Grado en Matemática Computacional

Estancia en Prácticas y Proyecto Final de Grado

Conceptos básicos del Análisis Estadístico Implicativo, su cálculo y aplicación a géneros cinematográficos

Autora:
Paula Ten López

Supervisor:
Fernando Gregori Pérez
Tutor académico:
Pablo Gregori Huerta

Fecha de lectura: Noviembre de 2020
Curso académico 2019/2020

Resumen

El presente documento incluye el proyecto llevado a cabo durante mi estancia en prácticas en PaynoPain Solutions S.L. y el desarrollo teórico de la asignatura *MT1030 – Practicas Externas y Proyecto de Final de Grado*, del Grado en Matemática Computacional en la Universitat Jaume I.

A lo largo de la estancia en prácticas implementé un panel de gestión para un proyecto de monederos electrónicos, tanto a nivel *backend* como la interfaz web.

En relación con la sección teórica, se presenta una descripción general del Análisis Estadístico Implicativo, método de análisis de datos dedicado a la extracción y estructuración de cuasi-implicaciones. Este informe es una síntesis que presenta brevemente el marco estadístico básico de la teoría y ejemplifica los cálculos de todos los conceptos involucrados en la teoría, a través de una pequeña investigación desarrollada por la autora.

Palabras clave

Implicación estadística, similaridad, cohesión, árbol jerarquico, nivel significativo, regla.

Abstract

This document includes the project carried out during my internship at PaynoPain Solutions S.L. and the theoretical development of the subject *MT1030 - External Practices and Final Degree Project*, of the Degree in Computational Mathematics in the Universitat Jaume I.

Throughout the internship I implemented a management panel for an electronic wallets project, both at the backend level and the web interface.

In relation to the theoretical section, a general description of Implicative Statistical Analysis is presented, a data analysis method dedicated to the extraction and structuring of quasi-implications. This report is a synthesis that briefly presents the basic statistical framework of the theory and exemplifies the calculations of all the concepts involved in the theory, through a small investigation developed by the author.

Keywords

Statistical implication, similarity, cohesion, hierarchical tree, significant level, rule.

Agradecimientos

Me gustaría dedicarle el presente trabajo a todas las personas que han confiado en mí durante estos 5 años de grado, a mis viejas y nuevas amistades, compañeros, profesores y sobre todo a mi familia y pareja. Este mérito no hubiera sido posible sin vosotros.

Para empezar quiero agradecer a mis padres por apoyarme en cada una de mis decisiones, por confiar más en mí que yo misma, por darme los mejores consejos que jamás aprenderé y por aguantarme en época de exámenes. Otro pilar fundamental durante mi viaje universitario ha sido mi abuela Nieves, ella y su fe me han dedicado cientos de velas para que yo aprobase cada uno de mis exámenes, pero sus llamadas y preocupación constantes son las que me han mantenido curso tras curso esforzándome por mis objetivos.

Al mismo tiempo, he tenido la suerte de contar con el apoyo de mis amigas de hace 11 años y de las nuevas amistades que he conocido en la carrera, compañeros de matemáticas e informática. Pero destaca una persona que ha estado acompañándome cientos de horas al otro lado de la pantalla mientras yo estudiaba: María, gracias por hacer del confinamiento un recuerdo más bonito. Hemos madrugado y trasnochado juntas mediante Skype estudiando cada una sus asignaturas, dándonos fuerza y ánimos cuando ninguna de las dos tenía.

Por último pero no menos importante está Carles, mi compañero de vida y mi conjunto complementario. Gracias por celebrar conmigo las alegrías y las tristezas, por ponerme los pies en el suelo, por enseñarme lo que es la paciencia, por hacerme ver siempre el lado bueno a las cosas y por regalarme los mejores recuerdos.

Índice general

Capítulo 1

Introducción

1.1. Contexto y motivación del proyecto

El aprendizaje del ser humano está constituido primordialmente de dos factores: los hechos y las reglas entre los hechos o entre las propias reglas. Por medio de la propia cultura y experiencia personal, el proceso de aprendizaje integra una elaboración progresiva de estas formas de conocimiento a pesar de regresiones, interrogantes o cambios que surgen de un confrontamiento decisivo. No obstante, las reglas formadas inductivamente se vuelven bastante estables cuando su número de éxito alcanza un cierto nivel de confianza, en cambio, cuando ese nivel no es logrado, la sensatez del sujeto hará que se resista a su rechazo. A decir verdad, es difícil sustituir una regla inicial por otra cuando se manifiestan pocos contraejemplos. Si aumentan, la confianza de la regla puede disminuir y la regla puede reformularse o incluso descartar. En cambio, cuando las verificaciones son numerosas y los contraejemplos son singulares, la regla se considera sólida y puede perdurar en nuestras mentes.

Por ejemplo, tomamos en consideración la regla robusta "Todos los Ferrari son rojos", aun habiendo uno o dos contraejemplos, la regla se mantiene y nuevamente se confirmará con nuevos ejemplos. Por esa razón, al contrario de lo que sucede en matemáticas, donde las reglas no permiten ninguna excepción, las reglas en ciencias humanas se consideran admisibles cuando el número de contraejemplos es tolerable. En análisis de datos, el problema es determinar un criterio de consenso que cuantifique el nivel de calidad de confianza de la regla de acuerdo con los requisitos del usuario. El enfoque que da del Análisis Estadístico Implicativo (ASI) se basa en tres supuestos epistemológicos: el criterio está establecido sobre bases estadísticas; no es lineal, es resistente al ruido (es decir, no está muy influenciado por los primeros contraejemplos); y pierde legitimidad si los contraejemplos reaparecen con asiduidad. En el presente Proyecto de Fin de Grado presentamos los fundamentos teóricos del ASI y su soporte computacional CHIC.

