ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume LVIII 36 Number 6, 2010

THE EVALUATION OF BINARY CLASSIFICATION TASKS IN ECONOMICAL PREDICTION

M. Pokorný

Received: September 6, 2010

Abstract

POKORNÝ, M.: *The evaluation of binary classification tasks in economical prediction.* Acta univ. agric. et silvic. Mendel. Brun., 2010, LVIII, No. 6, pp. 369–378

In the area of economical classification tasks, the accuracy maximization is often used to evaluate classifier performance. Accuracy maximization (or error rate minimization) suffers from the assumption of equal false positive and false negative error costs. Furthermore, accuracy is not able to express true classifier performance under skewed class distribution. Due to these limitations, the use of accuracy on real tasks is questionable. In a real binary classification task, the difference between the costs of false positive and false negative error is usually critical. To overcome this issue, the Receiver Operating Characteristic (ROC) method in relation to decision-analytic principles can be used. One essential advantage of this method is the possibility of classifier performance visualization by means of a ROC graph. This paper presents concrete examples of binary classification, where the inadequacy of accuracy as the evaluation metric is shown, and on the same examples the ROC method is applied. From the set of possible classification models, the probabilistic classifier with continuous output is under consideration. Mainly two questions are solved. Firstly, the selection of the best classifier from a set of possible classifiers. For example, accuracy metric rates two classifiers almost equivalently (87.7 % and 89.3 %), whereas decision analysis (via costs minimization) or ROC analysis reveal different performance according to target conditions of unequal error costs of false positives and false negatives. Secondly, the setting of an optimal decision threshold at classifier's output. For example, accuracy maximization finds the optimal threshold at classifier's output in value of 0.597, but the optimal threshold respecting higher costs of false negatives is discovered by costs minimization or ROC analysis in a value substantially lower (0.477).

binary classification, bankruptcy prediction, classifier performance evaluation, accuracy maximization, receiver operating characteristic (ROC)

In the area of economical research, much attention has been paid to development and improvement of many prediction methods and models so far. One of the typical tasks being solved is the bankruptcy and financial distress prediction and related binary classification task. Surprisingly, not so many studies pay attention to a more sophisticated classifier performance evaluation. Despite the fact the evaluation methodology critically affects the optimal classifier selection and its use, not so sufficient accuracy maximization (or error rate minimization) has been often used. Other non-financial literature criticizes accuracy maximization procedure as well. Detailed findings can be found in Pokorny (2009) within the present state analysis of this problem, where other authors are quoted.

In the area of medical research, the Receiver Operating Characteristic (ROC) method has been widely used, which has the potential to solve the lack of a sophisticated evaluation methodology, including consideration of different type I/II error costs. Moreover, it has the ability to visualize classifier's performance. Although the ROC is not completely unknown to the financial research, its application is rare.

The objective of this paper is to emphasize shortcomings of accuracy as an evaluation metric on concrete examples of the binary cost-sensitive classification task, and to show the use of ROC analysis as an alternative evaluation method. This paper is based on results of F. Provost's and T. Fawcett's research and tries to popularize this topic. Results of this study can be used as a guide for the economical research mentioned above.

MATERIAL AND METHODS

Binary classification

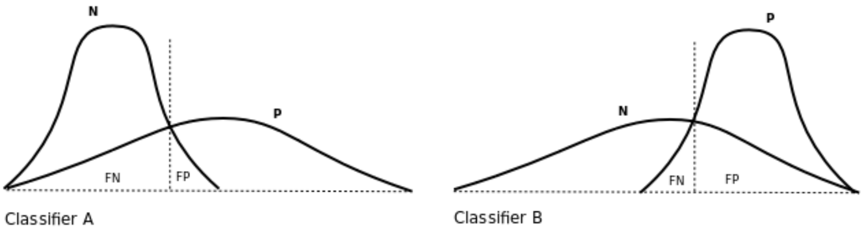
The *binary classification task* is confined to two class separation. The classifier's outcome can be treated as positive (P) or negative (N), and according to the true status of an instance being classified, the classification result can be true positive (TP), true negative (TN), false positive (FP, Type I Error) or false negative (FN, Type II Error). In case of bankruptcy prediction, a company could be rated as healthy (negative) or failing (positive). Many classification models applicable on this task exist, solid theoretical background can be found e.g. in Bishop (2006). This study focuses primarily on a so called probabilistic classifier with continuous output. The classifier's output is in the form of a probability or score – numeric value that represents the degree to which an instance is a member of a class. (Fawcett, 2004) The same author then follows: "These values can be strict probabilities, in which case they adhere to standard theorems of probability; or they can be general, uncalibrated scores, in which case the only property that holds is that a higher score indicates a higher probability.".

Having a test set[1] of negative and positive samples (with true status of class membership), and a set of potential classifiers, the goal is to choose the best classifier according to target conditions of application. Each of the classifiers is populated with the test set, and at its output produces the outcome usually in the form of a Gaussian distribution for either of the classes. Varying the decision threshold at the classifier output to a suitable position, the final form of the classifier is obtained – a discrete classifier (Fawcett, 2004), and so the second goal is to set the optimal decision threshold. Samples rated above this threshold are classified as positive, samples below are classified as negative. One example of a typical classifier of this type is the neural network having sigmoidal activation function on its output neuron. The fixed threshold of 0.5 would be an misleading approach because of relative score separation objective, as describes Fawcett's example (2004, p. 8).

