Metodología Basada en Agrupamiento y Visualización para el Fenotipado de Pacientes

J. M. Juarez, A. Lopez Martinez-Carrasco, M. Campos, A. Morales
Facultad de Informática
Universidad de Murcia
{jmjuarez|antonio.lopez31|manuelcampos|morales}@um.es

Francisco Palacios
Unidad de Cuidados Intensivos
Hosp. Universitario de Getafe
franciscodepaula@gmail.com

***Resumen*—El cuidado y tratamiento de pacientes hospitalarios depende en parte de la epidemiología local. La caracterización de grupos de población (fenotipado) a partir de la historia clínica es por tanto una tarea esencial que puede ser tratada con técnicas de aprendizaje computacional. A pesar del gran abanico de técnicas para identificación de grupos, los equipos asistenciales demandan la interpretabilidad de los procesos con el fin de darles una validez médica. En este trabajo proponemos una metodología que desarrolle este proceso de aprendizaje basada en agrupamiento y visualización con el fin de atender a los aspectos de reproducibilidad e interpretabilidad para el clínico. Finalmente demostramos la utilidad de la metodología con un caso de estudio en el campo de la resistencia antibiótica.**

***Palabras clave:*—Agrupamiento; Visualizacion; Subgrupos; Infecciones; Inteligencia Artificial Medicina**

I. Introducción

La caracterización de los conjuntos poblacionales en el ámbito de la salud es esencial para la mejora de la calidad asistencial. Así, en el ámbito hospitalario, la epidemiología local juega un papel esencial a la hora de tomar decisiones terapéuticas. Por ejemplo, para el problema de la resistencia a la antibioticoterapia, es clave contar con sistemas para la ayuda la identificación del fenotipo (características físicas y conductuales) de pacientes con una mayor pérdida de efectividad [1], [2].

Desde un punto de vista computacional, este tipo de problemas se traduce en la búsqueda de individuos con una serie de características comunes y formando conjuntos no disjuntos, es decir, un problema de búsqueda de subgrupos. El descubrimiento de subgrupos se define como un método descriptivo y exploratorio de minería de datos [3]. Hay un creciente interés por esta disciplina, proponiéndose un buen conjunto de algoritmos principalmente para datos cualitativos y binarios, realizando una búsqueda exhaustiva o aplicando heurísticas [4], [5]. Existen algunos antecedentes del uso de estas técnicas en el ámbito de la medicina. Por ejemplo, la librería VIKAMINE específica para descubrimiento de subgrupos se ha utilizado para la mejora en el diagnóstico con ultrasonidos [6]. Sin embargo, la búsqueda de grupos de interes suele medirse como la distribución inusual de cierta propiedad de interés, definiendo medidas de calidad de subgrupos. Estas medidas no son triviales y son altamente sensibles al problema específico y a los subgrupos seleccionados.

En el ámbito de la investigación médica existe cierta experiencia en el uso de técnicas de aprendizaje computacional. Por ejemplo, en problemas de clasificación, los algoritmos de árboles de decisión son bien conocidos, puesto que el modelo resultante es aplicable para toma de decisiones, es visual y se fundamenta en la partición de un conjunto de datos. Otras técnicas familiares al médico y que son potencialmente útiles en problemas de fenotipado son los métodos de agrupamiento (clústering) para clasificación no supervisada.

En la última década, debido a la eclosión de proyectos de data-science y la disponibilidad de paquetes estadísticos y de minería de datos, las soluciones para este tipo de problemas se centran en procesos de caja negra, dando poca opción al clínico a incorporar el conocimiento obtenido [7]–[9]. En oposición a esta aproximación, en los últimos años hay un creciente interés por la estrategia *human-in-the-loop* que consiste en involucrar al usuario en las tareas de selección, modelado y validación con el fin de refinar procesos de minería de datos y mejorar en la generación de conocimiento [7], [10]. En problemas del ámbito de la investigación médica, además, es imprescindible permitir la interpretabilidad del algoritmo, la trazabilidad vinculando el modelo obtenido con los pacientes concretos y la reproducibilidad del experimento para su validación clínica.

Las técnicas de visualización tienen el potencial de ayudar a los expertos a entender los modelos y la configuración de los algoritmos y sus resultados [11]. En concreto, la visualización exhaustiva de posibles resultados cuando hay un ajuste de parámetros aporta un gran ahorro de tiempo y costes. Por ejemplo, en [12], la interpretación visual de datos y patrones ha permitido mejorar el modelo obtenido en la obtención de reglas de asociación temporales en infecciones nosocomiales en una UCI.

