Centro de investigación en Computación

Clasificación inteligente de Patrones

Método de validación Hold-out

Autor:
Diego Noguez Ruiz

Profesor:
Dr. Cornelio Yáñez Márquez

18 de Marzo 2023

Índice general

Índice general **1**
1. Introducción 2
2. Desarrollo 2
   - 2.1. Propósito de la tarea 2
   - 2.2. Parte 1: 2
   - 2.3. 1.a) 2
   - 2.4. 1.b) 4
   - 2.5. 1.c) 4
   - 2.6. Parte 2: 6
   - 2.7. 2.a) 6
   - 2.8. 2.b) 6
   - 2.9. 2.c) 6
3. Conclusiones y trabajo futuro 7

1. Introducción

En las últimas semanas de nuestro curso de CIP, hemos empezado a manipular datasets y aprendimos acerca de los elementos básicos de un cip. En particular, estudiamos el método de validación, el cual es sumamente importante, ya que la elección que se haga en este paso puede influir de manera positiva o negativa en los resultados del clasificador.

En la tarea anterior, utilizamos el método de validación de partición fija, en el cual el desarrollador elige de manera arbitraria la partición del dataset. Este método de validación se considera simple y fue de mucha utilidad para entender este elemento básico. Sin embargo, surge la pregunta de si este método es efectivo. En general, no podemos considerarlo como una buena elección, ya que el desarrollador podría seleccionar la partición que mejor le convenga para potenciar los resultados de su clasificador.

Por ello, en esta tarea vamos a analizar un nuevo método de validación llamado Hold-Out, para compararlo con el método estudiado anteriormente y analizar sus ventajas y desventajas (si es que existen).

2. Desarrollo

2.1. Propósito de la tarea

Generar particiones 80-20 con el método de validación Hold-out y aplicar el Clasificador Euclidiano usando esas particiones.

2.2. Parte 1:

2.3. 1.a)

Instrucciones: Generar una partición 80-20 con el método de validación Hold-out en el dataset Nutt y aplicar el Clasificador Eulidiano con la partición generada.

Para esta parte usaremos un dataset de cáncer de cerebro, realizado por C. L. Nutt[1]. El dataset contiene 28 patrones, de los cuales la mitad pertenecen a la clase de glioblastoma clásico, mientras que los patrones restantes corresponden a la clase de glioblastoma no clásico. Cada patrón se compone de un total de 1070 atributos numéricos.[2]

Notemos que si C_i con $i = 0, 1$ es el conjunto que tiene como elementos a los patrones de la clase i, entonces, para este caso en particular:

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

Por lo tanto tenemos un dataset balanceado.

En vista de que emplearemos nuevamente el clasificador euclidiano, pero esta vez con patrones que constan de más de 2 atributos, procederé a generalizar las definiciones presentadas en la tarea anterior.

Definición 1. *Sean u y v dos vectores de dimensión n, entonces la distancia euclidiana se define como:*

$$d(u,v) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2}$$

Donde u_i y v_i son los componentes de los vectores u y v en la posición i, respectivamente.

A continuación recordaré la definición de centroide.

Definición 2. *El centroide* $\mathbf{c}$ *de un conjunto de m vectores $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ en $\mathbb{R}^n$ se define como:*

$$\mathbf{c} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{v}_i = \frac{1}{m}(\mathbf{v}_1 + \mathbf{v}_2 + \cdots + \mathbf{v}_m)$$

Adicionalmente, daré una definición de la asignación realizada por el clasificador euclidiano.

Definición 3. *Sean c_1 y c_2 los respectivos centroides de los patrones de la clase 1 y 2 del conjunto de entrenamiento. Sea p_i un patrón arbitrario del conjunto de prueba y sea*

$$d_c := min\{d(p_i, c_1), d(p_i, c_2)\}$$

entonces el clasificador euclidiano le asignará la clase c_j al patrón p_i sii $d_c = d(p_i, c_j)$.

A continuación los pasos que seguí para realizar una partición 80-20 usando el método de validación Hold-out.

1. Crear a partir del dataset D los respectivos datasets C_i, los cuales contienen solo los patrones de la clase i

2. Si $|C_i| = n_i$ entonces calcular el $k_i := .80(n_i)$ para conocer cuantos patrones se irán al conjunto de entrenamiento de la respectiva clase i. El número de patrones para el conjunto de prueba de la clase i será $n_i - k_i$. Nota: En caso de que k_i no sea un valor entero, le puedo aplicar la función piso $\lfloor k_i \rfloor$, para así tener un valor entero.

3. Para cada uno de los datasets C_i realizar un desordenamiento aleatorio de los patrones.

4. Del conjunto C_i tomar los primeros k_i patrones y construir el conjunto E_i de entrenamiento, los $n_i - k_i$ patrones restantes de C_i formarán el conjunto de prueba P_i. Los conjuntos E_i, P_i con $i = 0, 1$ son los respectivos conjuntos de entrenamiento y prueba de C_i.

Dado que en este caso no resulta nada trivial describir cada uno de los patrones debido a la gran cantidad de atributos que poseen, optaré por utilizar el índice de cada patrón para indicar el conjunto al que pertenece. Por ejemplo, i es el índice del patrón i, entonces si $i \in A$, entonces el patrón i es elemento del conjunto A. Una vez dicho esto, dado que $E := E_0 \cup E_1$ donde :

- $E_0 = \{10, 8, 12, 4, 1, 5, 7, 6, 0, 2, 3\}$
- $E_1 = \{26, 27, 14, 19, 20, 25, 16, 24, 17, 18, 22\}$

Y $P = D - E$ entonces tenemos la siguiente afirmación.

Afirmación: E y P forman una partición de D.

