Pattern Recognition Letters 136 (2020) 71–80

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset

Qiuming Zhu

Department of Computer Science, University of Nebraska at Omaha, Omaha, NE 68182, USA

ARTICLE INFO

Article history:
Received 28 September 2019
Revised 30 January 2020
Accepted 29 March 2020
Available online 29 May 2020

Keywords:
Matthews correlation coefficient
Classification accuracy measurement
Performance evaluation
Imbalanced dataset

ABSTRACT

The Matthews Correlation Coefficient (MCC) is one of the popular measurements for classification accuracy. It has been generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The study of this paper finds that this is not true. MCC deteriorates seriously when the dataset in classification are imbalanced. Experiment results and analysis show that MCC is not suitable for classification accuracy measurement on imbalanced datasets.

1. Introduction

While a confusion matrix records precisely the outcomes of a classification process, each of the components along does not properly represent the overall classification performance. Thus, a number of measurements derived from some combinations of the components have been proposed to represent the information in a single value and have been used to evaluate the performance of the classification processes [1,2]. However, among these derived measurements, some have been recognized as not suitable for measuring the classification results on imbalanced datasets, such as the Accuracy, Precision, G-mean of Precision and Recall, F1 score, etc. [3–6]. However, there existed some confusion on Matthews Correlation Coefficient (MCC) for whether it is suitable to imbalanced data or not [7,8].

MCC integrates the eight major derived ratios from the combinations of all the components of a confusion matrix, has been regarded as a good metric that represents the global model quality, and can be used even if the classes are of very different sizes [9–11].

This paper shows by experimental tests that MCC is inadequate for measuring classification outcomes on imbalanced datasets. The paper is organized as the following. Section 2 studies the dynamic characteristics of MCC in terms of its 1st-order derivatives with respect to TN and TP, two of the major components of a confusion matrix. The results of the derivatives indicate a close relationship of MCC with the ratios of sample sizes that are associated with the number of data records for the two classes in dataset. Section 3 shows the performance of MCC with respect to balanced and imbalanced datasets through a ROC analysis. Section 4 compares the dynamic behavior of MCC and its value distributions with other popular measurements and further illustrates its shortcomings for measuring classification accuracy on imbalanced datasets. Section 5 contains conclusion remarks.

2. The MCC measurement

The *Matthews correlation coefficient* (MCC) is defined as

$$\mathrm{MCC} = \frac{TP\ TN - FP\ FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2.1}$$

Which can also be expressed as a function of the derived ratios from a confusion matrix, such that [4]:

$$\begin{aligned}\mathrm{MCC} &= \sqrt{TPR\ TNR\ PPV\ NPV} - \sqrt{FPR\ FNR\ FDR\ FOR} \\ &= \sqrt{TPR\ TNR}\sqrt{PPV\ NPV} - \sqrt{FPR\ FNR}\sqrt{FDR\ FOR}.\end{aligned} \tag{2.2}$$

Footnote 1[1] below shows the definitions of these derived ratios and their aliases.

[1] From [https://en.wikipedia.org/wiki/Confusion_matrix]
1) sensitivity, recall, hit rate, or true positive rate (TPR) = $\frac{TP}{TP+FN}$ = 1-FNR
2) specificity, selectivity or true negative rate (TNR) = $\frac{TN}{TN+FN}$ =1-FPR
3) precision or positive predictive value (PPV) = $\frac{TP}{TP+FP}$ =1-FDR
4) negative predictive value (NPV) = $\frac{TN}{TN+FN}$ =1-FOR
5) miss rate or false negative rate (FNR) = $\frac{FN}{FN+TP}$ =1-TPR
6) fall-out or false positive rate (FPR) = $\frac{FP}{FP+TN}$ =1-TNR
7) false discovery rate (FDR) = $\frac{FP}{FP+TP}$ =1-PPV
8) false omission rate (FOR) = $\frac{FN}{FN+TN}$ =1-NPV

E-mail address: qzhu@unomaha.edu

https://doi.org/10.1016/j.patrec.2020.03.030
0167-8655/

Essentially, MCC is a measurement defined on the geometric means of the eight compounded measurements directly derived from the four components of a confusion matrix, with the geometric mean of four major measurement ratios on its positive side, and the geometric mean of the complemental measurement ratios on its negative side.

While MCC is expressed as a function of four variables: TP, FN, TN, and FP, as shown in expression (2.1), it can actually be represented by a function of two variables. It is because for any given dataset in a classification process, we have the number of samples in the majority class A = TN + FP, i.e., FP = A – TN, and the number of samples in the minority class B = TP + FN, i.e., FN = B – TP, where A, and B are the samples sizes for the two classes and are constant for any given dataset. Observing this fact, the MCC can be expressed as

$$\mathrm{MCC} = \frac{TP\ TN - (A-TN)(B-TP)}{\sqrt{(TP+A-TN)\ B\ A\ (TN+B-TP)}} = \frac{1}{\sqrt{AB}} \frac{TP\ TN - (A-TN)(B-TP)}{\sqrt{(TP+A-TN)\ (TN+B-TP)}}. \tag{2.3}$$

Further from (2.3), we have

$$\mathrm{MCC} = \frac{1}{\sqrt{AB}} \frac{TP\ TN - AB + BTN + ATP - TNTP}{\sqrt{(TP+A-TN)\ (TN+B-TP)}} = \frac{1}{\sqrt{AB}} \frac{-AB + BTN + ATP}{\sqrt{(TP+A-TN)\ (TN+B-TP)}} = \frac{-\sqrt{AB} + \sqrt{\frac{B}{A}}TN + \sqrt{\frac{A}{B}}TP}{\sqrt{(TP+A-TN)\ (TN+B-TP)}}. \tag{2.4}$$

It is seen from (2.4) that the MCC measurement takes the class sample size ratio $\sqrt{\frac{B}{A}}$ and $\sqrt{\frac{A}{B}}$ as two factors on its numerator components TN and TP, respectively. Obviously, when A = B, the factor does not have any effect to the roles of TN and TP in the MCC computation. However, when A is much greater than B (A ≫ B), i.e., when the data is imbalanced, the $\sqrt{\frac{B}{A}}$ and $\sqrt{\frac{A}{B}}$ serve as two weights on TN and TP, respectively, and place a bias on TP. The bias is vice versa for cases of B ≫ A.