Las contribuciones de este trabajo son:

- Una metodología para el fenotipado de pacientes dirigida por los principios de trazabilidad e interpretabilidad (Sección II-B)
- Una propuesta genérica de adaptación de técnicas de agrupamiento para resolver problemas de subgrupos (Sección II-A).
- Estudio de caso de aplicación de los puntos anteriores en el contexto médico real de las resistencias antibióticas (Sección III).

II. Subgrupos mediante agrupamiento

En esta sección describimos una propuesta para el descubrimiento de subgrupos de pacientes basada en técnicas de aprendizaje computacional. La propuesta se compone de dos aspectos fundamentales (ver Fig. 1). En primer lugar la adaptación de técnicas de agrupamiento, familiares en el ámbito clínico, para resolver el problema de subgrupos de forma automática. Así, la Sec. II-A establece el marco formal de esta adaptación. En segundo lugar se propone una metodología para la obtención de subgrupos basado en la estrategia human-in-the-loop (Sec. II-B) con el fin facilitar posteriores estudios en el campo de la investigación médica. Por este motivo, la propuesta se fundamenta en (i) el principio de trazabilidad, es decir, el modelo resultante debe tener una correspondencia clara con los individuos para su evaluación clínica, y (ii) la interpretabilidad del modelo que permitirá posteriormente aportar información experta.

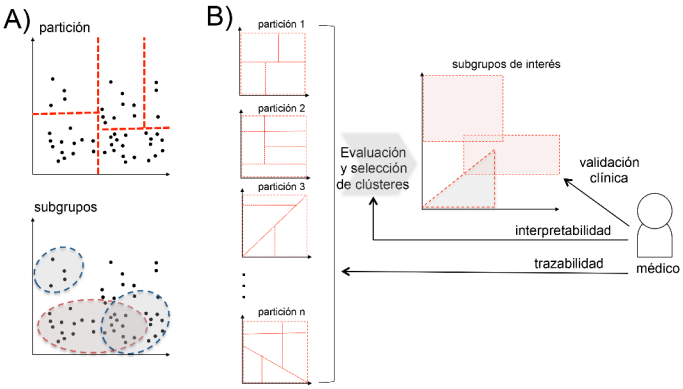

Figura 1. A) De la partición a los subgrupos; B) Trazabilidad, interpretabilidad y validación.

II-A. Marco formal

A continuación describimos una propuesta de descubrimiento de subgrupos relevantes mediante la utilización de algoritmos de agrupamiento. Esta propuesta parte de la hipótesis de que, tras aplicar los algoritmos de agrupamiento de forma iterada, los conjuntos de individuos que lleguen a permanecer juntos en esos clústeres son los candidatos a conformar los subgrupos que se desean encontrar. Por tanto, la propuesta se fundamentará en la evaluación y comparación de clústeres entre las diferentes particiones obtenidas tras la ejecución de algoritmos de agrupamiento.

Describiremos formalmente los principales elementos de este proceso.

Def. Partición: dado un conjunto de datos C, C_x es una *partición* de C si $C_x \subseteq \mathcal{P}(C)$ con $|C_x| = x$ donde $C_x = \{C_{x1}, \ldots, C_{xx}\}$ y $C_{x1} \cup \ldots \cup C_{xx} = C$.

Def. Clúster: a los elementos de C_x se les denominan *clústeres*, cumpliendo que $\forall C_{xi}, C_{xj} \in C_x, C_{xi} \cap C_{xj} = \emptyset$.

Es decir, denotamos como C_{xi} y C_{xj} a dos clústeres de una misma partición C_x, mientras que C_{xi} y C_{yi} son clústeres de que pueden ser de particiones diferentes si $x \neq y$.

Denotamos como $\mathcal{P}(C)_k$ al espacio de particiones de C con k clústeres.

Def. Funcion Agrupamiento: dado un conjunto de datos C y un valor entero positivo k la función de agrupamiento establece un partición de C obteniendo k clústeres.

$$Agrupamiento : C \times \mathbb{Z}^+ \rightarrow \mathcal{P}(C) \tag{1}$$

Es decir $Agrupamiento(C, k) \in \mathcal{P}(C)_k$.

Por simplicidad en el modelo, y sin pérdida de genericidad, asumimos el uso de algoritmos clásicos de agrupamiento, entendiendo que son aquellos cuyo objetivo es la partición del conjunto de datos en k subconjuntos (clústeres), siendo este parámetro establecido a-priori.