Capítulo 2

Estancia en prácticas

2.1. Empresa

Llegado el periodo de estancia en prácticas escogí a *PaynoPain Solutions S.L.*, empresa FinTech especializada en investigación y desarrollo tecnológico en medios de pago. Con sede en los Centros Europeos de Empresas Innovadores (CEEI) en Castellón de la Plana, esta organización lleva ofreciendo sus servicios desde el año 2011 y gracias a su esfuerzo consiguió el premio a PYME innovadora en 2013. Actualmente cuentan con dos soluciones de éxito: CHANGEiT y Paylands, en uso en más de 12 países. CHANGEiT es una tecnología *eWallet* que reúne infinidad de funcionalidades de pagos, envíos de dinero, aplicación Punto de Venta e incluso programas de fidelización. Por otra parte, Paylands es una pasarela de pago online de interfaz intuitiva que facilita la gestión de pagos y transacciones en cualquier parte del mundo.

PaynoPain está formado por un equipo de más de 30 personas de distintos perfiles, como son: directivos, desarrolladores de software, encargados de Sistemas, comerciales y personal de administración, marketing y comunicaciones. El supervisor que se me asignó es Fernando Gregori, CTO de la empresa, y mis compañeros de proyecto fueron un desarrollador Java, un encargado de proyectos, y un desarrollador Java y Web.

Figura 2.1: Logotipo de PaynoPain Solutions S.L.

2.2. Proyecto realizado en la empresa

2.2.1. Objetivos del proyecto formativo

Durante la estancia y bajo la supervisión de mi tutor de prácticas, Fernando Gregori, se me asignó el puesto de desarrolladora *FullStack* en el equipo de CHANGEiT con la finalidad de llevar a cabo un panel de gestión de monederos electrónicos, el cual se encargaría de gestionar distintas especificaciones del sistema de wallets para cada cliente, tales como sistemas de límites, notificaciones, comisiones, funcionalidades, etc.

El objetivo principal de este proyecto consiste en ampliar mis aptitudes en programación y bases de datos adoptadas en el grado y adquirir nuevas competencias en el ámbito del desarrollo y diseño web. Dado que la elaboración del panel establece una comunicación desde la parte más interna (*Back end*) hasta la interfaz web visible por el usuario (*Front end*), obtendré los conocimientos básicos de una programadora *FullStack*.

2.2.2. Metodología y definición de tareas

Las 4 partes principales que constituyen el desarrollo del panel que utilicé, y amplié, para dicho propósito son:

- Interfaz *Back end*
    1. *Core*: donde se encuentra toda la lógica de negocio.
    2. La interfaz de programación de aplicaciones (API) de la empresa: interfaz responsable de comunicarse con el panel para gestionar cada acción y enviarla al core.
    3. *Gateways* u objetos de acceso a datos (DAOs): capa que interactúa con la base de datos.
- Interfaz *Front end*
    4. Interfaz Web: parte encargada del diseño de la aplicación y las comunicaciones con la API de la empresa.

Mi trabajo durante esta estancia consistía en completar cada una de las funciones del panel dentro de las 4 partes del proceso de desarrollo descritas. Para llevar a cabo la gestión de monederos electrónicos debía de implementar en el panel 5 funcionalidades básicas: añadir un evento, obtener un evento, borrar un evento, listar los eventos existentes y editar la configuración de éstos.

Por otro lado, las metodologías y tecnologías que usé para dicho propósito fueron:

- Desarrollo de proyectos en estructura hexagonal: esta arquitectura separa por capas las diferentes responsabilidades del software del proyecto.
- Control de versiones con GIT: donde se lleva un registro de los cambios en archivos de código fuente y se coordina el trabajo entre varias personas sobre archivos compartidos.
- Metodología *Test-Driven Development* (TDD) en cada una de las partes del proyecto: para que el código funcione correctamente se desarrollan pruebas que verifican que cada unidad de código realiza la labor para la que ha sido diseñada.

- Ciclo de desarrollo con Integración continua y Despliegue continuo: esta manera de coordinar la creación de software mejora la eficiencia a la hora de incluir cambios.
- Arquitectura *Representational State Transfer* (REST) para establecer la comunicación entre *Front end* y *Back end*: interfaz entre sistemas que utiliza el protocolo HTTP para el intercambio y manipulación de datos.

2.2.3. Planificación temporal de las tareas

La duración de la estancia en prácticas consta de 290 horas presenciales. Estas horas se distribuyeron de manera que las pudiese compaginar con las clases del segundo semestre, haciendo así un total de 31 horas por semana. Este periodo empezó el 3 de febrero de 2020 y finalizó el 14 de abril del mismo año.

Durante estas 10 semanas debía acabar el panel por completo, es decir, implementar las funciones de éste (añadir, obtener, listar, editar y borrar) tanto en *Back end* como en *Front end*, para ello se decidió dedicar 5 semanas a cada parte. El apartado del desarrollo interno y la logística era el más costoso, pero al tener conocimientos previos sobre programación en Java el plazo era accesible. En cambio, era la primera vez que realizaba una tarea relacionada con el diseño web y se determinó dedicarle más tiempo para poder familiarizarme primero con el entorno y después acabar la labor.

2.2.4. Explicación detallada del proyecto

Mi primera tarea fue crearme cuentas en los distintos sitios web donde la empresa desempeña su trabajo: Slack es una herramienta de comunicación en equipo con todos miembros de la empresa, en GitLab es donde se suben y revisan las versiones del código creado, y Teamwork sirve para gestionar los proyectos y describir las tareas que cada empleado debe realizar.

Seguidamente se me explicó el funcionamiento de la logística que siguen, para que yo pudiera desarrollar el proyecto con las mismas directrices. Dado que cada tarea del panel será accesible a través de un *endpoint*[1], las llamaré así a partir de ahora.

Dentro de su API realizan peticiones HTTP para la comunicación entre cliente y servidor, y en este caso haremos uso de: DELETE, GET, PATCH, POST y PUT. Para estas comunicaciones siguen un esquema con las partes bien diferenciadas (Figura 2.2). Cada *endpoint* se compone de un controlador, un controlador de órdenes con su respectiva orden y, si la función lo requiere, una petición y una respuesta. Además, para cerciorarse de que se han programado de manera correcta las diferentes partes, se realizan test de los controladores simulando eventos no deseados para detectar fallos en futuras comunicaciones.

