Centro de investigación en Computación

Clasificación inteligente de Patrones

Elementos básicos de un clasificador inteligente de patrones (cip)

Autor:
Diego Noguez Ruiz

Profesor:
Dr. Cornelio Yáñez Márquez

4 de Marzo de 2023

Índice general

Índice de figuras

1. Introducción

Los datasets son la materia prima del curso CIP, por lo cual, durante las clases pasadas nos habíamos estado enfocando principalmente en ellos, iniciando por conocer repositorios para su búsqueda y obtención, estudiamos diversos tipos de datasets para determinadas tareas y analizamos sus peculiaridades y características.

Ahora ha llegado el momento de empezar a manipular los datasets. Dado que la filosofía de nuestro curso es ir de lo simple a lo complejo, hemos iniciado a trabajar con datasets que solo tienen atributos numéricos (no categóricos, ni valores perdidos), tienen una cardinalidad pequeña y son biclase, para así facilitar el entendimiento de lo que se está haciendo.

En esta tarea, se realizarán por primera vez de manera individual los 4 elementos básicos de un clasificador inteligente de patrones, para poder reforzar lo aprendido en clase.

2. Desarrollo

2.1. Propósito de la tarea

Proponer una partición arbitraria y desarrollar los cuatro elementos básicos de un cip sencillo.

2.2. Instrucciones

En el dataset del año 58 ya estudiado, proponer una partición fija de modo que en el conjunto de prueba queden 4 patrones de cada una de las dos clases (8 en total). Reportar el desarrollo completo y el valor de accuracy. NOTA: pueden hacerlo a mano, con calculadora o pueden programarlo.

2.3. Parte 1: Dataset

Para esta tarea usaremos el 'Haberman's Survival Data Set' [1] del repositorio UCI, debido a que este dataset está diseñado para la tarea de clasificación, cada patrón está asociado a una etiqueta de clase. Además, este dataset contiene datos relevantes para la actividad humana, ya que tiene que ver con pacientes que tuvieron una cirugía por cáncer de mama. De igual manera, como hemos visto en clase, este dataset satisface los requisitos de tener solo atributos numéricos, ser biclase y de cardinalidad relativamente pequeña (su cardinalidad será menor cuando sea filtrado); estas características lo hacen sencillo de manejar.

De manera particular, se nos solicita realizar el cip para el dataset del año 58. Por lo cual, primero, se requiere filtrar el dataset original y quedarnos solamente

con aquellos patrones que en su segundo atributo tengan el valor 58, quedando como resultado un nuevo dataset de cardinalidad 36, al cual le eliminaremos el atributo que representa el año, debido a que todos los patrones tienen el valor de 58 en su segundo atributo. Quedando así un dataset de dos atributos. Al analizar el nuevo dataset, notamos que hay 3 patrones indiscernibles (aparecen en las dos clases), por lo cual los eliminamos, ya que nosotros solo deseamos trabajar con single-label. De igual manera, encontramos 2 patrones redundantes, cada uno se repite una vez, por lo tanto, eliminaremos su repetición para no afectar la cardinalidad de las clases. Al eliminar todos estos patrones, nos queda un dataset de cardinalidad 28, donde las respectivas cardinalidades de las clases son:

- Clase 1: 20
- Clase 2: 8

Entonces, el IR=$20/8 = 2.5$, lo cual significa que el dataset es desbalanceado.

2.4. Parte 2: Método de validación

Sea D el dataset con el que vamos a trabajar, $E \subset D$ el conjunto de entrenamiento y $P \subset D$ el conjunto de prueba. Donde

- $P = \{[33, 10], [34, 30], [37, 0], [39, 0], [43, 52], [44, 9], [46, 2], [48, 11]\}$
- $E = D - P$

Nota: Los primeros 4 elementos de P son los primeros 4 patrones de la clase 1, los restantes corresponden a los primeros 4 de la clase 2.

Claramente E y P forman una partición de D.

2.5. Parte 3: Algoritmo

A continuación unas definiciones que son de utilidad para nuestro primer cip:

Definición 1. *Sean $P = (x_1, y_1)$ y $Q = (x_2, y_2)$ dos puntos en el plano, entonces la distancia euclidiana en el plano se define como:*

$$d(P, Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Definición 2. *Sea V un conjunto de n vectores en el plano, $V = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}\}$, cada vector tiene coordenadas (x_i, y_i), entonces el centroide $\mathbf{c}$ se define como:*

$$\mathbf{c} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} y_i \right)$$

Ahora procederemos a hacer la fase de entrenamiento con nuestros datos, el cual consiste en encontrar el centroide de cada una de las clases usando solo el conjunto de entrenamiento. A continuación una figura con lo obtenido.