The accuracy as an evaluation metric

Accuracy (error rate) is given by the ratio of correctly (incorrectly) classified instances to all instances in the test set, i.e. (TP + TN)/(P + N). This type of metric is criticized by authors for its inability to differentiate between type I and II error costs (both are assumed to be equal), but they do differ in most practical tasks. Furthermore, accuracy doesn't reflect true classifier's performance under skewed class distribution. In reality, classifiers often face to a grater number of negative instances compared to positive instances. (Obuchowski, 2003; Fawcett, 2004; Provost and Fawcett, 2001, 1997; Provost, Fawcett and Kohavi, 1998) Accuracy evaluates classifier's performance with one number for both of the classes and for one setting of target conditions.

Another shortcoming of the accuracy metric is the indistinguishable performance evaluation of two different scenarios, as depicted on Fig. 1 – example similar to Erkel and Pattynama (1998). The accuracy is the same for both of the situations, but for example, classifying almost all positives in the first one encompasses wrong classification of almost all negatives. On the other hand, classifying almost all positives within the second scenario encompasses approximately only half of false positives.



1: *Example of two distributions of the classifier's output class membership score*

ROC analysis

ROC analysis is a method frequently used in medical research. Essential ROC metrics are the sensitivity and specificity. Sensitivity (or true positive rate, TPR) is the proportion of correctly classified positives (TP) among all positives (P), i.e. TPR = TP/P.

[1] Not only one single test set should be used, for the purpose of variance measurement, several test sets should be used. (Fawcett, 2004). The same author then presents the so called ROC curves averaging.

Specificity (or true negative rate, TNR) is the proportion of correctly classified negatives (TN) among all negatives (N), i.e. TNR = TN/N. The opposite of specificity is the false positive rate (FPR = FP/N), where FP is the number of negatives incorrectly classified as positive. Similarly, false negative rate FNR = FN/P, where FN is the number of positives incorrectly classified as negative.

Varying the decision threshold at classifier's continuous output, the number of TP, FN, TN and FP changes, so does the sensitivity and specificity, but both in opposite direction. Higher sensitivity results in lower specificity and vice versa. Different thresholds constitute set of points [FPR, TPR] resulting in a *ROC Curve* and so called *ROC graph*. ROC curve generation algorithm can be found for example in Fawcett (2004). The X-axis in a ROC graph represents false positive rate, the Y-axis represents sensitivity. Discrete classifier (i.e. with the output of P or N only) is represented by one point in the graph. The more north-west the point lies, the better solution has been found. Ideal point is [0, 1], which means zero false positive rate and 100% sensitivity. The diagonal line represents the so called line of chance, in other words, no information is carried by the classifier, and real classifier should always be above this line. Classifiers laying under the line can easily be reversed to the upper left space by reverting their decisions. (Fawcett, 2004; Obuchowski, 2003; Provost and Fawcett, 2001, 1997; Provost, Fawcett and Kohavi, 1998; Erkel and Pattynama, 1998)

"ROC graphs illustrate the behavior of a classifier without regard to class distribution or error costs, and so they decouple classification performance from these factors". (Provost and Fawcett, 2001, 1997) Similarly, Obuchowski (2003) mentions ROC key characteristics. By treating both of the errors (FP, FN) separately, this method makes it possible to prioritize one type of error over another. Moreover, the ability to visualize the classifier's performance facilitates inherent analysis.

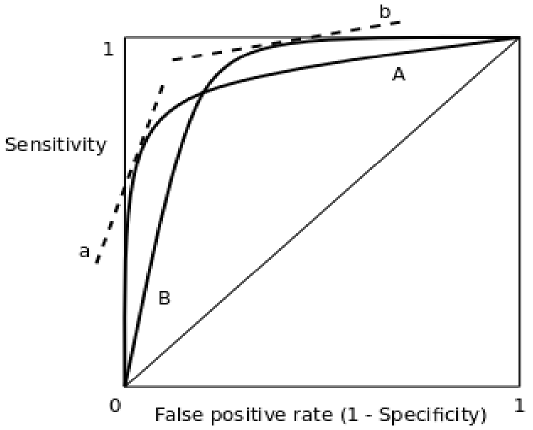
For the purpose of classifier performance comparison, it is usually easier to compare a single number then two values of sensitivity and specificity. The *Area Under the ROC Curve (AUC)* is an example of such metric and measures classifier discriminative power across all possible thresholds (or target conditions). Thereby AUC eliminates the influence of the decision threshold value on sensitivity and specificity. (Erkel and Pattynama, 1998). Perfect classifier has the AUC 1.0 (100 %), the area under the line of chance equals to 0.5 (50 %). For other interpretations of the AUC, see Obuchowski (2003, p. 5).

One major drawback associated with the classifier comparison based on the AUC is that usually only part of the curve is practically relevant. (Obuchowski, 2003; Erkel and Pattynama, 1998) For example, if we assume target conditions of high FN costs and relatively high occurrence of positive instances, the upper right part of the ROC graph is relevant. The target conditions can be visualized as a line in a ROC graph – according to Provost and Fawcett (2001, 1997), this line is called *iso-performance line* with a slope *s* given by the formula below, and all classifiers corresponding to points on the line have the same expected costs. Using the example above, the iso-performance line would have a small slope and as a tangent to a ROC curve would be located in the upper right part.

$$s = \frac{TP_2 - TP_1}{FP_2 - FP_1} = \frac{p(\mathrm{n})\mathrm{C(FP)}}{p(\mathrm{p})\mathrm{C(FN)}},$$

where *p*(p) is the prior probability of a positive example, *p*(n) = 1 – *p*(p) is the prior probability of a negative example, C(FP) and C(FN) are the costs of false positive and false negative errors.

