Resolución de Problemas Biomédicos mediante Técnicas de Extracción de Conocimiento

Oscar Reyes, Jose M. Luna, Jose M. Moyano, Eduardo Pérez y Sebastián Ventura
Dpto. Informática y Análisis Numérico, Universidad de Córdoba
Instituto Maimónides de Investigación Biomédica de Córdoba
Email: {ogreyes; jmluna; jmoyano; z72pepee; sventura}@uco.es

***Resumen*—En este trabajo se presenta el grupo "*Descubrimiento de Conocimiento y Sistemas Inteligentes en Biomedicina*" del Instituto Maimónides de Investigación Biomédica de Córdoba. Este grupo, de reciente creación, está integrado por varios investigadores interesados en las áreas de extracción de conocimiento y desarrollo de sistemas inteligentes, con especial interés en la resolucion de problemas de análisis de datos aplicados al ámbito de la biomedicina. A lo largo del documento, se describen brevemente algunas de las líneas de trabajo del grupo, así como algunos de los resultados alcanzados recientemente.**

I. Introducción

En los últimos años, las técnicas de inteligencia artificial se han revelado como una herramienta muy potente para la resolución de problemas complejos en el ámbito de la biomedicina [1]. De todas estas técnicas, merecen una mención especial el aprendizaje automático y la minería de datos, que han posibilitando la extracción automática de conocimiento útil a partir de bases de datos biomédicas de gran tamaño y complejidad [2, 3].

Este interés por una explotación de las distintas bases de datos existentes, que se generan como consecuencia tanto de la información masiva que generan las nuevas técnicas de diagnóstico [4] como de los cada vez más populares registros electrónicos de salud [5] está provocando un creciente interés por las disciplinas que integran la denominada ciencia de datos [6]. Los investigadores en biomedicina ya no solo saben estadística clásica, sino que empiezan a incorporar a sus estudios técnicas avanzadas de análisis de datos e incorporan a sus equipos especialistas en estas disciplinas que les ayuden a resolver los problemas que se plantean al intentar explotar estas nuevas fuentes de información. Un ejemplo de esta evolución se puede apreciar analizando los planes estratégicos de instituciones como el Instituto Maimónides de Investigación Biomédica de Córdoba[1] (IMIBIC), que contempla para el quinquenio 2016-2020 acciones para incorporar científicos de datos a sus equipos de trabajo, los cuáles proporcionarán este nuevo valor añadido al desarrollo de las investigaciones realizadas en la institución. Los expertos del IMIBIC reconocen que la ciencia de datos juega un papel fundamental hoy en día en el diagnóstico médico, especialmente con el desarrollo de la medicina de precisión, que está posibilitando la puesta a punto de estrategias inteligentes para la prevención, diagnóstico y tratamiento adaptados al perfil clínico, genético y molecular de cada paciente y cada enfermedad concreta.

Otra de las muestras del interés que suscita la aplicación de este tipo de técnicas entre los investigadores de biomedicina es la incorporación a estas instituciones de equipos, tanto técnicos como investigadores, especializados en el análisis de datos. Este es el caso del grupo *Descubrimiento de Conocimiento y Sistemas Inteligentes en Biomedicina*, al que pertenecen los autores del presente trabajo. Este es un grupo de investigación cuyos integrantes proceden del área de extracción de conocimiento y que, en los últimos años, tras su incorporación al instituto, han ido aumentando su interés por la resolución de problemas relacionados con el análisis de datos biomédicos, debido al interés que estos plantean desde el punto de vista aplicado y la complejidad de los mismos, que plantean interesantes retos desde el punto de vista teórico. El objetivo de este trabajo es presentar brevemente las líneas de investigación que desarrolla el grupo actualmente, así como algunos de los resultados alcanzados en colaboración con otros equipos de investigación del IMIBIC.

El resto del trabajo se organiza de la siguiente manera. En la Sección II se presenta la composición del grupo, se explican brevemente sus principales líneas de investigación y se mencionan las colaboraciones con otros grupos de investigación. Algunos de los estudios realizados por el grupo se presentan en la Sección III. Finalmente, en la Sección IV se presentan las conclusiones del presente trabajo.

