World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

A Comparative Study of Several Classification Metrics and Their Performances on Data

Jude Chukwura Obi *

Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Anambra State, Nigeria.

World Journal of Advanced Engineering Technology and Sciences, 2023, 08(01), 308–314

Publication history: Received on 06 January 2023; revised on 14 February 2023; accepted on 17 February 2023

Article DOI: https://doi.org/10.30574/wjaets.2023.8.1.0054

Abstract

Six classification metrics namely, Accuracy, Precision, Recall (Sensitivity), Specificity, F1-Score and Area Under the Curve have been studied in this work. A classification model based on the Support Vector Machine, was used to obtain a confusion matrix, which provided the needed information for calculating the different classification metrics. Twenty different datasets were used to assess the performances of the classification metrics. Accuracy and Area Under the Curve are the two metrics that consistently gave a classification result given each dataset used in the study. Although accuracy appears to be marginally better that AUC, it was discovered that in some cases where sensitivity is zero, accuracy yielded a high correct classification result. This goes further to implying that prior to choosing accuracy as a preferred metric for classification, investigation should be carried out to find out what sensitivity and specificity are. Where there are high values for sensitivity and specificity, the study shows that a choice of accuracy as a preferred classification metric leads to a high percentage of correct classification result.

Keywords: Classification Metrics; Machine Learning; Confusion Matrix; Support Vector Machines

1. Introduction

Classification can be defined as the prediction of class outcome variables (Johnson et al., 2002), using a classification function. Notable examples of a classification functions include the Fisher's Discriminant Analysis (FDA), Support Vector Machine (SVM) and the Logistic Regression, etc. A classification metric is constructed on any classification function and it is the outcome of discrimination. Discrimination on its part refers to the use of a set of labelled classes, otherwise called training set to construct a classifier (or allocation rule) that separate the predefined classes as much as possible (Izenman, 2008). Put differently, discrimination is concerned with the problem of class separation, whereas classification aims to allocate unlabeled input to a class it belongs. For instance, consider a training set $D = \{(x_i, y_i)\}_{i=1}^{n}$, and assuming that we partition it into K labelled classes c_k, where $k = 1,2, \cdots, K$. The goal of classification is to take an input vector $\mathbf{x}_i$, and assign it to one of the K classes (Bishop, 2007). The classes are assumed to be disjoint meaning that each $\mathbf{x}_i$ is assigned to one and only one class.

The problem often encountered in classification concerns making a suitable choice of a classification metric. The success of a classification metric is often tied to data, because different classification metrics perform differently given different datasets. A classification metric can be constructed using different methods, but in most cases, the merits and demerits of each one is often data dependent.

The classification metrics to be review will include the accuracy rate (conversely called the error rate), precision, recall (sensitivity or the true positive rate), F1-score, Specificity and the receiver operating characteristic (ROC) curve/area under the ROC curve.

* Corresponding author: Jude Chukwura Obi

Prior to reviewing these metrics, it is important to note that each of them has a connection to the confusion matrix. For this reason, I shall commence with reviewing what confusion matrix is.

1.1. Confusion Matrix

A confusion matrix is an $n \times n$ table with information about the predictions of a classification model, vis-à-vis the actual observations based on data. For a 2×2 table, a confusion matrix contains information about the true positive (TP), true negative (TN), false positive (FP) and the false negative (FN). A reference on a clearer definition of a confusion matrix can be found in (Wikipedia contributors, 2022a). Table 1 that follows is a schematic illustration of a confusion matrix.

Table 1 Schematic illustration of a confusion matrix

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

1.2. Accuracy Rate (opposite of the Error Rate)

The accuracy rate can be defined as the number of correct classifications over the entire test set or conversely the fraction of correct prediction of a classifier, over the entire test set. The accuracy rate is the opposite of the error rate and both rates give the same information about the strength or weakness of a classifier. Symbolically, (Zaki & Wagner, 2014) defined the rate as follows:

$$\text{Accuracy Rate} = \frac{1}{m}\sum_{i=1}^{m} I(y_i = \hat{y}_i), \tag{1.1}$$

Where I in an indicator function with value1 if the argument is true, otherwise it has value zero. m is the size of the test set. With regards to the error rate, (1.1) will be conversely stated as:

$$Error\ Rate = \frac{1}{m}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{1.2}$$

One disadvantage of the error rate, as well as the accuracy rate is that it does not provide information on the performance of a classifier in each class separately. It rather gives information on the overall performance of a classifier given the entire test set. As a result, one lacks the ability to assess how a classifier has performed in each c_k class.

