Extracción de factores relevantes en el análisis de datos biomédicos: una metodología basada en técnicas de aprendizaje supervisado

Oscar Reyes
Dpto. Informática y Análisis Numérico
Universidad de Córdoba
Instituto Maimónides de Investigación Biomédica de Córdoba
Email: ogreyes@uco.es

Jose M. Moyano
Dpto. Informática y Análisis Numérico
Universidad de Córdoba
Instituto Maimónides de Investigación Biomédica de Córdoba
Email: jmoyano@uco.es

Antonio Rivero-Juárez
Instituto Maimónides de Investigación Biomédica de Córdoba
Email: arjvet@gmail.com

Raúl M. Luque
Instituto Maimónides de Investigación Biomédica de Córdoba
Email: raul.luque@uco.es

Antonio Rivero
Instituto Maimónides de Investigación Biomédica de Córdoba
Email: ariveror@gmail.com

Justo Castaño
Instituto Maimónides de Investigación Biomédica de Córdoba
Email: justo@uco.es

Sebastián Ventura
Dpto. Informática y Análisis Numérico
Universidad de Córdoba
Instituto Maimónides de Investigación Biomédica de Córdoba
Email: sventura@uco.es

***Resumen*—La determinación del conjunto de variables que se diferencian significativamente entre los grupos de muestras presentes en un estudio biomédico es una tarea que comúnmente se realiza mediante el análisis de cada variable individualmente y/o utilizando técnicas no supervisadas que no tienen en cuenta directamente el criterio de los expertos. En este trabajo se presenta una metodología basada en técnicas de aprendizaje supervisado para guiar el análisis de datos biomédicos, que permite la extracción de subconjuntos de factores relevantes para una correcta clasificación de las muestras en los grupos definidos a priori por los expertos. La metodología propuesta consta de dos fases principales, en la primera se determina la importancia de los factores, mientras que la segunda fase se enfoca en la búsqueda de subconjuntos de factores relevantes mediante la construcción de modelos precisos que logran clasificar correctamente las muestras. La utilidad de la metodología propuesta se ilustra mediante dos casos de estudios reales, mostrando que mediante la aplicación de la misma se podrían detectar relaciones complejas entre los factores, y que favorece el análisis de datos biomédicos que tienen un elevado número de variables descriptoras.**

I. Introducción

Las técnicas de aprendizaje no supervisado, como el análisis de componentes principales y los algoritmos de clustering, son ampliamente utilizadas en el campo de la bioinformática [1]. Sin embargo, en sentido general este tipo de técnicas no toma en cuenta el criterio de los expertos, que previamente al análisis pudieron haber clasificado las muestras en grupos (cáncer vs sano, tumor maligno vs tumor benigno, etc.), lo que puede implicar una perdida significativa de información para la extracción del conocimiento en el análisis de datos biomédicos.

Las técnicas de aprendizaje supervisado, por otro lado, permiten que el conocimiento aportado por los expertos pueda guiar el análisis de los datos, mostrándole a los algoritmos cuáles son las conclusiones (salidas) a la cuales deben llegar. Por ejemplo, un algoritmo de clasificación de imágenes para el diagnóstico del melanoma tratará de aprender las relaciones que vinculan a los datos contenidos en las imágenes con las etiquetas asignadas [2]. De esta manera, los algoritmos de aprendizaje supervisado permiten, dado unos datos de entrada, encontrar una función que produce una salida lo más aproximada posible al conocimiento de los expertos.

Una de las tareas que comúnmente se realiza en el análisis de datos biomédicos es la determinación del conjunto de variables que se diferencian significativamente entre los grupos de muestras definidos por los expertos [3]. Por ejemplo, el *p-value* calculado por el t-test es ampliamente usado como indicador de la relevancia de un factor (en lo adelante se usa el termino "factores" para indicar el conjunto de variables que describe las muestras de un problema). Sin embargo, además de que los test paramétricos no deben ser usados en todas las situaciones (este tema se escapa del objetivo de este trabajo), se debe considerar que de esta manera el análisis que se realiza es univariante, desechándose así las relaciones estadísticas que normalmente existen entre los factores de un problema.

