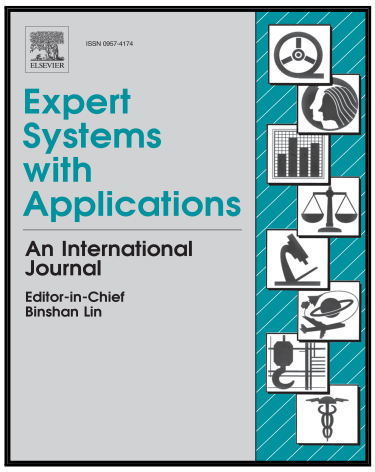

A Novel Weighted TPR-TNR Measure to Assess Performance of the Classifiers

Anil S. Jadhav

PII: S0957-4174(20)30215-3
DOI: https://doi.org/10.1016/j.eswa.2020.113391
Reference: ESWA 113391

To appear in: *Expert Systems With Applications*

Received date: 17 December 2019
Revised date: 23 February 2020
Accepted date: 15 March 2020

Please cite this article as: Anil S. Jadhav , A Novel Weighted TPR-TNR Measure to Assess Performance of the Classifiers, *Expert Systems With Applications* (2020), doi: https://doi.org/10.1016/j.eswa.2020.113391

Highlights

- We propose new single valued measure to assess performance of the classifiers.
- The proposed measure takes into account imbalanced ratio of the dataset.
- Performance of the proposed measure is more suitable when dataset is imbalanced.
- Results of the proposed measure are compared with the existing measures.

A Novel Weighted TPR-TNR Measure to Assess Performance of the Classifiers

Anil S. Jadhav
Symbiosis Centre for Information Technology
Symbiosis International (Deemed University), Pune - 411057, Maharashtra, India.
Email: a_s_jadhav74@yahoo.co.in

Abstract. Assessing performance of different classifiers and selecting the best one is one of the most important tasks in classification problem. The assessment of classifiers becomes more complex when dataset is imbalanced because most of the frequently used performance metrics can be misleading. Many real world classification problems such as fraud detection, churn prediction, medical diagnosis, and cyber-security suffer from the problem of imbalanced datasets. Therefore, in all such classification tasks it is very important to select the best classifier very carefully. In this study we propose new weighted TPR-TNR measure to assess performance of the classifiers. The proposed measure takes into consideration imbalance ratio of the dataset and assigns different weights to the TPR and TNR to assess classifiers performance. We have used five different datasets to assess performance of twelve different classifiers using weighted TPR-TNR measure and compared it with the existing measures. The experimental results show that the weighted TPR-TNR measure is more suitable to assess performance of the classifiers when dataset is imbalanced.

Keywords: Classification; Classifiers evaluation; Assess classifiers performance; Performance measures

1. Introduction

Classification is machine learning technique in which computer program learns from the past data and then uses the learning to predict class label for unseen data. For example, identifying whether i) a mail is spam or non spam; ii) a customer will default in loan repayment; iii) a customer will churn. There are several applications of classification problem in various domains. In general, classifiers are built using past data and are used to predict class label of unseen data. Usually, we build several classifiers and then select the best one based on performance of the classifiers. Therefore, assessing performance of the classifiers is an important task in classification problem (Hossin, 2015).

In the field of data mining and machine learning, when a new algorithm is proposed or when an existing algorithm is modified, implicit assumption is that the proposed algorithm gives better result than an existing algorithm. In order to prove or validate the assumption, performance of the proposed algorithm is compared with the existing algorithms with respect to a certain criteria (Demsar, 2006).

There exists several metrics in the literature to assess performance of the classification algorithms such as accuracy, sensitivity, specificity, true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), false negative rate (FNR), positive predictive value or precision, negative predictive value or inverse precision, likelihood ratio, F measure, area under receiver operating characteristics curve (AUC), geometric mean (GM) of TPR and TNR , Matthews Correlation Coefficient (Hand, 2012; powers, 2011; Sokolova, 2016; Tharwat, 2018, Matthews, 1975). When performance of a newly proposed or a modified algorithm is to be compared with an existing algorithm, researchers run the algorithms on selected test datasets and compare their performance using appropriate measures. Classifiers are compared using a single or multiple datasets. Assessing performance of different classifiers on multiple datasets is a common practice followed in the field of machine learning and data science (Brown, 2012; Zhu, 2017; Demsar, 2006).

Some researchers consider a single criterion while others use several criteria for evaluating performance of classifiers. Evaluating performance of multiple classifiers on several datasets using a single criterion can be easily done using a statistical test such as ANOVA or Friedman test statistics. Whereas, assessing performance of multiple classifiers on several datasets using multiple criteria is a difficult and challenging task (Demsar, 2006). Multi-criteria decision making methods such as Gray Relational Analysis (GRA) and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) methods have been used in

literature to evaluate performance of different classifiers using more than one performance measure (Ali et al. 2017; Kou et al. 2012).

Objectives of this study are i) to introduce existing performance measures used to assess performance of the classifiers; ii) to explain how to calculate scores for the existing performance measures; iii) to discuss pros and cons of the existing measures; iv) to propose a new weighted TPR-TNR measure to assess performance of the classifiers; and v) to compare proposed measure with the existing performance measures. Rest of the paper is organized as follows. Section 2 provides a theoretical background and related work. Section 3 describes two MCDM methods namely GRA and TOPSIS used to evaluate performance of the classifiers using more than one performance measure. In section 4 we present: i) proposed weighted TPR-TNR measure to assess performance of the classifiers; ii) experimental results of the proposed measure and its comparison with the existing performance measures. The paper is concluded in the section 5.

2. Theoretical background and related work

The most common question in the field of machine learning is to find which classifier gives better performance. There exist several literatures that talk about assessing performance of the classifiers. The performance of classifiers can be evaluated either by using a single evaluation criterion or several different criteria. In some literature, classifiers are assessed using a single evaluation criterion by taking its average performance over different datasets. While others apply statistical techniques to test whether difference in the performance of classifiers is real and not by chance. When the difference is real, it becomes necessary to understand which classifier performance is better. For evaluating performance of any two classifiers, either a t-test or Wilcoxon signed ranked test is used. When more than two classifiers are involved, their performance is evaluated using either ANOVA or the Friedman test statistics (Demsar, 2006).

A confusion matrix is the most common tool used to represent classification results to assess performance of the classifiers (Hossin, 2015; Tharwat, 2018; Sokolova, 2006; Saito, 2015; Powers 2011). The confusion matrix, as shown in Table 1, is first created to assess performance of the classifiers.

Table 1. Confusion Matrix

Actual Class	Predicted Class	
	Positive	Negative
Positive(P)	True Positive (TP)	False Negative (FN)
Negative(N)	False Positive (FP)	True Negative (TN)

The most common measures used to assess performance of the classifiers are summarized in Table 2. We divide these measures into three types: i) basic measures, ii) derived measures, and iii) graphical measures (Tharwat, 2018). The basic measures are classification results produced in the form of confusion matrix values namely: TP, TN, FP, and FN. The derived measures are derived from the values of the basic measures. Graphical measures are derived from computations based on the graphical representation of the derived measures such as precision, recall, sensitivity, and specificity (Tharwat, 2018).

Sokolova et al. presented a detailed systematic analysis of twenty four different performance measures used in the classification task (Sokolova, 2009). Tom Fawcett discussed in detail about the ROC curve with its common misconceptions, pitfalls, and guidelines when using it in practice (Fawcett, 2006). Garcia et al. presented a detailed analysis of performance measures for an imbalanced dataset (Garcia, 2010). There exist literatures that discuss in detail about the ROC and Precision-recall curve (Davis, 2006; Saito, 2015). The study conducted by Jeni et al. (2013) addressed the question – How does skewed distribution influence performance metrics for action unit detection? They found that except area under ROC curve, other measures namely Accuracy, F-score, Cohen's kappa, and Krippendorf's alpha attenuated by skewed distribution.

