RESEARCH NOTE

REVISED Optimal threshold estimation for binary classifiers using game theory [version 2; referees: 2 approved, 1 approved with reservations]

Ignacio Enrique Sanchez

Protein Physiology Laboratory, University of Buenos Aires, Buenos Aires, Argentina

v2 **First published:** 25 Nov 2016, **5**(ISCB Comm J):2762 (doi: 10.12688/f1000research.10114.1)

Second version: 15 Dec 2016, **5**(ISCB Comm J):2762 (doi: 10.12688/f1000research.10114.2)

Latest published: 08 Feb 2017, **5**(ISCB Comm J):2762 (doi: 10.12688/f1000research.10114.3)

Abstract

Many bioinformatics algorithms can be understood as binary classifiers. They are usually trained by maximizing the area under the receiver operating characteristic (*ROC*) curve. On the other hand, choosing the best threshold for practical use is a complex task, due to uncertain and context-dependent skews in the abundance of positives in nature and in the yields/costs for correct/incorrect classification. We argue that considering a classifier as a player in a zero-sum game allows us to use the minimax principle from game theory to determine the optimal operating point. The proposed classifier threshold corresponds to the intersection between the *ROC* curve and the descending diagonal in *ROC* space and yields a minimax accuracy of 1-FPR. Our proposal can be readily implemented in practice, and reveals that the empirical condition for threshold estimation of "specificity equals sensitivity" maximizes robustness against uncertainties in the abundance of positives in nature and classification costs.

This article is included in the International Society for Computational Biology Community Journal gateway.

This article is included in the ISCB Latin America Conference on Bioinformatics collection.

This article is included in the Machine learning: life sciences collection.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
REVISED version 3 published 08 Feb 2017		✓	✓ report
REVISED version 2 published 15 Dec 2016			? report
version 1 published 25 Nov 2016	✓ report	✓ report	

1 **Pieter Meysman**, University of Antwerp Belgium

2 **Luis Diambra**, Universidad Nacional de La Plata (UNLP-CONICET) Argentina

3 **Marcel Brun**, National University of Mar del Plata Argentina, **Arjen ten Have**, , Universidad Nacional de Mar del Plata Argentina

Discuss this article

Comments (1)

Corresponding author: Ignacio Enrique Sanchez (isanchez@qb.fcen.uba.ar)

Competing interests: No competing interests were disclosed.

How to cite this article: Sanchez IE. **Optimal threshold estimation for binary classifiers using game theory [version 2; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2016, **5**(ISCB Comm J):2762 (doi: 10.12688/f1000research.10114.2)

Grant information: ANPCyT [PICT 2012-2550]. IES is a CONICET career investigator.

First published: 25 Nov 2016, **5**(ISCB Comm J):2762 (doi: 10.12688/f1000research.10114.1)

REVISED **Amendments from Version 1**

We have prepared a revised version of the article that aims at answering the concerns raised by the reviewers. We think our work is improved thanks to the comments. A point by point response follows.

Reviewer Luis Diambra raises two main concerns.

1. This work should put the problem in the context of supervised versus non-supervised learning.

We have modified the first paragraph in the introduction to clarify this point.

2. The problem of generalizing a learning rule/score deduced from a training set.

We have modified the first paragraph of the discussion to take this point into account.

Reviewer Pieter Meysman raises two main concerns.

1. Loss of generality in the specific case a=d=1 and b=c=0.

Following the work of Peter Flach (reference 4), I would argue that the uncertainty in the values of a, b, c and d is equivalent to the uncertainty in the values of qP and qN, so considering only the class ratio is enough for the point made in the article. The second paragraph of the Theory section was modified accordingly.

2. Nature not being a conscious agent as assumed in game theory.

The reviewer is right. We clarified this in the third paragraph of the Theory section.

See referee reports

Introduction

Many bioinformatics algorithms can be understood as binary classifiers, as they are used to investigate whether a query entity belongs to a certain class[1]. Supervised learning trains the classifier by looking for the rule that gives the correct answers to a training set of question-answer pairs. On the other hand, score-based binary classifiers are trained in a non-supervised manner. Such classifiers use a scoring function that assigns a number to the query. During training, the scoring function is modified to give different scores to the positives and negatives in the training set. After training, the classifier is used by assigning a query to the class under consideration if its score surpasses a threshold. A minority of users are able to choose a threshold using their understanding of the algorithm, while the majority uses the default threshold.

