XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2018)

CAEPIA 9:
Clasificación y Agrupamiento

Learning Planning Action Models with Numerical Information and Logic Relationships using Classification Techniques*

*Note: The full contents of this paper have been published in the volume *Lecture Notes in Artificial Intelligence 11160* (LNAI 11160)

José Á. Segura-Muros, Raúl Pérez, Juan Fernández-Olivares
University of Granada
Granada, Spain
{josesegmur,fgr,faro}@decsai.ugr.es

***Abstract*—The task of constructing a planning domain is difficult and requires time and vast knowledge about the problem to be solved. This paper describes PlanMiner-O3 a planning domain learner designed to alleviate this problem, based on the use of a classification algorithm, able to learn planning action models from noisy and partially observed logic states. PlanMiner-O3 is able to learn continuous numerical fluents as well as simple logical relations between them. Testing was realized with benchmark domains obtained from the International Planning Competition and the results demonstrate PlanMiner-O3's capabilities of learning planning domains.**

***Index Terms*—**

Adapting Hierarchical Multiclass Classification to changes in the target concept*

*Note: The full contents of this paper have been published in the volume *Lecture Notes in Artificial Intelligence 11160* (LNAI 11160)

Daniel Silva-Palacios, Cesar Ferri, M. Jose Ramirez-Quintana
DSIC
Universitat Politècnica de València
Valencia, Spain
dasilpa@posgrado.upv.es, {cferri,mramirez}@dsic.upv.es

***Abstract*—Machine learning models often need to be adapted to new contexts, for instance, to deal with situations where the target concept changes. In hierarchical classification, the modularity and flexibility of learning techniques allows us to deal directly with changes in the learning problem by readapting the structure of the model, instead of having to retrain the model from the scratch. In this work, we propose a method for adapting hierarchical models to changes in the target classes. We experimentally evaluate our method over different datasets. The results show that our novel approach improves the original model, and compared to the retraining approach, it performs quite competitive while it implies a significantly smaller computational cost.**

***Index Terms*—Hierarchical, Classification, Adaptation, Novelty**

Clasificación ordinal de los grados de afectación de la enfermedad de Parkinson empleando imágenes de transportadores presinápticos de dopamina

J. Camacho-Cañamón * julio.camacho@uco.es
María-Victoria Guiote *UGC Medicina Nuclear* §
Antonio-María Santos-Bueno *UGC Medicina Nuclear* §
Ester Rodríguez-Cáceres *Equipo Provincial TICS* §
Elvira Carmona-Asenjo *UGC Medicina Nuclear* §
Juan-Antonio Vallejo-Casas *UGC Medicina Nuclear* § jantonio.vallejo.sspa@juntadeandalucia.es
P. A. Gutiérrez * pagutierrez@uco.es
C. Hervás-Martínez * chervas@uco.es

* *Dpto. de Informática y Análisis Numérico*
Universidad de Córdoba
Córdoba, España

§ *Hospital Universitario "Reina Sofía"*
Universidad de Córdoba. IMIBIC.
Córdoba, España

***Resumen*—La enfermedad de Parkinson se caracteriza por un descenso de la densidad de transportadores presinápticos de dopamina en los núcleos de la base. El método habitual de clasificación está basado en la observación y el análisis cualitativo de las imágenes obtenidas tras la administración de ^{123}I-ioflupano al paciente que se va a diagnosticar. De esta forma, las técnicas recientes de neuroimagen, como la imagen dopaminérgica utilizando tomografía computarizada por emisión de fotón único (SPECT-CT) con ^{123}I-ioflupano (DaTSCAN ®), han demostrado detectar la enfermedad, incluso en etapas tempranas. Sin embargo, los comités internacionales recomiendan un análisis cuantitativo, asociado a la construcción de modelos de apoyo que complementen el diagnóstico visual. El objetivo del presente estudio es establecer un sistema de apoyo a la decisión, mediante la clasificación ordinal de las imágenes asociadas a los diferentes grados de afectación de la enfermedad mediante técnicas de aprendizaje automático. La base de datos utilizada está formada por 316 estudios realizados a pacientes entre septiembre de 2015 y mayo de 2018, distribuidos en tres grupos: 191 no padecen la enfermedad de Parkinson, 55 la padecen con un nivel de afectación leve y 70 con un nivel de afectación grave. Tras la administración intravenosa de 5 mCi (185 MBq), se realizó una SPECT-CT, preprocesando y normalizando espacialmente las imágenes. Como clasificador ordinal utilizamos un método de regresión logística, que nos permite obtener las características (vóxeles de la imagen) más informativas para la tarea de clasificación. El mejor modelo alcanzó un error absoluto medio máximo (MMAE) de 0,4857, tras la aplicación de un diseño experimental de tipo 5-*fold*. El análisis de los vóxeles más informativos, de acuerdo con el modelo obtenido, destaca regiones del cerebro que no son consideradas habitualmente por los especialistas para el diagnóstico visual.**

***Index Terms*—Enfermedad de Parkinson, SPECT-CT, aprendizaje automático, clasificación ordinal, regresión logística ordinal, imagen médica**

Este trabajo ha sido realizado gracias al apoyo económico derivado de los proyectos TIN2017-85887-C2-1-P y TIN2017-90567-REDT del Ministerio de Economía, Industria y Competitividad de España y, también, por el apoyo de los fondos FEDER de la Unión Europea. También se agradece a la Universidad de Córdoba por haber premiado al proyecto "Clasificación y evaluación automática de los grados de párkinson" UCO-SOCIAL-INNOVA, en el marco del cual se realiza este trabajo, en el III Plan Propio Galileo de Innovación y Transferencia. Los autores agradecen a *NVIDIA Corporation* la cesión de recursos computacionales a través del *GPU Grant Program*.

I. Introducción

Una de las características neuropatológicas de la enfermedad de Parkinson es un sustancial descenso de dopamina en los núcleos basales (caudado y putamen) debido a la disminución progresiva de la densidad de transportadores presinápticos [1]. La densidad de transportador presináptico de dopamina se puede detectar mediante técnicas de neuroimagen, que actualmente constituyen una práctica ordinaria en el diagnóstico de trastornos neurodegenerativos como la enfermedad de Parkinson. El ^{123}I-ioflupano (DaTSCAN ®, General Electrics Healtcare Limited, Little Chalfont. Bucks HP79NA U.K.) es un radiofármaco, ampliamente empleado para este fin, que se une a los transportadores presinápticos de dopamina en el cuerpo estriado y permite evaluar la densidad de estos con alta sensibilidad [2].

Actualmente, se combinan sistemas de cuantificación semi-automáticos, que analizan las imágenes, con los diagnósticos visuales dependientes de un observador especializado y, juntos, son capaces de distinguir entre las clases: control, patológico con nivel de afectación leve y patológico con nivel de afectación grave. En el caso del diagnóstico de la enfermedad de Parkinson, se aborda el problema con un enfoque clásico, es decir, mediante la cuantificación de la pérdida de dopamina neuronal en el estriado [3]. En este sentido, el Comité de la Asociación Europea de Medicina Nuclear y Neuroimagen recomienda el análisis cuantitativo, asociado a la construcción de modelos computacionales de apoyo que complementen el diagnóstico visual [4]. Debido a que la enfermedad de Parkinson afecta, fundamentalmente, a las neuronas que manejan

dopamina como neurotransmisor principal, la mayoría de los modelos analíticos asociados al uso de ioflupano se centran en el cuerpo estriado. En este estudio, nos centraremos en el cuerpo estriado, pero incluiremos el resto de la imagen cerebral en el análisis, dado que una hipótesis fundamental de este trabajo plantea la existencia de otras zonas del cerebro en las que los efectos de la dopamina también pueden ayudar a realizar una clasificación más precisa.