Obuchowski (2003) states that "whenever the ROC curves of two tests cross (regardless of whether or not their areas are equal), it means that the test with superior accuracy (ie, higher sensitivity) depends on the FPR range; a global measure of accuracy, such as the ROC curve area, is not helpful here." Only if one model dominates in the whole ROC space (all other curves are below the curve of this test), this model can be said as the best one. (Provost and Fawcett, 2001, 1997; Provost, Fawcett, Kohavi, 1998) Obuchowski (2003) then suggests several alternatives for the situation of crossing ROC curves: use of the ROC curve to estimate sensitivity at a fixed false positive rate (or false positive rate at a fixed sensitivity), or use of the partial area under the ROC curve (the area between two false positive rates, or the area between two false negative rates). Similar recommendations can be found in Erkel and Pattynama (1998). Obuchowski (2003, p. 7, Fig. 4) or Erkel and Pattynama (1998, p. 92, Fig. 4) depict this situation, similarly Fig. 2 below shows two ROC curves corresponding to two classifiers from Fig. 1 of this paper. Obviously, the second classifier would be more appropriate in case of high FN costs and



2: *Example of two iso-performance lines corresponding to two target conditions* high prevalence of positives, its ROC curve lies more north-west in the relevant part of the ROC graph.

To sum up, the ROC curve represents classifier's behavior in every possible situation, the iso-performance line represents specific target conditions, by combining them, the optimal decision threshold can be found.

By the way, Provost and Fawcett invented a complex method called *ROC Convex Hull*, which employs principles of the ROC analysis, decision analysis and computation geometry, and which is able to identify set of methods that are potentially optimal under any cost and class distribution. (See Provost and Fawcett, 2001, 1997; Fawcett, 2004 for details.)

Data

To demonstrate the use of ROC analysis for the cost-sensitive classifier evaluation, six data sets were generated. Each data set represents classifier's output score (as explained in the Material and Methods chapter) in a form of Gaussian distribution from interval 0-1 separately for negative and positive instances on a test set. Such output can be obtained for example from a neural network having the sigmoidal activation function at the output unit.

Test 1 data set (i.e. the output score of classifier 1 for a given test set of samples) contains 100 negative and 100 positive instances (mean ± std. deviation negatives/positives: 0.2 ± 0.1/0.8 ± 0.1) and represents an ideal classifier with no overlap between both of the classes. Test 2 data set consists of 200 instances for each class (mean ± std. deviation negatives/positives: 0.35 ± 0.1/0.65 ± 0.1), having slight overlap. More realistic are the test 3 (mean ± std. deviation negatives/positives: 0.4 ± 0.12/0.6 ± 0.12) and test 4 (mean ± std. deviation negatives/positives: 0.45 ± 0.12/0.55 ± 0.12) data sets, where negative samples (3200) outnumber positive samples (800). Test 4 has considerable overlap compared to test 3. Last two tests, test 5 (mean ± std. deviation negatives/positives: 0.5 ± 0.15/0.75 ± 0.07) and test 6 (mean ± std. deviation negatives/positives: 0.25 ± 0.07/0.5 ± 0.15) with balanced classes of 1000 samples per class demonstrate the effect of reversed distribution of positive and negative samples on binary classification.

Statistical characteristics are shown in Tab. I–III, histograms (number of positive and negative samples from the test set being classified with the classifier's output score in one of 20 groups from interval 0–1) are depicted in Fig. 3–5.

RESULTS AND DISCUSSION

Two questions have to be solved before applying a new classifier. Firstly, all available classifiers have

I: *Classifier 1 and 2 statistical characteristics of the class output score*

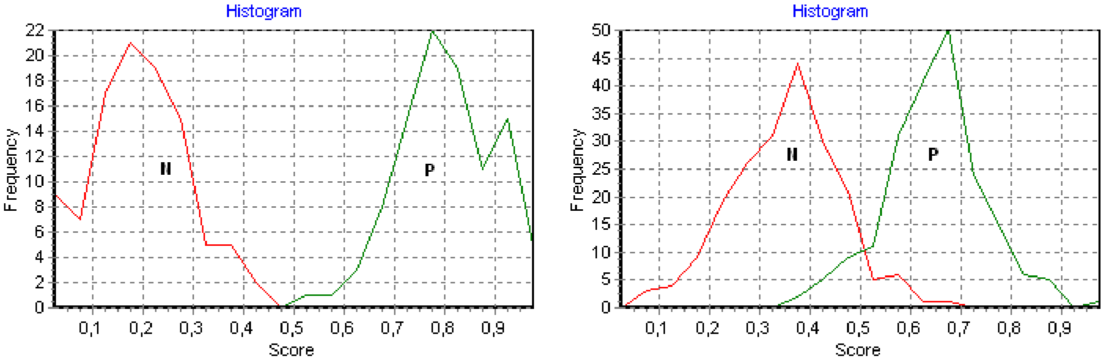
	Classifier 1		Classifier 2	
	Negatives n = 100	**Positives n = 100**	**Negatives n = 200**	**Positives n = 200**
Mean	0.192	0.802	0.351	0.647
Std. deviation	0.096	0.092	0.107	0.1
Median	0.189	0.8	0.36	0.652
Min	0.005	0.517	0.057	0.379
Max	0.431	0.970	0.656	0.978

