Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 149–156
DOI: 10.15439/2022F262
ISSN 2300-5963 ACSIS, Vol. 30

About Classifiers Quality Assessment: Balanced Accuracy Curve (BAC) as an Alternative for ROC and PR Curve

Aleksandra Weiss, Marcin Młyński
Scientific Circle of Robotics UWM in Olsztyn
ul. Słoneczna 54, 10-710 Olsztyn, Poland
Email: {aleksandra.weiss28, marcinmlynski920}@gmail.com

Piotr Artiemjew
University of Warmia and Mazury,
in Olsztyn
ul. Słoneczna 54, 10-710 Olsztyn, Poland
Email: artem@matman.uwm.edu.pl

***Abstract*—In this work, we propose a new parameter to study the effectiveness of classifiers - the AUC (area under curve) of the balanced accuracy curve (BAC) on data with different balance degrees - we compare its effectiveness with the popular AUC parameters for the ROC and PR curve. We use a global kNN classifier with typical metrics to verify the utility of the new parameter. BAC, ROC and PR curves generate similar results, the advantage of BAC is its simplicity of implementation and ease of interpretation of results.**

I. Introduction

CLASSIFICATION accuracy is the most natural parameter for assessing classification quality. Total accuracy fails when test data are unbalanced in terms of class sizes. With help comes a balanced version of accuracy, which is simply the efficiency of the average across all test classes. Thanks to this parameter, even small test classes are equally taken into account in the classification. In this work, we present a preliminary verification of whether the AUC of a balanced accuracy curve applied on training data balanced in different levels can create a valuable competitive parameter for PR and ROC curve. Let us turn to a brief review of the literature on the topic under discussion. First of all, the parameters that we discuss in this paper concern the evaluation of binary classifiers. Let us introduce the basic notation. The following symbols shall be used: $T = True, P = Positive, F = False, N = Negative$. TP is the number of test objects from the positive class that were correctly classified. FP is the number of test objects from the negative class that were classified into the positive class. FN is the number of objects in the positive class that have been classified in the negative class. A receiver operating characteristic curve (ROC) is a chart that shows the predictive capability of a binary classifier as its threshold of discrimination evolves. The method was initially designed for military radar receiver operators in 1941, resulting in its name [1], [2]. The ROC curve shows the ratio of TP to FP values through the prism of thresholds determining membership to a positive class. The best value of the ROC curve is closest to the upper left corner of the plot. A Precision Recall curve (PR) [4], [3] is basically a chart with the Precision values on the y axis and the Recall on the x axis. $precision = \frac{TP}{TP+FP}$ Precision is understood as the accuracy of the classification of a positive class - the percentage of correctly classified objects in that class out of those classified. $recall = \frac{TP}{TP+FN}$ The Recall parameter tells us about the relevance to the positive class - it specifies the percentage of correctly classified objects in the positive class in relation to all classified objects to the positive class. The best value of the PR curve is closest to the top right corner of the plot. An extensive introduction of the relationship of ROC and PR curves can be seen in papers [5] and [6]. In paper [7], the author leads a discussion on the imbalance of decision classes vs PR curve. The literature review related to classification quality assessment is enormous, we will not even attempt to review it in this paper, for further reading the reader is directed to e.g. paper [12].

In the following sections we have the following content. In Section II we present the research methodology. In section III we present the results of the experiments. In section IV we summarise the work and indicate further research plans. Let us move on to discuss the methodology used in this thesis.

II. Methodology

In this section we discuss how we implemented the ROC, PR, BAC curves and introduce information about the classifier used.

A. Basic classifier

In testing the effectiveness of AUC values for BAC, ROC and PR curve, we used the kNN classifier. Which does not mean that there is any restriction on the use of other classifiers. The one chosen is an initial reference point. The procedure used is as follows.

Step 1. We input a training decision system (U^{trn}, A, d) and a test decision system (U_{tst}, A, d), where A is a set of conditional attributes, d a decision attribute.

Step 2. The classification of test objects using training objects is carried out as follows.

Across all conditional attributes $a \in A$, training objects $v \in U^{trn}$ and test objects $u \in U_{tst}$, we calculate the importance weights $w(u, v)$ using selected metric.

In the version used, the obvious limitation k is the size of the training system.

