Measures and Characteristics of Classification Quality

Jiřina, Marcel
2005

cAk

CENTRUM APLIKOVANÉ KYBERNETIKY

České vysoké učení technické v Praze - fakulta elektrotechnická

Measures and Characteristics of Classification Quality

Technical report

Marcel Jiřina and Marcel Jiřina, jr.

www@c-a-k.cz

2005

Institute of Computer Science
Academy of Sciences of the Czech Republic

Measures and Characteristics of Classification Quality

Marcel Jiřina and Marcel Jiřina, jr

Technical Report No. V-936

May 2005

Abstract

Basic definitions and approaches to evaluation of performance of different types of classifiers are summarized for two-class problems and then two-classes classifiers. Local measures of classification quality are presented and the parametric dependence of these measures is discussed and mutual dependencies like ROC curve are shown. Local measures and functional dependencies do not characterize the behavior of a classifier as whole or in some broader region of possible applications. Thus some global measures and combined measures are presented.

Keywords:

Multivariate data, classification, classification error, ROC curve

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: +420 266 051 111, fax: +420 286 585 789

e-mail: ics@cs.cas.cz

Measures and Characteristics of Classification Quality

Marcel Jiřina
Institute of Computer Science, Pod vodarenskou vezi 2,
182 07 Prague 8 – Liben, Czech Republic

Marcel Jiřina, jr.
Faculty of Biomedical Engineering, Czech Technical University in Prague, Zikova 4,
166 36, Prague 6, Czech Republic

Contents

1 Introduction

We summarize the basic definitions and approaches to evaluation of performance of different types of classifiers. Basically, we consider the two-class problems and then the two-class classifiers will be discussed. It can be easily seen that there are two basic variables which describe the performance of the classifier. The one class (signal) efficiency and the other class (background) error. From these values all other parameters and characteristics can be derived. The two basic variables mentioned are usually dependent on one free parameter – the threshold – and thus all other values are, in fact, functions of this free parameter. In this case the other (internal) parameters which control the behaviour of classifier are considered as intrinsic part of the classifier and are not considered in this paper.

The work is organized as follows. First the basic data sets and the basic variables for evaluation of error are introduced. Then the local measures of classification quality are presented. After it, the parametric dependence of these measures is discussed and mutual dependencies like the ROC curve are shown. Local measures and functional dependencies do not characterize the behaviour of the classifier as whole or in some broader region of possible applications. Thus some global measures and combined measures are presented.

2 Basic data and basic error describing variables

In the two-class classification problem the task is to assign to a given samples a class to which it most probably belongs. The class we are looking for primarily is often denoted by number 1 and is called "signal". The other class is often denoted by number 0 (sometimes -1) and is called "background" or noise.

The classifier can be considered as a filter or a sieve which separates smaller particles (signal samples) from larger ones (background samples) as illustrated in Fig. 1.

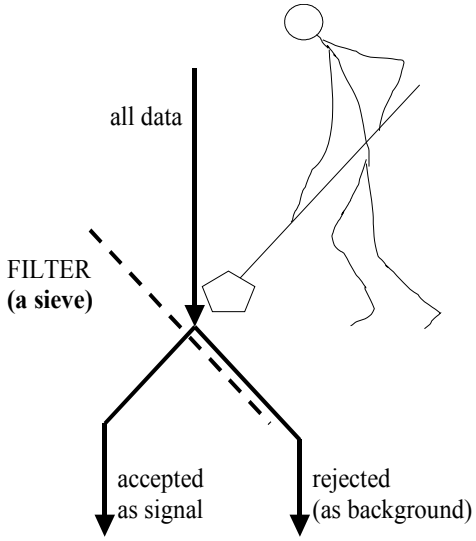

Fig. 1. Illustration of a two-class classifier as a sieve.

Let N samples be given, S samples of class 1, B samples of class 0. A classifier should separate from N samples, S samples as class 1 and B samples as class 0. This is an ideal case. In reality we get
S_1 samples of class 1 correctly recognized as class 1
S_0 samples of class 1 erroneously recognized as class 0
B_0 samples of class 0 correctly recognized as class 0
B_1 samples of class 0 erroneously recognized as class 1.

