Journal of Classification (2021) 38:16–26
https://doi.org/10.1007/s00357-020-09362-5

A New Performance Evaluation Metric for Classifiers: Polygon Area Metric

Onder Aydemir[1]

Published online: 25 January 2020

Abstract
Classifier performance assessment (CPA) is a challenging task for pattern recognition. In recent years, various CPA metrics have been developed to help assess the performance of classifiers. Although the classification accuracy (CA), which is the most popular metric in pattern recognition area, works well if the classes have equal number of samples, it fails to evaluate the recognition performance of each class when the classes have different number of samples. To overcome this problem, researchers have developed various metrics including sensitivity, specificity, area under curve, Jaccard index, Kappa, and F-measure except CA. Giving many evaluation metrics for assessing the performance of classifiers make large tables possible. Additionally, when comparing classifiers with each other, while a classifier might be more successful on a metric, it may have poor performance for the other metrics. Hence, such kinds of situations make it difficult to track results and compare classifiers. This study proposes a stable and profound knowledge criterion that allows the performance of a classifier to be evaluated with only a single metric called as polygon area metric (PAM). Thus, classifier performance can be easily evaluated without the need for several metrics. The stability and validity of the proposed metric were tested with the *k*-nearest neighbor, support vector machines, and linear discriminant analysis classifiers on a total of 7 different datasets, five of which were artificial. The results indicate that the proposed PAM method is simple but effective for evaluating classifier performance.

Keywords Classifier performance · Classification accuracy · Assessment metric · Polygon area metric

✉ Onder Aydemir
onderaydemir@ktu.edu.tr

[1] Department of Electrical and Electronics Engineering, Faculty of Engineering, Karadeniz Technical University, 61080 Trabzon, Turkey

1 Introduction

Pattern recognition and machine/deep learning have recently become active research areas due to their applications in a wide range of fields, including biomedical, smart device, and human-machine interface applications (Chu et al. 2011; Framinan et al. 2019). The success of such approaches strongly relates to the performance of the classifier, which can be assessed by various metrics. Classification accuracy (CA) is the most popular metric in pattern recognition. It works well if the classes have equal number of samples, but fails to evaluate the recognition performance of each class when the classes have different number of samples (Fawcett 2006). To overcome the limitation of the CA, researchers have developed other metrics including sensitivity (SE), specificity (SP), area under curve (AUC), Jaccard index (JI), kappa (K), and F-measure (FM) except CA. On the other hand, studies generally compare classifier performance via numerous metrics to determine the most suitable classifiers for specific problems. However, when comparing classifiers with each other, while a classifier might be more successful on a metric, it may have poor performance for the other metrics. Such kind of situation makes it difficult to determine the most successful classifier. For example, Aydemir and Kayikcioglu (2013) assessed the performances of five widely used classification algorithms in terms of four different metrics, including CA, SE, SP, and K, for low-dimensional feature vectors. They tested the classifiers using two real-world datasets, and the results of the metrics were given with a very large table. Comparing the performance of the classifiers at such a table is a challenging task. To do it easy, they calculated the average values of the performance metrics. However, they dramatically concluded that different classifiers achieved the best performance on different metrics. For instance, in a dataset they used, while support vector machines (SVM) obtained the best results on CA and K, *k*-nearest neighbor (*k*-NN) and naive Bayes achieved the best performance in terms of SE and SP, respectively. In another classifier-based study, Dixon and Brereton (2009) used six synthetic two-class datasets which consisted of an equal number of samples to compare five different classifiers. They only used the CA metric to evaluate the performances of classifiers. In another approach, Kim et al. (2017) aimed to develop machine learning models with strong prediction power and interpretability for the diagnosis of glaucoma based on retinal nerve fiber layer thickness and visual field. The dataset was recorded from patients who underwent optical coherence tomography. They tested four machine learning algorithms in terms of CA, SE, SP AUC, and likelihood ratio metrics. In order to determine the most suitable classifier of their proposed model, they required to assess the metrics with each other in detail, which might take some time. As a result, they concluded that random forest and SVM classifiers provided better performance than *k*-NN.

Existing performance measures have the relative advantage of being independent of class costs and prior probabilities. The aim of a classifier is to minimize the false-positive and false-negative rates or, similarly, to maximize the true-negative and true-positive rates. Unfortunately, there is a trade-off between false-negative rate and false-positive rate in most real-world applications and, similarly, between true-negative rate and true-positive rate. However, polygon area graphs can be used for analysis by showing six different metrics for a classifier with a single scalar.