II. Composición del grupo, líneas de investigación y colaboraciones

El grupo *Descubrimiento de Conocimiento y Sistemas Inteligentes en Biomedicina* se incorporó al IMIBIC en el año 2014. Dicho grupo está formado por investigadores del grupo de investigación *Knowledge Discovery and Intelligent Systems* (KDIS) de la Universidad de Córdoba[2], interesados en aplicar los algoritmos que llevan desarrollando desde el año 2009 a problemas biomédicos. El grupo actualmente está compuesto por 10 investigadores doctores y 5 estudiantes de doctorado. El investigador principal del grupo es el Dr. Sebastián Ventura Soto.

Los dos campos principales en los cuales se centran los estudios realizados por el grupo son: el descubrimiento de

[1] https://www.imibic.org

[2] http://uco.es/kdis

conocimiento y minería de datos, así como la aplicación de técnicas de inteligencia artificial para el desarrollo de sistemas inteligentes. La línea de trabajo que el grupo se propone desarrollar en los siguientes años se enfoca en el desarrollo de metodologías de análisis de datos para resolver problemas complejos de biomedicina de gran relevancia para la sociedad, como son la predicción de melanoma, el estudio de los factores de splicing alternativo, la predicción y descripción de patologías en hipertención arterial, entre otros.

II-A. Líneas de investigación

A continuación se presentan brevemente las líneas de investigación que desarrolla el grupo.

II-A1. Desarrollo de modelos predictivos: Las técnicas de aprendizaje supervisado 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 [7]. 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. Los modelos predictivos se pueden clasificar teniendo en cuenta el tipo de salida en modelos de clasificación (salida discreta) y modelos de regresión (salida continua). Por otra parte, los modelos de clasificación y regresión tradicionales producen una única salida a partir de un único vector de variables descriptoras. Sin embargo, en los últimos años la construcción de modelos a partir de representaciones de datos más flexibles (multi-instancia, multi-vista, multi-etiqueta, multi-salida) ha sido de gran interés en la comunidad científica.

Los estudios realizados por el grupo en este campo se basan en el desarrollo de modelos predictivos, tanto para problemas clásicos de predicción [8] como para problemas con representaciones más flexibles [9–12]. Alguno de estos estudios han sido aplicados directamente a problemas de Biomedicina; por ejemplo, en la predicción del riesgo de padecer diabetes y en el diagnóstico a partir de textos clínicos usando modelos de clasificación multi-etiqueta.

II-A2. Minería de patrones: Los patrones, como elemento clave en el análisis de datos, representan cualquier tipo de homogeneidad y regularidad en los datos, y por lo tanto estos sirven como descriptores de propiedades importantes presentes en los datos. Las técnicas de minería de patrones son comúnmente de carácter descriptivo y no supervisado, por lo que no se requiere incorporar conocimiento experto al comienzo de un estudio [13]. En ocasiones, dichas tareas descriptivas se enfocan en variables objetivo y, por tanto, tienen cierto carácter supervisado [14].

Los estudios realizados por el grupo en este campo se enfocan en la extracción de conocimiento a partir de datos originales y el descubrimiento de información útil asociada a variables específicas de interés. Se estudian diferentes tipos de patrones, incluyendo patrones frecuentes e infrecuentes [15], y patrones definidos sobre diferentes tipos de datos como relacionales, secuenciales, y en dominios ambiguos [13, 14]. Por otra parte, el grupo desarrolla algoritmos para la minería de patrones respecto a una (o múltiples) variable objetivo, incluyendo algoritmos de descubrimiento de sub-grupos [16] y algoritmos para modelos excepcionales [17], entre otros.

