Uso de un enfoque de aprendizaje automático para predecir éxito en el tratamiento de adicciones

Daniela Prina[1], Sabrina López[1], Stephan Arndt[2], and Laura Acion[1*]

[1] Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina
[2] Department of Psychiatry, University of Iowa, Iowa, Estados Unidos
[*] lacion@gmail.com

Resumen. Hay varios métodos para construir modelos de predicción. La riqueza de las técnicas de modelado actualmente disponibles generalmente obliga al investigador a juzgar, a priori, cuál será mejor. Super Learner (SL, también denominado "stacked ensemble") es una metodología que facilita esta decisión al combinar todos los algoritmos de predicción identificados pertinentes para un problema en particular. Bajo ciertos supuestos, SL genera un modelo final que es al menos tan bueno como cualquiera de los otros modelos considerados para predecir el resultado. El objetivo de este trabajo es mostrar el uso de SL según su implementación en H2O a través de su interfaz con R. Este trabajo compara resultados de: regresión logística, regresión penalizada, random forest, redes neuronales y SL para predecir el tratamiento exitoso de adicciones. Se usó una base de datos estadounidense que incluyó 99.013 pacientes con tratamiento para adicciones. Todos los algoritmos se evaluaron usando el área bajo la curva ROC en una muestra de prueba distinta a la muestra de entrenamiento. SL fue superior a todos menos uno de los algoritmos comparados.

1 Introducción

Existen varios métodos para construir modelos de predicción. Los modelos de predicción a menudo se generan mediante algún tipo de regresión lineal o logística. Más recientemente, otros algoritmos de aprendizaje como random forests (RF) o redes neuronales se están usando para la predicción en ciencias de la salud. Estas técnicas más nuevas podrían mejorar la predicción y, en consecuencia, las posibilidades de encontrar el tratamiento más efectivo para cada paciente.

La riqueza de las técnicas de modelado actualmente disponibles suele obligar a juzgar, a priori, cuál será el mejor método de predicción. Super Learner (SL) [1] es una metodología que facilita esta decisión al combinar todos los algoritmos de predicción identificados pertinentes para un problema en particular. Bajo ciertos supuestos, SL genera un modelo final que es al menos tan bueno como cualquiera de los otros modelos considerados para predecir el resultado. Esta propiedad de SL es teórica [1] y está sustentada empíricamente [2]. El objetivo de este trabajo es mostrar el uso de SL según su implementación en H2O a través de su interfaz para R [3] (para más detalles de este trabajo ver [4]).

2 Métodos

Datos. Para ilustrar diferentes enfoques analíticos, nos centramos en el tratamiento ambulatorio de adicciones para personas hispanas adultas. Se usaron datos administrativos públicos, habitualmente usados por investigadores de adicciones. Estos datos permiten ilustrar el uso de las metodologías dentro de un entorno realista. *Outcome.* Como es habitual en este tema, el tratamiento completado se consideró un éxito, las demás razones de alta terapéutica (por ej, "en contra del consejo profesional", "persona encarcelada") se consideraron como indicadores de tratamientos exitosos. *Predictores.* Se incluyeron 28 predictores en el análisis, incluyendo características de las personas, características de los tratamientos, tipo de adicción y problemas coexistentes de salud mental. *Modelado.* Se ajustaron

modelos de regresión logística, regresión penalizada (por ej, LASSO y ridge), random forest, redes neuronales de aprendizaje profundo y SL. Se usó R [5] y la interfaz con R de H2O [3] que optimiza estos métodos para bases de datos grandes. Todos los algoritmos se evaluaron usando el área bajo la curva ROC (AUC).

3 Resultados

SL mostró la mayor AUC. El rendimiento de SL fue seguido muy de cerca por RF. El algoritmo con el peor desempeño fue la regresión logística. La mejora relativa en el AUC de SL fue menor al 5% en comparación con el peor método de predicción. Además, la AUC para SL tuvo la varianza estimada más pequeña. El resto de los modelos considerados tuvieron varianzas estimadas para la AUC de hasta 20% más altas que SL.

Todos los modelos de regresión paramétrica, tanto los enfoques penalizados como los no penalizados, se comportaron de manera casi idéntica con respecto a AUC. Para los modelos que incluyeron todos los predictores e interacciones de 2 vías, LASSO superó a los otros tres modelos de regresión. Las AUC para LASSO fueron mayores que para las regresiones de ridge y las logísticas.

4 Conclusiones

Este trabajo comparó varios modelos para predecir el éxito del tratamiento de adicciones. Como era de esperar, SL mostró el mejor rendimiento predictivo. En este caso la superioridad de SL fue magra. Sin embargo, este resultado no era en absoluto evidente antes de realizar el análisis. Creemos que la falta de diferencias importantes en las soluciones de todos los métodos es un resultado relevante. La falta de diferencia sustantiva entre SL y el resto de los modelos usados significa que: a) cualquiera de estos métodos podría usarse para estos datos y b) no hay problemas importantes en los supuestos de los diferentes modelos utilizados. Esto nunca es evidente antes de analizar un conjunto de datos si se usa un único enfoque analítico, lo más habitual en ciencias de la salud. En este sentido, SL sirve como una herramienta para agilizar y mejorar el análisis de sensibilidad para predicción. Además, incluso pequeñas mejoras en la predicción pueden tener un alto impacto dependiendo de cada problema en particular. En este caso, una pequeña mejora en la predicción podría afectar significativamente la salud de los pacientes y los costos de tratamiento. En otros casos, pequeñas mejoras de predicción podrían salvar vidas.

Referencias

1. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Statistical applications in genetics and molecular biology. 2007;6(1).
2. Polley EC, Rose S, van der Laan MJ. Super learning. Targeted Learning: Springer; 2011. p. 43–66.
3. LeDell E, Gill N, Aiello S, Fu F, Candel A, Click C, Kraljevic T, Nykodym T, Aboyoun P, Kurka M, Malohlava M. h2o: R Interface for 'H2O'. R package version 3.22.1.1. 2019.
4. Acion L, Kelmansky D, van der Laan M, Sahker E, Jones D, Arndt S. Use of a machine learning framework to predict substance use disorder treatment success. PLoS One. 2017;12(4), e0175383.
5. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018.