Inv Ed Med. 2015;4(16):242–251

Investigación en Educación Médica

http://riem.facmed.unam.mx

METODOLOGÍA DE INVESTIGACIÓN EN EDUCACIÓN MÉDICA

Una breve introducción a los modelos de clases jerárquicas (HICLAS)

Iwin Leenen*

Instituto Nacional para la Evaluación de la Educación, México D.F., México

Recibido el 22 de septiembre de 2015; aceptado el 28 de septiembre de 2015
Disponible en Internet el 30 de octubre de 2015

PALABRAS CLAVE
Clases jerárquicas;
Clustering;
Datos binarios;
Modelos componenciales

Resumen En este artículo, se presenta una introducción no técnica a la familia de modelos de clases jerárquicas (HICLAS, abreviados por sus siglas en inglés). El modelo HICLAS original, propuesto por De Boeck y Rosenberg en 1988, permite analizar una matriz de datos binarios; es decir, analiza una tabla con valores de 0 o 1. Contrario a otros métodos de clasificación, HICLAS proporciona como resultado del análisis (a) *dos clasificaciones jerárquicas*, una de los «objetos» (las filas de la tabla) y otra de los «atributos» (las columnas), y (b) una *asociación* de dichas clasificaciones a partir de la cual se puede pronosticar (o reconstruir) el valor en cada celda de la tabla. Se ilustran los beneficios de los modelos HICLAS con ejemplos en el ámbito de la educación médica, explicando los diversos pasos en la aplicación de este método y con especial énfasis en la interpretación de los resultados.

KEYWORDS
Hierarchical classes;
Clustering;
Binary data;
Componential models

A brief introduction to hierarchical classes (HICLAS) models

Abstract This article presents a nontechnical introduction to the family of hierarchical classes (HICLAS) models. The original HICLAS model, proposed by De Boeck and Rosenberg in 1988, is suited for the analysis of a binary data matrix; that is, it analyses a two-way table with values equal to 0 or 1. Unlike other classification methods, the output of a HICLAS analysis includes (a) two hierarchical classifications, one for the ''objects'' (the table's rows) and another for the ''attributes'' (the columns), and (b) an association of these classifications from which it is possible to predict (or reconstruct) the value in each cell of the table. The advantages of HICLAS

* Autor para correspondencia: Instituto Nacional para la Evaluación de la Educación (INEE, 5.° piso); Av. Barranca del Muerto n.° 341, Colonia San José Insurgentes, Del. Benito Juárez; C.P. 03900 México, D.F.; Tel.: +5482 0900 Ext 42016.
Correo electrónico: iwin.leenen@gmail.com
La revisión por pares es responsabilidad de la Universidad Nacional Autónoma de México.

http://dx.doi.org/10.1016/j.riem.2015.09.003

models are illustrated with examples in the field of medical education, with an explanation of the different steps in the application of the model, as well as special emphasis on the interpretation of the results.

Clasificar es una de las tareas básicas de cualquier disciplina científica[1]; los ejemplos incluyen las clasificaciones en la biología que han llevado a las taxonomías para el reino animal y vegetal, el *clustering* de expresiones génicas en la medicina[2], y la agrupación de clientes en un contexto de mercadotecnia[3]. Las últimas décadas han presenciado un aumento exponencial en la popularidad de los métodos de clasificación[4], especialmente en las ciencias sociales, donde sirven como modelo estructural para una amplia gama de fenómenos psicológicos y educativos. Tanto el desarrollo de nuevos métodos, con modelos que permiten analizar los problemas de interés desde una perspectiva distinta, como la optimización de algoritmos y las computadoras cada vez más potentes, son factores que se relacionan con este incremento. Como nota al margen, es importante resaltar que, a pesar de los esfuerzos para poner orden en la jungla de los métodos de clasificación[4,5], la falta de un marco unificador sigue siendo un obstáculo significativo para los investigadores aplicados que buscan el método de clasificación idóneo para el análisis de sus datos; en particular, se encuentran con conceptos y modelos desarrollados en diferentes ámbitos sin contar con «puentes» que permiten entender las relaciones entre estos numerosos métodos y sus variantes. Para revisiones recientes del área de clustering y clasificación véase Hennig et al.[6] y Mirkin[7].

El objetivo de un *clustering* generalmente es derivar una clasificación de un conjunto de elementos según ciertos criterios de tal forma que cada uno de los grupos sea (relativamente) homogéneo. Aunque se pueden comparar los métodos de clasificación acorde con diferentes principios, el más común consiste en distinguir entre los siguientes tres tipos de *clustering*[5,8]:

(a) Los que implican una *partición* de los elementos, es decir, una clasificación en la cual cada elemento pertenece precisamente a un grupo; matemáticamente hablando, en una partición los grupos son exhaustivos y mutuamente excluyentes.
(b) Los que llevan a un *agrupamiento anidado*, lo cual quiere decir que se permite que los elementos pertenezcan a múltiples categorías; sin embargo, si la intersección entre dos categorías no es vacía, entonces una de ellas es un subconjunto de la otra. El *clustering jerárquico* se puede interpretar como un agrupamiento anidado, en el cual están presentes tanto el conjunto completo de todos los elementos como todos los conjuntos elementales (los que contienen solo un elemento).
(c) Los que resultan en *categorías superpuestas, no anidadas* (en inglés conocido como *overlapping clustering*); permiten que los grupos tengan elementos en común sin que estén anidados.

La figura 1 ilustra los tres tipos de clustering utilizando figuras geométricas como elementos. Más adelante se presentarán ejemplos propios del área de la educación médica.

De una forma u otra, los métodos de clasificación utilizan (una cuantificación de) la (di)similitud entre pares de los objetos que se desean clasificar. Varios métodos (como ADCLUS[9] y el *clustering* jerárquico de Johnson)[10] analizan directamente las similitudes entre los objetos, es decir, los datos para estos métodos se organizan en una tabla en la cual las filas y las columnas refieren a los mismos objetos y las celdas contienen, por ejemplo, las coocurrencias o confusiones entre pares de objetos[11] o juicios directos de su similitud[12]. Sin embargo, raras veces los datos primarios de un estudio toman la forma que se acaba de describir; es más común encontrarse con una tabla de datos donde las filas refieren a «observaciones» y las columnas a «variables». En el contexto de modelos de clasificación, se suelen utilizar, en vez de observaciones y variables, los nombres genéricos de *objetos* y *atributos*, respectivamente. Si se desea aplicar los métodos de clasificación anteriormente mencionados a datos de este tipo, se requiere un preprocesamiento de la tabla original para derivar la tabla de similitudes (o, como también se dice, las proximidades), por ejemplo, un análisis de correlación entre los objetos. De forma más general, se calcula para cada par de objetos la proximidad a través de un procedimiento que compara los valores de los dos objetos en los atributos[8].