II: *Classifier 3 and 4 statistical characteristics of the class output score*

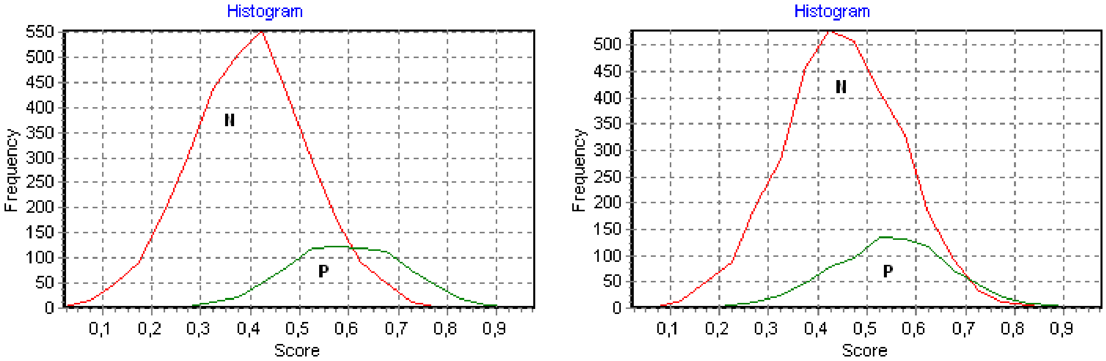
	Classifier 3		Classifier 4	
	Negatives n = 3 200	**Positives n = 800**	**Negatives n = 3 200**	**Positives n = 800**
Mean	0.399	0.594	0.449	0.55
Std. deviation	0.119	0.119	0.12	0.121
Median	0.401	0.594	0.448	0.55
Min	0.009	0.224	0.019	0.191
Max	0.773	0.943	0.874	0.895

III: *Classifier 5 and 6 statistical characteristics of the class output score*

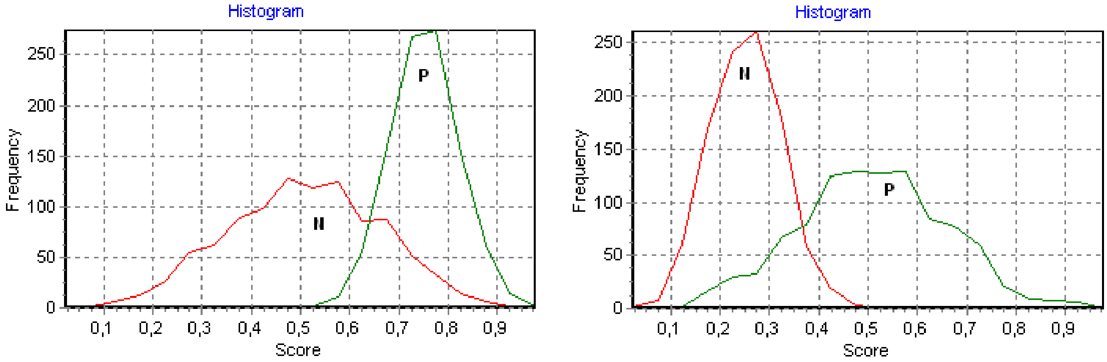
	Classifier 5		Classifier 6	
	Negatives n = 1 000	**Positives n = 1 000**	**Negatives n = 1 000**	**Positives n = 1 000**
Mean	0.508	0.752	0.253	0.509
Std. deviation	0.153	0.069	0.07	0.149
Median	0.511	0.752	0.253	0.508
Min	0.098	0.506	0.044	0.09
Max	0.959	0.995	0.487	0.936



3: *Histogram of the output score of classifier 1 (neg./pos.: 100/100; mean ± std. dev. 0.2 ± 0.1/0.8 ± 0.1), and classifier 2 (neg./pos.: 200/200; mean ± std. dev. 0.35 ± 0.1/0.65 ± 0.1)*



4: *Histogram of the output score of classifier 3 (neg./pos.: 3200/800; mean ± std. dev. 0.4 ± 0.12/0.6 ± 0.12), and classifier 4 neg./ pos.: 3200/800; mean ± std. dev. 0.45 ± 0.12/0.55 ± 0.12)*



5: *Histogram of the output score of classifier 5 (neg./pos.: 1000/1000; mean ± std. dev. 0.5 ± 0.15/0.75 ± 0.07), and classifier 6 (neg./ pos.: 1000/1000; mean ± std. dev. 0.25 ± 0.07/0.5 ± 0.15)*

to be compared according to their classification performance and the best one has to be chosen. Secondly, an optimal decision threshold (cut-off) must be set. Both tasks depend on target conditions, i.e. target error costs of FP and FN and target distribution of negatives and positives. Choosing the right metric for classifier performance evaluation is critical for both of the tasks.

Classifier 1 shows a classifier which is able to distinguish negatives from positives samples without any errors. In this situation, the setting of a decision threshold is straightforward and lies in the lowest positive instance 0.51735, or due to the classifier generalization, it could be more convenient to set the decision threshold to the middle of the interval between the highest negative and lowest positive instance. Any value lying above this threshold is classified as a positive instance, any value lying below this threshold is classified as a negative instance. No errors (FP, FN) are produced with this threshold. In this case, the accuracy (error minimization) is able to express true classifier performance and set the op-

IV: *Accuracy maximization vs. costs minimization, classifiers 1–4*

Classifier	Accuracy maximization (Error rate minimization)				Costs minimization		
	Accuracy (Error rate)	**Related threshold (s)**	**Related [FPR, TPR]**	**Related costs***	**Total costs***	**Related threshold**	**Related [FPR, TPR]**
Classifier 1	1 (0)	0.51735	[0, 1]	0	0	0.51735	[0, 1]
Classifier 2	0.9325 (0.0675)	0.51464; 0.48917	[0.055, 0.920]; [0.070, 0.935]	91; 79	57	0.47019	[0.110, 0.965]
Classifier 3	0.8565 (0.1435)	0.59754; 0.59679	[0.052, 0.489]; [0.053, 0.493]	2210; 2198	1412	0.47702	[0.251, 0.848]
Classifier 4	0.80975 (0.19025)	0.67699; 0.67682	[0.026, 0.153]; [0.026, 0.154]	3473; 3469	2329	0.47976	[0.390, 0.730]

* C(FP) = 1, C(FN) = 5

timal threshold, even if the target misclassification costs C(FP) and C(FN) were not the same.

