An approach to evaluate a classification model to predict a construction object's state

Olga Solovei[1,†], Bohdan Solovei[2,†], Yuliia Riabchun [3,†]

[1] *Kyiv National University of Construction and Architecture, Povitroflots'kyi Ave, 31, Kyiv, 03037, Ukraine*
[2] *Kyiv National University of Construction and Architecture, Povitroflots'kyi Ave, 31, Kyiv, 03037, Ukraine*
[3] *Kyiv National University of Construction and Architecture, Povitroflots'kyi Ave, 31, Kyiv, 03037, Ukraine*

Abstract

A classification model which is designed to predict technical conditions of the construction object must not make the mistakes in categorizing "unfit" for normal operation objects as "fit" as such mistakes could lead to the accidents with a wide range of severities. So, to guarantee the mode's results are correct, the scientists are doing model's evaluation by the metrics, which high score is the indicator of model's ability does not make the mistakes. However, on the question which metrics to be use to assess the model is not received a single answer as the researches' conclusions often contradict each other while recommending the metric. The goal of this study is to propose the approach on how to select the metric to assess a binary classification model for predicting the technical conditions of the construction object. To meet the goal in the research Matthews Correlation Coefficient formula and F-measure were described using maximized Youden index, which value is possible to obtain when model doesn't make the mistakes when predicting negative instance. The results of this work will provide the scientist the decision's support method which recommendation depends on the optimal cut-off point on the ROC curve, so it improves the accuracy of the received evaluation score.
.

Keywords

F-measure, Youden Index, Matthews Correlation Coefficient, optimal cut-off point on ROC curve

1. Introduction

Machine learning methods, specifically classification algorithms, are used to automate the process of the construction objects' technical inspection. Multiple classification model are built to put the objects depending on its' technical condition into one of four categories: "1" - normal; "2" - satisfactory; "3" - unfit for normal operation; "4" - emergency [1-2]. The ability of the classification model to determine correctly the technical conditions of the construction object is evaluated by metrics [3-6]. There are four metrics which are commonly accepted among the scientists, which formulas are the functions of the cells' values of a confusion matrix Table 1 (in case of supervise learning) or a matching matrix in the case of unsupervised learning:

ITTAP'2024: 4th International Workshop on Information Technologies: Theoretical and Applied Problems, October 23-25, 2024, Ternopil, Ukraine, Opole, Poland
[1*] Corresponding author.
[†] These authors contributed equally.
soloveiolga2@gmail.com (O. Solovei); bsolovei25@gmail.com (B. Solovei); Super.etsy@ukr.net (Y. Riabchun)
0000-0001-8774-7243 (O. Solovei); 0009-0008-0328-1123 (B. Solovei), 0000-0002-8320-4038 (Y. Riabchun)

1. Sensitivity (SEN) and Specificity (SPE) – the metrics to assess the model's ability to correctly identify positive and negative cases, respectively.
2. Positive Predictive Value (PPV) and Negative Predictive Value (NPV) – the metrics assess the model's error for positive and negative cases, respectively.

Table 1
2×2 confusion matrix

	Predicted positives	Predicted negatives	
True positives	TP	FN	P_i
True negatives	FP	TN	N_i
	P_j	N_j	

The values of other methods to evaluate the effectiveness of the classification model are derived from SEN, SPE, PPV, NPV, for example:

3. F-measure – the metric considers the model's ability to correctly identify positive cases and the model's error for positive cases. Its formula is a harmonic mean of SEN and PPV.
4. Matthews Correlation Coefficient (MCC) – this metric considers the model's ability to correctly identify positive cases and the model's errors for both positive and negative cases. Its formula considers all four parameters (SEN, SPE, PPV, NPV).

Recently, are submitted the studies which are showing that F-measure may not accurately reflect the model's true ability to classify objects as expected, and it is recommended to use the Matthews method for the evaluation instead [7-8]. However, such recommendation could not be used to evaluate model for prediction a construction object technical condition. Let's consider the confusion matrices of a binary classification model, create to categorize objects: "positive" - the objects fit for normal operations; "negative" – the objects unfit for normal operation. In Table 2 is considered 3 possible results of a binary classification and the notation to read Table 2 is:

1. True Positives (TP) - the number of "fit" objects whose conditions are correctly identified by the model.
2. True Negatives (TN) - the number of "unfit" objects whose conditions are correctly identified by the model.
3. False Positives (FP) - the number of "unfit" objects which are incorrectly identified by the model as "fit".
4. False Negatives (FN) - the number of "fit" objects which are incorrectly identified by the model as "unfit".

