INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

Tarea 12:

Métricas de Minkowski, complejidad de datos, conversión e imputación

Edgar Fernando Espinosa Torres

Clasificación inteligente de patrones

Dr. Cornelio Yáñez Márquez

Ciudad de México, 9 de diciembre de 2023.

Índice general

Índice de tablas

Introducción

Recordemos que uno de los métodos de validación recientemente estudiados es Leave-one-out, que en esta tarea se emplea junto con el algoritmo 3-NN de manera exhaustiva. Este método de validación consiste en utilizar cada patrón del dataset como un conjunto de prueba individual, mientras que el resto de los datos constituyen el conjunto de entrenamiento. Este proceso se repite tantas veces como patrones haya, garantizando que cada uno de ellos forme parte del conjunto de prueba en algún momento.

Hasta el momento solamente habíamos trabajado con la métrica euclidiana. Pero para esta tarea, se utilizó la métrica de $Bray - Curtis$ en el clasificador [1]:

$$Bray - Curtis = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} (x_i + y_i)}$$

Esta métrica fue desarrollada originalmente en el campo de la ecología, específicamente para comparar la composición de comunidades ecológicas. Los ecólogos J. Roger Bray y J. T. Curtis la introdujeron en 1957 con el propósito de cuantificar la similitud entre dos comunidades biológicas en términos de su composición de especies y abundancia relativa.

El principal objetivo de esta métrica era proporcionar un método robusto y significativo para comparar diferentes sitios ecológicos, basándose en los tipos y cantidades de especies presentes en cada uno. Esto era especialmente importante para estudiar cómo las comunidades ecológicas varían espacial y temporalmente, y para entender los efectos de factores ambientales y humanos en la biodiversidad.

También, se calculó la matriz de confusión y a partir de ella, se calculó $Recall$ y las medidas de desempeño presentadas por primera vez en una tarea: $Precision$, $F1$ y MCC.

Desarrollo y discusión

Propósitos de la tarea

Ejemplificar el uso de métricas diferentes a la Euclidiana en la aplicación de clasificadores de la familia $k-NN$; ejemplificar el cálculo de las medidas de desempeño: $Precision$, $F1$ y MCC; y comparar los resultados.

Parte 1

- **Con el método de validación Leave-one-out y con la métrica que se indica de manera individual, aplicar el algoritmo 3-NN en el dataset apendicitis (https://sci2s.ugr.es/keel/dataset.php?cod=183).**
- **Reportar en una tabla los valores de $Recall$, $Precision$, $F1$ y MCC.**

El dataset Appendicitis con 106 patrones, es un conjunto de datos relativamente pequeño, lo que es común en ciertos campos médicos donde la recolección de datos puede ser un reto. Todos los atributos son numéricos (reales) y varían entre 0.0 y 1.0 [2].

Contiene siete atributos numéricos ($At1$ a $At7$). Estas características son probablemente resultados de pruebas médicas o evaluaciones clínicas relacionadas con la apendicitis.

El objetivo es clasificar los patrones en dos clases, lo que indica que el conjunto de datos se utiliza para un problema de clasificación binaria.

A continuación, se presenta la matriz de confusión correspondiente a los resultados de la clasificación:

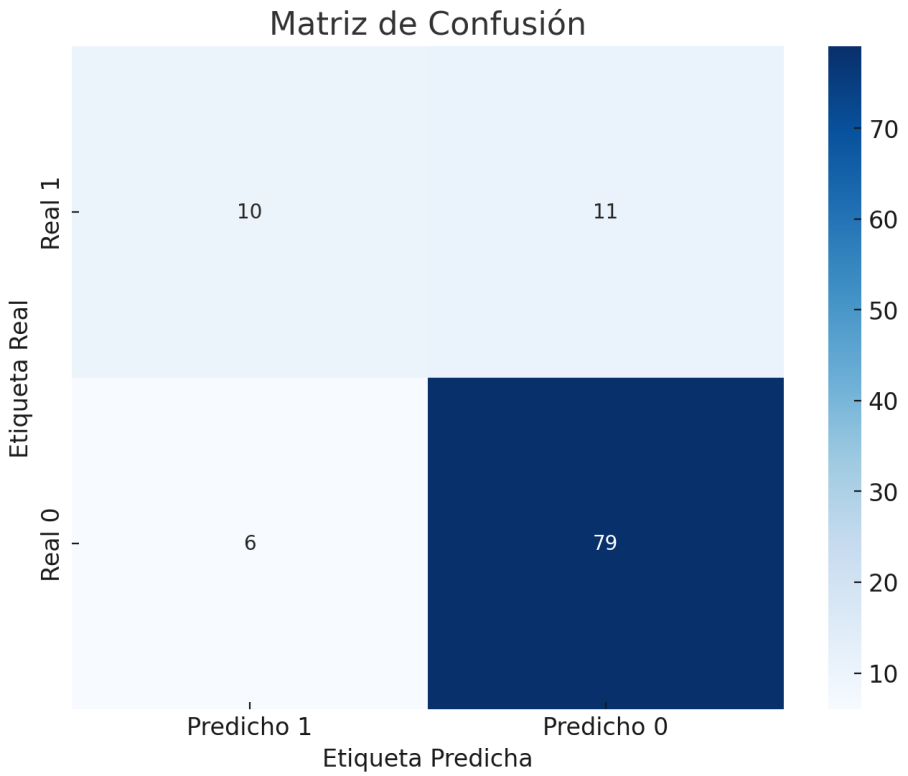

Tabla 1: Matriz de confusión para el dataset Appendicitis

El $imbalance\ ratio$ es $IR = \frac{85}{21} = 4.04 \geq 1.5$, es decir el dataset está desbalanceado.

Sin embargo, para cualquier valor de IR es válido calcular las siguientes medidas de desempeño:

$$Recall = \frac{TP}{TP + FN} = \frac{10}{10 + 11} = \frac{10}{21} = 0.476$$

$$Precision = \frac{TP}{TP + FP} = \frac{10}{10 + 6} = \frac{10}{16} = 0.625$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0.625 \times 0.476}{0.625 + 0.476} = 0.540$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$= \frac{(10 \times 79) - (6 \times 11)}{\sqrt{(10 + 6)(10 + 11)(79 + 6)(79 + 11)}} = 0.425$$

Porcentualmente, obtuvimos los siguientes resultados:

Medida de desempeño	Porcentaje (%)
Recall	47.6
Precision	62.5
F1	54.0
MCC	42.5

Tabla 2: Medidas de desempeño utilizadas

Conclusiones

Dado que el dataset está bastante desbalanceado es importante realizar las siguientes observaciones respecto a cada medida de desempeño:

Recall: Mide la proporción de positivos reales que se identificaron correctamente. Es especialmente útil en contextos donde es crítico no perderse los casos positivos (como en diagnósticos médicos). Un valor alto de recall es indicativo de que el modelo es efectivo en la identificación de la clase minoritaria. Esto no ocurre en los resultados obtenidos.

Precision: Mide la proporción de clasificaciones positivas que fueron realmente correctas. En datasets desbalanceados, ayuda a entender cómo el modelo maneja la clase minoritaria en términos de falsos positivos.

F1: Es la media armónica de Precision y Recall. Es especialmente útil en datasets desbalanceados porque toma en cuenta tanto los falsos positivos como los falsos negativos.

Coeficiente de Correlación de Matthews (MCC): Es una medida que captura la calidad de las clasificaciones binarias, considerando todas las categorías de la matriz de confusión (TP, TN, FP, FN). El MCC proporciona una visión más integral del desempeño del clasificador y es particularmente valioso en datasets desbalanceados.

En cuanto a los resultados obtenidos ($recall$ del 47.62%, $precision$ del 62.50%, $F1$ del 54.05% y MCC del 45.16%) indican un rendimiento del modelo de bajo a moderado. El recall es muy bajo, lo que sugiere que el modelo está perdiéndose una cantidad significativa de casos positivos (clase minoritaria), lo cual es muy grave en un contexto de detección de enfermedades.

Los valores tan bajos de las medidas de desempeño podrían deberse a la métrica utilizada. Es importante comparar estos resultados con los obtenidos usando métricas más tradicionales para comprender si el bajo rendimiento se debe al clasificador 3-NN o a la métrica Bray-Curtis.

Referencias bibliográficas

[1] Ehsani, R., & Drabløs, F. (2020). Robust Distance Measures for kNN Classification of Cancer Data. Cancer informatics, 19, 1176935120965542. https://doi.org/10.1177/1176935120965542

[2] KEEL: a software tool to assess evolutionary algorithms for data mining problems (regression, classification, clustering, pattern mining and so on). (s. f.). https://sci2s.ugr.es/keel/dataset.php?cod=183

Anexos

```python
from sklearn.model_selection import LeaveOneOut
from sklearn.neighbors import KNeighborsClassifier
from scipy.spatial.distance import braycurtis
import numpy as np

loo = LeaveOneOut()

# Preparamos los datos
X = data.iloc[:, :-1]  # features
y = data.iloc[:, -1]   # clase

# Función de distancia para Bray-Curtis
def bray_curtis_distance(x, y):
    return braycurtis(x, y)

# 3-NN con distancia de Bray-Curtis
knn = KNeighborsClassifier(n_neighbors=3, metric=bray_curtis_distance)

# Aplicamos Leave-one-out
for train_index, test_index in loo.split(X):
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]

    knn.fit(X_train, y_train)

    knn.predict(X_test)
```