Accuracy used for setting a decision threshold

Inappropriate use of accuracy for the purpose of setting an optimal decision threshold shows classifier 2. Accuracy maximization (or error rate minimization) sets the threshold in the intersection of the negative and positive class. In case of classifier 2, maximal accuracy of 93.25 % (minimal error rate of 6.75 %) is reached by setting the threshold to the value of 0.51464 or 0.48917. If we assume a situation of target costs e.g. C(FP) = 1 and C(FN) = 5, total misclassification costs would be 91 (for the first threshold) or 79 (for the second threshold). On the contrary, the costs minimization method sets the threshold to a lower value 0.47019, which produces lower total costs of 57. This clearly shows, that accuracy maximization (error rate minimization) doesn't set the optimal threshold in case of unequal error costs. The threshold set by this metric would be optimal in case of equal error costs.

Similar results can be shown on classifiers 3 and 4. The threshold set by accuracy maximization is too high compared to the threshold set be costs minimization. Costs minimization reflects higher costs of FN, thereby prefers correct classification of positives (higher true positive rate, TPR) at the expense of incorrect classification of many negatives (higher false positive rate, FPR). Tab. IV describes this phenomenon for classifiers 1–4.

One question may arise – why are there two optimal thresholds according to the accuracy maximization? Let's demonstrate this phenomenon on classifier 4 (similar effect can be found with classifiers 2, 3 and 6):

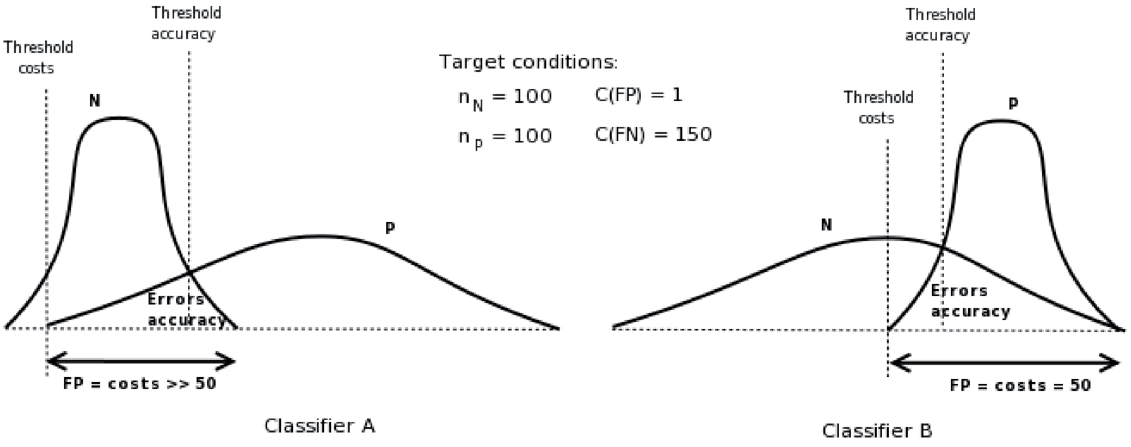
Score #	N/P	Score	FP	FN	Errors		
203	P	0.67791	83	680	763		
204	P	0.67734	83	679	762		
205	P	0.67699	83	678	761	⇒	min 1 (761/4000 = 0.19025)
206	N	0.67698	84	678	762		
207	P	0.67682	84	677	761	⇒	min 2 (the same error rate)
208	N	0.67593	85	677	762		
209	N	0.67570	86	677	763		

Instead of considering two optimal thresholds, it might be more correct to consider the average value of score lying between both of the minimums.

Accuracy used for classifier performance comparison

Inappropriate use of accuracy for the purpose of classifier performance comparison is demonstrated on two examplwes. Firstly, consider an example of two classifiers A and B whose output score distributions together with target conditions are shown on Fig. 6. Let's assume that this case is characteristic with high disproportion between FP and FN costs, e.g. C(FP) = 1 and C(FN) = 150, and the distribution is balanced N/P = 100/100 for simplicity. According to accuracy maximization, classifier A should be better than its counterpart because of its lower error rate. But then, if different error costs are taken into consideration, classifier B is obviously better than classifier A – see costs equal to FP in both situations. The reason why accuracy maximization doesn't work well in this case is that accuracy tacitly assumes equal error costs, i.e. C(FP) = C(FN).

