible, alternative classes. Applications such as automated cancer diagnosis, currency verification, speaker identification, and fraud detection fall in this category. Their common feature is that the classification result could have serious consequences: for this reason, the classifiers with binary outcomes (shortly, *binary classifiers*) used in these situations should ensure a very high reliability to avoid erroneous decisions. Unfortunately, in real world this is rarely the case because, when working on real data, the classifiers could easily encounter samples very different from those learned during the training phase. In these cases, the cost for a wrong classification could be so high that it should be convenient to suspend the decision and call for a further test, i.e. to reject the sample. Obviously, such a reject option should be defined by taking into account the requirements of the given application domain.

This topic has been addressed with reference to multi-class classifiers by Chow in [1,2]. The rationale of the Chow's approach relies on the exact knowledge of the *a posteriori* probabilities for each sample to be recognized. Under this hypothesis, the Chow's rule is optimal because minimizes the error rate for a given reject rate (or viceversa). However, the full knowledge about the distributions of the classes is ex-

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 611-620, 2000.

tremely difficult to obtain in real cases and thus the Chow's rule is rarely applicable "as it is". An extension to the Chow's rule when the a priori knowledge about the classes is not complete is proposed in [3] and in [4], while in [5] a reject option that does not require any a priori knowledge is proposed with reference to a Multi-Layer Perceptron. Although effective, these rules are applicable only with multi-class classifiers. As far as we know, a reject option specifically devised for binary classifiers has not been yet proposed.

The aim of this paper is to introduce an optimal reject rule for binary classifiers, based on the *Receiver Operating Characteristic* curve (ROC curve). ROC analysis is based in statistical decision theory and was first employed in signal detection problems. It is now common in medical diagnosis and particularly in medical imaging. Recently, it has been employed in Statistical Pattern Recognition for evaluating machine learning algorithms [6] and for robust comparison of classifier performance under imprecise class distribution and misclassification costs [7].

In the method here presented the information about the classifier performance provided by the ROC curve are employed to build an optimal reject rule. The rule is optimal since it maximizes a classification utility function *U*(.), defined on the basis of classification and error costs peculiar for the application at hand. Experiments performed with a data set publicly available confirmed the effectiveness of the proposed reject rule.

2 ROC Curve

In binary classification problems, a sample can be assigned to one of two mutually exclusive classes that can be generically called *Positive* (*P*) class and *Negative* (*N*) class. Let us assume that the classifier provides, for each sample, a value x in the range [0,1] which is a confidence degree that the sample belongs to one of the two classes, e.g. the class *P*. The sample should be consequently assigned to the class *N* if $x \to 0$ and to the class *P* if $x \to 1$. Operatively, a confidence threshold t is usually chosen, so as to attribute the sample to the class *N* if $x \leq t$ and to the class *P* if $x > t$. For a given threshold value t, some indices can be evaluated for measuring the performance of the classifier. In particular, the set of samples whose confidence degree is greater than t contains actually-positive samples correctly classified as "positive" and actually-negative samples incorrectly classified as "positive". It is thus possible to define the *True Positive Rate TPR*(*t*) as the fraction of actually-positive cases correctly classified and the *False Positive Rate FPR*(*t*), given by the fraction of actually-negative cases incorrectly classified as "positive".

If $f_N(x)$ and $f_P(x)$ are the density functions of the confidence degree for the class *N* and for the class *P*, respectively, *TPR*(*t*) and *FPR*(*t*) are given by (see fig. 1):

$$TPR(t) = \int_t^1 f_P(x)dx \qquad FPR(t) = \int_t^1 f_N(x)dx \tag{1}$$

In a similar way it is possible to evaluate (taking into account the samples with confidence degree less than *t*) the *True Negative Rate TNR*(*t*) and the *False Negative Rate FNR*(*t*), defined as:

$$TNR(t) = \int_0^t f_N(x)dx = 1 - FPR(t) \qquad FNR(t) = \int_0^t f_P(x)dx = 1 - TPR(t) \tag{2}$$

Since the four indices are not independent, as it is possible to note from eq. (2), the pair (*FPR*(*t*),*TPR*(*t*)) is sufficient to completely characterize the performance of the classifier when the decision threshold is set to *t*. Fig. 1 shows how these quantities can be represented on a plane having *FPR* on the X axis and *TPR* on the Y axis.