Table 2. Classifier Evaluation Measures

Measure	What is focus of measure?	Formula
	Basic Measures	
Sensitivity/Recall/True Positive rate (TPR)	Measures how good model is in correctly predicting positive cases	$\frac{TP}{TP+FN}$
Specificity/True Negative Rate (TNR)	Measures how good model is in correctly predicting negative cases	$\frac{TN}{TN+FP}$
False Positive Rate (FPR)/Type–I error/False alarm rate	Measures proportion of incorrectly classified negative cases	$\frac{FP}{FP+TN}$
False Negative rate/Type-II error (FNR)	Measures proportion of incorrectly classified positive cases	$\frac{FN}{FN+TP}$
Accuracy	Measures how good model is in correctly predicting both positive and negative cases	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision/Positive	Measures proportion of correctly	$\frac{TP}{TP+FP}$

Predictive Value	classified positive cases out of total positive predictions	
Inverse Precision/Negative Predictive Value	Measures proportion of correctly classified negative cases out of total negative predictions	$\frac{TN}{TN+FN}$
Jaccard coefficient	Measures similarity between actual and predicted values	$\frac{TP}{TP+FP+FN}$
Mathews Correlation coefficient (MCC)	Measures correlation between observed and predicted classifications	$\frac{TP \times TN - FP \times FN}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$
	Derived Measures	
F Measure	It is harmonic mean of precision and recall	$\frac{2 \times Precision \times Recall}{Precision + Recall}$
Geometric Mean	Geometric mean of sensitivity and specificity	$\sqrt{\mathrm{Sensitivity} \times \mathrm{Specificity}}$
Youden's Index	Measures discriminating power of the test i.e. ability of classifier to avoid misclassification	$\mathrm{Sensitivity} + \mathrm{Specificity} - 1$
Positive likelihood ratio (LR+)	It is ratio between TPR and FPR	$\frac{Sensitivity}{1 - Specificity}$
Negative likelihood ratio(LR-)	It is ratio between FPR and TNR	$\frac{1 - Sensitivity}{Specificity}$
	Graphical Measures	
Area under ROC	Measures area under plot of TPR against FPR	
Precision Recall Curve	Plot of Precision against Recall	

Limitations of the existing performance measures

In this section we discuss the limitations of the widely used classifiers performance measures namely: TPR, TNR, FPR, FNR, accuracy, precision, and inverse precision. In a classification task, the dataset can be balanced or imbalanced. A balanced dataset contains approximately equal number of both positive and negative cases. An imbalanced dataset on the other hand, contains unequal proportion of positive and negative cases where one class outnumbers the other class. Some of the widely used performance measures are not appropriate in evaluating a classifier's performance when dataset is imbalanced. For example, consider the confusion matrix shown in Table 3 and Table 4, representing a classifier's performance, for a balanced and imbalanced dataset respectively. The calculated values of performance measures for the balanced and imbalanced dataset are shown in Table 5.

Table 3. Classifier performance for balanced dataset

	Positive	**Negative**
Positive	TP (30)	FN (20)
Negative	FP (20)	TN (30)

Table 4. Classifier performance for imbalanced dataset

	Positive	Negative
Positive	TP (15)	FN (10)
Negative	FP (30)	TN (45)

Table 5. Example of classifiers Performance for balanced and imbalanced dataset

Metrics	TPR	TNR	FPR	FNR	Accuracy	Precision	Inverse Precision	F-Measure	Jaccard coefficient
Balanced dataset	0.6000	0.6000	0.4000	0.4000	0.6000	0.6000	0.6000	.6000	.4286
Imbalanced dataset	0.6000	0.6000	0.4000	0.4000	0.6000	0.3333	0.8182	.4286	.2727

From Table 5, it is observed that Precision, Inverse precision, F-measure, and Jaccard coefficient values get changed even though classifier's performance does not change. Therefore, these are misleading performance measures when dataset is imbalanced.

Consider another example, shown in Table 6, to demonstrate problems associated with commonly used performance measures when dataset is imbalanced. The top row in the Table 6 presents three sample cases of classifiers performance for: 1) balanced dataset, 2) imbalanced dataset with large number of negative cases as compared to positive cases, 3) imbalanced dataset with large number of positive cases as compared to negative cases. It also provides information about total number of positive cases, negative cases, and classifier results produced in the form of confusion matrix values TP, TN, FP, FN. The second row in the table indicates cost associated with misclassification of positive and negative cases. Rest all rows in the Table 6 represents performance measured using derived measures.

Table 6. Examples of performance measures for balanced and imbalanced dataset

Measure for evaluating performance of classification algorithm	Balanced Dataset: Actual positive cases: 500 Actual negative cases: 500 TP: 450, TN:455, FP:45, FN:50	Imbalanced Dataset: Actual positive cases: 100 Actual negative cases: 1000 TP: 5, TN:990, FP:10, FN:95	Imbalanced Dataset: Actual positive cases: 1000 Actual negative cases: 100 TP: 990, TN:5, FP:95, FN:10
	Cost associates with misclassification of positive and negative cases is same	Cost associated with misclassification of positive cases is higher than negative cases	Cost associated with misclassification of negative cases is higher than positive cases
Accuracy	0.9050	0.9045	0.9045
TPR (Sensitivity/Recall)	0.90	0.05	0.99
TNR (Specificity)	0.91	0.99	0.05
FPR (Type-I Error)	0.09	0.01	0.95
FNR (Type-II Error)	0.10	0.95	0.01
Precision/PPV	0.909	0.33	0.91
Inverse Precision/NPV	0.9010	0.91	0.33
Mathews	0.81	0.099	0.099

Correlation Coefficient			
F Measure	0.9045	0.08	0.949
Geometric Mean	0.9050	0.22	0.22
Jaccard Coefficient	.8257	0.04	0.90

From the Table 6 it is observed that:

i) When dataset is imbalanced, accuracy of the machine learning model is misleading because accuracy is very high even though performance of the model is very poor in correctly classifying minority cases. Therefore accuracy reflects true overall performance of the machine learning model only when dataset is balanced and cost associated with misclassification of positive and negative cases is same

ii) Sensitivity and specificity measures assess ability of classifier to correctly predicting positive and negative cases respectively. These measures are not appropriate in assessing overall performance of the classifier because we can obtain sensitivity value 1 by predicting all negative cases as positive. This is done at the cost of misclassifying all negative cases. Sensitivity and specificity are inversely related, increasing one will decrease the other. Therefore, sensitivity and specificity are not appropriate single valued measures to assess overall performance of classifier.

iii) FPR and FNR measures reflect proportion of incorrectly classified negative and positive cases respectively. Therefore, FPR and FNR are also not appropriate single valued measures to assess overall performance of the classifier

iv) Precision, F measure, and Jaccard coefficient are single valued measures to assess overall performance of the classifier but are not appropriate when dataset is imbalanced with large number of positive cases and small number of negative cases and objective of the machine learning algorithm is correctly predicting maximum number of negative cases without compromising much on incorrect prediction of positive cases. These measures ignore proportion of correct prediction of negative cases (TN). For example, consider third column where total number of positive cases are 1000; and negative cases are 100. Correctly classified positive cases are 990; and correctly classified negative cases are only 5. Jaccard coefficient value for this case is .9041 which is very high and misleading. F- measure value is .9496 which is again very high and misleading. Precision value is .91 which is also very high and misleading. In this scenario inverse precision measure is more suitable but it is inappropriate in opposite situation when there are large number of negative cases and small number of positive cases and objective of the machine learning algorithm is to correctly classifying maximum number of positive cases without compromising much on incorrect

prediction of negative cases. From the results of the performance measures for all the three cases it is observed that Matthew's correlation coefficient and Geometric mean are more informative single scores to assess overall performance of the classifiers.