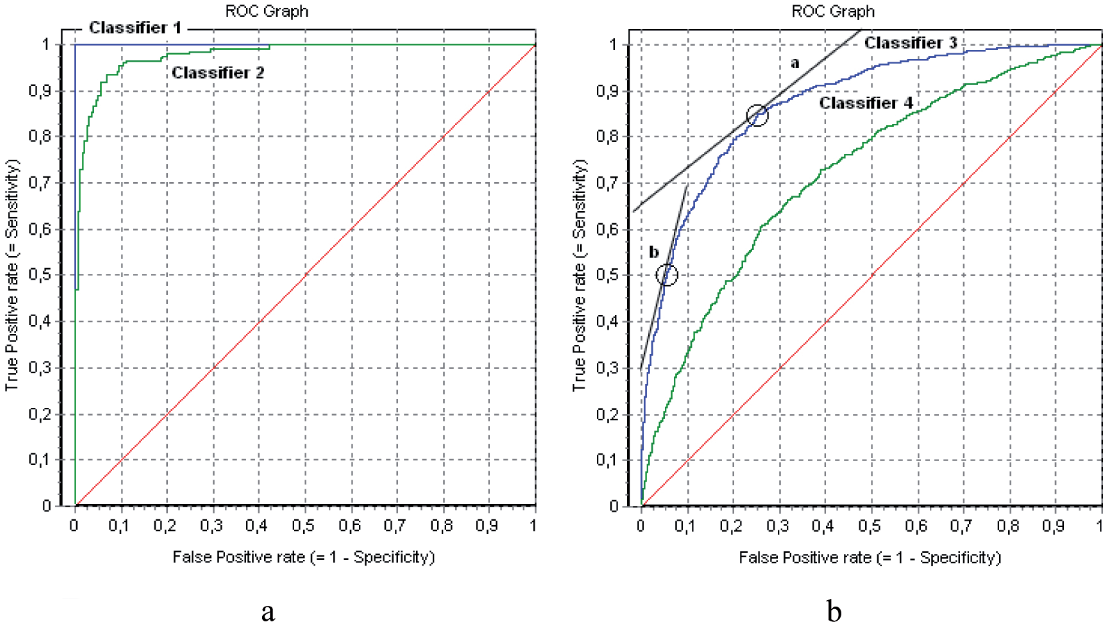
Similar behavior can be shown on classifiers 5 and 6. For the purpose of classifier comparison and selection, the accuracy is not usable here. According to the accuracy maximization, classifier 5 has accuracy of 87.7 % (in threshold 0.63527), classifier 6 has accuracy of 89.3 % (in threshold 0.37448 or 0.37381), and so both classifiers are rated almost equivalently. But if different error costs are taken into consideration, the situation is completely different. Having C(FP) = 1 and C(FN) = 5, classifier 5 has minimal costs of 339 (in threshold 0.6016), classifier 6 has minimal costs 641 (in threshold 0.31339) and is much worse classifier for this situation. Similarly, having C(FP) = 5 and C(FN) = 1, classifier 5 has mini-



6: *Inadequate use of accuracy for classifier performance comparison*

V: *Accuracy maximization vs. costs minimization, classifiers 5–6*

Accuracy maximization (Error rate minimization)			
Classifier	Accuracy (Error rate)	Related threshold (s)	Related [FPR, TPR]
Classifier 5	0.8765 (0.1235)	0.63527	[0.212, 0.965]
Classifier 6	0.8925 (0.1075)	0.37448; 0.37381	[0.043, 0.828]; [0.044, 0.829]
Costs minimization – C(FP) = 1, C(FN) = 5			
Classifier	Total costs	Related threshold	Related [FPR, TPR]
Classifier 5	339	0.60160	[0.274, 0.987]
Classifier 6	641	0.31339	[0.201, 0.912]
Costs minimization – C (FP) = 5, C(FN) = 1			
Classifier	Total costs	Related threshold	Related [FPR, TPR]
Classifier 5	720	0.72030	[0.081, 0.685]
Classifier 6	305	0.40996	[0.011, 0.750]



7: *ROC graphs for classifiers 1–4*

mal costs of 720 (in threshold 0.7203), classifier 6 has minimal costs of 305 (in threshold 0.40996) and is much better classifier for this situation. Tab. V contains further details (FPR, TPR).

ROC analysis in general

In addition to the methods discussed so far (accuracy maximization, costs minimization), the ROC analysis delivers visualization of classifier performance through a so called ROC graph. The ROC graph consists of a ROC curve that shows classifier performance in a form of [FPR, TPR] pairs across all possible decision thresholds.

On Fig. 7a, there are ROC curves of classifier 1 and classifier2. Classifier 1 curve is composed by three points [0, 0] – [0, 1] – [1, 1], and represents an ideal classifier with the area under the ROC curve (AUC) equal to 1 (100 %). Classifier 2 has lower curve with AUC 0.97837 (97.8 %). ROC graph for classifier 3 and 4 is shown on Fig. 7b. Classifier 3 curve is above classifier 4 curve in the whole ROC space, so we can conclude, that classifier 3 would be a better choice than classifier 4 in all possible decision thresholds and target conditions. Superiority of classifier 3 over classifier 4 can be expressed also with the AUC, classifier 3 has AUC of 87.5 %, classifier 4 AUC is 72.4 %.

ROC analysis and setting an optimal decision threshold

ROC analysis itself is not able to set an optimal decision threshold. But if it is combined with the iso-performance line, same results can be achieved as with the costs minimization, but with the luxury of visualization. The procedure is demonstrated on classifier 3 (Fig. 7b).

ROC curve of classifier 3 shows its behavior across all possible thresholds. Iso-performance line *a*, or its slope respectively, is given by target conditions, i.e. probability of a negative instance *p*(N) = 4/5 (3200/4000), probability of a positive instance *p*(P) = 1/5 (800/4000), and costs C(FP) = 1, C(FN) = 5. According to the formula in the theoretical section, the resulting slope equals to 4/5. The tangent of the iso-performance line *a* to the ROC curve of classifier 3 gives us the optimal decision threshold (emphasized with a circle). This point corresponds to the costs minimization result, i.e. the threshold 0.47702 with [FPR, TPR] = [0.251, 0.848] (see Tab. IV).

