Lab 2

Task A: Classification

Alejandro Silva Rodríguez

Marta Cuevas Rodríguez

Aprendizaje Computacional
Universidad de Málaga

Septiembre 2024

Contents

1 Introduction

In computational learning, applying classification algorithms to predict disease progression is critical for deriving meaningful insights from complex biomedical data. The primary focus of this project is to evaluate and compare the performance of various classification methods, utilizing established metrics like precision, recall, specificity, and accuracy. These metrics provide a quantitative basis to assess each model's strengths and limitations, enabling us to identify the most suitable algorithm for predicting disease outcomes.

This project will analyze a dataset relevant to disease classification, covering essential aspects such as the dataset's class distribution, balance, and overall characteristics. By exploring different methods through detailed metric calculations, we aim to determine the best-performing model for predicting outcomes in the dataset. Additionally, graphical representations will aid in visualizing and comparing each method's effectiveness, making it easier to assess their respective advantages. Overall, this project provides a systematic approach to classification model evaluation and supports informed decision-making in selecting models suited to biomedical data analysis.

2 Objectives

- **Implement an algorithm for performance metric calculations**: Develop an algorithm that calculates a range of classification performance metrics, including precision, recall, specificity, accuracy, and others, providing a comprehensive comparison of each method's effectiveness.
- **Analyze the dataset characteristics**: Examine key properties of the dataset, such as the number of samples, class distribution, and class balance. Understanding these characteristics will help us assess the dataset's impact on classification performance and ensure a fair comparison of methods.
- **Provide detailed definitions for each metric**: Define each performance metric in detail, including its mathematical formula, what it measures, its range, and an interpretation of whether higher or lower values are preferable for classification tasks.
- **Visualize results for clearer comparisons**: Generate visualizations, such as plots comparing False Positives and False Negatives, Precision and Recall, and Accuracy versus F-measure. These graphs will facilitate a better understanding of each method's strengths and weaknesses.
- **Identify the best classification method based on performance**: Conduct a thorough analysis of each metric to identify the method with the best overall performance, providing a rationale for why this model is most suitable for predicting disease outcomes in this dataset.

3 Methodology and Results

3.1 Dataset Description

The classification performance of each method, as detailed in Table 1, provides valuable insights into the dataset's characteristics. Specifically, we focus on four key metrics: True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN).

Given that there are only two classifications—positive and negative—we infer that the dataset contains two distinct classes.

Method A classifies all tuples as positive, resulting in 100 True Positives. This indicates that there are likely 100 instances belonging to the positive class. Additionally, the presence of 900 False Positives suggests that the remaining tuples are classified as negative, supporting the assumption that there are a total of 900 instances in the negative class. This conclusion is further corroborated by the performance of Method E,

which only classifies instances as negative, correctly identifying 900 True Negatives.

Overall, the dataset is significantly unbalanced, with a considerable disparity between the number of positive and negative instances. This imbalance poses challenges for classification, making it more difficult for the algorithms to accurately predict the positive class due to the overwhelming presence of negative instances.

Table 1: Performance Metrics for Each Classification Method

Method	**TP**	**FP**	**FN**	**TN**
A	100	900	0	0
B	80	125	20	775
C	25	25	75	875
D	50	50	50	850
E	0	0	100	900

3.2 Metrics Used for Performance Comparison

In this section, we will explore the various metrics employed to evaluate and compare the performance of different classification methods. Each metric provides insights into the model's strengths and weaknesses, helping us understand how well it performs in various aspects of classification.

- **Precision (PR)**: Precision is a key metric that measures the accuracy of the positive predictions made by the model. It represents the proportion of True Positives (TP) among all predicted positives, combining both correct and incorrect predictions.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Range: [0, 1].
Better values: Higher values indicate that a greater proportion of positive predictions are correct, which is desirable in many applications.