Debe ser claro que los métodos que operan de esta forma solo llevan a una clasificación de los objetos. Si bien es cierto que se puede aplicar la misma estrategia para calcular y analizar las proximidades entre los atributos (por ejemplo, calculando la correlación entre cada par de atributos) y, de esta forma, se obtiene una clasificación de los atributos, se trata de dos análisis separados, sin que los resultados obtenidos tengan una relación clara. En este artículo, presento una breve introducción a una familia de modelos, denominados *modelos de clases jerárquicas* (HICLAS, por sus siglas en inglés), que permiten analizar *directamente* la tabla de objetos por atributos (es decir, sin derivar medidas de la proximidad en un paso previo), y que proporcionan una clasificación *simultánea* de los objetos y atributos. En este sentido, el modelo es similar al método descrito por Hartigan[13,14]; sin embargo, como propiedad única, los modelos HICLAS, además de clasificar los objetos y atributos, también incluyen una representación de las relaciones jerárquicas entre las clases.

El modelo HICLAS original fue propuesto por Paul De Boeck y Seymour Rosenberg en un artículo en *Psychometrika* en 1988[15]. A partir de este modelo seminal, se han desarrollado variantes y extensiones de tal forma que hoy se habla de una *familia* de modelos HICLAS. En este artículo, se presenta primero el modelo original a través de un ejemplo

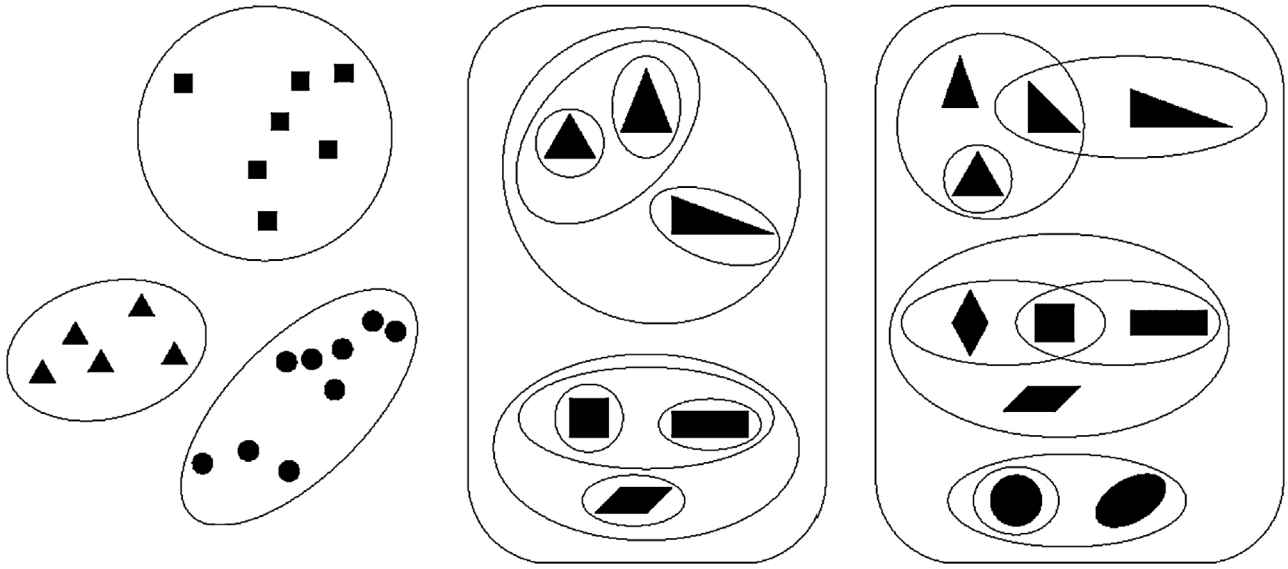

Figura 1 Ilustración de diferentes tipos de *clustering* de figuras geométricas. Izquierdo: partición; medio: agrupamiento anidado (*clustering* jerárquico); derecho: agrupamiento no anidado (*overlapping clustering*).

en el contexto de educación; en particular, se introducen las propiedades más importantes del modelo, su representación gráfica y el proceso de aplicar el modelo a datos empíricos. Como variante del modelo original, el cual, por razones que se explican más abajo, se conoce como disyuntivo, se propuso en 1995 el modelo conjuntivo[16]. En la segunda sección, se ilustran las propiedades de esta variante y se explica en qué difiere del modelo original. En la tercera sección, se describen brevemente algunos otros miembros de la familia HICLAS, desarrollados más recientemente. Se concluye este artículo con unos comentarios finales.

El modelo HICLAS disyuntivo de De Boeck y Rosenberg

Tres tipos de relaciones definidos sobre la tabla de datos

El modelo disyuntivo original analiza una tabla de objetos (filas) y atributos (columnas) cuyas celdas tienen valores iguales a 0 o 1. La tabla 1 muestra unos datos hipotéticos que servirán para ilustrar los conceptos en esta sección; representan las técnicas didácticas que ocho profesores del posgrado consideran adecuadas para utilizar en clase. En este ejercicio, se consideran las siguientes siete técnicas: lección magistral (LM), lectura y discusión grupal de artículos científicos (LD), enseñanza asistida por la computadora (EC), debates entre grupos de estudiantes (DG), ejercicios y problemas prácticos (EP), análisis de casos (AC), y trabajos en grupos pequeños (TG). Los profesores son los «objetos», las técnicas didácticas son los «atributos». Si un profesor considera adecuada una técnica, entonces la celda correspondiente muestra el valor de 1; en caso contrario, el valor es 0.

En una tabla de datos de este tipo se pueden definir tres tipos de relaciones: equivalencia, jerarquía y asociación. Estas relaciones se definen independientemente del modelo HICLAS (o cualquier otro modelo). Se trata de cualidades de los datos. A continuación, se explican estas relaciones.

Para los datos de la tabla 1, se define la *relación de equivalencia* entre dos (o más) profesores y, de forma similar, entre dos (o más) técnicas didácticas (en general, se definen relaciones de equivalencia entre los objetos y entre los atributos). Considerando dos profesores, existe una relación de equivalencia entre ellos si sus opiniones sobre cada una de las siete técnicas coinciden. En otras palabras, dos profesores son equivalentes si tienen los mismos (valores en los) atributos; es cuando sus filas en la tabla son idénticas. En nuestro ejemplo, los profesores Carlos y Hugo son equivalentes ya que ambos consideran adecuadas todas las técnicas excepto LM. De forma análoga, se define la relación de equivalencia entre las técnicas didácticas: dos o más técnicas son equivalentes si los mismos profesores las consideran adecuadas o, refiriéndonos a la tabla, si sus columnas son idénticas. Por ejemplo, las técnicas LD y DG en la tabla 1 son equivalentes.