In this study, we propose a novel, stable, and profound measure, called as polygon area metric (PAM), for evaluating the performance of a classifier using only a single scalar. It uses

the six existing metrics including CA, SE, SP, AUC, JI, and FM to generate a polygon, then calculates its area for PAM. The stability and validity of the PAM were tested with *k*-NN, SVM, and linear discriminant analysis (LDA) classifiers on a total of 7 different datasets, five of which were artificial.

This paper is organized as follows. Section 2 provides a description of the datasets. In Section 3, the performance evaluation metrics including CA, SE, SP, JI, AUC, and FM are introduced. After this section, the proposed polygon area metric is described. In Section 5, the results are presented. Multi-label polygon area metric is described in Section 6. Finally, in the last section, the paper concludes with a discussion of the results.

2 Description of Datasets

To approve the validity of the PAM, we used five artificially generated and two real-world datasets, which are described in the following subsections.

2.1 Artificially Generated Datasets

We utilized artificially generated data in two dimensions in order to illustrate graphically the selection of feature vectors. The distributions class 1 and class 2 samples were inspired by Dixon and Brereton (2009) and they are shown in Fig. 1. In this figure, plus points stand for samples of class 1 and circle points stand for the samples of class 2. The mean, variance, and number of samples (NoS) of each class are given in Table 1. It is worthwhile mentioning that we randomly selected half of the samples as training set and the rest of them as test set.

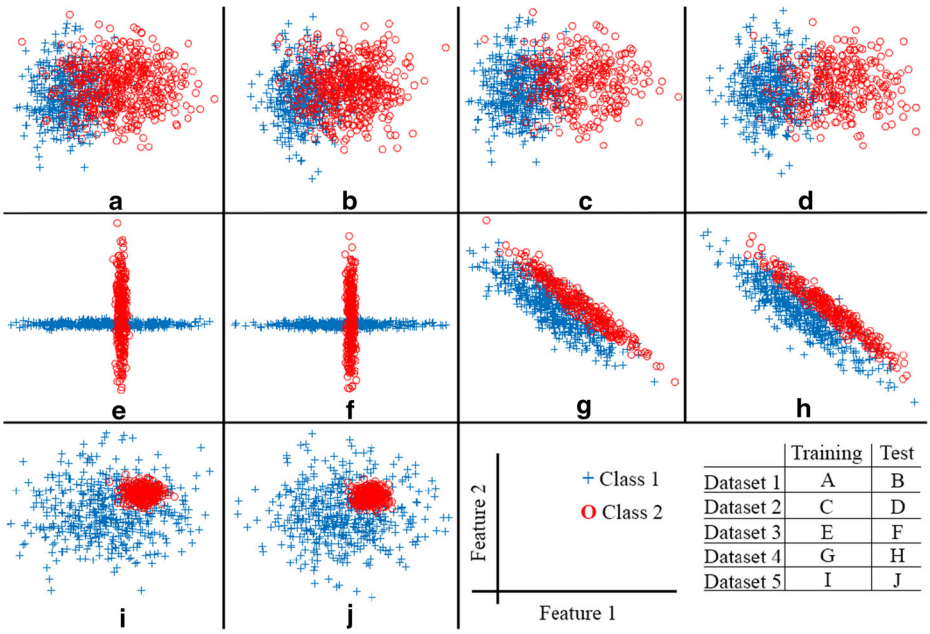

Fig. 1 The distributions of the datasets

Table 1 Distribution parameters of each dataset

Dataset	Class 1					Class 2				
	x-axis		*y*-axis		NoS	*x*-axis		*y*-axis		NoS
	Mean	Variance	Mean	Variance		Mean	Variance	Mean	Variance	
Dataset1	−2	16	3	36	1000	8	36	6	36	1000
Dataset2	−4	16	4	36	1000	8	36	6	36	600
Dataset3	0	64	0	0.25	1000	2	0.25	2	64	600
Dataset4	0	64	0	4	1000	4	64	4	1	600
Dataset5	0	16	1	16	1000	3	1	4	1	600

2.2 Real-world Datasets

2.2.1 Breast Cancer Dataset

The Wisconsin Breast Cancer Database (WBCD) dataset from the UCI Machine Learning Repository has been used as real-world dataset. It contains 569 samples taken from needle aspirates from human breast cancer tissue, of which 357 cases belong to benign class (class 1) and the remaining 212 of which are malignant (class 2) cases. Each sample has 32 features, the first two of which correspond to a unique identification number and diagnostic state (ID, diagnosis (benign/malignant), followed by 30 real-valued input features). The remaining 30 attributes are used for classification (William et al. 1995).