Based on this observation, we can set to study how the function MCC behaves with respect to the variables TN and TP under the given sample size conditions. This can be done by examining the 1st-order partial derivatives of MCC with respect to TN and TP. The 1st-order partial derivative of MCC, according to (2.3), with respect to TN can be calculated as

$$\mathrm{MCC}_{\mathrm{TN}}' = \frac{1}{\sqrt{AB}} \left(\frac{\partial\left(\frac{TP\ TN}{\sqrt{(TP+A-TN)\ (TN+FN)}} \right)}{\partial(TN)} - \frac{\partial\left(\frac{(A-TN)FN}{\sqrt{(TP+A-TN)\ (TN+FN)}} \right)}{\partial(TN)} \right)$$

$$= \frac{1}{\sqrt{AB}} \left(\left(\frac{\partial(TP\ TN)}{\partial TN} \frac{1}{\sqrt{(TP+A-TN)(TN+FN)}} + \frac{\partial\left(\frac{1}{\sqrt{(TP+A-TN)\ (TN+FN)}} \right)}{\partial TN} (TPTN) \right) \right.$$
$$\left. - \left(\frac{\partial((A-TN)\ FN)}{\partial TN} \frac{1}{\sqrt{(TP+A-TN)\ (TN+FN)}} + \frac{\partial\left(\frac{1}{\sqrt{(TP+A-TN)\ (TN+FN)}} \right)}{\partial TN} (A-TN)FN \right) \right)$$

$$= \frac{1}{\sqrt{AB}} \left(\left(\frac{TP}{\sqrt{(TP+A-TN)(TN+FN)}} + \left(-\frac{1}{2}\right) \frac{TP\ TN}{\sqrt[3]{(TP+A-TN)(TN+FN)}} \left(\frac{\partial((TP+A-TN)(TN+FN))}{\partial TN} \right) \right) \right.$$
$$\left. - \left(\frac{-FN}{\sqrt{(TP+A-TN)*(TN+FN)}} + \left(-\frac{1}{2}\right) \frac{(A-TN)FN}{\sqrt[3]{(TP+A-TN)(TN+FN)}} \left(\frac{\partial((TP+A-TN)(TN+FN))}{\partial(TN)} \right) \right) \right)$$

$$= \frac{1}{\sqrt{AB}} \left(\left(\frac{TP}{\sqrt{(TP+A-TN)(TN+FN)}} + \left(-\frac{1}{2}\right) \frac{TP\ TN}{\sqrt[3]{(TP+A-TN)(TN+FN)}} \left(\frac{\partial(TP+A-TN)}{\partial(TN)}(TN+FN) + \frac{\partial(TN+FN)}{\partial(TN)}(TP+A-TN) \right) \right) \right.$$
$$\left. - \left(\frac{-FN}{\sqrt{(TP+A-TN)(TN+FN)}} + \left(-\frac{1}{2}\right) \frac{(A-TN)FN}{\sqrt[3]{(TP+A-TN)(TN+FN)}} \left(\frac{\partial(TP+A-TN)}{\partial(TN)}(TN+FN) + \frac{\partial(TN+FN)}{\partial(TN)}(TP+A-TN) \right) \right) \right)$$

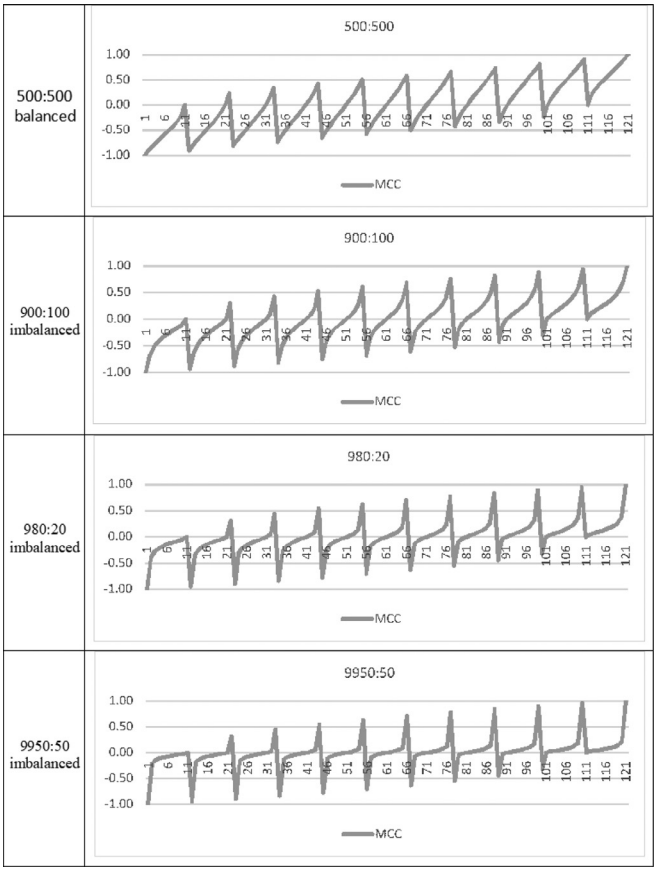
$$= \frac{1}{\sqrt{AB}} \left(\left(\frac{TP}{\sqrt{(TP+A-TN)(TN+FN)}} + \left(-\frac{1}{2}\right) \frac{TP\ TN}{\sqrt[3]{(TP+A-TN)(TN+FN)}} ((-1)(TN+FN) + (TP+A-TN)) \right) \right.$$
$$\left. - \left(\frac{-FN}{\sqrt{(TP+A-TN)(TN+FN)}} + \left(-\frac{1}{2}\right) \frac{(A-TN)FN}{\sqrt[3]{(TP+A-TN)(TN+FN)}} ((-1)(TN+FN) + (TP+A-TN)) \right) \right)$$

$$= \frac{1}{\sqrt{AB}} \left(\frac{TP}{\sqrt{(TP+A-TN)(TN+FN)}} + \frac{FN}{\sqrt{(TP+A-TN)(TN+FN)}} + \left(\frac{1}{2}\right) \frac{TP\ TN\ (TN+FN)}{\sqrt[3]{(TP+A-TN)(TN+FN)}} \right.$$
$$\left. + \left(-\frac{1}{2}\right) \frac{TP\ TN\ (TP+A-TN)}{\sqrt[3]{(TP+A-TN)(TN+FN)}} + \left(-\frac{1}{2}\right) \frac{(A-TN)FN\ (TN+FN)}{\sqrt[3]{(TP+A-TN)(TN+FN)}} - \left(-\frac{1}{2}\right) \frac{(A-TN)FN\ (TP+A-TN)}{\sqrt[3]{(TP+A-TN)(TN+FN)}} \right)$$

Table 2.1
Parameters of the test cases of the datasets.