Tabla 1 Datos hipotéticos de ocho profesores sobre si aplicarían (1) o no (0) siete técnicas didácticas

	LM	LD	EC	DG	EP	AC	TG
Prof. Abraham	1	0	0	0	1	0	0
Prof. Bruno	1	0	1	0	1	1	1
Prof. Carlos	0	1	1	1	1	1	1
Prof. Daniel	0	1	0	1	1	1	1
Prof. Elio	1	1	1	1	1	1	1
Prof. Flavio	1	0	0	0	1	0	0
Prof. Giovanni	1	0	0	0	1	0	0
Prof. Hugo	0	1	1	1	1	1	1

AC : análisis de casos; DG : debates entre estudiantes; EC : enseñanza asistida por la computadora; EP : resolución de ejercicios y problemas prácticos; LD : lectura y discusión grupal de artículos científicos; LM : lección magistral; TG : trabajos en grupos pequeños.

También las *relaciones de jerarquía* se definen entre los profesores (los objetos) y entre las técnicas didácticas (los atributos). En particular, se dice que el profesor A es jerárquicamente superior al profesor B si se cumplen las siguiente dos condiciones: (a) en el caso de que el profesor B considere adecuada una técnica didáctica, también el profesor A la considera adecuada y (b) el profesor A considera más técnicas como adecuadas que el profesor B. Por ejemplo, en la tabla 1, el Prof. Bruno es jerárquicamente superior al Prof. Abraham ya que (a) Bruno considera adecuadas todas las técnicas didácticas que Abraham considera adecuadas (LM y EP), y (b) hay técnicas (EC, AC, y TG) que Bruno considera adecuadas y Abraham no. Las relacionas jerárquicas entre las técnicas didácticas se definen de forma similar. Por ejemplo, en la tabla 1, la técnica AC es jerárquicamente superior a la técnica LD debido a que todos los profesores que encuentran adecuada LD también encuentran adecuada AC, y además, hay un profesor (Bruno) que considera adecuada la técnica AC, pero no la LD.

Es conveniente aclarar en este punto de la exposición tres elementos. Primero, la relación de equivalencia es simétrica, mientras que la relación de jerarquía es asimétrica. Por ejemplo, el Prof. Carlos es equivalente al Prof. Hugo, y, por lo tanto, se sabe que Hugo también es equivalente a Carlos. Por otro lado, el Prof. Bruno es jerárquicamente superior al Prof. Abraham; sin embargo, el Prof. Abraham es jerárquicamente *inferior* a el Prof. Bruno. Segundo, a partir de las relaciones de equivalencia se definen *clases* de profesores y clases de técnicas didácticas. Una clase es un conjunto de profesores equivalentes o técnicas equivalentes. Por ejemplo, los profesores Abraham, Flavio y Giovanni forman una clase; asimismo, las técnicas didácticas AC y TG forman una clase. Tercero, en el párrafo anterior se definió la relación de jerarquía entre profesores o técnicas didácticas individuales. Sin embargo, también es común hablar de relaciones jerárquicas entre las clases. Se dice, por ejemplo, que la clase de profesores {Bruno} (que consiste de un solo elemento) es jerárquicamente superior a la clase de profesores {Abraham, Flavio, Giovanni}.

La tercera relación que se define a partir de la tabla de datos es la *relación de asociación*. Contrario a las dos relaciones previas, que se definen entre profesores o bien entre técnicas didácticas, la relación de asociación conecta los profesores (objetos) y técnicas didácticas (atributos). La relación de asociación es la que naturalmente se lee en los datos: un profesor está asociado con una técnica didáctica, si la considera adecuada; es decir, si en la tabla la celda correspondiente al profesor y la técnica contiene un valor de 1. Aunque la relación de asociación se define para objetos y atributos individuales, se puede considerar también la asociación entre clases de objetos y atributos. Por ejemplo, en la tabla 1, la clase de profesores {Carlos, Hugo} está asociada con la clase de técnicas didácticas {AC, TG}; y la clase de profesores {Abraham, Flavio, Giovanni} no está asociada con la clase {LD, DG}.

El modelo HICLAS y la representación de los tres tipos de relaciones

Hasta el momento, todavía no hemos introducido el modelo HICLAS; únicamente se definieron relaciones estructurales en la tabla de datos. Sin embargo, lo típico del HICLAS es que

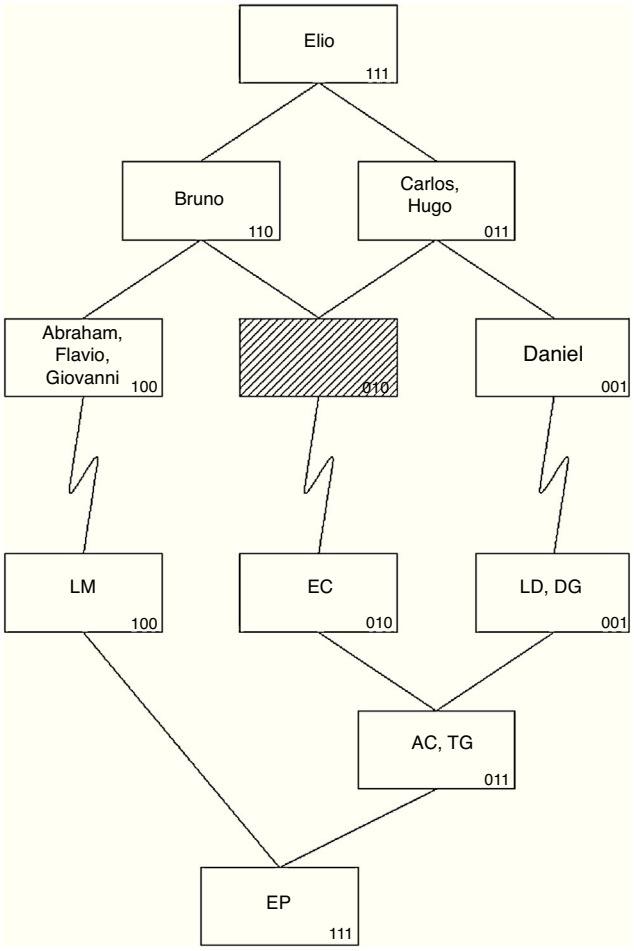

Figura 2 Representación gráfica del modelo HICLAS disyuntivo para los datos de la tabla 1. Para las abreviaturas utilizadas de las técnicas didácticas, véase la tabla.

representa claramente dichas relaciones de equivalencia, jerarquía, y asociación. Existen dos formas para representar un modelo HICLAS: la gráfica y la algebraica. Primero, nos enfocamos en la representación gráfica; se explicaría cómo los tres tipos de relaciones se representan en la gráfica asociada con el modelo HICLAS disyuntivo.