Table 2
A confusion matrix to evaluate a binary classification model

№	TP	FN	FP	TN
1	30	1	0	30
2	100	20	1000	30000
3	90000	10000	1	9

In row 1 of Table 2, is seen that the model incorrectly classified only one object, so model's score is high (>0.9) by all metrics (refer to Table 3). In row 2 of Table 2 - the model has a weak predictions in positive class PPV=100/(100+1000)=0.09 because 1000 "unfit" objects were treated as "fit" but model's ability to predict correctly negative is still high so NPV=30,000/(30,000+20)=0.99. As the result, the model score by normalized MCC is low but still may be accepted, however F-measure indicates the model's failed. In row 3 of Table 2, the model incorrectly classified 10,000 "fit" objects as "unfit", i.e. the model has a low ability to predict negative class as a result NPV=9/(9+10,000)=0.0 is low, how model's has a good prediction in positive class PPV = 9000/(9000+1)=1. As a result, F-measure score is high, but normalized MCC score dropped.

Table 3
Binary model's evaluation results

№	SEN	SPE	PPV	NPV	nMCC	F-measure
1	0.97	1.00	1.00	0.97	0.98	0.98
2	0.83	0.97	0.09	0.99	0.63	0.16
3	0.90	0.90	1.00	0.00	0.51	0.95

Since an effective binary classification model for prediction a construction object' technical conditions may make a mistake when "fit" object is considered as "unfit" (like 3) but must not consider "unfit" as "fit" (line 2), so value of F-measure is important and could not be ignored as recommended in [7-8] as in line 2 F-measure is much lower than MCC's value.

In current paper, will be proposed an approach to evaluate and recommend which nMCC or F-measure score to use while evaluation a binary classification model for predicting the technical conditions state of the construction object.

2. Literature Review

The relations between MCC and F-measure in researches [7-8] were studied with the values of TP, TN, FP, FN from the confusion matrix which is received when threshold equal to 0.5. However, cut-off threshold τ=0.5 is not obviously optimal and may never be selected by data scientists to get model's predictions for the given dataset. Therefore, the recorded differences in the scores of the evaluation metrics may not happen when the selected threshold corresponds to optimal cut-off point on ROC curve or difference's scale may be small. On Figure 1 is illustrated how the difference in the scores of normalized MCC and F-measure is changing depending on the ROC's curve threshold, where curves ware constructed with the result of 4 machine learning classification algorithms GaussianNB; Random Forest Classifier; Logistic Regression; Support Vector Classifier learned on the same data set and scaled by method specified in [9].

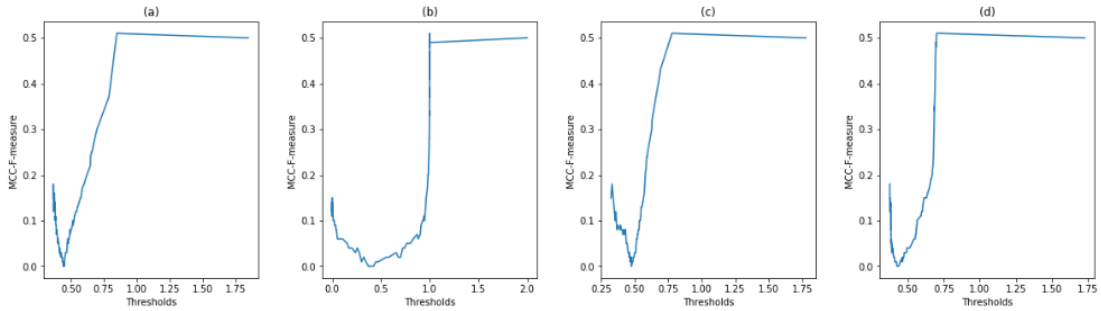

Figure 1: ROC curves of classifiers (a) – GaussianNB; (b) – Random Forest Classifier; (c) – Logistic Regression; (d) – Support Vector Classifier.

In our approach, instead of using the constant threshold of ROC curve to construct a confusion metric and get the values of normalized MCC and F-measure we will find an optimal point of ROC curve and use it to design a decision support algorithm.

3. Research Materials

Youden Index is frequently used to identify an optimal point of ROC curve [10-11]. When both SEN and SPE are given the same priority then Youden is calculated as the maximum difference between TPR and FPR (1).

$$Youden = \max_{\tau} \{SEN + SPE - 1\} = \max_{\tau} \{TPR + 1 - FPR - 1\} = \max_{\tau} \{TPR - FPR\} = mc \quad (1)$$

In equation (1), TP; FP; FN; TN are cell values of a confusion matrix, which values are used to calculate marginal values in Table 1, so that P_j= TP+ FP – is the number of positive instances which were predicted by model; N_j= FN+TN is the number of negative instances which were predicted by model; P_i=TP+FN is the number of positive instances in dataset; N_i=FP+TN is the number of negative instances in dataset.