Un aspecto esencial en el estudio de los algoritmos de agrupamiento es la evaluación de sus particiones mediante índices de validez de los clúster (CVI), como los índices Rand o Silhouette [13], [14]. Entre la evaluación directa de clústeres podemos encontrar diferentes criterios. En [15] se clasifican los CVI como: internos (propiedades de los elementos del clúster), relativos (evaluar la partición en su conjunto según un criterio como el número de individuos) y externos (estructura de la distribución de los individuos). Sin embargo, el método más habitual es definir una función para evaluar un clúster donde destacan métricas de compactación (cercanía entre indivíduos del clúster) y métricas de separación (separación respecto a indivíduos del resto de clústeres). En este último grupo, una de las métricas más extendidas es el coeficiente de Jaccard [16], que se define como:

$$J(C_{xi}, C_{yj}) = \frac{|C_{xi} \cap C_{yj}|}{|C_{xi} \cup C_{yj}|} \tag{2}$$

En este trabajo nos centraremos en medidas de evaluación de clústeres de diferentes particiones y con este fin generalizaremos este tipo de métricas a través de la función de coincidencia.

Def. Función Coincidencia: dadas dos particiones C_x y C_y la métrica de coincidencia entre sus clústeres C_{xi} y C_{yj} se define como la función que mide el grado de similitud entre clústeres, normalmente de distinta partición. Formalmente:

$$M : \mathcal{P}(C)_a \times \mathcal{P}(C)_b \rightarrow [0, 1] \tag{3}$$

Cumpliendo las siguientes dos propiedades:

$$M(C_{xi}, C_{yj}) = 1 \iff C_{xi} = C_{yj}. \tag{4}$$

$$M(C_{xi}, C_{yj}) = 0 \iff C_{xi} \cap C_{yj} = \emptyset. \tag{5}$$

En este trabajo, planteamos la idea intuitiva de traza como la tarea de seguimiento de los individuos de un clúster que se encuentran agrupados en los clústeres que otras particiones.

Def. Función de Traza: sea un conjunto de datos C y un conjunto de particiones $\{C_1, \ldots, C_{K-1}\}$ resultante de aplicar iterativamente un algoritmo de agrupamiento donde variamos el número de clústeres $(1 \ldots K)$. Dado un clúster (C_{Ki}) de la partición C_K, denominamos traza a un conjunto formado por el clúster de cada partición $C_2, \ldots, C_{K-1}$ (descartando $C_1 = C$) que maximiza la función de coincidencia en relación a dicho cluster C_{Ki}.

$$Traza : C_K \times \{\mathcal{P}(C)_1, \ldots, \mathcal{P}(C)_K\} \rightarrow C_{1i_1} \times \ldots \times C_{K-1i_k} \quad (6)$$

En este trabajo presentamos el Algoritmo 1 que implementa dicha función.

Algoritmo 1 Traza

Input C_{xi}: clúster ; $\{C_1, \ldots C_x\}$: conjunto particiones ; M: función de coincidencia
Output T *%vector de clústeres seleccionados*

```
T ← ∅
for k = x − 1 ... 2 do
    candidato ← C_k1
    for y = 1 ... k do
        if M(C_xi, C_ky) > M(C_xi, candidato) then
            candidato ← C_ky
        end if
    end for
    T_k ← candidato
end for
return T
```

Siendo M una función de coincidencia y T el vector de clústeres seleccionados como traza de C_{xi}. Cabe destacar que habiendo x particiones $(C_1, \ldots, C_x)$, $k \in [2, x-1]$. Esto es así puesto que: (1) C_1 es una partición con un único clúster y por tanto $(C_{x,i} \subseteq C_{11})$ y (2) C_{xi} es un clúster de C_x y por definición $C_{xi} \cap C_{xj} = \emptyset$ cuando $i \neq j$.

Por ejemplo, sean las particiones $C_1, \ldots C_5$ decimos que $Traza(C_{51}, \{C_1, C_2, C_3, C_4, C_5\}, M) =< C_{22}, C_{31}, C_{43} >$ para expresar que, de acuerdo con una métrica de coincidencia M, gran parte de los individuos del cluster C_{51} permanecen agrupados en los clústeres C_{22}, C_{31} y C_{43}.

Función M-Trazas: sea un conjunto de datos C y un valor entero K, la función $M - Trazas$ calcula una matriz de trazas a partir de las particiones de $C_1 \ldots C_K$, calculando los vectores a través de la función $Traza$ para los clústeres de C_{Ki}.

$$M - Trazas : C \times Z^+ \rightarrow C_1 \times \ldots \times C_{K-1} \quad (7)$$

El Algoritmo 2 presenta una implementación de dicha función. Siguiendo el ejemplo anterior $M - Trazas(C, 4) = T$, donde

Algoritmo 2 M-Trazas: Matriz de Trazas

Input C: conjunto de datos ; $K \in Z^+$, M: función de coincidencia
Output $\mathcal{T}$ *%matriz de clústeres seleccionados*

```
C = ∅
T ← ∅
for i = 1 ... K do
    C_i ← Agrupamiento(C, i)
    C = C ∪ {C_i}
end for
for i = 1 ... K do
    T_i ← Traza(C_Ki, C,M)
end for
return T
```

