I Workshop de Grupos de Investigación Españoles de IA en Biomedicina (IABiomed 2018)

SESIÓN 1

IASIS and BigMedilytics: Towards personalized medicine in Europe

Ernestina Menasalvas Ruiz, Alejandro Rodríguez González, Consuelo Gonzalo Martin
Centro de Tecnología Biomédica,
ETS Ingenieros Informticos,
Universidad Politécnica de Madrid
Pozuelo de Alarcón, Spain
{ernestina.menasalvas, alejandro.rg, consuelo.gonzalo}@upm.es

Massimiliano Zanin, Juan Manuel Tuñas
Centro de Tecnologa Biomédica,
Universidad Politécnica de Madrid
Pozuelo de Alarcón, Spain
{massimiliano.zanin, juan.tunas}@ctb.upm.es

Mariano Provencio, Maria Torrente, Fabio Franco, Virginia Calvo, Beatriz Nuñez
Servicio de Oncología Médica,
Hospital Universitario Puerta de Hierro
Majadahonda, Spain
{mariano.provencio, maria.torrente}@salud.madrid.org, f3franc@gmail.com, vircalvo@hotmail.com, beangarcia@gmail.com

***Abstract*—One field of application of Big Data and Artificial Intelligence that is receiving increasing attention is the biomedical domain. The huge volume of data that is customary generated by hospitals and pharmaceutical companies all over the world could potentially enable a plethora of new applications. Yet, due to the complexity of such data, this comes at a high cost. We here review the activities of the research group composed by people of the Universidad Politécnica de Madrid and the Hospital Universitario Puerta de Hierro de Majadahonda, Spain; discuss their activities within two European projects, IASIS and BigMedilytics; and present some initial results.**

***Index Terms*—Artificial Intelligence, Big Data, biomedical problems, Electronic Health Records, medical imaging**

I. Introduction

Since it was initially coined, the term Big Data is having an enormous impact in our society. It has gained such importance that governments around the world had to acknowledge its relevance in contexts such as politics, military, law, or management. Accordingly, the European Union followed this trend by creating specific associations and organizations dealing with the impact generated by Big Data and surrounding terms, such as Machine Learning, Artificial Intelligence, etc. If Big Data impacted several fields, the archetype is medicine, as it was soon understood that the incredible amount of routinely generated data could be used for very different purposes. Accordingly, the European Commission has launched several initiatives and calls aimed at funding projects with the objective of studying what are the insights that the extraction, analysis and use of medical data can provide.

The group of Mineria de Datos y Simulación (MIDAS) (Data Mining and Simulation) of the Universidad Politécnica de Madrid (UPM) has followed very closely these movements. With more than 20 years of experience in applying Data Mining (DM) techniques to several fields, the current MIDAS team started several years ago to move its research area to the biomedical domain. Its technical expertise has been complemented by the collaboration with the medical oncology department of the Hospital Universitario Puerta de Hierro de Majadahonda (HUPHM), Madrid, Spain. This has resulted in the involvement in several projects in the context of Big Data and Artificial Intelligence in the medical field, with two of them funded by the European Commission.

This contribution aims at describing the MIDAS / HUPHM team, the expertise of their members, and the techniques by them used. A special focus is given to the two European projects in which they participate, with an analysis of their objectives and characteristics. Some of the results that have been obtained so far in these projects are also presented, as well as other related initiatives.

II. The team

The meaningful application of Big Data techniques to the biomedical domain requires the convergence of two very different types of expertise: the knowledge of data managing and analysis on one hand and of the medical science on the other. This need buttressed the creation of collaboration between MIDAS and HUPHM, in which each partner contributes as described below.

A. Universidad Politécnica de Madrid

The MIDAS group is responsible of tasks related with the extraction of knowledge from medical unstructured data, both in the form of text (electronic health records) and image (CT/PET images). The main people involved are:

- Ernestina Menasalvas Ruiz: Prof. Ernestina Menasalvas is a Full Professor at the "Escuela Técnica Superior de Ingenieros Informáticos" in UPM. She is the principal investigator of the projects described below, and is in charge of the global supervision as well as of tasks in the context of data understanding of text data and of the validation of the results.
- Consuelo Gonzalo Martin: Prof. Consuelo Gonzalo is an Associate Professor at "Escuela Tcnica Superior de Ingenieros Informticos" in UPM. She is the leader of the sub-team inside MIDAS involved in image processing, analysis and understanding. Her main activities include the coordination and supervision of all tasks related to information extraction and structuring, as well as knowledge generation from CT/PET images.
- Alejandro Rodríguez González: Prof. Alejandro Rodrguez is an Associate Professor at the "Escuela Tcnica Superior de Ingenieros Informticos" in UPM and the Principal Investigator of the Medical Data Analytics Laboratory at Center for Biomedical Technology (CTB). He supervises all efforts related to Natural Language Processing (NLP) tasks.
- Massimiliano Zanin: Dr. Massimiliano Zanin is a post-doctoral researcher at Center for Biomedical Technology at UPM. His main tasks include the supervision of the technical team developing the technical pipeline, and support the work of Prof. Rodrguez and Prof. Menasalvas.
- Juan Manuel Tuñas: D. Juan Manuel Tuñas is a researcher responsible for the analysis and post-processing of the NLP pipeline results.

B. Hospital Universitario Puerta de Hierro-Majadahonda

Hospital Universitario Puerta de Hierro- Majadahonda is located in Madrid, Spain. This hospital, and more specifically, the medical oncology department, is in charge of providing the definition of the use cases and KPIs regarding the studied pathologies; the required data (electronic health records and images); and, more generally, the expertise necessary for the execution of the projects. The main people involved and their associated areas of responsibility are:

- Mariano Provencio: Medical Oncologist, Chief of Medical Oncology Department at Puerta de Hierro University Hospital, Full Professor, School of Medicine at Autónoma University of Madrid and Scientific Director of the Research Institute at Puerta de Hierro University Hospital. He is the principal investigator of the European projects described below, responsible for the Lung Cancer Pilot.
- Maria Torrente: Medical Doctor and PhD, responsible of international medicine programs in the Medical Oncology Department at Puerta de Hierro University Hospital. Associate Professor, School of Medicine, Francisco de Vitoria University, Madrid. Coordinator of national and international research projects focused on clinical oncology.
- Fabio Franco: Medical oncologist and PhD, within the Lung cancer group in the Medical Oncology Department at Puerta de Hierro University Hospital.
- Virginia Calvo: Medical oncologist and PhD, within the Lung cancer group in the Medical Oncology Department at Puerta de Hierro University Hospital.
- Beatriz Nuñez: Medical oncologist and attending physician in the Medical Oncology Department at Puerta de Hierro University Hospital.

III. The projects

As previously introduced, the MIDAS / HUPHM group is participating in various projects applying Big Data and Artificial Intelligence to the medical domain. Two of them, both funded by the H2020 programme, are described below.

A. IASIS

Integration and analysis of heterogeneous big data for precision medicine and suggested treatments for different types of patients (IASIS)[1] is a Research and Innovation Action (RIA) funded by European Commission, within its H2020 programme, and under the call "*SC1-PM-18-2016 - Big Data supporting Public Health policies*"[2]. This call targeted projects dealing with the problem of acquiring, managing, sharing, modeling, processing and exploiting huge amount of data within the medical domain, with the goal of developing solutions to support public health authorities. Aligned with the goal defined in the call, IASIS project "*seeks to pave the way for precision medicine approaches by utilizing insights from patient data. It aims to combine information from medical records, imaging databases and genomics data to enable more personalized diagnosis and treatment approaches in two disease areas - lung cancer and Alzheimer's disease*".

At the current stage of development, IASIS is primarily focusing in an application to lung cancer, being the one on the Alzheimer domain planned for the following months. For this reason, the detailed description of the project objectives is here focused on the lung cancer domain.

IASIS aims to provide answers that can be relevant and effective for the medical practitioners. The main aims regarding lung cancer include:

- Obtaining descriptive and predictive patterns to improve overall survival.
- Early detection of relapse and early palliative care initiation, and reducing overtreatments, comparing retrospective datasets with new datasets obtained from our EHR System.
- Implementation of algorithms that reduce drug-drug interactions.

[1] http://project-iasis.eu/

[2] https://cordis.europa.eu/programme/rcn/700320_en.html

- Risk stratification of lung cancer patients (treatment selection based on comorbidity index, family history, risk factors.).

In this context, the main use cases that have been defined in the lung cancer domain include:

- Identifying specific patterns in long surviving lung cancer patients, analysing all the key factors found that may associate to long survival, and compare long-survivors with the rest of the patients, in order to look for specific patterns (natural and family history, treatments, response to treatments, toxicities, comorbidities and molecular mechanisms).
- Search for risk and predictive factors for lung cancer in the study population.
- Analyse the effectiveness of tyrosin-kinase inhibitors (TKI) in mutated lung cancer patients (EGFR, ALKt, ROS-1), and look for a possible correlation between toxicities and type/duration of the TKI treatment.

The Project is coordinated by National Centre for Scientific Research "Demokritos" (NCSR) in Greece. Beyond UPM and HUPHM, additional partners include: the St. George's Hospital Medical School (UK); Alzheimer's Research (UK); Grupo español de investigación en cáncer de pulmón (Spain); the Centro de Regulación Genómica (CRG) (Spain); the university system of Maryland foundation (USA); and the Gottfried Wilhelm Leibniz Universitaet Hannover (Germany).

B. BigMedilytics

Big Data for Medical Analytics (BigMedilytics) is an Innovation Action (IA) funded by European Commission, within its H2020 programme, and under the call "*Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT)*"[3]. The projects funded under this call are also known as large-scale pilot projects. The aim of such initiative, in line with the flagship initiative 'Digital Agenda for Europe', includes "*enable[ing] Europe to support, develop and exploit the opportunities brought by ICT progress for the benefits of its citizens, businesses and scientific communities.*"

BigMedilytics goals include "*the transformation of Europe's Healthcare sector by using state-of-the-art Big Data technologies to achieve breakthrough productivity in the sector by reducing cost, improving patient outcomes and delivering better access to healthcare facilities simultaneously, covering the entire Healthcare Continuum - from Prevention to Diagnosis, Treatment and Home Care throughout Europe.*" BigMedilytics is coordinated by Philips (Netherlands) and the consortium is composed of 35 partners from 11 different countries. Due to the size of the project and the consortium, its organization is divided in several pilots, focusing on the following medical areas/diseases: comorbidities, kidney disease, diabetes, asthma/COPD, heart failure, prostate cancer, lung cancer, breast cancer, stroke, sepsis, asset management workflows and radiology workflows.

[3]https://www.cordis.europa.eu/programme/rcn/664147_en.html

As in the IASIS project, the UPM / HUPHM team is working in the lung cancer pilot. The aim and KPIs defined in this pilot differ from the IASIS project as in IASIS we are more focused in the disease and finding answers to clinical questions that may help us improve our daily clinical practice, while Bigmedilytics is focused in optimizing not only the patient's management, but also the medical oncology department's workflow by:

- Increase of early diagnosis: identification of patients at risk of developing lung cancer.
- Reducing the cost per patient (reduction of visits to ER, readmissions, reduction of toxicities).
- Reducing toxicity rates specially in complex patients.
- Improving the patients satisfaction: increasing patient's empowerement and information.

