

Experimento opción 4

De Fino - Solari Barrios - Wurzel

2023-08-22

Hipótesis pre-experimentación

Antes de llevar a cabo el experimento, planteamos varias posibles hipótesis sobre cómo las etiquetas ruidosas podrían afectar la performance de los árboles de decisión en diferentes conjuntos de datos. Las hipótesis iniciales incluyen:

1. Degradación de performance: Esperamos que a medida que aumenta la proporción de etiquetas con ruido, la performance de los árboles de decisión disminuya. A mayor cantidad ruido, se espera que el modelo tenga más dificultades para aprender patrones relevantes y, por lo tanto, su capacidad de predicción disminuya.
2. Mayor impacto en datasets pequeños: Suponemos que los efectos del ruido en las etiquetas podrían ser más pronunciados en conjuntos de datos más pequeños, como "Heart" y "Churn", en comparación con el conjunto de datos más grande, "CO2". Esto se debe a que en conjuntos pequeños, un número reducido de etiquetas erróneas podría tener un impacto más significativo.
3. Respuesta gradual en "CO2": Dado que el conjunto de datos "CO2" es el más grande por amplia diferencia (7385 observaciones contra 3150(Churn) y 918(Heart)), podría mostrar una respuesta más gradual al ruido en comparación con los otros conjuntos.