La mayoría de trabajos abordan el problema de clasificación binaria (pacientes de control y patológicos), utilizando imágenes funcionales, y empleando técnicas de aprendizaje automático basadas en regiones de interés (caudado y putamen) [5]. El análisis basado en estas regiones generalmente se justifica en que hay pérdida de actividad dopaminérgica en el putamen en relación con el caudado para los enfermos de Parkinson. Sin embargo, el rendimiento de modelos de clasificación binaria que utilizan solo estas regiones de interés es limitado. En primer lugar, no está claro si una variable aleatoria asociada a un vóxel (unidad volumétrica mínima de la imagen) es significativamente importante para la tarea de clasificación. Ciertos vóxeles, dentro de estas regiones, pueden sufrir una mayor pérdida de transportador de dopamina que otros y, por lo tanto, pueden ser más informativos. En segundo lugar, no se debe asumir que solamente el caudado y el putamen sean las zonas a partir de las cuáles se extraigan las características (vóxeles) que utilice el clasificador.

Se sabe que las estructuras extraestriadas están involucradas en la enfermedad de Parkinson. Y que, además, el globo pálido está involucrado en los subtipos de la enfermedad de Parkinson [6], así que esta región cerebral también se ha incluido en nuestro análisis. El hecho de incluir todos los vóxeles de estas regiones como características discriminantes reporta mayor interés que asumir *a priori* que ciertos vóxeles son los más importantes, según conocimiento especializado. Por lo tanto, nuestro estudio desarrolla un algoritmo basado en vóxeles en lugar de basado en las regiones de interés clásicas.

El modelo de aprendizaje automático utilizado para clasificación será una regresión logística ordinal regularizada [7]. Con la regularización del modelo se pretende minimizar el número de características que intervienen en la clasificación, es decir, vóxeles, mediante la minimización del valor absoluto de los coeficientes que tienen asociados. El modelo es una combinación lineal de los coeficientes asociados a los vóxeles, cuyo valor se emplea para calcular las probabilidades de pertenecer a las tres clases: normal (clase 1), patológico con nivel de afectación leve (clase 2) y patológico con nivel de afectación grave (clase 3). Así, mediante esta metodología estimaremos la probabilidad de pertenencia a una de las tres clases asociadas al nivel de afectación de la enfermedad de Parkinson. Además de la clasificación, otro de los objetivos de este estudio es reconocer cuáles son los "vóxeles informativos", término que empleamos para denotar aquellos que son significativamente útiles para la clasificación, es decir, más determinantes para construir el clasificador. Estos serán un subconjunto de todos los vóxeles analizados de la imagen completa y, previsiblemente, también un subconjunto de todos los vóxeles afectados por la enfermedad de Parkinson. Con el fin de crear modelos de clasificación aún más sencillos se evaluará el uso de un selector de características previo (ReliefF [8]).

Los resultados indican que el diagnóstico utilizando todos los vóxeles de la imagen es factible, obteniéndose un rendimiento aceptable para la complejidad del problema abordado. Además, se detectan vóxeles útiles para la tarea de clasificación en regiones que no habían sido previamente consideradas en la literatura. Por último, el uso del selector de características ReliefF no mejora el rendimiento del clasificador.

En la Sección II se realizará una revisión de los trabajos relacionados con el problema de clasificación de pacientes de párkinson. Posteriormente, en la Sección III, se detallarán los conjuntos de datos utilizados y su forma de obtención, así como la metodología empleada para generar los modelos de clasificación. Para concluir, en la Sección IV, se mostrarán los resultados obtenidos y se realizará una discusión de los mismos. La Sección V expondrá las conclusiones obtenidas tras este trabajo y las posibles futuras mejoras.

II. Estado del Arte

Son muchos los métodos basados en aprendizaje automático utilizados en los sistemas de apoyo al diagnóstico médico, como pueden ser las máquinas de vectores soporte (MVS), la regresión logística (RL), los árboles de decisión, las redes bayesianas o las redes neuronales. Todas ellas son importantes en tanto en cuanto puedan ayudar al médico clínico a realizar un diagnóstico temprano de la enfermedad, permitan planificar su tratamiento y posibiliten la monitorización de la progresión de la enfermedad [5]. Estas metodologías están siendo ampliamente utilizadas en neuroimagen [9], debido a las siguientes ventajas: permiten la clasificación a nivel de individuo más que a nivel de grupo, por lo que los resultados tienen un alto potencial de traslación a la práctica clínica; son técnicas multivariantes y supervisadas que tienen en cuenta las características de los patrones (imágenes volumétricas, en este caso) distribuidas en un espacio de características complejo de alta dimensionalidad para entrenar el modelo de clasificación y, una vez entrenado, poder clasificar nuevos pacientes en función del modelo obtenido.

Algunos trabajos en los que se utilizan las MVS como clasificador binario [10], [11] están basados en extraer los vóxeles correspondientes al estriado. En [10], los autores realizan una descomposición de datos utilizando mínimos cuadrados parciales (*Partial Least Squares*) seguida por la utilización del clasificador MVS. En [11], los autores utilizan, una vez extraídos los vóxeles, un clasificador MVS lineal, contando con 208 imágenes DaTSCAN y utilizando máscaras para la selección de vóxeles.

Por otra parte, el modelo de RL determina la probabilidad de tener la enfermedad de Parkinson para un paciente, utilizando como características explicativas los vóxeles de las imágenes normalizadas. Esta metodología puede ser útil también para clasificar pacientes en diferentes categorías de riesgo de padecer la enfermedad, como sugieren algunos estudios utilizando

SPECT-CT en imágenes con ^{123}I-ioflupano, que representan la progresión de degeneración dopaminérgica de la enfermedad de Parkinson [12].

III. Material y métodos

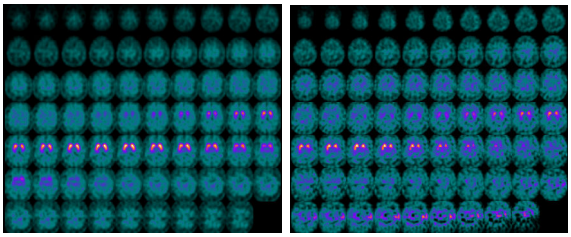
III-A. Material

Con el fin de evaluar la metodología propuesta en este trabajo, hemos utilizado los datos de 316 estudios con ^{123}I-ioflupano en pacientes derivados para evaluación de trastornos del movimiento realizados en el Hospital Universitario 'Reina Sofía' de Córdoba, en el periodo que va desde septiembre de 2015 hasta mayo de 2018. Los pacientes se distribuyen en tres clases: 191 no padecen la enfermedad de Parkinson (normal, clase 1), 55 padecen la enfermedad de Parkinson con un nivel de afectación leve (patológico, clase 2) y 70 padecen la enfermedad de Parkinson con un nivel de afectación grave (patológico, clase 3). La relación entre los valores muestrales asociados a padecer o no la enfermedad en cualquier de sus niveles de afectación se acercan a la relación poblacional entre dichas categorías. El diagnóstico definitivo se establece por la combinación de pruebas clínicas y complementarias realizadas por la Unidad de Neurología del centro hospitalario. La media de edad de los pacientes es de 70.46 años (35-89) con una desviación típica de 11,85 años. Las imágenes de SPECT-CT se adquieren según el protocolo indicado en [13].