T es una matriz 4×3 formado por las filas $T_1, \ldots, T_4$. Cada fila T_i es la traza para el clúster C_{4i}

II-B. Metodología

La metodología para la obtención de subgrupos propuesta está basada en el modelo de trazas de clústeres de la sección II-A. Esta metodología se resumen en la Fig. 2 y se compone de los siguientes pasos:

1. Extracción de datos y selección de parámetros.
2. Selección de algoritmo y parámetros de agrupamiento.
3. Subgrupos: preselección automática de clústeres.
4. Visualización: asistencia a selección de subgrupos.
5. Validación experta.

El primer paso consiste en la extracción, transformación y carga de las fuentes de datos clínicas. En nuestro caso este proceso se realiza con la herramienta WASPSS [2], que integra datos provenientes de los servicios de microbiología, farmacia, laboratorio y censos de un hospital. Una vez cargados, se procede al diseño de una vista minable, seleccionando los parámetros diana, en función de los objetivos clínicos del estudio. Este conjunto de datos lo denominamos C.

El segundo paso es la selección del algoritmo de agrupamiento y estimación del número de clústeres máximos esperables, denominados función $Agrupamiento$ y parámetro K respectivamente. Ambas decisiones dependerán de la naturaleza de los parámetros diana seleccionados.

En tercer lugar se pasará al cálculo de clústeres candidatos a ser seleccionados como subgrupos. Con este fin haremos uso de la función $M - Trazas$ (Algoritmo 2). Una vez decidido C, K y la función $Agrupamiento$ únicamente falta seleccionar una función de coincidencia M (Expr. 3). En este trabajo proponemos una medida específica basada en el índice de Jaccard (expr. 2), denominada $J2$ como sigue:

$$J2(C_{xi}, C_{yj}) = \frac{|C_{xi} \cap C_{yj}|}{|C_{yj}|} \quad (8)$$

Esta función está diseñada específicamente para procesos de comparación de un único clúster de poca cardinalidad frente a un conjunto de clústeres previsiblemente de mayor tamaño. Mientras que Jaccard obtiene el ratio entre los elementos comunes frente a todos los elementos, $J2$ obtiene el ratio frente al segundo clúster, ahorrando el cálculo del denominador. Además, $J2(a, b)$ respecto a $J(a, b)$ cumple la siguiente propiedad: si $|b| < |a| \Rightarrow J2(a, b) > J(a, b)$. Esta propiedad es útil en nuestro caso, ya que en la búsqueda de subgrupos se deben valorar los subconjuntos de menor cardinalidad que incluyan el mayor número de elementos de nuestro clúster de estudio (en el ejemplo a). En resumen, en el tercer paso se calculará la matriz de trazas $M - Trazas(C, K, J2) = \mathcal{T}$.

Una vez hecha la preselección de clústeres, el cuarto paso involucra al experto durante el proceso de elección del modelo computacional. Este paso se implementa mediante técnicas de visualización de los clústeres seleccionados con el fin de que el clínico de forma asistida pueda decidir cuáles son los subgrupos de estudio. En concreto, el objetivo es proporcionar una representación visual de la matriz de trazas $\mathcal{T}$ conteniendo para cada $\mathcal{T}_{xi}$ el cluster que más parecido a C_{Kx} cuando en C se hace una i-partición. En particular, el objeto de análisis se centrará en el estudio de cada una de sus filas $\mathcal{T}_x$ donde se representa la traza del C_{Kx} para cada tamaño de la partición. Proponemos resumir esta información construyendo

