Pontificia Universidad Javeriana de Cali

Facultad de Ingeniería de sistemas y computación

Taller 1 - Conjunto de datos "Sick"

Aprendizaje Automático y Análisis de Datos

Autor:
Ana María García

Marzo 4 del 2020

1 Conociendo el conjunto de datos

Este conjunto de datos es una recolección de datos de personas para identificar si padecen o no de problemas en la tiroides.

Este dataset cuenta con 3 tipos de archivos: data, names y test. Donde data es un conjunto que se usa como entrenamiento y test un conjunto de prueba. Sin embargo, para facilitar ciertas funciones, se decidió juntas ambos conjuntos de datos y, posteriormente, dividirlos en los conjunto de entrenamiento y prueba con una proporción de 75 y 25 respectivamente.

En este conjunto unido se encuentran los siguientes 30 atributos, donde f y t indican False y True respectivamente y "continuous" indica que es un atributo numérico:

age: continuous.
sex: M, F.
on thyroxine: f, t.
query on thyroxine: f, t.
on antithyroid medication: f, t.
sick: f, t.
pregnant: f, t.
thyroid surgery: f, t.
I131 treatment: f, t.
query hypothyroid: f, t.
query hyperthyroid: f, t.
lithium: f, t.
goitre: f, t.
tumor: f, t.
hypopituitary: f, t.
psych: f, t.
TSH measured: f, t.
TSH: continuous.
T3 measured: f, t.
T3: continuous.
TT4 measured: f, t.
TT4: continuous.
T4U measured: f, t.
T4U: continuous.
FTI measured: f, t.
FTI: continuous.
TBG measured: f, t.
TBG: continuous.
referral source: WEST, STMW, SVHC, SVI, SVHD, other.

Podemos identificar que hay atributos que se pueden agrupar en pares. Estos tienen un atributo en f o t y su respectivo atributo pareja es un dato numérico. Cada vez que en la primer columna aparece un dato en t, indica que ese mismo dato en su atributo pareja tiene un valor numérico. Cada vez que en la primer columna aparece un dato en f, indica que ese mismo dato en su atributo pareja no tiene ningún valor, es decir, es nulo.

Estas parejas de columnas son:

TSH measured - TSH
T3 measured - T3
TT4 measured - TT4
T4U measured - T4U
FTI measured - FTI
TBG measured - TBG

2 Plan para ajustar los datos

Según nos indica la función data.describe() y data.isnull().sum(), los datos del atributo TBG son en su totalidad nulos, lo que a su vez indica que todos los datos del atributo TBG measured están en f.

Debido a que no existe ninguna variedad en los datos de ambas columnas, se ha decidido eliminarlas, pues no aportan valor alguno para el análisis de los datos.

Por otro lado, ya que hay ciertos datos nulos en los atributos numéricos, no se pueden realizar ciertas operaciones para llevar a cabo el análisis. Por ende, estos datos nulos serán reemplazados por la moda.

Como se mencionó anteriormente, estas parejas tienen un indicador el cual es falso si el valor de su atributo compañero es nulo, ya que no habrán valores nulos entonces los datos representados en estas columnas no tendrían sentido y por ende no aportaría valor alguno para el análisis. Por esta razón, todos los atributos "measured" serán eliminados.

Ya con estos cambios, se pueden llevar a cabo las operaciones para analizar los datos y posteriormente ajustarlos.

Al no tener datos nulos, se puede generar una matriz de correlación, la cual nos indica qué tan parecidos son los datos de los atributos. Si este coeficiente se acerca más a 1, significa que los 2 atributos comparados son muy parecidos, de lo contrario, si se acercan a -1, significa que los atributos son muy diferentes.

En este caso, el coeficiente más grande es 0,78 entre los atributos TT4 y FTI. Considero que este valor no es lo suficientemente cercano a 1 como para decidir eliminar uno de estos atributos, pues aún hay diferencias entre estos. Por esta

razón, ningún atributo será eliminado.

El atributo clasificador es el nombrado como 'sick-negative.—classes', el cual indica si la persona a quien se le tomaron los datos está enferma (sick) o está sana (negative). Se puede ver que estos datos aparecen acompañados con ".—" y seguidos de unos números únicos. Debido a estos números, no se puede hacer una división para clasificar los datos, por lo que se toma la decisión de eliminarlos, pues los números no son relevantes para el análisis. Luego de esto ya podemos clasificar los datos.

Posteriormente se revisa si los datos están balanceados, sin embargo podemos ver que están muy desiguales, pues la cantidad de datos "negative" son aproximadamente 15 veces más que los datos "sick". Debido a esto se toma la decisión de replicar los datos sick para igualarlos con la cantidad de negative.

Luego de balancear los datos, se revisan los datos atípicos, en este caso, por diagramas de cajas. En el diagrama de "age", podemos ver que hay un dato cuya edad supera los 400 años, lo cual no tiene sentido, por ende eliminamos este dato, pues puede estar corrompiendo los demás datos.

Finalmente, convertimos los datos a numéricos para tener un mejor rendimiento en el procesamiento.

3 Análisis de resultados

Al analizar los modelos, se puede concluir que, en este caso, la métrica de desempeño más importante es Recall, pues al interpretar la matriz de confusión, esta se centra en revisar aquellos casos donde había enfermedad y efectivamente se predijo, pero también aquellos casos donde había enfermedad pero no se pudo predecir. Considero que este es el más relevante puesto que si hay una persona enferma y esto no se detecta, puede generar consecuencias graves. A diferencia del caso de Precisión donde se puede detectar a una persona como enferma cuando no tiene la enfermedad, en este caso esto no es tan delicado pues con unos estudios extra se pueden sacar conclusiones más precisas.

Por esta razón, el análisis se hará teniendo en cuenta principalmente el Recall.

Se hicieron ciertas pruebas ajustando los datos de 3 formas distintas:

1- REEMPLAZANDO LOS NULOS POR LA MODA Y REPLICANDO SOLO EL CONJUNTO MINIMAL: Se pudo ver que en todas las metricas se tuvo valores cercanos al 0.9, lo que indica un buen modelo en cuanto a la detección de enfermedades.

2- REEMPLAZANDO LOS NULOS POR LA MODA, REPLICANDO EL CONJUNTO MINIMAL Y MINIMIZANDO EL MAXIMAL: Se pudo ver que en todas las metricas se tuvo valores cercanos al 0.8, lo cual también indica un buen modelo, sin embargo continúa siendo mejor el modelo anterior.

3- REEMPLAZANDO LOS NULOS POR CERO, REPLICANDO EL CONJUNTO MINIMAL Y MINIMIZANDO EL MAXIMAL: Se pudo ver que en todas las metricas se tuvo valores cercanos al 0.7, lo cual indica modelo mucho menos preciso que los anteriores. Debido a esto esta es la solución menos eficaz en cuanto al procesamiento de los datos.

Por otra lado se puede ver que, al comparar las metricas de desempeño, en los 3 casos anteriores se cumple que el resultado al aplicar K vecinos es más alto que en las otras métricas.

Por esta razón, la métrica K vecinos obtiene mejor desempeño.