The additional partners involved in the lung cancer pilot of the BigMediliytics project include the National Centre for Scientific Research "Demokritos" (Greece), and the Gottfried Wilhelm Leibniz Universitaet Hannover (Germany).

IV. Data sources

The main dataset used in both projects is provided by HUPHM to UPM and includes an anonymized dataset containing data from the Electronic Health Records of 700 lung cancer patients (171.891 clinical notes and 7.021 clinical reports).

A. IASIS

The IASIS' lung cancer disease area involves the analysis of two different types of unstructured data:

Text: The text in the IASIS project came from the Electronic Health Records (EHR) provided by HUPHM, describing patients diagnosed with lung cancer.

Image: A basic set of images has been provided by HUPHM, for patients with nodules diagnosed as malignity/non-malignity - in a further step, the analysis will be extended to different kinds of malignity. In addition to this, several open access image data bases have been used, including the Lung Image Database Consortium image collection (LIDC-IDRI[4]), NSCLS-Radiomics[5], and LUNA[6].

B. BigMedilytics

In a similar manner, the lung cancer pilot in BigMedilytics will be executed with similar data (electronic health records provided in IASIS will be also available in BigMedilytics). The main differences in terms of data in BigMedilytics include:

- Image data are not provided: the analysis of the lung cancer information in BigMedilytics is not focused on the analysis of medical images.
- New structured data are provided, based on the specific goals and KPIs of the project:
  1) Oncology calls: The HUPHM provided a set of files containing information about a service for the

[4]https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI
[5]https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics
[6]https://luna.grand-challenge.org

telephonic attention of cancer patients. This service is not exclusively focused on lung cancer patients. The data recorded contain information about the number of calls performed by each patient and their dates, reasons, and related information.

2) Oncology app: HUPHM developed a mobile application[7] with the aim of providing cancer patients with information and personalized advices about their disease.

V. Technical work

As previously explained, both projects (IASIS and BigMedilytics) deal with three types of data:

- Textual information (Electronic Health Records),
- Medical images, and
- Structured data (Call service and mobile app).

For a better understanding of the main technological goals, the methodology associated to each one of these types is described below.

A. Textual information

The main goal of UPM in this project in terms of the textual information implies:

- To analyze structure free text. UPM is developing a framework called CliKES (Clinical Knowledge Extraction System) based on the Apache UIMA infrastructure. The system is in charge of performing most of the tasks of the classical NLP pipeline, starting with the clinical notes and reports of the patients' EHRs, to end up with a database containing all relevant information. The novelty behind the current work is based on the application of NLP techniques to EHR in Spanish, along with the creation of ad-hoc annotators focused on identifying terms referred to lung cancer domain (including treatments, mutations, etc.).
- To analyze the structured data yielded by the CliKES pipeline, to synthesize useful information for the physicians according to the use cases definition. Specifically, physicians aim at getting relevant information about possible co-occurrences in their lung cancer patients, as well as at trying to find specific correlations between them.
- To provide the results to the IASIS and BigMedilytics consortium, more specifically to the Hannover team, for the creation of knowledge graphs with the information of those patients. Both projects aim at creating a semantic-based version of the data, to allow complex queries with the processed patient data, medical literature knowledge provided by NCSR, and genetic information (in IASIS only, provided by CRG).

B. Medical images

- Structuring image data. Two types of features have been extracted from the available images: semantic and agnostic features. A python script module has been developed to extract the former ones from CT images in DICOM format. The principal tool used in this module has been Py-Radiomics [1], but also other proprietary libraries. Pre-trained Convolutional Neural Networks (CNN) models have been used for the extraction of agnostic features [3]. The most usual approach of feeding these models is by means of a sliding cube through the 3-D images; yet, this was here completely unviable from a computational point of view. In order to drastically reduce the volume of data to be processed, while minimizing the loss of information, this process has been implemented at a supervoxel level [2].
- In a future phase, the features extracted from images will be used to generate models allowing yielding useful knowledge for physician, such as predicting the survival time of lung cancer patients.
- Finally, provide the results to the IASIS consortium, more specifically to Hannover team, for the creation of knowledge graphs with the information of those patients.

C. Structured data

Two specific data sources were part of BigMedilytics' lung cancer: HUPHM Oncology call service and OncoApp mobile application. The aim of UPM is to integrate the data generated by these services (which is already in a structured form) to improve the analysis that will be performed; more specifically, to find correlations and patterns in the patients based on their clinical information and their behavior in using these services.

VI. Conclusions

As has been shown for all the types of data explained in this paper, a clear relationship is present between the aim of these projects and the application of Artificial Intelligence and Machine Learning techniques. On one hand, both projects have to deal with unstructured data (in image or text form), which require the application of complex techniques and strategies for their handling. In the case of text data, UPM is researching and developing a tool named CliKES, aimed at processing Electronic Health Records in Spanish, something that although has been under development and research by several groups in Spain, is still an on-going task. The nature of the data provided by each hospital and the corresponding processing, the problems associated to the narratives written by each physician, the identification of events, the detection of negation or acronyms in the correct context, the recognition of entities and the appropriate identification of information and the subject that belong to are, among other problems, still open problems in the field of Natural Language Processing, and this despite the large amount of work in the field. Here it is important to emphasize how the machine learning techniques play a very important role in several of the tasks of the NLP pipeline, and how important is to find accurate models to create accurate NLP systems.

Finally, both projects have to deal with structured data. In this context, UPM is mainly working on the application of Data Mining techniques to find important insights and

[7] https://play.google.com/store/apps/details?id=org.idiphim.oncoapp

evidences within the data. The amount of data, as well as its diversity (textual, image, call service, mobile application), requires huge efforts in terms of structuring, processing and cleaning. These efforts are done with the objective of having data with enough quality, to subsequently apply the correct data mining techniques and finding evidences based on the use cases, these latter defined by the physicians as well as the associated KPIs.

A comprehensive characterization of lung cancer tumor signatures is critical for a correct diagnosis and optimal treatments. As precision medicine is practiced more widely, one of the main challenges is the integration and analysis of clinical data, opening new opportunities for more accurate diagnosis, more sensitive and frequent disease monitoring and more personalized therapeutic strategies, at the level of the individual.

Acknowledgment

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 727658, project IASIS (Integration and analysis of heterogeneous big data for precision medicine and suggested treatments for different types of patients) and by the European Union's Horizon 2020 innovation programme under grant agreement No. 780495, project BigMedilytics (Big Data for Medical Analytics).

References

[1] J.J. Van Griethuysen, A. Fedorov, C. Parmar, N. Aucoin, V. Narayan, R.G.H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H.J.W.L. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype", Cancer Res., vol. 77, pp. e104–e107, November 2017.

[2] C. Gonzalo-Martín, A. García-Pedrero, M. Lillo-Saavedra and E. Menasalvasa, "Deep Learning for Superpixel-Based Classification of Remote Sensing Images", GEOBIA 2016: Solutions and Synergies, 2016.

[3] K.H. Cha, L. Hadjiiski, R.K. Samala, H. Chan, E.M. Caoili, and R.H. Cohan, "Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets", Med. Phys., vol. 43, pp. 1882–1896, 2016.

Tecnologías para el Modelado, Procesamiento y Gestión de Conocimiento

Jesualdo Tomás Fernández Breis
Universidad de Murcia, IMIB-Arrixaca
Murcia, España
jfernand@um.es

Marcos Menárguez Tortosa
Universidad de Murcia, IMIB-Arrixaca
Murcia, España
marcos@um.es

Catalina Martínez Costa
Medical Graz University
Graz, Austria
catalina.martinez@medunigraz.at

José Antonio Miñarro Giménez
Medical Graz University
Graz, Austria
jose.minarro-gimenez@medunigraz.at

María del Carmen Legaz García
FFIS, IMIB-Arrixaca
Murcia, España
mcarmen.legaz@ffis.es

Manuel Quesada Martínez
Universidad Miguel Hernández
Elche, España
mquesada@umh.es

Astrid Duque Ramos
Universidad de Antioquia
Medellín, Colombia
astrid.duquer@udea.edu.co

Ángel Esteban Gil
FFIS, IMIB-Arrixaca
Murcia, España
angel.esteban@ffis.es

Dagoberto Castellanos Nieves
Universidad de La Laguna
Tenerife, España
dcastell@ull.es

***Abstract*—El grupo de investigación *Tecnologías para el Modelado, Procesamiento y Gestión de Conocimiento* lleva quince años investigando, desarrollando y aplicando tecnologías semánticas en dominios biomédicos. En este trabajo se describen las líneas de investigación más relevantes en las que hemos trabajado en los últimos años, y cuyo objetivo principal es la consecución de interoperabilidad semántica entre sistemas de información sanitarios. Asimismo, también se comentarán objetivos de investigación para los próximos años en los que las técnicas inteligentes desempeñarán un papel importante.**

***Index Terms*—Artificial Intelligence, Knowledge Engineering, Electronic Medical Records**

I. Presentación del grupo de investigación

El grupo *Tecnologías para el Modelado, Procesamiento y Gestión de Conocimiento* (TECNOMOD) se creó en la Universidad de Murcia en el año 2003 constituyendo la ingeniería ontológica, la web semántica y las tecnologías del lenguaje sus áreas principales de investigación. Actualmente el grupo de investigación tiene 15 integrantes (profesores, becarios y contratados), y cuenta con colaboradores a nivel nacional e internacional, incluyendo antiguos doctorandos del grupo cuya contribución a los resultados que se presentan en este trabajo ha sido fundamental. La mayoría de los integrantes del grupo de investigación son informáticos, si bien disponemos de dos investigadores cuya formación y actividad es sanitaria y un biotecnólogo. Hemos aplicado nuestra investigación en diversos dominios como el turismo, las finanzas, la política o la educación, pero la biomedicina ha sido el área donde más investigación hemos realizado. En este documento nos ceñiremos a la investigación realizada exclusivamente en este ámbito, y especialmente destacaremos dos líneas: interoperabilidad semántica y aseguramiento de la calidad de ontologías y terminologías biomédicas.

TECNOMOD es miembro del Instituto Murciano de Investigación Biosanitaria (IMIB-Arrixaca), que está acreditado por el Instituto de Salud Carlos III. IMIB-Arrixaca está vinculado al Hospital Universitario Virgen de la Arrixaca. Tenemos colaboraciones en marcha con varios grupos de investigación del instituto y una colaboración permanente con la Plataforma de Informática Biomédica y Bioinformática del mismo.