ROC and precision-recall curve are graphical representation methods used to assess performance of different classifiers. ROC curve is a plot of true positive rate (sensitivity) against false positive rate (1-specificity). ROC has been widely used in many diagnostic and machine learning systems (Zou, 2002). It shows proportion of true positive and false positive cases for a given threshold value. The ROC curve is plotted by considering different threshold values. The proportion of correctly classified positive cases depends on actual threshold selected and not on different thresholds used to plot the ROC curve. If a model is perfect in predicting both positive and negative cases then the true positive rate would be 100 percent and false positive rate would be 0 percent for all thresholds. The area under ROC curve (AUC) is 1 when a model is perfect and it is 0.5 or less than 0.5 when a model is poor in discriminating positive and negative cases. AUC is averaged across all the thresholds and performance of the classifier is based on a selected threshold (Steve Halligan, 2015). Therefore, AUC does not reflect true performance of the classifier. Another problem with ROC AUC is that it gives equal importance to both positive and negative cases which means that it considers same cost for misclassification of both positive and negative cases. Therefore, in clinical diagnostic tests where cost associated with misclassification of positive cases is very high compared to misclassification of the negative cases, ROC AUC is comparatively not an appropriate measure to assess performance of different classifiers. The concept of precision-recall curve is similar to ROC curve. It shows the relationship between precision and recall. Precision-recall curve has limitations similar to ROC curve. A more detailed discussion about ROC and Precision-recall curve can be found in the study conducted by Tharwat and Davis (Tharwat, 2018; Davis, 2006)

3. MCDM Methods for evaluating performance of the classifiers

In this section, we discuss two existing MCDM methods that are used to evaluate performance of the classifiers. The GRA and TOPSIS methods are very widely used as multi criteria decision making methods for evaluating and ranking different alternatives (Behzadian, 2012; kuo, 2008). These methods are also used for evaluation and selection of the best classifiers (Ali, 20117; Kou, 2012). The process of evaluating performance of the

classifiers using these methods is discussed in this section. Whereas, the experimental results of evaluating classifiers using these methods are presented in the section 4.

3.1.TOPSIS Method

The technique for order of preference by similarity to ideal solution (TOPSIS) is a multi criteria decision making method for evaluating performance of various alternatives and eventually selecting the best one. It was originally developed by Ching-Lai Hwang and Yoon (1981) and was further enhanced by Chen and Hwang (1992). This technique is based on proximity measure and find out distance of each alternative from the positive ideal solution and Negative ideal solution. The basic principle of TOPSIS method is that the best alternative has shortest distance from the ideal solutions and farthest distance from the negative ideal solution.

The TOPSIS process for evaluating and selecting the best classifier is described as follows:

Step 1: Construct a decision matrix with m alternatives and n criteria. Each cell value x_{ij} in the matrix represents performance value of alternative 'i' for criteria 'j'. Where i= 1,2,…,m and j=1,2,…n.

Therefore first step in evaluating performance of different classifiers is to construct decision matrix where all classifiers to be evaluated are considered as alternatives and performance measures are considered as evaluation criteria.

Step 2: Normalize the decision matrix as follows:

$$n_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} xij^2}} \quad i=1,2,\ldots,m \quad j=1,2,\ldots,n$$

Where n_{ij} is normalized decision matrix

Step 3: Calculate weighted normalized decision matrix

$W_{ij}=n_{ij} * w_j$ where i=1,2,…,m and j=1,2,…,n w_j is weight of each criteria, and sum of all weights is 1

For evaluating performance of the classifiers we have assigned equal weight to each performance evaluation measure.

Step 4: Determine positive ideal solution and negative ideal solution.

The positive ideal solution is hypothetical alternative created by considering highest value among all the alternatives for each benefit criteria and lowest value among all the alternatives for the cost criteria.

The negative ideal solution is hypothetical alternative created by considering lowest value among all the alternatives for each benefit criteria and highest value among all the alternatives for the cost criteria.

Benefit criteria are the criteria for which highest value is best value, and cost criteria are criteria for which lowest value is best value. For evaluating performance of different classifiers, only FPR and FNR are the cost attributes while rest all performance measures are benefit attributes

Step 5: Calculate distance between each alternative and the positive ideal solution (d+); and distance between each alternative and the negative ideal solution (d-) using Euclidian distance method.

Step 6: Calculate similarity with the best solution as follows:

$$S = \frac{d-}{(d-)+(d+)}$$

If s=1, the alternative is best alternative, and if s=0 the alternative is worst alternative.

Step 7: Rank the alternatives based on similarity value. Highest value of similarity indicates the best alternative.

3.2. Gray Relational Analysis:

Gray relational analysis (GRA) is based on gray system theory introduced by J. L. Deng in 1982 (Deng, 1982). It was introduced to solve MCDM problem for small samples and missing information. It has been widely used to solve MCDM problem in various fields (Kuo, 2008; Lin, 2006). GRA is based on degree of similarity between alternative and ideal alternative. Stepwise GRA method for evaluating different classifiers using different performance measures is described as follows:

Step 1: Construct a decision matrix with m classifiers and n performance measures. Each cell value Xac in the matrix represents performance value of alternative 'a' for criteria 'c'. Where a= 1,2,…,m; and c=1,2,…n.

Step 2: Normalize decision matrix. Normalization can be done in three ways. For benefit attributes (i.e. attributes for which maximum value is best value), normalization is done using following formula.

$$Xac = \frac{Xac - \min(Xac)}{\max(Xac) - \min(Xac)}$$

For cost attributes (i.e. attributes for which minimum value is best value), normalization is done using following formula.

$$Xac = \frac{\max(Xac) - Xac}{\max(Xac) - \min(Xac)}$$

Attributes for which optimum value is best value, normalization is done using following formula.

$$Xac = \frac{Xac - Xoc}{\max(Xac) - \mathrm{Xoc}}$$

For evaluating performance of different classifiers only FPR and FNR are cost attributes and remaining performance measures are benefit attributes.

Step 3: Calculate gray relational coefficient: The first step in calculating gray relational coefficient is calculate difference between alternative series and reference series. Reference series is series having best value for each attribute. The second step is finding gray relational coefficient using following formula:

$$\gamma\mathrm{ac} = \frac{\Delta\min + \rho\,\Delta\max}{\Delta\mathrm{ac} + \rho\,\Delta\max}$$

where: Δmin is minimum difference between alternative and reference series for the attribute, and Δmax is maximum difference between alternative and reference series for the attribute. Δac is difference between alternative and reference series. The ρ is distinguishing coefficient and it is usually kept as 0.5; the range of ρ is between 0 to 1.

Step 4: Calculate gray relational degree: the formula to calculate gray relational degree is as follows:

$$^{\delta}\mathrm{a} = \frac{1}{n}\sum_{a=1}^{n}\gamma\mathrm{ac}$$

Step 5: Rank the alternatives based on gray relational degree. Alternative with highest value of gray relational degree is the best alternative.

4. Proposed method and Experimental results

In section 4.1 we discuss proposed measure to assess performance of the classifiers and then we present experimental results of proposed weighted TPR-TNR measure for evaluating performance of the classifiers, and compare it with the existing performance measures. The first experiment is conducted on 5 different datasets obtained from KEEL repository (https://sci2s.ugr.es/keel/imbalanced.php). The second experiment is conducted on synthetic data. Both the experiments are discussed in detail in the section 4.2 and 4.3 respectively.

4.1.Weighted TPR-TNR Measure

In this section we discuss the proposed weighted TPR-TNR measure to assess performance of the classifiers. In the previous section we have discussed basic as well as derived performance measures and its pros and cons in using it for assessing performance of the classifiers. We have also seen that most of the performance measures are not appropriate to assess performance of the classifier when dataset is imbalanced. When dataset is balanced, accuracy is the best single valued performance measure to assess performance of the classifier. But, when dataset is imbalanced, cost of misclassification of minority class (usually positive) instance is higher than the costs of misclassification of majority class (usually negative) instance.