Binary classifiers are often trained and compared under a unified framework, the receiver operating characteristic (*ROC*) curve[2]. Briefly, classifier output is first compared to a training set at all possible classification thresholds, yielding the confusion matrix with the number of true positives (*TP*), false positives (*FP*), true negatives (*TN*) and false negatives (*FN*) (Table 1). The *ROC* curve plots the true positive rate ($TPR = TP/(TP + FN)$), also called sensitivity,) against the false positive rate ($FPR = FP/(FP + TN)$), which equals 1-specificity) (Figure 1, continuous line). Classifier training often aims at maximizing the area under the *ROC* curve, which amounts to maximizing the probability that a randomly chosen positive is ranked before a randomly chosen negative[2]. This summary statistic measures performance without committing to a threshold.

Practical application of a classifier requires using a threshold-dependent performance measure to choose the operating point[1,3]. This is in practice a complex task because the application domain may be skewed in two ways[4]. First, for many relevant bioinformatics problems the prevalence of positives in nature $q_P = (TP + FN)/(TP + TN + FP + FN)$ does not necessarily match the training set q_P and is hard to estimate[2,5]. Second, the yields (or costs) for correct and incorrect classification of positives and negatives in the machine learning paradigm (Y_{TP}, Y_{TN}, Y_{FP}, Y_{FN}) may be different from each other and highly context-dependent[1,3]. Points in the *ROC* plane with equal performance are connected by iso-yield lines with a slope, the skew ratio, which is the product of the class skew and the yield skew[4]:

$$SKEW\ RATIO = \frac{q_N \cdot (Y_{FP} + Y_{TN})}{q_P \cdot (Y_{TP} + Y_{FN})} \quad (1)$$

The skew ratio expresses the relative importance of negatives and positives, regardless of the source of the skew[4]. Multiple threshold-dependent performance measures have been proposed and discussed in terms of skew sensitivity[3,4], but often not justified from first principles.

Table 1. Confusion matrix for training of a binary classifier. TP: Number of true positives. FP: Number of false positives. FN: Number of false negatives. TN: Number of true negatives.

		Training set	
		p	**n**
Classifier output	**p'**	*TP*	*FP*
	n'	*FN*	*TN*

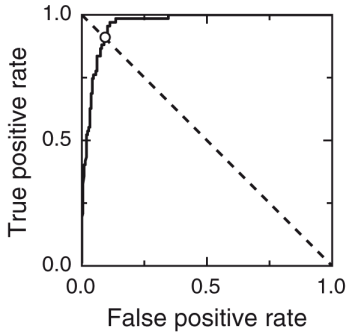

Figure 1. Optimal threshold estimation in *ROC* space for a binary classifier using game theory. The descending diagonal $TPR = 1 - FPR$ (dashed line) minimizes classifier performance with respect to q_P. The intersection between the receiver operating characteristic (*ROC*) curve (continuous line) and this diagonal maximizes this minimal, worst-case utility and determines the optimal operating point according to the *minimax* principle (empty circle).

Theory

Game theory allows us to consider a binary classifier as a zero-sum game between nature and the classifier[6]. In this game, nature is a player that uses a mixed strategy, with probabilities q_P and q_N=1-q_P for positives and negatives, respectively. The algorithm is the second player, and each threshold value corresponds to a mixed strategy with probabilities p_P and p_N for positives and negatives. Two of the four outcomes of the game, *TP* and *TN*, favor the classifier, while the remaining two, *FP* and *FN*, favor nature. The game payoff matrix (Table 2) displays the four possible outcomes and the corresponding classifier utilities *a*, *b*, *c* and *d*. The *Utility* of the classifier within the game is:

$$UTILITY = \frac{a \cdot TP + d \cdot TN + b \cdot FP + c \cdot FN}{TP + TN + FP + FN} \qquad (2)$$

