Universidad Tecnológica Nacional
ABRIL 2021 / Año 19- Nº 40

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

Clasificadores de Aprendizaje Automático aplicados a Datos RKI MALDI-TOF de Espectrometría de Masa de Bacterias

Machine Learning Classifiers applied to RKI MALDI-TOF Mass Spectrometry of Bacteria Data

Presentación: 02/02/2021

Aprobación: 29/03/2021

Andrea Rey
Centro de Procesamiento de Señales e Imágenes y Departamento de Ciencias Básicas Matemática, Facultad Regional Buenos Aires, Universidad Tecnológica Nacional - Argentina
arey@frba.utn.edu.ar

Resumen

En este trabajo evaluamos el rendimiento de diferentes métodos de aprendizaje automático, para la clasificación de bacterias a partir de espectros de masas disponibles en la base de datos RKI MALDI-TOF. La identificación de microorganismos empleando espectrometría de masas se ha convertido en una tecnología muy popular en los últimos años, especialmente en microbiología clínica, donde una clasificación adecuada es fundamental al momento de elegir un tratamiento correcto. Los modelos elegidos en nuestro estudio incluyen análisis por discriminantes, árboles de decisión, vecinos más cercanos y redes neuronales. Consideramos las siguientes medidas para el análisis: exactitud, coeficiente κ de Cohen, tasa de no información y tiempo consumido. Los resultados obtenidos nos permiten recomendar el análisis por discriminante lineal, con un desempeño levemente inferior a vecinos más cercanos, pero aventajándolo en términos de costo computacional.

Palabras claves: espectros de masas, clasificación, aprendizaje automático, RKI MALDI-TOF/MS.

Abstract

In this work we assess the performance of different machine learning methods, to classify bacteria from mass spectra available at RKI MALDI-TOF database. Microorganism identification employing mass spectrometry is a technology that has become very popular in the last years, especially in clinical microbiology, where an adequate classification is relevant to choose a proper treatment. The techniques selected in our study include discriminant analysis, decision trees, nearest neighbors and neuronal networks. We consider the following

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

measures for the analysis: accuracy, Cohen κ coefficient, no information rate and time consumed. The obtained results allow us to recommend linear discriminant analysis, with a slightly lower performance than nearest neighbors, but with advantages in computational cost.

Keywords: mass spectra, classification, machine learning, RKI MALDI-TOF/MS.

Introducción

Uno de los principios en microbiología consiste en la identificación de microorganismos, comenzando el rango taxonómico por familia, género y especie. En la familia de bacterias existen patógenos causantes de enfermedades infecciosas, por lo que la nomenclatura bacteriana es fundamental para encontrar relaciones entre la enfermedad del parásito y del huésped, diseñar planes terapéuticos y la investigación epidemiológica. Dentro de las técnicas de reconocimiento de microorganismos podemos mencionar pruebas morfológicas y bioquímicas o el enfoque tradicional que incluye métodos como microscopía, serología y cultivo.

Las secuencias de ARN ribosómico 16S aplicadas como cronómetro molecular (Olsen y Woese, 1993) tienen el propósito de agilizar una precisa clasificación e identificación de bacterias, aplicándose en microbiología clínica cuando otra técnica resulta imposible, difícil o muy costosa. Debido a que el gen del ARNr 16S consta de muchas regiones variables y conservadas, se crean los cebadores universales (Clarridge, 2004) para la comparación taxonómica. Estos cebadores se usan en el método de reacción en cadena de la polimerasa, conocido como PCR (Mullis, 1990), cuyos ensayos son extremadamente útiles para detectar enfermedades hereditarias e infecciosas. El costo y las limitaciones de su sensibilidad y éxito, ocasionadas por los componentes de la solución de extracción de ADN, la concentración de las mezclas de PCR, la temperatura y las condiciones de ciclo, son algunas desventajas (Rossen et al., 1992; Wilson, 1997).

Si bien se han reportado varios casos de identificación errónea de bacterias (Janda y Abbott, 2002; Hamilton-Miller, 1993; Hamilton-Miller y Shah, 1996), los sistemas comerciales y los laboratorios no tienen dificultad para identificar las especies de bacterias más comunes presentes en los agentes infecciosos de importancia médica y ofrecen un resultado rentable esperado. El objetivo de la fusión de prácticas de laboratorio y farmacia en un programa administrativo, es minimizar el contacto del paciente con antibióticos y los costos de estadías prolongadas en el hospital. Los laboratorios clínicos tienen la presión de producir resultados de ensayos rápidos, precisos y de calidad para mejorar la atención del paciente. Mientras que muchos métodos tradicionales basados en cultivos tardan días en producir resultados, la espectrometría de masas (MS del inglés mass spectrometry) proporciona información de diagnóstico en cuestión de horas, siendo su mayor ventaja la baja complejidad técnica del flujo de trabajo. En particular, la tecnología Matrix-Assisted Laser Desorption/Ionization Time Of Flight MS (MALDI-TOF/MS) utiliza pesos moleculares de proteínas y péptidos para identificar microorganismos. Los microbios se depositan en una misma placa y se mezclan con una matriz de ionización (Maier y Kostrzewa, 2007; Cherkaoui et al., 2010). Luego, todos los espectros de masas de proteínas se comparan con una base de datos de espectros recopilada de organismos conocidos, lo que permite que el programa identifique las bacterias según la mejor coincidencia. Este método distingue los microorganismos

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

mediante una única prueba, ya que no necesita cebadores específicos, reduciendo el número de procedimientos y reactivos, así como el conocimiento especializado en microbiología. Resulta asequible y fácil de ejecutar, incluso con poco entrenamiento, aunque se requiere cierto número de organismos para la identificación.

