KYBERNETIKA — VOLUME *50* (2014), NUMBER 5, PAGES 647–660

RELATIVE COST CURVES: AN ALTERNATIVE TO AUC AND AN EXTENSION TO 3-CLASS PROBLEMS

Olga Montvida and Frank Klawonn

Performance evaluation of classifiers is a crucial step for selecting the best classifier or the best set of parameters for a classifier. Receiver Operating Characteristic (ROC) curves and Area Under the ROC Curve (AUC) are widely used to analyse performance of a classifier. However, the approach does not take into account that misclassification for different classes might have more or less serious consequences. On the other hand, it is often difficult to specify exactly the consequences or costs of misclassifications. This paper is devoted to Relative Cost Curves (RCC) – a graphical technique for visualising the performance of binary classifiers over the full range of possible relative misclassification costs. This curve provides helpful information to choose the best set of classifiers or to estimate misclassification costs if those are not known precisely. In this paper, the concept of Area Above the RCC (AAC) is introduced, a scalar measure of classifier performance under unequal misclassification costs problem. We also extend RCC to multicategory problems when misclassification costs depend only on the true class.

Keywords: classifier, performance evaluation, misclassification costs, cost curves, ROC curves, AUC

Classification: 93E12, 62A10

1. INTRODUCTION AND MOTIVATION

A Receiver Operator Characteristic plot (for overviews see for instance [1, 3, 6, 9]) allows a classifier to be evaluated and optimised over all possible operating points. The Area Under ROC has become a standard performance evaluation criterion in two-class pattern recognition problems, used to compare different classification algorithms independent of operating points, prior, and costs. However, this became the main reason for criticising this approach: “The most fundamental shortcoming is the simple fact that a single, scalar performance measure cannot capture all aspects of the performance differences between two classifiers. An important example of this failing occurs when the cost of misclassifying examples in one class is much different than the cost of misclassifying examples in the other class, or when one class is much rarer than the other. A scalar measure can give expected performance given a probability distribution over costs and class ratios, but will not indicate for which costs and class ratios one classifier outperforms the other.” [2]

DOI: 10.14736/kyb-2014-5-0647

Existing well-known cost curve techniques (see [2, 4, 7]) evaluate the performance of a classifier by visualising its error rate across all possible values of probability of an object coming from one of the classes. Cost Curves generalised approach and Brier Curves technique [7] allow to analyse classifiers performance over a range of two operating conditions simultaneously. The curves depict normalised expected costs of a classifier over the full range of possible class distributions and misclassification costs.

In case unequal costs are the primary interest the Relative Cost Curves (RCC) technique, introduced in [8], offers a characterisation of the performance of a classifier in a more intuitive way. The method observes expected relative costs of a binary classifier as a function of miscalssification costs.

Section 2 introduces the notion of RCC, briefly describes the algorithm for evaluating expected costs and explains how to visualise them. The next section is devoted to Area Above RCC, a scalar measure of classifier performance. The idea of AAC is analogue to AUC, but has its own advantages and disadvantages. In Section 4 a simplified misclassification costs problem is generalised to the multi-class case. Well known data examples are analysed and compared in Section 5, where we also demonstrate the benefit of RCC. Final conclusions are presented in Section 6.

2. RELATIVE COST CURVES FOR BINARY CLASSIFICATION PROBLEMS

Relative cost curves were introduced in [8] and are based on the following ideas. We consider a classification problem with two classes. In the biomedical literature persons showing a given disease are usually classified as + and healthy as −. We will use such a notation, i. e. we are faced with a dataset where each instance is assigned to one of the two classes + or −. In addition, we have a score for each instance: $X = (x_1, x_2, \ldots, x_n)$, $x_i \in \mathbb{R},\ i = 1, \ldots, n$. Without loss of generality, we assume that a higher score speaks more in favour of the class +. We will denote all instances that come from the negative class as X_-, $|X_-| = k,\ k \leq n$. Analogously, X_+ stands for all instances from the positive class, $|X_+| = n-k$. The score can come from a classifier, i. e. a function – usually learned from data – that assigns score for the class or disease under consideration to an object or patient based other attributes like blood values. In this case, it the score is often the probability for the class + that the classifier has computed. For instance, naive Bayes classifiers or logistic regression provide such scores in the form of probabilities. But the score does not need to be a probability. It can also be the signed distance to the separating hyperplane of a support vector machine, the output of a neuron of a neural network or simply the value of a specific attribute, like a biomarker, usually the measurement of a certain biochemical component which is an indicator for a specific disease.

A simple decision rule of the form *"If $x_i < t$, then class −, otherwise class +"* will be used to make the classification decision. For every fixed threshold $t \in \mathbb{R}$, all instances can be divided into four groups as shown in Table 1, which is also known as contingency table or confusion matrix.