The payoff matrix for this zero-sum game corresponds directly to the confusion matrix for the classifier, and the game utilities *a*, *b*, *c*, *d* correspond to the machine learning yields Y_{TP}, Y_{FP}, Y_{FN}, Y_{TN}, respectively (Table 1). In our definition of the skew ratio, the uncertainty in the values of *a*, *b*, *c* and *d* is equivalent to the uncertainty in the values of q_P and q_N[4]. Thus, we can study the case a=d=1 and b=c=0 without loss of generality[4]. Classifier *Utility* within the game then reduces to the *Accuracy* or fraction of correct predictions[2–4]. In sum, maximizing the *Utility* of a binary classifier in a zero-sum game against nature is equivalent to maximizing its *Accuracy*, a common threshold-dependent performance measure.

We can now use the *minimax* principle from game theory[6] to choose the operating point for the classifier. This principle maximizes utility for a player within a game using a pessimistic approach. For each possible action a player can take, we calculate a worst-case utility by assuming that the other player will take the action that gives them the highest utility (and the player of interest the lowest). The player of interest should take the action that maximizes this minimal, worst-case utility. Thus, the *minimax* utility of a player is the largest value that the player can be sure to get regardless of the actions of the other player. In our case, nature is not a conscious player that chooses the action that gives them the highest utility. We instead understand our application of the minimax principle as the consideration of a worst-case scenario for the skew ratio.

Table 2. Payoff matrix for a zero-sum game between nature and a binary classifier.
a: Player I utility for a true positive. *b*: Player I utility for a false positive. *c*: Player I utility for a false negative. *d*: Player I utility for a true negative.

		Player II: Nature	
		p	**n**
Player I: Classifier	**p'**	*a*	*b*
	n'	*c*	*d*

In our classifier *versus* nature game, *Utility/Accuracy* of the classifier is skew-sensitive, depending on q_P for a given threshold[3,4]:

$$UTILITY = 1 - FPR + q_P \cdot (FPR + TPR - 1) \qquad (3)$$

The derivative of the *Utility* with respect to q_P is zero along the $TPR = 1 - FPR$ line in *ROC* space (Figure 1, dashed line). The derivative is negative below this line and positive above it, indicating that points along this line are minima of the *Utility* function with respect to the strategy q_P of the nature player. According to the *minimax* principle, the classifier player should operate at the point along the $TPR = 1 - FPR$ line that maximizes *Utility*. In ROC space, this condition corresponds to the intersection between the *ROC* curve and the descending diagonal (Figure 1, empty circle) and yields a *minimax* value of $1 - FPR$ for the *Utility*. It is worth noting that this analysis regarding class skew is also valid for yield/cost skew[4].

Discussion

We showed that binary classifiers may be analyzed in terms of game theory. From the *minimax* principle, we propose a criterion to choose an operating point for the classifier that maximizes robustness against uncertainties in the skew ratio, i.e., in the prevalence of positives in nature and in yield skew, i.e., the yields/costs for true positives, true negatives, false positives and false negatives. This can be of practical value, since these uncertainties are widespread in bioinformatics and clinical applications. However, it should be noted that this strategy assumes that a score optimized for a restricted training set is of general validity.

In machine learning theory, $TPR = 1 - FPR$ is the line of skew-indifference for *Accuracy* as a performance metric[4]. This is in agreement with the skew-indifference condition imposed by the *minimax* principle from game theory. However, to our knowledge, skew-indifference has not been exploited for optimal threshold estimation. Furthermore, the operating point of a classifier is often chosen by balancing sensitivity and specificity, without reference to the rationale behind[7]. Our game theory analysis shows that this empirical practice can be understood as a maximization of classifier robustness.

Competing interests

No competing interests were disclosed.

Grant information

ANPCyT [PICT 2012-2550]. IES is a CONICET career investigator.

Acknowledgements

I would like to thank Juan Pablo Pinasco and Francisco Melo for discussion.

