UCI Machine Learning Repository

Balloon databases

1. Detalles

Para poner en práctica nuestros conocimientos y determinar si una muestra es suficiente para realizar una tarea de minería de datos se hizo la elección de una base de datos relativamente pequeña llamada “Balloon”.

“Balloon” fue llevada a cabo para poner en práctica un problema de predicción de cuándo se inflará un globo con éxito. Teniendo en cuenta esta forma más débil de forma de conocimiento de fondo, fue posible producir varias explicaciones de por qué ocurrió un resultado particular. Dado que el conjunto de hipótesis que son consistentes con el conocimiento de fondo y los datos son mucho más que el conjunto de hipótesis que son consistentes sólo con los datos, se necesitaron menos ejemplos para aprender a hacer predicciones precisas.[1]

La base de datos cuenta con las siguientes especificaciones:

- 16 instancias
- 4 atributos
- 2 clases

Donde:

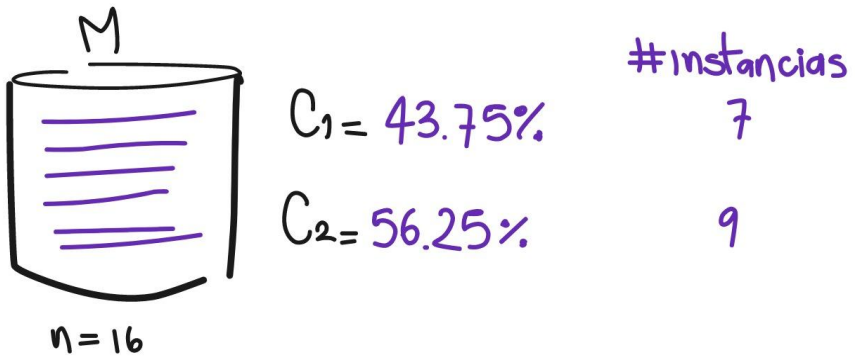
- Atributos
  - Color yellow, purple
  - size large, small
  - act stretch, dip
  - age adult, child

- Clases
  - Inflated True
  - Inflated False

Todos los datos ya se encuentran con la etiquetación adecuada y con el total de 4 atributos por cada instancia. Además, es importante destacar cada instancia pertenece a una clase u otra de acuerdo a lo siguiente:

- La instancia es de clase Inflated True si (*color=yellow and size = small) or (age=adult and act=stretch).*
- En caso de que lo anterior no se cumpla, entonces la instancia es de clase Inflated False.

Una vez teniendo conocimiento de lo anterior podemos definir:



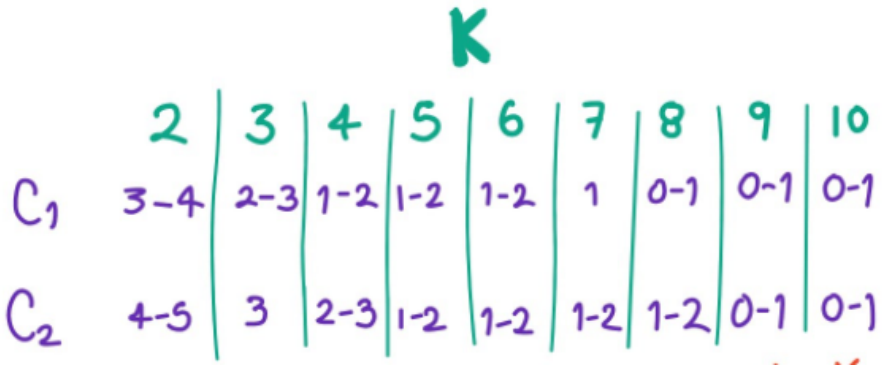
Donde:

- ***n*** Es el número de instancias de la base de datos
- C_1 Es la clase *Inflated True*
- C_2 Es la clase *Inflated False*.

Los porcentajes de cada clase, así como su número de instancias fueron calculados de acuerdo a los datos de las instancias almacenados en el archivo *yellow-small+adult-stretch* de la base de datos.

2. Determinación del valor de K

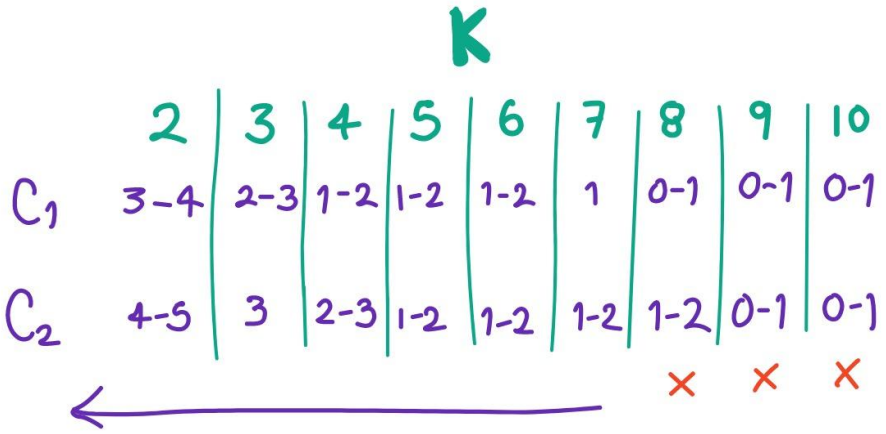
Se realizó el cálculo para cada valor de K partiendo del número de instancias de cada clase (C_1 = 7 y C_2 = 9).