2.2.2 Electrocorticogram-Based Brain-Computer Interface Dataset

The second real-world dataset was an electrocorticogram (ECoG)-based brain-computer interface (BCI) dataset. The original name of this dataset was the *BCI Competition 2005 Dataset I*, which was taken from an epilepsy subject on two different days with about 1 week of delay. In both sessions, the subject was asked to imagine of either the left small finger (class 1) or the tongue movement (class 2). The dataset consists of 278 training trials (139 trials for finger movements, 139 trials for tongue movements), performed during the first session and 100 test trials (50 trials for finger movements, 50 trials for tongue movements), performed from the second session. Each trial's duration was 3 s. Electrical brain activity was recorded by an 8×8 ECoG platinum electrode grid (totally from 64 points) placed on the contralateral (right) motor cortex (Lal et al. 2005). The purpose was to categorize the trials in the test set as finger or tongue movement imagery.

The features were extracted from only channel 12 and channel 39 by wavelet transform. After a variance normalization process was implemented to all the trials, we calculated the wavelet transform coefficients (WTCs) of the related channels. For the feature vector, we calculated the averages for channel 12 (feature 1) and the standard deviations for channel 39 (feature 2) of the absolute values of the WTCs. It is worth mentioning that the Morlet was used as mother wavelet function. The scale of the Morlet function was set to integer values between 1 and 90 with a step size of 3, as proposed in (Aydemir and Kayikcioglu 2011). The extracted features are shown in Fig. 2.

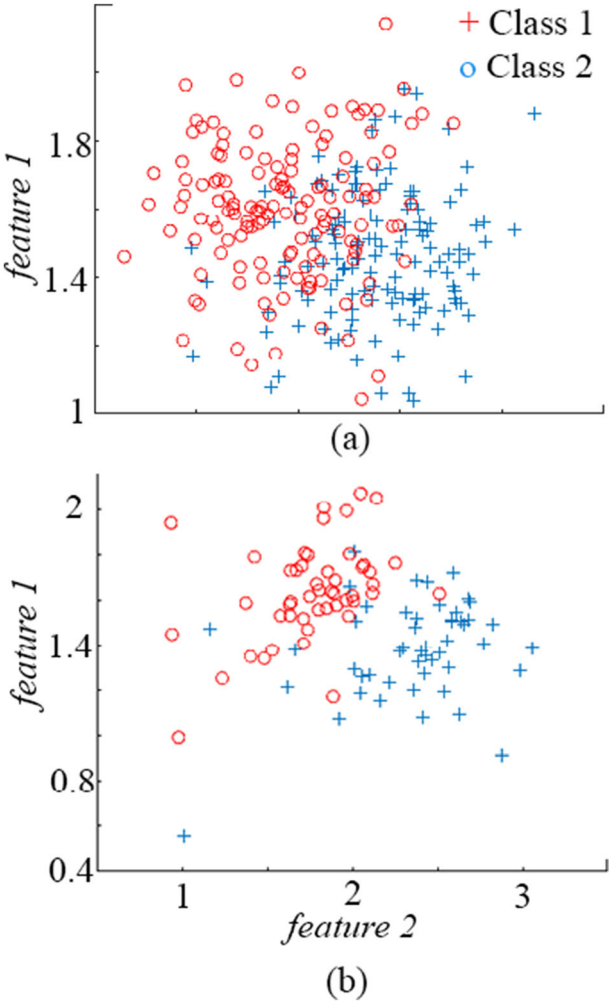

Fig. 2 Feature vectors. **a** Training dataset. **b** Test dataset

3 Performance Evaluation Metrics

One of the most informative ways to assess performance of classifiers is based on confusion matrix analysis (Ohsaki et al. 2017). Table 2 shows a confusion matrix for a two-class problem with class labels negative and positive.

