



INSTITUTO POLITÉCNICO NACIONAL

**CENTRO DE INVESTIGACIÓN
EN COMPUTACIÓN**

Tarea 07

Matriz de confusión: medidas de desempeño

Edgar Fernando Espinosa Torres

Clasificación inteligente de patrones

Dr. Cornelio Yáñez Márquez



Ciudad de México, 22 de octubre de 2023.

Índice general

Introducción	1
Desarrollo y discusión.....	2
Parte 1.....	2
Parte 2.....	3
Parte 3.....	4
Parte 4.....	5
Conclusiones	7
Referencias bibliográficas	8

Índice de figuras

Ilustración 1: Matriz de confusión para Haberman's Survival 1965 con partición fija.....	2
Ilustración 2: Matriz de confusión para la parte 2.	4

Introducción

Hasta el Recurso Didáctico 06 sólo se habían estudiado las medidas de desempeño *Accuracy* y *Error Rate*, las cuales son válidas para datasets balanceados. Es natural preguntarse qué sucede cuando nos enfrentamos a un dataset desbalanceado. Por lo cual, es necesario estudiar cuáles medidas de desempeño pueden utilizarse en esta situación.

En la parte 1 de la tarea, se calculan medidas de desempeño originadas a partir de los elementos básicos de la matriz de confusión. Por ejemplo: *accuracy* para datasets balanceados, *accuracy por clase* y *balanced accuracy* para datasets balanceados o desbalanceados para el dataset Haberman's Survival 1965 con partición fija. En las partes 2, 3 y 4 se reportan los resultados de una búsqueda de artículos JCR que utilicen estas medidas.

Desarrollo y discusión

Propósito de la tarea

Parte 1

Utilizamos las definiciones del recurso didáctico 07 [1].

1.1 Elegir algún ejemplo desarrollado previamente en el curso CIP

Se tomó el dataset Haberman's Survival [2] 1965 con partición fija.

1.2 Generar la matriz de confusión

3	1
0	4

Ilustración 1: Matriz de confusión para Haberman's Survival 1965 con partición fija.

1.3 Calcular *accuracy* (si es adecuado)

Es adecuado calcular *Accuracy* porque $IR = \frac{14}{13} = 1.08 \leq 1.5$

$$\begin{aligned} Accuracy &= \frac{\text{total de aciertos del clasificador}}{\text{total de patrones del conjunto } P} = \frac{TP + TN}{TP + FN + FP + TN} = \frac{3 + 4}{3 + 1 + 4} \\ &= \frac{7}{8} = 0.875 \end{aligned}$$

1.4 Calcular *accuracy* para la clase *Positive*

$$Accuracy \text{ para la clase } Positive = \frac{TP}{TP + FN} = \frac{3}{3 + 1} = \frac{3}{4} = 0.75$$

1.4 Calcular *accuracy* para la clase *Negative*

$$Accuracy \text{ para la clase } Negative = \frac{TN}{FP + TN} = \frac{4}{0 + 4} = 1$$

1.5 Calcular *Balanced accuracy*

Balanced Accuracy

$$= \frac{\text{accuracy para la clase Positive} + \text{accuracy para la clase Negative}}{2}$$

$$= \frac{0.75 + 1}{2} = 0.875$$

Parte 2

2.1 Buscar un artículo JCR donde se use *sensitivity*, *recall* o *TP Rate* en un problema biclase

Los autores de este artículo trabajaron con imágenes de rayos X y de tomografías computarizadas (CT) de pulmones. Mediante una estrategia de Deep learning basada en Convolutional Neural Network (CNN) se detectó e identificó la enfermedad de Covid-19 [3].

El aturo reporta *recall* de aproximadamente 99.84 % y refiere que esta medida de desempeño también es conocida como *sensitivity*.

2.2 Investigar si el autor reportó cuál es la clase *Positive*

Los autores mencionan que los pacientes positivos son los que tenían Covid-19 y los negativos aquellos que no tenían la enfermedad.

2.3 Investigar si el autor reportó la matriz de confusión

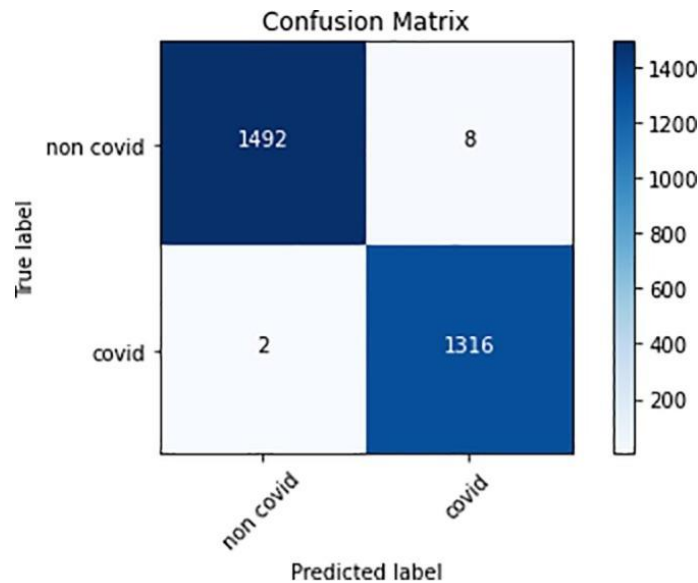


Ilustración 2: Matriz de confusión para la parte 2.

Es importante notar que los autores reportaron la matriz de confusión utilizando los cuadrantes para TP , FN , TN y FP en un orden distinto al que estamos utilizando en clase.

El modelo fue capaz de identificar con precisión a los pacientes con Covid-19, lo que se traduce en tasas elevadas de TP y TN .

