Centro de investigación en Computación

Clasificación inteligente de Patrones

Matriz de confusión: elementos iniciales

Autor:
Diego Noguez Ruiz

Profesor:
Dr. Cornelio Yáñez Márquez

18 de Marzo 2023

Índice general

1. Introducción

Los cuatro elementos básicos de un cip son sumamente importantes. Por esta razón, durante nuestras últimas sesiones del curso, hemos enfatizado su importancia y buscado formas de abordarlos de manera más efectiva. Por ejemplo, en la tarea del RD 05, prestamos especial atención a un nuevo método de validación llamado Hold-out.

Hasta el momento, solo se ha utilizado una medida para evaluar el rendimiento de un modelo, conocida como Accuracy. Sin embargo, es importante destacar que esta medida solo es aplicable cuando se trabaja con conjuntos de datos equilibrados. Por lo tanto, se requiere la introducción de nuevas medidas de desempeño que puedan ser utilizadas en conjuntos de datos no equilibrados. Para abordar esta cuestión, se ha comenzado a estudiar la matriz de confusión, ya que muchas de las nuevas medidas de desempeño se basan en ella.

En esta tarea, nos adentraremos en el concepto de la matriz de confusión con el objetivo de comprenderla mejor y calcularla para datasets reales.

2. Desarrollo

2.1. Propósito de la tarea

Calcular y representar los valores de las celdas en las matrices de confusión para problemas biclase.

2.2. Parte 1:

2.3. 1.a)

Instrucciones: Generar una partición 70-30 con el método de validación Hold-out en el dataset Nutt y aplicar el Clasificador Eulidiano con la partición generada.

Para esta parte usaremos el dataset Nutt [1].

Con anterioridad ya realizamos una función que genera una partición mediante el método de validación Hold-out, la cual se usará a continuación.

Recordemos que

$$|C_0| = |C_1| = 14 \longrightarrow IR = \frac{|C_0|}{|C_1|} = 1$$

Por lo tanto $IR = 1$ y hablamos de un dataset balanceado.

Dado que cada clase consta de 14 patrones y se desea aplicar un método Hold-out de 70-30, se procederá a seleccionar un 70 % de los patrones para entrenamiento y un 30 % para prueba de cada clase. En este caso, al realizar la

operación $14 \times 0.70 = 9.8$, se obtiene que se seleccionarán 10 patrones para entrenamiento y 4 patrones para prueba por clase.

Realizando el Hold-out 70-30 y usando la misma notación de la tarea pasada para describir los patrones que son elementos de un conjunto tenemos lo siguiente:

- $E_0 = \{1, 12, 9, 11, 10, 0, 8, 13, 4, 3\}$
- $E_1 = \{23, 14, 15, 17, 19, 18, 16, 26, 25, 20\}$

Si $E := E_0 \cup E_1$ y $P = D - E$ entonces E y P forman una partición de D.

Con el conjunto E entrené al clasificador euclidiano para este dataset.

A continuación los resultados obtenidos.

- El patrón 1 del conjunto de entrenamiento que pertenece a la clase 0:
  La distancia con el centroide 0 es: 6495.02
  La distancia con el centroide 1 es: 6543.11
  El punto está más cerca del centroide 0.
  La clasificación fue correcta.
- El patrón 2 del conjunto de entrenamiento que pertenece a la clase 0:
  La distancia con el centroide 0 es: 5882.16
  La distancia con el centroide 1 es: 7135.01
  El punto está más cerca del centroide 0.
  La clasificación fue correcta.
- El patrón 3 del conjunto de entrenamiento que pertenece a la clase 0:
  La distancia con el centroide 0 es: 7489.06
  La distancia con el centroide 1 es: 7544.91
  El punto está más cerca del centroide 0.
  La clasificación fue correcta.
- El patrón 4 del conjunto de entrenamiento que pertenece a la clase 0:
  La distancia con el centroide 0 es: 8974.21
  La distancia con el centroide 1 es: 9383.01
  El punto está más cerca del centroide 0.
  La clasificación fue correcta.

- El patrón 5 del conjunto de entrenamiento que pertenece a la clase 1:
  La distancia con el centroide 0 es: 4293.71
  La distancia con el centroide 1 es: 2057.54
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.
- El patrón 6 del conjunto de entrenamiento que pertenece a la clase 1:
  La distancia con el centroide 0 es: 4768.21
  La distancia con el centroide 1 es: 3030.72
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.
- El patrón 7 del conjunto de entrenamiento que pertenece a la clase 1:
  La distancia con el centroide 0 es: 4125.2
  La distancia con el centroide 1 es: 2494.03
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.
- El patrón 8 del conjunto de entrenamiento que pertenece a la clase 1:
  La distancia con el centroide 0 es: 4426.57
  La distancia con el centroide 1 es: 2066.46
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.

2.4. 1.b)

Instrucciones: Reportar accuracy.

Notemos que para todos los patrones de prueba el clasificador les asignó la clase correcta por lo tanto tiene un accuracy de 1.

2.5. 1.c)

Instrucciones: Reportar la matriz de confusión.

Para poder hablar de una matriz de confusión, es importante definir previamente cual será la clase positiva. En el caso de este dataset, sin pérdida de generalidad, se le puede asignar la clase positiva a los patrones de la clase 0 y la clase negativa a los patrones de la clase 1. Una vez dada esta asignación, se puede construir la matriz de confusión correspondiente, la cual tendrá la siguiente forma:

$$M = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$

2.6. 1.d)

Instrucciones: Calcular el valor de accuracy a partir de la matriz de confusión.