En primer lugar, me centré en el *endpoint* correspondiente a añadir un evento con la petición HTML POST: *AddApplication*. Siguiendo el esquema descrito en la Figura 2.2, programé cada parte de la comunicación y sus correspondientes test.

[1]Un *endpoint* es un extremo de un canal de comunicación. Cuando una API interactúa con otro sistema, los puntos de contacto de esta comunicación se consideran *endpoints*, que pueden incluir una URL de un servidor o servicio. Se trata la ubicación desde la cual las API pueden acceder a los recursos que necesitan para llevar a cabo su función.

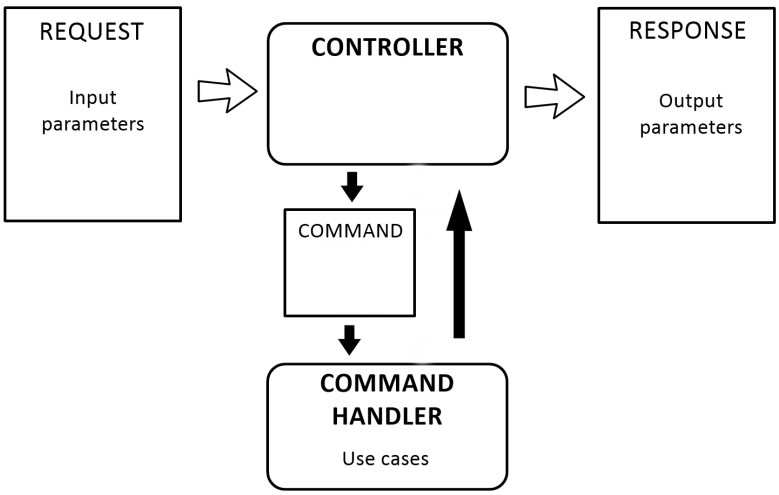

Figura 2.2: Funcionamiento de la logística interna.

Una vez pulido y finalizado el *endpoint AddApplication*, realicé el *endpoint* correspondiente a editar la configuración de un evento (*EditConfiguration*). Para esta tarea la llamada que realizaría sería PUT, el cliente que lo solicite podrá cambiar los parámetros de la configuración de su monedero electrónico. De igual forma elaboré los test para verificar que las partes de la comunicación funcionaban adecuadamente.

De igual modo ejecuté la función de listar eventos dado un cliente o propietario, por esta razón el *endpoint* se nombró *ListOwnerApplications*. Para concluir con la parte de la programación de aplicaciones desarrollé, por completo y de la misma manera que el resto, la opción de obtener la configuración de una cartera electrónica: *GetConfiguration*. La petición HTTP asociada a estas dos funciones es GET.

Llegado el ecuador de la estancia en prácticas mi labor cambió, tal y como estaba planificado desde un principio. Me familiaricé con la infraestructura *Angular* y el lenguaje de programación HTML para empezar con el diseño web del panel. A fin de desarrollar una aplicación web en Angular debía de llevar acabo una componente por cada funcionalidad del panel. Di comienzo a este proceso con la tarea de añadir una aplicación mediante la elaboración de un formulario (Figura 2.3).

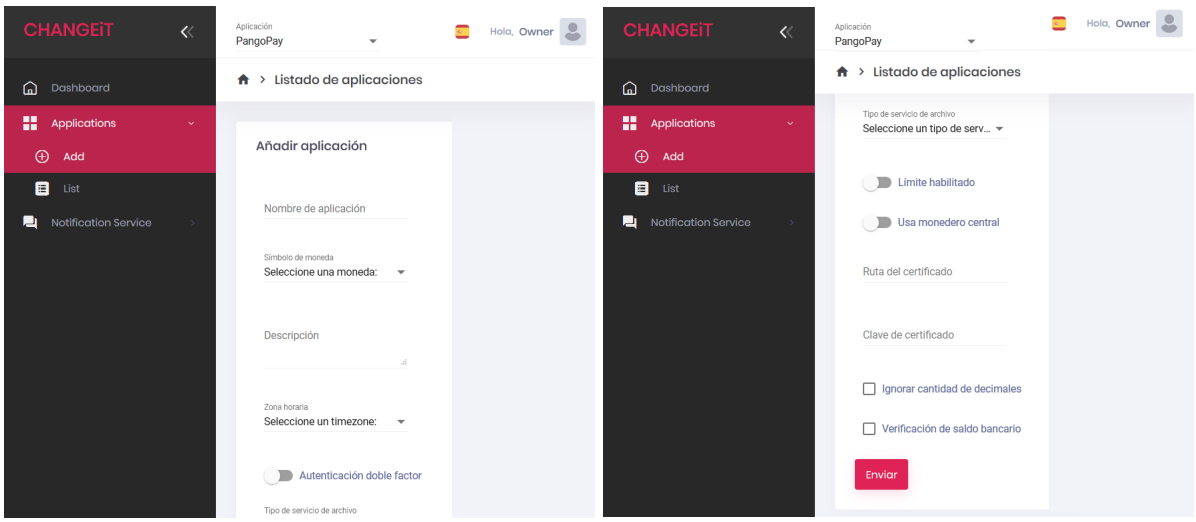

Figura 2.3: Formulario para añadir una nueva aplicación.

La siguiente labor a abordar fue listar las aplicaciones, para ello creé una tabla donde aparecían todas ellas (Figura 2.4). Para tal propósito hice uso de *ListOwnerApplications*.