Of course,

$$S_1 + S_0 = S \text{ and } B_0 + B_1 = B .$$

As a signal we get total $S_1 + B_1$ samples, as a background we get total $S_0 + B_0$ samples.

3 Types of classifiers

From the point of view of the output values, there are two kinds of classifiers. Some classifiers give just the class to which a particular sample should belong, others give a real number from some interval. In the second case, in order to classify, it is necessary to choose a fixed value θ, the threshold. Then, if the output value is equal to or larger than θ, the sample belongs to one class, usually the signal, otherwise it belongs to the other class, usually the background. Classification features depend on this threshold. All parameters above are then functions of the threshold.

Most classifiers have some other (intrinsic) parameters which must be properly tuned to get the best results for particular data. In some cases, part of these parameters are set up manually, some may be tuned during the learning process. In contrast to threshold θ, these parameters are fixed during testing or evaluation. The threshold is the only free variable.

4 Local measures for classification quality

Local measures depend on selection of particular threshold for separating classes.

4.1 Signal efficiency or signal acceptance

The ratio $S_e = S_1 / S$ gives the percentage of correctly recognized samples of class 1 and is often called signal efficiency or signal acceptance. In fact, it is *Pr*(accepted|signal) [3]. Sometimes this is called "sensitivity" [4] as it gives how sensitive the classifier is to the signal.
The signal error is given by $1 - S_e$.

4.2 Background error or background acceptance

The ratio $B_e = B_1 / B$ gives the percentage of erroneously recognized samples of class 0. It is often called background error or background acceptance, as it is in fact, *Pr*(accepted|background) [3]. The other term used is purity, which equals to $1 - B_e$. It says how "pure" is the output mixture considered as signal.

From these two basic variables other measures are derived.

4.3 Classification error (error rate)

Classification error gives a percentage of all errors of the classifier. The total number of erroneously classified samples is $S_0 + B_1$. The classification error is then $C_e = (S_0 + B_1)/N$.

The classification error or the weighted classification error (see the next paragraph) is used in [2] for comparison of classifiers. Sometimes this value is very sensitive to parametrization by threshold θ as illustrated in Figs. 2 and 3 .

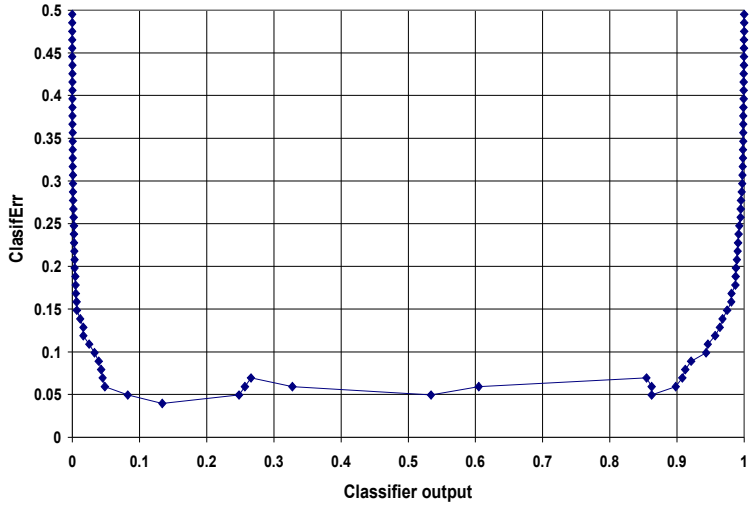

Fig.2. Dependence of the classification error on the threshold for IRIS data [2] and SFSloc7a [6] classifier. The minimal error is 0.0396 for threshold 0.134.