False positive rate will stand for false positive instances divided by k, true positive rate – for true positive divided by $n-k$. The higher t is chosen, the less false positives the classifier will produce, but at the same time the number of true positives will decrease. The choice of t depends on two aspects.

- The misclassification costs for false positives and for false negatives. It should be noted that misclassification costs for false positives and false negatives can be quite different. It makes a big difference whether a healthy person is wrongly classified as having a specific disease or a patient suffering from a specific disease is considered to be healthy.
- The prior distribution of the classes. Although the misclassification costs for false negatives might be high, it might still not be advisable to choose a low threshold t in order to reduce the number of false negatives. If the fraction of instances from the class + is very small, a small threshold t will lead to an extremely large number of false positives in comparison to the false negatives and the overall misclassification costs will be very high.

	predicted class	
true class	−	+
−	True Negative	False Positive
+	False Negative	True Positive

Tab. 1. Contingency table.

More generally speaking, we consider a normalised cost matrix as shown in Table 2. A correctly classified instance will cause no costs – the zeros in the diagonal – whereas a false positive will cause the misclassification costs 1 and a false negative the misclassification costs $c > 0$. Without loss of generality, we have assumed that the misclassification costs for false positives are normalised to 1. This is no restriction, since we have not specified the unit in which we measure the costs (Euros, Dollars, Cents,...). So we simply say that the cost unit corresponds to the misclassification costs of a false positive. When the classification is based on minimising the costs, it is sufficient to know the ration of the costs for a false positive and a false negative, but not their absolute values.

	predicted class	
true class	−	+
−	0	1
+	c	0

Tab. 2. A normalised cost matrix for a classification problem with two classes.

Given the value c, the average misclassification costs for threshold t are

$$g(t) \quad = \quad \frac{1}{n} \cdot \Big(\sum_{i=1}^{n} \mathbb{I}_{\{x_i \in X_+\}} \mathbb{I}_{\{x_i < t\}} + c \cdot \sum_{i=1}^{n} \mathbb{I}_{\{x_i \in X_-\}} \mathbb{I}_{\{x_i \geq t\}} \Big), \qquad (1)$$

where $\mathbb{I}_A$ is the indicator function for A. The optimal threshold t can now be determined by minimising Eq. (1).

In real applications, it is often difficult to specify exact misclassification costs. Therefore, the value c is usually not fixed, but considered to be variable or a probability distribution over c is assumed (compare [4]). Here, we do not make any assumption about c. We can represent the average misclassification costs as a function CC of c, i. e.

$$CC(c) \; = \; \min\{g(t) \mid t \in \mathbb{R}\}. \tag{2}$$

This cost curve has two disadvantages.

- The normalisation of the costs in Table 2 with respect to the false positives was a more or less arbitrary choice. We could have carried out the normalisation in the same way with respect to the false negatives. However, our normalisation causes an asymmetric situation. The cases where false negatives are considered worse than false positives correspond to the infinite interval $c \in (1, \infty)$, whereas the cases where false positives are considered worse than false negatives correspond to the finite interval $c \in (0, 1)$. Since RCC is a visualisation technique, we use a logarithmic scale for the costs to avoid asymmetry. In this way, cases where false negatives are considered worse than false positives correspond to the interval $c \in (-\infty, 0)$ and cases where false positives are considered worse than false negatives correspond to the interval $c \in (0, \infty)$. If we do not carry out this normalisation, the cost curve could look very unstable in the interval $(0, 1)$, since essentially the same things are represented in the finite interval $(0, 1)$ and the infinite interval $(1, \infty)$, except that the roles of the two classes are exchanged with respect to the costs.

- A low value of CC at costs c does not necessarily mean that the score is useful for the classification. For instance at $c = 1$, we can easily achieve low costs if the two classes are extremely imbalanced, say 99% of the instances belong to the class $-$. Choosing t large enough will assign all instances to the majority class $-$ and the average misclassification costs become in this case 0.01, corresponding to the 1% false negatives. But this value of 0.01 for the average misclassification costs is obtained even without considering the scores. This would mean that the scores are useless unless they lead to a misclassification rate significantly lower than 0.01. Therefore, the relative costs of the classifier using the scores compared to the naive classifier are considered. The naive classifier ignores the scores and assigns all instances to the class which yields the lower misclassification costs. The average misclassification costs for the naive classifier are

$$CC_{\text{naive}}(c) \; = \; \frac{1}{n} \min\{k, (n-k) \cdot c\}, \tag{3}$$

  where k of the n instances belong to the class $-$.