Based on Table 1;

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{Error Rate} = \frac{FP + FN}{TP + TN + FP + FN}.$$

1.3. Precision

Precision is the ratio of the true positive over all the positives observed. All the positives here include the true positives observed, plus some true negatives wrongly observed as positives (FP). Symbolically,

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{1.3}$$

Consider a model that gives a precision of 0.84, for instance; it means that the model is correct with its prediction (prediction of the true positive) 84% of the time. In fact, based on (1.3), precision gives a fraction of the true positive predicted given a classification model. It answers the question; of all the positive predictions, what fraction is truly positive.

1.4. Recall

Recall, also known as sensitivity or the true positive rate, is the ability of a model to correctly identify the true positive. It is mathematically defined as:

$$Recall = \frac{TP}{TP + FN} \tag{1.4}$$

Recall answers the question; of all the positives (all the positives here include the true positives plus the true positive wrongly observed as negatives (FN)), what percentage is truly positive? High recall means that the observed true positives are relatively higher than observed false negative and low recall means the opposite.

1.5. F1-Score

The F1-Score is the harmonic mean or the weighted average of precision and recall. By definition, it is a measure of a test's accuracy, and calculated from the precision and recall of the test. As mentioned already, the precision is the number of correctly identified positive results divided by the number of positive results, including those not identified correctly, and recall is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive (Wikipedia contributors, 2022b). For this reason, the F1-score takes both false positive and false negative into account. Compared to the accuracy, the F1-score is preferred particularly where cost is involved or uneven classes are involved. Where the false positive and false negative have similar costs, accuracy is preferred. Symbolically,

$$\text{F1 Score} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{1.5}$$

1.6. Specificity

Specificity is used to measure the fraction of true negatives correctly predicted as negatives by a given classifier. Consider, for instance, a model used to classify people who have headache and those without headache (negatives). Typically, if there is high specificity, it means that the model is a good one, because it has the ability to isolate people without the disease from those with the disease. Symbolically,

$$Specificity = \frac{TN}{TN + FP} \tag{1.6}$$

Based on (1.6), if $FP \rightarrow 0$, $Specificity \rightarrow 1$, and the model in question is a useful one.

1.7. Receiver Operating Characteristic (ROC) Curve/Area Under the ROC Curve

The ROC curve is a graph that shows the performance of a classification model at all classification thresholds. The curve consists of the plot of True Positive Rate (TPR) versus the False Positive Rate (FPR) for all classification thresholds. Note that:

$$TPR\ (Recall) = \frac{TP}{TP + FN}, \text{ and} \tag{1.7}$$

$$FPR = \frac{FP}{FP + TN} \tag{1.8}$$

The Area Under the ROC Curve (AUC) gives information on the overall performance of a classification model. The least AUC can assume is 0, meaning that a given classification model performed woefully badly, whereas the biggest value it assumes is 1. In this case, we have a highest performance rate of a given classification model. Figures 1 and 2 give the diagrams of ROC curve and AUC respectively.