Mientras que la secuenciación y los métodos convencionales de pruebas bioquímicas y determinación del fenotipo aplicados a la identificación de patógenos son costosos en dinero y tiempo (Zhu et al., 2015), MALDI-TOF/MS surge como una alternativa rápida, robusta y más económica. Trabajos recientes han demostrado el interés en el empleo de este método (Lasch et al., 2016; Avanzi et al., 2017; Manukumar y Umesha, 2017; Desaire y Hua, 2019). En los últimos años, MALDI-TOF/MS se ha convertido en la herramienta de identificación de primera línea de microorganismos.

En el contexto de aprendizaje automático, un problema de clasificación consiste en un enfoque de aprendizaje supervisado donde el algoritmo aprende de un conjunto de datos etiquetados, y usa este aprendizaje para clasificar nuevas observaciones. Esto diría que los nuevos datos sólo pueden clasificarse cuando se tiene un "conocimiento" sobre observaciones previas. Sin embargo, cuando entrenamos un clasificador supervisado, estamos entendiendo la estructura de los datos al tratar de comprender los atributos que los diferencian y sirven para estiman los valores de umbral que determinan las clases. Entonces, si estimamos estos umbrales de manera no supervisada, conocemos las cualidades compartidas de una categoría particular y, por lo tanto, podríamos clasificar nuevas observaciones. Es decir, estaríamos resolviendo un problema de clasificación mediante aprendizaje no supervisado.

Los espectros de la base de datos que utilizamos son representativos, seleccionados específicamente de una gran cantidad de datos, intensos y limpios. En experimentos reales, una mala extracción, la cristalización desigual o las interferencias e impurezas, afectarán la calidad de los espectros. Este marco tiene una gran influencia al seleccionar un modelo ya que el mismo depende fuertemente de los datos. Este trabajo pretende ahorrar a los investigadores esfuerzos en la elección del mejor algoritmo de clasificación, cuando los datos a analizar poseen una estructura similar a la base de espectros de masas RKI MALDI-TOF. Nuestro principal objetivo es comparar el rendimiento de técnicas de aprendizaje automático, aplicadas a una estructura de datos especial dada por la tecnología MALDI-TOF/MS. En particular, nos enfocamos en técnicas supervisadas como discriminantes y árboles de decisión, técnicas no supervisadas como vecinos más cercanos, y redes neuronales que pueden ser tanto supervisadas como no supervisadas.

Metodología

MODELOS DE CLASIFICACIÓN

Un modelo de clasificación asigna un vector de entrada a alguna de las clases disjuntas que se han establecido, mediante una división del espacio de entrada en regiones de decisión. Los algoritmos de aprendizaje automático construyen un modelo matemático basado en datos de entrenamiento, con el fin de predecir o decidir. En general, el conjunto de datos se divide en un porcentaje mayor de observaciones utilizadas para el entrenamiento y usando el resto para la validación. A continuación describimos brevemente los métodos utilizados y nos referimos a Friedman et al. (2001) y Bishop (2006) para más detalles.

Análisis por discriminante lineal y diagonal. El análisis por discriminante lineal es una generalización del discriminante lineal de Fisher, que encuentra una combinación lineal de cualidades para caracterizar o separar los datos en clases disjuntas. Una variable categórica

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

se explica con los valores de las variables independientes, bajo el supuesto de normalidad. Sea C_k una de las K clases, con n_k observaciones independientes de dimensión p, siguiendo una distribución normal multivariada con vector medio μ_k y matriz de covarianza Σ_k. Sea π_k la probabilidad a priori de observar un elemento en la clase, tal que $\pi_1+\cdots+\pi_K=1$. Decidimos $x=(x_1,\ldots,x_p)^T \in C_k$ si minimiza la siguiente función discriminante:

$$D_k(x) = (x-\mu_k)^T \sum_k^{-1}(x-\mu_k) + \log\left|\sum_k\right| - 2\log(\pi_k), \tag{1}$$

donde el supra índice T indica la transposición y $|\ |$ el determinante matricial. Si todas las matrices de covarianza son iguales a Σ, la Ec. (1) resulta:

$$d_k(x) = (x-\mu_k)^T \sum^{-1}(x-\mu_k) - 2\log(\pi_k), \tag{2}$$

dando lugar al análisis por discriminante lineal (ADL). Cuando las matrices de covarianza son diagonales, la regla discriminante está dada por contribuciones cuadráticas aditivas de cada clase, originando el análisis por discriminante cuadrático diagonal. En particular, si las densidades de clase comparten la misma matriz de covarianza diagonal, el método se conoce como análisis por discriminante diagonal (ADD).

