ALGORITMOS II, 2C 2024

MEJORA DE LA PRECISIÓN DEL RANDOM FOREST

Team: Fernandez Silvia, Rodrigez Javier, Molina Juana

INTRODUCCIÓN

- Explorar cómo mejorar el rendimiento de los bosques aleatorios mediante bootstrap sample > 1.0.

- Añadir el parámetro nivel peso (inspirado en Dijkstra), de manera tal que nos permita ajustar la probabilidad de seleccionar muestras.

- Optimizar tanto la precisión como la generalización de los modelos.

HIPÓTESIS

- Objetivo: Mejorar la precisión predictiva en una tarea de clasificación mediante una variante del algoritmo Random Forest, ya que el muestreo estándar puede no capturar la variabilidad especialmente en conjuntos de datos desbalanceados.

- Propuesta: Ajustar la Tasa de Bootstrap (BR), usar Niveles de Pesos personalizados para aumentar la variabilidad, definir el numero de árboles y niveles de profundidad.

¿QUÉ ES LA TASA DE BOOTSTRAP (BR)?

¿Qué es el Bootstrap Rate (BR)?

- Proporción de datos seleccionado aleatoriamente para entrenar cada árbol de decisión.
- BR Estándar: Normalmente se fija en 1 (100% de las muestras).
- BR > 1: Oversampling o sobreselección de algunas muestras, permitiendo que cada árbol vea variaciones adicionales de los datos.

¿Por qué BR > 1 mejora el modelo?

- Captura patrones complejos al darle a cada árbol una muestra que puede enfatizar diferentes aspectos de los datos.
- Aumenta la robustez del modelo al reducir el sobreajuste en características específicas, ya que cada árbol recibe un conjunto de datos más variado o con mayor énfasis.

¿QUÉ SON LOS PESOS POR NIVEL?

Definición de Pesos:

Peso(X) =(FrecuenciaX)^nivel_peso

P(X) = Peso(X)/Sum(PesosTot)

E= [A,A,B,C,C,C,C]

2A 1B 4C

Peso(A) = √2

Peso(B) = √1

Peso(C) =√4

P(A) = 0.30

P(B) =0.21

P(C) = 0.49

Ei =[A,B,C,C]

nivel_peso= 1/2

¿POR QUÉ PENSAMOS QUE MEJORARIA EL MODELO?

- El parámetro nivel peso introduce un ajuste en las probabilidades de muestreo basada en la frecuencia
- Esto permite balancear nuestro muestreo y mejorar el modelo en tareas de clasificación.
- En cada árbol se realiza un muestreo con el método bootstrap y éste va a estar ajustado por el parámetro nivel peso.

CÓDIGO

https://github.com/JaviCeRodriguez/algoritmos2_tpi

RESULTADOS

Dataset	BR	Nivel Peso	Nro Árboles	Max Depth	Precisión (Paper)	Precisión (Nuestro)
Diabetes (*)	0.85	0.919	20	5	-	**0.775**
Breast Cancer (*)	0.85	1.13	20	5	-	**0.974**
Abalon	2.0	0.8	500	5	**0.268**	**0.291**
Statlog	1.2	1.2	10	5	**0.872**	**0.877**
Horse Colic	2.0	0.8	10	5	**0.865**	**0.784**
Hepatitis	0.2	0.8	10	5	**0.847**	**0.774**
Adult	0.2	0.8	10	5	**0.865**	**0.505**

Tabla 1: Resultado de análisis hechos sobre cada dataset.
(*) Datasets de Scikit Learn, el resto son de UCI ML Repository

CONCLUSIÓN

💪 Eficiencia en clasificación binaria

🔗 Dependencia de las características del dataset

📈 Potencial de mejora adaptativa

¡GRACIAS!

(PUEDEN APLAUDIR 👏🏻)