II-A3. Desarrollo de Modelos Big Data: Hoy en día los sistemas de información producen colecciones masivas de datos que superan las capacidades de procesamiento y almacenamiento de los métodos de extracción de conocimiento tradicionales. Los problemas *Big Data* se caracterizan por grandes volúmenes de datos, que se generan comúmente a gran velocidad, con gran variedad de formatos, donde es necesario garantizar la veracidad de los datos y por último extraer el valor (conocimiento) oculto en ellos. [18]

En los últimos años, los investigadores se han enfocado principalmente en la mejora de la escalabilidad de los algoritmos para enfrentar correctamente el desafío que conlleva el tratamiento de grandes volúmenes de datos. Este desafío es especialmente acentuado en el campo de la biomédicina, donde podemos encontrar enormes bases de datos genéticos y bases de datos de historias clínicas. Sin embargo, una integración efectiva y eficiente de todos los datos biomédicos disponibles a partir de diferentes fuentes con el objetivo de extraer conocimiento útil y no trivial no es sencilla ni directa en la mayoría de los casos [19]. En este sentido, el grupo de investigación ha desarrollado algunos modelos [15, 16, 20, 21], los cuales pueden ser aplicados a problemas *Big Data* en el campo de la Biomedicina.

II-A4. Desarrollo de flujos de trabajo: Los flujos de trabajo o *workflows* son mecanismos de alto nivel que permiten automatizar y describir procesos como una serie de actividades interconectadas que producen una salida deseada. En el caso del análisis de datos, los *workflows* ofrecen una serie de pasos para conducir el análisis teniendo en cuenta las características de los dominios de aplicación, ocultando los requerimientos computacionales y detalles técnicos de las técnicas de análisis, y facilitando el desarrollo de procesos complejos para la extracción de conocimiento a partir de datos heterogéneos [22].

La aplicación de los *workflows* en ciencia de datos enfrenta varios desafíos, que no solo se relacionan con la descomposición de los métodos de extracción de conocimiento en procesos y actividades, sino también con la adaptación y disposición de procedimientos algorítmicos de bajo nivel para el análisis intensivo de datos. Por otro lado, los *workflows* para problemas *Big Data* requieren el análisis de métodos paralelizables de minería de datos, su ejecución en clusters o sistemas basados en la nube, la optimización de los procesos para la ejecución eficiente de tareas complejas, etc. En este campo, el grupo está trabajando en la construcción de soluciones basadas en *workflows* [23–25], con el objetivo principal de mejorar la aplicación y reusabilidad de las metodologías propuestas para el análisis de datos en los estudios biomédicos que se realizan en el IMIBIC.

II-B. *Colaboraciones*

El grupo de investigación colabora activamente con varios grupos del IMIBIC, entre los que podemos mencionar:

- Grupo GC-05 "*Enfermedades autoinmunes sistémicas-inflamatorias crónicas del aparato locomotor y tejido conectivo*" - Inv. principal Dra. Rosario López Pedrera. Se colabora en estudios para la determinación de los factores más relevantes en enfermedades autoinmunes y cardiovasculares, y además se analiza cómo estas enfermedades incrementan el riesgo de ictus y de mortalidad.
- GC-07 "*Nefrología. Daño celular en la inflamación crónica*" - Inv. principal Dr. Pedro Aljama García. Se está colaborando en la obtención de nuevos parámetros hemodinámicos ambulatorios y medicina de precisión.
- Grupo GC-08 "*Hormonas y Cáncer*" - Inv. principal Dr. Justo P. Castaño Fuentes. Se colabora en el estudio de los principios celulares y moleculares involucrados en los procesos naturales de la regulación neuroendocrino-metabólica y sus disfunciones en enfermedades tumorales y cáncer. Actualmente los estudios se centran principalmente en la detección de los factores de la maquinaria de *splicing* que más inciden en el desarrollo de diversas enfermedades, tales como el cáncer de próstata, tumores cerebrales y neuroendocrinos.
- Grupo GC-09 "*Nutrigenómica. Síndrome metabólico.*" - Inv. principal Dr. José López Miranda. Se ha colaborado en el desarrollo de modelos que detecten y expliquen los diferentes factores que influyen en el desarrollo de la diabetes mellitus tipo II.
- Grupo GC-26 "*Virología clínica y zoonosis*" - Inv. principal Dr. Antonio Rivero Román. Se realizan estudios para el diagnóstico y el diseño de estrategias de prevención de enfermedades virales (como la hepatitis) que tienen un alto riesgo en la salud de la población.
- Grupo GC-27 "*OncObesidad y metabolismo*" - Inv. principal Dr. Raúl M. Luque Huertas. Se colabora en el estudio de las bases celulares, moleculares y fisiopatológicas que influyen en el desarrollo y la progresión de enfermedades metabólicas, como la obesidad y la diabetes. Actualmente los estudios se centran principalmente en la detección de los factores de la maquinaria de *splicing* que más inciden en el desarrollo de esteatosis y diabetes mellitus tipo II.