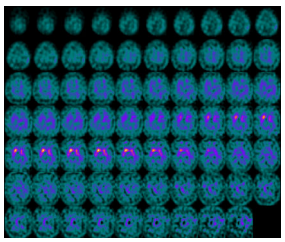
Tres especialistas de la Unidad de Medicina Nuclear, han procesado, interpretado y evaluado las imágenes. La evaluación visual se ha establecido considerando exclusivamente el criterio normal y patológico en función del grado basándose en criterios comunes preestablecidos [14], y después de llegar a un informe de consenso entre los tres especialistas. Se consideró que un paciente era normal cuando, en la imagen volumétrica correspondiente a su estudio, aparecía una simetría bilateral en los núcleos basales, caudado y putamen; enfermedad de Parkinson con un nivel de afectación leve, cuando había una asimetría o reducción completa en la actividad del putamen; y con un nivel de afectación grave cuando había una ausencia bilateral de actividad en el caudado y putamen. Véase la figura 1, donde se muestra una imagen de un paciente normal en 1(a), la de un paciente patológico con un nivel de afectación leve en 1(b) y la de un paciente patológico con un nivel de afectación grave en 1(c).

III-B. Preprocesado de la imagen

El preprocesado de las imágenes consta de dos partes: *normalización espacial* y *recuperación de información*. Las imágenes suministradas inicialmente están sin procesar, por tanto hay que normalizarlas espacialmente, de forma que los mismos vóxeles representen las mismas características en todos los estudios. Para ello se utiliza el *software* PETRA [15] basado en el *framework* SPM8 [16], que permite leer y reunir varios archivos DICOM para formar una sola imagen espacial. Para realizar la normalización espacial se coordinan los ejes tridimensionales con la comisura anterior cerebral de la imagen, cuyas coordenadas están almacenadas en el formato DICOM de la imagen. No obstante, el método de normalización espacial se explica con detalle en [15].



(a) Normal (b) Patológico afectación leve



(c) Patológico afectación grave

Figura 1. Imágenes de pacientes normal (a), patológico con nivel de afectación leve (b) y patológico con nivel de afectación grave (c).

Aplicando esta normalización a cada imagen, se obtiene una matriz tridimensional de 79×95×69 vóxeles, es decir, $V = 517{,}845$ vóxeles por cada estudio. Este elevado número de vóxeles implica que es necesario aplicar técnicas de aprendizaje automático preparadas para un gran número de características. Tras la normalización espacial, es necesario hacer una búsqueda de valores perdidos en los vóxeles de cada imagen. Los valores perdidos serán sustituidos por la media de los valores que tienen los vóxeles del eje horizontal en el que se encuentra el valor perdido. A este proceso se le conoce como recuperación de información. Una vez preprocesadas las imágenes utilizaremos todos los vóxeles de cada paciente como características de entrada al clasificador, como se explica en la siguiente sección.

III-C. Metodología

III-C1. Notación y terminología: En primer lugar, definimos algunos términos y notaciones. La imagen del paciente i-ésimo está representada por el vector $\mathbf{x}_i$, donde $i = 1, ..., N$, siendo N el número de estudios (316). Si v es la posición de un vóxel concreto en una imagen, $v = 1, ..., V$, entonces $\mathbf{x}_i(v)$ es el valor del vóxel v en la imagen del paciente i.

Leyendo la imagen de cada paciente i como una serie o conjunto de valores (vóxeles), la imagen $\mathbf{x}_i$ se puede entender como un vector de tamaño $V \times 1$. Por tanto, $\mathbf{x}_i(v)$ es la componente v del vector de dimensiones $V \times 1$. Dicho de otra forma, cualquiera de estos vectores (de tamaño $V \times 1$) serviría para reconstruir una imagen 79×95×69, la cual puede ser representada como una imagen 3D. De este modo, podremos considerar y visualizar un vector concreto.

Respecto a la clase de cada paciente i, la representaremos como una variable categórica y_i con tres categorías ordinales,

siendo $y_i = 1$ si el paciente i pertenece al grupo de control (normal), $y_i = 2$ si el paciente i padece enfermedad de Parkinson con un nivel de afectación leve, e $y_i = 3$ indicará que el nivel de afectación es grave.

Suponemos que este conjunto de patrones de entrenamiento es una realización de un conjunto de V variables aleatorias independientes e idénticamente distribuidas, así como de la variable asociada a la clase y_i de cada paciente. El conjunto de datos será representado por $D = \{(\mathbf{x}_i, y_i)\}, i = 1, ..., N$. El número de pacientes de cada una de las clases de interés (normal, afección leve y afección grave) será denotado como N_1, N_2 y N_3, respectivamente.

III-C2. Regresión logística ordinal regularizada: Consideraremos un modelo de RL ordinal [17] para estimar la probabilidad de que un paciente tenga un determinado nivel de afectación de la enfermedad de Parkinson, es decir, el modelo estima la probabilidad condicional $P(y_i = q|\mathbf{x_i}), q \in \{1, 2, 3\}$.

Suponemos que, al ser la variable dependiente ordinal, el mejor modelo que se ajusta a la nube de puntos es un modelo de RL ordinal. Como ya se mencionó, una aproximación similar, pero para clasificación binaria, [5] ha sido empleada para la detección de la enfermedad de Parkinson.

El modelo de RL ordinal (de forma más especifica, *Proportional Odds Model*, POM, [17]) es un modelo de umbral [18] que modela la respuesta categórica ordinal mediante una función lineal de proyección $f(\mathbf{x}_i, \boldsymbol{\theta})$ común para todas las clases y un vector de umbrales $\boldsymbol{\beta}$. Debido a que en nuestro problema hay tres clases, se cumple $\boldsymbol{\beta} = \{b_1, b_2\}$, siendo $b_1 < b_2$. La probabilidad *a posteriori* de pertenencia a una clase se obtiene modelando la probabilidad de pertenencia acumulada a toda clase menor o igual a la evaluada:

$$P(y_i \leq q|\mathbf{x}_i, \hat{\mathbf{s}}) = \sum_{j=1}^{q} P(y_i = j|\mathbf{x}_i, \hat{\mathbf{s}}),$$
$$P(y_i = q|\mathbf{x}_i, \hat{\mathbf{s}}) = P(y_i \leq q|\mathbf{x}_i, \hat{\mathbf{s}}) - P(y_i \leq (q-1)|\mathbf{x}_i, \hat{\mathbf{s}}),$$

donde $\hat{\mathbf{s}} = \{\boldsymbol{\theta}, \boldsymbol{\beta}\}$ incluye todos los parámetros libres del modelo. Las probabilidades acumuladas se aproximan mediante la inversa de la función `logit` (o función sigmoide):

$$P(y_i \leq q|\mathbf{x}_i, \hat{\mathbf{s}}) = \sigma\left(f(\mathbf{x}_i, \boldsymbol{\theta})\right) = \frac{1}{1 + \exp(-f(\mathbf{x}_i, \boldsymbol{\theta}))},$$

para $q = \{1, 2\}$. Por definición, $P(y_i \leq 0|\mathbf{x}_i, \mathbf{s}) = 0$ y $P(y_i \leq 3|\mathbf{x}_i, \hat{\mathbf{s}}) = 1$. La función de proyección es lineal, por lo que $f(\mathbf{x}_i, \boldsymbol{\theta}) = \boldsymbol{\theta}^{\mathrm{T}} \cdot \mathbf{x}_i$. Por ello, la expresión final del modelo es:

