Aprendizaje automático aplicado en área de la salud: Parte 2

Nicolás H. Quiroz[1], María Lourdes Posadas-Martínez[1,2,3], Emiliano Rossi[3,4], Diego H. Giunta[1,2] y Marcelo R. Risk[1]

1. Instituto de Medicina Traslacional e Ingeniería Biomédica (IMTIB), CONICET - Instituto Universitario del Hospital Italiano de Buenos Aires - Hospital Italiano de Buenos Aires. Buenos Aires, Argentina.
2. Área de Investigación en Medicina Interna, Servicio de Clínica Médica. Buenos Aires, Argentina.
3. Departamento de Investigación. Buenos Aires, Argentina.
4. Servicio de Cardiología (E.R.). Hospital Italiano de Buenos Aires. Buenos Aires, Argentina.

RESUMEN

En el artículo anterior se introdujo el tema y se desarrolló cómo es la recolección y análisis de datos, la selección y entrenamiento de modelos de aprendizaje automático supervisados y los métodos de validación interna que permiten corroborar si el modelo arroja resultados similares a los de otros conjuntos de entrenamiento y de prueba. En este artículo continuaremos con la descripción de la evaluación del rendimiento, la selección del modelo más adecuado para identificar la característica que se va a evaluar y la validación externa del modelo. Además, el artículo resume los desafíos existentes en la implementación del *Machine Learning* desde la investigación al uso clínico.

Palabras clave: aprendizaje automático, cuidado de la salud, historia clínica electrónica, desarrollo, investigación, colaboración transdisciplinaria.

Machine Learning In Healthcare. Part 2

ABSTRACT

In the previous article, we introduced topics such as data collection and analysis, selection and training of supervised machine learning models and methods of internal validation that allow to corroborate whether the model yields similar results to other training and test sets.

In this article, we will continue with the description of the performance evaluation, selecting the most appropriate model to identify the characteristic to evaluate and the external validation of the model. In addition, the article summarizes the actual challenges in the implementation of machine learning from research to clinical use.

Key words: machine learning, healthcare, electronic health record, development, research, transdisciplinary collaboration.

INTRODUCCIÓN

El aprendizaje automático (*Machine Learning*) engloba al conjunto de herramientas informáticas, matemáticas y estadísticas cuyo fin es encontrar patrones relevantes en los datos para extraer información, clasificar o realizar predicciones[1,2]. Sea cual fuere el campo de aplicación del *Machine Learning*, la viabilidad

Autor para correspondencia: nicolas.quiroz@hospitalitaliano.org.ar, Quiroz NH.

Recibido: 17/06/21 Aceptado: 7/03/22 En línea: 31/03/22

DOI: http://doi.org/10.51987/revhospitalbaires.v42i1.152

Cómo citar: Quiroz NH, Posadas-Martínez ML, Rossi E, Giunta DH y Risk MR. Aprendizaje automático aplicado en área de la salud: Parte 2 Rev. Hosp. Ital. B.Aires. 2022;42(1):56-58.

de su uso depende fundamentalmente de la calidad en el proceso de recopilación de los datos y de la fiabilidad de estos, rigiéndose por el principio de GIGO (*garbage in*, *garbage out*). A su vez, esta necesidad de calidad y comprensión de los datos está relacionada con lo que se conoce como *curse of dimensionality*, principio que afirma que la *performance* del *Machine Learning* disminuye cuando la relación entre la cantidad de instancias de los datos y la cantidad de características (dimensiones) no es suficiente. Para ejemplificar este principio, una base de datos de seguimiento de mil pacientes puede parecer un número más que suficiente para aplicar un modelo de predicción del curso de una patología pero, si para cada paciente se tienen más de 10 mil características, la *performance* de cualquier modelo puede ser baja debido a la alta redundancia en los datos y a la introducción de características poco representadas o con alta variabilidad en sus mediciones[3]. Por este motivo es necesario comprender de antemano qué reflejan las variables medidas[4] y definir cuáles son las variables que mejor describen nuestro conjunto de datos, recordando que –sobre la base de estas– el algoritmo realizará la tarea de sugerir diagnósticos y tratamientos y que –si no se hace de manera adecuada– podrían existir sesgos[4,5] y la sugerencia sería incorrecta[6].

En el artículo anterior se introdujo el tema y se desarrolló cómo es la recolección y análisis de datos, la selección y entrenamiento de modelos de aprendizaje automático supervisados y los métodos de validación interna que permiten corroborar si el modelo arroja resultados similares a los de otros conjuntos de entrenamiento y de prueba. En este artículo continuaremos con la descripción de la evaluación del rendimiento, la selección del modelo más adecuado para identificar la característica que se va a evaluar y la validación externa del modelo. Además, el artículo resume los desafíos existentes en la implementación del *Machine Learning* desde la investigación hasta el uso clínico.