La figura 2 muestra el modelo HICLAS para los datos de la tabla 1. En la parte superior de la gráfica se encuentran los profesores (y, como se explicaría en un momento, las relaciones de equivalencia y jerarquía entre ellos); en la parte inferior se observan las técnicas didácticas (y la representación de las relaciones de equivalencia y jerarquía entre estas). Miremos primero la parte superior: las cajas (rectángulos) representan las clases de profesores. Recuérdese que los profesores Abraham, Flavio y Giovanni formaban una clase; en la gráfica se observa que se encuentran en la misma caja.

De esta forma, el modelo HICLAS permite de manera fácil derivar cuáles son los profesores con las mismas opiniones sobre las técnicas didácticas. Nótese que la figura también incluye una clase vacía (el rectángulo sombreado); al final de esta sección, se explicará su función. Además, si dos clases están conectadas por una línea (siempre descendente y posiblemente pasando por la caja de otra clase), existe

Tabla 2 Representación algebraica (matrices características) del modelo HICLAS disyuntivo para los datos de la tabla 1

	Matriz característica de los profesores				Matriz característica de las técnicas didácticas		
	I	II	III		I	II	III
Prof. Abraham	1	0	0	LM	1	0	0
Prof. Bruno	1	1	0	LD	0	0	1
Prof. Carlos	0	1	1	EC	0	1	0
Prof. Daniel	0	0	1	DG	0	0	1
Prof. Elio	1	1	1	EP	1	1	1
Prof. Flavio	1	0	0	AC	0	1	1
Prof. Giovanni	1	0	0	TG	0	1	1
Prof. Hugo	0	1	1				

AC : análisis de casos; DG : debates entre estudiantes; EC : enseñanza asistida por la computadora; EP : resolución de ejercicios y problemas prácticos; LD : lectura y discusión grupal de artículos científicos; LM : lección magistral; TG : trabajos en grupos pequeños.

una relación jerárquica entre ellas. Por ejemplo, la caja con los profesores Carlos y Hugo está conectada con la caja de Daniel, que se encuentra más abajo en la gráfica, por lo cual derivamos que Carlos y Hugo son jerárquicamente superiores a Daniel. Recuérdese que esto significa que Carlos y Hugo consideran adecuadas todas las técnicas que Daniel considera adecuadas, y alguna(s) más.

Las equivalencias y relaciones jerárquicas entre las técnicas didácticas se leen de forma análoga en la parte inferior de la figura, con la importante diferencia de que la jerarquía se ha invertido. Es decir, las cajas representan las clases de técnicas didácticas y dos cajas conectadas por una línea indican que la clase que se encuentra visualmente arriba es jerárquicamente *inferior* a la que se encuentra abajo. Por ejemplo, la clase de las técnicas AC y TG, que en la figura se encuentra debajo de la clase de LD y DG es jerárquicamente *superior* a esta.

También la relación de asociación se lee fácilmente de la representación gráfica del modelo HICLAS: si existe un camino (siempre descendente, pasando por las líneas y las cajas) que permite llegar de (la clase de) un profesor a (la clase de) una técnica didáctica, entonces, están asociadas (las clases de) el profesor y la técnica. Por ejemplo, es posible llegar desde la caja del Prof. Bruno a la caja de la técnica LM (bajando por el lado de la caja de [Abraham, Flavio, Giovanni]), y entonces este profesor y técnica están asociados; asimismo, se lee que Bruno considera adecuadas las técnicas AC y TG, ya que existe un camino que conecta las respectivas cajas (bajando por la caja vacía y la de EC). Aplicando la misma estrategia, se deriva que el Prof. Bruno está asociado con las técnicas LM, EC, EP, AC y TG, un resultado que se puede verificar en la tabla de datos (tabla 1).

Cabe señalar que subyacen ciertas ecuaciones matemáticas a la figura 2. En particular, la representación algebraica implica una descomposición de la tabla de datos en dos *matrices características*, una para los profesores (los objetos) y otra para las técnicas didácticas (los atributos).

La tabla 2 muestra las dos matrices características del modelo para los datos de la tabla 1. Las columnas de las matrices se llaman *características*[*]. Aún sin entrar en los detalles matemáticos, es fácil ver que los tres tipos de relaciones introducidos anteriormente también están representados en las matrices características. Se observa que los profesores (o las técnicas didácticas) equivalentes tienen el mismo patrón en las características. Por ejemplo, los tres profesores equivalentes Abraham, Flavio, y Giovanni «cargan» únicamente en la primera característica. (Se dice que un objeto o un atributo carga en una característica, si la celda correspondiente de la matriz característica tiene el valor de 1).

Si dos profesores (o dos técnicas didácticas) están jerárquicamente relacionados, entonces, el que es jerárquicamente inferior carga en un subconjunto de las características en las que carga el jerárquicamente superior. Por ejemplo, Bruno, quien está debajo de Elio, carga solo en las características I y II, mientras que Elio carga en las tres. Similarmente, observamos que las técnicas didácticas LD y DG cargan en la característica III y que las técnicas AC y TG, que son jerárquicamente superiores, incluyen además de la característica III, también la II. Además, la relación de asociación se deriva a través de la siguiente regla: si un profesor y una técnica didáctica tienen una (o más) características en común, están asociados. Por ejemplo, el profesor Abraham comparte la característica I con las técnicas LM y EP y, por consiguiente, considera estas técnicas como adecuadas; Abraham no está asociado con las otras técnicas, debido a que estas no cargan en la característica I (la única con la que Abraham está asociado). Se puede verificar que estas y otras derivaciones coinciden perfectamente con los datos de la tabla 1.

Las clases que cargan en una sola característica se llaman *clases básicas*; en la representación gráfica, se encuentran en el nivel más bajo de la jerarquía. En el ejemplo de la figura 2, las clases básicas son {LM}, {EC} y {LD, DG} (técnicas didácticas) y {Daniel}, Ø y {Abraham, Flavio, Giovanni} (profesores). Debido a que ningún profesor está asociado de forma única con la característica II, la clase básica correspondiente está vacía (Ø, representado por el rectángulo sombreado en la figura). Cabe mencionar que la gráfica de un modelo HICLAS disyuntivo por lo general requiere que se representen todas las clases básicas, aunque estén vacías, ya que las dos jerarquías se conectan a través de las clases básicas (por otro lado, clases vacías en los niveles superiores de la

[*] Estas matrices se conocen como «*bundle matrices*» en la literatura anglosajona original, y las columnas refieren a «*bundles*».

jerarquía generalmente no se representan en la gráfica del modelo).