This leads to the definition of the relative cost curve

$$RCC(\log_2(c)) \; = \; \frac{CC(c)}{CC_{\text{naive}}(c)} \cdot 100\,\%. \tag{4}$$

A value of 100% for RCC at costs c would mean that the scores are of no use for these costs. The lower the value of RCC, the more the scores contribute to minimise the misclassification costs. At 0, we would have a perfect classifier without misclassifications. Values over 100% indicate that the assumption that higher scores always speak more in favour of the class + is not always valid. This might also simply be a sampling effect, since we always deal with a finite sample.

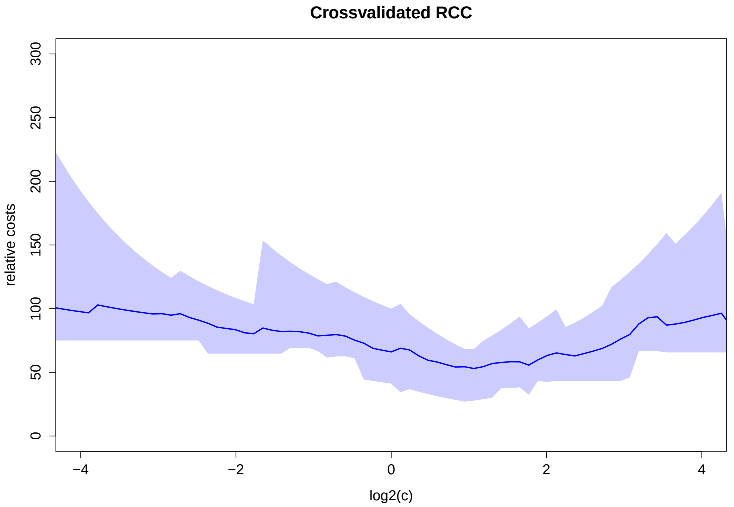

Fig. 1. An example for a Relative Cost Curve with a band based on cross-validation.

Figure 1 shows an example of an RCC. The curve is based on 10-fold cross-validation, i. e. the curve corresponds to the mean of 10 RCCs. The area around the curve corresponds to the standard deviation derived from the 10 RCCs.

Apart from the theoretical definition of RCCs, there is also a computational aspect, how these curves can be calculated to plot their graphs. A simple approach would evaluate Eq. (4) at equidistant points to draw the graph. However, one would have to fix a sampling rate for the equidistant points. A large sampling rate would lead to high computational costs, whereas a low sampling rate would result in a bad approximation quality of the true RCC.

Fortunately, the RCC can be calculated exactly by choosing a flexible sampling rate. The algorithm for this drawing algorithm is described in detail in [8] and basically needs a sorting of the instances in increasing order with respect to their scores.

3. AREA ABOVE CURVE

Letting the threshold t vary from ∞ to $-\infty$, ROC curves plot the false positive rate of a classifier on the x-axis against the true positive rate on the y-axis. Note that the ROC curve is independent of the prior distribution of the classes. AUC is defined in terms of ROC curves, summarising the ROC curve to a single performance value. It is simply the area under the ROC curve. A value of 0.5 for AUC corresponds to random guessing, whereas an AUC of 1 would mean that the classifier can perfectly separate the classes, at least for the given data. AUC can also be interpreted as the probability that a randomly selected instance from the class + has a higher score than a randomly selected instance from the class −.

We apply a similar idea and introduce the Area Above Curve (AAC) in terms of Relative Cost Curves. As stated before, RCC values generally lie below 100%. Therefore, we propose to measure classifier performance as the area between RCC and the $y = 100$ curve – the line corresponding to the performance of the naive classifier. Notice, that RCC is defined on an unbounded interval, therefore AAC needs to be restricted to a finite interval of misclassification costs to avoid infinite values (Figure 2). Essentially, AAC is the area between the RCC and the 100% line of the naive classifier. But since the RCC should normally be under the 100% line and for reasons of a simple name, we simply call it area above curve.