The test object u is classified using the weights computed for all training objects v. The weights are ordered in ascending order as,

$$w_1(u, v_1) \leq w_2(u, v_2) \leq \ldots \leq w_{|U_{trn}|}(u, v_{|U_{trn}|})$$

Using the calculated and sorted weights, the training decision classes vote with the following parameter, where c is over the decision classes in the training set,

$$Class_{weight_c}(u) = \sum_{i=1}^{k} w_i^c(u, v_i^c).$$

Eventually, the test object u is categorized into the class c with the smallest value $Class_{weight_c}(u)$.

Assume that the positive class is denoted by 1 and the negative class by 0. Once all test objects u have been classified, the quality parameter accuracy, acc is calculated, according to the equation

$$acc_{balanced} = \frac{acc_{class_1} + acc_{class_0}}{2}$$

$$acc_{class_c} = \frac{|correctly\ classified\ objects\ in\ class_c|}{|classified\ objects\ in\ class_c|}$$

B. BAC curve

We propose the AUC of the balanced accuracy curve as a new factor for cross-sectional assessment of classification quality. The curve is constructed from balanced accuracy values using training systems with varying levels of class balance. We divide the analyzed data into test and training datasets and determine which of the two possible classes is a positive class and which is a negative class. After classifying each decision class, we create a test set containing 30% of the total number of objects. From the remaining 70% (excluding the data contained in the test set), we create a training set. The splitting and classification procedure is carried out a hundred times - we include the average accuracy value as the final result. In the case of BAC, we use a 0.5 threshold for classification. The class balance levels (of training decision system) we use are: (10 : 90, 20 : 80, 30 : 70....90 : 10). The use of BAC is possible when the decision classes are adequately represented in terms of size, since we require the creation of subsets with different levels of balance used data should have access to the same number of objects in the classes. In other words, in the case of data that is highly unbalanced, where one of the important classes is small the use of BAC can be difficult.

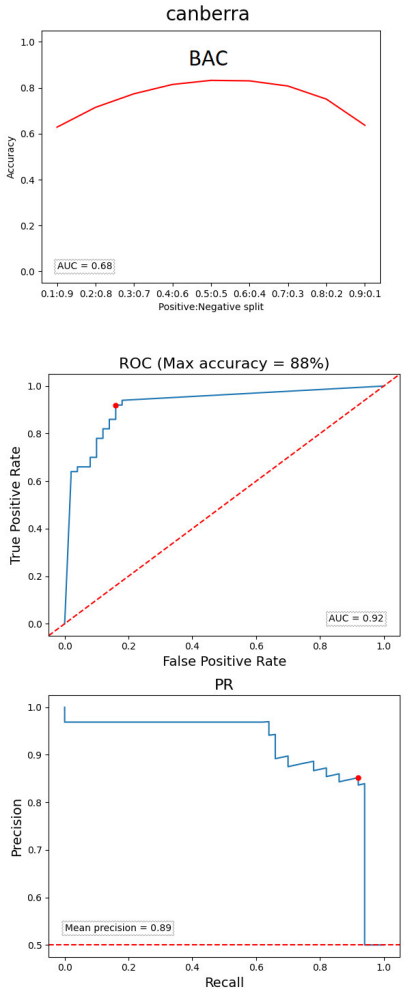

Fig. 1. An exemplary detailed result for Australian Credit decision system. From the top, BAC, ROC, and PR curves. The metric used is Canberra defined as follows: $d(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$

C. ROC and PR curve

We chose the AUC of the ROC and PR curves as reference parameters. We divide the analyzed data into test and training datasets and determine which of the two possible classes is the positive class and which is the negative class. After determining each decision class, we select 50 random objects from each class and combine the objects into one test set containing 100 objects. Then using the remaining objects from the entire dataset, we create a training set. We use the classification probabilities obtained after applying the kNN classifier to obtain the optimal threshold visualized by ROC curve and PR curve. We compare the probabilities with thresholds in the range from 0 to 1 with a step=0.001. If the probability is greater than or equal to the threshold, then we classify the object into the positive class, otherwise into the negative class. For each threshold, we calculate the coefficients of true positives (sensitivity), false positives and the accuracy