References

1. Swets JA, Dawes RM, Monahan J: **Better decisions through science.** *Sci Am.* 2000; **283**(4): 82–7.
PubMed Abstract | Publisher Full Text
2. Fawcett T: **An introduction to ROC analysis.** *Pattern Recognit Lett.* 2006; **27**(8): 861–874.
Publisher Full Text
3. Okeh UM, Okoro CN: **Evaluating Measures of Indicators of Diagnostic Test Performance: Fundamental Meanings and Formulars.** *J Biomet Biostat.* 2012; **3**(1): 132.
Publisher Full Text
4. Flach PA: **The geometry of ROC space: understanding machine learning metrics through ROC isometrics.** *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003).* 2003; 194–201.
Reference Source
5. Tompa M, Li N, Bailey TL, *et al.*: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol.* 2005; **23**(1): 137–44.
PubMed Abstract | Publisher Full Text
6. Von Neumann J, Morgenstern O: **Theory of games and economic behavior.** 6th ed., USA: Princeton university press. 1955.
Reference Source
7. Carmona SJ, Nielsen M, Schafer-Nielsen C, *et al.*: **Towards High-throughput Immunomics for Infectious Diseases: Use of Next-generation Peptide Microarrays for Rapid Discovery and Mapping of Antigenic Determinants.** *Mol Cell Proteomics.* 2015; **14**(7): 1871–84.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ✔ ?

Version 2

Referee Report 16 December 2016

doi:10.5256/f1000research.11243.r18290

? **Marcel Brun**[1] , **Arjen ten Have** [2]

[1] Digital Image Processing Group, School of Engineering, National University of Mar del Plata, Mar del Plata, Argentina

[2] Instituto de Investigaciones Biológicas-CONICET, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina

The author presents an interesting solution to the problem of selecting the best threshold for a binary classifier which defines a bioinformatics algorithm, when the proportion of positive and negative samples in nature is unknown. His approach consists on using the minimax approach, considering that the goal is to maximize the gain/accuracy (in the specific the case where the yields are 1 for TP and TN, and 0 for FP and FN) in a worst case scenario. For that, the author describes the utility as a function of both the threshold and the nature proportion of negatives q_n.

Based on the analysis of page 3, given any action of the player (a point in the ROC curve, defined by a threshold), its worst performance (lower utility) will be obtained for some q_n value. The idea is to maximize this minimal, across all possible actions (pairs of FPR,TPR across the ROC curve). This is obtained when the point in the ROC complies with TPR+FPR=1.

The article is original in the sense that it proposes a valid requirement for the design, to be robust against the unknown state of nature (proportion of positives and negatives), deriving a threshold selection that attains that condition (FPR+TPR=1). Despite the originality of the approach, the article needs minor revisions. In particular it seems the manuscript deals with aspects of pattern recognition from the perspective of Bioinformatics. Although understandable, this is a pattern recognition problem and the work would improve significantly when the role of ROC curves on classifier design, and the mathematics leading to the main result are explained with more detail and improved formalization.
Below follows a more detailed list of issues that need to be addressed:

1. The author presents this article in a bioinformatics context that is never specified. For example, in the introduction the authors says "Many bioinformatics algorithms can be understood as binary classifiers, ...", and then refers to "score based binary classifiers". At no part of the article the author provides a list, or some examples, of such algorithms in bioinformatics. Since this is presented as a bioinformatics article, it would be good to include some of such algorithms, which surely were the inspiration for this work. Alternatively, this might be presented in the more mathematical setting of pattern recognition, although this might require more profound changes.

2. Also, it is clear that the problem studied in this article is the one where the classifier model is already defined, producing one ROC curve, and the only adjusting parameter the user has is the

threshold. Although this is a rather typical solution in computational biology, it should be mentioned that if the user were able to adjust the classifier, better results might be obtained. There are many references in literature of robust classifier design, some are indicated below.

3. The author should review the statement that ROC curves are often being used for the training of classifiers. The training of classifiers is usually done by minimization or maximization of quality metrics, via the adjustment of the classifier parameters, once the model is defined. Few works are related to training classifiers based on ROC (Some examples are Blockeel, 2002[1] and Srinivasan, 2001[2]). On the other hand, classifier evaluation (after training) and model selection are often done via ROC curves.

   For example, in the second paragraph of the introduction, the author comments that "Binary classifiers are often trained and compared under ... the ROC curve", presenting a reference on the ROC curve, but not to classifier training. Here the term "training" may be confusing, since it is used, usually, for the process of selecting the parameters of the classifiers, from the data, once the model is already selected[3]. Usually the ROC is used for model selection, where different classifier models are trained (without considering the ROC), and then compared based on their ROC, as described in the reference 2 of the paper.