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i, \hat{\mathbf{s}}) =& \sigma(\boldsymbol{\theta}^{\mathrm{T}} \cdot \mathbf{x}_i - b_1), \\ P(y_i = 2|\mathbf{x}_i, \hat{\mathbf{s}}) =& \sigma(\boldsymbol{\theta}^{\mathrm{T}} \cdot \mathbf{x}_i - b_2) - \sigma(\boldsymbol{\theta}^{\mathrm{T}} \cdot \mathbf{x}_i - b_1), \\ P(y_i = 3|\mathbf{x}_i, \hat{\mathbf{s}}) =& 1 - \sigma(\boldsymbol{\theta}^{\mathrm{T}} \cdot \mathbf{x}_i - b_2). \end{aligned} \quad (1)$$

Para la estimación de los parámetros $\hat{\mathbf{s}} = \{\boldsymbol{\theta}, \boldsymbol{\beta}\}$, se utiliza el método de máxima verosimilitud, que minimiza la siguiente función de entropía cruzada:

$$L(\mathbf{s}, D) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{q=1}^{3} [[y_i = q]] \ln\left(P(y_i = q|\mathbf{x}_i)\right),$$

donde $[[c]] = 1$ si la condición c es cierta ($[[c]] = 0$, en caso contrario).

Es bien conocido que un sobre-ajuste del conjunto D puede incurrir en un modelo con varianza mayor, empobreciendo el rendimiento del conjunto de *test*. Una forma de controlar este fenómeno de sobre-entrenamiento es el uso de un término de regularización, que evita valores altos de los coeficientes, reduciendo la varianza y aumentando el sesgo, a costa de empeorar el ajuste al conjunto de entrenamiento. Para ello, modificamos la función de error de la siguiente forma:

$$L_2(\mathbf{s}, D) = L(\mathbf{s}, D) + \lambda \cdot \sum_{i=1}^{V} \theta_i^2, \quad (2)$$

donde λ es el parámetro de regularización. Dicho parámetro, que toma valores positivos, representa la importancia que se le da a la regularización frente a la minimización del error y debe ser ajustado por el investigador (normalmente en base a algún tipo de proceso de validación). Utilizando las Ecuaciones (1) y (2), el aprendizaje del modelo puede realizarse mediante las derivadas analíticas de $L_2(\mathbf{s}, D)$ con respecto a cada uno de los parámetros $s_i \in \{\boldsymbol{\theta}, \boldsymbol{\beta}\}$, mediante algún método de optimización (en el caso de la RL, se suelen emplear métodos de segundo orden). Estas derivadas pueden consultarse en [19]. Tal y como se discute en [20], la formulación de este modelo de RL ordinal se ajusta a una función de pérdida de tipo *Immediate-Threshold* (umbrales adyacentes), de forma que, para cada ejemplo etiquetado con y_i, solo se penaliza el error que cometen los umbrales que limitan el segmento correcto (es decir, b_{y_i} y $b_{y_i - 1}$).

A partir de este modelo de RL, se pueden obtener, lo que hemos denominado, *Componentes Logísticas Principales* (CLP), que son aquellos vóxeles cuyo respectivo coeficiente θ_i es más elevado en valor absoluto. Es decir, estos vóxeles son los que más han influido en la función $f(\mathbf{x}_i, \boldsymbol{\theta})$, por lo que son los vóxeles más importantes para clasificar el sujeto en cualquiera de las tres clases consideradas. Las CLP tienen una estructura de componentes binarias de dimensión $V \times 1$, donde 1 significa que ese vóxel es relevante y 0 que no lo es. Por lo tanto, como hemos explicado previamente, se pueden llegar a representar como una imagen 3D. Así, se puede apreciar qué zonas son las que más afectan a la clasificación según el nivel de afectación de la enfermedad de Parkinson, como se ilustrará en el apartado de resultados.

III-C3. Diseño experimental: Para la evaluación de la bondad del modelo, aplicamos un diseño experimental de validación cruzada de tipo *5-fold*. Los pliegues o *folds* se crearán de forma estratificada, es decir, manteniendo la proporción original de patrones de pacientes normales, leves y graves.

Por cada una de las 5 particiones, realizaremos el aprendizaje estimando el mejor valor para el parámetro λ mediante una validación cruzada anidada: el conjunto de entrenamiento (que incluye 4 pliegues) se divide, de nuevo, en 2 conjuntos estratificados, aplicando un diseño de tipo *2-fold* por cada valor del parámetro λ que va a ser explorado. Los valores explorados para λ son los siguientes $\lambda \in \{10^{-3}, 10^{-2}, ..., 10^{3}\}$.

A la hora de realizar el ajuste de λ elegimos el valor que produzca un menor error de validación. En concreto, utilizamos la función de error $MMAE$ (*Maximum Mean Absolute Error* [21], la cual es específicamente adecuada para problemas desequilibrados de clasificación ordinal). El MAE es una medida de error que tiene en cuenta la ordinalidad de la variable objetivo, $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i^*|$, donde $|y_i - y_i^*|$ es la distancia absoluta entre las etiquetas reales y predichas. El MAE varía desde 0 hasta 2 (que sería la máxima desviación en número de categorías en nuestro problema). Sin embargo, en problemas desequilibrados, las clases más frecuentes pueden dominar el error MAE, enmascarando un rendimiento pobre para las clases menos comunes. Es por ello que el $MMAE$ se define de la siguiente forma, $MMAE = \max\{MAE_1, MAE_2, MAE_3\}$, donde MAE_q es el error MAE teniendo en cuenta solo los patrones de la clase q: $MAE_q = \frac{1}{N_q}\sum_{i=1}^{N}|y_i - y_i^*|, q \in \{1,2,3\}$.

Con la intención de mostrar la mejora obtenida usando todos los vóxeles de las imágenes en lugar de usar solamente las regiones descritas por los especialistas, se han realizado pruebas con la misma experimentación, pero seleccionando previamente aquellas variables que los especialistas sitúan en las zonas estriatales del cerebro.

Igualmente, se han aplicado técnicas de selección de características de tipo filtro, en concreto, el método ReliefF [8]. Estas son aplicadas como paso previo al entrenamiento del modelo con el fin de vislumbrar aquellas características con las que el modelo podría obtener mejores resultados.