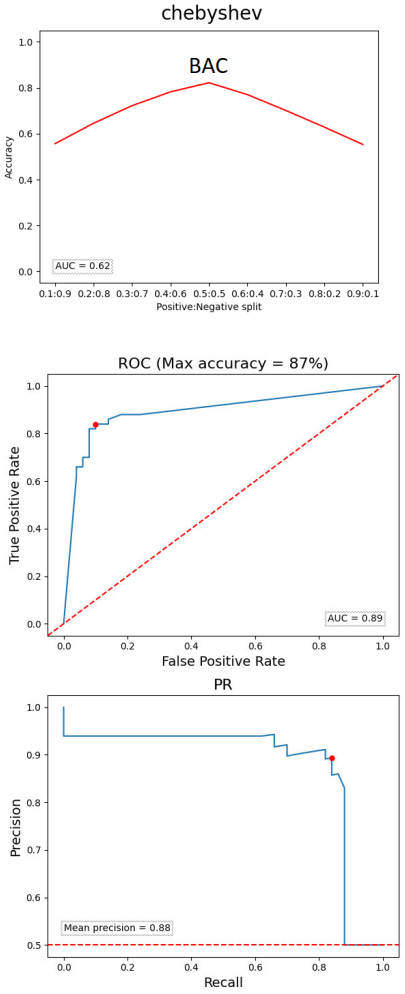

Fig. 2. An exemplary detailed result for Australian Credit decision system. From the top, BAC, ROC, and PR curves. The metric used is Chebyshev distance defined as follows: $d(x,y) = \max_i(|x_i - y_i|)$

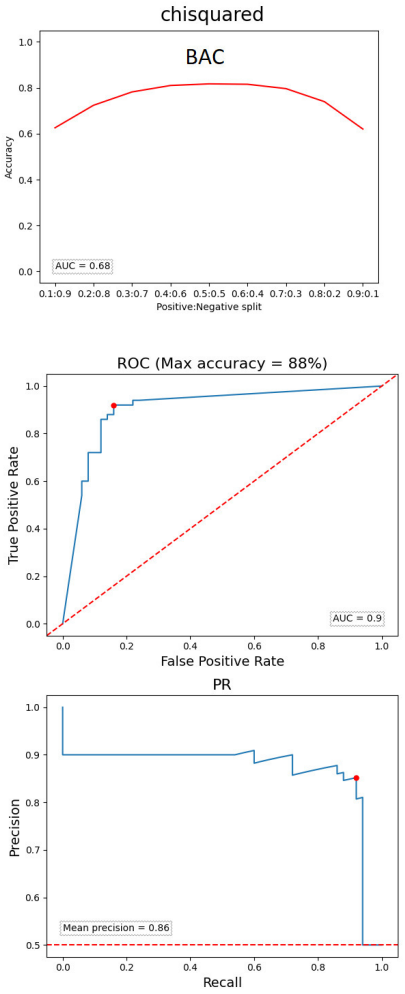

Fig. 3. An exemplary detailed result for Australian Credit decision system. From the top, BAC, ROC, and PR curves. The metric used is Chi-squared distance defined as follows: $d(x,y) = \sum_{i=1}^{n} \left(\frac{|x_i - y_i|}{|x_i| + |y_i|}\right)^2$

(precision) of classifying objects. The example detailed results for the ROC and PR curves, due to the difficulty of showing the mean result are from single classifications. The results summarising the performance of the curves are the average of a hundred times of the experiment.

III. Experimental session

In the experimental part we use the kNN method (See description in section II-A) with metrics: manhattan, euclidean, manhattan-maxmin, cosine, minkowski, chebyshev, canberra, chi-square, jaccard, epsilonHamming, sørensen. Definitions of each metric are provided in the headings of Figures 1 to 11. For the selected classifier, custom implementations for creating BAC, ROC and PR curves were written. We verify parameter performance on three decision systems selected from the UCI repository [15] - Australian Credit, Pima Indians Diabetes and Heart Disease datasets. Detailed results showing all three curves are presented for the Australian Credit decision system only - see figures from 1 to 11. A summary of metrics ranking generation based on AUC of BAC, ROC and PR curves for Australian Credit, Pima Indians Diabetes and Heart Disease systems can be seen in figures 12, 13 and 14.

A. Result summary

In Figures 12, 13 and 14 we have the rankings of the kNN method metrics based on the AUC of the BAC, ROC and PR curves - for Australian Credit, Heart Disease and Pima Indians Diabetes datasets respectively.

Results for Australian Credit decision system : AUC range of three positions for BAC and ROC curve agree for the metrics: canberra, euclidean, manhattan, minkowski, sorensen and epsilonHamming. AUC range of three positions for BAC and PR curve agree for the metrics: canberra,euclidean, minkowski, sorensen, epsilonHamming.