One of the main distinctions between different performance measures of the classifiers is whether or not they take the class size (more precisely imbalance ratio of the dataset) into account. To the best of our knowledge, none of the existing single valued performance measures takes into account the imbalance ratio of the dataset while assessing performance of the classifiers. Therefore, in the proposed measure we assign different weights to the i) TPR (Sensitivity) - correct prediction of the positive class instances and ii) TNR (Specificity) - correct prediction of negative class instances. The formula to calculate weighted TPR-TNR score is as follows:

$$\text{Weighted TPR-TNR} = \left(TPR * \frac{N}{P+N}\right) + \left(TNR * \frac{P}{P+N}\right)$$

Where P is total number of positive cases; N is total number of negative cases

$\frac{N}{P+N}$ is weight assigned to TPR; and $\frac{P}{P+N}$ is weight assigned to TNR

The weight assigned to sensitivity and specificity depends on the imbalanced ratio of the dataset. For example, consider imbalanced dataset having 5 positive cases (P=5) and 95 negative cases (N=95). The calculations of imbalanced ratio, weight of TPR, and weight of TNR are shown as follows:

Imbalanced ratio of the dataset = N/P = 95/5 = 19

Weight of sensitivity (TPR) = N / P+N= 95/100= 0.95

Weight of specificity (TNR) = P / P+N = 5/100 = 0.05

The calculation of weight for the above example shows that the weight of sensitivity is 19 times more than weight of specificity. This is because one correct prediction of positive case is equivalent to 19 correct predictions of negative cases. In other words one incorrect prediction of positive case is equivalent to 19 incorrect predictions of negative cases. The weight assigned to the sensitivity and specificity plays very important role in calculating combined overall accuracy of the classifier when dataset is imbalanced.

Consider another example given in Table 7 to demonstrate suitability of the proposed measure to assess performance of the classifier. The last row in the Table 7 shows calculated values of weighted TPR-TNR measure for all the three cases. The results show that the proposed measure is suitable for all the three cases irrespective of: i) whether dataset is balanced or not; ii) whether imbalanced dataset is having large number of positive cases or large number of negative cases. The calculations of the performance measures in Table 7 shows that only Mathews Correlation Coefficient (MCC), geometric mean (GM), and proposed weighted TPR-TNR measures are suitable single valued measures to assess overall performance of the classifiers when dataset is imbalanced. The comparison of these single valued performance measures is discussed in detail in the section 4.2 and 4.3.

Table 7. Examples of existing and proposed performance measures for balanced and imbalanced dataset

Measure for evaluating performance of classification algorithm	**Case 1: Balanced Dataset** Positive cases (P) : 500 Negative cases (N) : 500 TP: 450, TN:455, FP:45, FN:50	**Case 2: Imbalanced Dataset** Positive cases (P): 100 Negative cases (N): 1000 TP: 5, TN:990, FP:10, FN:95	**Case3:Imbalanced Dataset** Positive cases (P): 1000 Negative cases (N) : 100 TP: 990, TN:5, FP:95, FN:10
	Cost associated with misclassification of positive and negative cases is same	Cost associated with misclassification of positive cases is higher than negative cases	Cost associated with misclassification of negative cases is higher than positive cases
Accuracy	0.9050	0.9045	0.9045
TPR (Sensitivity/Recall)	0.90	0.05	0.99
TNR (Specificity)	0.91	0.99	0.05
FPR (Type-I Error)	0.09	0.01	0.95
FNR (Type-II Error)	0.10	0.95	0.01
Precision/PPV	0.909	0.33	0.91
Inverse Precision/NPV	0.9010	0.91	0.33
Mathews Correlation Coefficient	0.81	0.099	0.099
F Measure	0.9045	0.08	0.949
Geometric Mean	0.9050	0.22	0.22
Jaccard Coefficient	.8257	0.04	0.90
Weighted TPT-TNR	.9050	.1355	.1355

4.2.Experiment 1

In this experiment we use 12 different classification algorithms, and 5 different datasets to assess performance of the classifiers using proposed and existing measures. The datasets used in this experiment are downloaded from the KEEL repository (https://sci2s.ugr.es/keel/imbalanced.php). The information about these dataset is given in Table 8.

Table 8: Dataset Information

Dataset	No. of instances	No. of attributes	Imbalance Ratio
pima	768	9	1.87
yeast6	1484	8	41.4
ecoli4	336	7	15.8
page-blocks-1-3_vs_4	472	10	15.86
abalone-21_vs_8.dat	581	9	40.5

The information about 12 different classification algorithms used in this experiment is provided in Table 9. The detailed information about these classification algorithms can be found at https://cran.r-project.org/ or https://topepo.github.io/caret

Tables 9. Classifiers

Classification Algorithm	**Short description of algorithm**
J48 (C4.5)	Tree based model
C5.0	Tree based model
Knn	k-nearest neighbors
rf	Random forest –tree based ensemble
ctree	Conditional inference tree
svmradial	Support Vector Machines with Radial Basis Function Kernel
gbm	Stochastic gradient boosting
nb	Naïve bayes classifier
glm	Generalized linear model
rpart	Tree based model
svmlinear	Support Vector Machines with linear Kernel
svmpoly	Support Vector Machines with polynomial kernel

We have used 11 different widely used performance measures namely: Accuracy, TPR, TNR, FPR, FNR, precision, inverse precision, MCC, F-measure, geometric mean (GM), and Jaccard coefficients to assess classifiers performance. The information about these performance measures is provided in the Table 2 of section 2. Performance of the classifiers is also assessed using proposed weighted TPR-TNR measure and two MCDM methods namely: GRA and TOPSIS.

The procedure followed to assess performance of the classifiers is as follows:

1. The dataset was first divided into training and test subset with split percentage of 70 and 30 percent respectively

2. The classifier was built using training dataset with k-fold cross validation method
3. The classification results were obtained using the test dataset
4. The classifiers performance was measured using 11 different performance measures
5. GRA and TOPSIS methods were then used to rank each classifier using 11 different performance measures
6. The weighted TPR-TNR score was also calculated based on the classification result
7. Classifiers were then ranked according to scores of GRA, TOPSIS, MCC, GM and 'Weighted TPR-TNR' measure

The caret package in R programming is used to conduct the experiment. The above process is repeated for all the five datasets.

4.2.1. Experimental results for the pima dataset

Classification results of 12 different classifiers for pima dataset are shown in the Table 10. Classifiers performance using different measures for pima dataset is shown in the Table 11. The bold values in table indicate the best value for each performance measure. Table 12 shows rank of the classifiers obtained using proposed and existing performance measures for pima dataset. The ranks of the classifiers are calculated only for GRA, TOPSIS, MCC, GM, and proposed weighted TPR-TNR measure. The other measures are not considered to rank the classifiers because of its limitations discussed in the earlier section. The last column in the Table 12 indicates rank of the classifier based on authors judgement by taking into account classification results (TP,TN, FP, and FN values) and imbalance ratio of the dataset.

Table 10. Classification results for the pima dataset

Classifiers	P	N	TP	TN	FP	FN
J48	80	150	46	120	30	34
C5.0	80	150	34	139	11	46
Knn	80	150	39	127	23	41
rf	80	150	45	129	21	35
ctree	80	150	46	131	19	34
svmradial	80	150	43	131	19	37
gbm	80	150	44	125	25	36
nb	80	150	47	126	24	33
glm	80	150	44	130	20	36
rpart	80	150	52	116	34	28
svmlinear	80	150	42	129	21	38
svmpoly	80	150	40	133	17	40