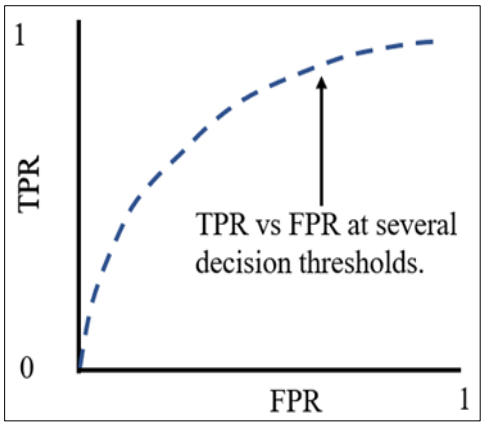

Figure 1 TPR vs FPR at different classification thresholds denoted by the dotted lines

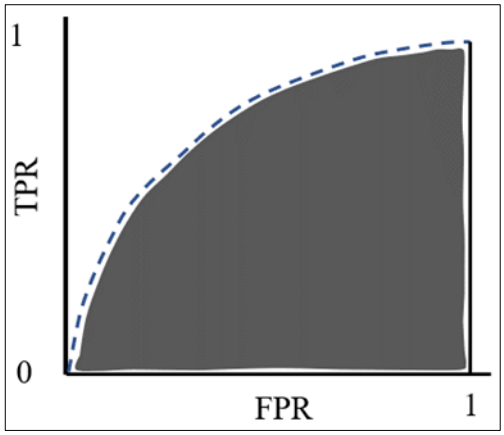

Figure 2 Area under the ROC curve

2. Research Aim and Objectives

The aim is to find out among the reviewed classification metrics, the one that comparatively is better. The objectives include the following:

- To create awareness that the error rate or alternatively the accuracy rate is not the only metric for measuring the performances of a classifier.
- To show how thorough understanding of a confusion matrix can aid in obtaining any desired classification metric.
- To show some appealing properties of the ROC curve and consequently the AUC.

3. Research Methodology

This research will involve the use of several classification datasets. Each dataset will be split into training set (70%) and test set (30%). A classification model (classifier) based on the support vector machine will be constructed using the training set and with the test set, the classifier is evaluated. Thereafter, a unique confusion matrix will be obtained for each dataset, which in turn forms a basis for calculation of the classification metrics, namely the accuracy rate, F1 score and the area under the ROC curve. The omission of precision and recall is due to the fact that F1 score is a function of both. Again, with F1 score, a higher score is a reflection that the classifier concerned performs better given the dataset.

Further discussions in this section will include the datasets to be used in this study, a classification model based on the support vector machine, and some inferential procedures.

3.1. Datasets

The datasets to be used include Appendicitis, Australia, Coil2000, Handheight, Heart, Heberman, Hepatitis, HVWN, Ionosphere, Magic, Mammographic, Parkinsons, Ringnorm, Saheart and WDBC. Detailed descriptions of the dataset are contained in (Jude, 2019). Additional datasets to use are Colon, Gisette, Prostate, Sonar and Twonorm.

3.1.1. Colon

Colon is a gene expression dataset from the microarray experiments of colon tissue samples (Alon et al., 1999). The dataset consists of 62 samples and 2000 genes (features). It has two classes namely tumour tissue with 40 samples, and normal tissue with 22 samples. It is contained in plsgenomics package in R.

3.1.2. Gisette

The dataset is one of five datasets used in the NIPS 2003 feature selection challenge, and was put together by (Guyon, 2003). It is also contained in the UCI Machine Learning repository.

3.1.3. Prostrate

Prostrate dataset is a gene expression dataset (Singh et al., 2002). The dataset is contained in R package spls and consists of two classes namely, 52 prostrate tumour and 50 normal classes. The number of genes involved is 6033.

3.1.4. Sonar

This dataset contains signals obtained from a variety of different aspect angles, spanning 90 degrees for mines and 180 degrees for rocks. Each pattern is a set of 60 numbers in the range 0.0 to 1.0, where each number represents the energy within a particular frequency band, integrated over a certain period of time. The output attribute contains the letter +1 if the object is a rock and −1 if it is a mine (metal cylinder). The source is UCI Machine Learning Repository.

3.1.5. Twonorm

This dataset is 20 dimensional, and consists of 2 classes. Each class is drawn from a multivariate normal distribution. Class +1 has mean (a, a, ..a) while Class −1 has mean (−a, −a, .. − a)..a = 2/sqrt(20). The dataset is contained in the KEEL dataset repository.