Table 2 Confusion matrix for a two-class problem

		Predicted label	
		Positive	Negative
Actual label	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

In this table TP, TN, FP, and FN are respectively defined as the number of positive samples correctly predicted, the number of negative samples correctly predicted, the number of positive samples incorrectly predicted, and the number of negative samples incorrectly predicted. The researchers calculate number of commonly used metrics from confusion matrix for evaluating machine learning systems performance, including CA, SE, SP, AUC, JI, and FM (Shiferaw et al. 2019). The mathematical definitions are respectively given as follows:

$$\mathrm{CA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{3}$$

$$\mathrm{JI} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{F} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{AUC} = \int_0^1 f(x)dx \tag{6}$$

where $f(x)$ is a receiver operating characteristic curve that the true-positive rate (SE) is plotted in function of the false-positive rate (1-SP) for different cut-off points. It is worth mentioning that SE refers to the ratio of correctly classified class 1 samples to the total population of class 1 samples, and SP is the ratio of correctly classified class 2 samples to the total population of class 2 samples.

4 Polygon Area Metric

The PAM is calculated using the area of the polygon that CA, SE, SP, AUC, JI, and FM points create in a regular hexagon, as illustrated in Fig. 3. It should be noted that the regular hexagon is made up of 6 equilateral triangles and the length of each side is equal to 1. Hence, it can be said that $|\mathrm{OA}| = |\mathrm{OB}| = |\mathrm{OC}| = |\mathrm{OD}| = |\mathrm{OE}| = |\mathrm{OF}| = 1$, while the area of the regular hexagon is equal to 2.59807. The lengths of |OA|, |OB|, |OC|, |OD|, |OE|, and |OF| represent the values of CA, SE, SP, AUC, JI, and FM, respectively. The PAM is calculated using the following formula:

$$\mathrm{PAM} = \frac{\mathrm{PA}}{2.59807} \tag{7}$$

where PA is the area of the polygon. It is worthwhile mentioning that in order to normalize the PAM into the [0, 1] interval the PA value is divided by 2.59807.

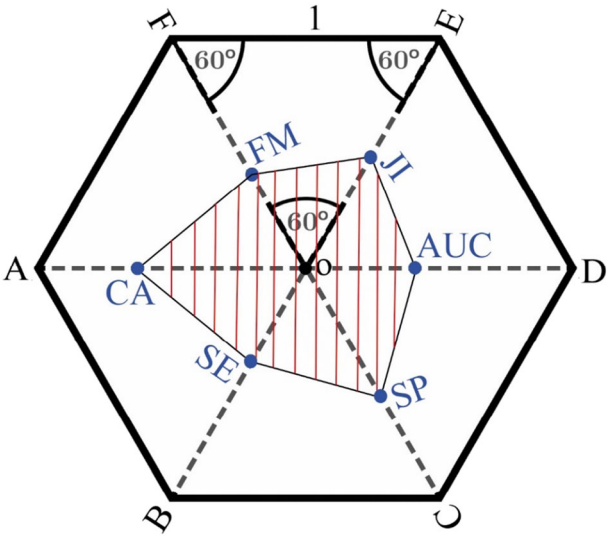

Fig. 3 The created polygon in a regular hexagon

5 Results

The performance of the proposed metric can be demonstrated by comparing its results with those of existing metrics. The comparison results of considered metrics were calculated for seven different datasets, five of which were artificial and two of which were real-world datasets. The obtained results for artificial and real-world datasets are tabulated in Tables 3 and 4, respectively. Additionally, the visual results are given in Figs. 4 and 5. As seen from the tables, the existing metrics, including CA, SE, SP, AUC, JI, and FM, have different values for each dataset and classifier. This may make it difficult to track the results, compare the classifiers, and evaluate their individual performances. However, by considering the PAM, it

Table 3 The results of artificial datasets

Dataset	Classifier	PAM	CA	SE	SP	AUC	JI	FM
Dataset1	SVM	0.69	0.85	0.89	0.80	0.85	0.74	0.85
	k-NN	0.66	0.83	0.87	0.79	0.83	0.72	0.84
	LDA	0.68	0.84	0.88	0.80	0.84	0.74	0.85
Dataset2	SVM	0.76	0.88	0.90	0.85	0.88	0.83	0.90
	k-NN	0.75	0.88	0.92	0.81	0.87	0.83	0.91
	LDA	0.76	0.88	0.91	0.84	0.87	0.83	0.91
Dataset3	SVM	0.92	0.96	0.98	0.94	0.96	0.94	0.97
	k-NN	0.93	0.97	0.98	0.95	0.97	0.95	0.98
	LDA	0.34	0.60	0.63	0.55	0.59	0.50	0.66
Dataset4	SVM	0.79	0.89	0.86	0.95	0.90	0.83	0.91
	k-NN	0.77	0.88	0.88	0.88	0.88	0.82	0.90
	LDA	0.79	0.89	0.83	0.98	0.91	0.82	0.90
Dataset5	SVM	0.85	0.92	0.90	0.97	0.93	0.88	0.94
	k-NN	0.84	0.92	0.90	0.95	0.93	0.88	0.93
	LDA	0.64	0.80	0.69	0.99	0.84	0.69	0.81