II. Contexto

La Web Semántica [2] es un espacio natural para la integración de datos y la interoperabilidad semántica entre sistemas, imponiendo un entorno de trabajo en el que cada sistema emplea el significado de los datos en diferentes contextos [10]. Las tecnologías semánticas posibilitan la descripción del contexto lógico de la información a intercambiar, mientras que se permite que cada sistema mantenga su máxima independencia. Las ontologías constituyen el nivel fundamental de la Web Semántica desde el punto de vista de representación formal del conocimiento, de ahí que parte del éxito de la Web Semántica recaiga en la calidad de las ontologías, por lo que el desarrollo de métodos que permita asegurar la calidad de las mismas es un objetivo crítico para entornos de interoperabilidad. En los últimos años, las tecnologías de la Web Semántica han ganado popularidad para la consecución de interoperabilidad semántica entre sistemas de información sanitarios, especialmente desde que el proyecto Semantic Health [24] recomendó su uso para dichos fines. Posteriormente, la FP7 Network of Excellence SemanticHealthNet[1] propuso que la formalización ontológica debería ser fundamental para permitir el intercambio y la cooperación entre los sistemas de historia clínica

[1] http://www.semantichealthnet.eu

electrónica (HCE) y los sistemas de ayuda a la decisión (SAD). Dicho rol de las ontologías en escenarios de interoperabilidad impone una serie de requisitos sobre las ontologías [8]:

- Facilitar la representación, compartición, reutilización de conocimiento para modelos de información y de inferencia.
- Clasificar y recuperar datos de HCE según las reglas establecidas en guías y protocolos.
- Tener garantizada su calidad formal y estructural.

En este trabajo agrupamos nuestras líneas de investigación en dos grupos:

- Interoperabilidad semántica en salud (Sección III): investigación orientada a conseguir el intercambio de información entre sistemas sanitarios, incluyendo HCE y SAD.
- Aseguramiento de calidad de ontologías (Sección IV): investigación orientada a evaluar el cumplimiento de los requisitos de interoperabilidad por parte de las ontologías y terminologías biomédicas.

La investigación que se describirá en las próximas secciones ha sido posible gracias a financiación fundamentalmente pública. A continuación se enumeran los proyectos más relevantes de los últimos diez años.

- **Plataforma para la adquisición y compartición de información y conocimiento para comunidades de investigación clínica en red II** (TSI2007-66575-C02-02). Ministerio de Educación y Ciencia. 01/10/2007-31/12/2010.
- **Herramientas inteligentes para enlazar historias clínicas electrónicas y sistemas de ensayos clínicos II** (TIN2010-21388-C02-02). Ministerio de Ciencia e Innovación. 01/01/2011-31/12/2014.
- **Modelos de información y conocimiento clínicos para enlazar los sistemas de historia clínica electrónica y de ayuda a la decisión clínica II** (TIN2014-53749-C2-2-R). Ministerio de Economía y Competitividad. 01/01/2015-31/12/2018.
- **Semantic Interoperability for Health Network**.Unión Europea. 03/10/2012-30/11/2014. Red de Excelencia del programa FP7 en la que miembros de TECNOMOD participaron como expertos externos en el paquete de trabajo 4 *"Harmonised resources for EHR/PHR and aggregation"*.
- **Gene regulation ensemble effort for the knowledge commons** (CA COST Action CA15205). Unión Europea. 08/09/2016 - 07/09/2020. Participación de TECNOMOD como líder del paquete de trabajo de ontologías y vocabularios controlados.
- **Unraveling in utero determinants predicting lung function in infants: a step for prenatal prevention of asthma** (PIE15/00051). Instituto de Salud Carlos III. 01/01/2016-31/12/2018.
- **Infraestructura y tecnologías de interoperabilidad para aplicaciones de Learning Health Systems I** (TIN2017-85949-C2-1-R). Ministerio de Economía, Industria y Competitividad. 01/01/2018 - 31/12/2020.

III. Interoperabilidad semántica en salud

A. Transformación de datos y modelos clínicos

Las recomendaciones del proyecto SemanticHealth incluyeron el uso combinado de estándares de historia clínica electrónica, ontologías y terminologías biomédicas para facilitar el intercambio significativo de datos. Nuestro grupo de investigación desarrolló PoseacleConverter [2] como demostrador de las posibilidades de intercambio de datos entre distintos formatos de modelos clínicos (openEHR, ISO 13606 o CEM) y de datos entre openEHR e ISO 13606.

El trabajo comenzó en el año 2008 y el primer objetivo fue la transformación de arquetipos (como tipo de modelo clínico) entre los estándares openEHR e ISO 13606 [14]. La transformación de modelos se basó en la formalización de correspondencias entre las ontologías que representaban los modelos de información de los estándares. Esto permitió definir unas reglas que guiaron la transformación estructural de los arquetipos. Este método se extendió para transformar datos entre dichos estándares explotando las correspondencias anteriores [13]. El uso de ontologías permitió también comprobar en ambos esfuerzos la consistencia lógica de los modelos y datos obtenidos mediante el uso de razonadores. La figura 1 describe el proceso seguido a nivel de datos. Se puede apreciar en la parte inferior el extracto de datos fuente ISO 13606, que es transformado en un extracto openEHR usando las correspondencias definidas a partir de las ontologías. Para la aplicación de la transformación de datos no es necesaria la existencia del arquetipo destino, el cual se crea automáticamente. Esto genera una traza de transformación que es empleada para la generación del extracto destino.

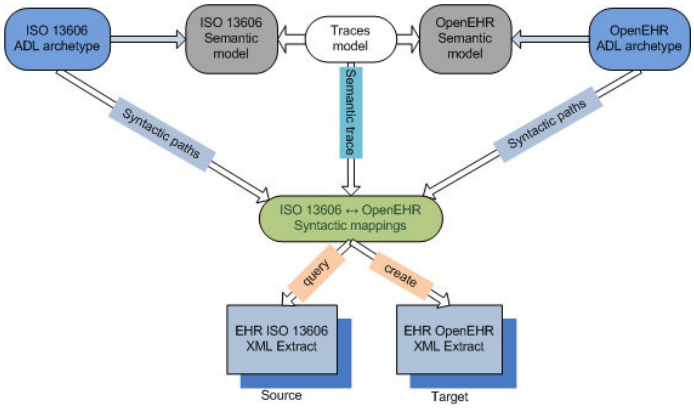

Fig. 1. Esquema de transformación de datos

Posteriormente el enfoque de transformación de modelos clínicos evolucionó cuando diseñamos el proceso para obtener arquetipos openEHR a partir de modelos clínicos CEM [11]. El nuevo enfoque se desarrolló exclusivamente con tecnologías OWL. En este caso las relaciones entre los modelos de

[2] miuras.inf.um.es/PoseacleConverter

referencia se expresaron como axiomas OWL y se crearon plantillas de transformación. De esta manera, la transformación se concibió como un proceso generativo de axiomas OWL, y el proceso estaba asistido por un razonador. La figura 2 ejemplifica este proceso de generación: (1) representación OWL del modelo clínico CEM; (2 y 3) plantillas de generación para los dos items incluidos en el modelo; (4) ontología del modelo de información openEHR, que se utiliza para la definición de las plantillas; (5, 6 y 7) representación parcial en OWL del modelo transformado a openEHR.

Estos métodos se evaluaron y validaron con las colecciones de arquetipos openEHR disponibles en CKM y modelos CEM disponibles en la librería de Intermountain disponibles en su momento.

B. Validación semántica de arquetipos

El intercambio automático de datos a través de sistemas informáticos basado en modelos clínicos, como pueden ser los arquetipos, requiere la corrección formal de los mismos. Nuestro grupo de investigación desarrolló Archeck[3] [15], que un método basado en OWL para la validación de arquetipos. Los arquetipos son representados como clases OWL y la validación semántica de los mismos se realiza mediante el uso de razonadores automáticos. Para ello usa una ontología que define los tipos de restricciones que se pueden definir para un arquetipo y que vienen especificados en el modelo de arquetipos. Además, Archeck extiende esta ontología para poder definir métricas de calidad para arquetipos basadas en OWL. ¡ El método Archeck se aplicó a dos repositorios públicos de arquetipos openEHR, el repositorio gestionado en la herramienta Clinical Knowledge Manager (CKM) y el repositorio de programa National Health Service (NHS) del Reino Unido. La evaluación de los repositorios se centró en los arquetipos que definen relaciones de especialización, 81 en CKM y 212 en NHS. La validación de los repositorios encontró que un 22,2% de los arquetipos especializados en el repositorio CKM y un 21,2% en NHS contenían errores. De ellos, el 3% de los errores en CKM fueron identificados gracias a la métrica de calidad descrita anteriormente. Si bien puede considerarse un valor bajo esto también se debió a que los muchos de los arquetipos evaluados no contenían enlaces terminológicos, lo cual es un indicador negativo de calidad.

C. Interoperabilidad de modelos de información, dominio e inferencia

El trabajo en esta línea es el fruto de proyectos coordinados con la Universidad Politécnica de Valencia (UPV) y la Universidad Jaime I (UJI). UPV es experto en modelos de información, UJI en modelos de inferencia y nuestro grupo en modelos de dominio. El trabajo realizado en esta línea pretende conseguir aplicar las guías clínicas computerizadas a datos estandarizados de la HCE. Para ello se presentan las ontologías como elemento mediador que proporcione la semántica compartida entre ambos modelos.

1) Flujos de transformaciones interoperables. Aplicación a la clasificación de pacientes: El primer trabajo en esta línea afrontó cómo abordar problemas de clasificación de pacientes a partir de protocolos estandarizados mediante las tecnologías de estándares de HCE y las ontologías [7]. Esto determinó un flujo de trabajo cuya entrada son datos no normalizados de HCE y la salida es la clasificación del paciente, tal y como se puede ver en la figura 3. En este caso trabajamos también en colaboración con el Programa de cribado de cáncer de colon y recto de la Región de Murcia. El objetivo específico era determinar el nivel de riesgo de los pacientes del programa. Para ello trabajamos con las guías europea y americana de cáncer de colon y recto. Las tecnologías semánticas contribuyen en dos etapas del proceso, como son la representación de los datos en formato semántico y la clasificación del paciente usando razonadores. Para ello también fue necesario formalizar las reglas de los protocolos en OWL. Este trabajo se sigue desarrollando a través de la plataforma CLIN-IK-LINKS [12], que busca generalizar y facilitar la definición y ejecución de transformaciones de datos interoperables.

2) Enriquecimiento de guías clínicas computerizadas: Una línea reciente de trabajo busca formalizar el conocimiento de las guías clínicas computerizadas (GCC) con conceptos existentes en ontologías y terminologías biomédicas, trabajo que se está realizando con colaboración con la UJI. El primer trabajo realizado ha sido identificar los conceptos de ontologías de BioPortal a partir de un conjunto de GCC seleccionado y disponible en formato PROforma [9]. SNOMED CT fue el recurso semántico con mayor número de resultados, por lo que es la que estamos usando en los estudios y evaluaciones preliminares del método desarrollado [22]. La figura 4 muestra cómo procesamos la GCC. Una vez extraído el texto a procesar, se buscan alineamientos usando métodos basados en OntoEnrich (ver Sección IV-B). El resultado es un conjunto de recomendaciones de conceptos asociados a la guía clínica.

IV. Aseguramiento de calidad de terminologías y ontologías

A. Evaluación de calidad de ontologías

OQuaRE[4] [5] es un *framework* para la evaluación de ontologías basado en la norma ISO/IEC 25000:2005 para la definición de requisitos y evaluación de la calidad de productos software, también conocida como SQuaRE [1]. OQuaRE realiza la evaluación de la calidad de ontologías con tres niveles de granularidad: características, subcaracterísticas y métricas de calidad. Actualmente el modelo de calidad de OQuaRE incluye 8 características, 29 subcaracterísticas y 19 métricas. Cada característica tiene un conjunto de subcaracterísticas asociadas que, a su vez, tienen asociadas un conjunto de métricas. Toda la información sobre el modelo de calidad está disponible en el sitio web de información sobre OQuaRE[5].