Por otro lado, es de destacar que muchos problemas de

biomedicina implican el análisis de un número considerable de factores [4], lo cual hace que la tarea anterior sea inviable de realizar si antes no se han filtrado los factores que son realmente relevantes para el estudio del problema. Ejemplo de esto se encuentra al realizar estudios que involucran el análisis de las expresiones de genes sobre un conjunto de muestras.

En este trabajo se presenta una metodología, la cual está basada en técnicas de aprendizaje supervisado, que permite la extracción de subconjuntos de factores relevantes para una correcta clasificación de las muestras en las clases definidas por los expertos (en lo adelante se usa el término "clase" para indicar la variable que describe la condición por la cual los expertos agrupan las muestras). Esta metodología consta de dos fases principales: (a) la determinación de la importancia de los factores, que permite determinar un ranking de importancia; y (b) la construcción de modelos de clasificación a partir de dicho ranking. El uso de esta metodología puede aportar varios beneficios al análisis de datos biomédicos, ya que no solo se pueden determinar subconjuntos de factores relevantes que influyen en la correcta clasificación de las muestras, sino que los métodos desarrollados también son capaces de detectar distribuciones conjuntas entre factores, e interacciones y dependencias complejas respecto a las clases.

El resto del este trabajo se organiza de la siguiente manera. En la Sección II se describe la metodología, explicando cada una de sus fases. La aplicación de la metodología propuesta se ilustra en la Sección III mediante dos casos de estudio reales, uno relacionado con el diagnóstico de tumores neuroendocrinos pulmonares y el otro con el aclaramiento espontáneo en Hepatitis C. Finalmente, en la Sección IV se presentan las conclusiones del presente trabajo.

II. Metodología

El esquema general de la metodología que se propone se muestra en la Figura 1. El preprocesamiento de los datos es un paso opcional, que no nos detendremos a analizar en profundidad en este trabajo. Sin embargo, hay que destacar que generalmente la calidad de los resultados en el análisis de datos biomédicos depende en gran medida de que se haya hecho un correcto preprocesamiento de los datos [5]. El preprocesamiento de datos abarca una amplia gama de métodos, que van desde la eliminación de outliers y la estimación de valores perdidos hasta el centrado, escalado y transformación de los datos. El uso de cada uno de los métodos de preprocesado debe tener una lógica y justificación correcta, ya que si bien es cierto que un correcto preprocesado de datos puede mejorar significativamente el análisis, también un preprocesamiento incorrecto puede conllevar a la obtención de conclusiones erróneas.

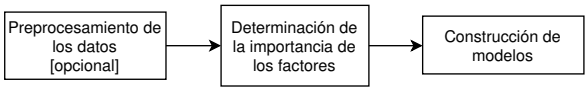

Figura 1. Esquema general de la metodología.

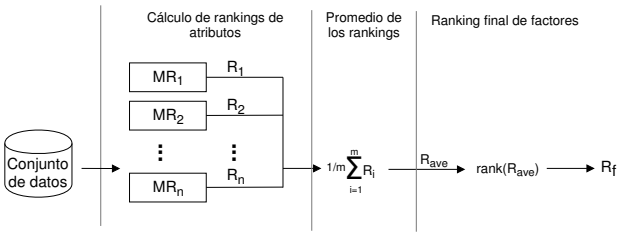

Figura 2. Cálculo del ranking final de factores.

II-A. *Determinación de la importancia de los factores*

La primera fase de la metodología se enfoca en la determinación de la importancia de cada uno de los factores del problema, y para ello se propone el uso de algoritmos supervisados de pesado de atributos o *feature weighting* (FW) [6]. La relevancia de un factor se determina mediante la asignación de un peso que representa la información que tiene este para la correcta separación de las muestras en las clases definidas por los expertos [7]. Un método de FW le asigna a cada factor un peso, siendo posible de esta manera obtener un ranking de factores directamente. El objetivo final de esta fase de la metodología es calcular un ranking donde están ordenados de mayor a menor importancia todos los factores.