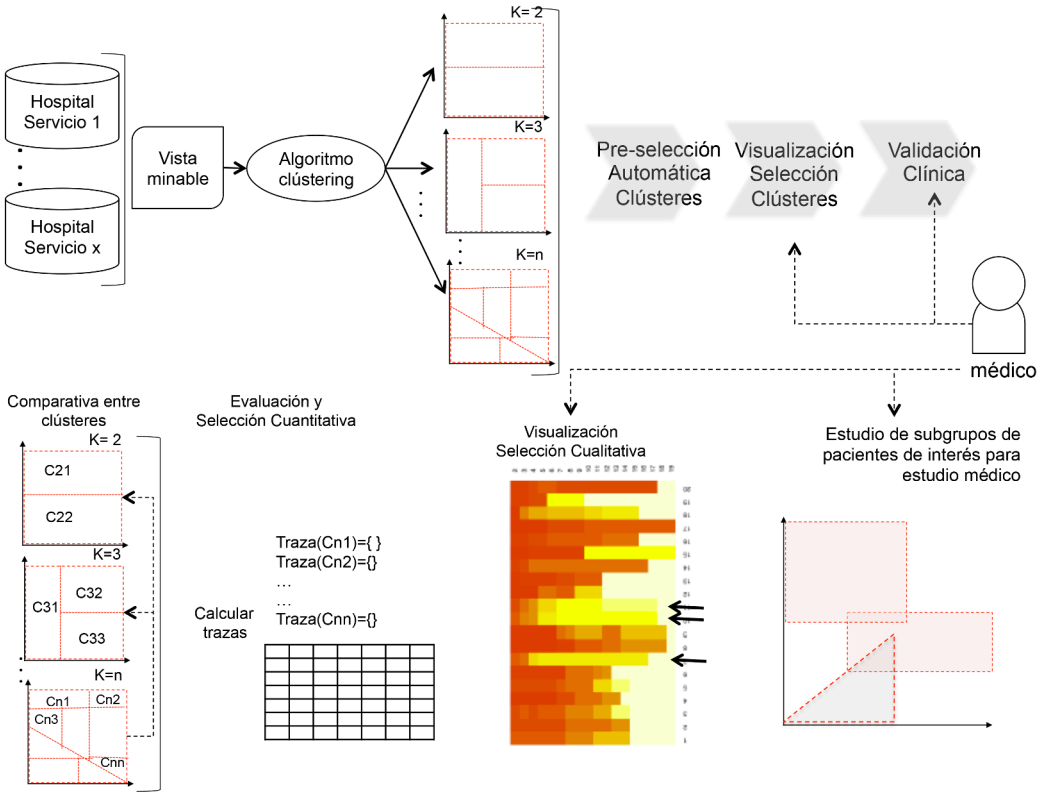

Figura 2. Detalles de la metodología.

la matriz $\mathcal{J}$ donde $\mathcal{J}_{xi} = J2(C_{Kx}, \mathcal{T}_{xi})$, es decir, su grado de coincidencia existente entre C_{Kx} con $\mathcal{T}_{xi}$.

Una vez resumida la información en la matriz $\mathcal{J}$ se procederá a la representación visual con el objetivo de facilitar la selección de subgrupos. Hay un amplio abanico de técnicas de visualización de datos matriciales, siendo el modelo de *heatmap* una forma efectiva de identificar grupos de valores que destacan por tener valores muy altos o muy bajos utilizando un código de colores. En [17], se demuestra la utilidad del modelo *heatmap* para representación de datos en particiones.

Aunque $\mathcal{J}_{xi}$ formalmente no cumple ninguna propiedad, en la práctica ocurre con frecuencia que $\mathcal{J}_{xi} > \mathcal{J}_{xj}$ cuando $i << j$. Esto sucede ya que la i-partición tiene menor número de clústeres que la j-partición y por tanto sus clústeres tendrán en muchos casos una mayor cardinalidad.

En la Figura 2 se muestra un ejemplo de visualización de la matriz $\mathcal{J}$ donde 3 clústeres de diferentes particiones han sido seleccionados como subgrupos de estudio.

El último paso de esta metodología es la validación de los clústeres seleccionados en el dominio médico. La actividad esencial es el estudio caso a caso de los pacientes incluidos en cada clúster con el fin de analizar la clínica del subgrupo. Al ser esta última etapa principalmente manual y con el fin de obtener resultados objetivos, proponemos adoptar técnicas de validación cuantitativa [18]. En el caso de contar con varios expertos, no gran número debido al dominio, sugerimos las medidas de asociación clásicas propuestas en [19].

III. Experimento

Este experimento se centra en mejorar el problema del ámbito del uso racional de antibióticos en el hospital. En concreto, el objetivo es identificar grupos de interés entre pacientes con sospecha de infección microbiana y las resistencias antibióticas. En concreto, se estudiará el tratamiento con Vancomicina y el antibiograma de dichos pacientes referente a las bacterias: Staphylococcus Aureus, Enterococcus Faecalis, Staphylococcus Epidermidis, MARSA, Staphilococcus Coagulasa Negativo y Enterococcus Faecium.

En el experimento realizado, se han recopilado datos provenientes de 4 fuentes:

- Historia Clínica: información demográfica: 4 atributos: $pk\_paciente$, $edad$, $sexo$, $tiempo\_ingreso$.
- Dep. Microbiología: cultivos realizados: 6 atributos con información de tiempos de cultivo (ej. $cultivosPrimeras72h$ o $cultivosDespues1dia$).
- Dep. Farmacia: información sobre tratamiento antibiótico: 12 atributos booleanos.
- Laboratorio: contexto de flora: 144 atributos (booleanos).