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP} = \frac{4 + 4}{4 + 4 + 0 + 0} = 1$$

2.7. 1.e)

Instrucciones: Verificar que son iguales los valores de los incisos 1.a) y 1.d).

Es trivial verificar que el accuracy es 1 en los dos incisos.

2.8. Parte 2:

Buscar un dataset (cardinalidad, dimensión e IR libres) con sólo atributos numéricos y sin valores perdidos.

Para esta parte usaremos el Titanic data set Data Set proporcionado por Keel [2].

2.9. 2.a)

Instrucciones: Realizar un proceso de limpieza de datos, a fin de eliminar los patrones redundantes y todos los patrones indiscernibles.

De inicio se tenia un total de 2201 patrones, después de la limpieza tuvimos un total de 14 patrones, donde 8 son de la clase 1 (no sobrevivió) y 6 de la clase -1 (si sobrevivieron)

2.10. 2.b)

Instrucciones: En el dataset del inciso 2.a), generar una partición 70-30 con el método de validación Hold-out y aplicar el Clasificador Eulidiano con la partición que se ha generado.

- El conjunto E_1 tendrá 6 patrones y $P1$ un total de 2
- El conjunto E_2 tendrá 4 patrones y $P2$ un total de 2

Los resultados al clasificar los patrones del conjunto de pruba obtengo lo siguiente:

- El patrón 1 del conjunto de entrenamiento que pertenece a la clase 1:
  La distancia con el centroide 0 es: 3.32
  La distancia con el centroide 1 es: 1.82
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.
- El patrón 2 del conjunto de entrenamiento que pertenece a la clase 1:
  La distancia con el centroide 0 es: 3.54
  La distancia con el centroide 1 es: 2.36
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.
- El patrón 3 del conjunto de entrenamiento que pertenece a la clase -1:
  La distancia con el centroide 0 es: 3.54
  La distancia con el centroide 1 es: 2.36
  El punto está más cerca del centroide 1.
  La clasificación fue INCORRECTA.
- El patrón 4 del conjunto de entrenamiento que pertenece a la clase -1:
  La distancia con el centroide 0 es: 2.06
  La distancia con el centroide 1 es: 3.67
  El punto está más cerca del centroide 0.
  La clasificación fue INCORRECTA.

2.11. 2.c)

Instrucciones: Reportar la matriz de confusión. Definamos a la clase de los sobrevivientes (-1) como la clase positiva, entonces la matriz de confusión esta dada por:

$$M = \begin{bmatrix} 0 & 2 \\ 0 & 2 \end{bmatrix}$$

2.12. 2.d)

Instrucciones: Si es adecuado, reportar el valor de accuracy. Explicar.

Para este caso en particular $IR = 8/6 = 1.3$ por lo que el dataset es balanceado y podemos hablar de accuracy. Usando la matriz de confusión notamos que el accuracy es de $2/4 = .5$ lo cual significa que solo la mitad de los patrones de prueba fueron clasificados correctos, de hecho, gracias a la matriz de confusión podemos observar que aquellos patrones clasificados a su clase original fueron los de la clase negativa, sin embargo los de la clase positiva fueron clasificados de manera errónea.

2.13. Parte 3:

Preguntas que podrían guiar el estudio de los elementos iniciales de una matriz de confusión.

- ¿Cómo se definen los conceptos TP, Tn, Fp y FN? True Positive, True Negative, False Positive y False Negative respectivamente
- ¿Qué hay en la fila superior? Número de patrones que originalmente son de la clase positive
- ¿Qué hay en la fila inferior? Número de patrones que originalmente son de la clase negative
- ¿Qué hay en la columna izquierda? Número de patrones que fueron clasificados como positive
- ¿Qué hay en la columna derecha? Número de patrones que fueron clasificados como negative
- ¿Qué hay en la intersección de la columna izquierda con la fila inferior? Los False Positive
- ¿Qué hay en la intersección del renglón superior y columna izquierda? Los True Positive
- ¿Qué hay en las diagonales? Los TP y TN
- ¿Dónde están los patrones que no son de la clase negative? En la primera fila
- ¿Dónde están los errores? ¿y los aciertos? Los errores están en la esquina superior derecha y en la esquina inferior izquierda. Por otro lado, los aciertos están en la diagonal principal de la matriz.
- ¿Cómo se calcula accuracy a partir de la matriz de confusión? $accuracy = \frac{TP+TN}{TP+TN+FN+FP}$

3. Conclusiones y trabajo futuro

De la tarea realizada he concluido que ha sido importante introducir el concepto de matriz de confusión, ya que será de mucha utilidad al momento de querer estudiar nuevas medidas de desempeño, ya que nos han comentado que la mayoría de estas se basan en la matriz de confusión.

En futuras clases iremos introduciendo nuevas medidas de desempeño que nos ayuden a analizar resultados de nuestros clasificadores para datasets generales.

Referencias

[1] B. Haznedar, M. T. Arslan, and A. Kalinli, "Microarray Gene Expression Cancer Data," vol. 4, Apr. 2017. Publisher: Mendeley Data.

[2] Keel, "Titanic data set." https://sci2s.ugr.es/keel/dataset.php?cod=189. Visitado en Marzo 2023.