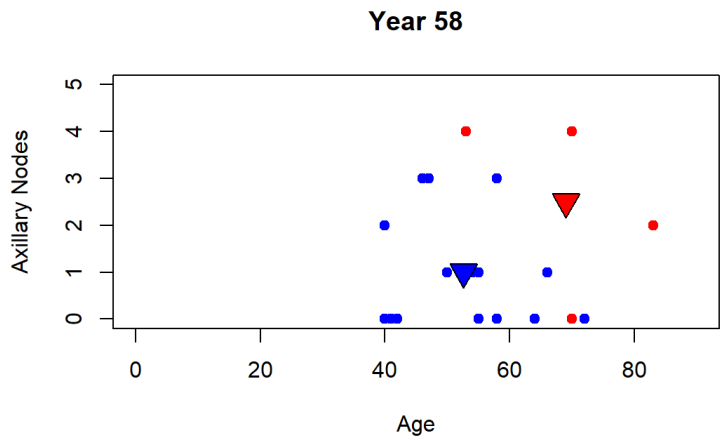

Figura 1: Fase de entrenamiento

Los puntos azules representa a los de la clase 1, y los rojos a los de la clase 2, además, los triángulos representan los centroides c_1, c_2 de cada una de las respectivas clases, los cuales tienen las siguientes coordenadas:

- $c_1 = [52.5625, 1]$
- $c_2 = [69, 2.5]$

El siguiente paso es probar el cip con los patrones del conjunto P. A continuación los resultados:

- El patrón analizado es [33, 10] que pertenece a la clase 1
  La distancia con el centroide 1 es: 21.53
  La distancia con el centroide 2 es: 36.77
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.
- El patrón analizado es [34, 30] que pertenece a la clase 1
  La distancia con el centroide 1 es: 34.43
  La distancia con el centroide 2 es: 44.51
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.

- El patrón analizado es [37, 0] que pertenece a la clase 1
  La distancia con el centroide 1 es: 15.59
  La distancia con el centroide 2 es: 32.1
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.
- El patrón analizado es [39, 0] que pertenece a la clase 1
  La distancia con el centroide 1 es: 13.6
  La distancia con el centroide 2 es: 30.1
  El punto está más cerca del centroide 1.
  La clasificación fue correcta.
- El patrón analizado es [43, 52] que pertenece a la clase 2
  La distancia con el centroide 1 es: 51.89
  La distancia con el centroide 2 es: 55.91
  El punto está más cerca del centroide 1.
  La clasificación fue INCORRECTA.
- El patrón analizado es [44, 9] que pertenece a la clase 2
  La distancia con el centroide 1 es: 11.72
  La distancia con el centroide 2 es: 25.83
  El punto está más cerca del centroide 1.
  La clasificación fue INCORRECTA.
- El patrón analizado es [46, 2] que pertenece a la clase 2
  La distancia con el centroide 1 es: 6.64
  La distancia con el centroide 2 es: 23.01
  El punto está más cerca del centroide 1.
  La clasificación fue INCORRECTA.
- El patrón analizado es [48, 11] que pertenece a la clase 2
  La distancia con el centroide 1 es: 10.99
  La distancia con el centroide 2 es: 22.66
  El punto está más cerca del centroide 1.
  La clasificación fue INCORRECTA.

Notemos que todos los patrones fueron asignados a la clase 1.

2.6. Parte 4: Medida de desempeño

La única medida de desempeño que conocemos al momento es el acuuracy. Sin embargo solo tiene sentido hablar de accuracy, cuando trabajamos con datasets balanceados, y en este caso en particular nuestro dataset es desbalanceado, por lo tanto no tiene sentido hablar de accuracy para este dataset.

3. Conclusiones

De la tarea realizada he concluido que cada uno de los 4 elementos básicos son sumamente importantes para la construcción de un clasificador inteligente de patrones, ya que las decisiones tomadas en cada elemento básico influirán en el desempeño del clasificador.

En particular, me ha gustado realizar este primer clasificador a pesar de ser considerado sencillo, debido a que ayuda a ejemplificar de manera clara los elementos básicos del cip.

Me encuentro intrigado y emocionado por conocer más clasificadores.

Referencias

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