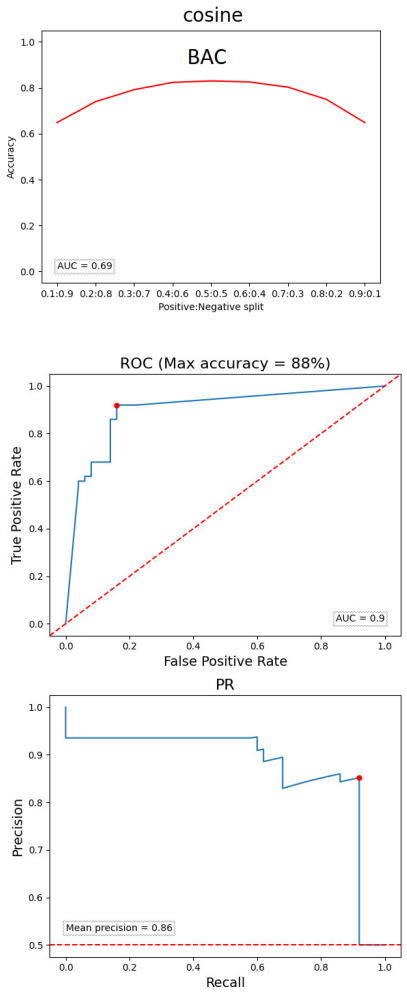

Fig. 4. An exemplary detailed result for Australian Credit decision system. From the top, BAC, ROC, and PR curves. The metric used is Cosine distance defined as follows: $d(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$

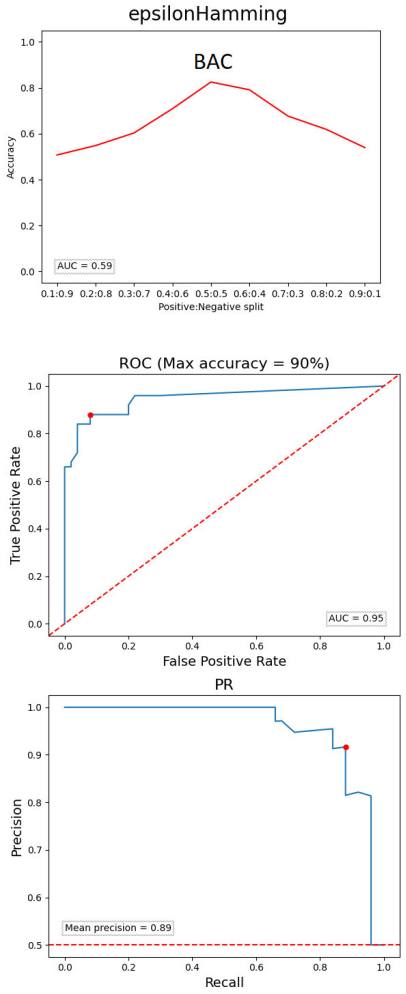

Fig. 5. An exemplary detailed result for Australian Credit decision system. From the top, BAC, ROC, and PR curves. The metric used is epsilonHamming distance defined as follows: $d(x,y) = |\{a \in A : dist(a(x_i), a(y_i)) \geq \varepsilon\}|, \varepsilon = 0.01$

Results for Heart Disease decision system: AUC range of three positions for BAC and ROC curve agree for the metrics: canberra, cosine, chisquared, manhattan, maxmin, euclidean,minkowski, chebyshev, epsilonHamming and sorensen. AUC range of three positions for BAC and PR curve agree for the metrics: jaccard, canberra, manhattan, maxmin, chisquared, minkowski, epsilonHamming, sorensen.

Results for Pima Indians Diabetes decision system: AUC range of three positions for BAC and ROC curve agree for the metrics: jaccard, cosine, chisquared, maxmin, sorensen and epsilonHamming. AUC range of three positions for BAC and PR curve agree for the metrics: jaccard, manhattan, chisquared, maxmin, minkowski, sorensen and epsilonHamming. The overall shape of the ranking indicates that the BAC, ROC and PR curves are comparable tools for assessing classification quality.

When analyzing the results for the three selected decision systems (Australian Credit, Heart Disease, and Pima Indians Diabetes), we see that the area under the balanced accuracy curve for the entire training system balance spectrum can be a competitive factor for determining the quality of classifiers to ROC and PR curve. The ranking results are similar to each other. Some metrics are mixed among themselves because the data are randomly selected, but there are clear common features to these rankings. The shape of the curve showing the ranking of metrics is similar. The main advantage of using the field under the accuracy curve is the simplicity of implementation and the ease of understanding the resulting factor. Which is simply the cross-sectional stability of the classifier for training knowledge balanced in different levels.