Por último, una vez estimado el error mediante el proceso *5-fold*, estimaremos las CLP. Para ello, repetiremos el entrenamiento del modelo, pero considerando todos los pacientes disponibles, con el fin de obtener el mejor modelo posible. De nuevo, se aplicará una validación cruzada anidada de tipo *2-fold* que ajuste el mejor valor de λ en la RL ordinal. En nuestro experimento, para establecer las CLP, escogemos el $0,50\,\%$ de los vóxeles (2590) cuyo θ_i es mayor en valor absoluto.

IV. Resultados

Cada uno de los subconjuntos de entrenamiento generados en el procedimiento *5-fold* se utiliza para calcular la matriz de confusión de su respectivo conjunto de generalización. La suma de las matrices de confusión generadas para cada subconjunto de generalización será el resultado final del modelo, y a partir de ella se calculará el *MMAE* completo. En la tabla I se muestran todos los resultados obtenidos empleando distintas metodologías: considerar todos los vóxeles (y dejar que la regularización seleccione los más importantes), considerar solo las regiones de interés determinadas por los expertos (caudado y putamen) y considerar el método ReliefF de selección de características manteniendo un $5\,\%$, $2\,\%$ y $1\,\%$ de los vóxeles originales. La tabla incluye una referencia a la subfigura de la Figura 2 que incluye la matriz de confusión correspondiente.

Los resultados demuestran que usando todos los vóxeles se consigue el mejor clasificador en *MMAE* y en *CCR*. También cabe destacar que usando las regiones de interés que los especialistas han indicado, se consigue un *MMAE* igual al que se consigue empleando ReliefF y seleccionando el $2\,\%$ de las características más representativas. Sin embargo, el número de vóxeles empleados analizando las regiones de interés es una tercera parte de los que se utilizan con el selector. Esto indica que ReliefF no es adecuado para este problema, posiblemente por estar pensado para clasificación nominal.

Tabla I
Análisis de resultados en función de los vóxeles considerados

Vóxeles	Figura	Nº Vóxeles	$MMAE\downarrow$	$CCR\uparrow$
Todos	2(a)	517845	**0,4857**	**0,7532**
Caudado y putamen	2(b)	3267	0,6727	**0,7532**
RelieF (5 %)	2(c)	25892	*0,6000*	*0,7468*
RelieF (2 %)	2(d)	10356	0,6727	0,7247
RelieF (1 %)	2(e)	5178	0,7091	0,7342

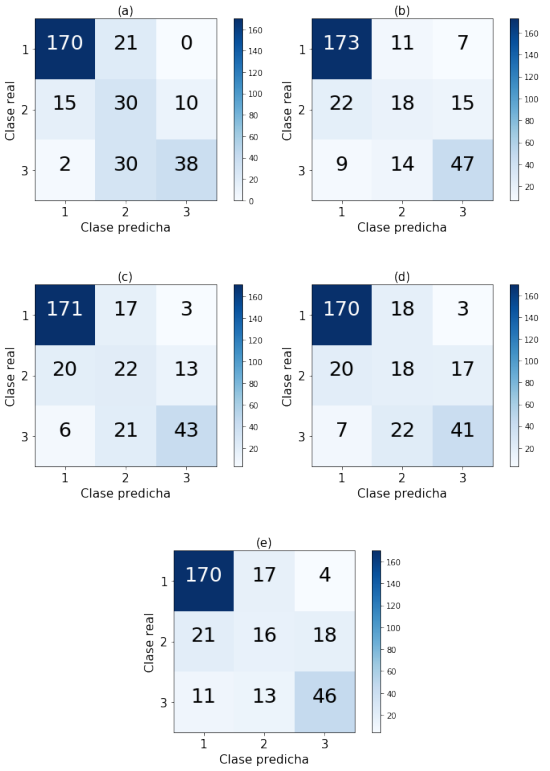
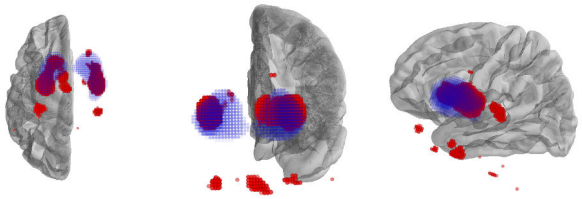


Figura 2. Matrices de confusión de los resultados de cada experimento.

Para finalizar, se visualizan los vóxeles más informativos de la imagen (CLP) tras entrenar los clasificadores con todos los datos y todos los vóxeles (ver Figura 3). Como puede apreciarse, existe una gran cantidad de vóxeles importantes agrupados en los propios núcleos basales, algo que era de esperar antes de realizar la representación. No obstante, debemos prestar especial atención a aquellas zonas que, sin pertenecer a las regiones de interés clásicas, también son muy relevantes para el modelo de RL ordinal en su tarea de clasificación. Estas lesiones extraestriatales coinciden con la tendencia de otros estudios (resonancia magnética, por ejemplo) a evidenciar alteraciones en áreas cerebrales fuera

del estriado. Se corresponden a pequeñas zonas corticales, de predominio temporal y en la región media del lóbulo parietal.



(a) Axial (b) Coronal (c) Sagital

Figura 3. Representación 3D de los vóxeles más importantes (CLP), utilizando PySurfer [22]. En azul se representan los núcleos basales, caudado y putamen. En rojo se representan las CLP.

V. Conclusión

Este artículo presenta un sistema de apoyo a la decisión mediante modelos de clasificación ordinal de pacientes de párkinson a partir de imágenes DaTSCAN. Para su desarrollo, se han empleado técnicas de aprendizaje automático, que incluye RL ordinal, y técnicas de normalización espacial y de recuperación de información. El modelo matemático computacional se considera más adecuado para el campo de la biomedicina dada su alta interpretabilidad.

Los resultados avalan que la metodología que utilizamos, empleando todos los vóxeles para la clasificación (junto con regularización en el modelo), obtiene un mejor rendimiento en la clasificación que si solamente se empleasen los vóxeles de las regiones de interés (núcleos basales) o si se emplease un selector de características nominal (ReliefF).

Respecto a la representación en 3D de las CLP, se confirma que los núcleos basales tienen una gran importancia en la tarea de clasificación de un paciente. Pero también se destacan otras zonas corticales que deberían de tenerse en cuenta para futuros estudios en la evaluación de la enfermedad de Parkinson. La densidad de transportadores presinápticos de dopamina tiene una gran importancia en la tarea de clasificación de un paciente, pero habría que tener en cuenta para futuros estudios otras zonas corticales que tendrían influencia en el diagnóstico diferencial de la enfermedad de Parkinson.

Como futuros trabajos se plantean realizar modificaciones en el método de selección de características nominal (ReliefF) para adaptarlo a una clasificación ordinal, de modo que la selección de características ayude a reducir la dimensionalidad del problema y mejorar los resultados. También se propone la aplicación de otros clasificadores como Procesos Gaussianos con núcleo lineal que permitan estimar todos los parámetros del modelo y calcular su incertidumbre.

Referencias

[1] R. Simon, D. Greenberg, and M. Aminoff, "Clinical neurology 5th edition," *Clinical Neurology*, 2009.

[2] J. Booij, J. B. Habraken, P. Bergmans, G. Tissingh *et al.*, "Imaging of dopamine transporters with iodine-123-fp-cit spect in healthy controls and patients with parkinson's disease," *The Journal of Nuclear Medicine*, vol. 39, no. 11, p. 1879, 1998.