2.4 Investigar si el autor reportó las razones que lo llevaron a utilizar esta medida de desempeño en su artículo

Los autores no reportaron cuáles fueron las razones que los llevaron a utilizar esta medida de desempeño.

Parte 3

3.1 Buscar un artículo JCR donde se use *specificity* o *TN Rate* en un problema biclase

En este artículo se presenta una comparación del rendimiento de siete métodos de machine learning (ML) para clasificar lesiones mamarias mediante ecografía. A partir de un conjunto de datos con 2032 casos (1341 benignos y 691 malignos), se extrajeron 137 características

morfológicas y de textura, con el objetivo de describir el léxico BI-RADS para masas [4]. Se evaluó *specificity* (*SPE*) y otras medidas de desempeño para Support Vector Machine (SVM), K-Nearest Neighbor (kNN), Radial Basis Function Network (RBFN), Linear Discriminant Analysis (LDA), Multinomial Logistic Regression (MLR), Random Forest (RF) y AdaBoost (Ada). Los resultados revelaron que LDA obtuvo el mejor rendimiento general de clasificación con $SPE = 0.93$.

3.2 Investigar si el autor reportó cuál es la clase *Positive*

Los autores reportaron a la clase *Positive* como aquella consistente de tumores malignos y a la clase *Negative* como aquella consistente de tumores benignos.

3.3 Investigar si el autor reportó la matriz de confusión

Los autores no reportaron la matriz de confusión.

3.4 Investigar si el autor reportó las razones que lo llevaron a utilizar esta medida de desempeño en su artículo

Los autores no reportaron cuáles fueron las razones que los llevaron a utilizar esta medida de desempeño.

Parte 4

4.1 Buscar un artículo JCR donde se use *Balanced accuracy*

Las herramientas *in silico* se refieren a los métodos y técnicas computacionales utilizados para simular, analizar y predecir comportamientos en biología y medicina. Este artículo altamente especializado se centra en la evaluación de 44 herramientas *in silico* frente a múltiples ensayos funcionales a gran escala de genes asociados con la susceptibilidad al cáncer. Las herramientas *in silico* son utilizadas para predecir la funcionalidad de las variantes genéticas [5].

En el estudio, las variantes missense (cambio de un solo nucleótido que resulta en un codón que codifica un aminoácido diferente) fueron clasificadas como no funcionales (deleterias,

DEL) o funcionales (toleradas, TOL) basándose en las especificaciones de los ensayos funcionales.

4.2 Investigar si el autor reportó cuál es la clase *Positive*

El autor menciona que las variantes de los genes no funcionales son consideradas la clase *Positive* mientras que las funcionales son consideradas la clase *Negative*.

4.3 Investigar si el autor reportó la matriz de confusión

Los autores no reportaron la matriz de confusión.

4.4 Investigar si el autor reportó las razones que lo llevaron a utilizar esta medida de desempeño en su artículo

Los autores mencionaron que utilizaron *Balanced Accuracy* para determinar el valor de corte óptimo en cada herramienta *in silico*, con el objetivo de clasificar las variantes genéticas como funcionales o no funcionales de manera precisa.

Balanced Accuracy es preferida en situaciones de clases desbalanceadas, ya que proporciona una evaluación más justa del rendimiento del clasificador para ambas clases. Esta métrica es crucial en el ámbito de la genómica médica para asegurar una clasificación precisa de las variantes genéticas, lo cual es fundamental para entender la susceptibilidad al cáncer y podría contribuir a mejorar el diagnóstico y tratamiento del cáncer.

Conclusiones

Aunque dos de tres artículos JCR encontrados no reportan las razones de porqué eligieron utilizar esas medidas de desempeño es importante mencionar lo siguiente:

La medida de desempeño *Accuracy* no es ideal para evaluar clasificadores en casos de clases desbalanceadas ya que puede dar una impresión engañosa de buen rendimiento, especialmente si la clase mayoritaria es predicha correctamente mientras que la clase minoritaria, que podría ser crucial como en el caso de la detección de enfermedades, es ignorada o predicha incorrectamente.

Mientras que el *accuracy por clase* y el *balanced accuracy* ofrecen una evaluación más justa y representativa del desempeño de clasificador en datasets desbalanceados, asegurando que todas las clases, independientemente de su tamaño, sean consideradas en la evaluación.

Referencias bibliográficas

- [1] Yáñez Márquez, C. (2023). RD 07: Matriz de confusión: medidas de desempeño [Diapositivas de clase]. Centro de Investigación en Computación.
- [2] UCI Machine Learning Repository. (s. f.-c).
<https://archive.ics.uci.edu/dataset/43/haberman+s+survival>
- [3] Thakur, S., & Kumar, A. (2021). X-ray and CT-scan-based automated detection and classification of covid-19 using convolutional neural networks (CNN). *Biomedical signal processing and control*, 69, 102920. <https://doi.org/10.1016/j.bspc.2021.102920>
- [4] González-Luna, F. A., Hernández-López, J., & Gomez-Flores, W. (2019). A Performance Evaluation of Machine Learning Techniques for Breast Ultrasound Classification. En 2019 *16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)* (pp. 1-5). Mexico City, Mexico: IEEE.
<https://doi.org/10.1109/ICEEE.2019.8884547>
- [5] Cubuk, C., Garrett, A., Choi, S. et al. (2021). Clinical likelihood ratios and balanced accuracy for 44 in silico tools against multiple large-scale functional assays of cancer susceptibility genes. *Genet Med*, 23(10), 2096–2104. <https://doi.org/10.1038/s41436-021-01265-z>