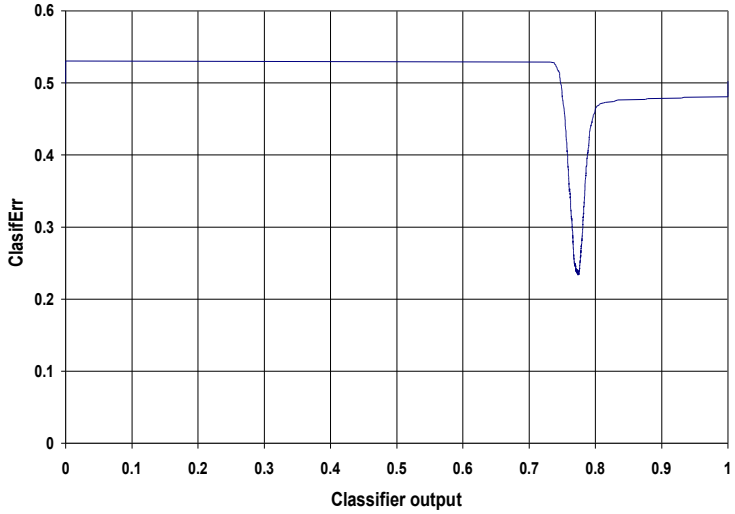

Fig. 3. Dependence of the classification error on the threshold for SPLICE data [2] and SFSloc7a classifier. The minimal error is 0.233 for threshold 0.77379. One can see that the minimum is very sharp here.

4.4 Weighted classification error

It is usual in financial and medical tasks that a false response is much more weighted than success because a fault could have serious or fatal consequences. For example, if background means that the patient is ill and signal means that the patient is healthy, we must maximize the signal efficiency, i.e. minimize the signal error, i.e. cases when the patient is considered healthy and he or she is, in fact, ill. In such a case the weight of the signal is much larger than the weight of the background. Let the weight of signal be W_s, say $W_s = 5$, and, at the same time, the weight of background $W_b = 1$. The classification error is then given by

$$C_e = (W_s.S_0 + W_b.B_1) / N .$$

Note that data in [2] are in some cases evaluated in this way.

Sometimes no weight is assigned to the background error and the classification error is evaluated (minimized) according to formula

$$C'_e = S_0 / N .$$

4.5 Enrichment factor (signal enhancement)

The enrichment factor or the signal enhancement, sometimes also signal-to-noise ratio enhancement are defined as

$$E = S_e / B_e .$$

Initially we have S signal samples in the mixture of N samples. The signal to noise ratio of the mixture is $S_c = S / B$. After the classifier, the filter is used, we have S_1 samples in the mixture of $S_1 + B_1$ samples recognized as a signal. The signal to noise ratio in the mixture after classification is $S_{cc} = S_1 / B_1$. The ratio S_{cc} / S_c is then

$$S_{cc} / S_c = E S_c \quad .$$

The signal to noise ratio in the mixture after the classifier is enlarged – enhanced - by factor E compared to the signal to noise ratio in the mixture before the classifier.

4.6 Quality factor

Important case in classification/filtering is the case of so called rare events. In this case there is lot of background (noise) events-samples and small number of signal samples so that $S << B$. Useful characteristic of the mixture of signal and background samples is ratio $q = S/\sqrt{B}$, where S is a number of signal samples, and B is a number of background samples. It is often denoted as the sensitivity or the quality factor (of data).
Let total N measurements be given. These N measurements may be repeated k times. The random variable we are interested in is the number of signal samples X. The mean value of X let be S, individual sets of data will have S_i, $i = 1, 2,.. k$ signal samples. The width of confidence interval for variable X is

$$K_0 = 2t(\alpha, N-1)\frac{\sigma}{\sqrt{N}}$$

and estimation of sigma is given by

$$\sigma^2 = \frac{1}{k}\sum_{i=1}^{k}(S_i - S)^2 = S^2\frac{1}{k}\sum_{i=1}^{k}\left(\frac{S_i}{S} - 1\right)^2 = S^2\sigma_x^2 \ ,$$

where σ_x^2 is the relative variance of the number of signal samples and it is supposed constant. We get

$$K_0 = \frac{S}{\sqrt{S+B}}2t(\alpha, N-1)\sigma_x \ .$$

The ratio $q = \frac{S}{\sqrt{S+B}}$, or $q = \frac{S}{\sqrt{B}}$ for $S << B$, is called the sensitivity or quality factor (of data).
Data after the classification/filtering is used should have this ratio at least as large as original data because we wish to get as much signal samples as possible. It is necessary to have data after classification/filtering process statistically at least as good as before.
The quality factor (of the classification/filtering process, not of data) is defined as

$$Q = S_e/\sqrt{B_e} \qquad \text{(Q)}$$

The Q larger than or equal to 1 says that the statistical quality q of the mixture after the classifier/filter is not worse than the statistical quality of the original mixture itself, namely for $S << B$ it holds that $q_{(\text{after filter})} = Q.q_{(\text{before filter})}$, hence Q can enhance ratio $S/\sqrt{B}$ of the data set.