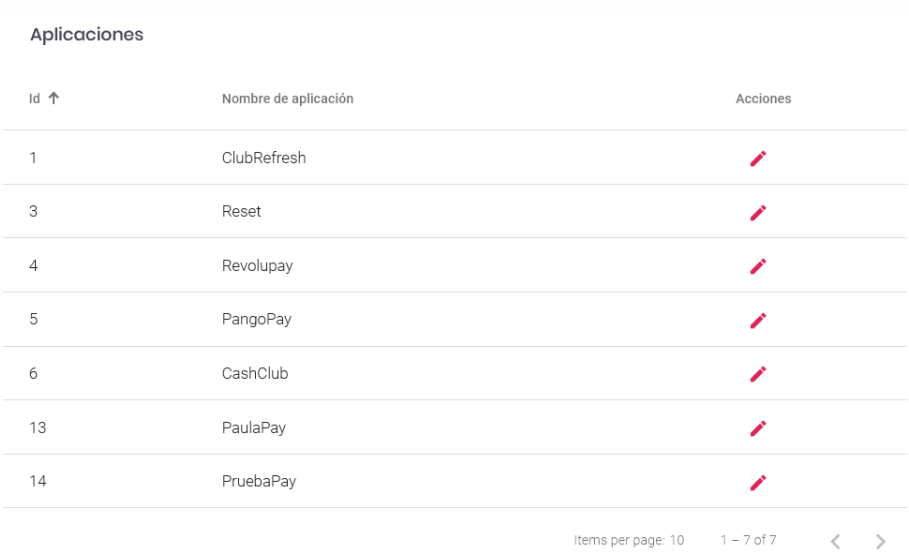

Figura 2.4: Tabla con el listado de aplicaciones de la base de datos interna.

Finalmente, elaboré un último formulario (Figura 2.5) para editar la configuración de cualquier monedero haciendo uso de *GetConfiguration* y *EditConfiguration*.

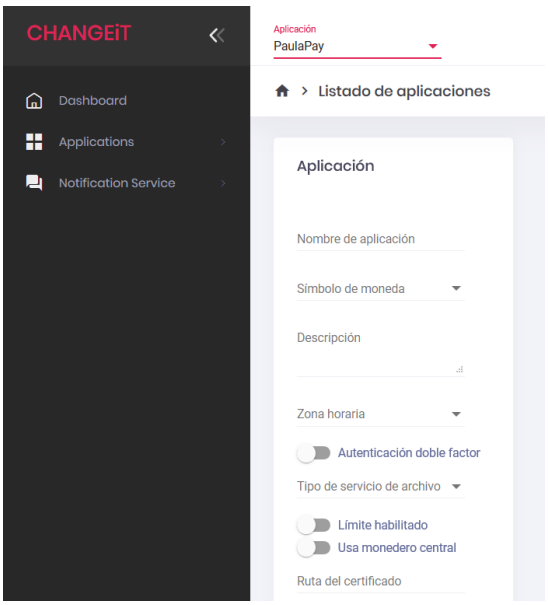

Figura 2.5: Formulario para editar la aplicación seleccionada.

2.2.5. Grado de consecución de los objetivos propuestos

La elaboración del panel de gestión de monederos electrónicos se completó al 80 %, de las 5 funciones iniciales que se habían planeado se llevaron a cabo 4 de ellas dejando por realizar la tarea de borrar un evento/monedero. Esto fue debido a la situación del estado de alarma en la que se comprometió España el 14 de marzo de 2020 a causa de la enfermedad del Covid-19.

Esta circuntancia obligó a seguir el desarrollo de manera telemática con el equipo que yo disponía en casa. Aparte de demorar el proceso unos días por la reinstalación de los programas, no disponía de las prestaciones necesarias para trabajar de manera initerrumpida.

Pese a esta nueva situación, los objetivos didácticos de ampliar mis conocimientos como programadora *FullStack* fueron logrados.

Capítulo 3

Memoria TFG: Conceptos fundamentales del Análisis Estadístico Implicativo

3.1. Motivación y Objetivos

Este trabajo engloba algunos conceptos fundamentales del Análisis Estadístico Implicativo (*Analyse Statistique Implicative*, ASI, en francés) y del programa informático CHIC (*Classification Hiérarchique Implicative et Cohésitive*), así como ejemplos que refuerzan la teoría.

La motivación inicial de este método fue dar respuesta a la pregunta: "si un objeto tiene cierto atributo, ¿posee algún otro?". Es por esto que el objetivo principal del ASI es la estructuración de datos, interrelacionando individuos con variables, estableciendo reglas de asociación a partir de las variables y, a partir de estas reglas, dar explicación y por consiguiente la posible previsión en distintas áreas del conocimiento. Esta propuesta se inició en la Tesis Doctoral [5] y en el artículo [12], y se fue desarrollando con nuevas metodologías como se describe en [7].

Por otra parte, CHIC permite utilizar la mayor parte de los métodos definidos en la teoría del ASI [2]. Su objetivo es descubrir las implicaciones más adecuadas organizándolas en forma de árbol de similaridad, árbol jerárquico o grafo implicativo.

3.2. Introducción

El descubrimiento de reglas, considerado un campo dentro de la minería de datos, y los criterios de evaluación, que permiten distinguir reglas interesantes de las demás, fueron el ámbito de investigación elegido por Régis Gras y sus colaboradores en la década de los 70 al desarrollar la teoría del Análisis Estadístico Implicativo. Su origen es la *modelización estadística de la cuasi-implicación*: cuando se observa la variable o la conjunción de variables "a" en la población, entonces generalmente también se observa la variable "b".

En comparación con otros métodos de análisis simétricos basados en una distancia o una correlación, el ASI se diferencia por utilizar una medida no lineal que satisface criterios importantes a la hora de crear reglas de asociación. Estructura conjuntos según diferentes características comunes complementarias, surgiendo conceptos como: intensidad de implicación, cohesión de clases, niveles significativos, etc. En primer lugar, esta medida se basa en la intensidad de la implicación que mide el nivel de asombro relacionado a una regla (a diferencia del índice propuesto en [13]). Por lo tanto, se suprimen las reglas triviales que son potencialmente evidentes y conocidas por el experto. Esta intensidad de implicación se puede afianzar con el grado de validez, en este caso la medida no tiene en cuenta sólo la legitimidad de la regla, sino su contra-recíproca. Ciertamente, cuando una regla de asociación se considera válida, esto es, el conjunto de elementos A está muy cerca del conjunto de elementos B, entonces es lícito e intuitivo esperar que la regla contra-recíproca sea válida, es decir, el conjunto de elementos que no son B está cerca del conjunto de elementos que no son A. Estas dos magnitudes se complementan con otra en función del tamaño del soporte de la regla. Así, al combinar estas tres medidas, se define una que posee las cualidades de las tres, tiene en cuenta la resistencia al ruido, la regla contra-recíproca y el rechazo de reglas triviales.

