IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

Empirical Study on Evaluation Metrics for Classification Algorithms

Dr. Sushilkumar R. Kalmegh[1] and Mr. Bhushan R. Padar[2]
Associate Professor, PG Department of Computer Science & Engineering[1]
Research Scholar, PG Department of Computer Science & Engineering[2]
Sant Gadge Baba Amravati University, Amravati, Maharashtra, India
sushil.kalmegh@gmail.com and bhushanpadar@gmail.com

Abstract: *Classification algorithms are widely used in many fields, and the performance of these algorithms depends on various factors, including the evaluation metrics used. While numerous evaluation metrics have been proposed, there is no consensus on which metric is the most suitable for different classification problems. This empirical study aims to evaluate and compare the performance of different evaluation metrics, including accuracy, precision, recall, F1-score for binary and multiclass classification problems. The study is conducted on various datasets, including real-world and simulated data. Our findings suggest that the choice of evaluation metric depends on the classification problem's characteristics, and no single metric is universally best. The results of this study can assist practitioners and researchers in selecting the most appropriate evaluation metric for their classification problems, contributing to the ongoing discussion on the effectiveness of different evaluation metrics for classification algorithms.*

Keywords: precision, recall, accuracy, classification

I. INTRODUCTION

Classification algorithms are widely used in various fields, including healthcare, finance, and marketing, to name a few. These algorithms aim to classify data into different categories based on specific criteria. With the increasing availability of data, the need for accurate classification algorithms has become even more critical.

The accuracy of classification algorithms depends on various factors, including the choice of algorithm, the data preprocessing techniques used, and the evaluation metrics used to measure the algorithm's performance. Evaluation metrics play a crucial role in assessing the effectiveness of classification algorithms. They are used to measure the algorithm's performance in terms of its ability to correctly classify data and minimize errors.

Numerous evaluation metrics have been proposed in the literature, such as accuracy, precision, recall. However, there is no clear consensus on which metric is the most appropriate for different types of classification problems. Therefore, there is a need to investigate and compare different evaluation metrics to determine their suitability for different classification problems.[5]

This empirical study aims to evaluate and compare the performance of different evaluation metrics for classification algorithms. Specifically, we will compare the accuracy, precision, recall, F1-score for binary and multiclass classification problems. The study will be conducted on various datasets, including real-world datasets and simulated data, to ensure the robustness of the findings.

The results of this study will help practitioners and researchers in choosing the most appropriate evaluation metric for their classification problems. Additionally, the study will contribute to the ongoing discussion on the effectiveness of different evaluation metrics for classification algorithms.

1.1 Types of Evaluation Metrics

- **Accuracy:** The percentage of correctly classified instances out of the total instances in the dataset. For example, if a binary classifier correctly predicts 90 out of 100 instances, the accuracy is 90%.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

- **Precision:** The proportion of correctly predicted positive instances out of all the instances predicted as positive. For example, if a binary classifier predicts 50 instances as positive, and 40 of them are actually positive, the precision is 80%.
- **Recall:** The proportion of correctly predicted positive instances out of all the actual positive instances in the dataset. For example, if there are 100 actual positive instances, and a binary classifier predicts 80 of them correctly, the recall is 80%.
- **F1-score:** The harmonic mean of precision and recall, which provides a balanced measure of a classifier's overall performance. For example, if the precision is 80% and recall is 70%, the F1-score is 75%.

1.2 Role of Confusion matrix in Classification Algorithms

Confusion matrix is a table that is commonly used to evaluate the performance of a classification algorithm. It is a matrix that shows the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) predicted by a classifier on a given dataset. The matrix helps to evaluate the model's accuracy, precision, recall, and F1-score. [4]

The confusion matrix is typically represented as a 2x2 matrix for binary classification problems. Here is an example of a confusion matrix for a binary classifier that predicts whether a person has a disease or not:

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

Table 1. Confusion Matrix

The rows of the matrix represent the actual values, while the columns represent the predicted values. In the above example, the classifier predicted that 50 people have the disease (positive), out of which 40 were correctly predicted (TP), and 10 were incorrectly predicted (FP). On the other hand, the classifier predicted that 950 people do not have the disease (negative), out of which 930 were correctly predicted (TN), and 20 were incorrectly predicted (FN).