Digamos que disponemos de m métodos de FW para lograr una mejor estimación del ranking final de factores. R_i representa el ranking calculado por el método i-ésimo, $R_i(f)$ representa el valor del factor f en el ranking R_i, y F es el conjunto de todos los factores existentes en el estudio. El ranking final de factores se calcula de la siguiente manera:

$$R_f{=}rank\left(\frac{1}{m}\sum_{i=1}^{m} R_i(f) : \forall f \in F\right), \qquad (1)$$

donde la función $rank(...)$ calcula el ranking final de factores a partir de los valores promedios de cada uno de los factores en los m ranking iniciales. La Figura 2 representa el cálculo del ranking final de factores.

Respecto a la cantidad de métodos de FW a utilizar en la estimación, cuanto mayor sea el número de métodos, más precisa será la estimación del ranking final. En este sentido, se recomienda el uso de métodos supervisados de FW que sean independientes de un clasificador para estimar la importancia de un factor, evitando de esta manera la introducción de sesgos y dependencias en el proceso de estimación. En su lugar se propone el uso de métodos de FW que calculen directamente medidas sobre los datos, como medidas de distancia, entropía o correlación. Estos métodos son conocidos en la literatura especializada como métodos filtros, y entre los más populares podemos encontrar a *Correlation Attribute Evaluation* [8] *Gain Ratio* [9], *Information Gain* [10] y *ReliefF* [11].

Es importante resaltar que para lograr una estimación precisa de la importancia de los factores, es necesario que cada uno de los m métodos de FW sean ejecutados mediante algún proceso de validación cruzada, el cual dependerá del tamaño del conjunto de datos analizado. Normalmente una validación

cruzada de 10 particiones repetidas varias veces es suficiente para lograr una buena estimación. Sin embargo, en el caso de que el conjunto de datos sea muy pequeño, se deberán considerar otras alternativas para la estimación, como una validación cruzada dejando uno fuera o *Leave One-out Cross Validation* (LOOC).

Por último, es importante destacar que en el ámbito de la biomedicina comúnmente la importancia de un factor se calcula agrupando las muestras por dos o más condiciones y se calcula la diferencia de este factor entre los diferentes grupos; por ejemplo el *p-value* calculado por el t-test es ampliamente usado como indicador de la relevancia de un factor. Sin embargo, además de que los test paramétricos no deben ser usados en todas las situaciones, se debe considerar que de esta manera el análisis que se realiza es univariante, desechándose así las relaciones estadísticas que normalmente existen entre varias variables descriptoras del problema. Esta característica principal es lo que distingue esta primera fase de la metodología propuesta en este trabajo. Los métodos filtros como *ReliefF*, son capaces de detectar distribuciones conjuntas entre variables, interacciones y dependencias complejas respecto a la clase, además de considerar como un todo el conjunto de factores F.

II-B. Construcción de modelos

Una vez estimado el ranking de factores, entonces se puede proceder a la determinación de los subconjuntos de factores que mejor logran predecir la clase añadida por los expertos. Sin embargo, esta no es una tarea fácil de realizar, ya que es complejo determinar un punto de corte a partir del cual los factores restantes se pueden considerar como irrelevantes para el análisis.

En lugar de realizar directamente un análisis sobre el ranking de factores R_f, en esta fase de la metodología se propone una búsqueda heurística guiada para encontrar el mejor subconjunto de factores; el método propuesto está inspirado en el algoritmo presentado por Reyes et al. [12]. En otras palabras, mediante esta fase se podrán determinar aquellos subconjuntos de factores a partir de los cuales se inducen modelos capaces de predecir efectivamente a qué clase pertenece cada muestra. La Figura 3 representa los pasos que sigue el algoritmo diseñado. Como puede observarse, es un proceso iterativo en el que, comenzando con el factor posicionado en el tope del ranking, en cada iteración se analiza si la inclusión del siguiente factor al subconjunto produce un mejor modelo. Finalmente el mejor subconjunto de factores será aquel sobre el cual se induce el mejor clasificador a lo largo de todas las iteraciones.