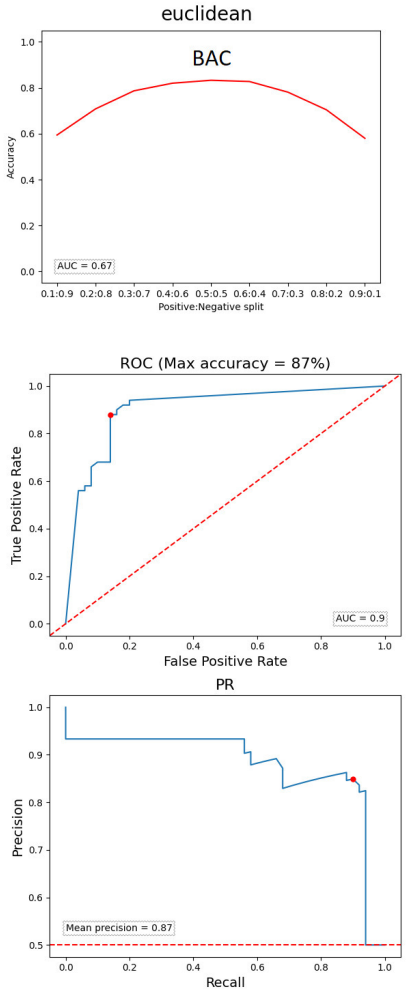

Fig. 6. An exemplary detailed result for Australian Credit decision system. From the top, BAC, ROC, and PR curves. The metric used is euclidean distance defined as follows: $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

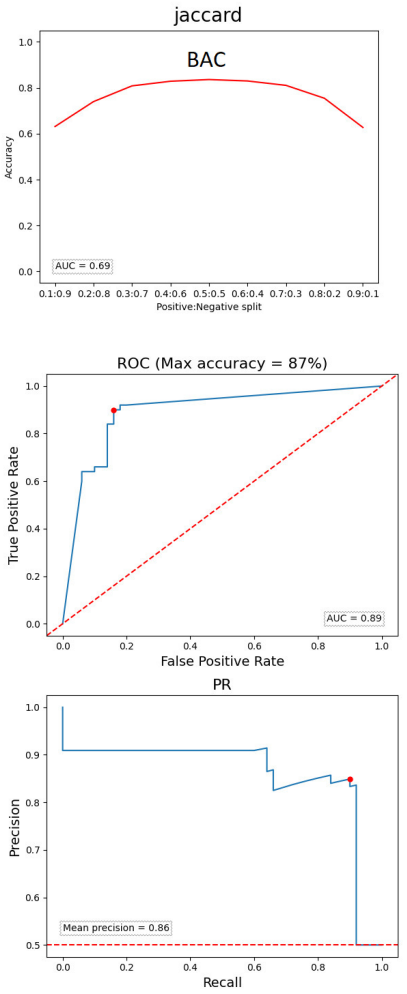

Fig. 7. An exemplary detailed result for Australian Credit decision system. From the top, BAC, ROC, and PR curves. The metric used is jaccard distance defined as follows: $d(x,y) = 1 - \left| \frac{\sum_{i=1}^{n}(x_i * y_i)}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n}(x_i * y_j)} \right|$

IV. Conclusion

This paper examines the applicability of the field under the accuracy curve-defined as the set of classification accuracies using training systems with different levels of decision class balancing-to the cross-sectional evaluation of the kNN classifier. We used the global method (k is selected from the entire system at once) with different metrics as reference kNN variants. By seeing the results, we can summarise that the specified factor is competitive with the ROC and PR curve, with its implementation and interpretation being much simpler and more understandable. Preliminary results show a similar gradation of methods, the difference in the performance of the metrics is due to the fact that the data were randomly selected, but the overall ranking looks similar. We consider our results as an initial step to open a discussion on the potential applicability of BAC to assess the stability of classifiers. We plan to extend our research to the entire spectrum of classification methods in the future.

Acknowledgment

This work has been supported by the grant from Ministry of Science and Higher Education of the Republic of Poland under the project number 23.610.007-000

References

[1] Woodward, P. M. (1953). Probability and information theory with applications to radar. London: Pergamon Press.
[2] Peterson, W., Birdsall, T., Fox, W. (1954). The theory of signal detectability, Transactions of the IRE Professional Group on Information Theory, 4, 4, pp. 171 - 212.
[3] Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. MIT Press
[4] Raghavan, V., Bollmann, P., & Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans. Inf. Syst., 7, 205–229.