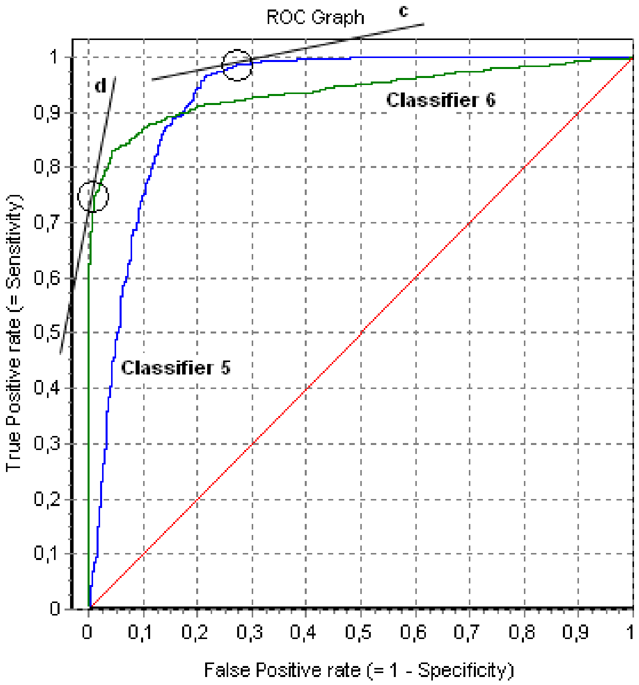
Besides the target situation of unequal error costs, also the situation of equal misclassification costs C(FP) = C(FN) is shown through the line *b*. Line slope is 4/1, the resulting point is at [FPR, TPR] corresponding to the threshold set by error minimization (see Tab. IV).

Again, costs minimization itself seems to be sufficient method for this problem, but what if the target conditions (e.g. misclassification costs) are not known, or are known only approximately? In this case, described visualization of the ROC curve and iso-performance line could be highly valuable.

ROC analysis and classifier performance comparison

In a ROC graph, the classifier performance can be compared visually, which is the first and quickest way. The more north-west the ROC curve is, the better (see Fig. 7 and 8).

The performance can be compared numerically according to the AUC. Higher AUC value means bet-



8: *ROC graph for classifiers 5–6*

VI: *Classifier performance numerical comparison*

Classifier	AUC	FPR (Specificity) at 95% sensitivity	Related threshold
Classifier 1	1.00000	0.00000 (1.00000)	0.66993
Classifier 2	0.97837	0.09500 (0.90500)	0.47712
Classifier 3	0.87484	0.50187 (0.49813)	0.40018
Classifier 4	0.72393	0.81125 (0.18875)	0.34555
Classifier 5	0.92656	0.20500 (0.79500)	0.64164
Classifier 6	0.93779	0.49100 (0.50900)	0.25521

ter performance. Tab. VI shows AUC of all classifiers discussed so far.

However, one considerable drawback about AUC must be emphasized – this metric is usually irrelevant when ROC curves cross. This is the case of classifier 5 and classifier 6 (see Fig. 8). The AUC rates both of the classifiers almost equivalently (92.7 % and 93.8 % for classifier 5 and classifier 6 respectively) because this type of metric measures the discriminative power across all possible thresholds, i.e. without regard to target conditions. But as was already shown, both classifiers are completely different, either successful in different target conditions (classifier 5 in a situation of higher C(FN), classifier 6 in a situation of higher C(FP)).

In this case, several other metrics are suggested. One of them could be the comparison according to the FPR (or specificity) at a fixed sensitivity level. This metric is useful in a situation of higher C(FN). Tab. VI compares all classifiers of this study not only with the AUC metric, but also with the FPR (specificity) at fixed sensitivity of 95 % (other values of sensitivity can be used as well, e.g. 92 %, 90 %). Here is obvious, that classifier 5 outperforms classifier 6 – specificity at fixed 95% sensitivity of classifier 5 (79.5 %) is higher than specificity of classifier 6 (50.9 %).

The same effect can be shown in a ROC graph. On Fig. 8, only the right-upper part of the ROC graph is practically relevant, and here the ROC curve of classifier 5 lies above the curve of classifier 6. Similarly, the case of higher C(FP) shifts the area of interest to the left-lower part of the ROC graph, where classifier 6 outperforms classifier 5.

Optimal decision threshold is set by the iso-performance line *c* touching the ROC curve of classifier 5, or line *d* touching the ROC curve of classifier 6 for either of the target conditions. Line *c* represents target conditions $p(\mathrm{P}) = p(\mathrm{N}) = 1/2$, C(FP) = 1, C(FN) = 5, so the line slope equals to 1/5, and resulting point on the ROC curve corresponds to the threshold 0.60160 with [FPR, TPR] = [0.274, 0.987] calculated by costs minimization. Line *d* represents target conditions $p(\mathrm{P}) = p(\mathrm{N}) = 1/2$, C(FP) = 5, C(FN) = 1, so the line slope equals to 5/1, and resulting point on the ROC curve corresponds to the threshold 0.40966 with [FPR, TPR] = [0.011, 0.750] calculated by costs minimization.

Software, platform and algorithms used

Receiver Operating Characteristic:

Own implementation based on Fawcett (2004, p. 13 – algorithm 2, p. 16 – algorithm 3), implemented in Borland Delphi 7 Professional.

Accuracy maximization/Error rate minimization, Costs minimization:

Own implementation based on ROC points – find minimum of errors (costs) produced by every possible classifier's threshold (classifier's output score), equally scored instances are treated as one instance (threshold considering errors/costs for all equally scored instances). Modified software application for ROC above.