3.2. Support Vector Machine (SVM)

The support vector machine (Cortes and Vapnik, 1995) is a binary classifier from the field of machine learning. It has a strong geometric appeal, with a concept based on the hyperplane. A hyperplane is a set of points $h(x)$ (Zaki and Wagner Meira, 2014), where $h(x)$ is a function of the hyperplane defined by

$$h(x) = \mathrm{W}^T\mathrm{x} + w_0 \tag{3.1}$$

Note that W is a p-dimensional weight vector and w_0 is a scalar. The SVM is implemented in R via the use of e1071 package. The function svm contained in the package is used to construct a svm model, which subsequently is used for classes allocations.

3.3. Some Inferential Procedures

The SVM will be used to construct a classification model for each dataset used in the study. With such model, a confusion matrix is obtained, which in turn helps to work out the three needed classification metrics on each dataset. Over all the datasets used in the study, efforts will be made to find out if there are differences among the results of the classification metrices or whether they are all essentially the same.

4. Results

4.1. Data Analysis/Result Presentation

The output of six classification metrics on twenty different datasets is contained on Table 1. The R codes used for data analysis is shown in the Appendix. Based on Table 1, the use of accuracy and AUC consistently output a result. With majority of the datasets, it seems that the output of accuracy is relatively higher, followed by AUC and both sensitivity and specificity appear to tally in output. It is possible to observe a direct dependence of Precision, Recall (sensitivity) and F1-Score on the TP prediction. If a model is not able to predict the TP, sensitivity will be zero, and there is no output for both Precision and F1-Score.

Apart from the dataset Hepatitis, it seems that whenever sensitivity is zero, the output of accuracy is small. Of all the datasets examined, the one that presented appealing output given all the classification metrics is Twonorm. Here, there is high output from all the classification metrices. This result may not be surprising because the dataset particularly follows a normal distribution, since it was simulated under a multivariate normal condition.

It is noteworthy that in as much as accuracy appears to be a better metric by virtue of its output, caution should be applied using it. For instance, with datasets Australia, Hepatitis, WDBC and Gisette; TP is zero, meaning that sensitivity is zero, yet there is output for accuracy. Now, the big question is, to what extent do we rely on a metric that could possibly not predict the TP at some times? For this reason, my view is that prior to considering accuracy as a choice metric, it may be necessary to know what sensitivity and specificity are. A model that outputs zero sensitivity may never be considered as a better model irrespective of its level of accuracy.

Table 1 Output of six classification metrics on twenty different datasets

S/No.	Dataset	Dim	Accuracy	Precision	Sensitivity	Specificity	F1 Score	AUC
1	Appendicitis	106 × 8	0.76	0.76	1.0	0.0	0.86	0.5
2	Austraria	482 × 15	0.56	NaN	0.0	1.0	NaN	0.5
3	Coil2000	6875 × 86	0.92	0.14	0.05	0.98	0.07	0.51
4	Handheight	117 × 3	0.86	1.0	0.74	1.0	0.85	0.87
5	Heart	176 × 14	0.55	0.55	1.0	0.0	0.71	0.5
6	Heberman	215 × 4	0.69	0.75	0.13	0.98	0.22	0.55
7	Hepatitis	44 × 20	0.79	NaN	0.0	1.0	NaN	0.5
8	HVWN	606 × 101	0.52	0.74	0.2	0.91	0.31	0.56
9	Ionosphere	245 × 33	0.96	1.0	0.89	1.0	0.94	0.94
10	Magic	13314 × 11	0.66	0.90	0.02	1.0	0.04	0.51
11	Mammographic	581 × 6	0.77	0.79	0.78	0.76	0.79	0.77
12	Parkinsons	137 × 23	0.83	0.67	0.25	0.97	0.36	0.61
13	Ringnorm	5180 × 21	0.51	0.51	1.0	0.0	0.67	0.5
14	Saheart	323 × 10	0.59	0.59	1.0	0.0	0.74	0.5
15	WDBC	398 × 31	0.65	NaN	0.0	1.0	NaN	0.5
16	Colon	43 × 2001	0.71	0.71	1.0	0.0	0.83	0.5
17	Gisette	6000 × 5001	0.49	NaN	0.0	1.0	NaN	0.5
18	Prostate	71 × 6034	0.9	1.0	0.8	1.0	0.89	0.9
19	Sonar	146 × 61	0.69	0.67	0.87	0.47	0.75	0.67
20	Twonorm	5180 × 21	0.98	0.97	0.98	0.97	0.98	0.98