- **Recall (RC)**: Also known as Sensitivity, Recall quantifies the model's ability to identify all relevant instances within the dataset. It is defined as the ratio of True Positives to the total actual positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Range: [0, 1].
Better values: Higher values are preferred, indicating that the model is effective at capturing as many actual positives as possible.

- **Specificity (SP)**: Specificity assesses how well the model identifies negative instances. It is calculated as the ratio of True Negatives (TN) to the total number of actual negatives.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Range: [0, 1].
Better values: Higher values suggest that the model is proficient at correctly identifying negatives, which is crucial for balanced performance.

- **False Negative Rate (FNR)**: The False Negative Rate indicates the proportion of actual positives that are incorrectly classified as negatives. It highlights the model's shortcomings in capturing positive instances.

$$\text{FNR} = \frac{FN}{TP + FN}$$

Range: [0, 1].
Better values: Lower values are preferable, as they suggest that fewer true positives are being missed by the model.

- **False Positive Rate (FPR)**: This metric quantifies the proportion of actual negatives that are mistakenly classified as positives. Understanding the FPR is important for assessing the model's performance in contexts where false alarms are costly.

$$\text{FPR} = \frac{FP}{FP + TN}$$

Range: [0, 1].
Better values: Lower values are preferred, indicating that the model generates fewer false positives.

- **Accuracy (ACC)**: Accuracy provides an overall assessment of the model's performance, reflecting the ratio of correctly predicted instances (both TP and TN) to the total instances in the dataset.

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

Range: [0, 1].
Better values: Higher values indicate a more effective model, although accuracy alone may not provide a complete picture in imbalanced datasets.

- **Spatial Accuracy (S) or Jaccard Index (J)**: This metric evaluates the similarity between the predicted positive instances and the actual positives. It is particularly useful for assessing how well the model performs in terms of agreement with the ground truth.

$$\text{Jaccard Index} = \frac{TP}{TP + FN + FP}$$

Range: [0, 1].
Better values: Higher values signify better alignment between predicted and actual positives.

- **F-measure (Fm)**: The F-measure provides a balanced score that incorporates both Precision and Recall. It is defined as the harmonic mean of these two metrics, making it a valuable indicator of model performance.

$$\text{F-measure} = \frac{2 \cdot PR \cdot RC}{PR + RC}$$

Range: [0, 1].
Better values: Higher values suggest that the model performs well in both precision and recall, making it effective for a wide range of applications.

To provide a concise overview of the metrics described earlier, Table 2 summarizes each metric, its definition, range, and whether higher or lower values are preferable. This table serves as a quick reference to understand how these metrics contribute to performance evaluation in classification tasks.

Table 2: Summary of Metrics Used for Performance Evaluation

Metric	Definition	Formula	Range	Better Values
Precision (PR)	Accuracy of positive predictions	$\frac{TP}{TP+FP}$	[0, 1]	Higher
Recall (RC)	Ability to identify relevant instances	$\frac{TP}{TP+FN}$	[0, 1]	Higher
Specificity (SP)	Identifying negative instances correctly	$\frac{TN}{TN+FP}$	[0, 1]	Higher
False Negative Rate (FNR)	Proportion of actual positives missed	$\frac{FN}{TP+FN}$	[0, 1]	Lower
False Positive Rate (FPR)	Proportion of actual negatives misclassified	$\frac{FP}{FP+TN}$	[0, 1]	Lower
Accuracy (ACC)	Overall correctness of predictions	$\frac{TN+TP}{TP+FN+FP+TN}$	[0, 1]	Higher
Jaccard Index (J)	Similarity between predicted and actual positives	$\frac{TP}{TP+FN+FP}$	[0, 1]	Higher
F-measure (Fm)	Balance between Precision and Recall	$\frac{2 \cdot PR \cdot RC}{PR+RC}$	[0, 1]	Higher

3.3 el punto tres ese

explicar esto

3.4 Performance Analysis

For each metric, the following analysis highlights the best and worst methods, with a brief explanation for each:

- **Precision (PR):**
  The best methods in terms of precision are Methods C and D (0.5000), showing that 50% of their positive predictions are correct. This is beneficial for scenarios where accurate positive identification is key. Method E, however, scores the lowest (0.0), as it makes no positive predictions.

