Statistical criterion for comparison of binary classifier accuracy

E V Strebkov¹, V S Zheltukhin¹,² and I A Borodaev²
¹Department of Mathematical Statistics, Kazan Federal University, 18 Kremlin street, Kazan, 420008, Russia
²Department of Plasma Technology and Nanotechnology of High Molecular Weight Materials, Kazan National Research Technological University, 68 Karl Marx street, Kazan 420015, Russia

E-mail: str9050258629@yandex.ru

Abstract. An ABC-method (Accuracy Binary Classifier) for a more accurate assessment of the binary classifier's quality compared to the ROC-method (Receiver Operating Characteristic) is proposed. The ABC-method is suitable for quantitative and qualitative measures in independent and dependent small samples and it is not limited by the laws of distribution of objects in a sample. The ABC-method is effective for classification issues in different scientific and applied fields: IT, physical, technical, medical.

1. Introduction

The effectiveness of computer decision support systems is determined by the quality of diagnostic algorithms, based on which decisions about the current state of the object are made. Binary classifiers (hereinafter referred to as BC), which classifies the object under study as a class with a positive or negative property, are used in many application fields, such as physical and technical problems [1], medical diagnostics [2, 3], machine learning [4–6].

A popular tool for analyzing of diagnostic systems is the method of Receiver Operating Characteristic (ROC). The ROC analysis principle is the comparison of the sensitivity and specificity of the test in the form of the ROC curve, and the area under this curve (AUC) is used to characterize the binary classifier's quality [7–10]. For analyzing a BC's quality, measures of two BCs are compared, for example, on training and test datasets. Series of tests are usually performed, and the mean values of their AUCs are compared with each other, for example, using the Student t-test [10] to compare two binary classifiers by means of the ROC method. Therefore, the AUC value is a generalized measure of the BC quality.

The binary classifier's accuracy evaluates based on the priority classes are important for the applied fields. In this paper, the Accuracy Binary Classifier (ABC) method is proposed for the analysis of the binary classifier's accuracy with different options of class priority.

2. Analysis of binary classifiers accuracy by ABC-method

A binary classifier refers the object to a class with a positive (P) or a negative (N) property, but a misclassification is possible.

The binary classifier is characterized by the main indicators (classes):
– TP (True Positives) - correctly classified positive cases;
– FN (False Negatives) - positive cases classified as negative (type I error);
– TN (True Negatives) - correctly classified negative cases;
– FP (False Positives) - negative cases classified as positive (type II error).

Let us consider the problem of comparing of BC1 and BC2 binary classifiers' indicators, for which \( k_i \) and \( m_i \) of cases for the main classes are given in Table 1.

| Classifier | TP | FN | TN | FP |
|------------|----|----|----|----|
| BC1        | \( k_1 \) | \( k_2 \) | \( k_3 \) | \( k_4 \) |
| BC2        | \( m_1 \) | \( m_2 \) | \( m_3 \) | \( m_4 \) |

Relative measures for BC’s are as follows:
– sensitivity is the percentage of correctly classified positive cases
  \[ Se = \frac{100 \cdot TP}{TP + FN} \text{ (\%)}. \]
– specificity is the percentage of correctly classified negative cases
  \[ Sp = \frac{100 \cdot TN}{TP + FP} \text{ (\%)}. \]

Often, some classes have a higher priority, for example, in such fields as diagnosis the presence of the disease, establishing a critical level in the regulation of physical and technological processes, etc. In these cases, the classifier's accuracy assessments based on the priority classes have great importance. The ABC method principle consists in the analysis of significance of differences between two binary classifiers' measures using the statistic test for different options of class priority.

Let us consider the application of the ABC method to different options for selecting priority classes: 1) TP and FN; 2) TN and FP; 3) TP and TN.