Análisis por discriminante usando predictores binarios. Sea $X=(X_1,\cdots,X_p)$ una variable aleatoria binaria multivariada con distribución Bernoulli multivariada y vector de esperanzas $\mu=(\mu_1,\cdots,\mu_p)$. La función de probabilidad conjunta con variables predictoras independientes es:

$$P(x) = \prod_{j=1}^{p} \begin{cases} 1-\mu_j & \textbf{si } x_j = 0, \\ \mu_j & \textbf{si } x_j = 1, \end{cases} \tag{3}$$

con matriz de covarianza diagonal dada por $Var(X)=(\mu_1(1-\mu_1),\cdots,\mu_p(1-\mu_p))$. Gibb y Strimmer (2015) emplean una regla de predicción bayesiana, llamada análisis por discriminante binario (ADB), y definen para cada clase C_k,

$$P(x|C_k) = \prod_{j=1}^{p} \begin{cases} 1-\mu_{C_k} & \textbf{si } x_j = 0, \\ \mu_{C_k} & \textbf{si } x_j = 1, \end{cases} \tag{4}$$

con probabilidad a priori π_k de observar un elemento en la clase C_k. La probabilidad a posteriori se deduce aplicando el teorema de Bayes, obteniéndose la función discriminante:

$$d_k(x) = \log P(C_k|x) = \log \pi_k + \log P(x|C_k) + C, \tag{5}$$

donde la constante C representa todos los términos que no dependen de C_k, ya que sólo se compara entre diferentes clases. La predicción de clase para una nueva observación se realiza eligiendo el grupo C_k que maximiza la Ec. (5).

Árbol de decisión. Produce una secuencia de reglas para clasificar datos, a partir de un

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

conjunto de atributos etiquetados. El algoritmo considera todas las características y realiza una división binaria para que los datos categóricos elijan el predictor con la mayor precisión, repitiendo este procedimiento recursivamente, hasta que los datos se dividen con éxito en todas las hojas posibles o se alcanza la profundidad máxima establecida. Si bien esta técnica es sencilla y posee ventajas de visualización, los árboles pueden ser muy complejos o inestables ante una pequeña variación de los datos. Dado un nodo *m*, representando la región R_m con N_m observaciones, la proporción de observaciones de la clase C_k en tal nodo es:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum x_i \in R_m I(y_i = k). \tag{6}$$

Los elementos en el nodo m se clasifican en la clase C_k si ésta es mayoritaria; es decir, maximizando la Ec. (6).

Vecinos más cercanos. Más conocido como *k*NN del inglés *nearest neighbors*, clasifica una observación por el voto mayoritario de sus vecinos en el espacio de los parámetros de entrada. Es un método no paramétrico, de fácil implementación, robusto para datos de entrenamiento ruidosos, y efectivo para conjuntos grandes de entrenamiento. Por el contrario, se requiere fijar la cantidad de clases y el costo computacional es alto, ya que la distancia se calcula entre cada instancia y todas las muestras de entrenamiento. La predicción es:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i, \tag{7}$$

donde N_k (x) es la vecindad de x definida como los k puntos más cercanos x_i en la muestra de entrenamiento. La cercanía implica la elección de una métrica. En otras palabras, *k*NN promedia las respuestas de las k observaciones más cercanas en el espacio de entrada.

Red neuronal. El término se debe a que su estructura computacional copia la forma en que funcionan las neuronas en un sistema biológico, a través de representaciones matemáticas del procesamiento de información. La red utiliza ejemplos de datos para inferir automáticamente reglas de reconocimiento de patrones realizando varios pasos de procesamiento, devolviendo una salida. Si el número de muestras de entrenamiento aumenta, la red puede aprender más sobre la configuración mejorando su precisión. De este modo, la complejidad de las redes neuronales se basa en el número de pasos de procesamiento y las características de cada paso. Su representación gráfica incluye secciones verticales (capas), y puntos de cálculo interno (nodos). Rosenblatt (1957) desarrolló una neurona artificial llamada perceptrón, que produce una única salida binaria a partir de varias entradas también binarias, ponderando la importancia de la entrada en la regla. Si la facilidad para alcanzar la unidad está representada por el sesgo b, la regla del perceptrón se puede escribir como:

$$\textbf{salida} = \begin{cases} 0 \ \textbf{si} \ w \cdot x + b \leq 0, \\ 1 \ \textbf{si} \ w \cdot x + b > 0, \end{cases} \tag{8}$$

donde $w \cdot x$ es el producto escalar entre los vectores de peso y entrada. Dada la necesidad de un algoritmo de aprendizaje para ajustar automáticamente los pesos y sesgos de la red,

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

se introduce un nuevo tipo de neurona artificial denominada sigmoidea, desarrollando un conjunto de técnicas de aprendizaje profundo basadas en el descenso de gradiente estocástico y la retropropagación.

BASE DE DATOS Y MÉTODOS

Para nuestro estudio, utilizamos la Versión 3 de la base de datos de espectros de masas MALDI-TOF de RKI (Lasch et al., 2018), segunda actualización de la base de datos original. Aunque en el informe se menciona un total de 6264 espectros de masas de varias bacterias, en su descarga hallamos 6341, correspondientes a 43 géneros según se observa en la Fig. 1.