Para la comparación de la efectividad de los modelos se puede utilizar cualquier medida de evaluación, como el área bajo la curva ROC (AUC, por sus siglas en inglés), ampliamente usada en el análisis de datos biomédicos. Por otro lado, es de destacar que este procedimiento se puede realizar solamente considerando el ranking R_f o para cada sub-ranking R_f^g : $\forall g \in R_f$; el sub-ranking de factores R_f^g está compuesto por el factor g en el tope y todos los subsecuentes factores en R_f. El

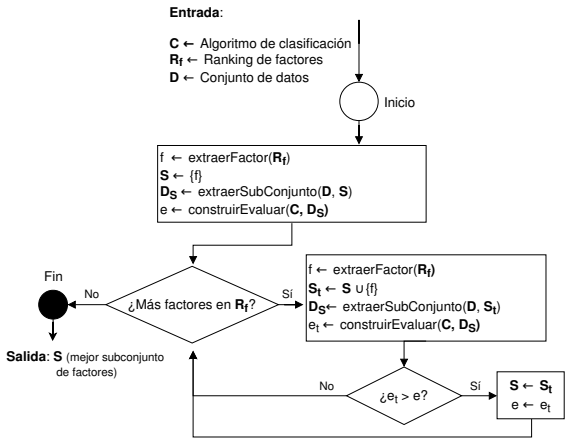

Figura 3. Búsqueda heurística del mejor subconjunto de factores.

primer caso claramente requeriría la construcción de un menor número de clasificadores ($O(|F|)$), mientras que la segunda opción requeriría la construcción de un número cuadrático de clasificadores ($O(|F|(|F|-1)/2)$); sin embargo, esta última opción es la que produce una mejor estimación. Además, al igual que en la fase anterior, es importante considerar que para lograr una estimación precisa del rendimiento de los clasificadores se debe utilizar un procedimiento de validación cruzada en el proceso de construcción de los mismos.

Respecto al criterio de parada del algoritmo, la Figura 3 ilustra un procedimiento que termina una vez que han sido evaluados todos los factores del ranking. Sin embargo, se pudieran definir otros criterios más flexibles que eviten evaluar completamente el ranking. Por ejemplo, los expertos pueden definir un umbral de aceptación de tal manera que el procedimiento se detenga a penas que se encuentre un modelo con un rendimiento superior a dicho umbral. Por otra parte, los expertos pueden estar interesados en no solo analizar el mejor modelo encontrado, sino los n mejores modelos construidos.

En esta fase de la metodología se puede utilizar cualquier algoritmo de clasificación para la construcción de los modelos, siempre y cuando este sea adecuado para el análisis. Por ejemplo, se puede usar cualquier algoritmo de clasificación binaria si solo se tienen dos posibles clases para las muestras, o cualquier algoritmo de clasificación multi-clase en caso de que se tengan más de dos clases. Por otra parte, es lógico que el rendimiento obtenido por los modelos dependerá de la efectividad y potencia que tenga el algoritmo de clasificación empleado; por ejemplo, es de esperar que un modelo de ensamblado como *Random Forest* [13] obtenga en promedio mejores resultados que otros modelos más sencillos como *KNN* [10] o *Naive Bayes* [10].

Por último, aclarar que también se pueden utilizar algoritmos de clasificación que tienen un proceso embebido de selección de atributos. En este último caso, es posible que el subconjunto de factores que finalmente utilice el modelo sea más pequeño que el subconjunto original sobre el cual se entrenó el algoritmo.

III. Casos de estudio

La metodología presentada en este trabajo es utilizada actualmente por varios laboratorios de investigación del Instituto Maimónides de Investigación Biomédica de Córdoba. A continuación se presentan dos casos de estudios reales que muestran la aplicación y utilidad de la propuesta.

III-A. Diagnóstico de tumores neuroendocrinos pulmonares

Los tumores neuroendocrinos pulmonares representan entre el 20 y el 30 % de todos los tumores neuroendocrinos [14]. La heterogeneidad, sus diferentes comportamientos clínicos, y la posibilidad de aparición recurrente y de hacer metástasis a largo plazo, enfatiza la importancia que tiene la identificación de nuevos marcadores de diagnósticos y terapéuticos, que pueden mejorar el diagnóstico, pronóstico y/o el tratamiento de los pacientes que sufren esta enfermedad [15].