Interpretación

La clasificación proporcionada por el modelo HICLAS se puede valorar desde dos perspectivas distintas. Por un lado, implica una partición del conjunto original tanto de los objetos como de los atributos. Por ejemplo, los ocho profesores se particionan en cinco clases, mutuamente excluyentes y conjuntamente exhaustivas. Por otro lado, se puede considerar la clasificación como un agrupamiento no anidado con clases superpuestas. Desde esta perspectiva, la clase que contiene el profesor Elio está anidada dentro de las otras clases, aunque por ejemplo la clase del profesor Bruno está superpuesta con la de los profesores Carlos y Hugo. Nótese la similitud con las gráficas en la figura 1.

Además del aspecto de clasificación (el cual el HICLAS comparte con otros métodos de *clustering)*, las relaciones de jerarquía pueden ser de mucho interés en la práctica. Suelen dar lugar a interpretaciones del tipo «si . . ., entonces . . .». En nuestro ejemplo, se puede derivar que: si algún profesor considera adecuada la técnica didáctica EC, entonces también las técnicas AC, TG, y EP (las cuales se encuentran jerárquicamente superiores). Ejemplos notables en la literatura donde estas interpretaciones son trascendentales incluyen la teoría del espacio de conocimiento[17], donde las relaciones jerárquicas se interpretan como caminos que un estudiante puede seguir durante un proceso de aprendizaje, y ciertas teorías en la psicología de la personalidad[18], donde se investigan diferencias entre personas respecto de su perfil situación-conducta (es decir, qué situaciones desencadenan qué conductas en diferentes tipos de personas).

A pesar de que las características (es decir, las columnas de las matrices características, véase la tabla 2) son variables abstractas que no necesariamente corresponden con ciertos aspectos de la realidad, en varias aplicaciones ha resultado interesante una interpretación de las mismas. En el ejemplo hipotético de los profesores y técnicas didácticas, la siguiente interpretación aplica a las tres características: la característica I corresponde con un estilo didáctico más tradicional (nótese que la técnica LM carga únicamente en esta característica); la característica II se refiere más a un estilo didáctico que incorpora el uso de herramientas de informática (la técnica EC carga exclusivamente en esta), y (c) la característica III corresponde más con un estilo de enseñanza moderno (con cargas de LD y DG en esta característica). Algunas técnicas didácticas (las que se encuentran en las clases superiores de la jerarquía) combinan varias características; por ejemplo, AC y TG son características tanto de un estilo didáctico moderno como basado en herramientas informáticas, y la técnica de EP se utiliza en cualquiera de los tres estilos didácticos. También los profesores tienen una carga en las mismas características, lo cual puede informar sobre los estilos de enseñanza con los que se sienten cómodos o que consideran adecuados para la enseñanza de su materia. Al interpretar la asociación entre objetos y atributos en HICLAS en términos de las características, el modelo indicará que un objeto y un atributo están relacionados solo si existe *por lo menos una* característica en la cual tanto el objeto como el atributo cargan. Según la lógica matemática, esta regla implica una disyunción y es la que hace que el modelo original se caracteriza como disyuntivo. En nuestro ejemplo, esta regla quiere decir que un profesor considera una técnica didáctica adecuada si existe (por lo menos) un estilo de enseñanza (a) con el cual el profesor se sienta cómodo y (b) que sea característica de la técnica.

Aplicación del modelo a datos empíricos

Para cualquier tabla con valores 0/1 existe un modelo HICLAS que perfectamente representa las relaciones de equivalencia, jerarquía y asociación[19]. Sin embargo, en general, con datos reales se requiere un modelo muy complejo para alcanzar la representación perfecta. La complejidad se suele expresar en términos del *rango* del modelo, el cual se define como el número de características (es decir, el número de columnas en las matrices características, o bien –lo cual resulta equivalente– el número de clases básicas en cada jerarquía). Para nuestro ejemplo hipotético, el rango es 3. Para representar datos reales, no obstante, el rango generalmente es muy alto.

Por lo tanto, es común buscar un modelo HICLAS de un rango bajo que provea una buena aproximación a los datos. Es decir, se permite que la tabla reconstruida con base en la regla de asociación (como se explicó al final de la sección *El modelo HICLAS y la representación de los tres tipos de relaciones*) muestre en algunas celdas un valor diferente al valor observado en la tabla de datos (un 0 en vez de un 1, o viceversa). En el lenguaje técnico de HICLAS, se llama a una celda de este tipo una *discrepancia*; conforme el modelo da lugar a menos discrepancias, se dice que tiene una mejor *bondad de ajuste* a los datos.

La bondad de ajuste de un modelo HICLAS se suele resumir a través de dos indicadores: el porcentaje de discrepancias y el índice de Jaccard. El porcentaje de discrepancias es el número de discrepancias relativo al número total de celdas en la tabla; se consideran valores aceptables los menores del 15%. El índice de Jaccard, cuya fórmula es algo más compleja, siempre se encuentra entre 0 y 1 e indica a mayor valor un mejor ajuste; generalmente se considera un índice de Jaccard superior a 0.70 como indicador de un ajuste aceptable.

Para encontrar un modelo HICLAS con un buen ajuste a los datos, se puede aplicar el algoritmo HICLAS original[15] o una variante mejorada[20]; ambos tienen como objetivo encontrar, para un rango determinado, el modelo con mejor ajuste en el sentido que tenga un porcentaje mínimo de discrepancias. La salida de estos programas incluye las dos matrices características (a partir de las cuales se puede obtener la representación gráfica), junto con información de la bondad de ajuste.

Pocas veces se fija el rango del modelo de antemano (excepto al utilizar estrategias confirmatorias)[21]; en la práctica, se ajustan modelos de diferentes rangos (por ejemplo, de 1 a 8) y se estudia la bondad de ajuste en función del rango del modelo, con el fin de encontrar un modelo de bajo rango y bondad de ajuste satisfactoria. En particular, se elige el modelo de rango más bajo cuyo ajuste no sea significativamente peor en comparación con modelos de rango más alto.

Tabla 3 Datos hipotéticos de diez estudiantes y su aprobación de las ocho materias del primer año de la carrera de Médico Cirujano

	Ana	Cel	BQ	Emb	SP1	SM	IB1	IBC
Andrea	0	1	0	1	0	0	0	0
Bianca	1	1	1	1	1	1	0	0
Carolina	1	0	0	0	0	0	0	0
Diana	0	1	0	1	1	1	1	0
Emma	0	1	0	1	1	1	1	0
Fernanda	0	0	0	0	1	1	0	0
Gabriela	1	1	1	1	1	1	1	1
Hilda	0	0	0	0	0	0	0	0
Isabel	0	0	0	0	1	1	0	0
Julia	0	1	0	1	1	1	1	0

Ana : Anatomía; BQ : Bioquímica; Cel : Biología Celular; Emb : Embriología; IB1 : Informática Biomédica I; IBC : Integración Básico-Clínica; SM : Salud Mental; SP1 : Salud Pública I.