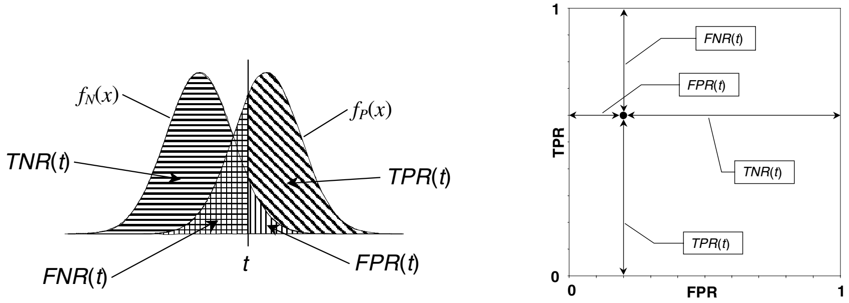


Fig. 1. The indices *TPR*, *FPR*, *TNR* and *FNR* evaluated on two bell-shaped confidence densities (*left*). The same quantities mapped on a (*FPR*, *TPR*) plane (*right*).

When the value of the threshold *t* varies between 0 and 1 the quantities in eq. (1) and eq. (2) vary accordingly, thus defining a set of operating points for the classifier, given by the pairs (*FPR*(*t*),*TPR*(*t*)). The plot of such points gives the ROC curve of the classifier (see fig. 2).

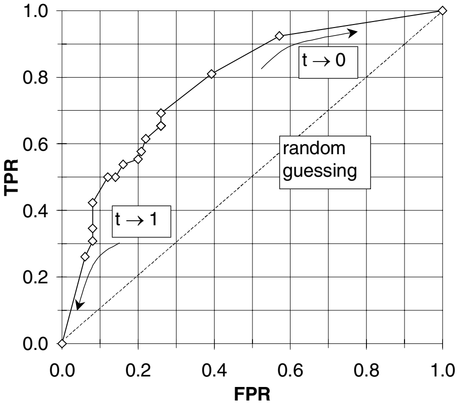


Fig. 2. A typical ROC curve.

It is worth noting that, when t approaches 0, both $TPR(t)$ and $FPR(t)$ approach 1, while the contrary happens when $t \to 1$. Informally, the nearer the curve to the upper left corner of the diagram, the better the performance obtained (higher *TPR* and/or lower *FPR*). An important reference is given by the line joining the points (0,0) and (1,1) which represents the case of a random guessing classifier.

3 The Reject Option

When a classifier is used in a real application, its outcomes have consequences to which is associated a benefit (in the case of success) or a loss (in the case of error). Thus, the effectiveness of the classifier in a given domain should be measured on the basis of both its absolute performance (correct classification rate and error rate) and the costs associated to the various outcomes. In the case of binary classes, such costs can be organized in the cost matrix shown in table 1.

Table 1. Cost matrix for a two-class problem

		Guess Class	
		N	*P*
True Class	*N*	*CTN*	*CFP*
	P	*CFN*	*CTP*

In the cost matrix, *CTN* and *CTP* are ≥ 0 since related to benefits, while CFN and CFP are ≤ 0.

In general, the cost matrix is not symmetrical, because the consequences of different errors are usually not equivalent. As an example, in the case of medical diagnosis a false negative outcome is much more costly of a false positive. Likewise, if the disease is rare, a true positive outcome might be much more appraised than a true negative outcome. Once the cost matrix has been established on the basis of the particular application requirements, it is possible to define a *classification utility* function $U(t)$ which measures, for a given decision threshold, the effectiveness provided by the binary classifier:

$$\begin{aligned} U(t) = p(P) \cdot \left[CTP \cdot TPR(t) + CFN \cdot FNR(t)\right] + \\ p(N) \cdot \left[CTN \cdot TNR(t) + CFP \cdot FPR(t)\right]. \end{aligned} \quad \textbf{(3)}$$

where $p(P)$ and $p(N)$ are the a priori probabilities of the positive and negative classes, respectively. In this way, the optimal decision threshold t_{opt} can be determined as:

$$t_{opt} = \arg\max_{t} U(t) \quad \textbf{(4)}$$

However, there can be real situations in which the cost of an error is so high that it is advisable to suspend the decision and to reject the sample if the outcome is considered unreliable. The rejection also involves a negative cost (indicated with *CR*), which is related to the price of a new classification with another system and has smaller magnitude with respect to the error costs.