III. Estudios y resultados

En esta sección se describen en más detalle algunos de los estudios ya realizados por el grupo y sus resultados principales, así como estudios que se están realizando y que no están concluidos.

III-A. *Metodología para la determinación de factores relevantes*

Se ha desarrollado una metodología 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. 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.

La metodología ha sido utilizada en diferentes estudios realizados en colaboración con varios de los grupos mencionados anteriormente en la Sección II-B. A continuación se describen brevemente tres de los estudios realizados que muestran la aplicación y utilidad de la metodología propuesta.

III-A1. Diagnóstico de tumores neuroendocrinos pulmonares: En colaboración con el grupo GC-08 "*Hormonas y Cáncer*"del IMIBIC se realizó un estudio para el diagnóstico de tumores neuroendocrinos pulmonares. La heterogeneidad, sus diferentes comportamientos clínicos, y la posibilidad de aparición recurrente y de 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 [26].

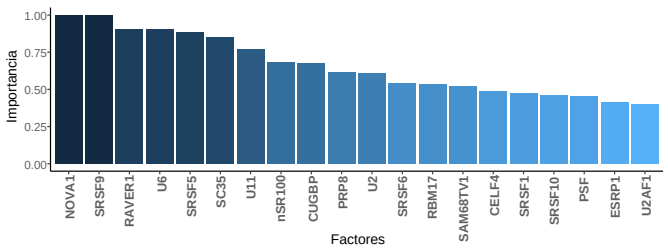

Figura 1. Ranking de factores para diferenciar entre muestras normales y tumorales.

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*. 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 1 muestra las importancia de los 20 primeros factores del ranking.

Posteriormente, en la segunda fase de la metodología 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. Tras realizar el análisis, los factores más relevantes encontrados fueron validados mediante pruebas de laboratorio.

III-A2. Aclaramiento espontáneo en Hepatitis C: En colaboración con el grupo GC-26 "*Virología clínica y zoonosis*"del IMIBIC se realizó un estudio para identificar factores o marcadores que ayuden a la predicción del aclaramiento espontáneo

o infección crónica del virus de Hepatitis C (VHC). Una vez que un paciente se infecta de 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. Sin embargo, se ha demostrado que un porcentaje menor de pacientes resuelven su infección de manera espontánea [27].

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 un aclaramiento espontáneo. Cada paciente estaba descrito por 43 marcadores distintos. A partir de la primera fase de la metodología, se obtuvo un ranking de factores, tal y como se mostró anteriomente en la Figura 1. Posteriormente, en la segunda fase de la metodología se utilizaron varios clasificadores (como árboles de decisión o clasificadores basados en reglas), obteniéndose en total casi 400 modelos distintos con un AUC $> 0{,}8$, lo cual permitió obtener una mejor estimación de la importancia de cada uno de los factores para el aclaramiento espotáneo del VHC. El hecho de utilizar árboles de decisión permitió además que los modelos resultantes fueran fácilmente interpretables por los expertos, pudiendo arribar a mejores conclusiones de una manera más sencilla. En la Figura 2 se muestra un ejemplo de los modelos obtenidos.

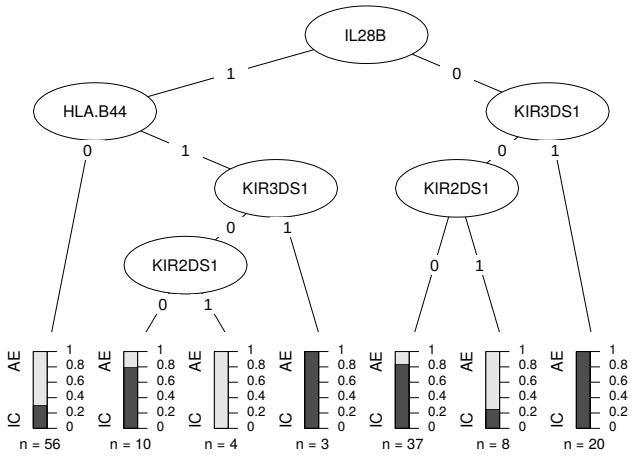

Figura 2. Árbol de decisión generado por uno de los modelos.