Como se puede observar en la imagen anterior los intervalos de instancias que pueden ser representativas para C_1 y C_2 varían para cada valor de K.

Cabe mencionar que para una validación cruzada es importante tomar el máximo valor de K y en caso de que esto no sea posible, el valor mínimo de K debe ser 2.

Análisis cuantitativo



Desde un punto de vista estrictamente cuantitativo, es decir, solo tomando en cuenta el número de instancias de cada clase, se tiene que los posibles valores para **K** pueden ser de 2 a 7. Esto se debe a que a partir de K = 8 el número de instancias pasa a ser de 0 a 1 y el realizar una validación cruzada con 0 instancias de una clase no daría, por ningún motivo, resultados correctos. Por lo tanto los valores de K que sean >= 8 quedan completamente descartados.

Aunque se tiene que el valor de K puede ser de 2 a 7 es necesario poner en duda las situaciones donde K = 4, 5, 6, o 7 ya que en sus intervalos se interpreta que se puede tomar solo 1 o de 1 a 2 instancias de una clase y a pesar de que no es imposible el trabajar con 1 sola instancia que sea la representativa de toda una clase, sí es una decisión riesgosa para la obtención de un resultado que se busca sea lo más preciso, pero confiable posible.

Con los puntos anteriormente expuestos entonces se tiene que los valores de K más adecuados son 2 o 3. De estos dos valores se destaca que en sus intervalos se manejan 2 o más instancias de clase lo que se traduce en que se tendrá una validación cruzada mucho más apropiada en comparación con la que se tendría con los valores de K que ya han sido descartados. Sin embargo, dado a que es recomendable usar el valor de K más alto posible, entonces para la presente base de datos el valor de **K** debería ser **3**.

Análisis cualitativo

Pasando a un punto de vista más enfocado al aspecto cualitativo que en pocas palabras toma muy en cuenta los atributos de cada instancia, se llega a una conclusión diferente al del análisis cuantitativo.

Para empezar se tiene que la muestra no sería suficiente o adecuada para las diferentes variaciones que tienen ciertos atributos de este. Ejemplificando la idea en la base de datos tenemos que el atributo *color* podría presentar variaciones en los tonos que este maneja, todo pese a que solo existen únicamente dos colores. Para un valor K = 3 se puede tomar 2 o 3 instancias de C_1, pero esas instancias bien podrían ser o todas amarillas o todas moradas o una combinación de ambos atributos. Y es que en términos de la materia y teorías del color a menudo el tono se confunde con el término “color” cuando en realidad el tono hace referencia a una parte identificativa del color. Cuando empleamos esta propiedad siempre debemos tener en cuenta que nos estamos refiriendo únicamente al nombre de un determinado color sin tener en cuenta la luminosidad o la saturación [2].

Si el anterior punto no fuera lo suficientemente acertado para hacernos llegar a esa conclusión, existe otro atributo que puede llegar a tener variaciones, pues se usan dos tipos de edad en las que los sujetos podrían encontrarse: niño y adulto. Siendo que estos no están indicando el rango de edad donde se encuentran específicamente, no se tiene la certeza de que el rango que ocuparon fuera uno que se mantuviera, por lo que este es un motivo para pensar que una mayor cantidad en la muestra inicial sería más óptimo para este problema de predicción, ya que al haber más rangos de edad los resultados que se obtengan tomarían valores y resultados más variados.

Conclusiones

Desde un punto de vista cuantitativo el tamaño de la muestra es adecuada para poder realizar un análisis de minería de datos con un manejo del valor de K = 3 para una variación cruzada. Sin embargo, si se quisiera tomar en cuenta y explorar los diferentes valores que puede tomar cada uno de los atributos de las instancias, entonces la muestra no sería la apropiada para poder darnos resultados verídicos.

Referencias

- [1] Pazzani, M. J. (1989, diciembre). *Explanation-based learning with weak domain theories*. CiteSeerX. Recuperado 15 de enero de 2022, de https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.7238&rep=rep1&type=pdf
- [2] Castillo, V. P. (2020, 2 septiembre). *Más propiedades del color: tono, brillo y saturación*. Victoria Pérez Castillo - Interiorista. Recuperado 15 de enero de 2022, de https://www.enverodeco.es/blog/propiedades-del-color-tono-brillo-y-saturacion