Variables

Inicialmente, tanto CHIC como el ASI fueron diseñados para manejar variables binarias. Posteriormente, el ASI se enriqueció con la introducción de otros tipos de variables, que ha incorporado CHIC. Actualmente, CHIC ofrece la posibilidad de tratar variables binarias, variables de frecuencia, variables definidas en intervalos y variables-intervalo. El caso de las variables binarias es obviamente el más sencillo, ya que solo se puede contestar afirmativa o negativamente, este tipo de variable es la que usaremos en este trabajo para mostrar manualmente la ejecución de las fórmulas presentadas. Las variables frecuenciales, en cambio, toman valores entre 0 y 1, expresando un grado modal. Las variables sobre intervalos y las variables-intervalo se utilizan para modelar situaciones complejas, se emplean para descomponer un intervalo numérico en otros más pequeños, mediante un criterio que minimiza la varianza, parecido al método de las k-medias (se conoce como Nubes Dinámicas [3, 4]). Por lo tanto, estas categorías más pequeñas están representadas por una variable binaria y un individuo pertenece a un único intervalo.

Datos

Presentamos ahora los datos que utilizaremos como ejemplo a lo largo de la memoria para comprender y visualizar las distintas fórmulas que irán apareciendo. La información que emplearemos emana de la investigación desarrollada por la autora del presente proyecto, donde se ha analizado el razonamiento de 20 individuos, cuyas edades están comprendidas entre los 20 y los 50 años. En esta investigación se ha trabajado con variables binarias, en vista de que en el cuestionario que le fue aplicado a los sujetos, se les preguntaba si les gustaba o no un determinado género cinematográfico: Acción (ACCI), Comedia (COME), Histórico (HIST), etc.

La información del estudio realizado por Paula Ten se encuentra en la Tabla 3.1. La explicación de las abreviaturas empleadas en los géneros cinematográficos se encuentra en la Tabla 3.2.

	ACCI	ANIM	AVEN	BELI	CFIC	COME	DOCU	DRAM	FANT	HIST	MIST	MUSI	POLI	ROMA	TERR
p1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	0
p2	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
p3	1	1	1	1	1	1	1	0	1	1	1	0	1	0	1
p4	1	0	1	0	1	1	1	0	1	1	1	0	1	0	1
p5	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0
p6	1	1	1	0	1	1	0	0	1	0	1	1	0	1	1
p7	1	0	0	0	1	1	1	0	1	1	1	1	0	1	1
p8	0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
p9	0	1	1	1	1	0	1	1	1	1	1	1	0	1	1
p10	1	1	1	0	0	1	1	0	0	1	1	0	1	0	0
p11	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0
p12	1	1	1	0	1	1	0	1	1	1	1	1	1	0	1
p13	1	1	1	1	1	1	0	0	0	0	1	0	1	1	0
p14	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
p15	1	1	1	0	1	1	0	0	1	0	1	1	1	1	1
p16	1	0	0	1	0	1	1	0	1	1	1	1	1	1	0
p17	1	1	1	1	1	1	1	0	1	1	1	0	1	1	0
p18	1	1	1	0	1	1	0	1	1	0	1	1	1	1	1
p19	1	1	1	1	1	1	1	0	1	1	1	0	1	1	0
p20	1	0	1	0	1	1	1	1	0	0	1	1	1	1	1

Tabla 3.1: *Matriz de datos de* 20×15.

ACCI: Acción	COME: Comedia	MIST: Misterio
ANIM: Animación	DOCU: Documental	MUSI: Musical
AVEN: Aventura	DRAM: Drama	POLI: Policíaco
BELI: Bélico	FANT: Fantasía	ROMA: Romántico
CFIC: Ciencia Ficción	HIST: Histórico	TERR: Terror

Tabla 3.2: *Leyenda para los géneros cinematográficos.*

3.3. Análisis Clasificatorio

El análisis clasificatorio tiene por objetivo juntar a aquellos objetos que comparten características similares, esta técnica está diseñada para desvelar las agrupaciones (*clusters*) naturales dentro de un grupo. A continuación se presentan los distintos conceptos que utiliza este análisis para evaluar las diferencias y similitudes entre objetos y poder determinar agrupamientos. Tomaremos como índice de similaridad el índice de Lerman [11].

La noción de similaridad es esencial, es una medida de correspondencia entre los objetos que van a ser asociados. La metodología más común se basa en medir la equivalencia en términos de la distancia entre pares de elementos, es decir, aquellos con distancias reducidas entre ellos son más parecidos entre sí, que aquellos cuyas distancias son mayores y se reunirán, en consecuencia, dentro del mismo cluster.

Consideramos un conjunto I formado por n sujetos y un conjunto A formado por p propiedades $A = \{a_1, a_2, ..., a_p\}$, se supone además que:

$$A_i = \{x \in I \mid a_i(x) = 1\}, \quad Card(I) = n \quad \text{y} \quad Card(A_i) = n_{a_i}.$$

Continuando con la formalización, consideramos dos conjuntos $A_i, A_j \subset I$, con $i \neq j$ y $1 \leq i, j \leq p$, y extraemos dos partes cualesquiera X_i e X_j de I, elegidas aleatoria e independientemente (ausencia de relación a priori) y de cardinales iguales a los de A_i y A_j respectivamente. El cardinal $Card(X_i \cap X_j)$ va a ser la variable de interés durante el análisis clasificatorio.

En comparación con otros métodos de clasificación, CHIC calcula los índices de similaridad en términos de una probabilidad para la variable $Card(X_i \cap X_j)$ y con posibilidad de elegir entre la ley de distribución Binomial o Poisson.