III-A3. Diagnóstico de diabetes mellitus tipo II: En colaboración con el grupo GC-09 "*Nutrigenómica. Síndrome metabólico*" del IMIBIC se realizó un estudio para identificar los factores que influyen en el desarrollo de diabetes mellitus tipo II. Este tipo de diabetes ha aumentado en los últimos años en todo el mundo y se ha determinado que afectó a más de 370 millones de personas en 2013. Si bien en las últimas décadas se ha producido un descendo de la mortalidad por esta enfermedad cardiovascular, la identificación de personas con alto riesgo de desarrollar diabetes es una tarea esencial. Entre los factores más conocidos se encuentran biomarcadores tradicionales no sanguíneos, factores glucémicos como glucosa en sangre y perfil de insulina y hemoglobina A1c (HbA1c), biomarcadores no glucémicos y biomarcadores genéticos.

Para este problema, se analizaron 1002 pacientes pertenecientes al ensayo clínico CORDIOPREV y se les siguió durante dos años. Se hicieron pruebas con diferentes modelos para identificar futuros pacientes con diabetes usando análisis de sensibilidad/especificidad y curvas ROC. En general, se obtuvieron modelos con niveles de predicción altos; curvas ROC con áreas superiores a 0.90. La Figura 3 muestra un árbol de decisión generado por uno de los modelos, donde se observa que Hba1c es una de las variables de predicción más importantes, más allá de los valores de glucosa en ayunas. Otra conclusión importante obtenida de los resultados es que el uso del índice IGI (función de células beta) permitió aumentar la sensibilidad y especificidad de los modelos obtenidos. Esta segunda conclusión lleva a pensar que la prueba OGTT, donde el índice IGI es obtenido, es fundamental para la correcta predicción de pacientes con alto riesgo de padecer diabetes tipo II.

III-B. Metodología para la extracción de patrones relevantes

La extracción de patrones en análisis de datos ha jugado un papel fundamental en diferentes dominios [13], pues sirven como descriptores de los datos. Estas descripciones son fundamentales para extraer información útil de los datos cuando no se posee conocimiento alguno. En ocasiones, las descripciones son realizadas sobre subconjuntos de datos dados en base a una o múltiples variables objetivo. La descripción de datos, ya sea sin conocimiento previo o basada en variables objetivo, es fundamental en Biomedicina, pues extrae relaciones desconocidas y que identifican inequívocamente a los datos.

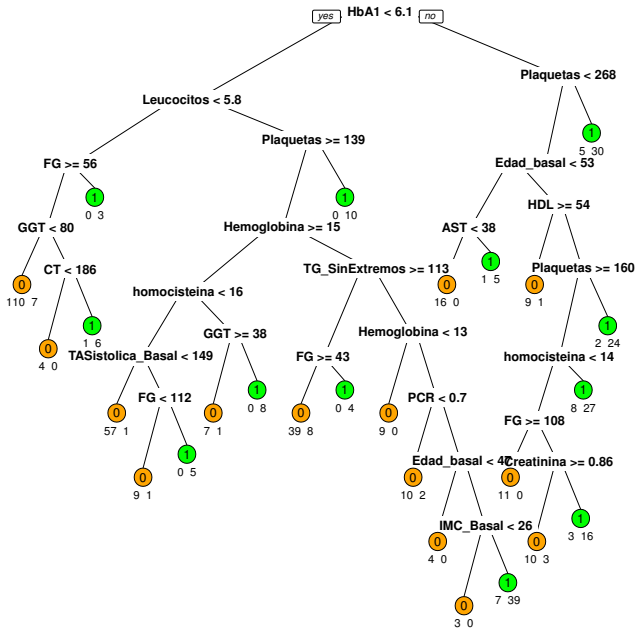

Figura 3. Árbol de decisión generado por uno de los modelos.