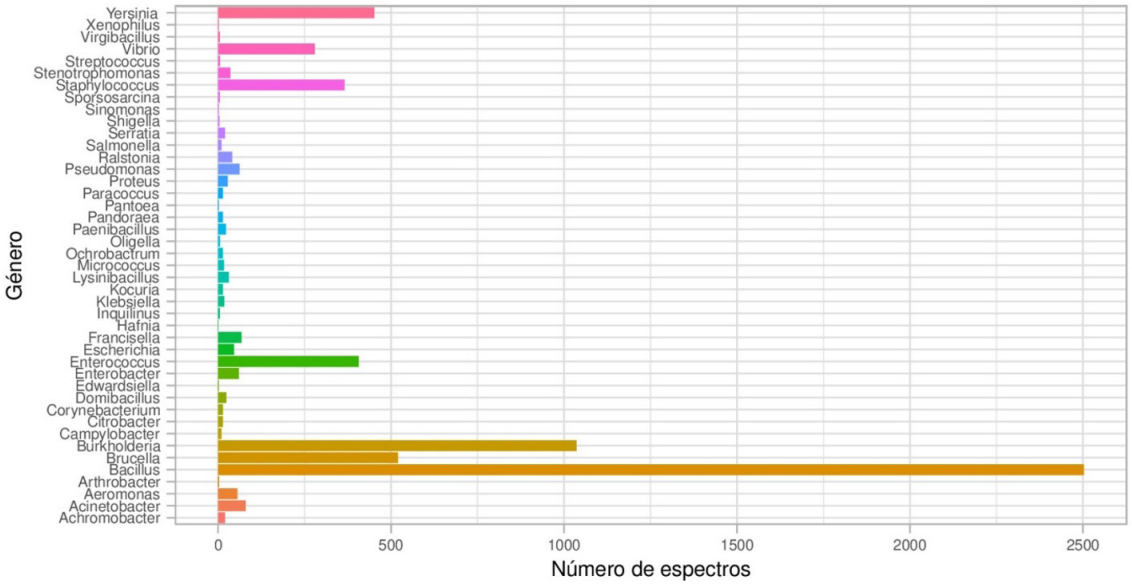

Fig. 1: Histograma de géneros en la base de datos.

El formato original de los espectros es el proporcionado por la plataforma estadounidense Bruker Daltonics, cuya importación se hizo con el paquete MALDIquantForeign del *software* libre R (R Development Core Team, 2013). Diferencias fundamentales entre especies, como por ejemplo, *Streptococcus pyogenes y Streptococcus agalactiae* causantes de enfermedades muy distintas, sugieren una clasificación por especie como más apropiada. Sin embargo, en una primera etapa decidimos comenzar con ensayos clasificando por género, con el fin de descartar algoritmos deficientes. Luego de aplicar los modelos a todos los datos, clasificando por género y especie, y considerando que la muestra original no está balanceada (ver Fig. 1), repetimos el proceso dividiendo por género al conjunto de datos de la siguiente manera: 31 grupos de tamaño 10, 15 grupos de tamaño 30 y 7 grupos de tamaño 100.

Para el preprocesamiento de los espectros de masas realizamos las siguientes tareas con el paquete MALDIquant de R. 1- Recorte de espectros a un rango común adecuado. 2- Transformación de la raíz cuadrada para evitar dependencia de la varianza de intensidad con la media. 3- Suavizado de las intensidades para eliminar ruido. 4- Aplicación del algoritmo SNIP (Ryan et al., 1988) para estimar y descartar la línea de base, causada por el ruido químico de los efectos de la matriz y la contaminación. 5- Normalización de la intensidad

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

en una calibración local. 6- Alineación de los valores de espectros de masas a un conjunto de referencia de picos obtenidos de una de las muestras, estableciendo 0.3 como valor de referencia para la frecuencia mínima de los picos totales. 7- Detección de picos para cada muestra, manteniendo sólo aquellos por encima de un umbral de ruido establecido en 2. 8- Cálculo de la matriz de intensidad, compuesta por valores de intensidad calibrados para picos identificados en todos los espectros y necesaria para el entrenamiento. En la Tabla 1 se muestra la cantidad de espectros y picos detectados en cada caso de estudio.

Criterio de clasificación	Tamaño muestral	Cantidad de clases	Cantidad de picos detectados
Por género	**6341**	**43**	**1128**
Por especie	**6341**	**171**	**1667**
Por especie, tamaño 10	**310**	**51**	**1155**
Por especie, tamaño 30	**150**	**33**	**968**
Por especie, tamaño 100	**700**	**19**	**766**

Tabla 1: Información de los datos en cada experimento.

Los modelos elegidos fueron entrenados usando los siguientes paquetes de R: sda para análisis por discriminante lineal y diagonal; binda para análisis por discriminante usando predictores binarios; party para árboles de decisión; class para vecinos más cercanos usando la distancia euclidiana; y nnet para redes neuronales.

Comenzamos el análisis identificando los picos discriminantes de clase más importantes utilizando t-scores. Agrupando rangos superiores a 10, 100 y 500, notamos una precisión aceptable de los modelos de clasificación ADL y ADD sólo para las 500 características de mayor rango, y ninguna diferencia significativa en el costo computacional al trabajar con esta restricción y con la matriz completa de características. En cuanto el algoritmo de vecinos más cercanos, variamos el valor de k, registrando menor precisión al aumentar dicho valor. La precisión de la red neuronal mejora a medida que crece su tamaño, consumiendo una innecesaria cantidad de tiempo extra. En consecuencia, seleccionamos los siguientes algoritmos para comparar su rendimiento: 1NN (vecinos más cercanos con k=1), AD (árbol de decisión), ADB, ADL, ADD, y RN (red neuronal con tres capas y un máximo de 500 iteraciones). La cantidad de neuronas en la capa oculta de la red fue 7 para la clasificación por género y 10 al clasificar por especie. El conjunto de datos se dividió aleatoriamente en un 70% para la etapa de entrenamiento y validando con el 30% restante.