Let the ratio between BC1 and BC2 for option 1 is given in table 2 with the following measures:
– percentage of correctly classified positive cases by BC1
  \[ P_1 = \frac{100 \cdot TP}{TP + FN} = \frac{100 \cdot k_1}{k_1 + k_2} \text{ (\%)}. \]
– percentage of correctly classified positive cases by BC2
  \[ P_2 = \frac{100 \cdot TP}{TP + FP} = \frac{100 \cdot m_1}{m_1 + m_2} \text{ (\%)}. \]

In Table 2, for Option 1, "Yes" means correct classification and "No" means misclassification of positive cases.

| Classifier | Yes | No |
|------------|-----|----|
| BC1        | \( P_1 \) | \( 100 - P_1 \) |
| BC2        | \( P_2 \) | \( 100 - P_2 \) |

To identify the significance of the difference between the percentages \( P_1 \) and \( P_2 \), let us consider the algorithm of the statistical (angular) \( F \)-ratio test [11].

Step 1. The largest percentage is chosen from the percentages \( P_1 \) and \( P_2 \), for example, it is \( P_1 \). The principal hypothesis \( H_0 = \{ P_1 \leq P_2 \} \) and alternative hypothesis \( H_1 = \{ P_1 > P_2 \} \) are formulated.

Step 2. The angles for \( P_1 \) and \( P_2 \) (in radians) are calculated by equations
  \[ \varphi_1 = 2 \arcsin \left( \frac{P_1}{100} \right)^{1/2}, \quad \varphi_2 = 2 \arcsin \left( \frac{P_2}{100} \right)^{1/2}, \quad (1) \]
and the empirical value $\varphi$ of the test is calculated by the equation

$$\varphi = (\varphi_1 - \varphi_2) \cdot \left(\frac{nm}{n+m}\right)^{1/2},$$  \hspace{1cm} (2)$$

where $n = k_1 + k_2, m = m_1 + m_2$.

**Step 3.** The critical value is chosen from the corresponding table [11] in accordance with the significance level $q$, for example,

$$\varphi_{cr} = \begin{cases} 1.64, & \text{if } q \leq 0.05, \\ 2.28, & \text{if } q \leq 0.01. \end{cases}$$

If $\varphi \geq \varphi_{cr}$, the $H_0$ hypothesis is rejected and, therefore, the inequality $P_1 > P_2$ is true at the significance level $q$.

In case of Option 2 with the priority classes $TN$ and $FP$, for negative cases according to table 1, the corresponding percentages are as follows

$$P_1 = \frac{100 \cdot TN}{TN + FP} = \frac{100 \cdot k_3}{k_3 + k_4} \%,$$

$$P_2 = \frac{100 \cdot TN}{TN + FP} = \frac{100 \cdot m_3}{m_3 + m_4} \%,$$

and in the equation (2) $n = k_3 + k_4, m = m_3 + m_4$.

In case of Option 3 with the priority classes $TP$ and $TN$, for the classification accuracy evaluated sample as a whole set, the corresponding percentages according to Table 1 are equal to

$$P_1 = \frac{(k_1 + k_3) \cdot 100}{n} \%,$$

$$P_2 = \frac{(m_1 + m_3) \cdot 100}{m} \%.$$  

Here and in equation (2) $n = k_1 + k_2 + k_3 + k_4, m = m_1 + m_2 + m_3 + m_4$.

The used statistical criterion is universal with simple limitations in its application [11]:

1) the measurement can be carried out in both quantitative and qualitative scale;
2) the sample characteristics can be both dependent and independent without limitation by the laws of their distributions;
3) the lower limit for the parameters $n$ and $m$ should be at least 5 objects without constraints to upper limit.

**Remark.** Statistical tests for comparing frequencies based on binomial distributions suggest an equal possibility of the studied property in the classification objects [12], which is not always feasible. For example, in medical diagnosis, the probability of diseases depends on the patients' sex and age.

The ABC method allows us a more accurate comparison of the measures of two binary classifiers with different options of class priority. In case of option 1, the BC accuracy is assessed relative to the classification of positive cases. In case of option 2, the BC accuracy is assessed relative to the classification of negative cases. In case of option 3, the correct BC classification accuracy is assessed relative to type I and II errors. The ABC method application for three options provides a system analysis of the BC's accuracy.