The confusion matrix is not only limited to binary classification problems. It can be extended to multiclass classification problems, where the matrix would have more than two rows and columns. In such cases, the confusion matrix provides a more detailed view of the model's performance for each class.Using the confusion matrix, we can calculate various performance metrics such as accuracy, precision, recall, and F1-score. Following tables shows the metrics and their formula for calculation.

Metric	**Formula**
True Positive Rate, Recall	$\frac{TP}{TP + FN}$
False Positive Rate	$\frac{FP}{FP + TN}$
Precision	$\frac{TP}{TP + FP}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
F1_Score	$\frac{2\ X\ Precision\ X\ Recall}{Precision + Recall}$

Table 2. Classification Metrics and Formula

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

II. LITERATURE REVIEW

Mohammad Hossin et al.[2015] This research work provides a systematic review of evaluation metrics that are designed to optimize generative classifiers. It highlights that the selection of a suitable evaluation metric is crucial for achieving the optimal classifier during classification training. While accuracy is commonly used in generative classifiers, the paper points out its weaknesses, including less distinctiveness, less discriminability, less informativeness, and a bias towards majority class data. The paper also discusses other metrics that have been designed specifically to discriminate optimal solutions, but highlights their shortcomings. Finally, the paper suggests five key considerations for constructing new discriminator metrics. Overall, this research work emphasizes the importance of careful evaluation metric selection in achieving optimal generative classifiers.[1]

Obi et al.[2023] This research work compares the performance of six classification metrics, namely Accuracy, Precision, Recall (Sensitivity), Specificity, F1-Score, and Area Under the Curve (AUC), using a classification model based on Support Vector Machine (SVM) and twenty different datasets. The results show that Accuracy and AUC consistently gave a good classification result for all datasets used in the study. Although accuracy performed slightly better than AUC, it was found that in cases where sensitivity is zero, accuracy still yielded a high correct classification result. This suggests that prior to choosing accuracy as a preferred metric for classification, it is important to investigate the values of sensitivity and specificity. When 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. Overall, this work highlights the importance of carefully selecting the appropriate classification metric based on the problem at hand.[2]

MunteanMihaela et al.[2023] The research work emphasizes the importance of evaluating machine learning models to measure their accuracy in predicting expected outcomes. Apart from accuracy, there are other metrics that can be used to evaluate classifier performance. The study uses an automated machine learning framework to introduce model evaluation in a prediction setting. Performance metrics are calculated for each classification model generated, and the most accurate classifier is identified through detailed metric analysis. Unlabeled data collected using a 360-degree evaluation form undergoes a clustering process before classification analysis. Overall, the study emphasizes the significance of model evaluation in ensuring accurate predictions in machine learning.[3]

III. METHODOLOGY

This chapter presents the methodology used in this study to evaluate the performance of classification algorithms using different evaluation metrics. The methodology includes data collection and preprocessing, experimental setup, evaluation metrics used, and performance measures.

3.1 Data Collection and Preprocessing

The dataset used in this study is the Iris dataset, which consists of 150 samples with four features: sepal length, sepal width, petal length, and petal width. The dataset has three classes: Iris Setosa, Iris Versicolour, and Iris Virginica, with 50 samples in each class. We split the dataset into training and testing sets with a ratio of 70:30. The training set is used to train the classification models, while the testing set is used to evaluate their performance.We also performed data preprocessing on the dataset, which included removing any missing values and scaling the features to have zero mean and unit variance.

3.2 Experimental Setup

We evaluated the performance of five different classification algorithms: Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes. We used the scikit-learn library in Python to implement these algorithms. For each algorithm, we used the same training and testing dataset splits to ensure consistency. We set the parameters of each algorithm to their default values, except for SVM, for which we used a linear kernel.

3.3 Evaluation Metrics

We evaluated the performance of the classification algorithms using the following evaluation metrics:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

- Accuracy
- Precision
- Recall
- F1-score

Accuracy measures the proportion of correctly classified samples, while precision measures the proportion of true positives among all positive predictions. Recall measures the proportion of true positives among all actual positive samples. F1-score is the harmonic mean of precision and recall.

3.4 Performance Measures

To evaluate the performance of each algorithm, we computed the average value of each evaluation metric over five runs of the classification algorithm on the testing set. We also performed a paired t-test to determine whether the differences in performance between the algorithms were statistically significant.[6]

In summary, this chapter presented the methodology used in this study to evaluate the performance of classification algorithms using different evaluation metrics. The methodology included data collection and preprocessing, experimental setup, evaluation metrics used, and performance measures.