Para este problema, los datos disponibles fueron de 26 muestras pareadas (muestras tumorales con su respectiva muestra de tejido normal adyacente), donde por cada muestra se tenía la expresión de 44 factores que regulan la maquinaria de *splicing*. El objetivo principal del estudio fue determinar subconjuntos de factores que caracterizaran claramente a las dos clases de muestras. En los datos originales no había datos perdidos, y en la etapa de preprocesamiento se eliminaron previamente aquellos factores que tenían varianza igual a cero, y además se centraron y escalaron los datos.

Mediante la primera fase de la metodología propuesta se obtuvo un ranking de factores que permitió determinar cuáles son en promedio los factores más relevantes para diferenciar las clases de muestras. La Figura 4 muestra las importancia de los 20 primeros factores del ranking.

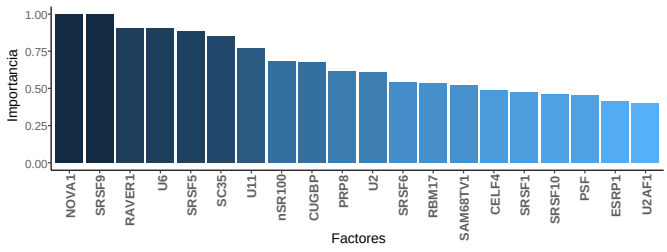

Figura 4. Ranking de factores para diferenciar entre muestras normales y tumorales.

Posteriormente, en la segunda fase de la metodología se utilizaron dos algoritmos de clasificación (*Logistic Regression* (LR) [16] y *Random Forest*) para la evaluación de subconjuntos de factores. Para la estimación de la precisión de los clasificadores se utilizó un procedimiento LOOC debido a que el número de muestras en el estudio era pequeño, y además se realizó una búsqueda *grid* de los mejores parámetros para el entrenamiento de los algoritmos. A partir de este análisis se encontraron 100 modelos con AUC mayor o igual a 0,85, arrojando subconjuntos de factores relevantes que aparecen generalmente en todos los modelos predictivos.

Por otra parte, aunque en la descripción de la metodología (véase Sección II) nos hemos limitado al uso de algoritmos de clasificación, es de destacar que en la segunda fase del procedimiento se podrían emplear algoritmos de clustering, siempre y cuando los resultados de agrupamiento sean analizados con medidas externas (como la pureza) que tienen en cuenta las clases definidas a priori por los expertos. De esta manera, a los efectos de la metodología el algoritmo de clustering empleado actuaría como si fuera un algoritmo supervisado. En este estudio, los resultados de la segunda fase de la metodología haciendo uso de un algoritmo de clustering jerárquico coinciden con los resultados obtenidos anteriormente por los algoritmos de clasificación, validando así la relevancia de los subconjuntos encontrados. La Figura 5 muestra un *heatmap* con uno de los subconjuntos de factores encontrados.

III-B. Aclaramiento espontáneo en Hepatitis C

Una vez que un paciente se infecta por el virus de Hepatitis C (VHC), se produce una hepatitis aguda que en la mayoría de los casos lleva a una infección crónica caracterizada por el avance gradual de fibrosis hepática, cirrosis y carcinoma hepatocelular [17]. Sin embargo, un porcentaje menor de pacientes resuelven su infección de manera espontánea. Por tanto, la identificación de factores o marcadores que ayuden a la predicción del aclaramiento espontáneo (AE) o infección crónica (IC) de VHC tendrían un alto impacto en la selección de la terapia que debería utilizarse para su tratamiento.

Para este problema, los datos disponibles fueron de 138 pacientes infectados con VHC, 81 de ellos con infección crónica y 57 en los que se produjo AE. Cada paciente estaba descrito por 43 marcadores distintos. En 43 muestras habían valores perdidos en algunos de sus marcadores, y se utilizó el algoritmo *knn-Imputation* [18] con $k = 3$ para estimar dichos valores.