Para los lectores familiarizados con el análisis de factorial exploratorio (AFE), es interesante destacar el paralelismo entre la forma de escoger el rango de un modelo HICLAS y la manera en que se suele escoger el número de factores en AFE. Similar al caso de HICLAS se sabe que, con un número elevado de factores, un análisis factorial puede explicar hasta el 100% de la varianza en los datos. Sin embargo, el investigador comúnmente elige una solución factorial más simple, con pocos factores que logran explicar la mayor parte de la varianza en los datos. Similar al porcentaje de discrepancias en HICLAS, el porcentaje de varianza explicada en AFE sirve como indicador de la bondad de ajuste. Para ambos métodos, habitualmente se estudia cómo la bondad de ajuste incrementa en función de la complejidad del modelo, por ejemplo, a través de una gráfica de sedimentación[22]. Cabe mencionar que en el caso de HICLAS no existe una indeterminación rotacional como en el caso del AFE (aunque, bajo ciertas condiciones, el modelo HICLAS también puede sufrir indeterminaciones)[23].

El modelo HICLAS conjuntivo de Van Mechelen, De Boeck y Rosenberg

La variante conjuntiva de HICLAS[16] se aplica al mismo tipo de datos que el modelo disyuntivo, es decir, una tabla de objetos por atributos con valores 0 y 1 en las celdas. Además, la variante conjuntiva representa las mismas relaciones de equivalencia, jerarquía y asociación en los datos. La diferencia entre ambos modelos se sitúa en la forma en que se representan dichas relaciones.

Para guiar el desarrollo en esta sección, se utilizarán los datos en la tabla 3, que indica para cada una de diez estudiantes del primer año de la carrera de Médico Cirujano de la UNAM, cuáles de las ocho asignaturas aprobó (digamos, a través de los exámenes parciales). Si la estudiante aprobó la materia, entonces el valor en la celda correspondiente es igual a 1; en caso contrario, se da el valor de 0.

Debido a que la variante conjuntiva representa las relaciones estructurales de una manera distinta al modelo disyuntivo, la forma de construir y leer la representación

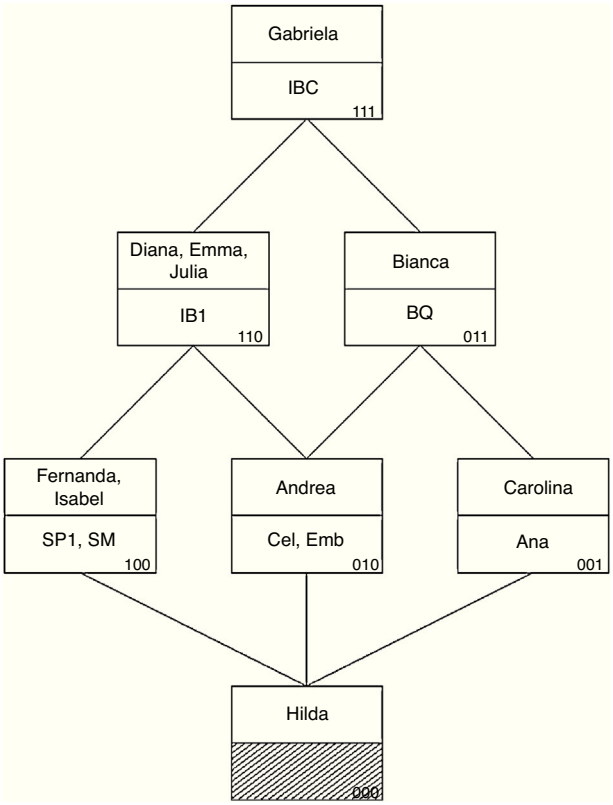

Figura 3 Representación gráfica del modelo HICLAS conjuntivo para los datos de la tabla 3. Para las abreviaturas utilizadas de las materias, véase la tabla.

gráfica difiere en algunos aspectos. En la figura 3, que muestra la gráfica del modelo HICLAS conjuntivo para los datos de la tabla 3, se observan pares de cajas, donde la caja superior de cada par representa una clase de los objetos (estudiantes) y la caja inferior una clase de los atributos (materias). Se deriva de esta figura que Diana, Emma y Julia forman una clase de estudiantes equivalentes (es decir, aprobaron las mismas materias, lo cual se puede verificar en la tabla 3). Asimismo, las materias equivalentes Biología Celular y Embriología se encuentran dentro de la misma caja en la figura 3. Respecto de las relaciones jerárquicas, se observa que la clase de estudiantes {Diana, Emma, Julia} se encuentra jerárquicamente superior a las estudiantes {Fernanda, Isabel}, lo cual significa que las primeras aprobaron todas las materias que aprobaron Fernanda e Isabel (SP1 y SM) y además aprobaron otras (Cel, Emb y IB1). Las relaciones jerárquicas entre los atributos se leen de forma similar, tomando en cuenta que las (clases de) materias que se encuentran arriba en la representación gráfica, son jerárquicamente *inferiores*. Por ejemplo, IBC, que se representa en la parte superior de la gráfica, es jerárquicamente inferior a todas las otras materias (es la materia *menos* aprobada de todas). Finalmente, la relación de asociación entre estudiantes y materias se lee fácilmente como sigue: la estudiante está asociada con todas las materias que se encuentran en posiciones inferiores a ella en la gráfica (ya sea dentro de la misma caja o en una caja inferior a la cual se puede bajar a través de las líneas). Por ejemplo, se lee que Diana, Emma

y Julia aprueban las materias de Informática Biomédica I, Salud Pública I, Salud Mental, Biología Celular y Embriología.

La tabla 4 muestra las matrices características correspondientes al modelo HICLAS en la figura 3. Se observa que se trata de un modelo de rango 3, ya que dichas matrices tienen tres columnas. En cuanto a la representación de las relaciones estructurales, se destaca lo siguiente: similar al caso del modelo disyuntivo, los objetos y atributos equivalentes tienen el mismo patrón en las tres características, por ejemplo, las tres estudiantes equivalentes Diana, Emma y Julia cargan en las características I y II. También la relación jerárquica para los objetos (estudiantes) se lee de la misma forma que en el modelo disyuntivo: si la estudiante A es jerárquicamente superior a la estudiante B, entonces las características en las que B carga forman un subconjunto de las características de A. Por ejemplo, Fernanda e Isabel cargan solo en la característica I (un subconjunto de las características de Diana, Emma y Julia). Sin embargo, la representación de la relación jerárquica para los atributos en un modelo conjuntivo es distinta en comparación con el modelo disyuntivo: en el contexto de nuestro ejemplo, si la materia A es jerárquicamente *superior* a la materia B, entonces las características en las que carga A son un *subconjunto* de las características de B. Por ejemplo, la materia de Anatomía, la cual es jerárquicamente superior a Bioquímica (tabla 3), carga solo en la característica I, mientras que Bioquímica carga tanto en la característica I como en la II.