Esta base de datos consta de 169 atributos con un total de 1.778 registros referentes a los cultivos realizados.

La recopilación y almacenamiento de datos se realiza a través de la plataforma WASPSS de vigilancia antimicrobiana [2].

En relación a la limpieza y transformación de datos se han tenido en cuenta las siguientes cuestiones:

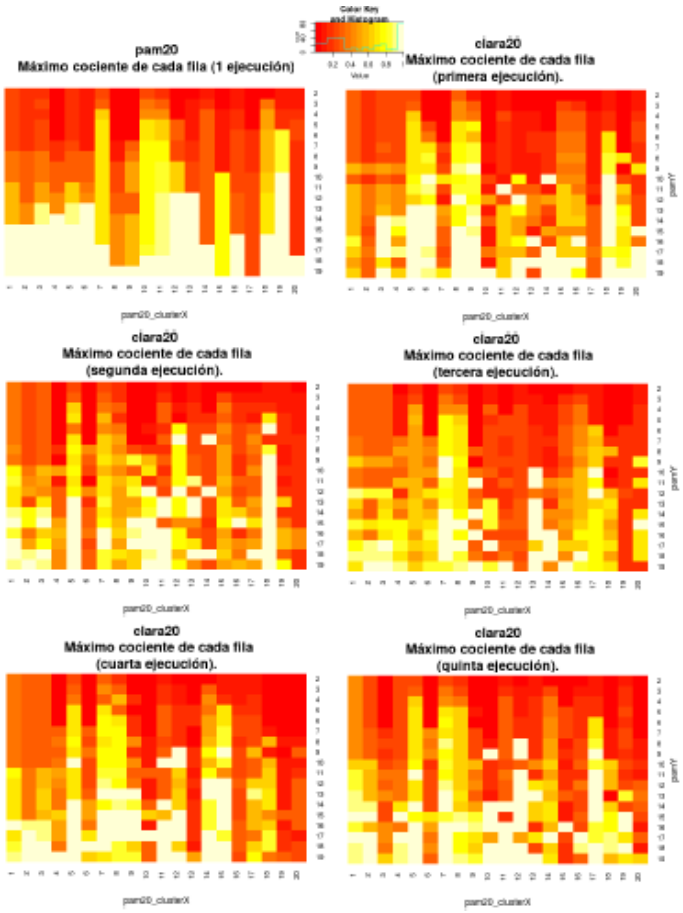

Figura 3. Heatmap de matrices $\mathcal{J}$.

- Revisión de filas/columnas duplicadas, atributos vacíos o de valor único.
- Tranformación de atributos y creación de características: los datos temporales (fechas) por sugerencia clínica pasan a dos atributos (mes/año) para analizar estacionalidad.
- Dependencia de atributos temporales: ej. $cultivosPrimeras72h$ y $cultivosDespues1dia$ (booleanos) pasan a 1 único atributo $periodoCultivo$ multivaluado ($< 3, 3-10, > 10$).
- No se aplican técnicas de reducción dimensional debido a la necesidad de interpretabilidad durante el proceso por parte del médico.

En referente a la discretización de atributos, se han tenido en cuenta métodos de discretización tanto supervisada como no supervisada. Sin embargo, la discretización en datos clínicos tiene un gran impacto en los modelos aprendidos [20] y por tanto se ha guiado por conocimiento médico. Por ejemplo, la edad ha sido discretizada de acuerdo con la clasificación clínica estándar (noenatal, pediátrica, adulta y anciana).

Tras este proceso la vista minable de la base de datos se compone de 1.768 filas y 83 atributos que resumimos en el Cuadro I.

El segundo paso de la metodologías es la selección de algoritmos de agrupamiento y elección de parámetros. En este experimento hemos seleccionado algoritmos $k-medioids$: PAM y CLARA [21]. Se han elegido algoritmos clásicos al existir antecedentes en la literatura clínica y con el fin de facilitar la interpretación del proceso por parte del médico. De acuerdo con el número total de pacientes y con la epidemiología local, se fijó un número máximo de posibles subgrupos, eligiendo un parámetro $K = [1, \ldots, 20]$. La función $M - Trazas$ ha sido desarrollada en *R* (versión 3.3.2) usando las implementaciones de las funciones de $Agrupamiento$ con el paquete *cluster* (versión 2.0.5). En este ejemplo ilustrativo, dado el caracter no determinista del algoritmo CLARA, éste se ha calculado 5 veces mientras que PAM únicamente 1 vez. Esto significa la obtención de un total de 1.254 clústeres, resultantes de analizar 120 clústeres para particiones para $K = 20$ con 1.134 clústeres para las particiones con $K = [2, \ldots, 19]$.