Las métricas de OQuaRE tienen una función $f(x)$ cuyo dominio es una ontología, pero el rango de cada métrica puede

[3] miuras.inf.um.es/archeck

[4] http://sele.inf.um.es/oquare

[5] http://miuras.inf.um.es/oquarewiki

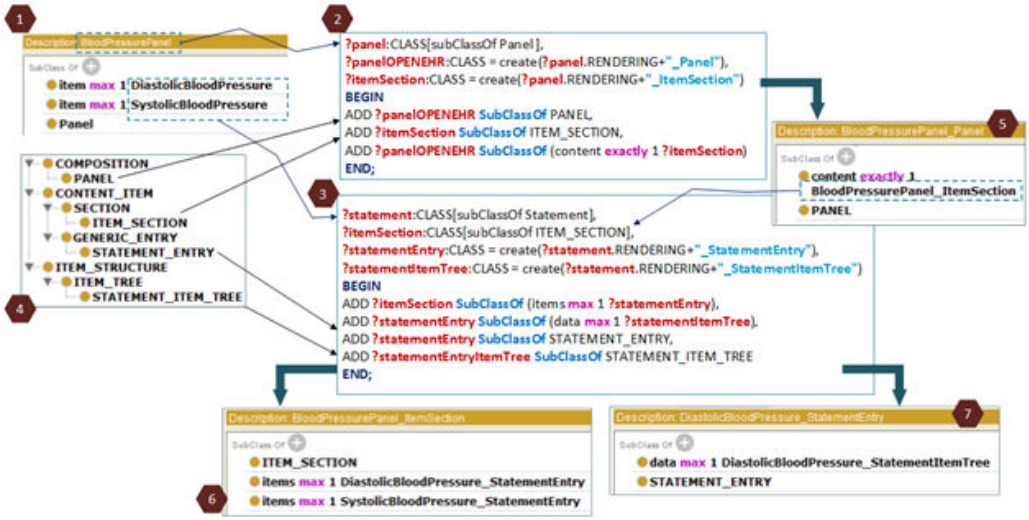

Fig. 2. Esquema de transformación de datos

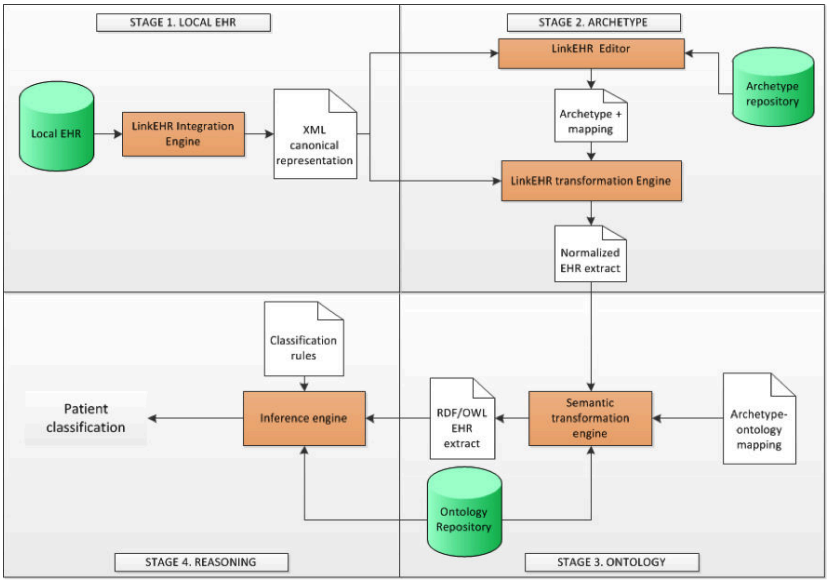

Fig. 3. Flujo de transformación de datos

variar. Tener una visión global de la evaluación de una ontología requiere combinar los resultados de todas las métricas. Es por ello que OQuaRE define dos tipos de funciones de escalado de métricas, que permiten abstraer al método de los distintos rangos de cada métrica:

- **Estática**, basada en recomendaciones y buenas prácticas. $n(f(x))$ está predefinida y está basada en intervalos fijos continuos que particionan el rango de $f(x)$ en k categorías.
- **Dinámica**, que usa datos experimentales como referencia. $n(f(x))$ aplica el método de clustering k-means para particionar el rango de $f(x)$ en k intervalos continuos no prefijados que contienen todas las observaciones incluidas en los datos experimentales. Para maximizar la compactación de las ontologías de cada categoría, minimizando la varianza intra-cluster, y para maximizar la separación entre las categorías, maximizando la varianza inter-cluster en cada iteración se recalculan los nuevos k centroides y la asignación de cluster se realiza asociando cada R_{θ_j} al centroide más cercano. El algoritmo de clustering requiere la información de si los valores altos de $f(x)$ se corresponden con las categorías más altas del factor. Esto se debe a que valores altos de una métrica no tienen por qué representar siempre una buena propiedad de una ontología.

Además, cabe mencionar que la versión actual de OQuaRE usa $k = 5$ para todas las métricas. La escala estática de OQuaRE ha sido aplicada para analizar ontologías [3], mientras que la dinámica ha sido empleada para estudiar la evolución de ontologías. Para ello se han usado como datos

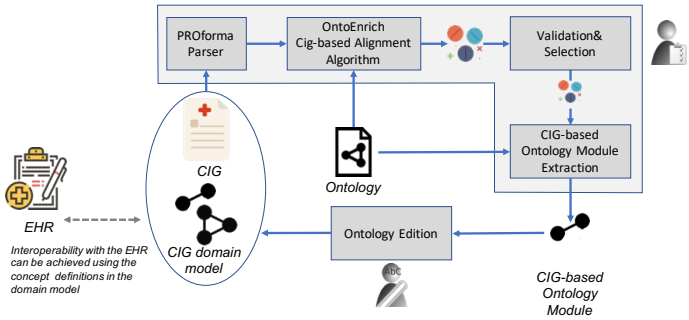

Fig. 4. Diagrama de procesamiento del contenido de la guía clínica computerizada

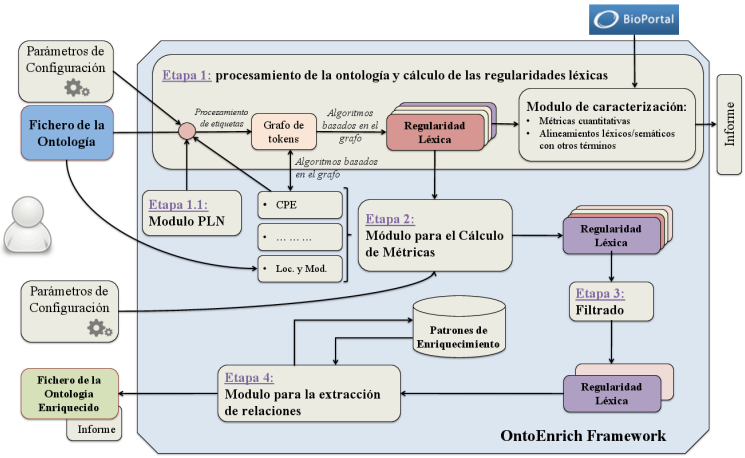

Fig. 5. Descripción de la metodología aplicada por OntoEnrich

experimentales los valores de las métricas para cada versión de la ontología procesada [6], [17]. El conjunto de datos experimentales también puede estar compuesto por una versión de cada ontología, sirviendo entonces para medir similitudes y diferencias entre ellas.

B. Enriquecimiento de ontologías

OntoEnrich [6] es un método para el enriquecimiento de ontologías biomédicas basada en el análisis léxico de sus etiquetas [16], ya que las ontologías son ricas en contenido léxico para humanos, el cual no se pone siempre a disposición de las máquinas en forma de axiomas lógicos. Por tanto, el objetivo principal de OntoEnrich es explotar la semántica oculta de las ontologías [25]. OntoEnrich se basa en el concepto de regularidad léxica (RL), que se define como un conjunto de tokens consecutivos repetidos en diferentes etiquetas de la ontología. Cabe mencionar que la entrada habitual a OntoEnrich es una ontología en formato OWL, si bien en distintos casos de uso la herramienta que implementa la metodología ha sido adaptada para aceptar otros tipos de entradas, como ficheros de texto donde cada línea incluya cada etiqueta a analizar. La figura 5 muestra las principales etapas de la metodología:

- **Procesamiento de la ontología y cálculo de regularidades léxicas**. En primer lugar permite obtener el conjunto completo de regularidades léxicas (RL) utilizando como parámetro de entrada un *threshold* mínimo que permite podar las búsquedas. También se establecen relaciones super-sub entre regularidades léxicas. Se realiza también una descripción cuantitativa de cada RL. Se buscan alineamientos a partir de las RD en la propia ontología o en otras externas para promover la reutilización de conceptos entre la comunidad biomédica, y estos algoritmos buscan clases que contengan la RL o cuya etiqueta coincida con la RL [21].
- **Cálculo de métricas avanzadas**. En esta etapa se calculan métricas avanzadas que relacionan las RL con diferentes aspectos semánticos de la ontología, como localización o modularidad de las RL y los productos cruzados, que informa sobre la potencialidad de enriquecimiento de una RL.
- **Filtrado de regularidades**. OntoEnrich permite utilizar las métricas para definir filtros que reducen el conjunto de RL a aquellas que cumplen ciertas propiedades. Se podría filtrar a aquellas RL que poseyeran determinado valor de modularidad, localización, etc. También se pueden filtrar las subregularidades para quedarnos con las regularidades más largas, ya que esto quiere decir que son más específicas.
- **Extracción de relaciones y axiomas**. En este paso se permite la definición de patrones axiomáticos a partir de las RL anteriores. El método es capaz de crear automáticamente patrones a partir de relaciones de subclase. Como resultado de la ejecución de dichos patrones se obtendría la ontología enriquecida.

C. Aplicaciones basadas en OQuaRE y OntoEnrich

Los métodos OQuaRE y OntoEnrich se han aplicado para analizar las ontologías de repositorios de ontologías biomédicas como OBO Foundry y BioPortal con el objeto de caracterizar dichas ontologías desde el punto de vista de optimalidad para su uso en escenarios de interoperabilidad. Los aspectos específicos que han contribuido a estudiar han sido:

- Evaluación de ontologías [4], [6], [17].
- Evaluación de la aplicación de principios de diseño de ontologías, como delineación de contenido, riqueza de relaciones y nombrado sistemático [20], [21].
- Análisis de reutilización de axiomas y axiomas ocultos en ontologías [18].
- Enriquecimiento de la Gene Ontology [23].
- Detección de axiomas potencialmente erróneos y detección de posibles axiomas en SNOMED CT [19], [26].