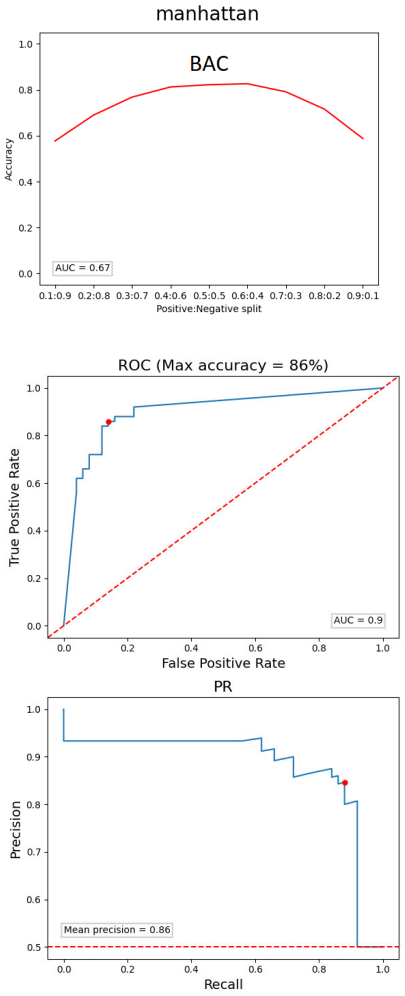

Fig. 8. An exemplary detailed result for Australian Credit decision system. From the top, BAC, ROC, and PR curves. The metric used is manhattan distance defined as follows: $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$

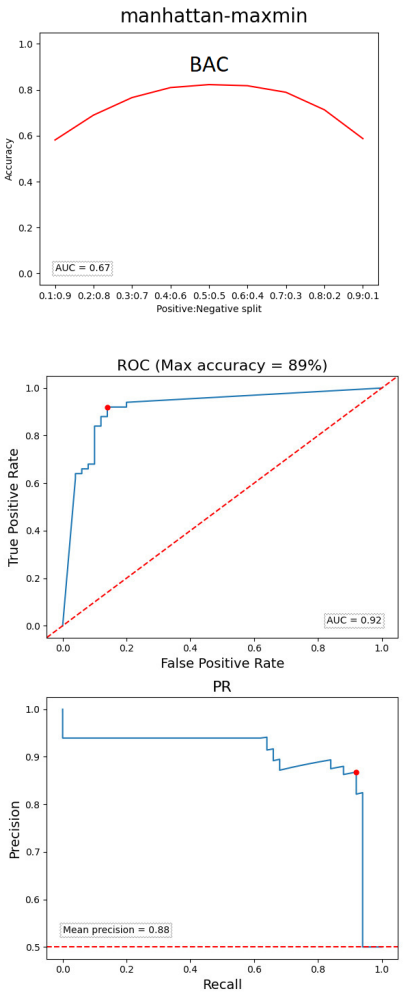

Fig. 9. An exemplary detailed result for Australian Credit decision system. From the top, BAC, ROC, and PR curves. The metric used is manhattan-maxmin distance defined as follows: $d(x,y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{|x_{ij} - y_{ij}|}{maxmin_j}$

[5] Davis, J., Goadrich, M.: 2006. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (ICML '06). Association for Computing Machinery, New York, NY, USA, 233–240. https://doi.org/10.1145/1143844.1143874

[6] Saito T., and Rehmsmeier M. 2015. "The Precision-Recall Plot Is More Informative Than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." PLoS ONE. 10(3): e0118432

[7] Williams, C.K.I. 2021. "The Effect of Class Imbalance on Precision-Recall Curves." Neural Computation 33(4): 853–857.

[8] Morzy, Tadeusz. Eksploracja danych. Red. . Warszawa: Wydawnictwo Naukowe PWN, 2013, 566 s. ISBN 978-83-01-17175-9

[9] Hastie T., Friedman J., Tibshirani R. (2001) The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY.

[10] Qimin Cao, Lei La, Hongxia Liu, and Si Han. Mixed Weighted KNN for Imbalanced Datasets [J]. Int J Performability Eng, 2018, 14(7): 1391-1400.

[11] L., Polkowski, P., Artiemjew, "Granular Computing in Decision Approximation - An Application of Rough Mereology," in: Intelligent Systems Reference Library 77, Springer, ISBN 978-3-319-12879-5, **2015**, pp. 1-422.