[3] A. Antonini, K. L. Leenders, P. Vontobel, R. P. Maguire, J. Missimer, M. Psylla, and I. Günther, "Complementary pet studies of striatal neuronal function in the differential diagnosis between multiple system atrophy and parkinson's disease." *Brain: a journal of neurology*, vol. 120, no. 12, pp. 2187–2195, 1997.

[4] J. Darcourt, J. Booij, K. Tatsch, A. Varrone, T. Vander Borght, Ö. L. Kapucu, K. Någren, F. Nobili, Z. Walker, and K. Van Laere, "Eanm procedure guidelines for brain neurotransmission spect using 123i-labelled dopamine transporter ligands, version 2," *Eur J Nucl Med Mol*, vol. 37, no. 2, pp. 443–450, 2010.

[5] R. Prashanth, S. D. Roy, P. K. Mandal, and S. Ghosh, "Automatic classification and prediction models for early parkinson's disease diagnosis from spect imaging," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3333–3342, 2014.

[6] A. Rajput, H. Sitte, A. Rajput, M. Fenton, C. Pifl, and O. Hornykiewicz, "Globus pallidus dopamine and parkinson motor subtypes clinical and brain biochemical correlation," *Neurology*, vol. 70, no. 16 Part 2, pp. 1403–1410, 2008.

[7] F. Pedregosa-Izquierdo, "Feature extraction and supervised learning on fmri: from practice to theory," Ph.D. dissertation, Université Pierre et Marie Curie-Paris VI, 2015.

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

[9] B. Magnin, L. Mesrob, S. Kinkingnéhun, M. Pélégrini-Issac, O. Colliot, M. Sarazin, B. Dubois, S. Lehéricy, and H. Benali, "Support vector machine-based classification of alzheimer's disease from whole-brain anatomical mri," *Neuroradiology*, vol. 51, no. 2, pp. 73–83, 2009.

[10] F. Segovia, J. Górriz, J. Ramírez, I. Alvarez, J. Jiménez-Hoyuela, and S. Ortega, "Improved parkinsonism diagnosis using a partial least squares based approach," *Medical physics*, vol. 39, no. 7, pp. 4395–4403, 2012.

[11] I. Illán, J. Górriz, J. Ramírez, F. Segovia, J. Jiménez-Hoyuela, and S. Ortega Lozano, "Automatic assistance to parkinson's disease diagnosis in datscan spect imaging," *Medical physics*, vol. 39, no. 10, pp. 5971–5980, 2012.

[12] A. Winogrodzka, P. Bergmans, J. Booij, E. Van Royen, A. Janssen, and E. C. Wolters, "[123 i] fp-cit spect is a useful method to monitor the rate of dopaminergic degeneration in early-stage parkinson's disease," *Journal of neural transmission*, vol. 108, no. 8, pp. 1011–1019, 2001.

[13] A. V. García, J. C. Vaamonde, V. G. Poblete, S. M. Rodado, M. R. Cortés, S. S. Ruiz, R. A. Ibáñez, and A. C. Soriano, "Utility of dopamine transporter imaging (123-i ioflupane spect) in the assessment of movement disorders," *Revista espanola de medicina nuclear*, vol. 23, no. 4, pp. 245–252, 2004.

[14] G. Perlaki, S. Szekeres, G. Orsi, L. Papp, B. Suha, S. A. Nagy, T. Doczi, J. Janszky, K. Zambo, and N. Kovacs, "Validation of an automated morphological mri-based 123i-fp-cit spect evaluation method," *Parkinsonism & related disorders*, vol. 29, pp. 24–29, 2016.

[15] F. Segovia, I. Á. Illán, D. Salas-Gonzalez, F. J. Martínez-Murcia, C. Phillips, C. G. Puntonet, J. R. P. de Inestrosa, and J. M. G. Sáez, "Petra: Multivariate analyses for neuroimaging data." in *IWBBIO*, 2014, pp. 1302–1312.

[16] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*. Academic press, 2011.

[17] P. McCullagh, "Generalized linear models," *European Journal of Operational Research*, vol. 16, no. 3, pp. 285–292, 1984.

[18] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernandez-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: survey and experimental study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, 2016.

[19] P. A. Gutiérrez, P. Tiňo, and C. Hervás-Martínez, "Ordinal regression neural networks based on concentric hyperspheres," *Neural Networks*, vol. 59, pp. 51–60, 2014.

[20] J. D. Rennie and N. Srebro, "Loss functions for preference levels: Regression with discrete ordered labels," in *Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling*. Kluwer Norwell, MA, 2005, pp. 180–186.

[21] M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero, and P. A. Gutiérrez, "Metrics to guide a multi-objective evolutionary algorithm for ordinal classification," *Neurocomputing*, vol. 135, pp. 21–31, 2014.

[22] Waskom, Gramfort, Burns, Luessi, and Larson, "Pysurfer," 2017. [Online]. Available: http://pysurfer.github.io

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.

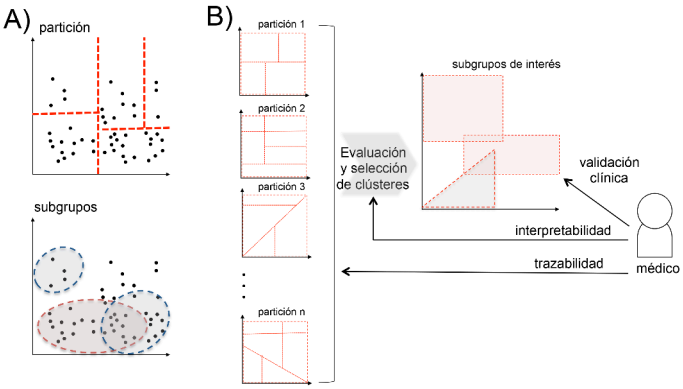


Figura 1. A) De la partición a los subgrupos; B) Trazabilidad, interpretabilidad y validación.