V. Próximos objetivos

En estas líneas de investigación pretendemos aplicar a corto plazo técnicas de inteligencia artificial para los siguientes

[6]http://sele.inf.um.es/ontoenrich

objetivos:

- Desarrollar el concepto de *Learning Health Systems* incluyendo técnicas de machine learning dentro de los flujos de procesamiento de datos interoperables.
- Analizar ontologías a partir de conjuntos de datos de HCE.
- Extracción de módulos de ontologías para el enriquecimiento de guías clínicas computerizadas y de otras ontologías.
- Aprendizaje del valor óptimo de k para cada métrica empleada en OQuaRE en la función de escalado dinámica.
- Clasificación y analizar tipos y grupos de ontologías a partir de los valores de sus métricas.
- Identificación de relaciones entre métricas de las ontologías y principios de diseño.
- Aprendizaje automático de patrones axiomáticos a partir de regularidades léxicas.

Agradecimientos

Este trabajo ha sido posible gracias a la financiación del Ministerio de Educación y Ciencia (TSI2007-66575-C02-02), Ministerio de Ciencia e Innovación (TIN2010-21388-C02-02), Ministerio de Economía, Industria y Competitividad (TIN2014-53749-C2-2-R, TIN2017-85949-C2-1-R), así como el Fondo Europeo de Desarrollo Regional (FEDER) a través de los proyectos citados.

Referencias

[1] ISO/IEC 25000:2005, Software Engineering - Software Product Quality Requirements and Evaluation (SQuaRE) - Guide to SQuaRE, 2005. [Online; accessed 01-June-2017].

[2] T. Berners-Lee, J. Hendler, O. Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[3] A. Duque-Ramos, M. Boeker, L. Jansen, S. Schulz, M. Iniesta, and J. T. Fernández-Breis. Evaluating the good ontology design guideline (goodod) with the ontology quality requirements and evaluation method and metrics (oquare). *PLOS ONE*, 9(8):1–14, 08 2014.

[4] A. Duque-Ramos, J. T. Fernández-Breis, M. Iniesta, M. Dumontier, M. E. Aranguren, S. Schulz, N. Aussenac-Gilles, and R. Stevens. Evaluation of the oquare framework for ontology quality. *Expert Systems with Applications*, 40(7):2696–2703, 2013.

[5] A. Duque-Ramos, J. T. Fernández-Breis, R. Stevens, and N. Aussenac-Gilles. OQuaRE: A SQuaRE-based approach for evaluating the quality of ontologies. *Journal of Research and Practice in Information Technology*, 43(2):159–176, 2011.

[6] A. Duque-Ramos, M. Quesada-Martínez, M. Iniesta-Moreno, J. T. Fernández-Breis, and R. Stevens. Supporting the analysis of ontology evolution processes through the combination of static and dynamic scaling functions in oquare. *Journal of Biomedical Semantics*, 7(1):63, 2016.

[7] J. T. Fernández-Breis, J. A. Maldonado, M. Marcos, M. d. C. Legaz-García, D. Moner, J. Torres-Sospedra, A. Esteban-Gil, B. Martínez-Salvador, and M. Robles. Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. *J Am Med Inform Assoc*, 20(e2):e288–96, Dec 2013.

[8] J. T. Fernández-Breis, M. Quesada-Martínez, and A. Duque-Ramos. Can existing biomedical ontologies be more useful for ehr and cds? In *International Workshop on Knowledge Representation for Health Care*, pages 3–20. Springer, 2016.

[9] J. Fox, N. Johns, and A. Rahmanzadeh. Disseminating medical knowledge: the proforma approach. *Artif Intell Med*, 14(1-2):157–81, 1998.

[10] C. Goble and R. Stevens. State of the nation in data integration for bioinformatics. *Journal of biomedical informatics*, 41(5):687–693, 2008.

[11] M. d. C. Legaz-García, M. Menárguez-Tortosa, J. T. Fernández-Breis, C. G. Chute, and C. Tao. Transformation of standardized clinical models based on owl technologies: from cem to openehr archetypes. *J Am Med Inform Assoc*, 22(3):536–44, May 2015.

[12] J. A. Maldonado, M. Marcos, J. T. Fernández-Breis, E. Parcero, D. Boscá, M. D. C. Legaz-García, B. Martínez-Salvador, and M. Robles. A platform for exploration into chaining of web services for clinical data transformation and reasoning. *AMIA Annu Symp Proc*, 2016:854–863, 2016.

[13] C. Martínez-Costa. *Modelos de representación y transformación para la interoperabilidad semántica entre estándares de Historia Clínica Electrónica basados en arquitectura de modelo dual*. PhD thesis, Universidad de Murcia, 2011.

[14] C. Martínez-Costa, M. Menárguez-Tortosa, and J. T. Fernández-Breis. An approach for the semantic interoperability of iso en 13606 and openehr archetypes. *J Biomed Inform*, 43(5):736–46, Oct 2010.

[15] M. Menárguez-Tortosa and J. T. Fernández-Breis. Owl-based reasoning methods for validating archetypes. *Journal of biomedical informatics*, 46(2):304–317, 2013.

[16] M. Quesada-Martínez. *Methodology for the enrichment of biomedical knowledge resources*. PhD thesis, Depto. de Informática y Sistemas. Univ. de Murcia, 2015.

[17] M. Quesada-Martínez, A. Duque-Ramos, M. Iniesta-Moreno, and J. T. Fernández-Breis. Preliminary analysis of the obo foundry ontologies and their evolution using oquare. *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, 235:426, 2017.

[18] M. Quesada-Martínez and J. T. Fernández-Breis. Studying the reuse of content in biomedical ontologies: An axiom-based approach. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 3–13. Springer, 2017.

[19] M. Quesada-Martínez, J. T. Fernández-Breis, and D. Karlsson. Suggesting missing relations in biomedical ontologies based on lexical regularities. *Stud Health Technol Inform*, 228:384–8, 2016.

[20] M. Quesada-Martínez, J. T. Fernández-Breis, and R. Stevens. Lexical characterization and analysis of the bioportal ontologies. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 206–215. Springer Berlin Heidelberg, 2013.

[21] M. Quesada-Martínez, J. T. Fernández-Breis, and R. Stevens. Lexical characterisation of bio-ontologies by the inspection of regularities in labels. *Current Bioinformatics*, 10(2):165–176, 2015.

[22] M. Quesada-Martínez, M. Marcos, F. Abad-Navarro, B. Martínez-Salvador, and J. T. Fernández-Breis. Towards the semantic enrichment of computer interpretable guidelines: a method for the identification of relevant ontological terms. In *AMIA Annual Symposium Proceedings*, 2018.

[23] M. Quesada-Martínez, E. Mikroyannidi, J. T. Fernández-Breis, and R. Stevens. Approaching the axiomatic enrichment of the gene ontology from a lexical perspective. *Artificial intelligence in medicine*, 65(1):35–48, 2015.

[24] V. Stroetman, D. Kalra, P. Lewalle, A. Rector, J. Rodrigues, K. Stroetman, G. Surjan, B. Ustun, M. Virtanen, and P. Zanstra. Semantic Interoperability for Better health and Safer Healthcare [34 pages]. (January), 2009.

[25] A. Third. Hidden semantics: what can we learn from the names in an ontology? In *Proceedings of the Seventh International Natural Language Generation Conference*, pages 67–75. Association for Computational Linguistics, 2012.

[26] P. van Damme, M. Quesada-Martínez, R. Cornet, and J. T. Fernández-Breis. From lexical regularities to axiomatic patterns for the quality assurance of biomedical terminologies and ontologies. *J Biomed Inform*, Jun 2018.

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].

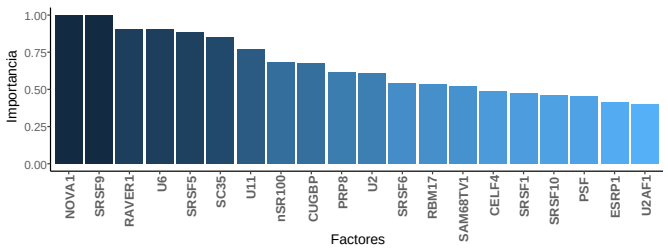

Figura 1. Ranking de factores para diferenciar entre muestras normales y tumorales.

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.

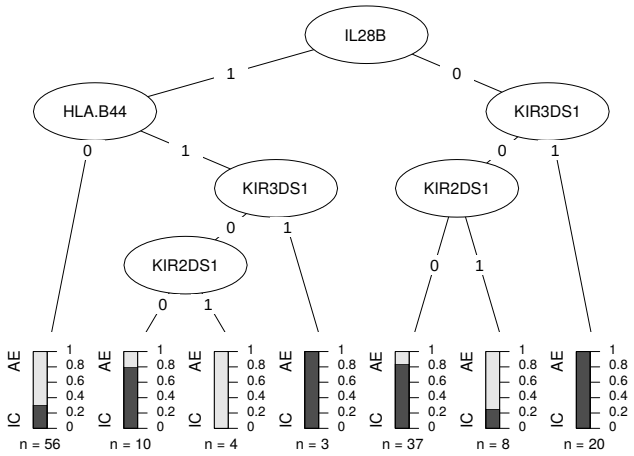

Figura 2. Árbol de decisión generado por uno de los modelos.

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.

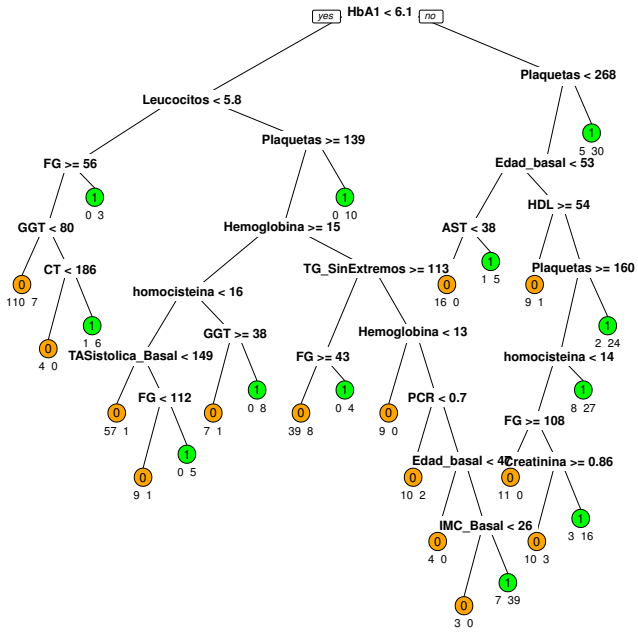

Figura 3. Árbol de decisión generado por uno de los modelos.

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.