[12] Japkowicz, N., & Shah, M. (2011). Evaluating Learning Algorithms: A Classification Perspective. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511921803

[13] Metrics definition: manhattan, euclidean, canberra, cosine https://www.itl.nist.gov/div898/software/dataplot/homepage.htm

[14] epsilonHamming Metric definition: In: Polkowski, L., Artiemjew, P.: Granular Computing in Decision Approximation - An Application of Rough Mereology, In: Intelligent Systems Reference Library 77, Springer, ISBN 978-3-319-12879-5, pp. 1–422 (2015).

[15] UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/index.php. Last accessed 12 Apr 2022

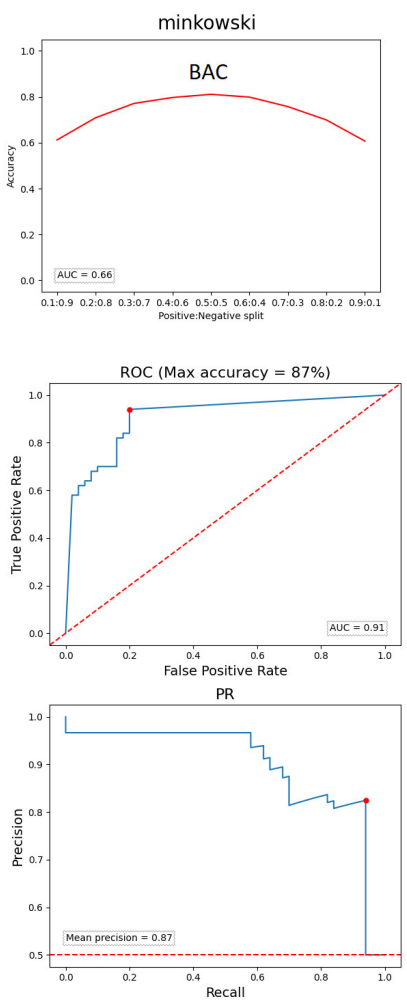

Fig. 10. An exemplary detailed result for Australian Credit decision system. From the top, BAC, ROC, and PR curves. The metric used is Minkowski distance defined as follows: $d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}, p = 3$

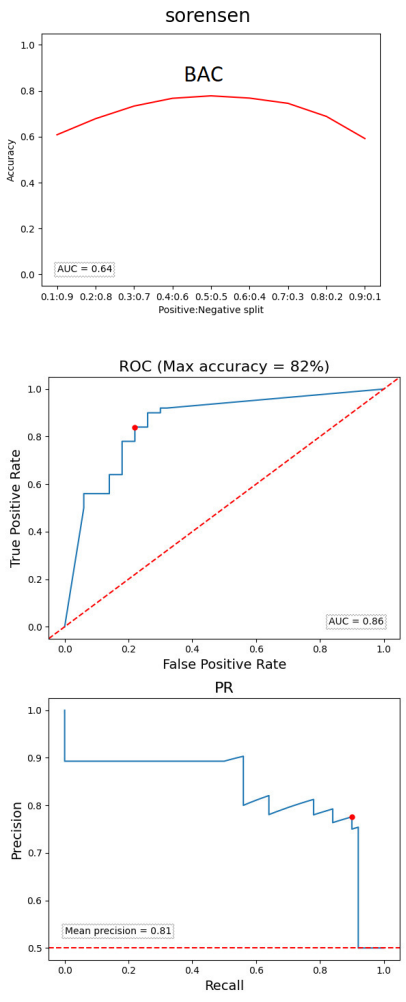

Fig. 11. An exemplary detailed result for Australian Credit decision system. From the top, BAC, ROC, and PR curves. The metric used is sorensen distance defined as follows: $d(x,y) = 1 - \left|\frac{2*\sum_{i=1}^{n}(x_i*y_i)}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2}\right|$