Table 4 The results of real-world datasets

Dataset	Classifier	PAM	CA	SE	SP	AUC	JI	FM
WBCD	SVM	0.96	0.98	1.00	0.95	0.98	0.97	0.99
	k-NN	0.95	0.98	0.99	0.95	0.97	0.96	0.98
	LDA	0.93	0.97	1.00	0.92	0.96	0.96	0.98
BCI	SVM	0.82	0.92	0.88	0.96	0.92	0.85	0.92
	k-NN	0.74	0.88	0.84	0.92	0.88	0.78	0.88
	LDA	0.78	0.90	0.84	0.96	0.90	0.81	0.89

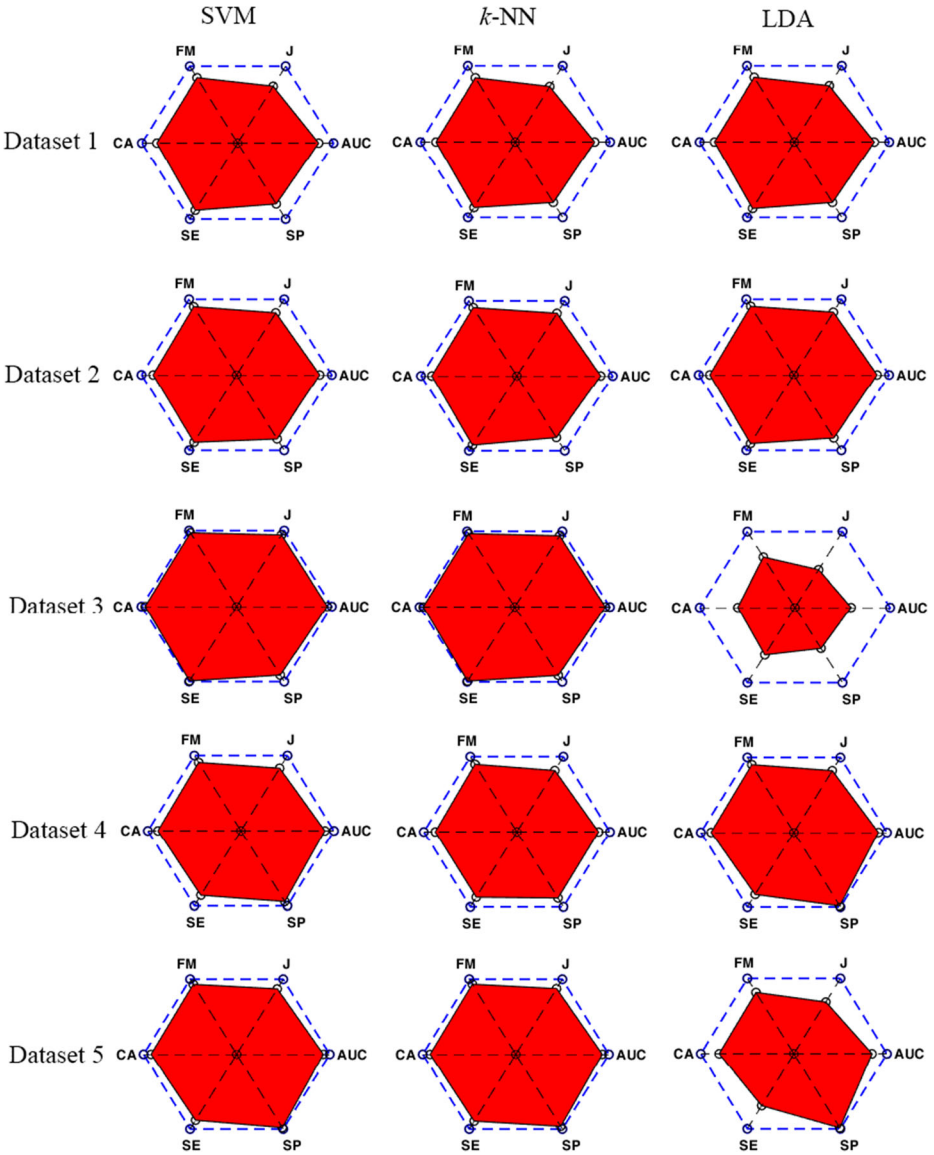

Fig. 4 Artificial dataset polygon area graphs

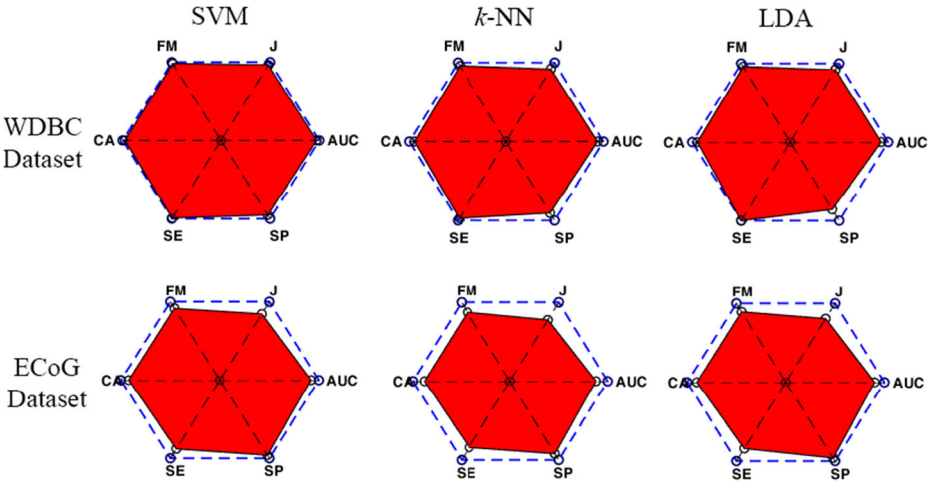

Fig. 5 Real-world dataset polygon area graphs

is more efficient to assess the performance of the classifiers. Additionally, the visual graphs can be examined for detailed evaluation.

In addition to the artificially generated and the real-world datasets, we also calculated PAM for the conditions where all samples predicted randomly, completely correct (the best condition), completely incorrect (the worst condition), as class 1 (all C1) and as class 2 (all C2). These results are given in Table 5. Because the number divided by zero is undefined, we obtained Not-A-Number (NaN) for the worst and all C2 conditions of FM. Hence, we could not calculate PAM value for those conditions. On the