Esta propiedad de la variante conjuntiva de representar inversamente la jerarquía entre atributos es muy vinculada a la forma en que se deriva la relación de asociación con base en las matrices características y a la interpretación de las características. A partir de la información en la tabla 4 se puede derivar la asociación entre estudiantes y materias como sigue: una estudiante aprueba una materia solo si la estudiante carga *en todas las características* en las que carga la materia. Por ejemplo, Gabriela y Bianca son las únicas estudiantes que aprueban Bioquímica ya que son las únicas que cargan en las características II y III, las cuales son las «requeridas» por esta materia. Efectivamente, las características en un modelo conjuntivo a menudo refieren a ciertos requisitos de los atributos que los objetos deben satisfacer. En nuestro ejemplo hipotético, las tres características se pueden interpretar como ciertas habilidades requeridas por las materias: la característica I refiere a la capacidad de ver relaciones entre diferentes partes del área de estudio; la II a ciertos conocimientos físico-matemáticos y la III a una habilidad de poder memorizar grandes cantidades de información. Retomando el ejemplo anterior, podemos concluir ahora que Gabriela y Bianca aprueban Bioquímica porque (a) dominan los aspectos físico-matemáticos y además (b) son capaces de memorizar grandes cantidades de información como se requiere para esta materia. Nótese que el requerimiento de satisfacer *todas* las características puestas por los atributos explica por qué el modelo se llama conjuntivo.

Otros miembros de la familia HICLAS

Después de la publicación de los modelos HICLAS disyuntivos y conjuntivos, se propusieron varias generalizaciones y variantes de los mismos. En esta sección, se describen a vuelo de pájaro algunos de estos nuevos miembros de la familia HICLAS.

Modelos para datos que se organizan en una estructura tridimensional

Supongamos que a los profesores del primer ejemplo se hubiese preguntado su opinión sobre las técnicas didácticas en diferentes momentos de su carrera. En este caso, se dispondría de múltiples tablas de objetos (profesores) por atributos (técnicas didácticas). Los datos que se obtienen de esta forma se podrían organizar en una tabla tridimensional; se conocen en la literatura como datos de tres vías[24]. Los modelos de INDCLAS y Tucker3-HICLAS[25–27] son una extensión del modelo HICLAS original para este tipo de datos: después de una generalización de la definición de las relaciones de equivalencia, jerarquía y asociación, dichos modelos representan las relaciones estructurales entre los elementos de las tres vías a través de tres matrices características. Los autores propusieron también una adaptación de la representación gráfica del modelo que capta adecuadamente la información generada por el modelo.

Modelos para datos con valores ordinales o reales

Todos los modelos discutidos hasta ahora requieren que los valores en la tabla de datos sean binarios. En caso de que los datos asumiesen valores ordinales (por ejemplo, al aplicar un cuestionario utilizando el formato tipo Likert) o reales (cuando se registran variables continuas como el tiempo de reacción o la intensidad de expresiones génicas), se aplicaba, previo al análisis con los modelos originales, una dicotomización de los datos. Esta transformación implica que todos los valores iguales o mayores que cierto punto de corte (escogido de forma más o menos arbitraria) se convierten a 1 y los valores inferiores a 0. Aunque hay numerosas publicaciones que han aplicado con éxito los modelos HICLAS después de una dicotomización, es claro que este procedimiento implica una pérdida de información. Los modelos más recientes desarrollados en el grupo de investigadores de Van Mechelen resuelven este inconveniente[28,29].

Variantes probabilísticas

Casi todos los modelos de la familia HICLAS son *deterministicos*. Esto quiere decir que se postula, a través de la aplicación de cierta fórmula a las matrices características, un valor fijo para cada celda en la tabla. Como vimos en las secciones anteriores, en los modelos disyuntivos y conjuntivos originales se aplica una regla de asociación que resulta en un valor 0 o 1 para cada celda. Aunque el análisis HICLAS considera la posibilidad de que la tabla derivada a partir de esta fórmula o regla muestre discrepancias con la tabla de datos observados (con algoritmos que buscan minimizar el porcentaje de discrepancias), al nivel del modelo no se explica cómo o por qué el valor en una celda difiera del valor observado. Este enfoque contrasta con los modelos probabilísticos, los cuales en vez de especificar un único valor para cada celda en la tabla de datos, consideran una distribución de valores (por ejemplo, asignando una probabilidad de

Tabla 4 Representación algebraica (matrices características) del modelo HICLAS conjuntivo para los datos de la tabla 3

	Matriz característica de los estudiantes				Matriz característica de las materias		
	I	II	III		I	II	III
Andrea	0	1	0	Ana	0	0	1
Bianca	0	1	1	Cel	0	1	0
Carolina	0	0	1	BQ	0	1	1
Diana	1	1	0	Emb	0	1	0
Emma	1	1	0	SP1	1	0	0
Fernanda	1	0	0	SM	1	0	0
Gabriela	1	1	1	IB1	1	1	0
Hilda	0	0	0	IBC	1	1	1
Isabel	1	0	0				
Julia	1	1	0				

Ana : Anatomía; BQ : Bioquímica; Cel : Biología Celular; Emb : Embriología; IB1 : Informática Biomédica I; IBC : Integración Básico-Clínica; SM : Salud Mental; SP1 : Salud Pública I.

observar valores 0 y 1). Probablemente, el inconveniente más importante de los modelos determinísticos es que no proporcionen una base para llevar a cabo contrastes estadísticos (por ejemplo, para comparar estadísticamente la bondad de ajuste de diferentes modelos y decidir que un modelo tenga o no un ajuste «significativamente» mejor). Para solucionar este conveniente, se propuso una variante probabilística, la cual reconsidera el modelo original en un marco bayesiano[30], moviendo así los modelos HICLAS a la estadística convencional con todas las ventajas que implica.

Consideraciones finales

En este artículo se presentó una breve introducción a los modelos de la familia HICLAS. Lo que une a todos los miembros de esta familia es que proporcionan una clasificación jerárquica, simultáneamente para los objetos y atributos de una tabla, a través de la representación de las relaciones de equivalencia, jerarquía y asociación presentes en los datos.