For classifiers which give a real number from some interval as the output the quality factor depends on the threshold.

4.7 Rejection factor

The rejection factor should be called a background rejection factor as is given by $R = 1/B_e$. This value shows how many times the background is suppressed – rejected without respect to any signal efficiency. It has its sense in cases when a massive part of the background should be removed under the assumption that an essential part of signal samples remains retained.

5 Parameter dependencies

As has been said earlier, in classifiers which give a real number from some interval as the output the parameters mentioned can depend on a free parameter, threshold θ. Thus in such a case all parameters are functions of threshold θ. Using this parametrization, one can find several interesting and illustrative dependences.

5.1 ROC curve and related measures

We can have the ROC curve as a very general way of depicting a classifier's behaviour with respect to a particular problem. On the vertical axis there is signal efficiency (signal acceptance), and on the horizontal axis the background error (background acceptance). The threshold is a parameter of the curve shown for example in Fig. 4.

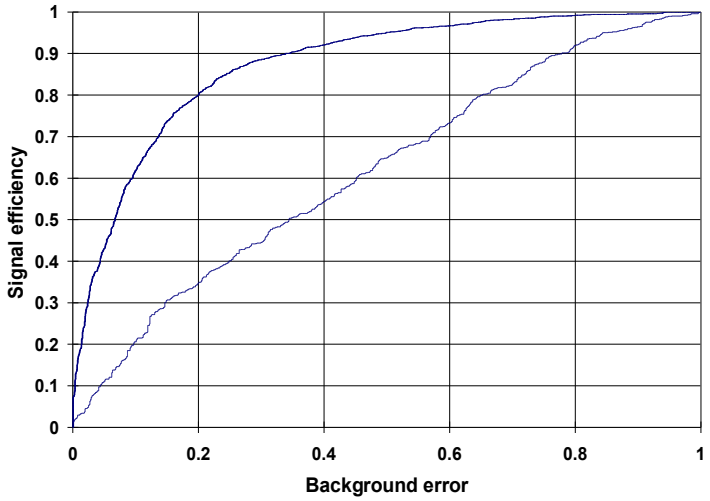

Fig. 4. Example of ROC curve - "good" data, i.e. relatively easily separable (upper line) and "bad" data, i.e. difficult to separate (bottom line).

Receiver Operator Characteristic (ROC) curves [5] were developed in the 1950's as a by-product of research into making sense of radio signals contaminated by noise [4]. This diagram is known also under the name Neyman–Pearson diagram or decision quality diagram. In statistical terms, the ROC curve shows the probability of a false alarm on the x-axis and the probability of detection on the y-axis. The assumption is that samples of events or probability density functions are available both for signal (authentic) and background (imposter) events; a suitable test statistic is then sought which optimally distinguishes between the two. Using a given test statistic (or discriminant function), one can introduce a cut which separates acceptance region (dominated by the signal events) from a rejection region (dominated by the background). The Neyman-Pearson diagram plots contamination (misclassified background events, i.e. classified as signals) against losses (misclassified signal events, i.e. classified as the background), both as fractions of the total number of samples of the corresponding class.
An ideal test statistic causes the curve to pass close to the point where both losses and contamination are zero, i.e. the acceptance is one for signals, and zero for the background. Different decision strategies choose a point of the closest approach,

where a “liberal“ strategy favours the minimal loss (i.e. high acceptance of signals), a “conservative“ strategy favours the minimal contamination (i.e. high purity of the signal). For a given test (fixed cut parameter), the relative fraction of losses (i.e. the probability of rejecting good events, which is the complement of acceptance), is also called the significance or the cost of the test; the relative fraction of contamination (i.e. the probability of accepting background events) is denominated by the power or purity of the test [3].