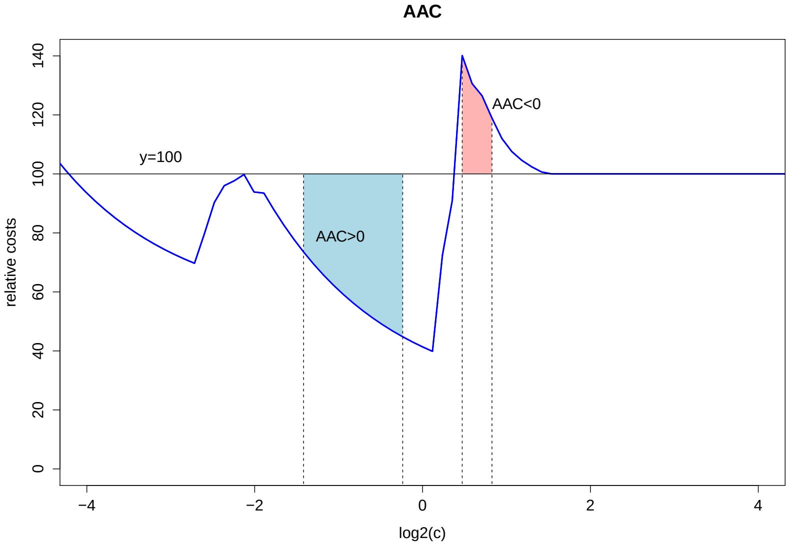

Fig. 2. Area Above Curve on two different intervals of misclassification costs.

The AAC value is obtained by standardising the area with respect to the area of the naive classifier on a given interval:

$$AAC[a, b] = 1 - \frac{\int_a^b RCC(\log_2(c))\mathrm{d}(\log_2 c)}{100(\log_2(b) - \log_2(a))}.$$

Perfect classification results in a large area. The AAC value is bounded by 1. This upper bound of 1 corresponds to a perfect classifier with no misclassification. Poor classification results in a small area. Even negative values of AAC are possible if a classifier performs worse than the naive classifier. This can happen, especially when cross-validation is used and the classifier by chance yields sometimes worse results than the naive classifier. We recommend to use RCC as a primary analysis, but conclusions might be supported by the AAC value.

For example, both classifiers in Figure 3 perform similarly, but the AAC value measured on a given interval points out which one is better for this range of costs.

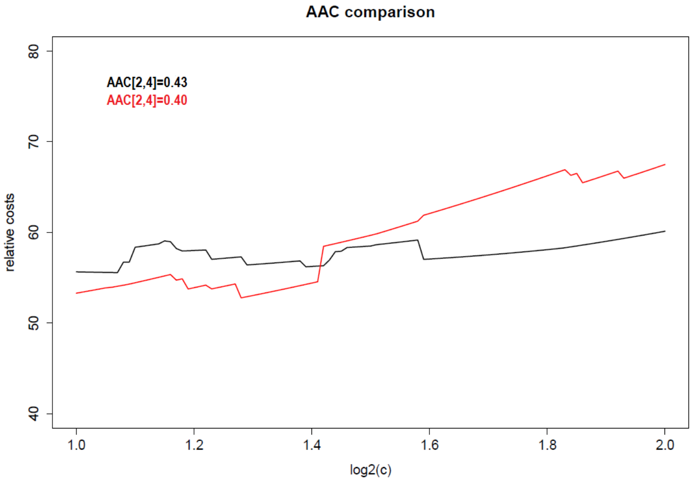

Fig. 3. Comparison of the performance of two classifiers.

4. MULTICLASS PROBLEMS

Since we have adopted the ideas from AUC for the introduction of AAC, the application is restricted to classification problems with two classes. There are extensions of ROC curves and AUC to classification problems with three [12] or more classes [5, 10]. Here we also extend the idea of AAC to classification problems with three classes under certain assumptions on the cost matrix. First consider the following example:

A company wants to classify potential customers in order to send them an appropriate offer. Classes depend on age (student, senior, working age) or region (districts, cities, countries etc.) or any other factor. Based on current client classes contribution levels, approximate costs of loosing a potential customer from a particular class might be calculated. Notice, that for the company there is no interest if a “Prague citizen” has been classified as an “Ostrava citizen” or as a “Brno citizen”, in both cases the company looses a potential “Prague citizen” client, by sending an inappropriate marketing offer.

A similar situation can occur in medical diagnosis. If we do not only distinguish between two classes, i. e. healthy and a specific disease, but between a number of different diseases, the wrong diagnosis means that the patient is not treated correctly and will suffer from the consequences of the undiscovered and untreated disease.

Now consider a classification problem with three classes. We will denote them by 0, 1 and 2. Assume that the misclassification costs only depend on the true class but not on the class to which an object is assigned wrongly, e. g. classifying an instance from class 0 to class 1 has the same consequences as classifying the instance to class 2. The general structure of such a misclassification costs matrix is given in Table 3. On the right hand side of Table 3 is normalised with respect to the costs of the first class. This cost matrix is a simplified extension to three classes of a cost matrix for two classes in Table 2.

	predicted class		
true class	0	1	2
0	0	c_0	c_0
1	c_1	0	c_1
2	c_2	c_2	0

	predicted class		
true class	0	1	2
0	0	1	1
1	$\tilde{c}_1$	0	$\tilde{c}_1$
2	$\tilde{c}_2$	$\tilde{c}_2$	0

Tab. 3. Cost matrix for a simplified 3-class problem (left) and a normalised version (right).