Con el fin de evaluar el desempeño de los clasificadores, calculamos tres medidas. La exactitud, dada por el cociente entre el número de observaciones clasificadas correctamente y el número total de observaciones, mide la proporción de predicciones correctas entre todas las predicciones, la cual podría ser poco confiable en el caso de un conjunto de datos donde el número de observaciones varía mucho entre las diferentes clases. El coeficiente κ de Cohen, robusto en cuanto al buen desempeño del clasificador en comparación con un funcionamiento azaroso, está definido como el cociente entre la diferencia de proporciones de predicciones correctas y esperadas por azar, y el complemento de esta última proporción. La tasa de no información (TNI) mide la proporción de clases que se etiquetan correctamente cuando se asignan al azar, necesaria para saber si el modelo realmente está haciendo algo útil. Observemos que la TNI debe ser superada por la exactitud del modelo

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

para demostrar que el mismo produce resultados significativos.

Al analizar el comportamiento de una clase específica, comparamos la misma contra todas las clases restantes. Fijado un género, definimos un resultado como positivo si la observación pertenece a la clase dada por este género; de lo contrario, será negativo. Replicamos el entrenamiento de cada modelo cincuenta veces y calculamos la media de las siguientes medidas: sensibilidad (fracción de positivos predichos correctamente), especificidad: (fracción de negativos predichos correctamente), precisión (relación entre verdaderos positivos y número total de positivos predichos), y exactitud equilibrada (media entre la sensibilidad y la especificidad), adecuada en los casos en que las clases son muy diferentes en tamaño.

Al elegir un modelo de aprendizaje automático, debemos considerar no sólo su rendimiento, sino también su costo computacional en función del tiempo de entrenamiento y aplicación del algoritmo. Trabajamos con un procesador Intel (R) Core (TM) i7-6700 @ 3.40GHz 3.41 GHz, 16.0 GB RAM. Replicamos los procesos de entrenamiento y validación 50 veces para cada modelo seleccionado y registramos la suma de los tiempos de usuario y sistema.

Resultados

La Fig. 2 muestra los boxplots obtenidos para la exactitud, el coeficiente κ y la TNI de cada modelo de clasificación aplicado a todos los datos. Descartamos AD en la clasificación por especie debido a su bajo rendimiento. Observamos que en el resto de los métodos seleccionados, la clasificación por género es levemente más exacta que la clasificación por especie. Los resultados de las tres medidas cuando los modelos se usan en muestras balanceadas se presentan en la Fig. 3. En particular, ADD y ADB mostraron una mejora en las predicciones al trabajar con muestras equilibradas, independientemente del tamaño muestral. Los valores del coeficiente κ son similares a la exactitud del modelo en todos los casos de estudio presentados. De igual modo en todos los casos, la TNI es menor para la clasificación por especies que para la clasificación por género, disminuyendo directamente proporcional al tamaño de la muestra equilibrada.

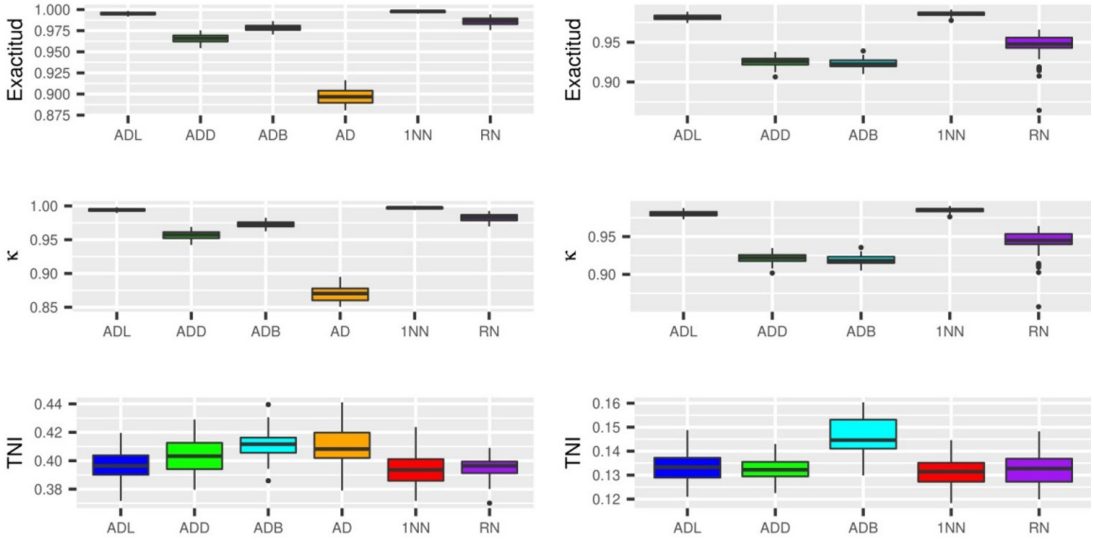

Fig. 2: Boxplots de exactitud, coeficienteκ y TNI, para la clasificación por género (izquierda) y por especie (derecha).