A partir de la primer fase de la metodología, se obtuvo un ranking de factores, del cual en la Figura 6 se muestran los 20 primeros. De esta manera, se puede observar de manera simple la importancia de cada uno de los factores en el problema de VHC. El primer factor en el ranking tiene una importancia cercana a 1, lo que significa que en el proceso de estimación todos los métodos de FW le asignaron en promedio una alta importancia a dicho factor.

Posteriormente, en la segunda fase de la metodología se utilizaron varios clasificadores como *C4.5* [10] , *PART* [19], *Random Forest*, *Sparse Discriminant analysis* (*sparseLDA*) [20] y *Logistic Model Trees* (*LMT*) [21]. Para la ejecución de cada modelo se repitió 3 veces una validación cruzada en 10 particiones, evaluando en cada caso sobre el conjunto de *test* correspondiente, y promediando así los valores entre un total de 30 ejecuciones. Además, para la búsqueda de los parámetros de los algoritmos se realizó una búsqueda aleatoria de parámetros entre 30 combinaciones distintas.

Para este problema, se obtuvieron en total casi 400 modelos distintos con un AUC $> 0{,}8$; donde 126 tenían un AUC $> 0{,}85$; y 30 modelos con un AUC $> 0{,}87$. Posteriormente, para aquellos modelos con AUC $> 0{,}85$ se midió el número de veces que aparece cada uno de los atributos entre dichos

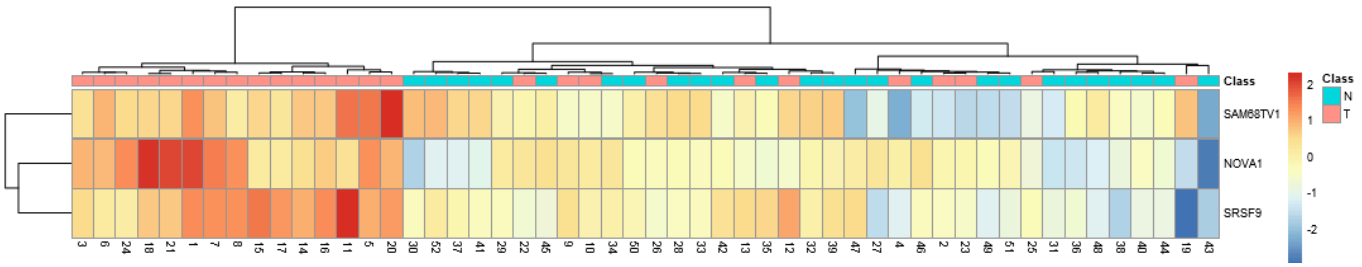

Figura 5. Heatmap generado a partir de un subconjunto de tres factores.

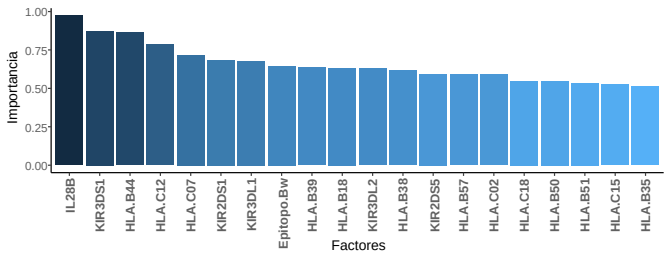

Figura 6. Ranking de factores para el problema de aclaramiento espontáneo en VHC.

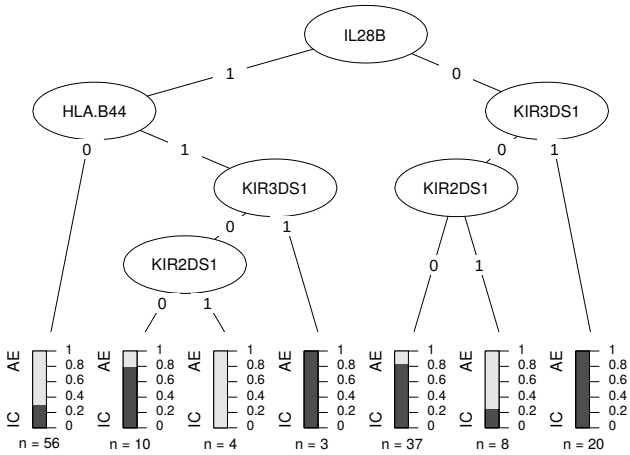

Figura 7. Árbol de decisión generado por uno de los modelos de C4.5 para el problema de VHC.