At the same time the maximum Youden score can be achieved also when FP is zero. In this case, according to Table 1 TP=P_j and (1) is rewritten as (2).

$$Youden_{fp-¿0} = \frac{P_j}{P_i}, \quad (2)$$

As both equations (1) and (2) express the maximum value of Youden then they both equally identify the optimal point of the ROC curve, so in our proposed approach we take equation (2) and express MCC and F-measure using $Youden_{fp \to 0}$.

MCC equation (3) when FP is zero is rewritten as equation (4) according to which MCC is a function of two arguments: maximized Youden and a ration of the negatives in dataset to predicted negatives.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (3)$$

$$MCC_{fp-¿0} = \frac{P_j \cdot N_i}{\sqrt{P_j \cdot N_i \cdot P_i \cdot N_j}} = \sqrt{\frac{P_j \cdot N_i}{P_i \cdot N_j}} = \sqrt{\frac{P_j}{P_i}} \cdot \sqrt{\frac{N_i}{N_j}} = \sqrt{Youden_{fp-¿0}} \cdot \sqrt{\frac{N_i}{N_j}} \quad (4)$$

F-measure equation (5) when FP is zero is rewritten as equation (6) according to which F-measure is a function of maximized Youden.

$$F_1 = \frac{2TP}{2TP+FP+FN} \quad (5)$$

$$F_{1fp-¿0} = \frac{2P_j}{2P_j+P_i-P_j} = \frac{2P_j}{P_j+P_i} = \frac{2}{1+\frac{1}{Youden_{fp \to 0}}} \quad (6)$$

Having formally defined MCC and F-measure via maximized Youden gave us possibility to obtain objective differences in metric scores and use them in the proposed approach to select the most appropriate metric (Figure 2).

The algorithm includes 3 blocks: 1. “Define and Calculate measures”; 2. “Assess Significance of Differences”; 3. “Rank Metrics Based on Significance”. In the 1st block are executed calculations of the metrics and their differences compared to Youden are are saved. In the 2nd block – the differences saved as a result of execution of block 1 one by one is compared with the defined in block 1 the maximum expected difference, which is denoted as “threshold”. When the difference is less then the “threshold” it is marked as “notSignificant”. In last block all differences which are marked by flag “notSignificant” will be ranged and the metric with the smallest difference value will proposed by the algorithm. In case, when the algorithm did find any metric with flag “notSignificant” it will stop with no suggested metric to use.

4. Conclusions

In the current research the formal descriptions of the metrics: MCC and F-measure were described as the functions of Youden index, which is calculated on the confusion matrix when a binary model doesn’t make a mistake in predicting negative instances.

The obtained definitions gave us the possibility to calculate unbiased differences in the metric pair scores instead of using the differences which are received when ROC curve cut-off is equal to 0.5, which is not obviously optimal and may never be selected by data scientists to get model’s predictions.

The defined calculations are included in the proposed algorithm to support the scientists to select a metric in order to evaluate a binary classification model for predicting the technical conditions state of the construction object.

The novelty of the proposed approach compared to recommendations from [8-9] is that it recommends the metric considering optimal cut-off point of the ROC curve, so its recommendation depends on the leaning algorithm’s results, so it improves the accuracy of model’s evaluation.

Further work will be to extend the proposed algorithm by including more measures, such as balanced accuracy, diagnostic odds ratio (DOR) and others.