5.2 Classifier error

Suppose that there are non-equal numbers of signal and background samples so that $N = S + B$, where B is the number of samples of class 0 (background), and S is the number of samples of class 1 (signal). Moreover let errors be weighted. Then it is possible to derive that

$$C_e = \frac{SW_S(1-S_e) + BW_bB_e}{N} \quad .$$

From this equation it follows that

$$S_e = 1 - \frac{N}{SW_S}C_e + \frac{BW_b}{SW_S}B_e \quad .$$

This is a linear equation with respect to variable B_e. In the most frequent case there is $S = B$ and $W_s = W_b = 1$, and then

$$S_e = 1 - 2C_e + B_e \quad .$$

The set of straight lines for different constant values of the classification error is shown in Fig. 5.

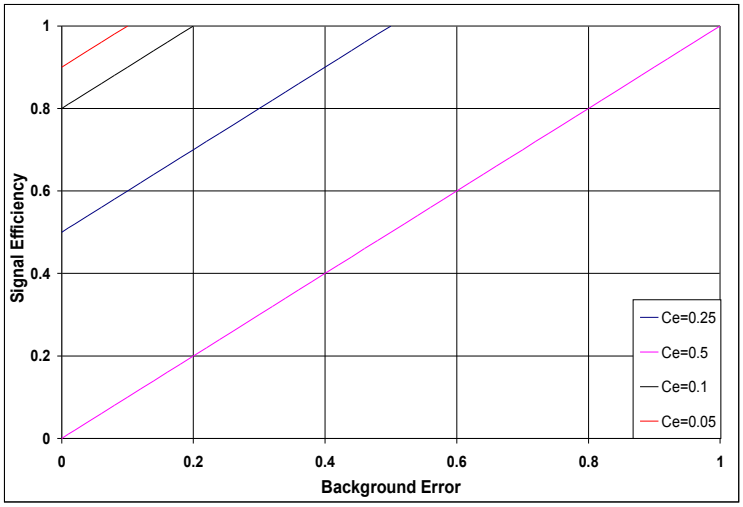

Fig. 5. Lines of the constant classification error in the ROC diagram.

5.3 Enrichment factor (signal enhancement)

Direct dependence of the enrichment factor on the threshold θ need not be clear enough. More useful is the dependence of the enrichment factor on the background error. This dependence can be compared with ROC curve. An example is shown in Fig.6.

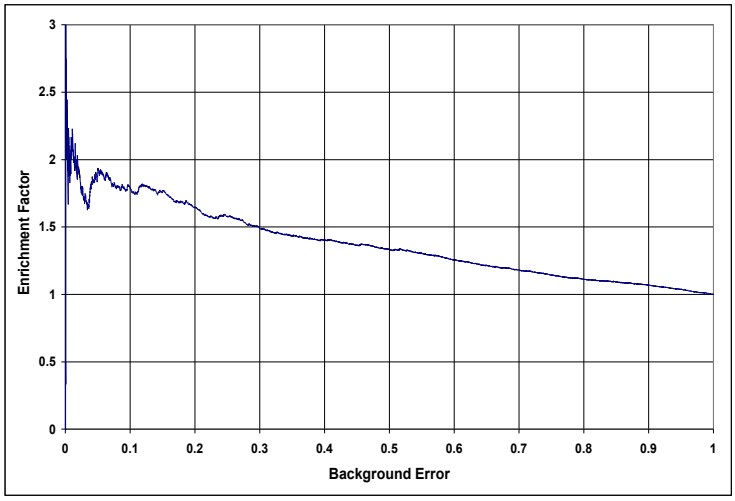

Fig. 6. Example of dependence of the enrichment factor on the background error for "bad" data. This figure is comparable to the bottom ROC curve in Fig. 4.

5.4 Quality factor

The dependence of the quality factor on the background error can be compared with the ROC curve. As the quality factor can be less than 1, there are important intervals, where $Q \geq 1$ and data after classification are statistically at least as good as before. Sometimes it need not be essential. An example is shown in Fig.7.