other hand, for the best condition, we obtained 1.00 for all metrics. Additionally, the table shows that the PAM value individually has a potential to assess the classification performance for the random and All C1 conditions. As a result, it can be said that PAM is a very powerful metric for assessing the performance of a classifier.

The computational time for calculating the PAM for 1000 test samples was measured as 8.2 ms. All runtime experiments were conducted on a desktop PC with an Intel Core i7 CPU at 1.73 GHz with 4 GB of RAM.

Table 5 The results of specific conditions

Condition	Number of samples		PAM	CA	SE	SP	AUC	JI	FM
	Class 1	Class 2							
Random	100	100	0.22	0.50	0.50	0.50	0.50	0.33	0.50
	400	100	0.26	0.50	0.50	0.50	0.50	0.44	0.62
The best	100	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	400	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00
The worst	100	100	NaN	0.00	0.00	0.00	0.00	0.00	NaN
	400	100	NaN	0.00	0.00	0.00	0.00	0.00	NaN
All C1	100	100	0.24	0.50	1.00	0.00	0.50	0.50	0.67
	400	100	0.44	0.80	1.00	0.00	0.50	0.80	0.89
All C2	100	100	NaN	0.50	0.00	1.00	0.50	0.00	NaN
	400	100	NaN	0.20	0.00	1.00	0.50	0.00	NaN