To accomplish the reject option in a binary classifier, the decision rule for a generic sample with confidence degree x should be changed into:

$$\begin{array}{ll} \text{assign the sample to } N & \text{if } x < t_1 \\ \text{assign the sample to } P & \text{if } x > t_2 \\ \text{reject the sample} & \text{if } t_1 \le x \le t_2 \end{array} \qquad \mathbf{(5)}$$

where t_1 and t_2 are two decision thresholds (with $t_1 \le t_2$) fixed so as to maximize the utility function.

As a consequence, the rates defined in eq. (1) and eq. (2) are modified in:

$$\begin{array}{ll} TPR(t_2) = \int_{t_2}^{+\infty} f_P(x)dx & FPR(t_2) = \int_{t_2}^{+\infty} f_N(x)dx \\ TNR(t_1) = \int_{-\infty}^{t_1} f_N(x)dx & FNR(t_1) = \int_{-\infty}^{t_1} f_P(x)dx \end{array} \qquad \mathbf{(6)}$$

while the reject rates relative to negative samples, $RN(t_1,t_2)$, and to positive samples, $RP(t_1,t_2)$, are given by:

$$\begin{array}{l} RN(t_1,t_2) = \int_{t_1}^{t_2} f_N(x)dx = 1 - TNR(t_1) - FPR(t_2) \\ RP(t_1,t_2) = \int_{t_1}^{t_2} f_P(x)dx = 1 - TPR(t_2) - FNR(t_1) \end{array} \qquad \mathbf{(7)}$$

Accordingly, the utility function becomes:

$$\begin{aligned} U(t_1,t_2) = p(P)\cdot CFN \cdot FNR(t_1) + p(N)\cdot CTN \cdot TNR(t_1) + \\ p(P)\cdot CTP \cdot TPR(t_2) + p(N)\cdot CFP \cdot FPR(t_2) + \\ p(P)\cdot CR \cdot RP(t_1,t_2) + p(N)\cdot CR \cdot RN(t_1,t_2) . \end{aligned} \qquad \mathbf{(8)}$$

If we take into account the relations given in eq. (7), the utility function can be written as:

$$U(t_1,t_2) = U_1(t_1) + U_2(t_2) + CR . \qquad \mathbf{(9)}$$

where:

$$U_1(t_1) = p(P)\cdot CFN' \cdot FNR(t_1) + p(N)\cdot CTN' \cdot TNR(t_1) . \qquad \mathbf{(10)}$$

$$U_2(t_2) = p(P)\cdot CTP' \cdot TPR(t_2) + p(N)\cdot CFP' \cdot FPR(t_2) . \qquad \mathbf{(11)}$$

and

$$CTP' = CTP - CR \quad CFN' = CFN - CR \quad CTN' = CTN - CR \quad CFP' = CFP - CR$$

In this way, the optimal thresholds (t_{1opt}, t_{2opt}) which maximize $U(t_1,t_2)$ can be separately evaluated by maximizing $U_1(t)$ and $U_2(t)$:

$$t_{1opt} = \arg\max_t \; p(P) \cdot CFN' \cdot FNR(t) + p(N) \cdot CTN' \cdot TNR(t) \,. \tag{12}$$

$$t_{2opt} = \arg\max_t \; p(P) \cdot CTP' \cdot TPR(t) + p(N) \cdot CFP' \cdot FPR(t) \,. \tag{13}$$

By taking into account the relations introduced in eq. (2), the optimization problem in eq. (12) is equivalent to:

$$t_{1opt} = \arg\min_t \; p(P) \cdot CFN' \cdot TPR(t) + p(N) \cdot CTN' \cdot FPR(t) \,. \tag{14}$$