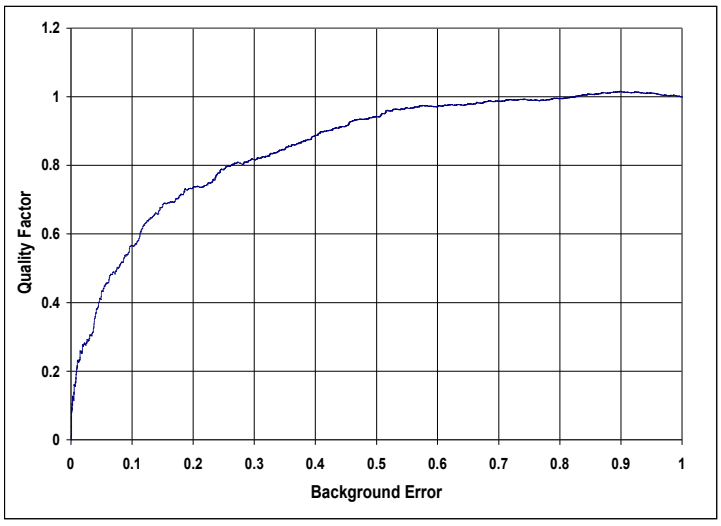

Fig. 7. Example of dependence of the quality factor on the background error for "bad" data. This figure is comparable to the ROC curve for "bad" data in Fig. 4. One can see that $Q \geq 1$ for the background error larger than 0.7147. In this interval, the statistical

validity remains retained. From the dependence of the enhancement factor on the background error it follows that relatively low values of the enrichment factor can be reached, $E_{\max.} = 1.182$ for $B_e = 0.7147$ and $\theta = 0.498$. Note that this example is a truly bad case but serves well for illustration.

5.5 Acceptable region

The quality factor as well as the enrichment factor define the region of the ROC curve where the ROC curve should lie. In cases when some statistical data processing follows we need $Q \geq 1$. Similarly, for an enrichment factor smaller than 1 filtering or classification has no sense. These areas are shown in Figs. 8 and 9 for "good" and "bad" data (or filtering) respectively.

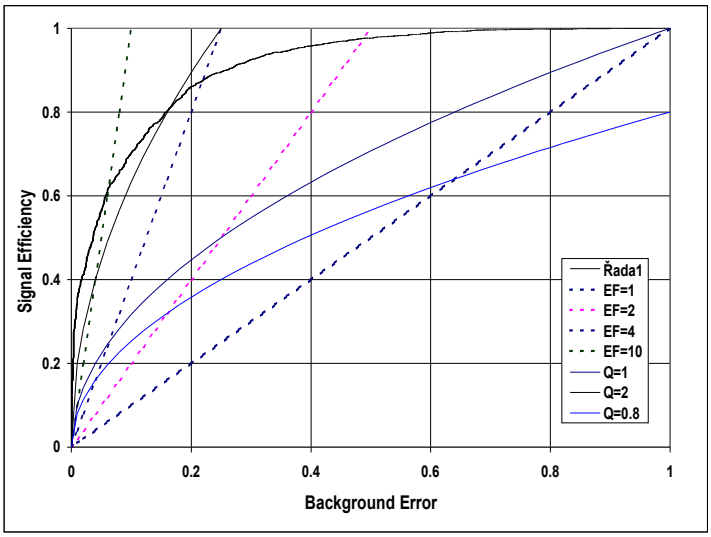

Fig. 8. The ROC curve for "good" data with parabolic lines of constant quality factor and straight lines of constant enrichment factor. We can see that the whole ROC curve lies above the line for $Q = 1$ and that there is no problem to reach an enrichment factor larger than 10.