En la actualidad, el grupo está trabajando en dos problemas diferentes y sobre los que no se tienen aún resultados plausibles o destacados, pues se encuentran en un estado muy prematuro.

III-B1. Extracción de patrones de expresiones génicas para describir tipos de cáncer: En colaboración con el grupo GC-08 "*Hormonas y Cáncer*" del IMIBIC se está realizando una serie de estudios para identificar y describir diferentes tipos de cáncer. El objetivo de este estudio es demostrar cómo técnicas de *Supervised Descriptive Pattern Mining* [14] pueden ser útiles en la descripción de tumores y cáncer. Las técnicas utilizadas no parten de un conocimiento previo, sino que analizarán todas y cada una de las variables existentes, pudiendo dar relaciones desconocidas e imposibles de obtener por técnicas clásicas comúnmente utilizadas en Biomedicina. Los primeros estudios realizados han demostrado que, sobre bases de datos ampliamente estudiadas en la literatura, las nuevas técnicas son capaces de obtener información ya conocida, lo cual demuestra la efectividad y validez de estos métodos. Así pues, se está trabajando en la aplicación de estas mismas técnicas sobre nuevos conjuntos de datos donde las técnicas clásicas están limitadas (requieren conocimiento previo de cuáles genes deben ser analizados).

III-B2. Extracción de patrones para describir variables hemodinámicas: En colaboración con el grupo GC-07 "*Nefrología. Daño celular en la inflamación crónica*" del IMIBIC se está realizando una serie de estudios para la obtención de nuevos parámetros hemodinámicos ambulatorios y personalizados de tratamientos antihipertensivos. La hipótesis fundamental de trabajo consiste en que la aplicación de los principios de la medicina de precisión en el campo de la enfermedad cardiovascular abrirán nuevas vías de intervención individualizadas que permitirá mejorar el pronóstico de los pacientes optimizando la prescripción racional del medicamento. Se usarán nuevas técnicas diagnósticas no invasivas así como técnicas de análisis de datos que permitirán identificar relaciones no conocidas hasta el momento entre variables hemodinámicas, tratamientos y pronóstico del paciente.

IV. Conclusiones

En este trabajo se ha presentado el grupo "*Descubrimiento de Conocimiento y Sistemas Inteligentes en Biomedicina*" del IMIBIC, cuya línea de trabajo principal radica en el desarrollo de metodologías de análisis de datos para resolver problemas complejos de Biomedicina de gran relevancia para la sociedad. Se han descrito brevemente las líneas de investigación que actualmente desarrolla el grupo, y además se han presentado alguno de los estudios biomédicos en los cuales el grupo ha colaborado o que actualmente se están desarrollando, demostrando la importancia que tiene hoy en día la aplicación de técnicas modernas de ciencias de datos en los estudios biomédicos. Se espera que próximamente el grupo pueda extender su campo de acción a otros grupos de investigación biomédica del IMIBIC, así como fortalecer la colaboración con otros grupos externos.

Agradecimientos

Este trabajo ha sido financiado por el proyecto TIN2017-83445-P del Ministerio de Economía y Competitividad y Fondos FEDER.

Referencias

[1] A. Kocheturov, P. M. Pardalos, and A. Karakitsiou, "Massive datasets and machine learning for computational biomedicine: trends and challenges," *Annals of Operations Research*, pp. 1–30, 2018.

[2] C. Cao, F. Liu, H. Tan, D. Song, W. Shu, W. Li, Y. Zhou, X. Bo, and Z. Xie, "Deep learning and its applications in biomedicine," *Genomics, proteomics & bioinformatics*, 2018.

[3] N. Tempini and S. Leonelli, "Genomics and big data in biomedicine," in *Routledge Handbook of Genomics, Health and Society*. Routledge, 2018, pp. 44–51.

[4] S. M. et al., "Intelligent and effective informatic deconvolution of "big data" and its future impact on the quantitative nature of neurodegenerative disease therapy," *Alzheimer's & Dementia*, 2018.

[5] Y. Essa, G. Attiya, A. El-Sayed, and A. ElMahalawy, "Data processing platforms for electronic health records," *Health and Technology*, pp. 1–10, 2018.