Tabla I
Datos

Paciente y Antibiograma (4 y 3 atributos)			
Atributo	**Contenido**	**Atributo**	**Contenido**
Sexo	1140/628	Edad	60/27/652/1029
T. Ingreso	2015-2016	Microorg.	$[Aureus, \ldots, Faec]$
Susceptib.	Res(21)/Sen(1747)	CMI	$[0{,}25, \ldots, 4]$
Cultivo (6 atributos)			
Atributo	**Contenido**	**Atributo**	**Contenido**
Tipo	$[Sangre, \ldots, LCR]$	Realización	2015-2016
Per.Cultivo	$[< 3, 3-10, > 10]$	Servicio	$[UCI, \ldots, URG]$
Contexto Tratamiento (12 atributos)			
Atributo	**Contenido**	**Atributo**	**Contenido**
Vanco_year	yes/no	Vanco_days	yes/no
⋮	⋮	⋮	⋮
Contexto Flora (144 atributos)			
Atributo	**Contenido**	**Atributo**	**Contenido**
MARSA_year	yes/no	Faec_year	yes/no
⋮	⋮	⋮	⋮

Debido al número de clústeres, se han preseleccionado los clústeres $C_{x,20}$ cuya $Mean(\sum_2^K \mathcal{J}_{x,i}) > 0{,}7$ y $Median(\sum_2^K \mathcal{J}_{x,i}) > 0{,}7$. Esta medida adicional ha permitido evitar al experto el estudio manual de gran número de clústeres de total irrelevancia, reduciendo un 92 % el número de clústeres a estudiar.

La Fig. 3 muestra el resultado visual de las matrices $\mathcal{J}$ para su selección por parte del experto. Finalmente, han sido seleccionados para el estudio los clúseres: *pan20_cluster7*, *pan20_cluster10*, *pan20_cluster11*, *pan20_cluster19*, *clara20_cluster18* (ejec. 1), *clara20_cluster7* (ejec. 2), *clara20_cluster7* (ejec. 3), *clara20_cluster15* (ejec. 4) y *clara20_cluster12* (ejec. 5).

Actualmente se están analizando los grupos de pacientes de estos clústeres elegidos para determinar su relevancia clínica.

IV. Conclusiones

En este trabajo se aborda el problema de la fenotipado de pacientes para el tratamiento antibiótico. En concreto, se propone una metodología para la búsqueda de subgrupos de individuos con características comunes mediante la adaptación

de algoritmos de agrupamiento y visualización de datos. Se ilustra la utilidad de esta metodología en un caso clínico real.

Desde el punto de vista computacional, la principal contribución es la propuesta de utilización de algoritmos de agrupamiento, donde se evalúan los clústeres para conformar sugbrupos. Mientras que existe una gran tradición en el estudio de validez de clústeres mediante CVI [13], [15] para obtener la mejor partición posible, en este trabajo utilizamos dichas técnicas para extraer subgrupos.

Debido a la fuerte componente aplicada al dominio médico, un aspecto esencial de la metodología es la implicación del experto durante todo el proceso [7], [10]. Por tanto, el proceso de obtención de subgrupos debe ser interpretable, utiLizando algoritmos que permitan la trazabilidad de los modelos (identificando el paciente original) y apoyado mediante técnicas de visualización.

Hay que indicar que el incremento del número de ejecuciones de los algoritmos de agrupamiento aumenta el número de clúster candidatos a analizar. Esto podría llegar a suponer un cuello de botella para el posterior análisis semiautomático. No obtstante el uso de técnicas de visualización y métodos de estadística descriptiva ayudan a descartar un gran número de candidatos. En el experimento descrito en la Sec. III, se computaron 114 particiones obteniendo 1.254 clústeres, pero únicamente los expertos deben revisar un 8 % de los mismos.

Entre las líneas de trabajo futura destacamos la exploración y evaluación de técnicas de visualización de trazas y el análisis de otros algoritmos de partición para gestionar el problema del desbalanceo de datos. Desde un punto de vista aplicado, se seguirá desarrollando la metodología propuesta con el fin de identificar nuevos fenotipos en el ámbito de las resistencias antibióticas.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio de Economía y Competitividad y fondos FEDER a través del proyecto WASPSS (Ref: TIN2013-45491-R).

Referencias

[1] F. Palacios, M. Campos, J. Juarez, S. Cosgrove, E. Avdic, B. Canovas-Segura, A. Morales, M. Martinez-Nunez, T. Molina-Garcia, P. Garcia-Hierro, and J. Cacho-Calvo, "A clinical decision support system for an antimicrobial stewardship program," in *HEALTHINF 2016 - 9th International Conference on Health Informatics, Proceedings; Part of 9th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2016*. SciTePress, 2016, pp. 496–501.