	#1	#2	#3	#4	#5	#6	#7	#8
A:B	500:500	750:250	900:100	950:50	980:20	990:10	9950:50	9990:10
$\frac{B}{A+B}$	0.5	0.25	0.1	0.05	0.02	0.01	0.005	0.001

Table 2.2
Plots of MCC values as TN increases from minimum to maximum at 11 different TP levels on test datasets.



$$= \frac{1}{\sqrt{AB}}\left(\frac{TP+FN}{\sqrt{(TP+A-TN)(TN+FN)}} + \frac{1}{2}\left(\frac{(TN+FN)-(TP+A-TN)}{(TP+A-TN)(TN+FN)}\right)\left(\frac{TP\ TN}{\sqrt{(TP+A-TN)(TN+FN)}} - \frac{(A-TN)FN}{\sqrt{(TP+A-TN)(TN+FN)}}\right)\right). \quad (2.5)$$

Note TP + FN = B and the expression of MCC in (2.3), then the above can be rewritten as

$$\mathrm{MCC_{TN}}' = \frac{1}{2}\frac{(TN+FN)-(TP+A-TN)}{(TP+A-TN)(TN+FN)}\mathrm{MCC} + \sqrt{\frac{B}{A}}\frac{1}{\sqrt{(TP+A-TN)(TN+FN)}} \quad (2.6)$$

The expression (2.6) indicates that the rate of changes of the MCC values with respect to the changes of TN values under the condition of fixed TP levels, i.e., the dynamic characteristics of MCC with respect to TN, take the class size ratio $\sqrt{\frac{B}{A}}$ as a factor.

We experimented with a set of test cases with imbalance rate of 0.5 (1:1 balanced), 0.25, 0.1, 0.05, 0.02, 0.01, 0.005, and down to 0.001, respectively, as shown in Table 2.1.

Confusion matrices were constructed on possible outcomes of classifications (in terms of percentage of samples in TP, FP, TN, and FN respectively) for each of the test cases. Table 2.2 shows the plots of the MCC values with respect to TN increase at different levels of TP (for illustration purpose only the plots of datasets #1, #3, #5, and #7 are shown). Each saw-tooth shaped interval in the horizontal direction of the plot represents how the MCC value changes as TN increases from its minimum to its maximum, while TP keeps fixed at each level and with a 10% increment between each interval along the horizontal axis. For example, the last saw-tooth interval at the far right side of the plot represents the MCC values as TN increases from 0 to its maximum while TP is kept at its maximum level. Note that the shape of the curve for each saw-tooth section changes along with the increase of the TP levels, especially with the more imbalanced datasets.

By the symmetricity of the MCC components with respect to TP and TN, the 1st-order partial derivative of MCC with respect to TP can be expressed as

$$\mathrm{MCC_{TP}}' = \frac{1}{\sqrt{AB}}\left(\frac{\partial\left(\frac{TP\ TN}{\sqrt{(TP+FP)\ (TN+B-TP)}}\right)}{\partial(TP)} - \frac{\partial\left(\frac{FP(B-TP)}{\sqrt{(TP+FP)\ (TN+B-TP)}}\right)}{\partial(TP)}\right)$$

$$= \frac{1}{\sqrt{AB}}\left(\left(\frac{\partial(TP\ TN)}{\partial TP}\frac{1}{\sqrt{(TP+FP)(TN+B-TP)}} + (TPTN)\frac{\partial\left(\frac{1}{\sqrt{(TP+FP)\ (TN+B-TP)}}\right)}{\partial TP}\right)\right.$$
$$\left.-\left(\frac{\partial(FP(B-TP))}{\partial TP}\frac{1}{\sqrt{(TP+FP)\ (TN+B-TP)}} + FP(B-TP)\frac{\partial\left(\frac{1}{\sqrt{(TP+FP)\ (TN+B-TP)}}\right)}{\partial TP}\right)\right)$$

$$= \frac{1}{\sqrt{AB}}\left(\left(\frac{TN}{\sqrt{(TP+FP)(TN+B-TP)}} + \left(-\frac{1}{2}\right)\frac{TP\ TN}{\sqrt[3]{(TP+FP)(TN+B-TP)}}\left(\frac{\partial((TP+FP)(TN+B-TP))}{\partial TP}\right)\right)\right.$$
$$\left.-\left((-FP)\frac{1}{\sqrt{(TP+FP)\ (TN+B-TP)}} + \left(-\frac{1}{2}\right)\frac{FP(B-TP)}{\sqrt[3]{(TP+FP)(TN+B-TP)}}\left(\frac{\partial((TP+FP)(TN+B-TP))}{\partial(TP)}\right)\right)\right)$$

$$= \frac{1}{\sqrt{AB}}\left(\left(\frac{TN}{\sqrt{(TP+FP)(TN+FN)}} + \left(-\frac{1}{2}\right)\frac{TP\ TN}{\sqrt[3]{(TP+A-TN)(TN+FN)}}\left(\frac{\partial(TP+FP)}{\partial(TP)}(TN+B-TP) + \frac{\partial(TN+B-TP)}{\partial(TP)}(TP+FP)\right)\right)\right.$$
$$\left.-\left(\frac{-FP}{\sqrt{(TP+FP)(TN+B-TP)}} + \left(-\frac{1}{2}\right)\frac{FP(B-TP)}{\sqrt[3]{(TP+FP)(TN+B-TP)}}\left(\frac{\partial(TP+FP)}{\partial(TP)}(TN+B-TP) + \frac{\partial(TN+B-TP)}{\partial(TP)}(TP+FP)\right)\right)\right)$$

$$= \frac{1}{\sqrt{AB}}\left(\left(\frac{TN}{\sqrt{(TP+FP)(TN+B-TP)}} + \left(-\frac{1}{2}\right)\frac{TP\ TN}{\sqrt[3]{(TP+A-TN)(TN+FN)}}((TN+B-TP)+(-1)(TP+FP))\right)\right.$$
$$\left.-\left(\frac{-FP}{\sqrt{(TP+FP)(TN+B-TP)}} + \left(-\frac{1}{2}\right)\frac{FP(B-TP)}{\sqrt[3]{(TP+FP)(TN+FN)}}((TN+B-TP)+(-1)(TP+FP))\right)\right)$$

$$= \frac{1}{\sqrt{AB}}\left(\frac{TN}{\sqrt{(TP+FP)(TN+B-TP)}} + \frac{FP}{\sqrt{(TP+FP)(TN+B-TP)}} + \left(-\frac{1}{2}\right)\frac{TP\ TN\ (TN+B-TP)}{\sqrt[3]{(TP+FP)(TN+B-TP)}}\right.$$
$$\left.+\left(-\frac{1}{2}\right)(-1)\frac{TP\ TN\ (TP+FP)}{\sqrt[3]{(TP+FP)(TN+B-TP)}} - \frac{1}{2}(-1)\frac{FP(B-TP)\ (TP+FP)}{\sqrt[3]{(TP+FP)(TN+B-TP)}}\right)$$

$$= \frac{1}{\sqrt{AB}}\left(\frac{TN+FP}{\sqrt{(TP+FP)(TN+B-TP)}} - \frac{1}{2}\left(\frac{(TN+B-TP)-(TP+FP)}{(TP+FP)(TN+B-TP)}\right)\left(\frac{TP\ TN}{\sqrt{(TP+FP)(TN+B-TP)}} - \frac{FP(B-TP)}{\sqrt{(TP+FP)(TN+B-TP)}}\right)\right). \tag{2.7}$$