MÉTRICAS DE EVALUACIÓN PARA CLASIFICACIÓN

Existen múltiples métricas para evaluar y comparar los diferentes modelos entrenados con los mismos datos. Entre ellos, por lo común de su utilización, podemos mencionar: la matriz de confusión, precisión, exhaustividad, valor-F, exactitud, área bajo la curva *F1-score* y el área bajo la curva ROC.

Matriz de confusión (*matrix confusion*): esta matriz es una de las métricas más intuitivas y descriptivas para conocer el rendimiento de modelos de clasificación (binaria o multiclase) comparando las clases predichas por el algoritmo con las reales. A partir de esta matriz se obtienen, entre otras, las siguientes métricas.

Precisión (*precision*): corresponde a la proporción de individuos clasificados dentro de una clase que normalmente se llama positiva, aunque debe ser definido de antemano que realmente pertenecen a esta condición de positividad.

Exhaustividad (*recall*): mide, sobre todas las observaciones que se conocen positivas, cuántas de ellas han sido predichas como tales por el algoritmo.

Valor-F (*F1-score*): es una métrica que combina por igual las métricas de precisión y exhaustividad para calcular el rendimiento del modelo.

Exactitud (*accuracy*): es la fracción de predicciones que el modelo realizó correctamente, tanto para la clase positiva y negativa para la clasificación binaria como para cualquier clase en caso de clasificación múltiple. Es una de las métricas más utilizadas con el fin de evaluar el desempeño de los modelos pero, para que su valor posea importancia real, debe corroborarse previamente que no exista un marcado desequilibrio de clases en los datos. En caso contrario, se incurre en lo que se conoce como la paradoja de exactitud (*accuracy paradox*), obteniéndose altos valores de esta métrica solo por detectar la clase más representada en el conjunto de datos.

Área bajo la curva (AUC, *area under the curve*): esta gráfica representa la proporción de verdaderos positivos (vertical) vs. falsos positivos (en horizontal) para diferentes umbrales de clasificación que determinan si una muestra es de una u otra clase. El área bajo la curva ROC refleja cuán bueno es un modelo para discriminar clases positivas y negativas, donde un área de 1 corresponde a un clasificador perfecto y un área de 0,5 corresponde a un clasificador que asigna aleatoriamente las muestras a una u otra clase.

SELECCIÓN DEL MODELO MÁS ADECUADO

A grandes rasgos, la selección del modelo de clasificación más adecuado se basa en identificar aquel que tenga mayor relevancia clínica, menor costo computacional, mayor sencillez para su aplicación y obtención de las mejores métricas de evaluación[7-9]. Como se explicó en el artículo anterior, la validación interna del modelo se realiza mediante técnicas como *bootstrapping* o *cross validation*. De esta forma, un mismo modelo puede desempeñarse de manera diferente en cada subconjunto con el que fue entrenado y validado, por lo que es posible obtener una *performance* promedio y sus respectivos desvíos. En el caso de que varios modelos obtengan una *performance* similar, pueden realizarse tests estadísticos para probar si un modelo es efectivamente mejor que otro o si las diferencias se deben al azar. No obstante, pueden existir otros criterios además de la *performance* para elegir un modelo. Por ejemplo, puede elegirse el modelo que resulte más simple para aplicar o aquel con menor costo computacional y mejor rendimiento en términos de velocidad o escalabilidad en los datos, en caso de que se busque su aplicación *online*. Otro criterio no menor es elegir un modelo que posea alta interpretabilidad, esto es, la capacidad para explicar –al profesional que lo utilizará– los criterios para la elección de una u otra clase en el proceso de clasificación. Muchas veces modelos poco interpretables (también conocidos como modelos de "caja negra") pueden tener una *performance* ligeramente superior, pero no es posible identificar

los patrones en los datos que derivan en ella y, por ende, no es posible identificar fácilmente en qué casos pueden estar cometiéndose errores debidos a sesgos en los datos, errores espurios, etc.[10].