Table 11. Classifiers performance for the pima dataset

Classifiers	Accuracy	TPR	TNR	FPR	FNR	Precision	Inverse precision	MCC	F Measure	GM	Jaccard	Weighted TPR-TNR
J48	0.7217	0.5750	0.8000	0.2000	0.4250	0.6053	0.7792	0.3797	0.5897	0.6782	0.4182	0.6533
C5.0	0.7522	0.4250	**0.9267**	**0.0733**	0.5750	**0.7556**	0.7514	0.4222	0.5440	0.6276	0.3736	0.5995
knn	0.7217	0.4875	0.8467	0.1533	0.5125	0.6290	0.7560	0.3587	0.5493	0.6425	0.3786	0.6124
rf	0.7565	0.5625	0.8600	0.1400	0.4375	0.6818	0.7866	0.4449	0.6164	0.6955	0.4455	0.6660
ctree	**0.7696**	0.5750	0.8733	0.1267	0.4250	0.7077	0.7939	**0.4742**	**0.6345**	**0.7086**	**0.4646**	**0.6788**
svmradial	0.7565	0.5375	0.8733	0.1267	0.4625	0.6935	0.7798	0.4410	0.6056	0.6851	0.4343	0.6543
gbm	0.7348	0.5500	0.8333	0.1667	0.4500	0.6377	0.7764	0.3984	0.5906	0.6770	0.4190	0.6486
nb	0.7522	0.5875	0.8400	0.1600	0.4125	0.6620	0.7925	0.4408	0.6225	0.7025	0.4519	0.6753
glm	0.7565	0.5500	0.8667	0.1333	0.4500	0.6875	0.7831	0.4428	0.6111	0.6904	0.4400	0.6601
rpart	0.7304	**0.6500**	0.7733	0.2267	**0.3500**	0.6047	**0.8056**	0.4167	0.6265	0.7090	0.4561	0.6929
svmlinear	0.7435	0.5250	0.8600	0.1400	0.4750	0.6667	0.7725	0.4112	0.5874	0.6719	0.4158	0.6415
svmPoly	0.7522	0.5000	0.8867	0.1133	0.5000	0.7018	0.7688	0.4265	0.5839	0.6658	0.4124	0.6345

Table 12. Classifiers rank for the pima dataset

Classifiers	GRA rank	TOPSIS rank	MCC rank	GM Rank	Weighted TPR-FPR rank	Rank based on authors own judgement
J48	7	7	7	7	7	7
C5.0	1	1	12	12	12	11
knn	11.5	11.5	11	11	11	12
rf	2.5	2.5	4	4	4	4
ctree	7	7	2	2	2	1
svmradial	11.5	11.5	6	6	6	6
gbm	10	9	8	8	8	8
nb	9	10	3	3	3	3
glm	4.5	4.5	5	5	5	5
rpart	7	7	1	1	1	2
svmlinear	4.5	4.5	9	9	9	9
svmpoly	2.5	2.5	10	10	10	10

The results of classifiers performance for pima dataset and its ranks based on the proposed and existing measures shows that: i) The ranks of the classifiers based on GRA and TOPSIS method are far different from the ranks based on authors judgement and not appropriate. For example, GRA and TOPSIS methods assign rank '1' to the 'C5.0' classifier which is not correct because results clearly shows that performance of the 'C5.0' is poor than 'ctree', 'rpart', 'nb' classifiers; ii) The ranks of the classifiers based on MCC, GM, and Weighted TPR-TNR measure are exactly same and are also similar to the ranks based on authors judgement except first two and last two ranks.

4.2.2. Experimental results for the yeast dataset

Classification results of 12 different classifiers for yeast dataset are shown in the Table 13. Classifiers performance using different measures for yeast dataset is shown in the Table 14. The bold values in the table indicate the best value for each performance measure. Table 15 shows ranks of the classifiers obtained using proposed and existing measures for the yeast dataset.

Table 13. Classification results for yeast dataset

Classifiers	P	N	TP	TN	FP	FN
J48	10	434	2	433	1	8
C5.0	10	434	0	434	0	10
knn	10	434	4	433	1	6
rf	10	434	5	432	2	5
ctree	10	434	5	431	3	5
svmradial	10	434	3	434	0	7
gbm	10	434	4	433	1	6
nb	10	434	0	434	0	10
glm	10	434	2	433	1	8
rpart	10	434	5	431	3	5
svmlinear	10	434	0	434	0	10
svmPoly	10	434	1	434	0	9

Table 14. Classifiers performance for yeast dataset

Classifiers	Accuracy	TPR	TNR	FPR	FNR	Precision	Inv Precision	MCC	F Measure	GM	Jaccard	Weighted TPR-TNR
J48	0.9797	0.2	**0.9977**	0.0023	0.8	0.6667	0.9819	0.3581	0.3077	0.4467	0.1818	0.2180
C5.0	0.9775	0	1	**0**	1	0	0.9775	0	0	0	0	0.0225
knn	**0.9842**	0.4	**0.9977**	0.0023	0.6	**0.8**	0.9863	0.5592	0.5333	0.6317	0.3636	0.4135
rf	**0.9842**	**0.5**	0.9954	0.0046	**0.5**	0.7143	**0.9886**	**0.5901**	**0.5882**	**0.7055**	**0.4167**	0.5112
ctree	0.982	**0.5**	0.9931	0.0069	**0.5**	0.625	0.9885	0.55	0.5556	0.7047	0.3846	0.5111
svmradial	**0.9842**	0.3	1	**0**	0.7	1	0.9841	0.5434	0.4615	0.5477	0.3	0.3158
gbm	**0.9842**	0.4	**0.9977**	0.0023	0.6	**0.8**	0.9863	0.5592	0.5333	0.6317	0.3636	0.4135
nb	0.9775	0	1	**0**	1	0	0.9775	0	0	0	0	0.0225
glm	0.9797	0.2	**0.9977**	0.0023	0.8	0.6667	0.9819	0.3581	0.3077	0.4467	0.1818	0.2180
rpart	0.982	**0.5**	0.9931	0.0069	**0.5**	0.625	0.9885	0.55	0.5556	0.7047	0.3846	0.5111
svmlinear	0.9775	0	1	**0**	1	0	0.9775	0	0	0	0	0.0225
svmPoly	0.9797	0.1	1	**0**	0.9	1	0.9797	0.313	0.1818	0.3162	0.1	0.1203

Table 15. Classifiers rank for yeast dataset

Classifiers	GRA rank	Topsis rank	MCC rank	GM Rank	Weighted TPR-FPR rank	Rank based on authors own judgment
J48	8.5	7.5	7	7	7	7
C5.0	11	11	10	10	10	10
Knn	5.5	1.5	2	4	4	4
rf	1	4	1	1	1	1
ctree	3.5	5.5	4	2	2	2
svmradial	2	3	6	6	6	6
gbm	5.5	1.5	2	4	4	4
nb	11	11	10	10	10	10
glm	8.5	7.5	7	7	7	7
rpart	3.5	5.5	4	2	2	2
svmlinear	11	11	10	10	10	10
svmpoly	7	9	9	9	9	9

The result of classifiers performance for the yeast dataset and its ranks based on the proposed and existing measures shows that: i) The ranks of the classifiers based on weighted TPR-TNR, and GM measure are exactly similar to the ranks based on authors judgement; ii) The ranks of the classifiers based on GRA, TOPSIS, and MCC measure are different than the ranks based on authors judgement; iii) Careful examination of classification results (TP, TN, FP, FN) shows that ranks based on TPR-TNR and GM measure are more appropriate compared to the ranks based on GRA, TOPSIS, and MCC measures. For example, the results clearly shows that ‘ctree’ and ‘rpart’ classifier performance is better than ‘gbm’ and ‘knn’

classifier but it is not reflected in the ranks based on MCC and TOPSIS measures. Further, ranks of GRA measure shows that 'svmradial' is better than 'rpart' and 'ctree' which is also incorrect.

4.2.3. Experimental results for the ecoli dataset

Classification results of 12 different classifiers for the ecoli dataset are shown in the Table 16. Classifiers performance using different measures for the ecoli dataset is shown in Table 17. The bold values in the table indicate the best value for each performance measure. Table 18 shows rank of classifiers performance obtained using proposed and existing methods for ecoli dataset.