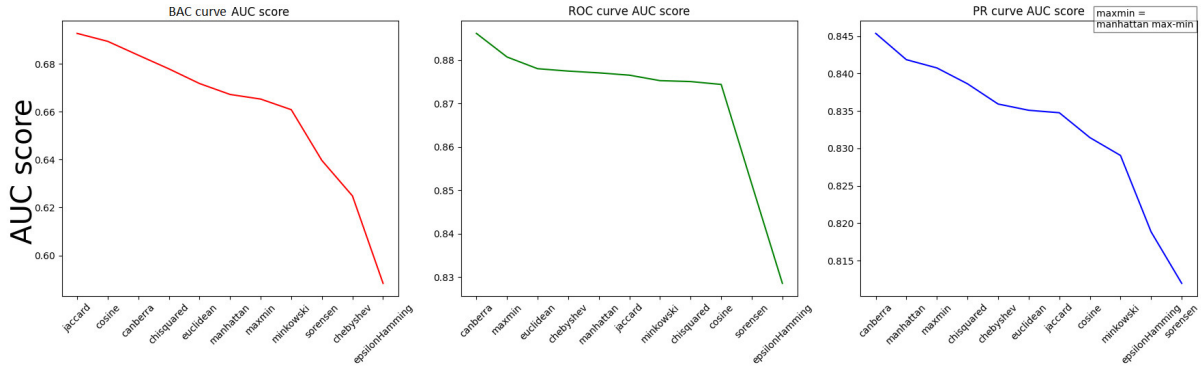

Fig. 12. Ranking of metrics for the Australia Credit data set. Comparison of rankings for AUC of BAC, ROC and PR curves.

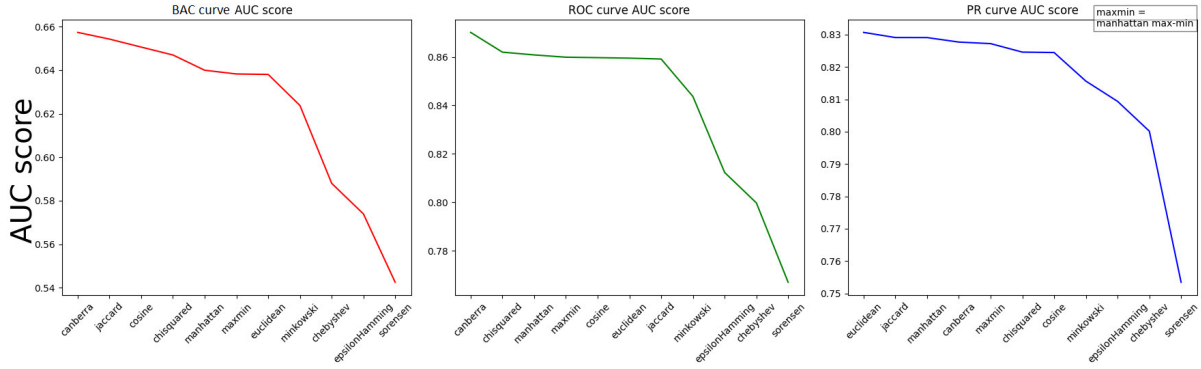

Fig. 13. Ranking of metrics for the Heart Disease data set. Comparison of rankings for AUC of BAC, ROC and PR curves.

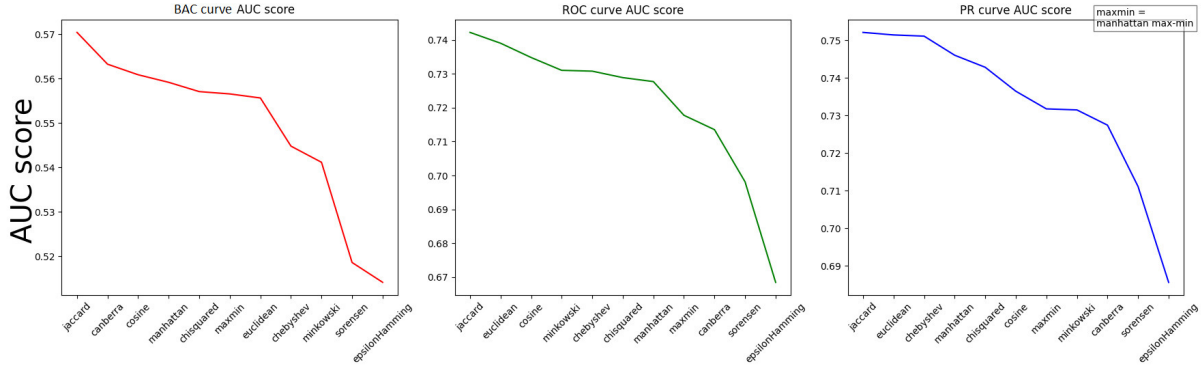

Fig. 14. Ranking of metrics for the Pima Indians Diabetes data set. Comparison of rankings for AUC of BAC, ROC and PR curves.