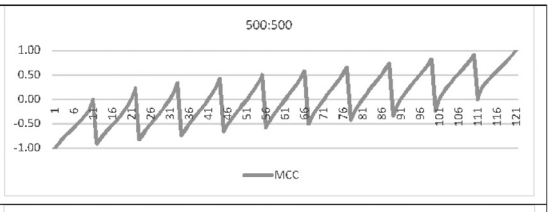
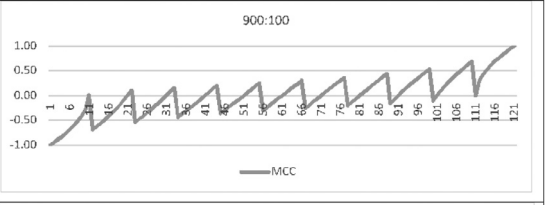
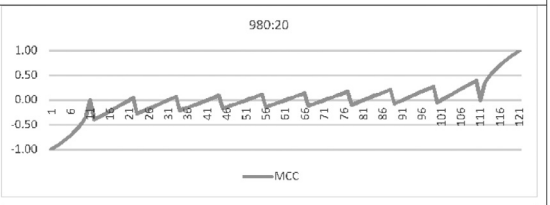
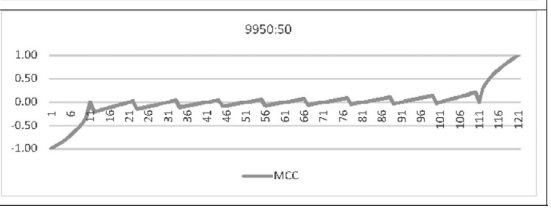
Note the MCC in above expression, it can be written as

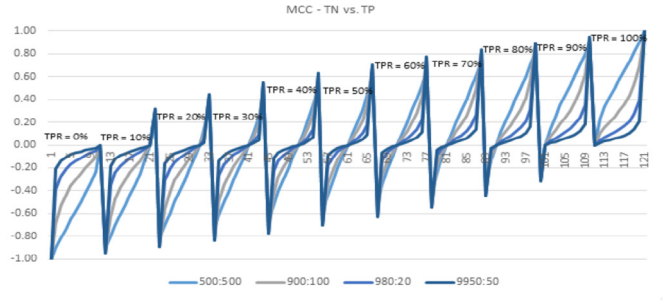
$$\mathrm{MCC_{TP}}' = \frac{1}{\sqrt{AB}}\frac{TN+FP}{\sqrt{(TP+FP)(TN+B-TP)}} - \frac{1}{2}\frac{(TN+B-TP)-(TP+FP)}{(TP+FP)(TN+B-TP)}\mathrm{MCC} \tag{2.8}$$

Note TN + FP = A, we then have

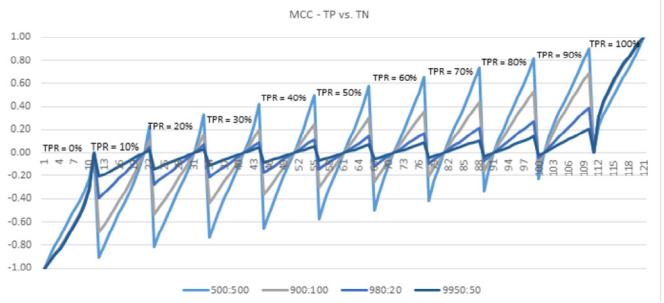
$$\mathrm{MCC_{TP}}' = -\frac{1}{2}\frac{(TN+B-TP)-(TP+FP)}{(TP+FP)(TN+B-TP)}\mathrm{MCC} + \frac{1}{\sqrt{AB}}\frac{A}{\sqrt{(TP+FP)(TN+B-TP)}}$$
$$= -\frac{1}{2}\frac{(TN+B-TP)-(TP+FP)}{(TP+FP)(TN+B-TP)}\mathrm{MCC} + \sqrt{\frac{A}{B}}\frac{1}{\sqrt{(TP+FP)(TN+B-TP)}} \tag{2.9}$$

Table 2.3
Plots of MCC curves as TP increases at different TN levels on the test datasets.

500:500 balanced	
900:100 imbalanced	
980:20 imbalanced	
9950:50 imbalanced	



(a)



(b)

Fig. 2.1. MCC measurements for test cases of different imbalance ratios plotted together for a comparison, where (a) TN increases from 0 to maximum w.r.t. each fixed TP level, and (b) TP increases from 0 to maximum w.r.t. each fixed TN level.

Similar to MCC_{TN}', the MCC_{TP}' values with respect to the changes of TP under the condition of fixed TN levels take the class size ratio $\sqrt{\frac{A}{B}}$ as a factor. Table 2.3 shows the plots of the $MCC_{TP'}$ for the test cases of #1, #3, #5, and #7, respectively.

The expressions of (2.6) and (2.9) indicate that, in addition to the TN and TP values, the dynamic characteristics of MCC is also affected by the sample sizes A and B, especially the class imbalance ratios A:B. The higher the imbalance ratio, the MCC measurements tend to be more skewed and behave nonlinearly with respect to the linear increases of TP and TN values. This fact can be seen clearly from the shapes and slop changes of the curves with respect to the different imbalance ratios of the test cases as shown in the plots of Tables 2.2 and 2.3.

It is seen from the plots that the behavior of MCC for the imbalanced cases is quite different from that for the balanced cases. Using the 500:500 case as a reference, the characteristics shown in the curves for the imbalanced cases were seriously skewed. The higher (e.g., 9950:50 case is higher than 980:20 case) of the imbalance ratio, the more skew is for the characteristic curve. For a better comparison, the plots of Fig. 2.1 place the MCC measurements of the above test cases together, where (a) and (b) are with respect to the increases of TN and TP, respectively. The changes of the shapes for the performance curves of MCC can be better compared in these figures for the different imbalance ratios of the test cases.

3. ROC space analysis

3.1. MCC in the ROC space

In the following, we present the MCC measurements on the test cases in the ROC space [12,13], and study the cut-off (threshold) points and boundaries for the performance evaluation of MCC itself. It is noted that a cut-off (threshold) values in the measurement corresponds to a boundary line in the ROC space that divides the space into two performance regions. Typically, it is the TPR = FPR line that serves as the boundary line and bisects the ROC space into two performance sections, a below average performance section under the TPR = FPR line and an above average performance section over the TPR = FPR line [14].