Table 16. Classification results for ecoli dataset

Classifiers	P	N	TP	TN	FP	FN
J48	6	94	6	93	1	0
C5.0	6	94	6	92	2	0
Knn	6	94	6	93	1	0
rf	6	94	6	91	3	0
ctree	6	94	5	93	1	1
svmradial	6	94	6	93	1	0
gbm	6	94	6	92	2	0
nb	6	94	5	94	0	1
glm	6	94	6	92	2	0
rpart	6	94	5	93	1	1
svmLinear	6	94	0	94	0	6
svmPoly	6	94	0	94	0	6

Table 17. Classifiers performance for ecoli dataset

Classifiers	Accuracy	TPR	TNR	FPR	FNR	Precision	Inverse Precision	MCC	F Measure	GM	Jaccard	Weighted TPR-TNR
J48	**0.99**	**1**	0.9894	0.0106	**0**	0.8571	**1**	**0.9209**	**0.9231**	**0.9947**	**0.8571**	0.9994
C5.0	0.98	**1**	0.9787	0.0213	**0**	0.75	**1**	0.8568	0.8571	0.9893	0.75	0.9987
knn	**0.99**	**1**	0.9894	0.0106	**0**	0.8571	**1**	0.9209	0.9231	**0.9947**	**0.8571**	0.9994
rf	0.97	**1**	0.9681	0.0319	**0**	0.6667	**1**	0.8034	0.8	0.9839	0.6667	0.9981
ctree	0.98	0.8333	0.9894	0.0106	0.1667	0.8333	0.9894	0.8227	0.8333	0.908	0.7143	0.8427
svmradial	**0.99**	**1**	0.9894	0.0106	**0**	0.8571	**1**	0.9209	**0.9231**	**0.9947**	**0.8571**	0.9994
gbm	0.98	**1**	0.9787	0.0213	**0**	0.75	**1**	0.8568	0.8571	0.9893	0.75	0.9987
nb	**0.99**	0.8333	**1**	**0**	0.1667	**1**	0.9895	0.9081	0.9091	0.9129	0.8333	0.8433
glm	0.98	**1**	0.9787	0.0213	**0**	0.75	**1**	0.8568	0.8571	0.9893	0.75	0.9987
rpart	0.98	0.8333	0.9894	0.0106	0.1667	0.8333	0.9894	0.8227	0.8333	0.908	0.7143	0.8427
svmlinear	0.94	0	**1**	**0**	1	0	0.94	0	0	0	0	0.0600
svmpoly	0.94	0	**1**	**0**	1	0	0.94	0	0	0	0	0.0600

Table 18. Classifiers rank for ecoli dataset

Classifiers	GRA rank	TOPSIS rank	MCC rank	GM Rank	Weighted TPR-TNR rank	Rank based on authors own judgment
J48	3	3	1	1	1	1
C5.0	6	8	5	4	4	4
knn	3	3	1	1	1	1
rf	10	10	10	7	7	7
ctree	8.5	5.5	8	9	9	9
svmradial	3	3	1	1	1	1
gbm	6	8	5	4	4	4
nb	1	1	4	8	8	8
glm	6	8	5	4	4	4
rpart	8.5	5.5	8	9	9	9
svmlinear	11.5	11.5	11	11	11	11
svmpoly	11.5	11.5	11	11	11	11

The results of classifiers performance for ecoli dataset and its ranks based on the proposed and existing measures shows that: i) The ranks of the classifiers based on weighted TPR-TNR, and GM measure are exactly similar to the ranks based on authors judgement; ii) The ranks of the classifiers based on GRA, TOPSIS measure are different than the ranks based on authors judgement and are not appropriate. For example, classification results clearly shows that performance of the 'J48' and 'svmradial' is better than the 'nb' classifier but it is not reflected in the ranks of GRA and TOPSIS methods; iii) the ranks based on MCC measure are also not appropriate for the 'rf' and 'nb' classifiers because results clearly shows that 'rf' classifier is better than 'nb' classifier.

4.2.4. Experimental results for the page-block dataset

Classification results of 12 different classifiers for the page-block dataset are shown in the Table 19. Classifiers performance using different measures for the page-block dataset is shown in Table 20. The bold values in the table indicate the best value for each performance measure. Table 21 shows rank of classifiers performance obtained using proposed and existing methods.

Table 19. Classification results for the page-block dataset

Classifiers	P	N	TP	TN	FP	FN
J48	8	133	8	133	0	0
C5.0	8	133	8	133	0	0
knn	8	133	7	131	2	1
rf	8	133	8	133	0	0
ctree	8	133	7	132	1	1
svmradial	8	133	2	133	0	6
gbm	8	133	8	133	0	0
nb	8	133	4	127	6	4
glm	8	133	4	133	0	4
rpart	8	133	4	133	0	4
svmlinear	8	133	3	132	1	5
svmpoly	8	133	4	132	1	4

Table 20. Classifiers performance for the page-block dataset

Classifiers	Accuracy	TPR	TNR	FPR	FNR	Precision	Inv Precision	MCC	F Measure	GM	Jaccard	Weighted TPR-TNR
J48	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	1.0000
C5.0	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	1.0000
knn	0.9787	0.875	0.985	0.015	0.125	0.7778	0.9924	0.8138	0.8235	0.9284	0.7	0.8812
rf	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	1.0000
ctree	0.9858	0.875	0.9925	0.0075	0.125	0.875	0.9925	0.8675	0.875	0.9319	0.7778	0.8817
svmRadial	0.9574	0.25	1	0	0.75	1	0.9568	0.4891	0.4	0.5	0.25	0.2926
gbm	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	1.0000
nb	0.9291	0.5	0.9549	0.0451	0.5	0.4	0.9695	0.41	0.4444	0.691	0.2857	0.5258
glm	0.9716	0.5	1	**0**	0.5	1	0.9708	0.6967	0.6667	0.7071	0.5	0.5284
rpart	0.9716	0.5	1	**0**	0.5	1	0.9708	0.6967	0.6667	0.7071	0.5	0.5284
svmLinear	0.9574	0.375	0.9925	0.0075	0.625	0.75	0.9635	0.5121	0.5	0.6101	0.3333	0.4100
svmPoly	0.9645	0.5	0.9925	0.0075	0.5	0.8	0.9706	0.616	0.6154	0.7044	0.4444	0.5279

Table 21. Classifiers rank for the page-block dataset

Classifiers	GRA Rank	TOPSIS Rank	MCC rank	GM Rank	Weighted TPR-TNR rank	Rank based on authors own judgment
J48	**2.5**	**2.5**	1	1	1	1
C5.0	**2.5**	**2.5**	1	1	1	1
knn	6	6	6	6	6	6
rf	**2.5**	**2.5**	1	1	1	1
ctree	5	5	5	5	5	5
svmradial	9	10	11	12	12	12
gbm	**2.5**	**2.5**	1	1	1	1
nb	12	12	12	10	10	10
glm	7.5	7.5	7	7	7	7
rpart	7.5	7.5	7	7	7	7
svmlinear	11	11	10	11	11	11
svmpoly	10	9	9	9	9	9

The result of classifiers performance for the page-block dataset and its ranks based on the proposed and existing measures shows that: i) The ranks of the classifiers based on Weighted TPR-TNR and GM measure are exactly similar to the ranks based on authors judgement; ii) The ranks of the classifiers based on GRA, TOPSIS, and MCC measure are different than the ranks based on authors judgement and not appropriate. For example, classification results clearly shows that performance of the 'nb' classifier is better than the 'svmradial' but it is not reflected in the ranks of GRA, TOPSIS, and MCC measures.

4.2.5. Experimental results for the abalone dataset

Classification results of 12 different classifiers for abalone dataset are shown in the Table 22. Classifiers performance using different measures for abalone dataset is shown in Table 23. The bold values in the table indicate the best value for each performance measure. Table 24 shows rank of the classifiers performance obtained using proposed and existing measures.