Para este caso en particular, dado que ambas clases tienen 14 patrones, cada uno de los conjuntos de entrenamiento tiene 11 patrones y los conjuntos de prueba tienen 3 patrones cada uno.

Con los conjuntos E_i entrené el clasificador euclidiano.

2.4. 1.b)

Instrucciones: Reportar accuracy.

Dado que el dataset con el que estamos trabajando es balanceado, entonces tiene sentido hablar de accuracy. Recordemos la definición de accuracy.

Definición 4.

$$Accuracy := \frac{N\acute{u}mero\ de\ predicciones\ correctas}{N\acute{u}mero\ total\ de\ predicciones}$$

Para este caso obtuve $accuracy = \frac{4}{6}$

2.5. 1.c)

Instrucciones: Reportar los detalles de los aciertos y errores.

A continuación los resultados que se obtuvieron al aplicar el clasificador a los patrones del conjunto de prueba.

- El patrón 1 del conjunto de prueba que pertenece a la clase 0:

  La distancia con el centroide 0 es: 9004.73

La distancia con el centroide 1 es: 9985.84

El punto está más cerca del centroide 0.

La clasificación fue correcta.

- El patrón 2 del conjunto de prueba que pertenece a la clase 0:
  La distancia con el centroide 0 es: 6773.78
  La distancia con el centroide 1 es: 5969.28
  El punto está más cerca del centroide 1.
  La clasificación fue INCORRECTA.
- El patrón 3 del conjunto de prueba que pertenece a la clase 0:
  La distancia con el centroide 0 es: 5284.35
  La distancia con el centroide 1 es: 4273.86
  El punto está más cerca del centroide 1.
  La clasificación fue INCORRECTA.
- El patrón 4 del conjunto de prueba que pertenece a la clase 1:
  La distancia con el centroide 0 es: 4098.33
  La distancia con el centroide 1 es: 3293.38
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.
- El patrón 5 del conjunto de prueba que pertenece a la clase 1:
  La distancia con el centroide 0 es: 3733.2
  La distancia con el centroide 1 es: 3207.04
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.
- El patrón 6 del conjunto de prueba que pertenece a la clase 1:
  La distancia con el centroide 0 es: 4324.91
  La distancia con el centroide 1 es: 3903.65
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.

2.6. Parte 2:

2.7. 2.a)

Instrucciones: Generar una partición 80-20 con el método de validación Hold-out en el dataset Haberman's Survival Data Set del año 58 y aplicar el Clasificador Eulidiano con la partición generada.

Para esta parte usaremos el Haberman's Survival Data Set del año 58 [3].

El conjunto de datos consta de 28 patrones en total, de los cuales 20 pertenecen a la clase 1 y 8 a la clase 2. Esto indica claramente que se trata de un conjunto de datos desequilibrado, con un IR de 2.5.

Al aplicar el método de validación Hold-out procedí de manera análoga a la parte 1, obteniendo lo siguiente:

- $E_1 = \{18, 0, 14, 5, 12, 4, 7, 13, 19, 8, 17, 10, 9, 3, 11, 15\}$
- $E_2 = \{25, 21, 26, 24, 22, 23\}$

Recordemos que $E := E_1 \cup E_2$, entonces $P := D - E$, claramente E y P forman una partición para D.

2.8. 2.b)

Instrucciones: Reportar accuracy.

Como hemos dicho en 2.a) el dataset es desbalanceado, por lo cual no podemos hablar de accuracy.

2.9. 2.c)

Instrucciones: Reportar los detalles de los aciertos y errores.

- El patrón 1 del conjunto de prueba que pertenece a la clase 1:
  La distancia con el centroide 0 es: 7.29
  La distancia con el centroide 1 es: 3.47
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.
- El patrón 2 del conjunto de prueba que pertenece a la clase 1:
  La distancia con el centroide 0 es: 9.98
  La distancia con el centroide 1 es: 15.02
  El punto está más cerca del centroide 0.
  La clasificación fue INCORRECTA.

- El patrón 3 del conjunto de prueba que pertenece a la clase 1:
  La distancia con el centroide 0 es: 33.18
  La distancia con el centroide 1 es: 32.76
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.
- El patrón 4 del conjunto de prueba que pertenece a la clase 1:
  La distancia con el centroide 0 es: 13.95
  La distancia con el centroide 1 es: 18.84
  El punto está más cerca del centroide 0.
  La clasificación fue INCORRECTA.
- El patrón 5 del conjunto de prueba que pertenece a la clase 2:
  La distancia con el centroide 0 es: 51.17
  La distancia con el centroide 1 es: 48.55
  El punto está más cerca del centroide 1.
  La clasificación fue INCORRECTA.
- El patrón 6 del conjunto de prueba que pertenece a la clase 2:
  La distancia con el centroide 0 es: 32.13
  La distancia con el centroide 1 es: 27.99
  El punto está más cerca del centroide 1.
  La clasificación fue INCORRECTA.

3. Conclusiones y trabajo futuro

De la tarea realizada he concluido que aunque el método de validación Hold-out es un poco mejor que el de partición fija, aún no se puede considerar como un método de validación útil en general, debido a que dependiendo de la permutación realizada en los patrones, obtendremos un resultado distinto al momento de clasificarlos.

Conjeturo que pronto veremos algún otro método de validación que no tenga los defectos de los métodos estudiados hasta el día de hoy.

Referencias

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

[2] C. Yáñez-Márquez, "Toward the bleaching of the black boxes: Minimalist machine learning," *IT Professional*, vol. 22, no. 4, pp. 51–56, 2020.

[3] UCI ML, "Haberman's survival data set." https://archive.ics.uci.edu/ml/datasets/haberman's+survival. Visitado en Marzo 2023.