In addition to looking at the bisections of the MCC values in the ROC space, we also studied the performance of the MCC within each of the bisections. That is, we further looked at the performances of MCC at the lower percentile regions and upper percentile regions of the ROC space. It is because from the study of the last section we found that the MCC curves are more skewed within these regions and the performances of the MCC become more deteriorated on datasets with higher imbalance ratios.

Following the convention of ROC space, the horizontal axis is designated as the FPR and the vertical axis as the TPR for the ROC planes discussed in this section. The same axis designation is adopted in the plots of the tables below.

Plots in Table 3.1 are generated with three cut-off lines of the MCC value percentiles for each test case.

(1) Rows 1 and 2 of Table 3.1 represent the lower 10% and 5% percentile cut-off lines in the below average performance section corresponding to the MCC values of -0.2 and -0.1, respectively. The 10% line is set for the 500:500 and 900:100 test cases, and the 5% line for the 980:20 and 9950:50 test cases. The reason for the difference is that when the dataset becomes more imbalanced, such as for the 980:20 and 9950:50 cases, the 10% below average cut-off (0.2 in this case) puts the corresponding boundary line at the extreme corner of the ROC space. That is, it results a splitting that places most of the MCC measurements at one side of the splitting boundary.
(2) Rows 3 and 4 represent the middle cut-off corresponding to the line TPR = FPR, i.e., where the MCC = 0.0.
(3) Rows 5 and 6 represent the higher 10% or 5% percentile cut-off lines in the above average performance section, where 10% (MCC = 0.2) is set for the 500:500 and 900:100 cases and 5% (MCC = 0.1) is set for the 980:20 and 9950:50 cases, with the

Table 3.1
Plots of MCC values in the ROC space with respect to the measurement boundaries on balanced and imbalanced test cases.

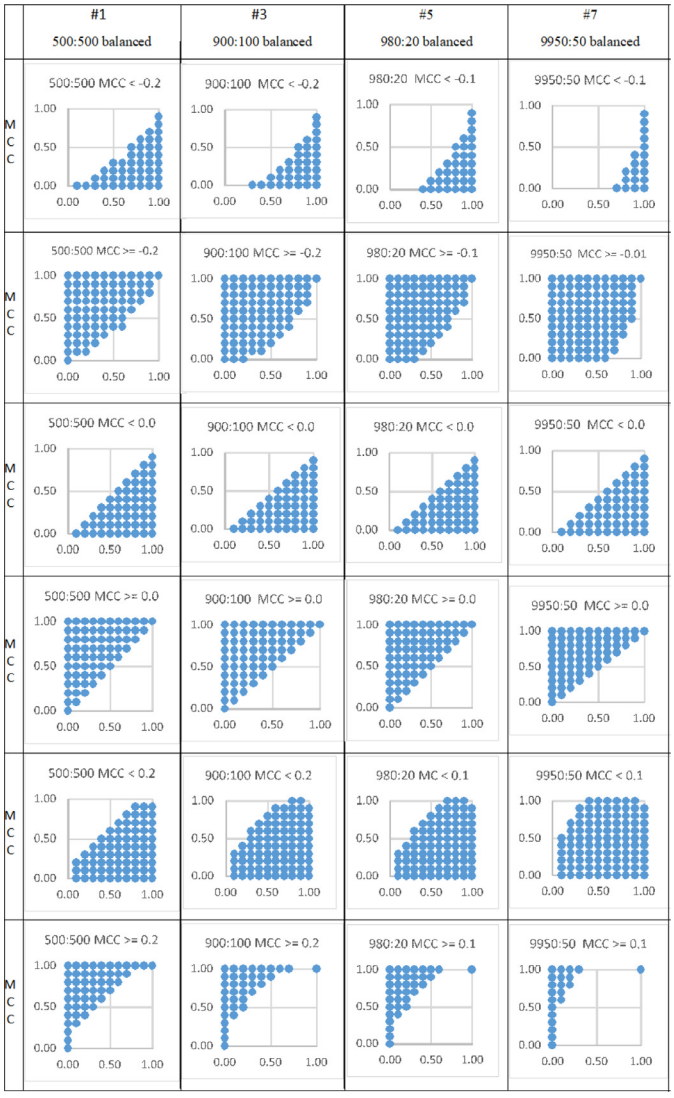


Table 3.2
Cut-off points of the ROC space for the test cases.

	#1: 500:500 balanced	#3: 900:100 imbalanced	#5: 980:20 imbalanced	#7: 9950:50 imbalanced
TPR = FPR	0.0	0.0	0.0	0.0
TPR < FPR	-0.2	-0.2	-0.1	-0.1
TPR > FPR	0.2	0.2	0.1	0.1

same reason as for the below average performance section of rows 1 and 2 discussed above.

Note that the 10% and 5% cut-off lines are set purely based on the MCC value range, not the percentage of the MCC measurement data points (the number of measurements). In fact, we will see in the next section that the percentage of the MCC measurements falling within each of these cut-off ranges is much larger than the 10% or 5% cut-off of their overall distributions.

As shown in the plots of Table 3.1, while the cut-off points for the TPR = FPR lines (shown on rows 3 and 4) are at the 0.0 for all the test cases, the cut-off points for the 10% or 5% TPR < FPR lines (shown on rows 1 and 2) and the 10% or 5% TPR > FPR lines (shown on rows 5 and 6) become quite disparate. The numbers are listed in Table 3.2 for a better comparison of the plots for the balanced (500:500), less imbalanced (900:100), and more imbalanced (980:20 and 9950:50) test cases.

Note that the #5 and #7 cases have cut-off points at -0.1 and 0.1, while the #1 and #3 are at -0.2 and 0.2. It can be seen from the plots that at the -0.1 to 0.1 cut-off interval, which is the 5% up and down of the MCC's middle splitting point for the 9950:50 case, the actual MCC measurements fall in this interval is much heavier than that for the 500:500 case though it has a four times larger interval, i.e., with -0.2 and 0.2 as its cut-off points. Table 3.3 shows the amount of the MCC measurements within the corresponding middle cut-off sections for the test cases.

The plots of Table 3.3 demonstrates the MCC measurement perform differently when the dataset becomes imbalanced. It also

Table 3.3
Plots of MCC measurements in ROC space where the MCC values fall in the 20% and 10% middle intervals on test cases.