Table 22. Classification results for the abalone dataset

Classifiers	P	N	TP	TN	FP	FN
J48	4	170	4	163	7	0
C5.0	4	170	4	169	1	0
knn	4	170	0	170	0	4
rf	4	170	4	168	2	0
ctree	4	170	4	163	7	0
svmradial	4	170	0	170	0	4
gbm	4	170	2	165	5	2
nb	4	170	4	153	17	0
glm	4	170	4	167	3	0
rpart	4	170	4	163	7	0
svmlinear	4	170	4	167	3	0
svmpoly	4	170	4	168	2	0

Table 23. Classifiers performance for abalone dataset

Classifiers	Accuracy	TPR	TNR	FPR	FNR	Precision	Inverse precision	MCC	F Measure	GM	Jaccard	Weighted TPR-TNR
J48	0.9598	1.0000	0.9588	0.0412	0.0000	0.3636	1.0000	0.5905	0.5333	0.9792	0.3636	0.9991
C5.0	0.9943	1.0000	0.9941	0.0059	0.0000	0.8000	1.0000	0.8918	0.8889	0.9971	0.8000	0.9999
knn	0.9770	0.0000	1.0000	0.0000	1.0000	0.0000	0.9770	0.0000	0.0000	0.0000	0.0000	0.0230
rf	0.9885	1.0000	0.9882	0.0118	0.0000	0.6667	1.0000	0.8117	0.8000	0.9941	0.6667	0.9997
ctree	0.9598	1.0000	0.9588	0.0412	0.0000	0.3636	1.0000	0.5905	0.5333	0.9792	0.3636	0.9991
svmRadial	0.9770	0.0000	1.0000	0.0000	1.0000	0.0000	0.9770	0.0000	0.0000	0.0000	0.0000	0.0230
gbm	0.9598	0.5000	0.9706	0.0294	0.5000	0.2857	0.9880	0.3589	0.3636	0.6966	0.2222	0.5108
nb	0.9023	1.0000	0.9000	0.1000	0.0000	0.1905	1.0000	0.4140	0.3200	0.9487	0.1905	0.9977
glm	0.9828	1.0000	0.9824	0.0176	0.0000	0.5714	1.0000	0.7492	0.7273	0.9911	0.5714	0.9996
rpart	0.9598	1.0000	0.9588	0.0412	0.0000	0.3636	1.0000	0.5905	0.5333	0.9792	0.3636	0.9991
svmLinear	0.9828	1.0000	0.9824	0.0176	0.0000	0.5714	1.0000	0.7492	0.7273	0.9911	0.5714	0.9996
svmPoly	0.9885	1.0000	0.9882	0.0118	0.0000	0.6667	1.0000	0.8117	0.8000	0.9941	0.6667	0.9997

Table 24. Classifiers rank for abalone dataset

Classifiers	GRA rank	TOPSIS rank	MCC rank	GM Rank	Weighted TPR-FPR rank	Rank based on authors own judgment
J48	7	7	6	6	6	6
C5.0	1	1	1	1	1	1
knn	11.5	11.5	11	11	11	11
rf	2.5	2.5	2	2	2	2
ctree	7	7	6	6	6	6
svmradial	11.5	11.5	11	11	11	11
gbm	10	9	10	10	10	10
nb	9	10	9	9	9	9
glm	4.5	4.5	4	4	4	4
rpart	7	7	6	6	6	6
svmlinear	4.5	4.5	4	4	4	4
svmpoly	2.5	2.5	2	2	2	2

The results of classifiers performance for abalone dataset and its ranks based on the proposed and existing measures shows that: i) The ranks of the classifiers based on Weighted TPR-TNR, GM, MCC, and GRA measure are exactly similar to the ranks based on authors judgement; ii) The ranks of the classifiers based on TOPSIS method are slightly different than the ranks based on the authors judgement and not appropriate. For example, classification results show that performance of the 'nb' classifier is better than 'gbm', but it is not reflected in the ranks of TOPSIS method.

Thus, the classification result of all the five datasets shows that: i) ranks of the classifiers based on weighted TPR-TNR and GM measures are exactly similar to the ranks based on authors judgement for four datasets. The small change is observed in the first two and last two ranks only for one (pima) dataset; ii) ranks of the classifiers based on TOPSIS method are different than the ranks based on authors judgement for all five datasets; iii) ranks of the classifiers based on GRA method are different than the ranks based on authors judgement for four datasets; iv) ranks of the classifiers based on MCC measure are similar to the ranks based on authors judgement only for two datasets.

Therefore, on the basis of experimental results of all five datasets we can conclude that the ranks of the classifiers based on weighted TPR-TNR and GM measure are more appropriate than the ranks of the classifiers based on GRA, TOPSIS, and MCC measures.

4.3.Experiment 2

The second experiment is conducted using synthetic data. The reason for considering synthetic data is that we know in advance performance of the classifier and that gives us better understanding about suitability of the proposed weighted TPR-TNR method for evaluating and selecting the best classifier. We found that synthetic data has been used in the literature to validate the results [Koyejo, 2014; Boughorbel, 2017]. The Table 25 provides synthetic results of 15 different classifiers.

Table 25. Synthetic Classification results

Classifier	P	N	TP	TN	FP	FN
1	100	1000	95	995	5	5
2	100	1000	95	950	50	5
3	100	1000	90	980	20	10
4	100	1000	90	950	50	10
5	100	1000	85	875	125	15
6	100	1000	80	850	150	20
7	100	1000	75	900	100	25
8	100	1000	70	800	200	30
9	100	1000	40	850	150	60
10	100	1000	35	800	200	65
11	100	1000	30	900	100	70
12	100	1000	20	980	20	80
13	100	1000	15	950	50	85
14	100	1000	10	990	10	90
15	100	1000	5	990	10	95

Where: P: total positive cases; N: total negative cases; TP: True positives; TN: True negatives; FP: False positives; FN: False Negatives

Classifiers performance and rank of the classifiers obtained using proposed and existing measure for the synthetic data is shown in Table 26 and Table 27.

Table 26. Performance of the classifiers for synthetic data

Classifiers	Accuracy	TPR	TNR	FPR	FNR	Precision	Inverse precision	MCC	F Measure	GM	Jaccard	Weighted TPR-TNR
1	0.9909	0.95	0.995	0.005	0.05	0.95	0.995	0.945	0.95	0.9722	0.9048	0.954091
2	0.95	0.95	0.95	0.05	0.05	0.6552	0.9948	0.7648	0.7755	0.95	0.6333	0.95
3	0.9727	0.9	0.98	0.02	0.1	0.8182	0.9899	0.8433	0.8571	0.9391	0.75	0.907273
4	0.9455	0.9	0.95	0.05	0.1	0.6429	0.9896	0.7332	0.75	0.9247	0.6	0.904545
5	0.8727	0.85	0.875	0.125	0.15	0.4048	0.9831	0.5303	0.5484	0.8624	0.3778	0.852273
6	0.8455	0.8	0.85	0.15	0.2	0.3478	0.977	0.4595	0.4848	0.8246	0.32	0.804545
7	0.8864	0.75	0.9	0.1	0.25	0.4286	0.973	0.5109	0.5455	0.8216	0.375	0.763636
8	0.7909	0.7	0.8	0.2	0.3	0.2593	0.9639	0.334	0.3784	0.7483	0.2333	0.709091
9	0.8091	0.4	0.85	0.15	0.6	0.2105	0.9341	0.1901	0.2759	0.5831	0.16	0.440909
10	0.7591	0.35	0.8	0.2	0.65	0.1489	0.9249	0.1052	0.209	0.5292	0.1167	0.390909
11	0.8455	0.3	0.9	0.1	0.7	0.2308	0.9278	0.1781	0.2609	0.5196	0.15	0.354545
12	0.9091	0.2	0.98	0.02	0.8	0.5	0.9245	0.2764	0.2857	0.4427	0.1667	0.270909

13	0.8773	0.15	0.95	0.05	0.85	0.2308	0.9179	0.1219	0.1818	0.3775	0.1	0.222727
14	0.9091	0.1	0.99	0.01	0.9	0.5	0.9167	0.1936	0.1667	0.3146	0.0909	0.180909
15	0.9045	0.05	0.99	0.01	0.95	0.3333	0.9124	0.0992	0.087	0.2225	0.0455	0.135455

Table 27. Classifiers rank for synthetic data

Classifier	GRA rank	TOPSIS rank	MCC rank	GM Rank	Weighted TPR-TNR rank	Rank based on authors own judgment
1	1	1	1	1	1	1
2	3	3	3	2	2	2
3	2	2	2	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	7	7	7	6	6	6
7	6	6	6	7	7	7
8	11	9	8	8	8	8
9	14	14	11	9	9	9
10	15	15	14	10	10	10
11	13	13	12	11	11	11
12	8	8	9	12	12	12
13	12	12	13	13	13	13
14	9	10	10	14	14	14
15	10	11	15	15	15	15

From the Table 25 it is observed that imbalance ratio of the dataset is 10. It means that the cost of one incorrect prediction of the positive instance is approximately equivalent to cost of 10 incorrect predictions of negative instances. Considering this fact and using classifications results (TP, TN, FP, FN values in table 25), we have ranked classifiers and the ranks are shown in last column of the Table 27.

