Algoritmo de clasificación AdaBoost para la predicción de sobrevivientes de pasajeros del Titanic

Técnicas emergentes

Vásquez V., C.R.A.

EPG - UNALM

05/12/2021

UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA

VII. Referencias

I. Introducción

I. Introducción

La aplicación de técnicas y algoritmos de Machine Learning (Aprendizaje Automático) se están volviendo cada vez más comunes, el uso en problemas poco habituales, diversos y hasta creativos se están popularizando. Problemas que van desde la clasificación de textos al reconocimiento de objetos o formas en imágenes, pasando por la predicción/detección de enfermedades y análisis de conductas, entre muchísimos otros usos. Las librerías scikit-learn y tensor flow para el lenguaje Python son algunas de las más utilizadas para crear modelos de clasificación y predicción, estas poseen implementaciones de los algoritmos más utilizados para las tareas mencionadas, por esta razón, utilizar estas herramientas reduce en gran medida el armado de los modelos. Como ya se adelantó, se realizó una comparación del algoritmo clasificador Ada Boost antes y después de la calibración de hiperparámetros y se obtuvo el mejor desempeño luego de la calibración.

Adaptive Boosting, o AdaBoost, parte de un vector de características y emplea un árbol de decisión como clasificador de base. Un comité de subclasificadores se forma a partir de un árbol de decisión. Con cada iteración del subclasificador, las instancias de entrenamiento que fueron etiquetadas erróneamente son más pesadas para asegurar que esas instancias reciban mayor atención a medida que se generen subsecuentes subclasificadores, de esta manera el algoritmo se adapta y logra obtener mejores resultados. Si bien este tipo de clasificador ha sido muy utilizado para la detección de automóviles, también fue implementado con éxito para identificar el sexo de una persona a partir de una imagen de su cara en escala de grises con baja resolución, o en la detección de puertas para robots móviles, o en la identificación de cabezas de ganado.

El presente estudio tuvo como objetivo predecir la sobrevivencia de pasajeros en la base de datos Titanic, haciendo uso del algoritmo de clasificación AdaBoost.

II. Marco teórico

II. Marco teórico

Las estrategias de Boosting pretenden elevar el desempeño de un algoritmo de aprendizaje débil combinando varias hipótesis adecuadamente generando un algoritmo de aprendizaje fuerte. El algoritmo AdaBoost genera un clasificador fuerte con clasificadores débiles. Los clasificadores débiles son los miembros del clasificador conjunto que utiliza el algoritmo AdaBoost. AdaBoost crea un comité de clasificadores débiles miembros ajustando adaptativamente los pesos en cada ciclo. Los pesos de las muestras de entrenamiento que han sido clasificadas erróneamente por un clasificador débil actual se incrementan, mientras que los pesos de las muestras de entrenamiento que han sido clasificadas correctamente por un clasificador débil actual se reducen. AdaBoost es un buen algoritmo para construir clasificadores de conjunto, pero no siempre puede generar clasificadores de conjunto para minimizar el error medio de generalización.

Se utiliza el algoritmo AdaBoost para seleccionar un pequeño número de características visuales críticas de un conjunto muy de un conjunto muy grande de características potenciales. AdaBoost proporciona un algoritmo de aprendizaje eficaz y fuertes límites en rendimiento de la generalización. Utilizaron el algoritmo AdaBoost para buscar un pequeño número de buenas características que que, sin embargo, tienen una variedad significativa. El algoritmo AdaBoost restringe el aprendiz débil al conjunto de funciones de funciones de clasificación, cada una de las cuales depende de una sola característica. Los clasificadores débiles que utilizan (umbral características únicas) pueden verse como árboles de decisión de un solo nodo.

III. Métodología

A. Datos utilizados

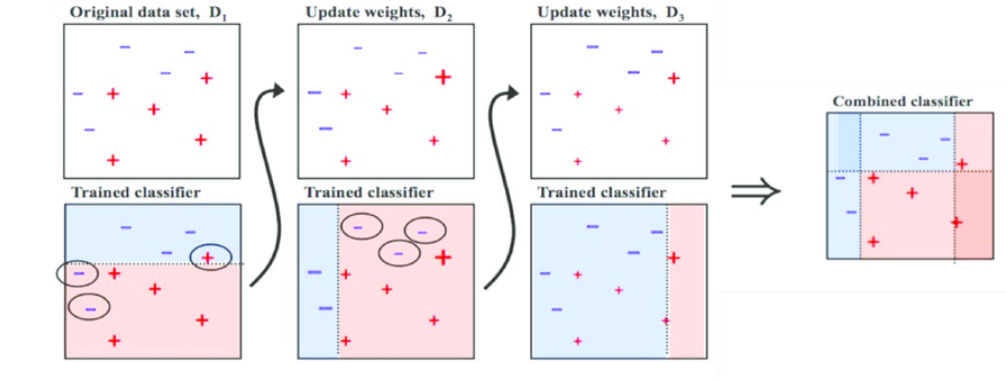
Se emplearon los datos originales de la competición Titanic de Kaggle (Reina, 2016). Los datos se han dividido en dos grupos: - Conjunto de entrenamiento (train.csv) - Conjunto de prueba (test.csv)