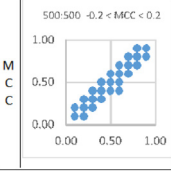
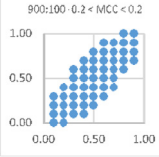
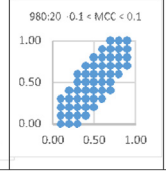
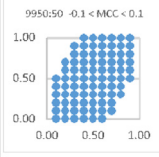
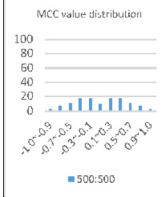
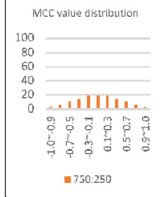
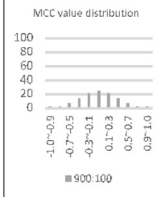
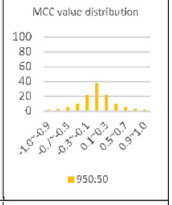
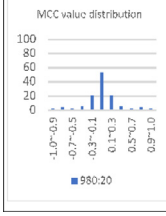
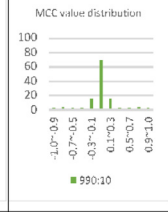
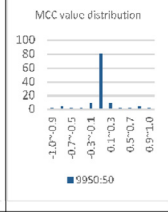
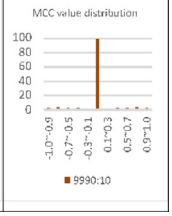
	#1 500:500 balanced	#3 900:100 balanced	#5 980:20 balanced	#7 9950:50 balanced
MCC				

Table 3.4
Distributions of the MCC measurements in equal sized intervals of its value range on the eight test cases.

#1 500:500 balanced	#2 750:250 imbalanced	#3 900:100 imbalanced	#7 950:50 imbalanced
			
#5 **980:20 imbalanced**	**#6** **990:10 imbalanced**	**#7** **9950:50 imbalanced**	**#8** **9990:10 imbalanced**
			

tends to be unevenly distributed over the value range when the dataset is imbalanced. This fact is reveled in more detail below.

3.2. The distributions of MCC

In the following we show the MCC measurement distributions for all the 8 test cases applied in our experiments. Table 3.4 contains the plots of the distributions of all possible MCC measurements, with respect to the given TN and TP value coverages (0 to A and 0 to B). The distributions are counted within each of the 10 equal sized intervals spanning the entire MCC value range from -1.0 to 1.0. The horizontal axis of the plots shows these value intervals and the vertical axis the percentage of distributions, i.e., the amount of MCC measurements, in each of these intervals.

It is seen from the plots that the MCC measurement distributions become more crowded at the center intervals, and more concentrated when the dataset becomes more imbalanced. The distributions of MCC measurements for imbalanced dataset are much denser around the middle of the MCC value range, and correspondently the distributions of the measurement are rather sparse at the two end sides of the value range. The plots show that a larger percentage of MCC measurement data points are located in the -0.1 to 0.1 interval for datasets with high imbalance ratio (e.g., 990:10, 9950:50 and 9990:10). The observation is consistent with the numeric results shown in Table 3.5.

The above situation is further illustrated in the Fig. 3.1 where the MCC distributions for all eight test cases are sorted in terms of the 9990:10 case so as to give a layered view of the data and their distribution trends.

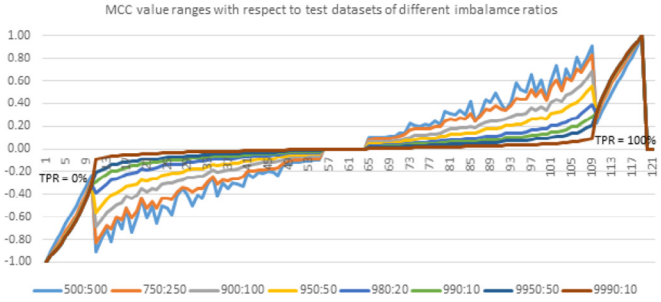


Fig. 3.1. Distribution range of MCC measurements of eight test cases with data sorted on the 990:10 test case.

From the Fig. 3.1 we see that a large portion of the distribution curve near the zero line for the 9990:10 case, and the drawing line is nearly flat. To see the details of the numbers, Table 3.5 further lists the quantitative values of the test cases for all the intervals evaluated. The data is also illustrated in Fig. 3.2.

As shown in Table 3.5 and Fig. 3.2, for the MCC value interval of from -0.1 to 0.1, i.e., a 10% interval of MCC value range, the MCC measurements runs up to 58% for the test case of imbalanced rate of 0.01 (ratio 990:10), 68% for the test case of imbalanced rate of 0.005 (ratio 9950:50), and 83% for the test case of imbalanced rate

Table 3.5
Distributions of MCC values over the 10% intervals for all eight test cases.

Percentage of distribution of MCC values w.r.t. different dataset imbalance rate											
	−1.0~−0.9	−0.9~−0.7	−0.7~−0.5	−0.5~−0.3	−0.3~−0.1	−0.1~0.1	0.1~0.3	0.3~0.5	0.5~0.7	0.7~0.9	0.9~1.0
500:500	0.03	0.06	0.09	0.14	0.14	**0.08**	0.14	0.14	0.09	0.06	0.03
750:250	0.02	0.05	0.08	0.12	0.15	**0.16**	0.15	0.12	0.08	0.05	0.02
900:100	0.02	0.03	0.06	0.12	0.18	**0.21**	0.18	0.12	0.06	0.03	0.02
950:50	0.02	0.03	0.04	0.08	0.18	**0.31**	0.18	0.08	0.04	0.03	0.02
980:20	0.02	0.03	0.02	0.04	0.17	**0.45**	0.17	0.04	0.02	0.03	0.02
990:10	0.02	0.03	0.02	0.02	0.13	**0.58**	0.13	0.02	0.02	0.03	0.02
9950:50	0.02	0.03	0.02	0.02	0.08	**0.68**	0.08	0.02	0.02	0.03	0.02
9990:10	0.02	0.03	0.02	0.02	0.00	**0.83**	0.00	0.02	0.02	0.03	0.02

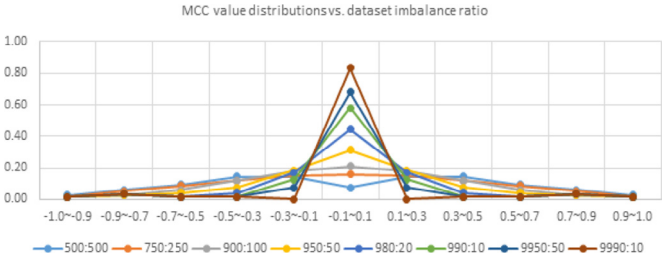


Fig. 3.2. Percentages of the MCC measurement data distributed in equal sized intervals over its value range for the eight test cases.

of 0.001 (ratio 9990:10). The distributions are quite skewed toward the center of the horizontal axis.

In an additional point of view, for the imbalanced dataset of 9990:10 case, a MCC value falls in the interval of 0.1 to 0.3 places the classification result at a 91% high accuracy region, though a measurement value in the range of 0.1 to 0.3 seems quite low conventionally. It is the same for the imbalanced dataset of 9950:50 case, a MCC value of 0.3 places the classification result at the top 10% range (i.e., >90% accuracy level) of the performance metric. That is, when MCC is applied to imbalanced datasets, it tends to give a low performance measurement value though the classification accuracy could possibly be high (near or above the 90 percentile line) already.