Performance of classifiers for the synthetic data and its ranks based on the proposed and existing measures shows that: i) The ranks of the classifiers based on weighted TPR-TNR and GM measures are exactly similar to the ranks based on authors judgement; ii) The ranks of the classifiers based on GRA, TOPSIS, and MCC method are different than the ranks of the classifiers based on the authors judgement and not appropriate. For example, Performance of the classifier 12 is better than the classifiers 9, 10, and 11 but it is not reflected in the ranks of GRA, TOPSIS, and MCC measures.

Thus, the synthetic data also shows that weighted TPR-TNR measure gives more appropriate ranks than the GRA, TOPSIS and MCC measures. It is also found that there is no difference in the ranks of the classifiers based on GM and Weighted TPR-TNR measures and these ranks are similar to the ranks of the classifiers based on author's judgement. The problem with GM measure is that its score could be zero when sensitivity (TPR) or specificity (TNR) value is zero. Therefore, we can conclude that proposed weighted TPR-TNR measure is more suitable in evaluating and selecting the best classifier when dataset is imbalanced.

5. Conclusion

Evaluation and selection of the best classifier is one of the most important tasks in solving classification problem. Evaluation of the classifiers using a single performance measure is easy and simple, but relying on a single performance measure is sometimes misleading when dataset is imbalanced. Therefore, MCDM methods such as TOPSIS and GRA have been used in literature to evaluate classifiers performance using different performance measures instead of relying on a single measure. In this study we have evaluated performance of different classifiers using GRA and TOPSIS methods by considering 11 different performance measures. The experimental results shows that the ranks of the classifiers calculated using GRA and TOPSIS methods are not as accurate as ranks of the classifiers calculated using proposed weighted TPR-TNR measure.

In this study we have also discussed: i) existing performance measures, its meaning, and how to assess performance of the classifiers using existing measures; ii) limitations of the existing performance measures when dataset is imbalanced; iii) new weighted TPR-TNR measure for evaluating classifiers performance that overcomes the limitations of the existing single valued performance measures.

The experimental results shows that the ranks of the classifiers based on the proposed weighted TPR-TNR measure are more appropriate than the existing measures because: i) it takes into account imbalanced ratio of the dataset while assigning weights to the TPR and TNR; ii) it takes into account the fact that cost of misclassification of positive and negative cases are different by assigning different weights to TPR and TNR; iii) it works well for both balanced as well as imbalanced dataset; iv) it gives appropriate results irrespective of whether majority cases in the imbalanced dataset are positive or negative.

We believe that although all performance measures assists decision makers to make more informed and appropriate decision, expertise of the decision makers always plays very important role in the final decision making.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit Author Statement

As there are no co-authors to this paper, all work is done my Anil S. Jadhav.

References

1. Ali, R., Lee, S., & Chung, T. C. (2017). Accurate multi-criteria decision making methodology for recommending machine learning algorithm. Expert Systems with Applications, 71, 257-278.
2. Behzadian, Majid, S. Khanmohammadi Otaghsara, Morteza Yazdani, and Joshua Ignatius (2012). A state-of the-art survey of TOPSIS applications. Expert Systems with applications 39, no. 17: 13051-13069.
3. Bogaert, Matthias, Justine Lootens, Dirk Van den Poel, and Michel Ballings (2019). Evaluating Multi-Label Classifiers and Recommender Systems in the Financial Service Sector. European Journal of Operational Research.
4. Brazdil, Pavel B., and Carlos Soares (2000). A comparison of ranking methods for classification algorithm selection. In European conference on machine learning, pp. 63-75. Springer, Berlin, Heidelberg.
5. Brown, Iain, and Christophe Mues (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications* 39, no. 3: 3446-3453.
6. Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS one, 12(6).
7. Chen, Shu-Jen, and Ching-Lai Hwang (1992). Fuzzy multiple attribute decision making methods. In Fuzzy multiple attribute decision making, pp. 289-486. Springer, Berlin, Heidelberg.
8. Davis, Jesse, and Mark Goadrich (2006). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning, pp. 233-240. ACM.
9. Demšar, Janez (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research 7, 1-30.

10. Deng, Ju-Long (1982). Control problems of grey systems. Sys. & Contr. Lett. 1, no. 5 , 288-294.
11. Fawcett, Tom (2006). An introduction to ROC analysis. Pattern recognition letters 27, no. 8, 861-874.
12. Garcia, Vicente, Ramon A. Mollineda, and J. Salvador Sanchez (2010). Theoretical analysis of a performance measure for imbalanced data. In 2010 20th International Conference on Pattern Recognition, pp. 617-620. IEEE.
13. Halligan, Steve, Douglas G. Altman, and Susan Mallett (2015). Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. European radiology 25, no. 4, 932-939.
14. Hand, David J (2012). Assessing the performance of classification methods. International Statistical Review 80, no. 3, 400-414.
15. Hwang, Ching-Lai, and Kwangsun Yoon (1981). Methods for multiple attribute decision making. In Multiple attribute decision making, pp. 58-191. Springer, Berlin, Heidelberg.
16. Hossin, Mohammad, and M. N. Sulaiman (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process 5, no. 2.
17. He, Haibo, and Edwardo A. Garcia (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, no. 9, 1263-1284.
18. Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013, September). Facing imbalanced data--recommendations for the use of performance metrics. In 2013 Humaine association conference on affective computing and intelligent interaction (pp. 245-251). IEEE.
19. Kuo, Yiyo, Taho Yang, and Guan-Wei Huang (2008). The use of grey relational analysis in solving multiple attribute decision-making problems. Computers & industrial engineering 55, no. 1, 80-93.
20. Koyejo, O. O., Natarajan, N., Ravikumar, P. K., & Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics. In Advances in Neural Information Processing Systems (pp. 2744-2752).
21. Lin, Chin-Tsai, Che-Wei Chang, and Chie-Bein Chen (2006). The worst ill-conditioned silicon wafer slicing machine detected by using grey relational analysis. The International Journal of Advanced Manufacturing Technology 31, no. 3-4, 388-395.

22. Luque, Amalia, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition 91, 216-231.
23. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, 405(2), 442-451.
24. Powers, David Martin (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.
25. Saito, Takaya, and Marc Rehmsmeier (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one 10 (3).
26. Sheskin, David J (2003). Handbook of parametric and nonparametric statistical procedures. Chapman and Hall/CRC.
27. Sokolova, Marina, Nathalie Japkowicz, and Stan Szpakowicz (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pp. 1015-1021. Springer, Berlin, Heidelberg.
28. Sokolova, Marina, and Guy Lapalme (2009). A systematic analysis of performance measures for classification tasks. Information processing & management 45, no. 4, 427-437.
29. Tharwat, Alaa (2018). Classification assessment methods. *Applied Computing and Informatics*.
30. Yin, Mu-Shang (2013). Fifteen years of grey system theory research: A historical review and bibliometric analysis. Expert systems with Applications 40, no. 7, 2767-2775.
31. Zhu, Bing, Bart Baesens, and Seppe KLM vanden Broucke (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information sciences* 408, 84-99.
32. Zou, Kelly H (2002). Receiver operating characteristic (ROC) literature research. *On-line bibliography available from:< http://splweb. bwh. harvard. edu* 8000.