Una aproximación a la interpretación del electrocardiograma desde la perspectiva de la Inteligencia Artificial

Paulo Félix
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela
15782 Santiago de Compostela, SPAIN
paulo.felix@usc.es

***Resumen*—En este trabajo se revisa la trayectoria científica reciente del Centro Singular de Investigación en Tecnoloxías da Información de la Universidade de Santiago de Compostela en la interpretación electrocardiográfica desde la perspectiva de los métodos, técnicas y herramientas asimilables a la Inteligencia Artificial.**

***Index Terms*—Series temporales estocásticas, Electrocardiograma, Agrupamiento, Clasificación, Regresión, Abducción**

I. Introducción

El electrocardiograma (ECG) es una prueba diagnóstica de bajo coste que consiste en el registro de la actividad eléctrica del corazón a partir de la medición de la diferencia de potencial entre un conjunto de electrodos colocados en la superficie de la piel del paciente. La interpretación del ECG plantea numerosos retos absolutamente fascinantes desde múltiples puntos de vista. Desde un punto de vista *asistencial*, se pretende la generalización de su uso en la prevención, diagnóstico y seguimiento de la enfermedad cardiovascular, primera causa de muerte en el mundo [1]. Desde un punto de vista *científico*, en tanto que se ha convertido en una fuente todavía sin agotar de información sobre multitud de procesos fisiopatológicos que se manifiestan de alguna manera en el ECG, y así, disciplinas como la neumología, la obstetricia, la neurología o incluso la psiquiatría, buscan evidencias en el ECG que permitan un abordaje sencillo y precoz en múltiples y variados trastornos. Desde un punto de vista *tecnológico*, en busca de una instrumentación más pequeña, más fiable, con mayor capacidad de interpretación, más eficiente y más autónoma, que pueda realizar un registro de la manera más inadvertida posible, como parte de los dispositivos que por su comodidad pueden considerarse vestibles. Desde un punto de vista *cognitivo*, en tanto que la interpretación del ECG que realiza un experto cardiólogo ha de poner en juego un conjunto de procesos mentales como la percepción, la memoria, el aprendizaje o la adaptación que son objeto de constante revisión con la experiencia, en un problema que podríamos calificar como de parcialmente observable, y para el que no existen criterios de consenso, tal y como pone de manifiesto la diferente interpretación que realizan distintos expertos sobre el mismo registro [2].

Son estos retos cognitivos los que aquí interesan, desde la perspectiva de la Inteligencia Artificial, con el objetivo de dotar a la tecnología de una mayor capacidad de interpretación mediante nuevos métodos que permitan hacer computable el razonamiento, y simultáneamente, proporcionen nuevas herramientas para modelar el comportamiento de un sistema como sistema físico, Conviene revisar en este punto algunas de las dificultades que plantea el problema: 1) la variabilidad de los procesos fisiológicos y fisiopatológicos que subyacen en el trazo electrocardiográfico, variabilidad que se observa entre múltiples pacientes e incluso en el mismo paciente a lo largo del tiempo (Fig. 1); 2) la naturaleza estocástica de dichos procesos; 3) la ocurrencia simultánea de múltiples procesos fisiológicos que interaccionan entre sí de múltiples maneras; 4) la presencia de ruido y artefactos en la señal obtenida, ocultando la percepción de los procesos de interés; o 5) la ausencia de un modelo preciso del miocardio; y 6) el conocimiento tácito, subjetivo y difícilmente formalizable que forma parte de la experiencia del cardiólogo.

A continuación se presentan y discuten algunas de las propuestas que un mismo equipo científico ha realizado en los últimos años en este problema. Cabe decir que el ECG no ha sido el objeto de estudio, sino el banco de pruebas con el que experimentar nuevos modelos y técnicas de representación y razonamiento para los que la aspiración ha sido plantear soluciones generales a problemas generales. En esta trayectoria consideramos fundamental el papel de la iniciativa Physionet[1], porque más allá de compilar múltiples colecciones de datos fisiológicos de referencia en los procesos de validación científica, constituye un agente dinamizador que ha asumido el papel de estimular y alinear el interés de la comunidad científica con los desafíos que afronta la medicina actual [3]. También queremos destacar nuestra colaboración

This work was partly funded by the Spanish MINECO under projects TIN2014-55183-R and TIN2009-14372-C03-03, from the Consellería de Cultura, Educación e Ordenación Universitaria da Xunta de Galicia and the European Regional Development Fund (ERDF) under Grant No. 2016-2019-ED431G/08

[1] www.physionet.org

con el Servicio de Cardiología del Complejo Hospitalario Universidade de Santiago de Compostela.

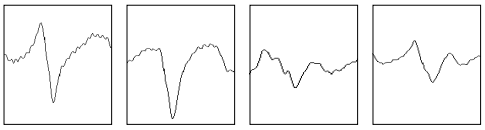

Figura 1. Evolución de la morfología de la clase de normalidad de un paciente a lo largo del tiempo.

II. El agrupamiento morfológico de latidos

Los recientes desarrollos tecnológicos en sensores y dispositivos móviles han facilitado la aparición de nuevos escenarios para una monitorización del ECG de larga duración, lo que permite la detección temprana de determinados eventos relevantes, que con frecuencia ocurren de manera esporádica. A medida que el período de registro aumenta (considerando habitual que sea de 24 horas) la tarea de interpretación exige más tiempo, y parece ineludible demandar nuevas herramientas de apoyo a la decisión. Su objetivo es realizar un resumen efectivo del registro electrocardiográfico, llamando la atención sobre las anomalías detectadas.

Entre todos los posibles hallazgos que requieren atención sobre el ECG, las arritmias cardiacas son las más relevantes. Se distinguen dos tipos de arritmias: 1) aquéllas en cuyo origen hay un trastorno del automatismo, esto es, un conjunto de alteraciones en el foco de activación del latido, ya sea en el lugar del miocardio donde se origina o en la frecuencia de activación; o 2) un trastorno de la conducción, esto es, una propagación anormal del frente de onda del latido a lo largo del tejido cardiaco. Ambos son reconocibles en el ECG, bien porque afectan a la morfología del latido, o a su ritmo de aparición.

Dos son las estrategias que se encuentran en la bibliografía científica para abordar la identificación de arritmias: la clasificación y el agrupamiento. La clasificación, fundamentalmente clasificación de latidos, busca asignar a cada latido una etiqueta que identifica su naturaleza fisiológica. De manera mayoritaria, esto se logra mediante alguna técnica de aprendizaje automático entrenada sobre un conjunto de entrenamiento. La principal dificultad de esta aproximación estriba en su dependencia de la diversidad morfológica presente en el conjunto de entrenamiento, lo que conduce a menudo a resultados decepcionantes sobre nuevos registros de ECG [4]. Además, las clases sólo proporcionan información sobre el origen del latido, dejando al margen la información sobre el camino de conducción, lo que impide distinguir entre las múltiples familias morfológicas que comparten el mismo origen y, por tanto, pertenecen a la misma clase, como ocurre en el caso de las arritmias multifocales.

El agrupamiento, en cambio, se limita a dividir el registro de ECG en un conjunto de grupos de latidos, de modo que cada uno de ellos preserva algunas propiedades de semejanza. Aquí tradicionalmente se han venido publicando propuestas en las que se fija a priori un número máximo de grupos, y el agrupamiento se realiza en diferido sobre la totalidad del registro. Esto tiene la ventaja de lograr una razonable robustez frente al ruido, pero por contra, suele distribuir los ejemplos de una misma morfología entre distintos grupos, y penaliza la identificación de morfologías raras, que acaban ocultas en grupos ajenos. Por otra parte, esta aproximación obvia el carácter dinámico del ECG y, en particular, ignora la evolución temporal de las distintas morfologías. Además, la detección de eventos de carácter crítico se pospone demasiado como para proporcionar una respuesta oportuna.

En [5] proponemos un método adaptativo para el agrupamiento de latidos, con el potencial de servir de paso previo a una técnica de clasificación, o de resumen sobre aquellas morfologías presentes en un cierto período de tiempo, su evolución temporal y su variabilidad.

Adoptamos una estrategia inspirada en la percepción visual, que reduce el contorno de cada imagen a un conjunto de puntos de máxima curvatura respecto al contexto, puntos que concentran la información más relevante y reconocible de la imagen, y así extraemos del trazo electrocardiográfico un conjunto de puntos dominantes a partir de los cuales se realiza una caracterización de las ondas constituyentes del latido cardiaco. La similitud entre latidos se calcula mediante la técnica de *Dynamic Time Warping*, lo que permite por un lado realizar un alineamiento no lineal entre el latido actual y el representante de cada uno de los grupos, y por otro, reducir la variabilidad estocástica de la señal con el fin de obtener una asignación correcta del latido al grupo más semejante. El método hace uso de un conjunto reducido de parámetros, pero todos ellos toman valores cuya justificación es exclusivamente fisiológica, ajenos a la necesidad de responder adecuadamente a algún conjunto de entrenamiento.

El método propuesto emula el comportamiento del experto cardiólogo en tanto que explota la información presente en el contexto temporal de cada latido con el fin de realizar la asignación de cada nuevo latido al grupo más apropiado. Así, los distintos grupos se adaptan continuamente a la evolución temporal de las morfologías de los nuevos latidos, pudiendo crearse dinámicamente nuevos grupos, modificarse grupos previos, o fusionarse varios grupos en uno solo, de modo que el número de grupos resultante se adapta a la variabilidad morfológica del registro analizado. El método es eficiente y puede ejecutarse en tiempo real en una computadora de propósito general.

Se ha realizado una validación de este método con la *MIT-BIH Arrhythmia database*, base de datos de referencia en la bibliografía científica para problemas de clasificación y agrupamiento en electrocardiografía. El método proporciona una medida de pureza del 98.56 %, ligeramente superior al mejor resultado previo de la bibliografía, de 98.49 %, que realiza un agrupamiento en diferido, con un número de grupos fijado a priori, y excesivo en el caso de muchos registros de la base de datos [7]. Como se ha dicho, los parámetros del método propuesto no surgen de un proceso de entrenamiento, ni se han ajustado a la base de datos de validación. El método muestra una escasa sensibilidad a pequeñas variaciones de

dichos parámetros, y aunque se pueden obtener mejoras en los resultados mediante un ajuste fino de los parámetros a cada base de datos, no es ese el objetivo del trabajo, sino mostrar la validez de la propuesta en un problema de monitorización continua en el que las características de los grupos evolucionan con el tiempo.

III. Análisis de procesos estocásticos

El análisis de la Variabilidad de la Frecuencia Cardiaca (VFC) constituye un dominio de estudio por sí mismo, concitando en los últimos años un gran interés en la comunidad científica. El término se refiere en general al estudio de la serie de las distancias temporales entre latidos consecutivos, medida como diferencia entre las correspondientes ondas R. El interés en la VFC surge de su capacidad para mostrar información sobre los mecanismos reguladores del sistema nervioso autónomo: sistema simpático, parasimpático y sistema renina-angiotensina, y de ahí su utilidad en el estudio de múltiples patologías, cardiacas y no cardiacas [8].

En el análisis de la VFC se observa en primer lugar un espectro de banda ancha, característico de los procesos con memoria a largo plazo, así como una estructura autosemejante, característica de los procesos fractales. De hecho, algunas de las características fractales de la VFC se han mostrado como eficaces predictores de fallo cardiaco e infarto de miocardio [9]. Así todo, resulta evidente que el corazón responde de una manera determinista en situaciones caracterizadas por una particular demanda: aumentando la frecuencia del latido ante la activación del sistema simpático, y reduciéndola ante la activación del parasimpático. Por otra parte, existen acoplamientos bien conocidos entre el miocardio y otros sistemas del organismo, y así por ejemplo, resulta evidente la traza de la respiración (y todas sus anomalías) en la VFC.