- **Recall (RC):**
  Method A achieves the highest recall (1.0000), identifying all actual positives correctly, which is ideal in cases where detecting all positives is essential. Method E performs the worst (0.0), as it fails to classify any positives.

- **Specificity (SP):**
  Method E has the highest specificity (1.0000), correctly classifying all negatives without any false positives. This is useful for minimizing false alarms. Method A, by contrast, has the lowest specificity (0.0), as it incorrectly classifies all negatives as positives.

- **False Negative Rate (FNR):**
  Method A has the lowest FNR (0.0), indicating it captures all positives without missing any. In

Metric	A	B	C	D	E
TP	**100.0000**	80.0	25.0	50.0	0.0
FP	**900.0000**	125.0	25.0	50.0	0.0
FN	0.0	20.0	75.0	50.0	**100.0000**
TN	0.0	775.0	875.0	850.0	**900.0000**
Precision (PR)	0.1	0.3902	**0.5000**	0.5000	0.0
Recall (RC)	**1.0000**	0.8	0.25	0.5	0.0
Specificity (SP)	0.0	0.8611	0.9722	0.9444	**1.0000**
False Negative Rate (FNR)	**0.0000**	0.2	0.75	0.5	1.0
False Positive Rate (FPR)	1.0	0.1389	0.0278	0.0556	**0.0000**
Accuracy (ACC)	0.1	0.855	**0.9000**	0.9	0.9
Jaccard Index (J)	0.1	**0.3556**	0.2	0.3333	0.0
F-measure (Fm)	0.1818	**0.5246**	0.3333	0.5	0.0

Table 3: Performance metrics for each method with the best values highlighted in bold

contrast, Method E has the highest FNR (1.0), as it misses all positive cases, leading to poor detection of positives.

- **False Positive Rate (FPR):**
  The lowest FPR is achieved by Method E (0.0), as it avoids misclassifying any negatives as positives. Method A has the highest FPR (1.0) because it classifies all negatives as positives, resulting in many false positives.

- **Accuracy (ACC):**
  Methods C, D, and E exhibit the highest accuracy (0.9000), showing an effective balance between correctly identifying positives and negatives. Method A, with an accuracy of 0.1, performs the worst, as it misclassifies almost all negative instances.

- **Jaccard Index (J):**
  Method B achieves the highest Jaccard Index (0.3556), indicating a favorable overlap between predicted and actual positives. Method E, however, scores 0.0 on this metric due to its lack of positive predictions.

- **F-measure (Fm):**
  The highest F-measure is obtained by Method B (0.5246), showing a solid balance between precision and recall. Method E, again, scores the lowest (0.0), as it fails to identify any positives.

In summary, Method B demonstrates the most balanced performance across metrics, effectively combining precision and recall. In contrast, Method E consistently ranks the lowest across metrics that measure positive classification, only excelling in identifying negatives.

4 todoList

~~1. Information about the dataset. Number of samples, number of classes, number of samples per class, if the dataset is balanced or unbalanced...~~

~~2. Information about the metrics you will use to compare the performance of the methods. Name, what it represents, its definition, range of the value provided by that metric, if lower/higher values are better. . .~~

3. A table with the yielded performance of each method for each metric. You can distribute the methods along the columns and the metrics along the rows. Highlight best results in bold.

4. Analysis of the performance. For each metric, describe briefly which method is the best and the worst, and why.

5. Conclusion. According to the analysis, determine which method, in general, is the best and why

following figures: - FN against FP - PR against RC - ACC against Fm

5 Conclusion

6 Repository Access

All additional information, including the source code and full documentation of this project, is available in the GitHub repository [1].

References

[1] Marta Cuevas. Lab2_computational_learning. `https://github.com/martacuevasr/Lab2_Computational_learning`, 2024. Último acceso: 1 noviembre 2024.