It is worth noting that the objective functions in eq. (14) and eq. (13) define on the ROC plane two sets of level curves having parametric equations:

$$p(P) \cdot CFN' \cdot TPR(t) + p(N) \cdot CTN' \cdot FPR(t) = k_1 \,. \tag{15}$$

and

$$p(P) \cdot CTP' \cdot TPR(t) + p(N) \cdot CFP' \cdot FPR(t) = k_2 \,. \tag{16}$$

Each set is composed by parallel straight lines. The slopes associated to the sets are, respectively:

$$m_1 = -\frac{p(N) \cdot CTN'}{p(P) \cdot CFN'} \qquad m_2 = -\frac{p(N) \cdot CFP'}{p(P) \cdot CTP'} \,. \tag{17}$$

Since the set of feasible points for both the objective functions is given by the ROC curve, the optimal threshold t_{1opt} can be determined by searching the point on the ROC curve belonging also to the line defined by eq. (15) which intersects the ROC and has minimum k_1. In a similar way can be found t_{2opt}, with the only difference that we must consider the line that satisfies eq. (16), intersects the ROC curve and has maximum k_2. It can be simply shown that, in both cases, the searched line is the level curve that intersects the ROC and has largest *TPR*-intercept. Such a line lies on the *ROC Convex Hull* [7], i.e. the convex hull of the set of points belonging to the ROC curve (see fig. 3).

In particular, the line could share with the ROC convex hull only one point (a vertex of the convex hull) or an entire edge. In the first case, the optimal threshold is given by the value of *t* associated to the point. In the second case, either of the two vertices of the segment can be chosen; the only difference is that the left vertex will have lower *TPR* and *FPR*, while the right vertex will have higher *TPR* and *FPR*.

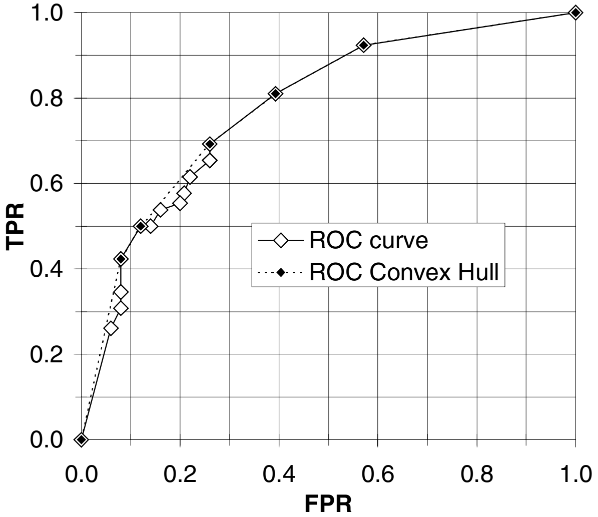


Fig. 3. A ROC curve with its convex hull.

To give an operative method for finding the optimal thresholds, let us call V_0, V_1, …, V_n the vertices of the ROC convex hull, with $V_0 \equiv (0,0)$ and $V_n \equiv (1,1)$; moreover, let s_i be the slope of the edge joining the vertices V_{i-1} and V_i and assume that $s_0 = \infty$ and that $s_{n+1} = 0$. If m is the slope of the level curve of interest, the list $\{s_i\}$ should be searched to find a value s_k such that $s_k = m$ or $s_k > m > s_{k+1}$: in the first case, the level curve and the edge are coincident and thus either of the vertices V_{k-1} and V_k can be chosen. In the second case, the level curve touches the ROC convex hull in the vertex V_k, which provides the optimal threshold.

It is important to recall that t_{1opt} must be less than t_{2opt} to achieve the reject option. For this reason, the slopes m_1 and m_2 defined in eq. (17) must be such that $m_1 < m_2$, otherwise the reject option is not practicable.