Parece por tanto razonable conjeturar que la VFC es el resultado de un conjunto de procesos deterministas y estocásticos que concurren en el control del miocardio. Desde esta premisa, conjeturamos un modelo de superposición en el que la serie RR es la suma $RR[n] = x[n] + B[n]$, $n = 1, ..., N$ de una componente determinista $x[n]$ limitada en banda y una componente fractal estocástica $B[n]$ que responde a las características del movimiento fraccional browniano [9]. Dado que el movimiento fraccional browniano se caracteriza por ser no estacionario y, por tanto, variable en el tiempo, y autosemejante, esto es, con las mismas propiedades estadísticas en distintas escalas, se propone analizar la distribución de energía mediante la transformada *wavelet* de la serie original. Nuestra propuesta parte de una observación relevante: la energía de los procesos autosemejantes cambia con la escala siguiendo una relación en forma de ley de potencia, y las desviaciones respecto a este comportamiento permiten la estimación del proceso determinista superpuesto $x[n]$.

Se propone por tanto un método para la estimación simultánea de las componentes deterministas y estocásticas de naturaleza fractal en series temporales no estacionarias [9], y se aplica a un problema bien conocido, la identificación del Síndrome de Apnea-Hipopnea Obstructiva del Sueño (SAOS) a partir únicamente de la serie RR. El SAOS es un trastorno del sueño caracterizado por el cese total (apnea) o parcial (hipopnea) del flujo respiratorio durante el sueño del paciente. En episodios prolongados de apnea, el ritmo cardiaco muestra una secuencia característica de bradicardia-taquicardia, bradicardia en el comienzo del cese del flujo y taquicardia durante la recuperación del flujo.

El método propuesto se ha aplicado a la *Apnea-ECG database* [10], base de datos formada por un conjunto de registros de ECG de pacientes con SAOS, y anotados con los episodios de ocurrencia de apnea, proporcionando una exactitud en la identificación de episodios de apnea del 87.62 %, por debajo del resultado del mejor clasificador publicado, del 92.62 %. En aquellos registros de la base de datos en los que el método resulta eficaz la eliminación de la componente fractal facilita el reconocimiento de la secuencia bradicardia-taquicardia (Fig. 2).

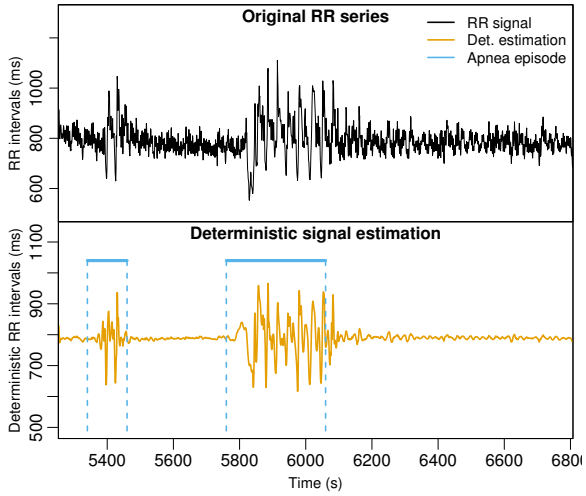

Figura 2. Estimación de la componente determinista sobre una serie RR y anotación de las apneas presentes en el registro.

En estos casos se pondría de manifiesto que durante el episodio de apnea las componentes deterministas del sistema nervioso toman el control para responder a la hipoxia que se produce durante la apnea; y por contra, en condiciones de reposo, la componente estocástica prevalecería. Sin embargo, en algunos registros de la base de datos nuestro método proporciona un resultado incorrecto, lo que sugiere que el problema reside en el carácter aditivo del modelo de partida, que ignora las posibles interacciones entre las componentes estocásticas y deterministas. En dichos registros parece que la aparición de ciertas componentes armónicas (como sucede en el acoplamiento con la respiración) inhibiría el proceso fractal estocástico, invalidando nuestra hipótesis inicial.

Probado este método en algunas series de datos de carácter económico, como la serie de precios de gasolina, se demuestra que permite una separación de componentes que hace aflorar comportamientos deterministas útiles en el análisis, como el llamado “efecto lunes” que compensa con bajadas de precios las subidas de los fines de semana [9].

Esta línea de trabajo ha continuado con el planteamiento del análisis de series temporales mediante ecuaciones diferenciales estocásticas, debido a su carácter interpretable. Dichas ecuaciones conjugan una ecuación determinista de movimiento con la existencia de fluctuaciones que interfieren en la dinámica del sistema en forma de ruido. Nuestra propuesta realiza una estimación de los coeficientes de la ecuación a partir de una serie temporal densa, haciendo uso de un procedimiento de regresión no paramétrica basada en procesos gausianos dispersos y en un marco de razonamiento bayesiano [11]. Esta propuesta se ha aplicado con buenos resultados a series de datos económicas y a problemas de paleoclimatología, pero todavía no con acierto a series temporales fisiológicas o, en particular, de cardiología.

IV. Interpretación abductiva del ECG

La comunidad cientifica ha dedicado una gran atención y esfuerzo al desarrollo de paradigmas, estrategias, metodologías y técnicas destinadas a la clasificación de series temporales. Sin embargo, y a pesar del amplio catálogo de propuestas para el diseño de clasificadores, ya sea a partir de alguna representación del conocimiento del dominio, o mediante inducción a partir de un conjunto de observaciones, el clasificador resultante se comporta siempre como un sistema deductivo. La línea de investigación que aquí se expone parte de la hipótesis de que algunas de las más importantes debilidades de los actuales clasificadores proceden de su naturaleza deductiva, y que una aproximación abductiva puede superar algunas de esas debilidades.

Se propone un nuevo paradigma abductivo de interpretación de series temporales [12], y se escoge como ámbito de aplicación el ECG, en tanto que sobre él se proyecta el comportamiento eléctrico de aquellos procesos fisiológicos que concurren en el miocardio y, en su evolución a lo largo del tiempo, el cardiólogo encuentra evidencia para la identificación y caracterización de los fenómenos cardiacos.

La propia finalidad de la interpretación pone de manifiesto la primera limitación de la deducción como inferencia. Recordemos que una deducción contiene en sus conclusiones información que ya está implícitamente contenida en las premisas, y de ese modo podríamos decir que preserva la verdad. En este sentido, un clasificador simplemente asigna una etiqueta o un conjunto de etiquetas a las observaciones disponibles. Esta etiqueta puede nombrar un proceso o un mecanismo que subyace a lo observado, pero no es más que un término que resume el conjunto de premisas que satisfacen las observaciones. Por contra, la abduccción, o inferencia de la mejor explicación, realiza el camino que va de las observaciones disponibles a las hipótesis que mejor las explican. Las conclusiones de la abducción contienen nueva información no contenida en las premisas, información con capacidad de predecir nueva evidencia, aunque de manera falible. La abducción, por tanto, amplía la verdad, llevándonos desde el lenguaje de las observaciones al lenguaje de los procesos y mecanismos subyacentes [13], y respondiendo así al reto de interpretación del ECG antes expuesto.

Ciertamente, el resultado proporcionado por un clasificador puede considerarse una conjetura, pero siempre desde un agente externo a él, ya que un clasificador, como sistema lógico, es monótono, y no puede refutar sus propias conclusiones. Las agrupaciones de clasificadores tratan de superar los errores de clasificadores individuales mediante la combinación de resultados. Así todo, su planteamiento sigue siendo de abajo a arriba, de los datos a las clases, y a partir de cierto nivel de distorsión en los datos el fallo es inevitable. Si la clasificación se realiza en un conjunto de niveles de abstracción, la monotonicidad del razonamiento propaga los errores a medida que aumentamos el nivel de abstracción, y con ello aumenta su impacto en el resultado. En cambio, desde la lógica de la abducción, cada nivel de abstracción representa un lenguaje de observación diferente, de modo que cada nueva observación se conjetura como la mejor explicación de las observaciones que se disponen en los niveles inferiores, y en el contexto de la información de los niveles superiores. La no monotonicidad de la abducción permite retractar cualquier observación en cualquiera de los niveles de abstracción a medida que se dispone de nueva información, en un proceso que integra un flujo de razonamiento de abajo a arriba y otro de arriba a abajo. Como consecuencia, la abducción permite conjeturar la ocurrencia de un proceso fisiológico a partir de un fragmento de ECG corrompido por ruido.

Por otra parte, un clasificador se construye sobre la hipótesis de que cada una de las clases resultantes constituyen categorías mutuamente excluyentes. En el lenguaje de procesos, se excluye la superposición de dos o más procesos, que tendría que representarse mediante un nuevo proceso, una nueva categoría diferente a las anteriores, y que conduce a una casuística artificiosa que añade una complejidad adicional a la interpretación de resultados. Por contra, la abducción puede alcanzar una conclusión a partir de la disponibilidad parcial de evidencia, y refinar el resultado conforme se dispone de nueva información. Esto permite inferir un conjunto de procesos como concurrentes en el tiempo, en tanto que explican la evidencia disponible en un fragmento de ECG y no son incompatibles entre ellos.

Siguiendo con el argumento anterior, en un clasificador la verdad de una conclusión se sigue de la verdad de todas sus premisas, y la ausencia de datos requiere de alguna estrategia de imputación que resulta ser una conjetura: una suerte de abducción para proseguir la deducción. En cambio, una interpretación abductiva se plantea como un ciclo de hipótesis y test, en el que la ausencia de información se integra de forma natural, en tanto que una hipótesis puede ser evocada por la más simple pieza de evidencia que resulta por ella explicada, de modo que los datos pueden ser incorporados de manera incremental en el razonamiento. Esta característica resulta completamente oportuna para el tipo de análisis variable en el tiempo que requiere la interpretación del ECG, donde datos futuros pueden exigir cambios en conclusiones previamente alcanzadas, y donde se puede exigir un resultado de la interpretación en cualquier momento del proceso de razonamiento, como la mejor explicación con la información

disponible.

El paradigma propuesto se apoya en la definición de un conjunto de patrones de abstracción. Un patrón de abstracción $P = \langle h, M_P, C_P, \Theta_P \rangle$ representa un conjunto de restricciones C_P que debe satisfacer la evidencia, representada por un conjunto de hallazgos M_P, para poder conjeturar la observación de un determinado proceso h, junto con un procedimiento de observación $\Theta_P()$ que proporciona un conjunto de medidas para las características del proceso observado. Un patrón de abstracción permite modelar un proceso de abstracción de manera abductiva, basado en la relación conjetural $m \leftarrow h$ [14], que podemos leer como 'la observación de un hallazgo $m \in M_P$ nos permite conjeturar la observación del proceso h como una posible hipótesis explicativa'.

La definición de patrón de abstracción fija a priori el conjunto de hallazgos a partir de los que se puede conjeturar la observación de h. En general, sin embargo, un proceso no tiene una duración determinada a priori ni, por tanto, un número prefijado de hallazgos. Pensemos por ejemplo en el número indeterminado de latidos que constituye un episodio de ritmo normal en un individuo. Para soslayar esta limitación, se define una gramática de abstracción, basada en la teoría de lenguajes formales, como una gramática con atributos que permite generar dinámicamente un conjunto de patrones de abstracción, de manera similar a como las gramáticas formales generan las cadenas de un lenguaje. Disponemos así de un proceso constructivo que permite añadir de manera incremental nuevas restricciones a medida que se dispone de nuevos hallazgos, proporcionando una forma sistemática de ensamblar conocimiento mediante mecanismos bien conocidos de generación de lenguaje.