5. Conclusion

I have so far considered six classification metrics, namely Accuracy, Precision, Sensitivity, Specificity, F1-Score and the AUC. I have equally demonstrated how the computation of each of the metrics is dependent on information from the confusion matrix. In other words, the moment the confusion matrix is obtained, any of the six given classification metrics can easily be calculated.

The accuracy appeared to have shown improved performance over AUC, although the two metrics consistently outputs a result even on datasets where others could not give any result. In spite of this marginal gain over AUC, caution is still advised prior to using it. The reason is because in some cases where sensitivity is zero, we still see accuracy outputting a higher classification result, and example here is with the dataset Hepatitis. Now, if you consider that with this dataset, the classification model could not predict any positive case (those individuals who have Hepatitis), but only predicted majority of those without the disease, you cannot claim that such model is entirely a good one.

Based on this information, I am of the view that prior to choosing accuracy as a preferred classification metric, it is important to know what sensitivity and specificity of the given classification problem are. If there is high correct classification result with both sensitivity and specificity, accuracy can be adjudged a preferred metric. Example here is seen with the dataset Twonorm, because here, we have high correct classification result with sensitivity and specificity, and accuracy also gave a high correct classification result of 98%.

Compliance with ethical standards

Acknowledgments

I heartily acknowledge the support from my family that offered me the opportunity to complete this research on schedule. My students at COOU are invaluable assets academically because my regular interaction with them often provoke several research topics, of which this one is one of them. Last, but not the least, I appreciate the editorial board of World Journal of Advanced Engineering Technology and Sciences, for accepting to publish my work on their website.

References

[1] Bishop, C. (2007). Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn. Springer, New York.

[2] Izenman, A. J. (2008). Modern Multivariate Statistical Techniques (Vol. 1). Springer.

[3] Johnson, R. A., Wichern, D. W., & others. (2002). Applied Multivariate Statistical Analysis. Prentice hall Upper Saddle River, NJ.

[4] Jude, C. O. (2019). On The Effectiveness of Lasso Discriminant Analysis (LADA) In Variable Selection. COOU Journal of Physical Sciences, 2(8), 228.

[5] Wikipedia contributors. (2022a). Confusion matrix — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Confusion_matrix&oldid=1107701525

[6] Wikipedia contributors. (2022b). F-score — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=F-score&oldid=1120355370

[7] Zaki, M. J., & Wagner, M. Jr. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.

Appendix

```
rm(list = ls())
load(file.choose())
ls()
df[1:3, ]
names(df)[1] = "Class"
set.seed(1)
sample = sample(c(TRUE, FALSE), nrow(df), replace=TRUE, prob=c(0.7,0.3))
train = df[sample, ]
test = df[!sample, ]
library(e1071)
Svm.Mod = svm(Class ~ ., data = train, type = "C-classification",
scale = FALSE, kernel = "radial", cost = 5)
Pred = predict(Svm.Mod, test)
CM = table(Prediction = Pred, Actual = test$Class); CM
accuracy = sum(CM[1], CM[4])/sum(CM[1:4]); accuracy
precision = CM[1]/sum(CM[1], CM[3]); precision
sensitivity = CM[1]/sum(CM[1], CM[2]); sensitivity
specificity = CM[4]/sum(CM[4], CM[3]); specificity
f1score = (2 * (sensitivity * precision))/(sensitivity + precision); f1score
## Calculate Area Under the Curve
## https://www.projectpro.io/recipes/calculate-area-under-curve-r
library(pROC)
PredN = as.numeric(levels(Pred))[Pred]
AUC_Mod = roc(test$Class, PredN)
auc(AUC_Mod)
dim(df)
```