We now assume that two threshold values t_1 and t_2 must be specified. Scores below t_1 are assigned to class 0, scores between t_1 and t_2 to class 1 and scores above t_2 to class 2. The main advantage of the above mentioned assumptions is that both thresholds can be found independently.

Let F_0 denote all misclassified instances from class 0 (Figure 4). Instances misclassified from class 1 to class 0 are denoted by F_{10} and correspondingly instances from class 1 wrongly assigned to class 2 are written as F_{12}. Misclassified instances from class 2 are denoted as F_2. The values depend on the choice of the thresholds.

Applying the same logic as previously, the misclassification costs can be computed in the following way.

$$h(t_1) = \{F_0(t_1) + \tilde{c}_1 F_{10}(t_1)\},$$

$$f(t_2) = \{\tilde{c}_1 F_{12}(t_2) + \tilde{c}_2 F_2(t_2)\}.$$

The minimal misclassification costs are

$$CM(\tilde{c}_1, \tilde{c}_2) = \min\{h(t_1) \mid t_1 \in \mathbb{R}\} + \min\{f(t_2) \mid t_2 \in \mathbb{R}\}.$$

Thus, the 3-class problem is reduced to two 2-class problems and the algorithm introduced in [8] can be applied directly. This approach can be extended in a straight forward manner to find the optimal thresholds for classification problems with k classes given that misclassification costs depend only on the true class but not on the wrongly assigned class.

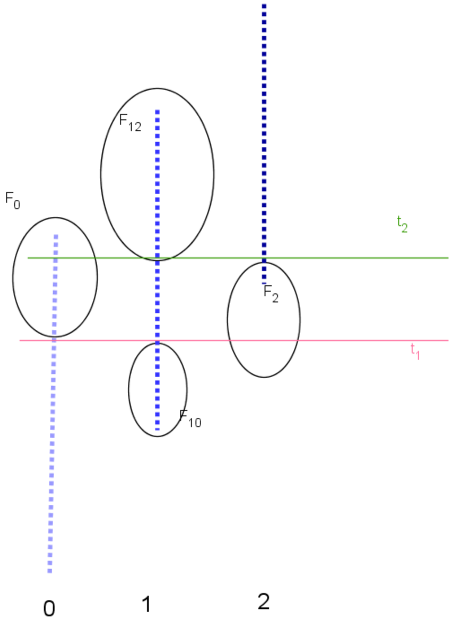

Fig. 4. Threshold choice for a 3-class problem.

Of course, it is impossible to draw corresponding RCCs for three or more classes, since the misclassification costs of a classifier would be a function of three or more arguments, i. e. as many classes we have. As in the case of two-class problems, one can carry out a normalisation by setting one of the costs c_i to 1. In this way, we could get rid of one argument. This would only help for the visualisation for 3-class problems. In this case, we could draw the misclassification costs as a 3D-view of a function of two arguments. But in any case, we can extend our concept of AAC to an arbitrary number of classes. Instead of the area above a curve, we would have to consider a (hyper-)volume above a manifold in the same way as AUC is extended to HUM ((hyper-)volume under manifold) in [10]. For the above described 3-class problem we obtain

$$RCM(\log_2(\tilde{c}_1), \log_2(\tilde{c}_2)) = \frac{CC(\tilde{c}_1, \tilde{c}_2)}{\min\{l\tilde{c}_1 + m\tilde{c}_2; k + m\tilde{c}_2; k + l\tilde{c}_1\}} \cdot 100\%,$$

where k, l, m are the number of instances from classes 0, 1 and 2, respectively. For given lower $(\tilde{c}_0^L,\ \tilde{c}_1^L,\ \tilde{c}_2^L)$ and upper $(\tilde{c}_0^U,\ \tilde{c}_1^U,\ \tilde{c}_2^U)$ bounds for each misclassification cost, the AAC analogue – (hyper-) Volume Above Manifold – is

$$HAM = 1 - \frac{\int\limits_{\tilde{c}_1^L}^{\tilde{c}_1^U} \int\limits_{\tilde{c}_2^L}^{\tilde{c}_2^U} RCM(\log_2(\tilde{c}_1), \log_2(\tilde{c}_2))\, \mathrm{d}(\log_2 \tilde{c}_2)\, \mathrm{d}(\log_2 \tilde{c}_1)}{\left(\log_2\left(\tilde{c}_1^U\right) - \log_2\left(\tilde{c}_1^L\right)\right)\left(\log_2\left(\tilde{c}_2^U\right) - \log_2\left(\tilde{c}_2^L\right)\right) \cdot 100}.$$

5. EXAMPLES

5.1. Generated data

As our first example, we consider artificial data where scores of the instances from the class $-$ follow a standard normal distribution $N(0,1)$ and scores of the instances from the class $+$ follow the normal distribution $N(d,1)$, where $d > 0$ is assumed. The prior probability of the class $+$ is $p_+ \in (0,1)$, i. e. a fraction of p_+ instance scores will be generated from the class $+$ and a fraction of $(1-p_+)$ from the class $-$.