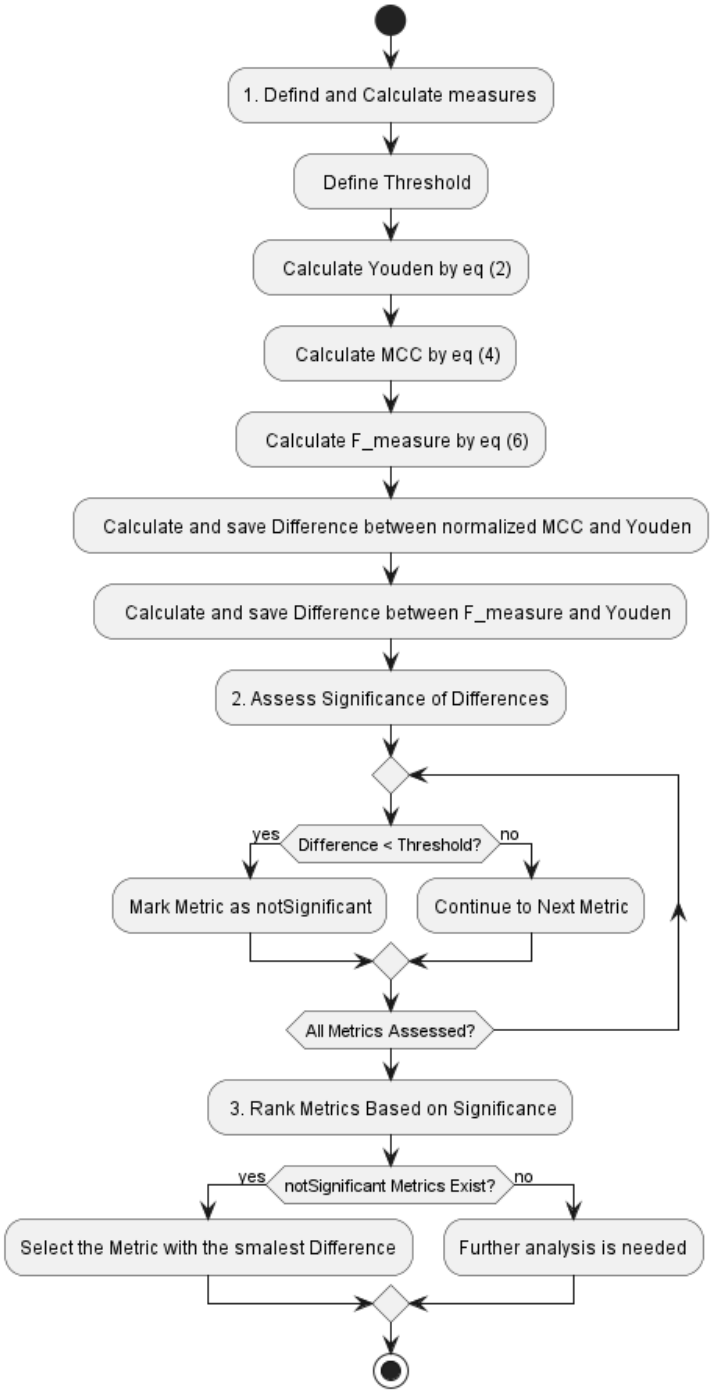

Figure 2: An algorithm to select the metric to evaluate a binary model

References

[1] G. Ryzhakova, O. Malykhina, V. Pokolenko, O. Rubtsova, O. Homenko, I. Nesterenko, T. Honcharenko, Construction project management with digital twin information system. International Journal of Emerging Technology and Advanced Engineering, volume 12(10), 2022, pp. 19-28. DOI: 10.46338/ijetae1022_03.

[2] D. Chernyshev; S. Dolhopolov; T. Honcharenko; H. Haman; T. Ivanova; M. Zinchenko, Integration of Building Information Modeling and Artificial Intelligence Systems to Create a Digital Twin of the Construction Site, IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), 2022, pp. 36-39. DOI: 10.1109/CSIT56902.2022.10000717.

[3] A. Tharwat, Classification assessment methods. Applied computing and informatics, volume 17(1), 2020, pp.168-192.

[4] M. Warrens, Kappa coefficients for dichotomous-nominal classifications. Advances in Data Analysis and Classification, volume 15(1), 2021, pp. 193-208.

[5] T. Honcharenko, R. Akselrod, A. Shpakov, O. Khomenko, Information system based on multi-value classification of fully connected neural network for construction management, IAES International Journal of Artificial Intelligence, volume 12(2), 2023, pp. 593-601.

[6] D. Dhamnetiya, RP. Jha, S. Shalini, K. Bhattacharyya, How to Analyze the Diagnostic Performance of a New Test? Explained with Illustrations. Journal of laboratory physicians, volume 14(01), 2021, pp 90-98.

[7] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, volume 21(1), 2020, pp. 1-13.

[8] D. Chicco, N. Tötsch, G. Jurman, The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData mining, volume 14(1), 2021, pp. 1-22.

[9] T. Honcharenko, O. Solovei, Optimal bin number for histogram binning method to calibrate binary probabilities. CEUR Workshop Proceedings, vol-3628, 2023, pp. 126–135.

[10] K. Garg, S. Campolonghi, A Step-by-Step Guide to Selecting an Optimal Cut-Off Value Based on the Receiver Operating Characteristic and Youden Index in Methods Designed to Diagnose Lyme Disease. InBorrelia burgdorferi: Methods and Protocols, New York, NY: Springer US., 2024, pp. 69-76.

[11] M. Hassanzad, K. Hajian-Tilaki, Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review. BMC Medical Research Methodology, volume 24(1), 2024.