[2] B. Cánovas-Segura, M. Campos, A. Morales, J. M. Juarez, and F. Palacios, "Development of a clinical decision support system for antibiotic management in a hospital environment," *Progress in Artificial Intelligence*, vol. 5, no. 3, pp. 181–197, Aug 2016. [Online]. Available: https://doi.org/10.1007/s13748-016-0089-x

[3] A. Martin, "Subgroup discovery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 1, pp. 35–49, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1144

[4] N. Lavrač, B. Cestnik, D. Gamberger, and P. Flach, "Decision support through subgroup discovery: Three case studies and the lessons learned," *Mach. Learn.*, vol. 57, no. 1-2, pp. 115–143, Oct. 2004. [Online]. Available: http://dx.doi.org/10.1023/B:MACH.0000035474.48771.cd

[5] M. Mampaey, S. Nijssen, A. Feelders, and A. Knobbe, "Efficient algorithms for finding richer subgroup descriptions in numeric and nominal data," in *2012 IEEE 12th International Conference on Data Mining*, Dec 2012, pp. 499–508.

[6] M. Atzmueller, "Profiling examiners using intelligent subgroup mining," in *In Proc. 10th Intl. Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2005*, 2005, pp. 46–51.

[7] J. Lu, W. Chen, Y. Ma, J. Ke, Z. Li, F. Zhang, and R. Maciejewski, "Recent progress and trends in predictive visual analytics," *Frontiers of Computer Science*, vol. 11, no. 2, pp. 192–207, Apr 2017. [Online]. Available: https://doi.org/10.1007/s11704-016-6028-y

[8] T. Mühlbacher, H. Piringer, S. Gratzl, M. Sedlmair, and M. Streit, "Opening the black box: Strategies for increased user involvement in existing algorithm implementations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1643–1652, 2014.

[9] J. Krause, A. Perer, and E. Bertini, "Using visual analytics to interpret predictive machine learning models," *arXiv*, vol. abs/1606.05685, 2016. [Online]. Available: http://arxiv.org/abs/1606.05685

[10] P. Tamagnini, J. Krause, A. Dasgupta, and E. Bertini, "Interpreting black-box classifiers using instance-level visual explanations," in *Proceedings of the 2Nd Workshop on Human-In-the-Loop Data Analytics*, ser. HILDA'17. New York, NY, USA: ACM, 2017, pp. 6:1–6:6.

[11] T. von Landesberger, D. W. Fellner, and R. A. Ruddle, "Visualization system requirements for data processing pipeline design and optimization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 8, pp. 2028–2041, Aug 2017.

[12] H. Ltifi, E. Benmohamed, C. Kolski, and M. B. Ayed, "Enhanced visual data mining process for dynamic decision-making," *Knowledge-Based Systems*, vol. 112, pp. 166 – 181, 2016. [Online]. Available: https://doi.org/10.1016/j.knosys.2016.09.009

[13] Y. Lei, J. C. Bezdek, S. Romano, N. X. Vinh, J. Chan, and J. Bailey, "Ground truth bias in external cluster validity indices," *Pattern Recognition*, vol. 65, pp. 58 – 70, 2017.

[14] B. Kim, H. Lee, and P. Kang, "Integrating cluster validity indices based on data envelopment analysis," *Applied Soft Computing*, vol. 64, pp. 94 – 108, 2018.

[15] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. The address: Academic Press, 2008.

[16] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2, pp. 107–145, Dec 2001.

[17] T. Mühlbacher and H. Piringer, "A partition-based framework for building and validating regression models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 1962–1971, Dec 2013.

[18] E. Mosqueira-Rey and V. Moret-Bonillo, "Validation of intelligent systems: a critical study and a tool," *Expert Systems with Applications*, vol. 18, no. 1, pp. 1 – 16, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417499000457

[19] M. Kendall and J. Gibbons, *Correlation Methods*, 5th ed. Oxford University Press, 1990.

[20] I. J. Casanova, M. Campos, J. M. Juarez, A. Fernandez-Fernandez-Arroyo, and J. A. Lorente, "Impact of time series discretization on intensive care burn unit survival classification," *Progress in Artificial Intelligence*, vol. 7, no. 1, pp. 41–53, Mar 2018. [Online]. Available: https://doi.org/10.1007/s13748-017-0130-8

[21] X. Jin and J. Han, *K-Medoids Clustering*. Boston, MA: Springer US, 2017, pp. 697–700. [Online]. Available: https://doi.org/10.1007/978-1-4899-7687-1_432