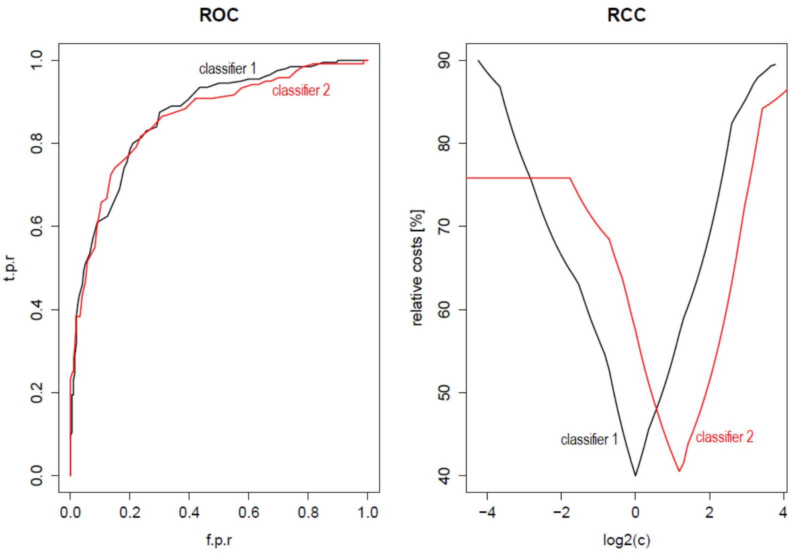

Fig. 5. Classifier comparison with ROC and RCC.

Let us compare classifier 1 where $d = 1.5$ and $p_+ = 0.5$ with classifier 2 where $d = 1.5$ and $p_+ = 0.3$. Classifier 1 has equal proportions of positive and negative examples, classifier 2 has 30% positive instances and 70% negative ones. Figure 5 shows ROC curves on the left hand side and RCCs on the right hand side. On the one hand, the ROC plot indicates that the classifiers have a similar performance. But on the other hand, our RCC analysis reveals the difference between the classifiers which comes mainly from the fact that we have used different prior distributions. The prior class distribution does not have any influence on ROC curves. From the RCCs it can be seen, that when misclassifications of instances from the class $+$ have more serious consequences than misclassifying negative instances, it is better to use classifier 2.

RCC mainly depends on two factors: prior and scores. To understand the RCC's behaviour we will observe the curve analysing both factors independently. Figure 6 shows how the RCC changes depending on the prior probability of the class $+$. We fixed $d = 1.5$ and generated 3 datasets with $p_+ = 0.9$, $p_+ = 0.5$ and $p_+ = 0.1$. The corresponding RCCs are shown in the figure.

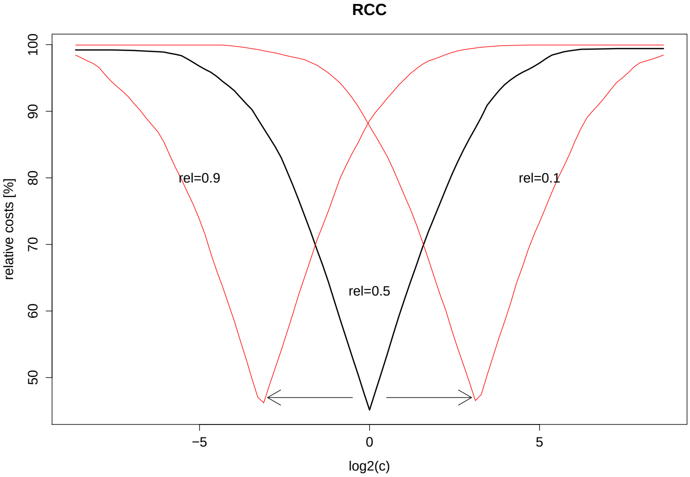

Fig. 6. RCC's behaviour depending on the prior probability p_+ of the class $+$.