   Based on the previous comment, the author should also review the sentence, in the same paragraph, where he states "Classifier training often aims at maximizing the area under the ROC". Actually, the training aims, usually, to minimize error, or cost, or other measures of accuracy. Reference 4 of the article, in the first paragraph of the introduction, also makes clear the difference between model building (training), and model evaluation, where the ROC is used.

4. In general, the assumption on a=d=1, c=b=0 could start earlier, simplifying the concepts and equations 1 and 2.

5. The mathematical proof that the minimax threshold is obtained by FPR+TPR=1 requires some elaboration. In one hand, given a classifier trained from data, with a variable threshold t, both FPR and TPR are functions of the threshold, so they would be FPR(t) and TPR(t). From this, the utility is dependent of two variables, q_n and t, so it would be Utility(q_n,t). Now, the problem of finding a minimax value of t would be a bit more complex than the solution described in the article. Still more, one should impose the restriction on the curve (TPR(t),FPR(t)) to be increasing, so that if t2>=t1, then TPR(t2)>=TPR(t1) and FPR(t2)>=TPR(t1). If such condition is held (which is expected on a properly designed ROC curve), then the proof that FPR(t)+FNR(t)=1 is the minimax solution is almost straightforward.

   Without that condition defined above (which is not defined in the paper), one could have a pair (0.8,0.2) which complies with FPR+TPR=1 and (0.81,0.18) that does not comply with that rule, and where the minimum utility (as function of q_n) for the first pair is 0.8, and for the second pair is 0.81, which disagree with the mathematical result presented in the paper. The key here is the increasing nature of the ROC curve. An increasing ROC curve (where TPR increases when FPR increases) could not contain these two points, since an increase in TPR (0.8 to 0.81) should be accompanied by an increase on the FPR (but it goes from 0.2 to 0.18).

6. Equation 1 uses q_n, which is only defined later, after the equation.

7. The author presents the work in the general context of zero-sum game, with general yield values, but for the main theoretical result, the problem is reduced to the specific situation of a=d=1 and b=c=0. Therefore, the result is not general, only specific to that situation, so it is a little superfluous to maintain all these generic yield values across the paper. For example, Equation 1, when using the costs a=d=1, c=b=0, studied in the theory section, reduces to skew ratio = q_n/q_p, which is more informative for the rest of the paper.

8. In the discussion the author talks about maximizing robustness, but actually the goal is to design "a robust classifier"[4]. The discussion section also states that the robustness is obtained against a) skew ratio and b) yield skew. This last part (b) is not proved in the paper, since the theoretical results show a minimax results (robustness) against skew ratio, for the specific case of a=d=1 and b=c=0.

   It seems true that no other methods for choosing the operating point present a good rationale beyond mathematical considerations. To strengthen the impact I suggest the author should at least name some of such approaches, like the three ones described by Song, 2014.

9. In a situation of continuous ROC curve, this point should always exist, since it is an increasing curve from (0,0) to (1,1). In the discrete ROC case, which is the usual one when the classifier and the ROC are based on a finite amount of data, this intersection may not occur, and the thresholds may have to be selected using an appropriate interpolation technique. This needs to be clarified.

10. Finally there is a question about the approach. Despite the theoretical attractiveness of the proposed approach, it is not clear how the maximization of expected utility is a desirable quality for a classifier being used in bioinformatics. There is a lot of study and discussion about whether is it better to optimize accuracy, precision, recall, etc, and how each application may require a different goal, but there is not so much discussion about why a classifier should maximize utility in a zero-sum game. The author could address this issue, with a brief statement on the advantages of this measure (the utility) in the context of bioinformatics.