VALIDACIÓN EXTERNA

Consiste en probar el modelo en una nueva muestra de pacientes hasta el momento desconocida y realizar las métricas de evaluación. Por ejemplo, este conjunto de pacientes puede proceder de diferentes centros de salud, diferentes zonas geográficas o períodos de tiempo. La importancia del proceso de validación externa radica en el sesgo que suelen interiorizar los modelos basados en *Machine Learning* para clasificar pacientes. Estos sesgos provienen de aprender características que son particulares del centro de salud donde se entrenaron, como por ejemplo los datos que se eligieron (criterios de inclusión y exclusión), la forma de llenado de las historias clínicas electrónicas (HCE) y las características demográficas de los pacientes. Por este motivo, es deseable que los modelos, para poder replicar una buena *performance* en escenarios diferentes de aquel de su entrenamiento sean producto de esfuerzos colaborativos multicéntricos dentro de un país o con un enfoque internacional[11].

CONCLUSIÓN

Como se ha explicado en esta serie de artículos, el *Machine Learning* en salud puede impactar en las estrategias de anticipación de eventos y en la construcción de una medicina personalizada con la potencialidad de elevar la calidad de vida del paciente. Sin embargo, para llegar a este objetivo se deberán superar distintas etapas y cuestiones que están asociadas a la implementación de las nuevas tecnologías[12].

En la actualidad, la eficacia de los modelos se ha estudiado en su mayoría para modelos retrospectivos y son pocos los ensayos prospectivos en la aplicación del *Machine Learning*. En muchos casos, los modelos resultantes tienen una *performance* superior a la de otras técnicas estadísticas tradicionales, pero no resulta sencillo entender cómo el algoritmo capta las relaciones y brinda una respuesta. Adicionalmente, desplegar un sistema de estas características insume tiempo, recursos profesionales y dinero, lo que produce en muchos casos fuertes expectativas que deben ser cumplidas para lograr un retorno de la inversión y la adopción del modelo sin generar resistencias en los usuarios[13,14]. Además, los datos deben estar actualizados, respaldados y protegidos para asegurar cumplir pautas éticas y de ciberseguridad[15].

Nos espera un futuro prometedor con nuevas tecnologías en desarrollo, aunque deberemos superar múltiples instancias previas para alcanzar el éxito. Es por ello que deberemos crear grupos multidisciplinarios capaces de trabajar coordinadamente. Las computadoras no están para desplazar al equipo de salud, sino para darle soporte.

Conflictos de interés: los autores declaran no tener conflictos de interés.

REFERENCIAS

1. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis. 2018;66(1):149-153. https://doi.org/10.1093/cid/cix731.
2. Park C, Took CC, Seong JK. Machine learning in biomedical engineering. Biomed Eng Lett. 2018;8(1):1-3. https://doi.org/10.1007/s13534-018-0058-3.
3. Barbour DL. Precision medicine and the cursed dimensions. NPJ Digit Med. 2019;2:4. https://doi.org/10.1038/s41746-019-0081-5.
4. Clarke R, Ressom HW, Wang A, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer. 2008;8(1):37-49. https://doi.org/10.1038/nrc2294.
5. Finlayson SG, Bowers JD, Ito J, et al. Adversarial attacks on medical machine learning. Science. 2019;363(6433):1287-1289. https://doi.org/10.1126/science.aaw4399.
6. Miotto R, Wang F, Wang S, et al. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236-1246. https://doi.org/10.1093/bib/bbx044.
7. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574. https://doi.org/10.1177/0272989X06295361.
8. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019;19(1):64. https://doi.org/10.1186/s12874-019-0681-4.
9. Uddin S, Khan A, Hossain ME, et al. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019;19(1):281. https://doi.org/10.1186/s12911-019-1004-8.
10. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206-215. https://doi.org/10.1038/s42256-019-0048-x.
11. van Soest JPA, Dekker ALA, Roelofs E, et al.. Application of machine learning for multicenter learning. En: El Naqa I, Ruijiang L, Murphy LJ, eds. Machine learning in radiation oncology. Cham: Springer; 2015. p. 71-97. https://doi.org/10.1007/978-3-319-18305-3_6.
12. Obermeyer Z, Weinstein JN. Adoption of artificial intelligence and machine learning Is increasing, but irrational exuberance remains. NEJM Catalyst. 2020;1(1). https://doi.org/10.1056/CAT.19.1090.
13. Sendak M, Gao M, Nichols M, et al. Machine learning in health care: a critical appraisal of challenges and opportunities. EGEMS (Wash DC). 2019;7(1):1. https://doi.org/10.5334/egems.287.
14. Silcox C, Dentzer S, Bates DW. AI-enabled clinical decision support software: a "trust and value checklist" for clinicians. NEJM Catalyst. 2020;1(6). https://doi.org/10.1056/CAT.20.0212.
15. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. https://doi.org/10.1038/s41591-018-0300-7.