We take a closer look at the RCC which corresponds to the classifier with $p_+ = 0.5$. The curve is non-increasing on the interval $c \in (-\infty, 0)$. Relative costs are minimal at 0 when misclassifying negative instances has the same consequences as misclassifying positive instances. The curve is non-decreasing on $c \in (0, +\infty)$. All RCCs with $p_+ = 0.5$ will show this behaviour for the following reason. Consider given misclassification costs $c \in (0, \infty)$. Let us abbreviate the sums in Eq. (1), i. e. the number of false positives and false negatives, by

$$s_+ \;=\; \sum_{i=1}^{n} \mathbb{I}_{\{x_i \in X_+\}} \mathbb{I}_{\{x_i < t\}} \tag{5}$$

and

$$s_- \;=\; \sum_{i=1}^{n} \mathbb{I}_{\{x_i \in X_-\}} \mathbb{I}_{\{x_i \geq t\}}, \tag{6}$$

repsectively. Given that $n/2$ of the data belong to the $+$ and $n/2$ of the data to class $-$, the relative costs are

$$\begin{aligned} RCC(\log_2(c)) &= \frac{\min\limits_{t \in \mathbb{R}} \left\{\frac{1}{n} \cdot (s_+ + c \cdot s_-)\right\}}{\frac{\min\{c,1\}}{2}} \\ &= \frac{\min\limits_{t \in \mathbb{R}} \left\{\frac{1}{n} \cdot \left(\max\{\frac{1}{c}, 1\} \cdot s_+ + \max\{1, \frac{1}{c}\} \cdot s_-\right)\right\}}{\frac{1}{2}} \\ &\geq \frac{\min\limits_{t \in \mathbb{R}} \left\{\frac{1}{n} \cdot (s_+ + s_-)\right\}}{\frac{1}{2}}. \end{aligned} \tag{7}$$

Eq. (7) corresponds exactly to the relative costs for $c = 1$, i. e. $\log_2(c) = 0$.

For $p_+ \neq 0.5$ the RCC shifts to the right when $p_+ < 0.5$ and to the left when $p_+ > 0.5$.

Another example shows how RCC behaves when the parameter d is changed where we assume equal prior probabilities for both classes. The parameter indicates how well two classes are separated, the bigger d, the better classes are separated. In Figure 7 three RCCs for $d = 0.5$, $d = 1.5$ and $d = 4.5$ are shown. An ideal classifier separating the classes perfectly with no misclassifications would have the x-axis as its RCC. Similarly, a classifier where scores provide no information on the classes would have $y = 100$ as its RCC, since random guessing is the best strategy.

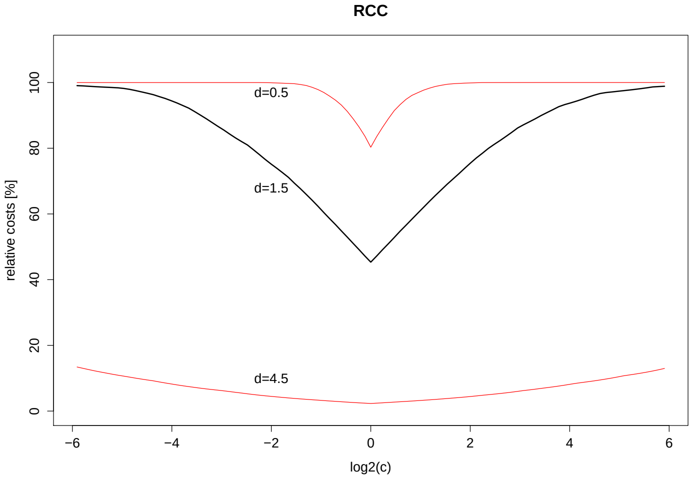

Fig. 7. RCC's behaviour depending on the parameter d, i. e. how well classes are separated.

5.2. Breast cancer data

The breast cancer dataset [11] was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. He assessed biopsies of breast tumours for 699 patients up to 15 July 1992; each of nine attributes has been scored on a scale of 1 to 10, and the outcome is also known. There are 699 instances: 241 of class malignant and 458 of class benign. The dataset is also known as the "biopsy dataset".

We compare two classifiers, i. e. two attributes or biomarkers: marginal adhesion and bland chromatin. Figure 8 shows the ROC curves on the left hand side. It seems that bland chromatin clearly outperforms marginal adhesion. The right plot shows the RCCs based on 10-fold cross-validation. On the one hand, in general terms bland chromatin outperforms marginal adhesion here. This conclusion is supported by an AAC value of 0.19 for bland chromatin and 0.02 for marginal adhesion. But on the other hand, for $c \in [0.0625, 1.1]$ (or $\log_2(c) \in [-4, 0.1]$) bland chromatin is unstable, marginal adhesion performs better.