6 Multi-label Polygon Area Metric

Although the PAM is mostly suitable for binary classification problems, it could be extended to multi-label (PAM_{ML}) classification approaches. To do this, it is necessary to carry out pairwise binary comparison (one class versus all other classes). While reference class is assigned as the positive, all other classes are assigned as the negative class. Therefore, for given K classes, K different $PAM(k_i)$ ($i = 1, 2, \ldots K$) values are calculated, one for each reference class. In order to be more sensitive to the performance for individual classes, each $PAM(k_i)$ is multiplied by a weight $w(k_i)$, which is calculated for every class such that $\sum_{i=1}^{K} w(k_i) = 1$. Then, the multiplication results are summed as shown in Eq. 8:

$$\mathrm{PAM_{ML}} = \sum_{k \in K} \mathrm{PAM}(k_i) \times w(k_i) \tag{8}$$

Note that the weight is obtained as follows:

$$w(k_i) = \frac{N(k_i)}{M(K)} \tag{9}$$

where $N(k_i)$ is the number of observations of class (k_i) and $M(K)$ is the total number of observations of all classes. It should be mentioned that the higher the value of $w(k_i)$for an individual class, the greater is the effect of observations from that class on the $\mathrm{PAM_{ML}}$.

7 Conclusion

In this paper, we have introduced an objective PAM for easily assessing the performance of classifiers. The performance of the proposed metric was validated by comparing its results with state-of-the-art metrics against the same set of benchmark datasets. The results indicated that although the PAM is a single value, it includes more information from CA, SE, SP, AUC, JI, and FM metrics. The simple and effective nature of PAM makes it promising for the evaluation of the performance of classifiers in pattern recognition and machine/deep learning applications. In conclusion, the proposed PAM can be able to evaluate the performance of a classifier with or without the use of existing metrics.

There are two main limitations of PAM, which should be addressed. Firstly, PAM produces the NaN value when any of the considered metrics (CA, SE, SP, AUC, JI, and FM) is equal to NaN. Moreover, it is not known which of the metric has NaN value. But it is worth mentioning that this is clearly revealed by polygon area graph. Secondly, unlike confusion matrix, it does not provide information about exact values of TP, TN, FP, and FN, which could be important to figure out the lack of pattern recognition model. Although it has a few drawbacks, I believe that the PAM contribute to pattern recognition and machine/deep learning community for better classifier evaluation.

References

Aydemir, O., & Kayikcioglu, T. (2011). Wavelet transform based classification of invasive brain computer interface data. *Radioengineering, 20*(1), 31–38.

Aydemir, O., & Kayikcioglu, T. (2013). Comparing common machine learning classifiers in low-dimensional feature vectors for brain computer interface applications. *International Journal of Innovative Computing Information and Control, 9*(3), 1145–1157.

Chu, C., Ni, Y., Tan, G., Saunders, C. J., & Ashburner, J. (2011). Kernel regression for fMRI pattern prediction. *NeuroImage, 56*(2), 662–673.

Dixon, S. J., & Brereton, R. G. (2009). Comparison of performance of five common classifiers represented as boundary methods: Euclidean distance to centroids, linear discriminant analysis, quadratic discriminant analysis, learning vector quantization and support vector machines, as dependent on data structure. *Chemometrics and Intelligent Laboratory Systems, 95*(1), 1–17.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Framinan, J. M., Perez-Gonzalez, P., & Fernandez-Viagas, V. (2019). Deterministic assembly scheduling problems: a review and classification of concurrent-type scheduling models and solution procedures. *European Journal of Operational Research, 273*(2), 401–417.

Kim, S. J., Cho, K. J., & Oh, S. (2017). Development of machine learning models for diagnosis of glaucoma. *PLoS One, 12*(5), e0177726.

Lal, T. N., Hinterberger, T., Widman, G., Schröder, M., Hill, N. J., Rosenstiel, W., Elger, C. E., Schölkopf, B., & Birbaumer, N. (2005). Methods towards invasive human brain computer interfaces. *Advances in Neural Information Processing Systems (NIPS) 17, MA, Cambridge:MIT Press*, 737–744.

Ohsaki, M., Wang, P., Matsuda, K., Katagiri, S., Watanabe, H., & Ralescu, A. (2017). Confusion-matrix-based kernel logistic regression for imbalanced data classification. *IEEE Transactions on Knowledge and Data Engineering, 29*(9), 1806–1819.

Shiferaw, H., Bewket, W., & Eckert, S. (2019). Performances of machine learning algorithms for mapping fractional cover of an invasive plant species in a dryland ecosystem. *Ecology and Evolution, 9*(5), 2562–2574.

William H. Wolberg W. H., Street W. N. and Mangasarian O. L., (1995) UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).