Un conjunto de gramáticas de abstracción permiten definir un modelo de abstracción, y junto con un conjunto de observaciones iniciales, que en nuestro caso se corresponderían con un framento de ECG, definen un problema de interpretación. Se define una interpretación como un conjunto de hipótesis que explican las observaciones iniciales. Y se define la solución del problema de interpretación como el conjunto de todas las interpretaciones mínimas que recubren las observaciones iniciales. Se ha de cumplir que una interpretación no puede incluir hipótesis mutuamente excluyentes.

Como era de esperar, se demuestra que la búsqueda de la solución de un problema de interpretación es un problema NP-completo. El objetivo se traslada a la búsqueda de una solución aproximada, para lo que se introduce un conjunto de principios que operan como heurísticas que permiten discriminar entre distintas interpretaciones: 1) un *principio de cobertura*, que muestra su preferencia por aquellas interpretaciones que explican más observaciones iniciales; 2) un *principio de simplicidad*, que muestra su preferencia por interpretaciones con un menor número de hipótesis; 3) un *principio de abstracción*, que muestra su preferencia por interpretaciones con un mayor nivel de abstracción; y 4) un *principio de predictibilidad*, que muestra su preferencia por interpretaciones que predicen adecuadamente la evidencia futura.

Se propone y diseña el algoritmo CONSTRUE() para la búsqueda de soluciones en problemas de interpretación, que utiliza las heurísticas antes comentadas. Este algoritmo muestra las siguientes ventajas respecto a aproximaciones previas basadas en clasificadores deductivos: 1) evita la necesidad de construir una interpretación sobre una casuística exhaustiva de patrones (Fig. 3); 2) es capaz de proporcionar el resultado 'no sé', lo que resulta conveniente en ciertos fragmentos aquejados de ruido; 3) es capaz de sugerir o predecir la aparición de evidencia de la que no hay constancia clara; 4) la solución del problema de interpretación resulta explicable.

El algoritmo CONSTRUE() se ha aplicado en distintos problemas relacionados con la interpretación del ECG. Destacamos aquí dos aplicaciones que han sido objeto de publicación.

Por un lado, se ha realizado una clasificación de latidos según su origen, y se ha validado con la *MIT-BIH Arrhythmia database*. A partir de la abstracción realizada por CONSTRUE() sobre cada uno de los registros, se ha añadido una etapa final formada por un sencillo clasificador basado en reglas, para proporcionar las etiquetas con las que está anotada la base de datos. El resultado es el mejor de la bibliografía respecto a otros clasificadores automáticos, y está entre los mejores entre aquéllos que son asistidos por un experto para el etiquetado [15]. Resulta digno de mención que el conocimiento que utiliza CONSTRUE() no está ajustado a la base de datos de validación, es el mismo que se puede encontrar en un manual de electrocardiografía.

Por otro lado, se ha utilizado CONSTRUE() para resolver el reto planteado en 2017 de manera conjunta por la iniciativa *Physionet* y *Computing in Cardiology* para la identificación de fibrilación auricular en registros de corta duración en una derivación. En este caso la disparidad de criterios y la falta de consenso en la anotación de la base de datos de validación obligó a complementar los resultados proporcionados por CONSTRUE() con técnicas de aprendizaje automático, consiguiendo así el mejor resultado de todos los presentados a concurso [16].

V. Discusión

Se ha presentado un conjunto de trabajos de investigación que tienen como preocupación común el proporcionar soluciones para el análisis e interpretación de series temporales, mostrando su posible efectividad sobre la señal de ECG. En líneas generales, distinguimos dos estrategias diferenciadas en nuestra trayectoria y que tendrán continuidad en nuestro trabajo futuro: 1) Una estrategia de modelado de la serie temporal entendida como resultado de un conjunto poco conocido de procesos, deterministas y estocásticos, que probablemente interaccionan entre sí de múltiples maneras todavía por conocer. Creemos que mediante la construcción de modelos es como podemos alcanzar un conocimiento profundo sobre cómo funcionan los sistemas que observamos mediante la evolución en el tiempo de algunas de sus características medibles. La construcción de modelos es un proceso laborioso e iterativo, en el que la inteligencia artificial, o el aprendizaje automático en particular, pueden proporcionar herramientas valiosas de estimación a partir de los datos disponibles. 2) Una

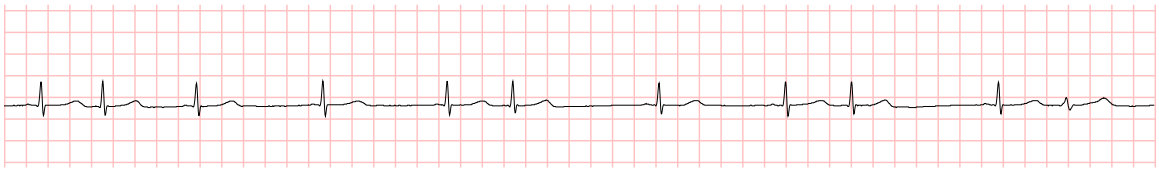

Figura 3. Los detectores conocidos de fibrilación auricular utilizan criterios que hacen de este registro un firme candidato a esa etiqueta. Sin embargo, CONSTRUE() explica apropiadamente los primeros cinco latidos como una bradicardia sinusal, compatible con la presencia de un latido prematuro ectópico en la segunda posición, seguida de un patrón de trigeminismo durante seis latidos, y finalmente otro latido ectópico con un cambio en la morfología. [Fuente: Proyecto Mobiguide, registro privado]

estrategia de modelado del razonamiento que consideramos que forma parte de las funciones cognitivas del ser humano, y que le permiten interpretar el comportamiento de un sistema a partir de los datos disponibles.

Los retos científicos alineados con ambas estrategias son numerosos. En la primera de ellas destaca naturalmente la búsqueda de modelos estocásticos apropiados para series temporales en las que el ruido constituye una componente intrínseca de la dinámica. Siguiendo la aproximación clásica, el objetivo sigue siendo obtener una representación de la dinámica del sistema en el espacio de fases, pero en problemas en los que no se satisface la propiedad de markovianidad asociada a las ecuaciones diferenciales estocásticas usuales, con el fin de modelar procesos caracterizados por una memoria a largo plazo.

Respecto a la segunda de las estrategias planteadas, el reto más relevante es sin duda incorporar el aprendizaje al ciclo hipótesis-test, e integrar así en un mismo esquema formal los tres tipos inferenciales reconocidos por Peirce [14]: deducción, abducción e inducción. La primera forma de incorporación, la más sencilla, consiste en la adaptación dinámica del conocimiento disponible a la evolución particular de la serie temporal, para lo que puede resultar útil un agrupamiento morfológico que evoluciona en el tiempo e incorpora un factor de memoria que determina el compromiso necesario entre plasticidad y estabilidad en función de la capacidad de cambio del sistema [17]. Una segunda forma de incorporación del aprendizaje ha de permitir ampliar el modelo de interpretación con nuevos patrones de abstracción que se pueden proponer como respuesta a determinados errores en la cobertura que proporciona un modelo de intepretación sobre la evidencia. Resulta intuitivo pensar que si una morfología de señal sobre el ECG ocupa sistemáticamente el lugar de un latido pero no se reconoce como tal a partir del conocimiento disponible, este conocimiento debe ampliarse para incorporar tal novedad. Por último, una tercera forma de incorporación del aprendizaje debería plantear el descubrimiento de patrones y gramáticas de abstracción en series temporales, sin ningún tipo de supervisión, sólo con la orientación de un conjunto mínimo de principios que organicen la búsqueda, como es la frecuencia de ocurrencia como medida de interés [18], o el patrón mínimo discernible significativo: por ejemplo, en el caso del ECG, la onda o deflexión de la señal, bajo una restricción de amplitud mínima para ser manifestación de algún fenómeno fisiológico.

Referencias

[1] World Health Organization, "WHO methods and data sources for country-level causes of death 2000-2015", 2017.

[2] G. Begg, K. Willan, K. Tyndall, C. Pepper, M. Tayebjee, "Electrocardiogram interpretation and arrhythmia management: a primary and secondary care survey," British Journal of General Practice, vol. 66, no. 646, pp. e291–e296, 2016.

[3] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, pp. e215–e220, 2000.

[4] Y. H. Hu, S. Palreddy, W. J. Tompkins, "A patient-adaptable ECG beat classifier using a mixture of experts approach," IEEE Transactions on Biomedical Engineering, vol. 44, no. 9, pp. 891–900, 1997.

[5] D. Castro, P. Félix, J. Presedo, "A method for context-based adaptive QRS clustering in real time," IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 5, pp. 1660–1671, 2015.

[6] G. Moody, R. Mark, "The impact of the MIT-BIH Arrhythmia Database," IEEE Eng. Med. Biol. Mag., vol. 20, no. 3, pp. 45–50, 2001.

[7] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, L. Sörnmo, "Clustering ECG complexes using Hermite functions and self-organizing maps," IEEE Transactions on Biomedical Engineering, vol. 47, no. 7, pp. 838–848, 2000.

[8] X. A. Vila, F. Palacios, J. Presedo, M. Fernández-Delgado, P. Félix, S. Barro, "Time frequency analysis of heart rate variability," IEEE Engineering in Medicine and Biology Magazine, vol. 16, no. 5, pp. 119–126, 1997.

[9] C. A. García, A. Otero, P. Félix, J. Presedo, D. G. Márquez, "Simultaneous estimation of deterministic and fractal stochastic components in non-stationary time series," Physica D: Nonlinear Phenomena, vol. 374-375, pp. 45–57, 2018.

[10] T. Penzel, G. Moody, R. Mark, A. Goldberger, J. Peter, "The apnea-ECG database," in Computers in Cardiology, pp. 255–258, 2000.

[11] C. A. García, A. Otero, P. Félix, J. Presedo, D. G. Márquez, "Non parametric estimation of stochastic differential equation with sparse gaussian processes," Physical Review E, vol. 96, no. 2, pp. 022104, 2017.

[12] T. Teijeiro, P. Félix, "On the adoption of abductive reasoning for time series interpretation," Artificial Intelligence, Aceptado, 2018.

[13] J. R. Josephson, S. G. Josephson, Abductive inference. Computation, philosophy, technology. Cambridge University Press, 1994.

[14] C. Hartshorn et al. Collected papers of Charles Sanders Peirce. Harvard University Press, 1931.

[15] T. Teijeiro, P. Félix, J. Presedo, D. Castro, "Heartbeat classification using abstract features from the abductive interpretation of the ECG," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 2, pp. 409–420, 2018.

[16] T. Teijeiro, C. A. García, D. Castro, P. Félix, "Arrhythmia classification from the abductive interpretation of short single-lead ECG records, " in Computing in Cardiology, pp. 1–4, 2017.

[17] D. G. Márquez, A. Otero, P. Félix, C. A. García, "A novel and simple strategy for evolving based clustering," Pattern Recognition, vol. 82, pp. 16–30, 2018.

[18] M. R. Álvarez, P. Félix, P. Cariñena, "Discovering metric temporal constraint networks on temporal databases," Artificial Intelligence in Medicine, vol. 58, pp. 139–154, 2013.