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

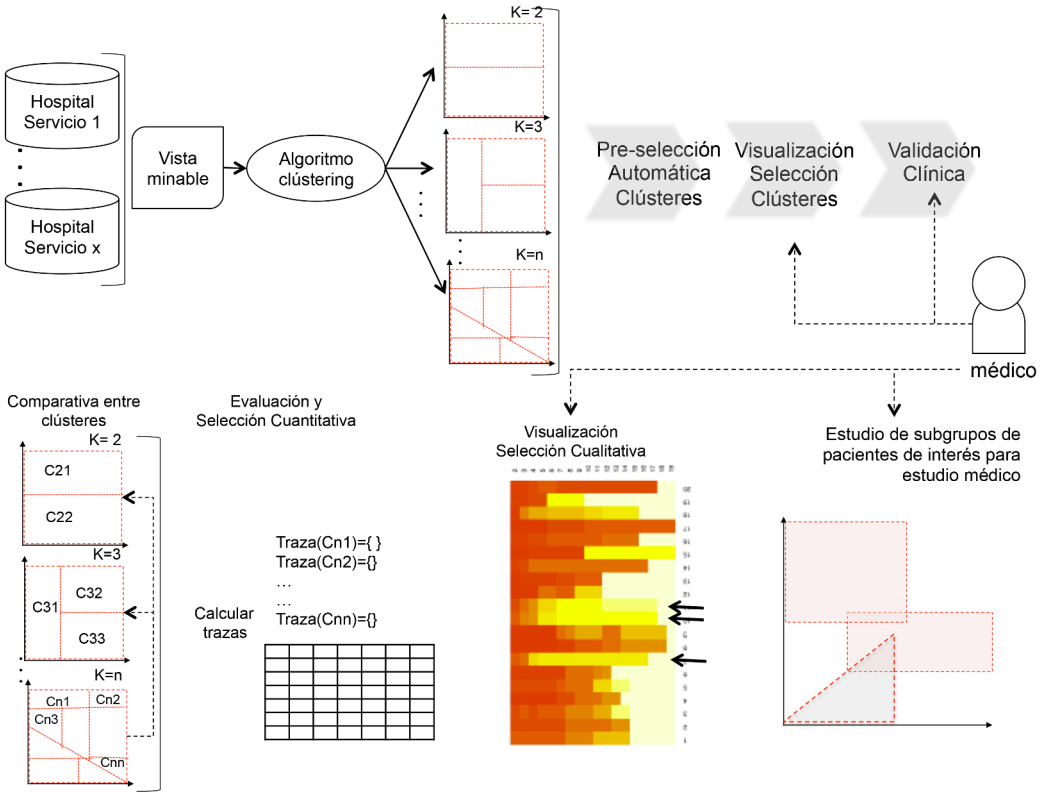


Figura 2. Detalles de la metodología.

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:

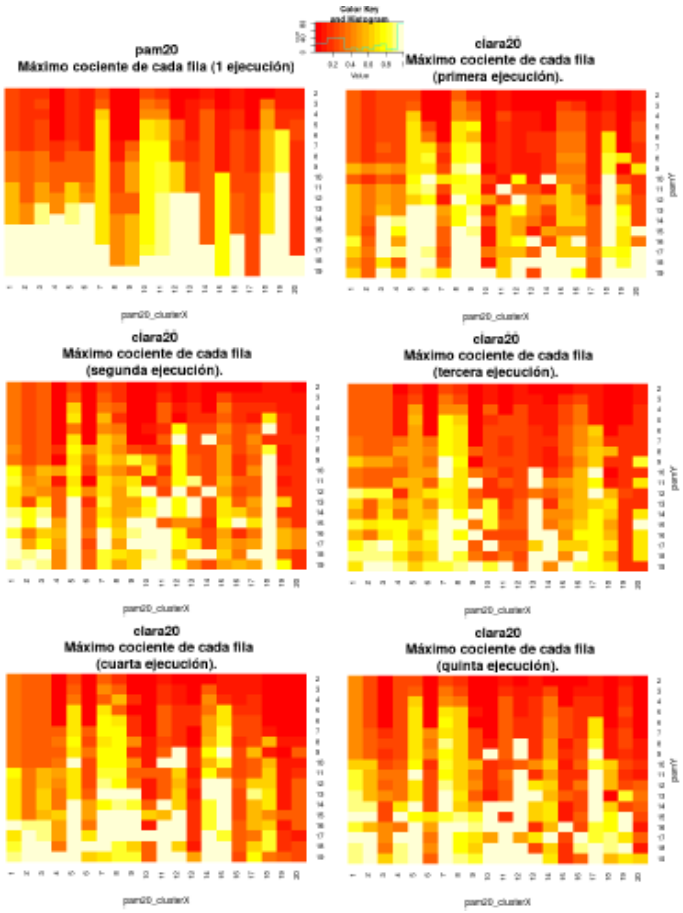


Figura 3. Heatmap de matrices $\mathcal{J}$.

- 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

A boundary-point approach applied to gene selection in gene expression data

Juan Ramos
IBSAL/BISITE Research Group
University of Salamanca
Salamanca, Spain
juanrg@usal.es

José A. Castellanos-Garzón
IBSAL/BISITE Research Group
University of Salamanca
Salamanca, Spain
jantonio@usal.es

Juan F. de Paz
IBSAL/BISITE Research Group
University of Salamanca
Salamanca, Spain
fcofds@usal.es

Juan M. Corchado
BISITE Research Group
University of Salamanca, Osaka Institute of Technology
Salamanca, Spain
corchado@usal.es

***Abstract*—In recent years there has been an increasing interest in using hybrid-technique sets to face the problem of meaningful gene selection; nevertheless, this issue remains a challenge. In work *A data mining framework based on boundary-points for gene selection from DNA-microarrays: Pancreatic Ductal Adenocarcinoma as a case study*, we propose a novel hybrid framework based on data mining techniques applied to the problem of meaningful gene selection and the search for new biomarkers. For this purpose, the framework deals with approaches such as statistical significance tests, cluster analysis, evolutionary computation, visual analytics and boundary points. The latter is the core technique of the proposal, which allows us to define two alternative methods of gene selection. Moreover, the framework has added a variable to the study (as the age), which is studied with respect to gene expression levels.**

***Index Terms*—Feature selection, Gene selection, Data mining, Cluster analysis, Genetic algorithm, Boundary point. DNA-microarray.**

I. Introduction

While significant efforts have been placed in the development of new methods and strategies to discover informative genes, the problem remains a challenge today since there is not a single technique able to solve all the underlying issues and adapt to different situations and problems.

We have used hybrid techniques to build a data mining framework for gene selection tasks, because they provide more robust and stable solutions than simple methods [1]–[3]. Generally, simple methods of gene selection assume that some criterion should be met in data, which does not have to be true for all data types. Hence, hybrid techniques fusion different simple methods to reach solutions holding more than one criterion, making solutions more stable with respect to variations in data. On the other hand, hybrid techniques are more flexible to changes in user needs and allow us to replace the methods taking place in the overall proposal without carrying out meaningful changes. Finally, we want to stress that this research has been published in [4].

The research of Juan Ramos González has been co-financed by the European Social Fund and Junta de Castilla y León (Operational Programme 2014-2020 for Castilla y León, BOCYL EDU/602/2016). This work has also been supported by project MOVIURBAN: Máquina social para la gestón sostenible de ciudades inteligentes: movilidad urbana, datos abiertos, sensores moviles. SA070U 16. Project cofinanced with Junta Castilla y León, Consejería de Educación and FEDER funds.

II. Hybrid framework

Since HybridFrame is based on data mining techniques, we have focused our efforts on the combination of areas such as evolutionary computation, visual analytics, and cluster analysis, among others to develop a methodology to follow in the domain of gene expression data, Figure 1.

Statistical filtering module (SFM) This module is responsible for a preliminary data processing and the first gene filtering processes based on statistical significance. Thus, the first process in this module consists of a data treatment by removing control probes, standardizing, and applying algorithms of missing data treatment if needed.

The first applied filter method is the Mann-Whitney test [5]. Then a second filter method is selected in relation to user goals. In this case, the module implemented five filter methods, although new methods can be added. In our case study we used Kruskal-Wallis, which can be used by when introducing a variable measurement external to the dataset to filter out genes related to the variable of interest.