Platform:

MS Windows XP Professional, Linux Mandriva 2009.1

SUMMARY

In the area of economical classification tasks, the accuracy maximization is often used to evaluate classifier performance. Accuracy maximization (or error rate minimization) suffers from the assumption of equal false positive and false negative error costs. Furthermore, accuracy is not able to express true classifier performance under skewed class distribution. Due to these limitations, the use of accuracy on real tasks is questionable. In a real binary classification task, the difference between the costs of false positive and false negative error is usually critical. To overcome this issue, the Receiver Operating Characteristic (ROC) method in relation to decision-analytic principles can be used. One essential advantage of this method is the possibility of classifier performance visualization by means of a ROC graph. This paper presents concrete examples of binary classification, where the inadequacy of accuracy as the evaluation metric is shown, and on the same examples the ROC method is applied. From the set of possible classification models, the probabilistic classifier with continuous output is under consideration. Mainly two questions are solved. Firstly, the selection of the best classifier from a set of possible classifiers. For example, accuracy metric rates two classifiers almost equivalently

(87.7 % and 89.3 %), whereas decision analysis (via costs minimization) or ROC analysis reveal different performance according to target conditions of unequal error costs of false positives and false negatives. Secondly, the setting of an optimal decision threshold at classifier's output. For example, accuracy maximization finds the optimal threshold at classifier's output in value of 0.597, but the optimal threshold respecting higher costs of false negatives is discovered by costs minimization or ROC analysis in a value substantially lower (0.477).

SOUHRN

Evaluace binárních klasifikačních úloh v ekonomické predikci

V oblasti ekonomických klasifikačních úloh je maximalizace přesnosti často používanou metrikou pro hodnocení klasifikačního výkonu. Maximalizace přesnosti (resp. minimalizace chybovosti) trpí předpokladem rovných nákladů chyb typu falešná pozitivita a falešná negativita. Kromě toho není přesnost schopna vyjádřit pravý výkon klasifikátoru v situaci nerovnoměrného rozložení tříd. Vzhledem k těmto omezením je použití přesnosti v reálných úlohách diskutabilní. V reálné binární klasifikační úloze je rozdíl mezi náklady falešné pozitivity a falešné negativity obvykle kritický. K překonání tohoto problému je použita metoda ROC ve spojení s principy rozhodovací analýzy. Jednou z podstatných výhod této metody je možnost vizualizace klasifikačního výkonu prostřednictvím ROC grafu. Tato studie prezentuje konkrétní příklady binární klasifikace, kde je ukázána neadekvátnost přesnosti jako evaluační metriky, a na stejných příkladech je dále aplikována metoda ROC. Z množiny dostupných klasifikačních modelů je uvažován pravděpodobnostní klasifikátor se spojitým výstupem. Zejména jsou řešeny dvě otázky. Za prvé výběr nejlepšího klasifikátoru z množiny dostupných klasifikátorů. Například metrika přesnosti hodnotí dva klasifikátory téměř ekvivalentně (87,7 % a 89,3 %), zatímco rozhodovací analýza (prostřednictvím minimalizace nákladů) nebo ROC analýza odhalují rozdílný výkon podle cílových podmínek nerovných nákladů falešných pozitivit a falešných negativit. Za druhé nastavení optimální rozhodovací hraniční hodnoty na výstupu klasifikátoru. Například maximalizace přesnosti nachází optimální hraniční hodnotu na výstupu klasifikátoru v hodnotě 0,597, avšak optimální hraniční hodnota respektující vyšší náklady falešných negativit je nalezena nákladovou minimalizací nebo ROC analýzou v hodnotě podstatně nižší (0,477).

binární klasifikace, predikce bankrotu, hodnocení výkonu klasifikátoru, maximalizace přesnosti, metoda ROC

REFERENCES

BISHOP, C. M., 2006: *Pattern Recognition and Machine Learning*. New York: Springer, 738 p. ISBN 0-387-31073-8.

ERKEL, A. R., PATTYNAMA, P. M. T., 1998: Receiver operating characteristic (ROC) analysis: Basis principles and applications in radiology. *European Journal of Radiology*, 27: 88–94. ISSN 0720-048X.

FAWCETT, T., 2004: *ROC Graphs: Notes and Practical Considerations for Researchers*. HP Labs Tech Report HPL-2003-4. 2nd version. Kluwer Academic Publishers. [online]. on the Internet: [quoted 2010-09-01]. Accessible.
<http://home.comcast.net/˜tom.fawcett/public html/papers/ROC101.pdf>.

OBUCHOWSKI, N. A., 2003: Receiver operating characteristic curves and their use in radiology. *Radiology*, 229: 3–8. ISSN 1527-1315.

POKORNÝ, M., 2009: *The application of neural networks and ROC method in classification tasks of economical prediction*. Dissertation theses. Brno, Mendel University, Faculty of Business and Economics, Department of Informatics.

PROVOST, F., FAWCETT, T., 1997: Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. In: HECKERMAN, D., PREGIBON, D., UTHURUSAMI, R. (ed.) *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*. ISBN 978-1-57735-027-9.

PROVOST, F., FAWCETT, T., 2001: Robust Classification for Imprecise Environments. *Machine Learning Journal*, 42, 3: 203–231. ISSN 0885-6125.

PROVOST, F., FAWCETT, T., KOHAVI, R., 1998: The Case Against Accuracy Estimation for Comparing Induction Algorithms. In: SHAVLIK, W. J. (ed.) *Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers. ISBN 1-55860-556-8.

Address

Ing. Martin Pokorný, Ph.D., Ústav informatiky, Mendelova univerzita v Brně, Zemědělská 1, 613 00, Brno, Česká republika, e-mail: martinp@pef.mendelu.cz