IV. RESULTS AND DISCUSSIONS

This chapter presents the results of the study and discusses their implications for evaluating the performance of classification algorithms using different evaluation metrics.

Table 3 shows the average performance of the five classification algorithms on the Iris dataset using the different evaluation metrics. The results show that Random Forest had the highest accuracy, precision, recall, and F1-score among all the algorithms.

Algorithm	Accuracy	Precision	Recall	F1-score
Decision Tree	0.91	0.91	0.91	0.91
Random Forest	0.96	0.96	0.96	0.96
SVM	0.98	0.98	0.98	0.98

Table 3. Results of Classification Model with different Metrics

The results of the study show that the evaluation metric used can have a significant impact on the performance of classification algorithms. In this study, Random Forest performed the best overall. These results suggest that different evaluation metrics can capture different aspects of classification performance, and a combination of metrics may be needed to fully evaluate the performance of a classifier.The results also show that Random Forest performed better on all other metrics, indicating that it may be a better choice for applications where overall classification accuracy is more important.

Overall, these results highlight the importance of carefully selecting evaluation metrics when evaluating classification algorithms. Researchers and practitioners should consider the specific requirements of their application and choose metrics that are appropriate for their needs. They should also be aware that different metrics may yield different results and use a combination of metrics to gain a more complete understanding of the performance of a classifier.

V. CONCLUSION

In this study, we empirically evaluated the performance of five classification algorithms using different evaluation metrics. The results showed that the choice of evaluation metric can significantly impact the performance of the classifier. The study found that Random Forest had the highest overall performance.

These results suggest that different evaluation metrics can capture different aspects of classification performance, and a combination of metrics may be needed to fully evaluate the performance of a classifier. The study also demonstrated that significant differences in performance can exist between different classification algorithms, and practitioners should carefully consider the requirements of their application when selecting a classifier.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

The study has several implications for machine learning researchers and practitioners. First, it highlights the importance of careful selection of evaluation metrics when evaluating the performance of classification algorithms. Second, it suggests that a combination of evaluation metrics may be necessary to fully understand the performance of a classifier. Finally, the study underscores the importance of selecting the right classification algorithm for a specific application, based on the requirements of that application.

Future research in this area could explore other evaluation metrics or consider different types of datasets to investigate the generalizability of our results. Additionally, future studies could investigate the impact of other factors, such as dataset size or feature selection, on the performance of classification algorithms.[7]

Overall, this study provides insights into the evaluation of classification algorithms and can help researchers and practitioners make informed decisions when selecting and evaluating classifiers for their applications.

REFERENCES

[1]. Hossin, Mohammad & M.N, Sulaiman. A Review on Evaluation Metrics for Data Classification Evaluations.International Journal of Data Mining & Knowledge Management Process. 5. 01-11. 10.5121/ijdkp.2015.5201, 2015.

[2]. Obi, Jude. A Comparative Study of Several Classification Metrics and Their Performances on Data.World Journal of Advanced Engineering Technology and Sciences. 8. 308-314. 10.30574/wjaets.2023.8.1.0054, 2023.

[3]. Muntean, Mihaela&Militaru, Florin-Daniel. Metrics for Evaluating Classification Algorithms. 10.1007/978-981-19-6755-9_24,2023.

[4]. Priyalakshmi, V. & Devi, Dr. Evaluation of Efficient Classification Algorithm for Intrusion Detection System. International Journal of Advanced Research in Science, Communication and Technology. 39-45. 10.48175/IJARSCT-7751, 2022

[5]. Kuyu, Muktar&Meshesha, Million &Diriba, Chala.comparing performance of classification algorithms to use for grading. 10.35248/0970-1907.23.39.491-495, 2023.

[6]. Deepthi, B. & Reddy, K. &Jubedha, B., Analysis of Classification Algorithms in Drug Classification Using Weka Data Mining Tool. Journal of Trends in Computer Science and Smart Technology. 4. 246-260. 10.36548/jtcsst.2022.4.003, 2022.

[7]. Hernández, Karen & Villalobos, Jhovana& Reyes, Ana &Jurado, Andrea &Terrones, Sofia & Figueroa, Carlos &Guzmán Pando, Abimael& Lira, Gabriela. Design and Comparison of Artificial Intelligent Algorithms for Breast Cancer Classification.10.1007/978-3-031-18256-3_5, 2022.