B. Algoritmo empleado

1. Algoritmo de clasificación AdaBoost

Fig 1.

Interación del algoritmo AdaBoost.

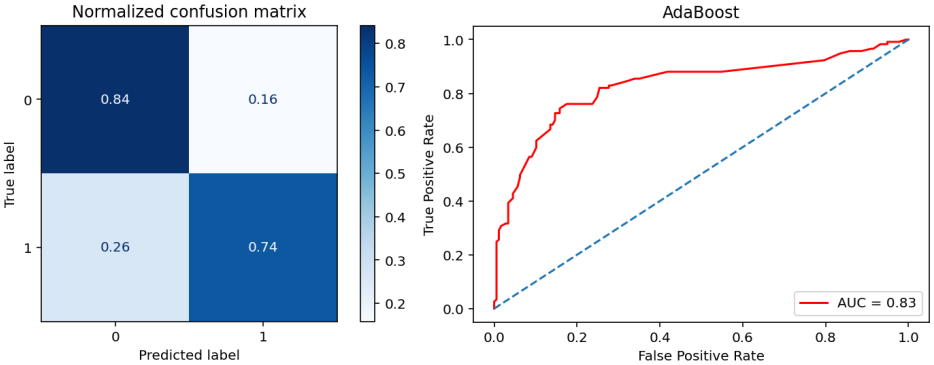


IV. Resultados

IV. Resultados

Fig 2.

Matriz de confusión normalizada, Curva ROC y AUC.



Al realizar predicciones de sobrevivencia de pasajeros de la base de datos Titanic empleando el algoritmo de clasificación AdaBoost con hiperparámetros calibrados (ratio de aprendizaje de 0.15 y 50 estimadores), se tiene una precisión de 79.93 %, sensibilidad de 73.5 % y especifidad de 84.18 % y un AUC de 0.8279.

V. Discusión

V. Discusión

Las ventajas del algoritmo de clasificación AdaBoost son:

- Consigue un mayor rendimiento que el bagging cuando los hiperparámetros se ajustan adecuadamente.
- Puede utilizarse para la clasificación y la regresión por igual.
- Maneja fácilmente tipos de datos mixtos.
- Puede utilizar funciones de pérdida "robustas" que hacen que el modelo sea resistente a los valores atípicos.

Sus desventajas son:

- Dificultad y largo tiempo para ajustar adecuadamente los hiperparámetros.
- No se puede paralelizar como el bagging (mala escalabilidad cuando hay grandes cantidades de datos).
- Más riesgo de sobreajuste en comparación con el bagging.

VI. Conclusión

VI. Conclusión

Con el algoritmo AdaBoost luego de calibrar hiperparámetros se obtiene 79.93 % de la precisión, con una sensibilidad de 73.5 % y una especificidad de 84.18 % para predecir la sobrevivencia de pasajeros con la data Titanic.

VII. Referencias

VII. Referencias

1. García, E. (s.f.). Adaboost aplicado a clasificación de fonemas. Universidad de los Andes. Recuperado de: https://elkingarcia.github.io/Papers/CWCAS06.pdf
2. Masum, R. (2019). AdaBoost with Titanic Dataset. Recuperado de: https://www.kaggle.com/masumrumi/adaboost-with-titanic-dataset/notebook
3. Medrano, J, (2018). Comparación de algoritmos de clasificación para predecir la condición de una persona como Fumador/No Fumador, a partir de una encuesta. Universidad Nacional de Jujuy. https://doi.org/ 10.13140/RG.2.2.11401.70246
4. Reina, T. (2016). Titanic with AdaBoost. Kaggle. Recuperado de: https://www.kaggle.com/treina/titanic-with-adaboost/data

5. Tae, A. & Moon, K. (2010). A New Diverse AdaBoost Classifier. 2010 International Conference on Artificial Intelligence and Computational Intelligence. IEE Computer Society. DOI: https://doi.org/ 10.1109/AICI.2010.82