[6] L. Garmire, S. Gliske, Q. Nguyen, J. Chen, S. Nemati, H. Van, D. John, J. Moore, C. Shreffler, and M. Dunn, "The training of next generation data scientists in biomedicine," in *Pacific Symposium on Biocomputing*. World Scientific, 2017, pp. 640–645.

[7] 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.

[8] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, "Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records," *BioMed research international*, vol. 2014, 2014.

[9] J. M. Luna, A. Cano, V. Sakalauskas, and S. Ventura, "Discovering useful patterns from multiple instance data," *Information Science*, vol. 357, pp. 23–38, 2016.

[10] E. Gibaja, J. M. Moyano, and S. Ventura, "An ensemble-based approach for multi-view multi-label classification," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 251–259, 2016.

[11] O. Reyes, C. Morell, and S. Ventura, "Effective lazy learning algorithm based on a data gravitation model for multi-label learning," *Information Sciences*, vol. 340, pp. 159–174, 2016.

[12] O. Reyes, A. Cano, H. Fardoun, and S. Ventura, "A locally weighted learning method based on a data gravitation model for multi-target regression," *International Journal of Computational Intelligence Systems*, vol. 11, no. 1, pp. 282–295, 2018.

[13] S. Ventura and J. M. Luna, *Pattern mining with evolutionary algorithms*. Springer, 2016.
[14] ——, *Supervised Descriptive Pattern Mining*. Springer, 2018.
[15] J. M. Luna, F. Padillo, M. Pechenizkiy, and S. Ventura, "Apriori versions based on mapreduce for mining frequent patterns on big data," *IEEE Transactions on Cybernetics*, vol. 99, pp. 1–15, 2017.
[16] F. Padillo, J. M. Luna, and S. Ventura, "Exhaustive search algorithms to mine subgroups on big data using Apache Spark," *Progress in Artificial Intelligence*, vol. 6, no. 2, pp. 145–158, 2017.
[17] J. M. Luna, M. Pechenizkiy, and S. Ventura, "Mining exceptional relationships with grammar-guided genetic programming," *Knowledge and Information Systems*, vol. 47, no. 3, pp. 571–594, 2016.
[18] S. Ventura, J. M. Luna, and A. Cano, *Big Data on Real-World Applications*. InTech, 2016.
[19] R. Salado-Cid, A. Ramîrez, and J. R. Romero, "On the need of opening the big data landscape to everyone: challenges and new trends." Berlin, Heidelberg: Springer Berlin Heidelberg, 2018, pp. 675–687.
[20] F. Padillo, J. M. Luna, and S. Ventura, "Subgroup discovery on big data: exhaustive methodologies using map-reduce," in *Proceedings of the 2016 IEEE Trustcom/BigDataSE/ISPA*, 2016, pp. 1684–1691.
[21] ——, "An evolutionary algorithm for mining rare association rules: A big data approach," in *2017 IEEE Congress on Evolutionary Computation, CEC 2017, Donostia, San Sebastián, Spain, June 5-8, 2017*, 2017, pp. 2007–2014.
[22] R. Salado-Cid and J. R. Romero, "Enabling the definition and reuse of multi-domain workflow-based data analysis," in *16th International Conference on Intelligent Systems Design and Applications (ISDA'16)*, 2016.
[23] R. Salado-Cid, J. R. Romero, and S. Ventura, "Metaherramienta para la generación de aplicaciones cientîficas basadas en workflows," in *X Jornadas de Ciencia e Ingenierîa de Servicios (JCIS'14)*, 2014, pp. 96–105.
[24] R. Salado-Cid, G. Luque, and J. R. Romero, "Sistema de gestión de flujos de trabajo para la definición visual de aplicaciones basadas en algoritmos evolutivos," in *XVI Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA'15)*, 2015, pp. 261–270.
[25] R. Salado-Cid and J. R. Romero, "Lenguaje especîfico para el modelado de flujos de trabajo aplicados a ciencia de datos," in *XXI Jornadas en Ingeniería del Software y Bases de Datos (JISBD'16)*, 2016, pp. 227–240.
[26] 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.
[27] 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.