Hierarchical clustering method module (HCMM) The dataset resulting from the module above is divided into subsets (clusters) in order to move the complex gene selection task from the whole current dataset to smaller gene subsets. The idea consists of applying data clustering methods to divide the complex task of gene selection from a big dataset into small subsets (divide and conquer strategy), identified by their gene similarity. Although this module does not really perform a gene filtering task, it partitions the data for the following stages.

Visual analytics module (VAM) This module selects the most suitable clustering from each input dendrogram. Internal

measures of cluster validity (such as homogeneity, separation and silhouette width) are applied to input dendrograms to estimate level ranges with high quality clusterings. [6], [7], which are applied to each level of a dendrogram to select the one with the best score. Then consists of choosing and visually validating a level from each level interval computed in the process above. For this propose, each dendrogram is explored from its level interval through a linked visualization set, supporting heatmaps, dendrograms, parallel coordinates, 3D-scatterplots and boundary gene visualizations.

Clustering boundary module (CBM) This module carries out a filtering process by extracting out the boundary genes for each cluster given from input clusterings to the module. The boundary point algorithm used for this purpose is focused on the ClusterBoundary algorithm given in [8].

The final stage of the framework consists in two alternative selection methods:

- **Clustering intersection method (CIM)**: the CIM method is based on the idea of boundary intersections coming from different clustering methods. We assume that boundary genes achieved from the intersection of different clustering boundaries coming from different methods, which develop different cluster strategies on data, are the main candidates to be informative genes.
- **Evolutionary hierarchical clustering method (ECM)**: The second method leading to discover informative genes in this framework is ECM, as shown in Fig 1. This method is based on the evolutionary model for clustering (EMHC) given in [9], [10]. We propose that, since dendrograms given as ECM solutions inherit, alter, recombine and even improve part of the genetic code (high quality clusters) of good solutions given by others methods, then it is expected that genes located on the boundary of such clusters are strong candidates to be informative genes.

III. Case study on pancreatic ductal adenocarcinoma

As a case study to apply and validate our proposal, we have focused our attention on the tissue sample study of pancreatic ductal adenocarcinoma (PDAC) through microarray technology, given that PDAC has been identified as one of the most aggressive types of existing cancer [11], [12]. Although every cancer has a strong relation to age due to several cellular processes, but for PDAC, this relation appears to be more remarkable than other cancers. In fact, 85% of pancreatic cancer cases involve patients older than 65-years old with a diagnosis mean age of 73-years old [13]. For that reason, this research introduces the age factor for further analysis of its influence in cancer patients.

After applying the methodology above, two sets of genes were obtained, one for each filter method, identifying informative genes of each method. Information about each concrete gene has previously been identified in other research and/or databases as a PDAC related gene. Information provided by both tables was consulted in PED

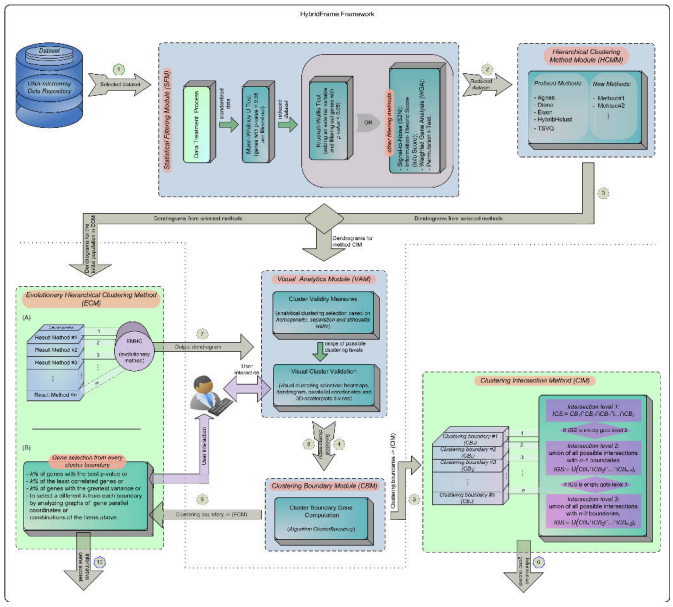


Fig. 1. Chart representing the data mining framework HybridFrame for gene selection.

(http://www.pancreasexpression.org/) and Pancreatic Cancer Database (http://pancreaticcancerdatabase.org/index.php). Selected genes have a larger relation to normal and tumor tissue samples of PDAC and are highly age-related. Moreover, 10 genes from these tables were identified by both methods (genes in the intersection are highlighted in both tables), meaning they could be even more meaningful for PDAC than the rest. In summary, according to the whole discovery process of informative genes given by Hybridframe, we assume that selected genes can be considered for further pharmaceutic research.

Acknowledgment

We would like to thank Dr. Liviu Badea from Bioinformatics research group, National Institute for Research in Informatics (Rumania) for provided additional information on the dataset used in this research.

References

[1] I. Guyon, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[2] J. Jager, R. Sengupta, and W. Ruzzo, "Improved gene selection for classification of microarrays," in *Pacific Symposium on Biocomputing (UW CSE Computational Biology Group), PMID: 12603017*, 2003.

[3] C. Lazar, J. Taminau, D. Meganck, S.and Steenhoff, A. Coletta, V. Molter, C.and deSchaetzen, H. Duque, R.and Bersini, and A. Nowé, "A survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM Transactions On Computational Biology And Bioinformatics*, vol. 9, no. 4, pp. 1106–1118, 2012.

[4] J. Ramos, J. A. Castellanos-Garzón, J. F. de Paz, and J. Corchado, "A data mining framework based on boundary-points for gene selectionfrom DNA-microarrays: Pancreatic Ductal Adenocarcinoma as a case study," *Engineering Applications of Artificial Intelligence, Elsevier*, vol. 70, pp. 92–108, 2018.

[5] P. Weiss, "Applications of generating functions in nonparametric tests," *The Mathematica Journal*, vol. 9, no. 4, pp. 803–823, 2005.

[6] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, 2004.

[7] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data. An Introduction to Clustering Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2005.

[8] J. Castellanos-Garzón, C. García, P. Novais, and F. Díaz, "A visual analytics framework for cluster analysis of DNA microarray data," *Expert Systems with Applications, Elsevier*, vol. 40, pp. 758–774, 2013.

[9] J. Castellanos-Garzón, "Evolutionary framework for DNA microarray cluster analysis," Ph.D. dissertation, Department of Computer Science, University School of Computer Science, University of Valladolid, 2012.

[10] J. A. Castellanos-Garzón and F. Díaz, "An evolutionary computational model applied to cluster analysis of DNA microarray data," *Expert Systems with Applications, Elsevier*, vol. 40, pp. 2575–2591, 2013.

[11] L. Badea, V. Herlea, S. Olimpia, T. Dumitrascu, and I. Popescu, *Combined Analysis of Whole-Tissue and Microdissected PDAC*, Bioinformatics group, National Institute for Research in Informatics, Bucharest 011455, Romania, 2008.

[12] ——, "Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia," *Hepato-Gastroenterology*, vol. 88, pp. 2015–2026, 2008.

[13] J. Koorstra, S. Hustinx, G. Offerhaus, and A. Maitra, "Pancreatic carcinogenesis," *Pancreatology*, vol. 8, no. 2, pp. 110–125, 2008.