Árbol de similaridad

Para cada pareja de variables (a_i, a_j) se calculan los Índices de Proximidad o Similaridad [10, 11]:

$$s(a_i, a_j) = Pr\lfloor Card(X_i \cap X_j) \leq K \rfloor \tag{3.1}$$

donde $K := Card(A_i \cap A_j)$. El valor de estas probabilidades está sujeto a la ley asumida, no obstante, CHIC realiza la aproximación de éstas a la distribución Normal y devuelve:

$$s(a_i, a_j) = Pr\left[\frac{Card(X_i \cap X_j) - \frac{n_{a_i} * n_{a_j}}{n}}{\sqrt{\frac{n_{a_i} * n_{a_j}}{n}}} \leq K_c\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{K_c} e^{-\frac{1}{2}x^2}\, dx, \tag{3.2}$$

donde $K_c := \frac{K - \frac{n_{a_i} * n_{a_j}}{n}}{\sqrt{\frac{n_{a_i} * n_{a_j}}{n}}}$.

Los $s(a_i, a_j)$ obtenidos mediante (3.1) representan las similaridades de cada pareja (a_i, a_j) al nivel cero de la jerarquía.

Ejemplo

Para la ejemplificación de los cálculos mostraremos los resultados ligados a las variables: BELI, DRAM, MUSI, ROMA y TERR. Para estas variables $n_{BELI} = 9$, $n_{DRAM} = 10$, $n_{MUSI} = 14$, $n_{ROMA} = 13$, $n_{TERR} = 11$ y $n = 20$. En la Tabla 3.3 se muestran los valores de las variables seleccionadas y en la Tabla 3.4 las copresencias ($Card(A_i \cap A_j)$), las copresencias estandarizadas (K_c) y los índices de similaridad $s(a_i, a_j)$, calculados mediante la aproximación de la distribución Normal.

	BELI	DRAM	MUSI	ROMA	TERR
p1	0	1	1	1	0
p2	1	1	1	0	1
p3	1	0	0	0	1
p4	0	0	0	0	1
p5	0	1	1	1	0
p6	0	0	1	1	1
p7	0	0	1	1	1
p8	0	1	1	0	1
p9	1	1	1	1	1
p10	0	0	0	0	0
p11	1	1	1	1	0
p12	0	1	1	0	1
p13	1	0	0	1	0
p14	1	1	1	0	0
p15	0	0	1	1	1
p16	1	0	1	1	0
p17	1	0	0	1	0
p18	0	1	1	1	1
p19	1	0	0	1	0
p20	0	1	1	1	1

Tabla 3.3: *Submatriz de la matriz de datos.*

Variables		$K = Card(A_i \cap A_j)$	K_c	$s(a_i, a_j)$
BELI	DRAM	4	-0.24	0.406832
BELI	MUSI	5	-0.52	0.302253
BELI	ROMA	6	0.06	0.524726
BELI	TERR	3	-0.88	0.190390
DRAM	MUSI	10	1.13	0.871580
DRAM	ROMA	6	-0.20	0.422260
DRAM	TERR	6	0.21	0.584415
MUSI	ROMA	10	0.30	0.617281
MUSI	TERR	9	0.47	0.680282
ROMA	TERR	6	-0.43	0.333570

Tabla 3.4: *Valores de copresencias, copresencias estandarizadas e índices de similaridad.*

Con la finalidad de obtener las agrupaciones se hace uso de la clasificación clásica:

- Se forma una matriz con los Índices de Similaridad $s(a_i, a_j)$, obtenidos a partir de las combinaciones de todas las variables.

  Nivel cero de la jerarquía:

	BELI	DRAM	MUSI	ROMA	TERR
BELI	1.000000	0.406832	0.302253	0.524726	0.190390
DRAM	0.406832	1.000000	**0.871580**	0.422260	0.584415
MUSI	0.302253	**0.871580**	1.000000	0.617281	0.680282
ROMA	0.524726	0.422260	0.617281	1.000000	0.333570
TERR	0.190390	0.584415	0.680282	0.333570	1.000000

- Se obtienen de nuevo los Índices de Similaridad al combinar cada variable con la clase $s(a_i, a_j)$ de mayor índice [10], de modo que:

  - con cada variable aislada: $s((a_i, a_j), a_k) = Max[s(a_i, a_k), s(a_j, a_k)]^2$,
  - con clases:

    $$s(C_1, C_2) = \{Max[s_i \mid s_i = s(a_j, a_k) \ \ \forall a_j \in C_1, a_k \in C_2]\}^{Card(C_1)\times Card(C_2)}$$

    donde C_1 y C_2 son dos clases previamente formadas [10].

  Nivel uno de la jerarquía: se unen las variables DRAM y MUSI ya que tienen el mayor índice de similaridad (0.871580) y se recalcula la matriz de similaridad.

	BELI	[DRAM, MUSI]	ROMA	TERR
BELI	1.000000	0.165512	**0.524726**	0.190390
[DRAM, MUSI]	0.165512	1.000000	0.381036*	0.462784
ROMA	**0.524726**	0.381036*	1.000000	0.333570
TERR	0.190390	0.462784	0.333570	1.000000

  * Este valor se obtiene de la siguiente manera:

$$\begin{aligned} s((DRAM, MUSI), ROMA) &= \{Max[s(DRAM, ROMA), s(MUSI, ROMA)]\}^2 \\ &= \{Max[0.422260, 0.617281]\}^2 = \{0.617281\}^2 = 0.381036. \end{aligned}$$

  Nivel dos de la jerarquía: se unen las variables BELI y ROMA ya que tienen el mayor índice de similaridad (0.524726) y se recalcula la matriz de similaridad.

	[BELI, ROMA]	[DRAM, MUSI]	TERR
[BELI, ROMA]	1.000000	0.145188**	0.111269
[DRAM, MUSI]	0.145188**	1.000000	**0.462784**
TERR	0.111269	**0.462784**	1.000000

** Este valor se obtiene de la siguiente manera:

$$s((BELI, ROMA), (DRAM, MUSI)) =$$
$$= \left\{ Max \begin{bmatrix} s(BELI, DRAM), & s(BELI, MUSI), \\ s(DRAM, ROMA), & s(MUSI, ROMA) \end{bmatrix} \right\}^{2*2} =$$
$$= \{Max[0.406832, 0.302253, 0.422260, 0.617281]\}^4 = \{0.617281\}^4 = 0.145188.$$

Nivel tres de la jerarquía: por último se unen la clase [DRAM, MUSI] con la variable TERR, puesto que tienen el mayor índice de similaridad (0.462784) y se recalcula la matriz de similaridad.