Contrary to ABC method, AUC is a generalized characteristic of the BC's quality in the ROC method. For estimating the AUC, it is necessary using a quadrature with a small number of nodes, which leads to errors [13]. Besides, for comparing the quality of two binary classifiers by the ROC method, a series of tests are usually performed, and then the mean values of their AUCs are compared with each other, for example, using the Student $t$-test [10]. It is necessary to take into account the conditions of the applicability of the comparison criteria for the mean values. For example, the Student $t$-test is applicable only for two independent samples with a general characteristic, obeying the normal
law with the same variances [14]. Similar requirements to homogeneity criteria limit the application of the ROC-method for comparing the quality of binary classifiers.

3. Conclusions

The advantage of the proposed ABC method is the possibility of obtaining an estimate of the accuracy of binary classifiers in different priority classes. This allows us to perform a system analysis of the binary classifier's quality.

Contrary to the ROC method, the ABC method is universal due to following specifics:
1) a more accurate analysis of the binary classifier quality for quantitative and qualitative measures is allowed, while the ROC method is applicable for quantitative characteristics only [15-17];
2) it is visual and does not require users to have additional knowledge of analytical statistics and computational mathematics;
3) it has no limits in distribution law of sample objects in contrast to binomial statistical criteria.
4) it is suitable for both independent and dependent small samples.

The binary classifier's accuracy assessments based on the priority classes are currently important for the applied issues. The ABC method is effective in designing and different training options for binary classifiers due to its advantages [10].

The ABC method universality is extended possibilities of the binary classifier's quality analysis in important scientific and applied fields, for example, information technology [4-6], physical, engineering and medical issues [1-3].

Acknowledgments

This paper was supported by the Ministry of Education and Science of the Russian Federation (Program to Improve the Competitiveness of Kazan Federal University).

References

[1] Birger I A 1979 Technical diagnostics (Moscow: Mashinostroenie) 240 (In Russian)
[2] Vlasov V V 1988 Diagnostic testing effectiveness (Moscow: Meditsina) 256 (In Russian)
[3] Lasted L 1971 Introduction to a problem of a decision making in medicine (Moscow: Mir) 282 (In Russian)
[4] Mitchel T M 1997 Machine Learning (New York: McGraw-Hill) 432
[5] Vapnik V N 1998 Statistical learning theory (New York: John Wiley & Sons, Inc.) 768
[6] Witten I H and Eibe F 2000 Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (San Diego, CA: Morgan Kaufmann) 525
[7] Metz C E 2000 Fundamental ROC analysis Progress in Medical Physics and Psychophysics Handbook of Medical Imaging vol 1 (Bellingham WA: SPIE Press) 754–769
[8] Spackman K A 1989 Signal detection theory: Valuable tools for evaluating inductive learning Proc. of the 6th Int. Workshop on Machine Learning 160–163
[9] Ferri C Hernández-Orallo J and Salido M A 2003 Volume Under the ROC Surface for Multiclass Problems. Exact Computation and Evaluation of Approximations Technical Report Dep. Sistemes Informàtics i Computació, Univ. Politècnica de València (Spain) 36–43
[10] Vezhnevets V 2006 Classifier performance evaluation Computer graphics and multimedia Issue 4 (1) http://cgm.computergraphics.ru/content/view/106 (In Russian)
[11] Gubler E V 1978 Computational methods for analysis and recognition of pathological effects (Leningrad: Meditsina) 296 (In Russian)
[12] Gmurman V E 2003 Theory of probability and mathematical statistics (Moscow: Vyshaya shkola) 479 (In Russian)
[13] Samarsky A A and Gulin A V 1989 Numerical methods (Moscow: Nauka) 432 (In Russian)
[14] Brandt S. 2014 Data Analysis. Statistical and Computational Methods for Scientists and Engineers (Berlin, New York: Springer) 523
[15] Fawcett T 2004 *ROC Graphs: Notes and Practical Considerations for Researchers* (Netherland: Kluwer Academic Publisher) 38

[16] Drummond C and Holte R 2004 What ROC curves can't do (and cost curves can) *ROCAI* 19–26

[17] Flach P and Wu S 2003 Repairing concavities in ROC curves *Proc. UK Workshop on Company Intel.* 38–44