modelos. Esta medida proporciona otra aproximación de la importancia de cada uno de los factores en la predicción de la clase, pues un factor que aparezca en un gran número de modelos previsiblemente será más importante en la predicción de la clase que otro factor que aparezca con menor frecuencia.

Por otro lado, el hecho de utilizar modelos de árboles de decisión o reglas de asociación, como *C4.5* o *PART* respectivamente, hace que los modelos resultantes sean fácilmente interpretables por los expertos, pudiendo ver de manera sencilla cómo los factores discriminan para determinar si se predice una u otra clase. En la Figura 7 se muestra un ejemplo de los modelos de árbol obtenidos en el análisis. En la figura se observa como, para un nuevo paciente, dependiendo del valor de cada uno de los factores seleccionados, el modelo descenderá en el árbol hasta predecir la clase a la cual pertenece el paciente (nodo hoja). En estas hojas se puede observar cuál es el porcentaje de pacientes de cada una de las clases que cumplen las condiciones de los factores de los nodos superiores. Por ejemplo, para un paciente donde el factor *IL28B* valga 1 y el factor *HLA.B44* valga 0, el modelo asignará la clase AE, ya que en torno al 70 % de los pacientes observados con esa combinación de factores pertenecen a dicha clase. Cabe destacar también que, como se puede observar, los factores que aparecen en este modelo de árbol se encontraban en las primeras posiciones del ranking obtenido en la primera fase de la metodología, siendo la raíz del árbol precisamente el factor con mayor importancia en el ranking.

Por último, se comparan los resultados de los modelos generados mediante la metodología propuesta con modelos que son construidos considerando todos los factores del problema. La Tabla I muestra los resultados de esta comparación. Para cada

Tabla I
Mejora del rendimiento predictivo al ejecutar los distintos algoritmos utilizando todos los factores (AUC_0) o un subconjunto de los mismos determinados por la metodología propuesta (AUC_{sub}).

Algoritmo	AUC_0	AUC_{sub}	f_{sub}	% mejora
C4.5	0,766	0,842	10	9,92 %
PART	0,742	0,869	11	17,12 %
Random Forest	0,825	0,882	14	6,91 %
sparseLDA	0,839	0,880	8	4,89 %
LMT	0,803	0,872	7	8,59 %

uno de los algoritmos utilizados se muestra el valor de AUC obtenido utilizando los 43 factores del problema (AUC_0), y el mejor valor de AUC obtenido generando el modelo con subconjuntos de factores (AUC_{sub}); para este último caso, se indica además el número de factores utilizados para la construcción del modelo (f_{sub}). En la última columna de la tabla se incluye el porcentaje de mejora en rendimiento predictivo, calculado como $\frac{AUC_{sub}-AUC_0}{AUC_0}$. A partir de los resultados se puede observar como mediante la metodología propuesta se pueden obtener mejoras considerables en las predicciones de la clase; por ejemplo en el caso de *PART* se obtiene una mejora de un 17 % considerando solo 11 factores de los 43 existentes.

IV. Conclusiones

En este trabajo se ha propuesto una metodología para la extracción de factores relevantes en datos biomédicos mediante el uso de técnicas de aprendizaje supervisado. Dicha metodología se divide en dos partes principales, la creación de un ranking de factores que determine la importancia de cada uno de ellos, y la búsqueda de subconjuntos de factores que permitan construir modelos con una alta precisión para predecir el tipo de muestra. Mediante esta metodología es posible detectar relaciones complejas que existen entre los factores que describen a las muestras de un estudio, superando de esta manera el análisis de factores individuales que comúnmente se emplea en biomedicina.