	[BELI, ROMA]	[[DRAM, MUSI], TERR]
[BELI, ROMA]	1.000000	0.055322
[[DRAM, MUSI], TERR]	0.055322	1.000000

Por último:

$$s((BELI, ROMA), ((DRAM, MUSI), TERR)) =$$
$$= \left\{ Max \begin{bmatrix} s(BELI, DRAM), & s(BELI, MUSI), & s(BELI, TERR), \\ s(ROMA, DRAM), & s(ROMA, MUSI), & s(ROMA, TERR) \end{bmatrix} \right\}^{2*3} =$$
$$= \left\{ Max \begin{bmatrix} 0.406832; & 0.302253; & 0.190390; \\ 0.422260; & 0.617281; & 0.333570 \end{bmatrix} \right\}^6 = \{0.617281\}^6 = 0.055322.$$

La representación gráfica de estos resultados calculados se muestra en la Figura 3.1.

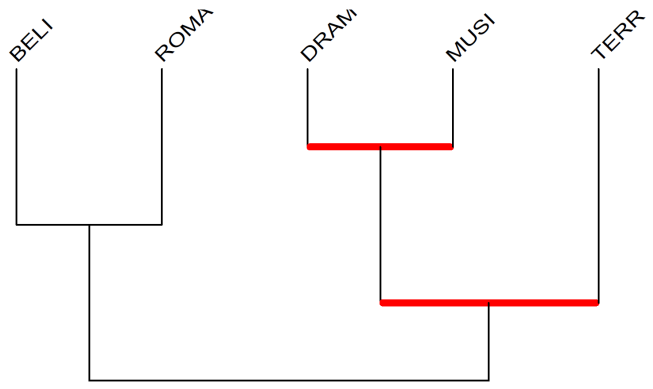

Figura 3.1: *Árbol de similaridad.*

Nodos significativos

Los nodos significativos de un árbol de similaridad son aquellos correspondientes a una clasificación compatible lo mejor posible con los valores y la calidad de los valores de similaridad.

Definición 1 *Se llama* ***preorden inicial y global*** Ω *sobre* $A \times A$*, al preorden inducido por la aplicación similaridad* S *sobre* $A \times A$*.*

$$G_s(\Omega) = \{((a,b);(c,d)) \mid s(a,b) < s(c,d)\}$$

Llamamos $S\Pi_k$ al conjunto de pares separados al nivel k y $R\Pi_k$ al conjunto de pares que ya se han reunido en el nivel k. La intersección de los conjuntos $G_s(\Omega) \cap [S\Pi_k \times R\Pi_k]$ está formada por los pares de parejas que en la altura k respetan el preorden inicial y que además la primera pareja se encuentra separada mientras que la segunda está reunida. Por ejemplo, si $((a,b),(c,d)) \in G_s(\Omega) \cap [S\Pi_k \times R\Pi_k]$ se tiene que $s(a,b) < s(c,d)$ y que las variables (a,b) se encuentran separadas y (c,d) reunidas al nivel k.

No obstante, el cardinal de este conjunto varía en función de la altura k. Al cardinal de $G_s(\Omega) \cap [S\Pi_k \times R\Pi_k]$ se le asocia el índice aleatorio $G_s(\Omega^*) \cap [S\Pi_k \times R\Pi_k]$, donde Ω^* es un preorden aleatorio provisto de una probabilidad uniforme. Este índice tiene [8]:

- por esperanza: $\frac{1}{2}s_k r_k$, y
- por varianza: $\frac{s_k r_k (s_k + r_k + 1)}{12}$,

siendo $s_k = Card[S\Pi_k]$ y $r_k = Card[R\Pi_k]$. El índice centrado se define como:

$$S(\Omega,k) = \frac{Card[G_s(\Omega) \cap [S\Pi_k \times R\Pi_k]] - \frac{1}{2}s_k r_k}{\sqrt{\frac{s_k r_k (s_k + r_k + 1)}{12}}}.$$

Definición 2 *Se llama* ***nivel significativo*** *a todo nivel que corresponde a un máximo local de* $S(\Omega,k)$ *durante la construcción de la jerarquía. Se llama* ***nodo significativo*** *a cualquier nodo formado a un nivel que corresponde con un máximo local de* $v(\Omega,k)$ *donde:*

$$v(\Omega,k) = S(\Omega,k) - S(\Omega,k-1).$$

CHIC determina los nodos significativos para la clasificación obtenida, los cuales se muestran gráficamente con trazos gruesos y rojos (Figura 3.1). Para el ejemplo desarrollado, aparecen los nodos significativos en los niveles 1 y 3. Examinemos cómo se determinaron estos nodos, para ello utilizaremos las siguientes notaciones:

- $t = Card(A)$, $I_k = \{j \mid p_j \in R\Pi_k\}$,
- p_j son los elementos del preorden inicial Ω, con $j = 1, ..., m$ con $m = Card(\Omega)$,
- $P_l = \{p_j \in \Omega \mid s(p_j) = s_l\}$, donde s_l es un valor cualquiera que puede tomar el índice de similaridad y $l = 1, ..., d$, siendo d la cantidad de valores diferentes que toma este índice,
- $f_i^k = Card[p_j \mid j \notin I_k, j < i, s(p_j) = s(p_i)]$.