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

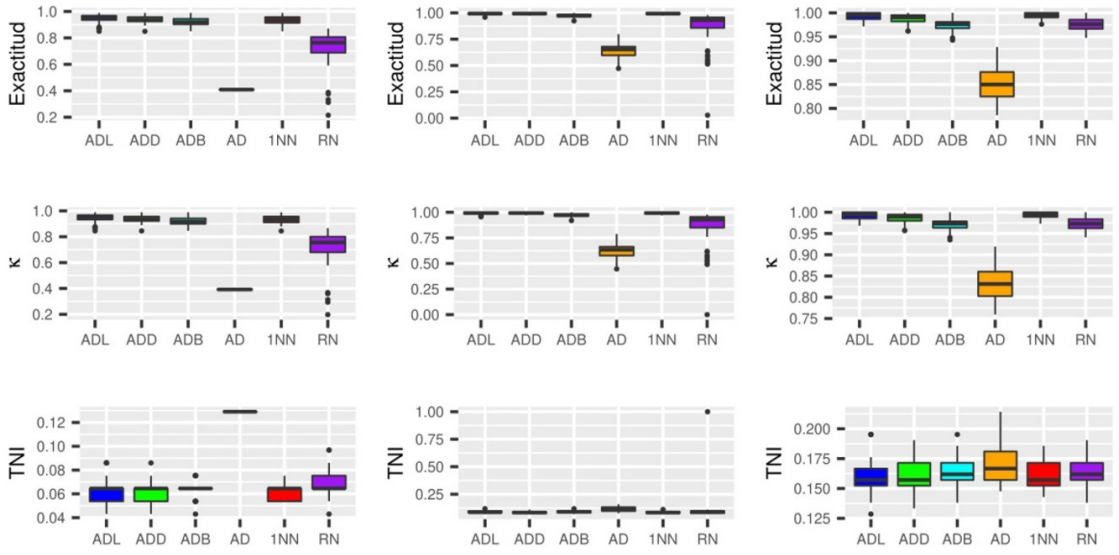

Fig. 3: *Boxplots* de exactitud, coeficienteκ y TNI, para la clasificación por especie de la muestra con clases balanceadas de tamaño 10 (izquierda), 30 (centro) y 100 (derecha).

Debido a las escasas muestras en algunos géneros de la base de datos, restringimos nuestro estudio por género a aquellas clases que pueden compararse en todos los modelos seleccionados, a excepción de RN. Tales géneros son: *Acinetobacter, Aeromonas, Bacillus, Brucella* y *Enterobacter*. Los resultados obtenidos se muestran en la Fig. 4, pudiendo observar que *Acinetobacter* logra la mejor sensibilidad, *Brucella* obtiene la mayor precisión, mientras que *Acinetobacter* y *Aeromonas* tienen la exactitud equilibrada más favorable. La especificidad es superior al 99% en todos los casos, excepto *Bacillus* para todos los modelos (siendo 69 el porcentaje más bajo para ADB). Dado que la especificidad indica la probabilidad de verdaderos negativos, un valor inferior a 0.95 implica que 5 de cada 100 pacientes infectados no tendrá un diagnóstico correcto, privándolos de un tratamiento y pudiendo provocar su muerte. Los valores de especificidad de *Bacillus* son así inadmisibles, especialmente aplicando ADB. Las bacterias que alcanzan las mejores puntuaciones son *Brucella* y *Acinetobacter*. En general, la precisión es baja. En relación con los modelos, podemos apreciar que ADL, AD y 1NN poseen los mejores desempeños para *Enterobacter, Brucella* y *Acinetobacter*, respectivamente. El uso de ADB es inaceptable para *Bacillus*.

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

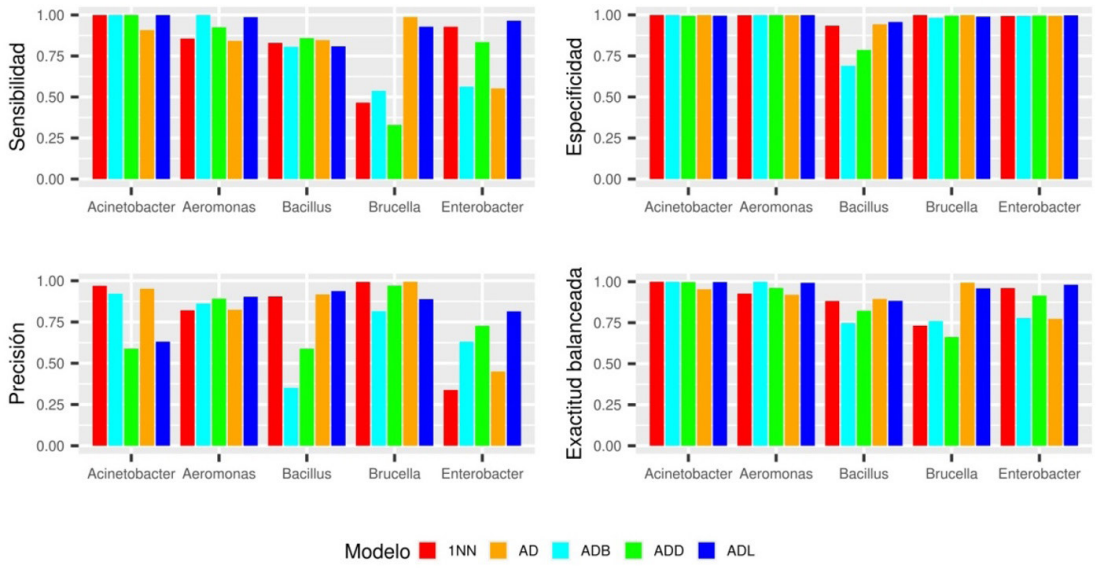

Fig. 4: Sensibilidad, especificidad, precisión y exactitud balanceada de los clasificadores entre algunos géneros.