Experimentación

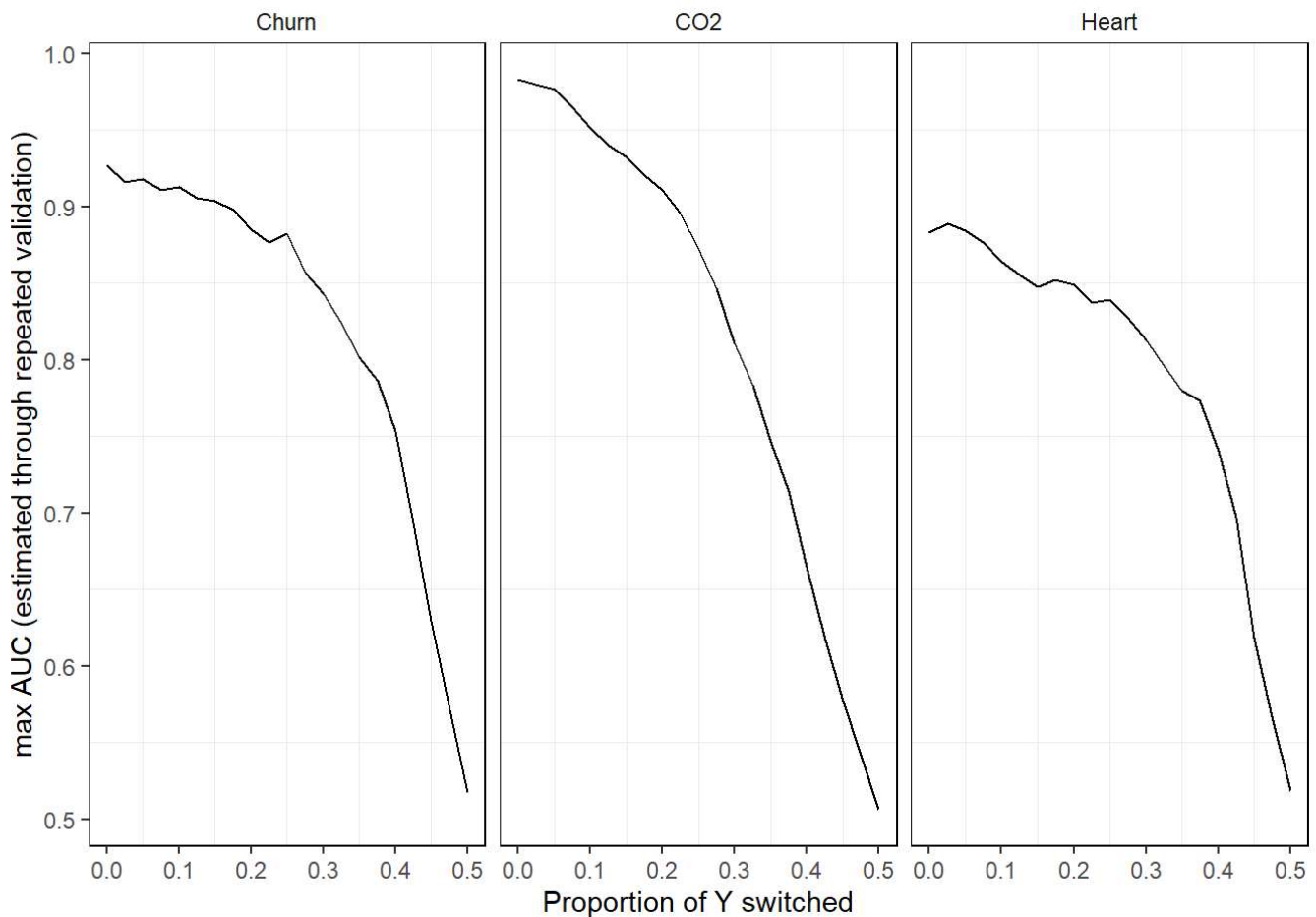
Importamos el script y el gráfico que este realiza:

```
#Cargamos el script
suppressWarnings(source("exp_4.R"))
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##   filter, lag
```

```
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```



Análisis de hipótesis

1. La experimentación revela que la introducción de ruido en las etiquetas (mediante el ajuste de `prop_switch_y`) tiende a deteriorar la performance general de los árboles de decisión en los tres conjuntos de datos. A medida que la proporción de etiquetas con ruido aumenta, el área bajo la curva (AUC) disminuye, indicando una disminución en la calidad de predicción de los modelos. Este patrón es consistente en todos los conjuntos de datos, sugiriendo que los árboles de decisión son sensibles a las noisy labels y que este factor puede tener un impacto significativo en la calidad de las predicciones.
2. Los resultados del experimento respaldan la idea de que el impacto del ruido en las etiquetas es más pronunciado en conjuntos de datos más pequeños. Tanto en el conjunto "Churn" como en "Heart", observamos una disminución más marcada en la performance a medida que aumenta la proporción de etiquetas con ruido. En contraste, en el conjunto más grande "CO2", el efecto del ruido es más gradual y menos pronunciado. Esta tendencia sugiere que en conjuntos de datos más pequeños, incluso una pequeña cantidad de etiquetas ruidosas puede tener un impacto más notable en la calidad de las predicciones de los árboles de decisión.
3. Inicialmente, consideramos la posibilidad de que en el conjunto de datos "CO2" hubiera una disminución del AUC gradual, con una pendiente no tan pronunciada. Sin embargo, el análisis de los resultados revela un gráfico que podría aproximarse a una lineal salvo por ligeras variaciones en la velocidad de disminución de la performance. Hasta `prop_switch_y=0.25` la disminución es más gradual y horizontal, lo que sugiere una menor sensibilidad al ruido. Sin embargo, pasado este punto, la pendiente se vuelve más pronunciada indicando un aumento más rápido en la degradación de la performance a medida que se introduce más ruido.

En general, estos resultados enfatizan la importancia de manejar y abordar el ruido en las etiquetas al trabajar con árboles de decisión y otros modelos similares. El ruido en las etiquetas puede deteriorar significativamente la calidad de las predicciones y tener un impacto desproporcionado en conjuntos de datos más pequeños.

Una observación extra

Observando los gráficos detenidamente notamos que en los tres conjuntos de datos se presenta un patrón común relacionado con la velocidad de degradación de la performance a medida que aumenta la proporción de etiquetas con ruido. En todos los casos, hasta una cierta proporción crítica de ruido, la disminución en el AUC es gradual y relativamente constante. Sin embargo, pasada esta proporción crítica, la curva de degradación adquiere una pendiente más pronunciada, indicando un rápido deterioro en la calidad de las predicciones. Este patrón puede estar relacionado con la manera en que los árboles de decisión toman decisiones en función de las etiquetas ruidosas, llegando a un punto en el que la calidad de las predicciones se ve comprometida drásticamente.

Conclusión

En este experimento, hemos demostrado que la introducción de etiquetas ruidosas tiene un impacto negativo en la performance de los árboles de decisión. A medida que la proporción de etiquetas ruidosas aumenta, el valor máximo de AUC disminuye, lo que indica una pérdida en la capacidad predictiva. Además, hemos observado que esta sensibilidad al ruido puede variar según el conjunto de datos y que existe un límite en la cantidad de ruido que el modelo puede manejar antes de que el rendimiento se degrade significativamente. Por lo tanto, al construir modelos de árboles de decisión, es crucial tener en cuenta la calidad de las etiquetas y considerar estrategias para mitigar el impacto del ruido.