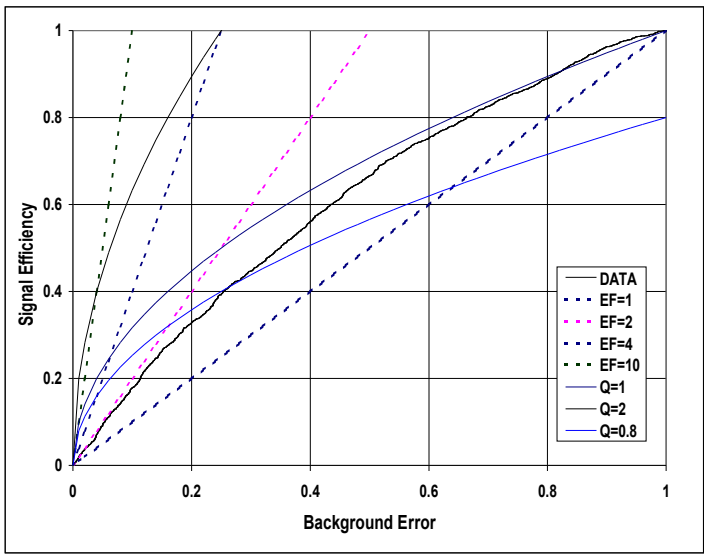

Fig. 9. ROC curve for "bad" data (or a bad classifier!) with lines of the constant quality factor and the constant enrichment factor. It is seen that nearly the whole ROC the bold curve lies below the parabolic curve for Q = 1 and that there is a problem to reach the enrichment factor a little bit larger than 1. If it is not necessary to have a quality factor larger than or equal to 1, it is possible to get enrichment factor 2 – in this case the best value of the quality factor is Q = 0.8 .

6 Global measures

By global measures we try to evaluate the whole ROC curve or the whole range of the classifier and get a single measure for classification quality. Possibly, a part of the ROC curve can be used.
The classifier usually gives some response to each sample. Let this response be real number between 0 and 1 (it may include 0 and 1 as well).

6.1 Global measure with sorting

Let the samples of the testing set be sorted according to the response of the classifier. Let each sample x_i be assigned to its order number j_i and let variable z_i be 1 for the signal and 0 for the background.
Let us define measure

$$A = \sum_{i=1}^{N} j_i z_i$$

where N is the number of all samples. The larger the A the better. The maximal value of A is

$$A_{\max} = \sum_{i=N_B+1}^{N} i \quad .$$

It can be advantageous to relate results to maximal value using $a = A/A_{\max}$.
An ideal case is $a = 1$. For "good" data illustrated in Fig. 2 $a = 0.9163$, for bad data in the example above $a = 0.7456$. The random classification gives value 0.5. Thus, the larger the a, the better the classification.

6.2 Area below ROC curve

The better classification, the larger is the area below the ROC curve. Simply, this area can serve as global measure of the classification quality. Naturally, the best possible classification gives value 1, and the random classification gives value 0.5.

It is also possible to use a part of the area below ROC curve, e.g. the left half of the diagram from the background error 0 to the background error 0.5 and similarly.

7 Combined measures

Sometimes it may be useful to use several values of some local measure for several values of the threshold or of the signal efficiency or of the background error. Several values thus define the region of interest, i.e. region of application thought about. One can use e.g. mean value, a simple sum or a weighted sum. Such an approach has been used in comparison of classifiers in [1].

8 Conclusions

Evaluation of a classification task depends on the problem solved. According to it, local measure, global measures or functional dependences are used. We have shown that especially a minimal classification error may be very sensitive to a particular value of the threshold in classifiers which give a real number from some interval as the output. It is possible to deduce that in such a case results will depend on the particular selection of testing and also of a learning set.

Acknowledgement

This work was supported by the Ministry of Education of the Czech Republic under projects No. 1M684077004 Center of Applied Cybernetics, and No. MSM6840770012 Transdisciplinary Research in the Field of Biomedical Engineering II .

References

[1] Bock,R.K. et al.: Methods for multidimensional event classification: a case study using images from Cherenkov gamma-ray telescope. Nuclear Instruments and Methods in Physics Research A 516 (2004), pp. 511-528.
[2] UCI Machine Learning Repository.
http://www.ics.uci.edu/~mlearn/MLSummary.html.
[3] Bock,R.K.: Data Analysis BriefBook, 1998
http://rkb.home.cern.ch/rkb/AN16pp/node185.html .
[4] Van Schalkwyk, Jo: The magnificent ROC (Receiver Operating Characteristic curve). http://www.anaesthetist.com/mnm/stats/roc/ .
[5] Masters. T.: Practical Neural Network Recipes in c++. Morgan-Kaufman, San Diego, USA, 1993, Chap. 19.
[6] Jiřina, M.: Distribution Mapping Exponent for Multivariate Data Classification. Proc. of the Eight World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2004), Orlando, Florida(USA), July 18-21, 2004, Vol. V., pp.103-108, CD ROM ISBN 980-6560-14-0.