4 Experimental Results

For testing the proposed reject rule, a medical dataset (the *Pima Indians Diabetes* dataset), publicly available from the UCI Machine Learning Repository [8], has been considered. This dataset involves the diagnosis of diabetes diseases on the basis of the results of several tests. The data were collected by the National Institute of Diabetes and Digestive and Kidney Diseases. All of the patients were females at least 21 years old of Pima Indian heritage. The class variable has the values 0 (healthy) and 1 (diabetes). The dataset contains 768 labeled cases (500 healthy and 268 diabetes), each including 8 continuously valued inputs.

The classifier adopted is a Multi Layer Perceptron with 8 input units, 4 hidden units and 1 output unit, implemented in C language using the SPRANNLIB library [9]. The network has been trained for 20,000 epochs using the back propagation algorithm with a learning rate of 0.01 and a momentum of 0.2. The set used for the training contained the 80% of the samples of the whole dataset. The remaining 20% were split into two different sets, the first one for evaluating the ROC curve, while the second one was adopted as test set.

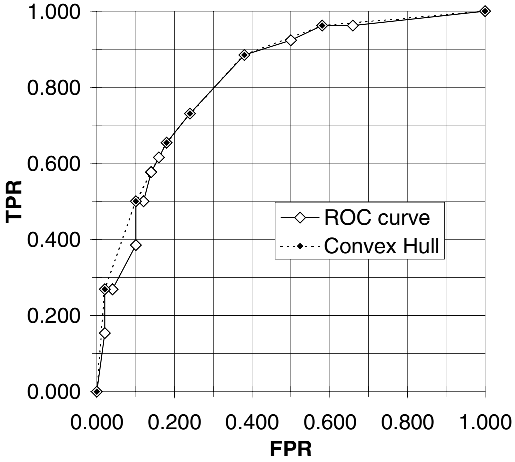


Fig. 4. The obtained ROC curve together with its convex hull.

In fig. 4, the ROC curve obtained is shown together with its convex hull. The coordinates of the vertices with the respective threshold values and the slopes of the edges of the ROC convex hull are listed in tables 2 and 3.

Table 2. The ROC convex hull vertices

	ROC convex hull Vertices	*t*
0	(0.00 0.00)	1.00
1	(0.02 0.27)	0.90
2	(0.10 0.50)	0.70
3	(0.18 0.65)	0.30
4	(0.24 0.73)	0.25
5	(0.38 0.88)	0.20
6	(0.58 0.96)	0.10
7	(1.00 1.00)	0.00

Table 3. The ROC convex hull slopes

	ROC Convex hull Edge Slopes
0	∞
1	13.45
2	2.89
3	1.92
4	1.28
5	1.10
6	0.38
7	0.09
8	0.00

The costs considered for the experiments are shown in table 4: seven different cost combinations (denoted by **a-g**) have been chosen which reflect different situations. The reject cost has been assumed constant.

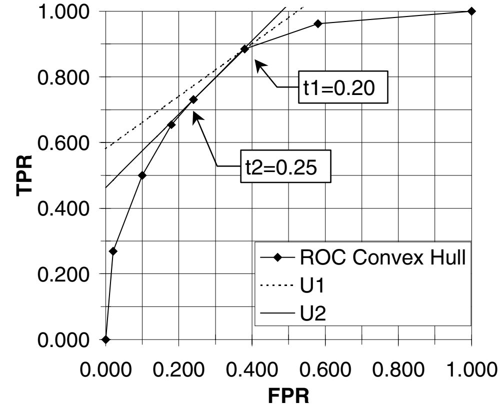
Table 4. The combinations of costs

	CFN	*CFP*	*CTN*	*CTP*	*CR*
a	-50	-25	200	400	-12.5
b	-50	-25	100	200	-12.5
c	-50	-25	50	100	-12.5
d	-50	-25	25	50	-12.5
e	-100	-50	25	50	-12.5
f	-200	-100	25	50	-12.5
g	-400	-200	25	50	-12.5

Table 5. The slopes and the thresholds

	m_1	m_2	t_{1opt}	t_{2opt}
a	10.570	0.057	-	-
b	5.596	0.110	-	-
c	3.109	0.207	-	-
d	1.865	0.373	-	-
e	0.799	1.119	0.20	0.25
f	0.373	2.611	0.10	0.70
g	0.181	5.596	0.10	0.90

Table 5 shows the values for m_1 and m_2 evaluated for each cost combination. It is worth noting that the reject option is achievable only in the last three cases, where $m_1 < m_2$. The relative optimal thresholds, which can be deduced by looking at the tables 2 and 3, are reported in the last two columns. As an example, figure 5 shows the optimal level curves of U_1 and U_2 for the cost combinations **e** and **g**.