La aplicación de la metodología se ilustró mediante dos casos de estudios reales, mostrando la utilidad y potencial de la misma. En estos problemas, gracias a la metodología propuesta se pudieron identificar subconjuntos de factores relevantes que permiten con una alta precisión clasificar las muestras en las clases definidas a priori por los expertos. Se espera que el uso de la presente metodología se pueda extender a otros grupos de investigación biomédica, facilitando el análisis de datos, así como la creación de biomarcadores para el tratamiento temprano de enfermedades patológicas.

Agradecimientos

Este trabajo ha sido financiado por el proyecto TIN2017-83445-P del Ministerio de Economía y Competitividad y Fondos FEDER. También ha sido financiado por la ayuda FPU del Ministerio de Educación FPU15/02948.

Referencias

[1] P. Baldi and S. Brunak, *Bioinformatics: the machine learning approach*. MIT press, 2001.

[2] I. Giotis, N. Molders, S. Land, M. Biehl, M. F. Jonkman, and N. Petkov, "MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images," *Expert Systems with Applications*, vol. 42, no. 19, pp. 6578 – 6585, 2015.

[3] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[4] M. Gahete, M. del Rio-Moreno, E. Alors-Perez, O. Reyes, A. Camargo, J. Delgado-Lista, J. Lopez-Miranda, J. P. Castaño, and R. M. Luque, "Identification of an altered spliceosome-associated fingerprint as an early, predictive event for the development of type 2 diabetes in high-risk patients," in *100th Endocrine Society (ENDO) annual meeting*, 2018.

[5] R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: improving the biological information content of metabolomics data," *BMC genomics*, vol. 7, no. 1, p. 142, 2006.

[6] D. Wettschereck, D. W. Aha, and T. Mohri, "A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms," *Artificial Intelligence Review*, vol. 11, no. 1-5, pp. 273–314, 1997.

[7] O. Reyes, C. Morell, and S. Ventura, "Evolutionary feature weighting to improve the performance of multi-label lazy algorithms," *Integrated Computer-Aided Engineering*, vol. 21, no. 4, pp. 339–354, 2014.

[8] M. A. Hall, "Correlation-based feature selection for machine learning," Tech. Rep., 1999.

[9] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2016.

[10] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[11] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[12] O. Reyes, C. Morell, and S. Ventura, "Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context," *Neurocomputing*, vol. 161, pp. 168–182, 2015.

[13] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[14] A. Fisseler-Eckhoff and M. Demes, "Neuroendocrine tumors of the lung," *Cancers*, vol. 4, no. 3, pp. 777–798, 2012.

[15] A. D. Herrera-Martínez, M. D. Gahete, R. Sánchez-Sánchez, R. O. Salas, R. Serrano-Blanch, A. Salvatierra, L. J. Hofland, R. M. Luque, M. A. Gálvez-Moreno, and J. P. Castaño, "The components of somatostatin and ghrelin systems are altered in neuroendocrine lung carcinoids and associated to clinical-histological features," *Lung Cancer*, vol. 109, pp. 128–136, 2017.

[16] S. K. Shevade and S. S. Keerthi, "A simple and efficient algorithm for gene selection using sparse logistic regression," *Bioinformatics*, vol. 19, no. 17, pp. 2246–2253, 2003.

[17] M. Frias, A. Rivero-Juárez, D. Rodriguez-Cano, A. Camacho, P. López-López, M. Risalde, B. Manzanares-Martín, T. Brieva, I. Machuca, and A. Rivero, "HLA-B, HLA-C and KIR improve the predictive value of IFNL3 for Hepatitis C spontaneous clearance," *Scientific Reports*, vol. 8, no. 1, p. 659, 2018.

[18] L. Torgo, *Data Mining with R, learning with case studies*. Chapman and Hall/CRC, 2010. [Online]. Available: http://www.dcc.fc.up.pt/ ltorgo/DataMiningWithR

[19] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," in *Fifteenth International Conference on Machine Learning*, 1998, pp. 144–151.

[20] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[21] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine learning*, vol. 59, no. 1-2, pp. 161–205, 2005.