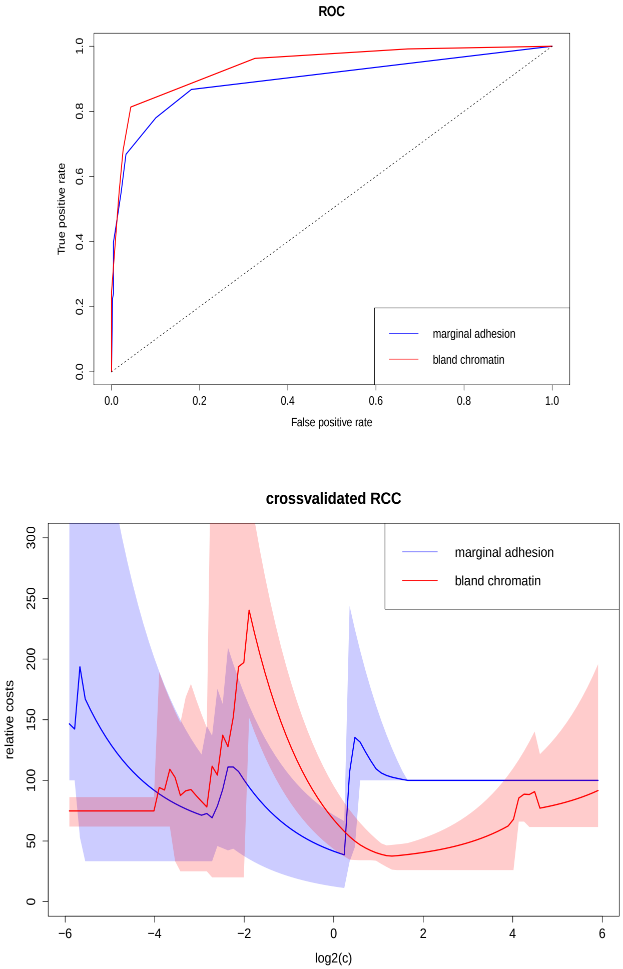

Fig. 8. Comparing of marginal adhesion and bland chromatin classifiers.

6. CONCLUSIONS

We have demonstrated in this paper that our proposed RCC and AAC analysis can complement ROC and AUC analysis. RCC and AAC reveal properties of the classifiers that cannot be seen from ROC curves, since neither the prior distribution of the classes nor the consequences (costs) for misclassifications are taken into account whereas these aspects play a crucial role for RCC and AAC.

ACKNOWLEDGEMENTS

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115523-2 – COMBACTE, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution.

(Received July 30, 2013)

REFERENCES

[1] A. P. Bradley: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition *30* (1997), 1145–1159.

[2] C. Drummond and R. C. Holte: Cost curves: An improved method for visualizing classifier performance. Machine Learning *65* (2006) 95–130.

[3] T. Fawcett: An introduction to roc analysis. Pattern Recognition Lett. *27* (2006), 861–874.

[4] D. J. Hand: Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning *77* (2009), 103–123.

[5] D. J. Hand and R. J. Till: A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning *45* (2001), 171–186.

[6] J. A. Hanley: Receiver operating characteristic (ROC) methodology: the state of the art. Critical Reviews in Diagnostic Imaging *29* (1989), 307–335.

[7] J. Hernández-Orallo, P. Flach, and C. Ferri: Brier curves: a new cost-based visualisation of classifier performance. In: Proc. 28th International Conference on Machine Learning (ICML-11) (L. Getoor and T. Scheffer, eds.), ACM, New York 2011, pp. 585–592.

[8] F. Klawonn, F. Höppner, and S. May: An alternative to ROC and AUC analysis of classifiers. In: Advances in Intelligent Data Analysis X, (J. Gama, E. Bradley, and J. Hollmén, eds.), Springer, Berlin 2011, p. 210–221.

[9] W. J. Krzanowski and D. J. Hand: ROC Curves for Continuous data. Chapman and Hall, London 2009.

[10] J. Li and J. P. Fine: ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. Biostatistics *9* (2008), 566–576.

[11] P. M. Murphy and D. W. Aha: Uci repository of machine learning databases. 1992. Avaible: `http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`.

[12] D. Mossman: Three-way ROCs. Medical Decision Making *19* (1999), 78–89.

Olga Montvida, University of Latvia, Department of Mathematics, Zellu street 8, Riga, LV-1002. Latvia.
e-mail: olgamontvida@yahoo.com

Frank Klawonn, Bioinformatics and Statistics, Helmholtz Centre for Infection Research, Inhoffenstr. 7, D-38124 Braunschweig, Germany and Department of Computer Science, Ostfalia University of Applied Sciences, Salzdahlumer Str. 46/48, D-38302 Wolfenbuettel. Germany.
e-mail: f.klawonn@ostfalia.de