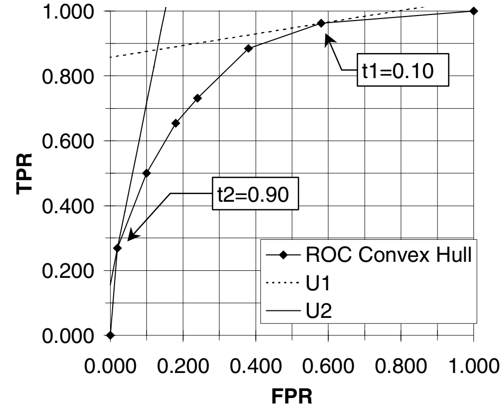


Fig. 5. The optimal level curves of U_1 and U_2 for the cost combinations **e** (*above*) and **g** (*below*). The optimal points on the ROC convex hulls are also highlighted.

Table 6 resumes the results obtained on the test set with and without the reject option. The first six columns contain the rates obtained: these values are costant in the rows **a-d** because the optimal point on the ROC for the utility function without reject is given by *FPR* = 0.38 and *TPR* = 0.88 for all the cost combinations.

Table 6. Results obtained on the test set

	FPR	*TPR*	*FNR*	*TNR*	*RP*	*RN*	*U*	U_{rej}
a	0.38	0.88	0.62	0.12	-	-	196.079	-
b	0.38	0.88	0.62	0.12	-	-	93.944	-
c	0.38	0.88	0.62	0.12	-	-	42.876	-
d	0.38	0.88	0.62	0.12	-	-	17.343	-
e	0.27	0.70	0.15	0.55	0.15	0.18	9.151	16.125
f	0.12	0.45	0.06	0.39	0.49	0.49	-7.231	-4.000
g	0.04	0.24	0.06	0.39	0.70	0.57	-39.996	-26.125

In the rows **e-g**, where the reject option is possible, the rates are those given by t_{1opt} and t_{2opt}. It is also possible to observe how the value of the utility obtained with the reject option (last column) is sensibly better than in the case of classification without reject option (seventh column).

References

1. Chow, C.K.: An Optimum Character Recognition System Using Decision Functions. IRE Trans. Electronic Computers EC-6 (1957) 247-254
2. Chow, C.K.: On Optimum Recognition Error and Reject Tradeoff. IEEE Trans. Inf. Th. IT-10 (1970) 41-46
3. Dubuisson, B., Masson, M.: A Statistical Decision Rule with Incomplete Knowledge about Classes. Pattern Recognition 26 (1993) 155-165
4. Muzzolini, R., Yang, Y.-H., Pierson, R.: Classifier Design with Incomplete Knowledge. Pattern Recognition 31 (1998) 345-369
5. Cordella, L.P., De Stefano, C., Tortorella, F., Vento, M.: A Method for Improving Classification Reliability of Multilayer Perceptrons. IEEE Trans. Neur. Net. 6 (1995) 1140-1147
6. Bradley, A.P.: The use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition 30 (1997) 1145-1159
7. Provost, F., Fawcett, T.: Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD-97)
8. Blake, C., Keogh, E., Merz, C.J.: UCI Repository of machine learning databases, [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1998
9. Hoekstra, A., Kraaijved, M.A., de Ridder, D., Schmidt, W.F., Ypma, A.: The complete SPRLIB & ANNLIB. Statistical Pattern recognition and Artificial Neural Network Library. 2nd edn. Version 3.1. User's Guide and Reference Manual, Pattern Recognition Group, Faculty of Applied Physics, Delft University of Technology, Delft, The Netherlands (1998)