Un inconveniente de los modelos HICLAS es que los algoritmos de estimación no se han incorporado en los paquetes de software comunes como SAS, SPSS, STATA y R. Prácticamente, la única forma para conseguir los programas para correr (una variante específica de) un análisis HICLAS es ponerse en contacto con los autores originales. A pesar de que estos autores generalmente ponen los programas a la disposición de los investigadores sin ningún costo, un obstáculo adicional en algunos casos es que el software no es muy amigable y/o requiere una licencia de un software externo (por ejemplo, MATLAB). Afortunadamente, los autores suelen ser bastante benévolos para guiar a los investigadores en la utilización del software; sin embargo, esta limitación ha sido una de las razones por las que el HICLAS no se ha diseminado a gran escala.

Otro factor relevante al respecto ha sido la escasez de publicaciones no técnicas, que introducen los modelos HICLAS a investigadores no expertos en métodos estadísticos; especialmente en el ámbito español casi no existe literatura sobre el HICLAS. En este sentido, se espera que este artículo contribuya a aumentar la accesibilidad de este método interesante para el análisis de datos y que los investigadores se entusiasmen y exploren las posibilidades de esta herramienta.

Responsabilidades éticas

Protección de personas y animales. Los autores declaran que para esta investigación no se han realizado experimentos en seres humanos ni en animales.

Confidencialidad de los datos. Los autores declaran que en este artículo no aparecen datos de pacientes.

Derecho a la privacidad y consentimiento informado. Los autores declaran que en este artículo no aparecen datos de pacientes.

Financiación

Ninguna.

Conflicto de intereses

El autor declara no tener conflicto de intereses.

Agradecimientos

El autor agradece a José Daniel Morales Castillo y Uri Torruco García sus valiosas sugerencias a una versión anterior de este manuscrito.

Referencias

1. Bailey KD. Typologies and taxonomies: An introduction to classification techniques. Thousand Oaks, CA, EE.UU.: Sage; 1994.
2. Lukk M, Kapushesky M, Nikkilä J, Parkinson H, Goncalves A, Huber W, et al. Global map of human gene expression. Nat Biotechnol. 2010;28:322–4.

3. Reutterer T, Mild A, Natter M, Taudes A. Dynamic segmentation approach for targeting and customizing direct marketing campaigns. J Interact Marketing. 2006;20:43–57.
4. Van Mechelen I. Classipedia: A road map to help traverse the classification jungle. Discurso Presidencial en la Conferencia de la International Federation of Classification Societies, Tilburg. 2013.
5. Van Mechelen I, Bock HH, De Boeck P. Two-mode clustering methods: A structured overview. Stat Methods Med Res. 2004;13:363–94.
6. Hennig C, Meila M, Murtagh F, Rocci R, editores. Handbook of cluster analysis. Boca Raton, FL, EE. UU.: CRC Press; 2015.
7. Mirkin B. Clustering: A data recovery approach. 2.ª ed. Boca Raton, FL, EE. UU.: CRC Press; 2013.
8. Arabie P, Hubert LJ. An overview of combinatorial data analysis. En: Arabie P, Hubert LJ, De Soete G, editores. Clustering and Classification. River Edge, NJ, EE. UU.: World Scientific; 1996. p. 5–63.
9. Shepard RN, Arabie P. Additive clustering: Representation of similarities as combinations of discrete overlapping properties. Psychol Rev. 1979;86:87–123.
10. Johnson SC. Hierarchical clustering schemes. Psychometrika. 1967;32:241–54.
11. Shepard RN. Multidimensional scaling, tree-fitting, and clustering. Science. 1980;210:390–8.
12. Pakhomov SVS, Pedersen T, McInnes B, Melton GB, Ruggieri A, Chute CG. Towards a framework for developing semantic relatedness reference standards. J Biomed Inform. 2011;44:251–65.
13. Hartigan JA. Clustering algorithms. Nueva York, NY, EE.UU.: Wiley; 1975.
14. Rosenberg S, Van Mechelen I, De Boeck P. A hierarchical classes model: Theory and method with applications in psychology and psychopathology. En: Arabie P, Hubert LJ, de Soete G, editores. Clustering and Classification. River Edge, NJ, EE.UU.: World Scientific; 1996, p. 123-55.
15. De Boeck P, Rosenberg S. Hierarchical classes: Model and data analysis. Psychometrika. 1988;53:361–81.
16. Van Mechelen I, De Boeck P, Rosenberg S. The conjunctive model of hierarchical classes. Psychometrika. 1995;60:505–21.
17. Falmagne JC, Koppen M, Villano M, Doignon JP, Johannesen L. Introduction to knowledge spaces: How to build, test and search them. Psychol Rev. 1990;97:201–24.
18. Vansteelandt K, van Mechelen I. Individual differences in situation-behavior profiles: A triple typology model. J Pers Soc Psychol. 1998;75:751–65.
19. Leenen I, Van Mechelen I, De Boeck P. A generic disjunctive/conjunctive decomposition model for *n*-ary relations. J Math Psychol. 1999;43:102–22.
20. Leenen I, Van Mechelen I. An evaluation of two algorithms for hierarchical classes analysis. J Classif. 2001;18:57–80.
21. Ceulemans E, Van Mechelen I, Kuppens P. Adapting the formal to the substantive: Constrained Tucker3-HICLAS. J Classif. 2004;21:19–50.
22. Cattell RB. The meaning and strategic use of factor analysis. En: Cattell RB, editor. Handbook of multivariate experimental psychology. Chicago IL, EE.UU.: Rand McNally; 1966. p. 174–243.
23. Ceulemans E, Van Mechelen I. Uniqueness of *N*-way *N*-mode hierarchical classes models. J Math Psychol. 2003;47: 259–64.
24. Tucker LR. The extension of factor analysis to three-dimensional matrices. En: Frederiksen N, Gulliksen H, editores. Contributions to mathematical psychology. Nueva York, NY, EE.UU.: Holt, Rinehart, & Winston; 1964. p. 109–27.
25. Leenen I, Van Mechelen I, De Boeck P, Rosenberg S. INDCLAS: A threeway hierarchical classes model. Psychometrika. 1999;64:9–24.
26. Ceulemans E, Van Mechelen I, Leenen I. Tucker3 hierarchical classes analysis. Psychometrika. 2003;68:413–33.
27. Leenen I, Ceulemans E. Three-way hierarchical classes: A comparison of the INDCLAS and Tucker3-HICLAS models. Appl Multivariate Res. 2009;13:43–76.
28. Van Mechelen I, Lombardi L, Ceulemans E. Hierarchical classes modeling of rating data. Psychometrika. 2007;72:475–88.
29. Schepers J, Van Mechelen I, Ceulemans E. The real-valued model of hierarchical classes. J Classif. 2011;28:363–89.
30. Leenen I, Van Mechelen I, Gelman A, De Knop S. Bayesian hierarchical classes analysis. Psychometrika. 2008;73:39–64.