RECONOCIMIENTO AUTOMATIZADO DE PATRONES COMO HERRAMIENTA DIAGNÓSTICA EN EL LUPUS ERITEMATOSO SISTÉMICO (LES)

Ana Carpio[1], Alejandro Simón[1], Alicia Torres[1], Luis F. Villa[2]

1 Universidad Complutense de Madrid, ana_carpio@mat.ucm.es, alejsimo@ucm.es, alitor01@ucm.es

2. Servicio de Reumatología, Hospital Universitario Puerta de Hierro de Madrid, luisfernando.villa@salud.madrid.org

1. Introducción

La inteligencia artificial (IA) es útil para el aprendizaje automático en contextos clínicos con grandes cantidades de datos. Se emplean técnicas de clasificación supervisada y no supervisada en el estudio de la expresión/transcripción génica en la enfermedad y de respuesta terapéutica. Sin embargo, la cantidad de datos disponibles en muchas situaciones médicas es escasa, por lo que es preciso calibrar muy bien la técnica de análisis a emplear para que el rendimiento sea adecuado. Dada la complejidad, el polimorfismo clínico del LES y sus diversas complicaciones, el reconocimiento automatizado de patrones evolutivos de variables puede ser de ayuda en la clínica.

2. Objetivo

Evaluar la utilidad de un sistema automatizado de diagnóstico en la detección de brotes y complicaciones en los pacientes con LES.

3. Pacientes y Métodos

Se crea una base de datos con los valores analíticos anonimizados e irreversiblemente disociados de Id. de 20 pacientes con lupus eritematoso sistémico (LES), diagnosticado conforme a criterios ACR/EULAR (no es precisa evaluación por CEIC según Leyes 15/1999 y 41/2002). Se asimilan los datos a una matriz

$$M = (m_{i,j}),\ i = 1,...,I,\ j = 1,...,J,$$

que contiene datos para I variables en J momentos diferentes. Normalizamos los resgistros siguiendo dos estrategias distintas para obtener caracterizaciones temporales, por una parte, y para identificar la naturaleza de los brotes según patrones observados por otra.

4. Normalización y búsqueda de clusters

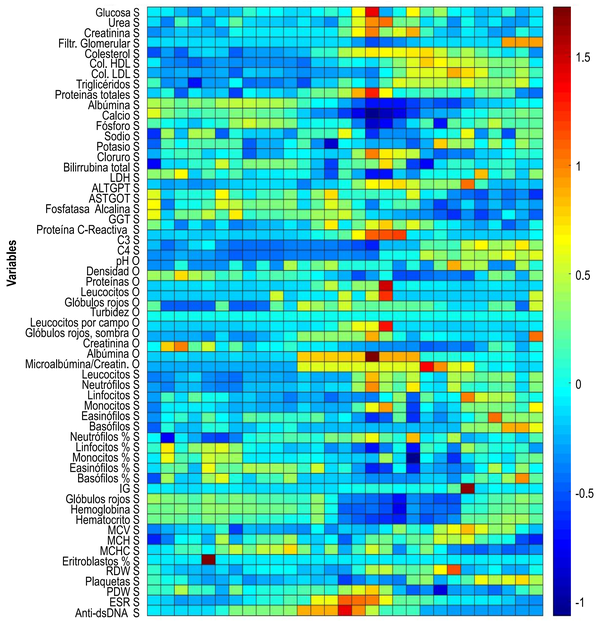

Fig1. Medidas registradas a lo largo del tiempo

Mapa de calor ilustrando la variación con el tiempo (columnas) del estado de las variables (filas) normalizadas. Las regiones azul oscuro quedan muy por debajo de las media usual en el paciente, mientras que las regiones rojas quedan muy por encima.

Normalizamos los datos calculando para cada variable v_i, i=1,..., I, la media μ_i y su desviación estándar σ_i, de las medidas disponibles en distintos tiempos $m_{i,j}$, j=1,...,J. A continuación, construimos la matriz con datos normalizados

$$\underline{m}_{i,j} = (m_{i,j} - \mu_i)/(3\sigma_i), \qquad i=1,...,I,\ j=1,...,J,$$

la mayoría de los cuales están comprendidos entre -1 y 1 [1,2]. Para el análisis se silencian las variables en las que faltan más del 50% de los registros y aquellas que mantienen esencialmente el mismo valor en el intervalo de análisis.

Al visualizar el resultado con un mapa de calor para un paciente concreto, detectamos en la Figura 1 periodos de fuertes variaciones con respecto al estado normal del paciente, que se corresponden con brotes de la enfermedad [1,2,4].

Calculando la distancia Euclidea entre los valores que toman las variables normalizadas en los distintos tiempos

$$d(m_{\cdot,j} - m_{\cdot,k}) = (\Sigma_{i=1,...,I} | m_{i,j} - m_{i,k} |^2)^{1/2},\ \ j,k=1,...,J,$$

podemos implementar un algoritmo de agrupamiento jerárquico que conduce a un dendrograma. El dendrograma visualiza cómo se agrupan los distintos registros temporales en clusters (brote, remisión, tratamiento...).

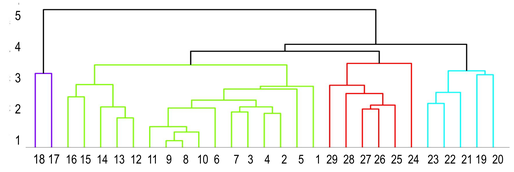

Fig 2. Dendrograma correspondiente al mapa de calor.