References

1. Blockeel H, Struyf J: Deriving Biased Classifiers for Better ROC Performance. *Slovene Society Informatika*. 2002; **26** (1).
2. Srinivasan A*Machine Learning*. 2001; **44** (3): 301-324 Publisher Full Text
3. Jain AK, Duin RPW, Mao J: Statistical Pattern Recognition: A Review. *IEE Transactions on Pattern Analysis and Machine Intelligence*. 2000; **22** (1).
4. Dougherty E, Hua J, Xiong Z, Chen Y: Optimal robust classifiers. *Pattern Recognition*. 2005; **38** (10): 1520-1532 Publisher Full Text

Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Version 1

Referee Report 07 December 2016

doi:10.5256/f1000research.10895.r17994

Luis Diambra
Centro Regional de Estudios Genómicos, Universidad Nacional de La Plata (UNLP-CONICET), La Plata, Argentina

The author presents a criterion to choose the operating point for a binary classifier. This criterion is analyzed in term of the game theory. By using the mininax principle author proposes to use as classifier threshold the intersection between the ROC curve and the descending diagonal in ROC space. This operating point for the classifier could maximizes the robustness against some bias in the training set. I found some novelty in the fact to consider such bias for an optimal threshold estimation. The paper is well written and organized but I think also that it could be improved by incorporating some general considerations that helping readers to a better understanding of the problem and the present proposition [1,2].

In the binary classification problem one is trying to deduce the answers to new questions, rather than just recall the answers to old ones. In order to do that we need to train the classifier from question-answer pairs (the training set). This is called supervised learning, because it requires a teacher, knowing the rule, which gives the correct answer to the example questions. In the case here, the author consider score-based binary classifiers, which does not need such learning stage. Could the author put the problem in the context supervised vs. no-supervised?

In the supervised learning context the classifier threshold is a parameter that is found during the learning stage. Training the classifier maximizing the area under ROC curve is an strategy for the classifier learn the training set. Consequently, the proposed strategy could be considered as a "learning rule". However, the performance over new examples is not guaranteed. Other point which can improve the manuscript would be to consider the ability of generalization of the proposed strategy. Could the author add a discussion in this sense?

I believe that this manuscript is of an acceptable scientific standard, and that it will be of interest to a wide audience; however, the manuscript could be revised, as outlined above.

References

1. Watkin T, Rau A, Biehl M: The statistical mechanics of learning a rule. *Reviews of Modern Physics*. 1993; **65** (2): 499-556 Publisher Full Text
2. Diambra L, Fernández J: Information theory approach to learning of the perceptron rule.*Phys Rev E Stat Nonlin Soft Matter Phys*. 2001; **64** (4 Pt 2): 046106 PubMed Abstract | Publisher Full Text

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 06 December 2016

doi:10.5256/f1000research.10895.r17996

Pieter Meysman
Department of Mathematics and Computer Science, University of Antwerp, Edegem, Belgium

The article by Ignacio Enrique Sanchez concerns a common problem in machine learning, namely the selection of the optimal classification threshold, and provides a mathematical solution based on the principles of game theory. The main concern of the article deals with the unknown distribution of positive and negative samples in the 'real world' or 'nature', thus beyond the provided training data set. The provided derivation is very elegant, and luckily for those researchers in the field the solutions turns out to be to select a threshold where sensitivity and specificity are equal in the training data set.

The biggest concern from the perspective of game theory is that 'nature' is not a conscious agent, and thus will not mischievously choose a positive/negative fraction where the classifier will perform the worst. However as stated in the article, this is to simulate the worst case scenario. However this also means that the threshold calculation may only be optimal in this worst case scenario, but suboptimal in all other cases. It is therefore still not the final word in threshold optimisation, and still leaves machine learning researchers the flexibility to choose other thresholds.

However I do have a minor comment on the derivation, that I expect can be addressed with small clarifications to the text:

1. The *Accuracy* equals the *Utility* as defined by the payoff matrix in the specific case a=d=1 and b=c=0, which is stated without a loss in generality. However in my understanding, this step makes the assumption that the cost for a false negative and the cost for a false positive is equal, which may not be the case for all classifiers. Thus it is unclear if this specific case can be transposed to all classifiers in general.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Discuss this Article

Version 1

Reader Comment 12 Dec 2016

Matúš Kalaš, Computational Biology Unit, Department of Informatics, University of Bergen, Norway

Just a note, perhaps an interesting complement is the article Saito & Rehmsmeier 2015 and website Classifier evaluation with imbalanced datasets.

Competing Interests: No competing interests were disclosed.