From another point of view, the data in Table 3.5 and Fig. 3.2 show that for the MCC value range of from 0.1 to 1.0, which counts for almost 45% of the total MCC value range, only 17% of the MCC measurements fall in this interval for the 9950:50 case, and 9% of for the 9990:10 case, as in contrary that same interval counts for 46% of the measurements for the balanced dataset of 500:500. The same is true for the lower -1.0 to -0.1 interval for both the balanced and imbalanced cases. This means that the overall distributions of the MCC measurements for classification results of imbalanced datasets become quite sharply non-uniform and unevenly distributed than for the dataset of balanced cases.

4. Comparisons of measurement metrics

It is clear from the experiment results of the test cases presented in the last two sections that the behavior of MCC differs quite significantly with respect to balanced and imbalanced datasets. In this section we focus on a comparison of the performance characteristics of the MCC with some other known classification measurements.

The Table 4.1 places together the plots of the measurements performed on the four test cases, with the same axis denotations as those shown in table (2.2). The measurement metrics compared here include the MCC, the Bookmaker Informedness (BM), the Markedness (MK), the g-mean of TPR and TNR (GBA), the g-mean of PPV and TPR (GPR), the F1 score, and the Cohen's Kappa (CK) [4,15]. Again, the data are generated and plotted with respect to the increases of TN values in each saw-tooth interval where the TP value keeps constant with an increment of 10% between the intervals (that together forms the 11 sections of the measurements in the horizontal direction).

Explanations and comparisons of these measurement metrics plotted in the Table 4.1 are discussed more detail in the following.

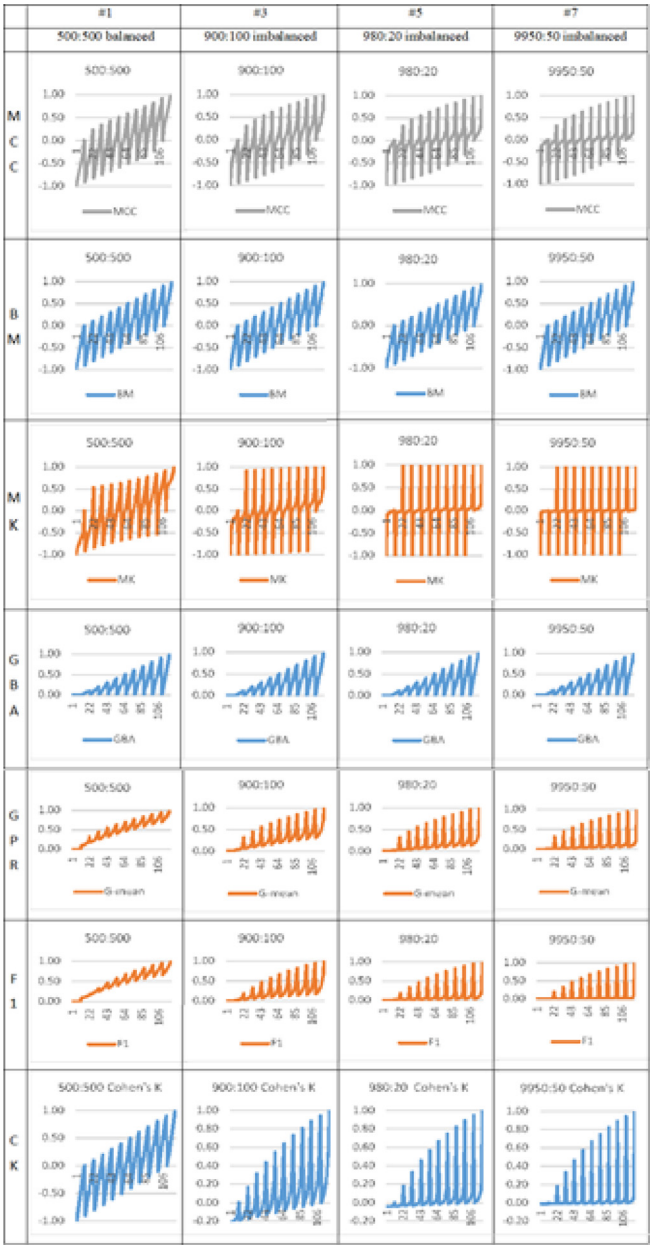
4.1. BM and MK

The Bookmaker Informedness measurement (a.k.a. Youden's index, Youden's J statistic) is defined as BM = TPR + TNR − 1. The Markedness measurement is defined as MK = PPV + NPV − 1 [9]. The measurements have the same −1.0 to 1.0 value range as MCC. However, the BM and MK have distinct performance behavior with respect to balanced and imbalanced datasets. It is seen from the plots that the BM has a linear performance and quite consistent on both balanced and imbalanced datasets in all the intervals. The MK becomes quite irresponsible to the classification outcomes (in terms of the variations of the TN and TP values in the intervals) when the dataset becomes more imbalanced. Unfortunately, MCC is somehow more leaning toward the performance of MK rather than toward the BM regarding its responses to both balanced and imbalanced dataset, as the shapes of the plots on the tested datasets show.

4.2. GBA and GPR

The g-mean of TPR and TNR can be expressed as GBA = g-mean(TPR, TNR) = $\sqrt{TPR\ TNR}$. The g-mean of PPV (which is also called Precision) and TPR (also called Recall) can be expressed as GPR = g-mean(PPV, TPR) = $\sqrt{PPV\ TPR}$. Actually, viewing MCC as consisting of two square-rooted component groups connected with a subtraction sign, as shown in Eq. (2.2), either GBA or GPR can be considered as one component of the square-rooted multiplication at the right (positive) side of MCC. It is noted from the plots that the GBA performs consistently for both balanced and imbalanced datasets, and shows a harmonic correspondence with the TN and TP value increases along the horizontal axis in the 11 sections. This property is very much desired for a classification accuracy measurement. Especially, the GBA performs uniformed for both balanced and imbalanced datasets. On the other hand, the curves of GPR in each interval skew significantly when looking from the balanced dataset to imbalanced datasets. It is obviously not suitable for performance measurement of imbalanced datasets. The performance of MCC again falls closer to that of GPR, except the value range which is from -1.0 to 1.0 rather than GPR's 0.0 to 1.0. It is interest to note that the shape of GPR resembles very closely to that of MCC at the top half of the curve, i.e., for the MCC value range of from 0.0 to 1.0. This is probably not a coincidence, and makes us to reflect on the role the GPR component plays in the characteristics and behavior changes of the MCC with respect to imbalanced datasets.