El costo del preprocesamiento de espectros, inherente a la metodología empleada, impacta en todos los clasificadores. Si bien los cálculos necesarios para procesar un solo espectro no son particularmente costosos, al necesitar muchas muestras para alcanzar una alta precisión, el costo computacional se ve extremadamente afectado por la preparación de los datos. La suma de los tiempos de usuario y sistema consumidos por el proceso de datos, la detección de picos y el cálculo de la matriz de características fue de 205.44 segundos. El tiempo de entrenamiento puede variar ampliamente según el algoritmo. Los costos computacionales, en términos de tiempo de usuario y sistema, se muestran en la Fig. 5.

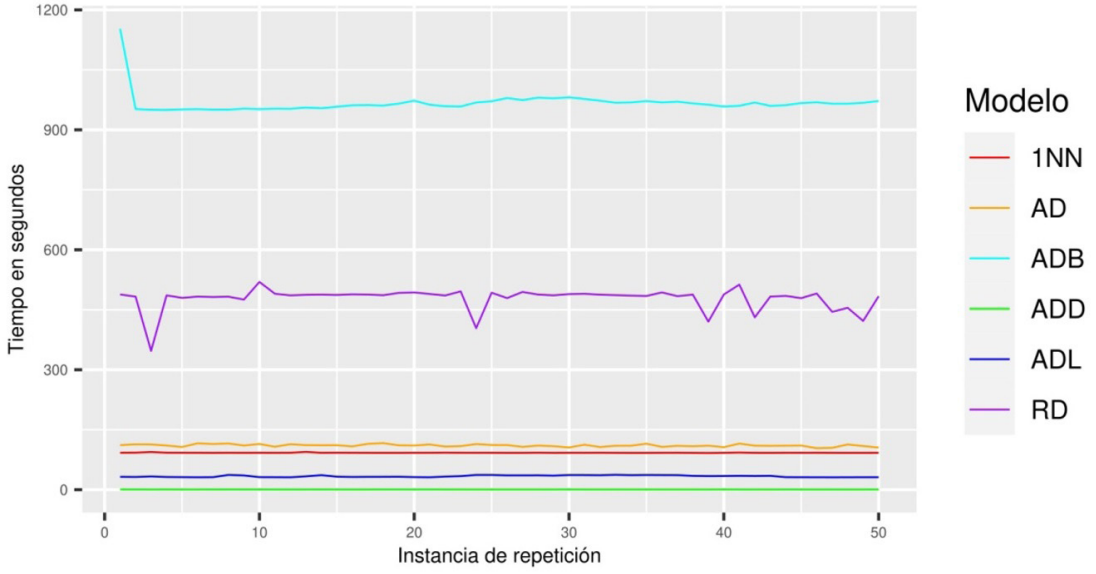

Fig. 5: Tiempo de procesamiento consumido por cada clasificador.

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

Observamos que las técnicas de análisis por discriminante demostraron ser uno de los métodos más rápidos de ejecución. Estos modelos utilizan directamente la matriz de características para hacer predicciones, como consecuencia, el tiempo de entrenamiento es casi nulo y las predicciones son inmediatas. Más aún, el consumo de ADD es incluso más barato que el de ADL. El tiempo de entrenamiento en el método de vecinos más cercanos tampoco necesita cálculos, por lo que el entrenamiento también es rápido. Sin embargo, existe un costo relevante cuando las predicciones necesitan encontrar los vecinos más cercanos para una muestra, en especial al aplicar el método de fuerza bruta. Por el contrario, las predicciones en el modelo de árbol de decisión son rápidas, pero el entrenamiento requiere una cantidad importante de cálculos. Las tareas completas de entrenamiento y validación en AD requieren aproximadamente la misma cantidad de tiempo que el enfoque *k*NN. Es sabido que entrenar redes neuronales es una tarea muy costosa, incluso al usar una sola capa oculta con sólo 7 neuronas, la cantidad de características y clases empleadas hace que el entrenamiento requiera una cantidad considerable de tiempo en relación con otros clasificadores. Al utilizar predictores binarios, ADB debe dicotomizar la matriz de intensidad, lo que implica encontrar el umbral óptimo para cada valor de masa antes de establecer los nuevos valores. Este es un cálculo bastante costoso convirtiendo a este clasificador en el más lento de nuestras pruebas.

Conclusiones

Es importante mencionar que el desarrollo de los experimentos presentados en este trabajo, necesita un gran uso de memoria RAM. Esto se debe a que MALDIquant replica una gran cantidad de datos de la ya numerosa cantidad de espectros de masas. Como optimización, sugerimos buscar picos de forma avanzada y almacenarlos considerando su alineación en el entrenamiento. Observemos que una misma máquina brinda siempre la misma intensidad relativa en relación con el pico base del espectro, por lo que no es necesario alinear las características para muestras obtenidas por un solo equipo.

Aunque todas las pruebas se realizaron utilizando un solo hilo de ejecución, existen niveles que pueden paralelizarse de forma sencilla. Por ejemplo, todos los pasos del preprocesamiento de espectros de masas admiten un tratamiento en paralelo, excepto el cálculo de la matriz de intensidad y el agrupamiento de picos. Las distancias a cada uno de los puntos conocidos en *k*NN también pueden calcularse en paralelo.