La Figura 2 agrupa los registros en 4 grupos principales: brote (17-18, fuertes desviaciones en varias variables), tratamiento (19-23), remisión (24-29), pre-brote (1-16). El periodo previo al brote muestra asimismo subgrupos, por ejemplo, los registros 15-16 ya anuncian inestabilidad con un pico de Anti-dsDNA.

Análogamente se podría agrupar variables en lugar de registros.

5. Normalización y búsqueda de patrones

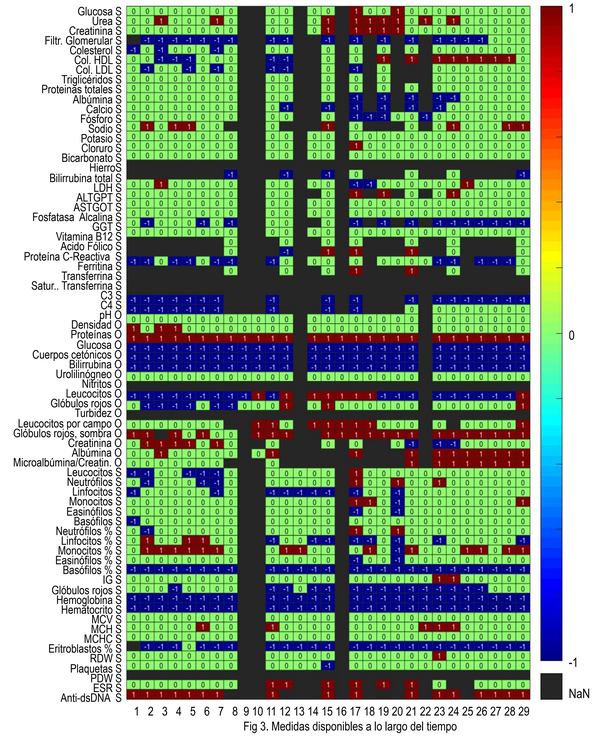

Fig 3. Medidas disponibles a lo largo del tiempo

Mapa de calor que indica las variables que toman valores dentro de su rango de normalidad (0), por debajo (-1) o por encima (1). Los cuadros negros indican que no se conoce el rango de normalidad o que falta el registro. Se identifica las variables que están fuera de rango de forma habitual, y aquellas que lo hacen puntualmente. Los patrones de -1,0,1 observados en cada registro temporal permiten identificar patologías.

Una normalización alternativa proporciona información complementaria. Conocido el rango de normalidad para cada variable medida, proporcionado por el Laboratorio de Análisis Clínicos, reemplazamos su valor por

0 si queda dentro del rango de normalidad,

-1 si queda por debajo,

1 si queda por encima,

véase la Figura 3.

Numerosas patologías se caracterizan por combinaciones de 1, 0, -1 en determinadas variables. Por ejemplo, la anemia normocítica corresponde a -1 hemoglobina en sangre y 0 velocidad corpuscular media.

Se compara el estado de las variables en determinadas columnas correspondientes a días específicos con patrones de enfermedad conocidos definidos por secuencias de -1, 0, 1 utilizando la distancia de Hamming, ya que realizando un análisis previo de clasificación del modelo Plackett-Luce observamos que en este contexto es la más eficiente [1,3].

. Para las rutinas de cálculo se emplea Matlab_R2018b.

El sistema automatizado es capaz de identificar, en los pacientes: (1) Secuencias sospechosas de brote indicado por elevación de anti-dsDNA e hipocomplementemia C3 y C4, (2) Disminución de la tasa de filtrado glomerular y (3) Proteinuria. Asimismo, se incluye en el algoritmo detección automática de patrones definitorios de hemocitopenias, elevación de LDL-colesterol (combinable con monitorización de variables de riesgo vascular), o sospechosos de inicio de síndrome hemofagocítico. Considerando el paciente analizado en las Figuras 1 y 2, observamos que la hipocomplementemia y la elevación de Ac. anti-dsDNA anteceden a la elevación de creatinina en los días 15-17 y a la proteinuria, heráldicos del desarrollo de daño renal.

6. Conclusiones

Este sistema de análisis es extrapolable a matrices constituidas por valores de constantes vitales, síntomas o signos físicos, parámetros obtenidos de la lectura automatizada de imágenes y patrones de activación génica, transcriptómica y proteómica y constituye una ayuda potencial a la toma de decisiones en el seguimiento de los pacientes con LES.

7. Referencias

[1] A. Carpio, A. Simón, A. Torres, L.F. Villa, Pattern recognition in data as a diagnosis tool, Journal of Mathematics in Industry 12, 3, 2022

[2] A. Simón, Métodos Bayesianos para comparar el funcionamiento de algoritmos sobre un conjunto de datos biomédicos, Trabajo de fin de Doble Grado de Informática y Matemáticas, UCM-UGR 2020

[3] A. Torres, Técnicas matemáticas para datos médicos: desórdenes autoimmunes, Trabajo de fin de Master de Ingeniería Matemática, UCM 2021

[4] A. Carpio, A. Simón, L.F. Villa, Clustering methods and Bayesian inference for the analysis of the evolution of immune disorders, arXiv:2009.11531, 2020

8. Agradecimientos

Investigación financiada parcialmente por los proyectos MTM2017-84446-C2-1-R y PID2020-112796RB-C21. Se agradece al Hospital Universitario Puerta de Hierro de Madrid el acceso a datos anonimizados.