Table 4.1
Plots of MCC with other measurements on test cases.



4.3. F1

Expressed as $F1 = 2\frac{PPV \cdot TPR}{PPV+TPR} = 2\frac{TP}{2*TP+FN+FP}$, the F1 score behaves very similar to the GPR in terms of the shapes of the curves for both balanced and imbalanced datasets. In this regards, MCC also behaves on the side aligning with F1. It is known that F1 falls into the existing category of not suitable measurements for imbalanced datasets [4]. We also notice that both F1 and GPR involve the use of PPV and their performance curves resemble.

4.4. CK

The Cohen's Kappa (CK) [15,16] is also not a measurement suitable for evaluating classification results of imbalanced data. One to remark is that CK has the same value range as MCC of from -1.0 to 1.0. CK has the very similar measurement behavior as the MCC for balanced dataset. Like MCC, the CK measurement becomes more skewed when the data becomes more imbalanced. However, it is noted that while the MCC values remains in the range of from -1.0 to 1.0 for all the tested datasets, the CK tends to divert from range of -1.0 to 1.0 to the range of 0.0 to 1.0 when the datasets become more imbalanced. Besides this, the MCC and CK could be considered to be in the same category of biased measurement for imbalance dataset.

Overall viewing from the plots of Table 4.1, the characteristics of the seven measurements compared in this section resemble to each other at some degree with balanced datasets. While the BM and GBA remain consistent on imbalanced datasets, the other five become quite different when dealing with the imbalanced datasets. It is noticed that classification accuracy measurements involving the use of PPV and NPV (and their complements FDR and FOR) tend to behave undesirably in the case of the sample size imbalance. The uneven and biased performance of PPV and NPV on im-

balanced dataset is rooted at their formulation. The TPR and TNR are independent to sample size imbalance by its nature. This property of TPR and TNR leads directly to the balanced behavior of the BM and GBA measurements.

5. Conclusion

Imbalanced datasets pose a big challenge to both the selection of proper classification algorithms and the use of proper metrics for the evaluation of the accuracy in the classification outcomes [17,18]. Though there are a number of algorithmic approaches that have been studied and used in machine learning and data analytics practices for dealing with imbalanced datasets, such as the resampling techniques, there is still a need of a reliable evaluation metrics, for example, to compare the results from applying different resampling rates to determine the best parameter settings for the given application problems.

Most of the classification measurements discussed in this paper all work well for balanced datasets. Distinctions come when they are applied to measure the classification results on imbalanced datasets. The major problem of those measurements that divert away from balanced measurements is the skewing of their distributions when dealing with imbalanced datasets. Such skewing behavior makes the resolution and sensitivity uneven and non-uniformly distributed over the range of their measurement values, thus not properly reflecting the variations of the classification accuracy on the classification model applied to the imbalanced dataset. The performance of these measurements also deteriorates more when the imbalance ratio of the dataset increases.

MCC was said to be “generally regarded as a balanced measure which can be used even if the classes are of very different sizes.” [4,6]. The illustrations and analyses of our experimentations show that this statement is not really true. The MCC falls into the same category as the GPR, F1, MK, and CK that are not suitable for directly applying to the measurement of classification accuracy on imbalanced datasets.

Declaration of Competing Interest

None.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patrec.2020.03.030.

References

[1] N. Chawla, V. Nitesh, Data mining for imbalanced datasets: an overview, in: Oded Maimon, Abel Browarnik (Eds.), Data Mining and Knowledge Discovery Handbook, Springer Science+Business Media, January 2005, pp. 853–867, doi:10.1007/978-0-387-09823-4_45. (Chapter 40)ISBN 978-0-387-09822-7.

[2] G. Cheng, W.J. Poon, A new evaluation measure for imbalanced datasets, in: *Proceedings of the Seventh Australasian Data Mining Conference* (AusDM 2008), 87, November 2008, pp. 27–32.

[3] Q. Gu, L. Zhu, Z. Cai, Evaluation measures of the classification performance of imbalanced data Sets, in: *Computational Intelligence and Intelligent Systems, International Symposium on Intelligence Computation and Applications* (ISICA 2009), 51, 2009, pp. 461–471.

[4] D. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation, J. Machine Learn. Technol. 2 (1) (2011) 37–63.

[5] Q. Zou, S. Xie, Z. Lin, M. Wu, Y. Ju, Finding the best classification threshold in imbalanced classification, J. Big Data Res. 5 (September 2016) 2–8 Elsevier.

[6] D. Chicco, Ten quick tips for machine learning in computational biology, BioData Mining 10 (35) (December 2017) Published: 08 Available at https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0155-3 , doi:10.1186/s13040-017-0155-3. August 2019.

[7] L.A. Jeni, J.F. Cohn, F. De La Torre, Facing imbalanced data - recommendations for the use of performance metrics, in: *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. ACII '13, IEEE Computer Society, 2013, pp. 245–251.

[8] Y. Liu, J. Cheng, C. Yan, X. Wu, F. Chen, Research on the Matthews correlation coefficients metrics of personalized recommendation algorithm evaluation, Int. J. Hybrid Inf. Technol. 8 (1) (2015) 163–172 http://dx.doi.org/10.14257/ijhit.2015.8.1.14.

[9] Wikipedia, Matthews correlation coefficient, Available at https://en.wikipedia.org/wiki/Matthews_correlation_coefficient, August 2019.

[10] J. Lever, M. Krzywinski, N. Altman, Classification evaluation, Nat. Methods 13 (8) (2016) 603–604 Available at http://fortinlab.bio.uci.edu/FortinLab/Teaching_files/Stats/POS_Classification_evaluation.pdf . August 2019.

[11] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using Matthews correlation coefficient metric, PLOS ONE 12 (6) (2017) June 2,Available at https://doi.org/10.1371/journal.pone.0177678 . August 2019.

[12] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (8) (2006) 861–874.

[13] J. Hallinan, Assessing and comparing classifier performance with ROC curves, Machine Learning Process (2014) Nov. 26available at https://machinelearningmastery.com/assessing-comparing-classifier-performance-roc-curves-2/ . August 2019.

[14] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recogn. 30 (6) (1997) 1145–1159 Elsevier.

[15] D. Powers, The problem with Kappa, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, April 2012, pp. 345–355.

[16] M.J. Warrens, Five ways to look at Cohen's Kappa, J. Psychol. Psychother. 197 (2015) 1–4 5.

[17] K. Madasamy, M. Ramaswami, Data imbalance and classifiers: impact and solutions from a big data perspective, Int. J. Comput. Intell. Res. 13 (9) (2017) 2267–2281 ISSN 0973-1873.

[18] N. Tomašev, M. Dunja, Class imbalance and the curse of minority hubs, Knowl.-Based Syst. 53 (2013) 157–172.