Si bien intuimos que redes más complejas pueden mejorar potencialmente el rendimiento del modelo en comparación con los métodos 1NN y ADL, esta ventaja significa un incremento importante al ya elevado tiempo de ejecución. A pesar de que el algoritmo de vecinos más cercanos es una de las técnicas más simples, 1NN se convierte en el modelo más exacto en la clasificación de espectros de masas bacterianas, tanto por género como por especie. Le sigue en desempeño ADL con una diferencia promedio menor a 5×10^{-3} en todos los indicadores considerados. En general, RD no admite el análisis por clase. Por el contrario, ADL presenta excelentes valores para las métricas consideradas, seguido de AD y 1NN.

En caso de trabajar en un contexto similar al presentado, no recomendamos el uso de AD debido a su bajo rendimiento. Por otro lado, sugerimos enfáticamente aplicar ADL ya que, a pesar de que el comportamiento de kNN es ligeramente mejor, el costo computacional consumido por el primer método es aproximadamente un 60% menor.

Agradecimientos

La autora agradece especialmente a Pablo Paglilla por sus valiosos comentarios.

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

Referencias

Ahdesmaki, M., Zuber, V., Gibb, S., y Strimmer, K. (2015) sda: Shrinkage discriminant analysis and CAT score variable selection. R package, 1.3.7. https://CRAN.R-project.org/package=sda

Avanzi, I.R., Gracioso, L.H., Baltazar, M.d.P.G., Karolski, B., Perpetuo, E.A., y do Nascimento, C.A.O. (2017). Rapid bacteria identification from environmental mining samples using MALDI-TOF MS analysis. Environmental Science and Pollution Research 24, 3717-3726.

Bishop, C.M. (2006) Pattern recognition and machine learning. Springer.

Cherkaoui, A., Hibbs, J., Emonet, S., Tangomo, M., Girard, M., Francois, P., y Schrenzel, J. (2010). Comparison of two matrix-assisted laser desorption ionization-time of flight mass spectrometry methods with conventional phenotypic identification for routine identification of bacteria to the species level. Journal of Clinical Microbiology 48, 1169-1175.

Clarridge, J.E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clinical Microbiology Reviews 17, 840-862.

Desaire, H., y Hua, D. (2019). The Aristotle classifier provides accurate identifications of highly similar bacteria analyzed by MALDI-TOF MS. Analytical Chemistry.

Friedman, J., Hastie, T., y Tibshirani, R. (2001) The elements of statistical learning. Springer Series in Statistics New York.

Gibb, S., y Strimmer, K. (2015). Differential protein expression and peak selection in mass spectrometry data by binary discriminant analysis. Bioinformatics 31, 3156-3162.

Hamilton-Miller, J. (1993). A possible pitfall in the identification of Pasteurella spp. with the API system. Journal of Medical Microbiology 39, 78-79.

Hamilton-Miller, J., y Shah, S. (1996). Anomalous but helpful findings from the BBL Crystal ID kit with Haemophilus spp. Letters in Applied Microbiology 23, 47-48.

Janda, J.M., y Abbott, S.L. (2002). Bacterial identification for publication: when is enough enough? Journal of Clinical Microbiology 40, 1887-1891.

Lasch, P., Grunow, R., Antonation, K., Weller, S.A., y Jacob, D. (2016). Inactivation techniques for MALDI-TOF MS analysis of highly pathogenic bacteria–A critical review. TrAC Trends in Analytical Chemistry 85, 112-119.

Lasch, P., Stämmler, M., y Schneider, A. (2018) Version 3 (20181130) of the MALDI-TOF Mass Spectrometry Database for Identification and Classification of Highly Pathogenic Microorganisms from the Robert Koch-Institute (RKI).

Maier, T., y Kostrzewa, M. (2007). Fast and reliable MALDI-TOF MS-based microorganism identification. Chimica Oggi 25, 68.

Manukumar, H., y Umesha, S. (2017). MALDI-TOF-MS based identification and molecular characterization of food associated methicillin-resistant Staphylococcus aureus. Scientific Reports 7, 1-16.

Revista Tecnología y Ciencia
DOI: https://doi.org/10.33414/rtyc.40.75-87.2021 - ISSN 1666-6933

Mullis, K.B. (1990). The unusual origin of the polymerase chain reaction. Scientific American 262, 56-65.

Olsen, G.J., y Woese, C.R. (1993). Ribosomal RNA: A key to phylogeny. The FASEB Journal 7, 113-123.

R Development Core Team. (2013) R: A language and environment for statistical computing. R Foundation for Statistica Computing.

Rosenblatt, F. (1957) The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory.

Rossen, L., Nørskov, P., Holmstrøm, K., and Rasmussen, O.F. (1992). Inhibition of PCR by components of food samples, microbial diagnostic assays and DNA-extraction solutions. International Journal of Food Microbiology 17, 37-45.

Ryan, C., Clayton, E., Griffin, W., Sie, S., y Cousens, D. (1988). SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms 34, 396-402.

Wilson, I.G. (1997). Inhibition and facilitation of nucleic acid amplification. Applied and Environmental Microbiology 63, 3741.

Zhu, W., Sieradzki, K., Albrecht, V., McAllister, S., Lin, W., Stuchlik, O., Limbago, B., Pohl, J., and Rasheed, J.K. (2015). Evaluation of the Biotyper MALDI-TOF MS system for identification of Staphylococcus species. Journal of Microbiological Methods 117, 14-17.