En función de los índices de similaridad que se muestran en la Tabla 3.4, se puede establecer el preorden inicial y global Ω sobre $A^* \times A^*$, donde $A^* \subset A$ y
$A^* = \{BELI, DRAM, MUSI, ROMA, TERR\}$, el cual representamos mediante diez pilas, cada una de las cuales comprende a los pares de variables que tienen el mismo índice de similaridad:

$$\Omega = \left\{ \begin{bmatrix} (BELI,TERR) \\ (TERR,BELI) \end{bmatrix} < \begin{bmatrix} (BELI,MUSI) \\ (MUSI,BELI) \end{bmatrix} < \begin{bmatrix} (ROMA,TERR) \\ (TERR,ROMA) \end{bmatrix} < \right.$$
$$< \begin{bmatrix} (BELI,DRAM) \\ (DRAM,BELI) \end{bmatrix} < \begin{bmatrix} (DRAM,ROMA) \\ (ROMA,DRAM) \end{bmatrix} < \begin{bmatrix} (BELI,ROMA) \\ (ROMA,BELI) \end{bmatrix} <$$
$$< \begin{bmatrix} (DRAM,TERR) \\ (TERR,DRAM) \end{bmatrix} < \begin{bmatrix} (MUSI,ROMA) \\ (ROMA,MUSI) \end{bmatrix} < \begin{bmatrix} (MUSI,TERR) \\ (TERR,MUSI) \end{bmatrix} <$$
$$\left. < \begin{bmatrix} (DRAM,MUSI) \\ (MUSI,DRAM) \end{bmatrix} \right\}, \quad (3.3)$$

$$G_s(\Omega) = \left\{ \begin{array}{l} ((BELI,TERR),(BELI,MUSI)); ...; ((BELI,TERR),(MUSI,DRAM)); \\ ((TERR,BELI),(BELI,MUSI)); ...; ((TERR,BELI),(MUSI,DRAM)); ...; \\ ((TERR,MUSI),(DRAM,MUSI)); ((TERR,MUSI),(MUSI,DRAM)) \end{array} \right\},$$

$t = Card(A^*) = 5$, $d = 10$, $m = Card(\Omega) = 2\binom{t}{2} = 2\binom{5}{2} = 20$ y

$$Card(G_s(\Omega)) = \frac{m(m-1)}{2} - \sum_{l=1}^{d} \frac{Card(P_l)(Card(P_l)-1)}{2} =$$
$$= \frac{20(20-1)}{2} - \sum_{l=1}^{10} \frac{Card(P_l)(Card(P_l)-1)}{2} = 190 - 10 = 180.$$

El $Card[G_s(\Omega) \cap [S\Pi_k \times R\Pi_k]]$ se puede extraer aplicando la fórmula siguiente:

$$Card[G_s(\Omega) \cap [S\Pi_k \times R\Pi_k]] = \sum_{j=1}^{r_k} i_j - \frac{r_k(r_k+1)}{2} - \sum_{j=1}^{r_k} f_j^k \ .$$

Nivel 1 de la jerarquía: a esta altura se han reunido las variables DRAM y MUSI, por tanto $R\Pi_1 = \{(DRAM, MUSI)\}$,

$$S\Pi_1 = \left\{ \begin{array}{l} (BELI,TERR);(TERR,BELI);(BELI,MUSI);(MUSI,BELI); \\ (ROMA,TERR);(TERR,ROMA);(BELI,DRAM);(DRAM,BELI); \\ (DRAM,ROMA);(ROMA,DRAM);(BELI,ROMA);(ROMA,BELI); \\ (DRAM,TERR);(TERR,DRAM);(MUSI,ROMA);(ROMA,MUSI); \\ (MUSI,TERR);(TERR,MUSI);(MUSI,DRAM) \end{array} \right\},$$

$r_1 = Card(R\Pi_1) = 1$, $s_1 = Card(S\Pi_1) = 19$, $I_1 = \{19\}$, $f_1^1 = 0$ y
$Card[G_s(\Omega) \cap [S\Pi_1 \times R\Pi_1] = \sum_{j=1}^{r_1} i_j - \frac{r_1(r_1+1)}{2} - \sum_{j=1}^{r_1} f_j^1 = 19 - 1 - 0 = 18.$

Nivel 2 de la jerarquía: se incorpora el par $(BELI, ROMA)$ a los reunidos y en consecuencia $R\Pi_2 = \{(DRAM, MUSI); (BELI, ROMA)\}$,

$$S\Pi_2 = \left\{ \begin{array}{l} (BELI, TERR); (TERR, BELI); (BELI, MUSI); (MUSI, BELI); \\ (ROMA, TERR); (TERR, ROMA); (BELI, DRAM); (DRAM, BELI); \\ (DRAM, ROMA); (ROMA, DRAM); (ROMA, BELI); (DRAM, TERR); \\ (TERR, DRAM); (MUSI, ROMA); (ROMA, MUSI); (MUSI, TERR); \\ (TERR, MUSI); (MUSI, DRAM) \end{array} \right\},$$

$r_2 = 2$, $s_2 = 18$, $I_2 = \{19, 11\}$, $f_1^2 = 0$, $f_2^2 = 0$ y $Card[G_s(\Omega) \cap [S\Pi_2 \times R\Pi_2] = 27$.

Nivel 3 de la jerarquía: se construye la clase $((DRAM, MUSI), TERR)$ y por ende $R\Pi_3 = \{(DRAM, MUSI); (BELI, ROMA); (DRAM, TERR); (MUSI, TERR)\}$,

$$S\Pi_3 = \left\{ \begin{array}{l} (BELI, TERR); (TERR, BELI); (BELI, MUSI); (MUSI, BELI); \\ (ROMA, TERR); (TERR, ROMA); (BELI, DRAM); (DRAM, BELI); \\ (DRAM, ROMA); (ROMA, DRAM); (ROMA, BELI); (TERR, DRAM); \\ (MUSI, ROMA); (ROMA, MUSI); (TERR, MUSI); (MUSI, DRAM) \end{array} \right\},$$

$r_3 = 4$, $s_3 = 16$, $I_3 = \{19, 11, 13, 17\}$, $f_1^3 = 0$, $f_2^3 = 0$, $f_3^3 = 0$, $f_4^3 = 0$ y
$Card[G_s(\Omega) \cap [S\Pi_3 \times R\Pi_3] = 50$.

Nivel 4 de la jerarquía: a este nivel se forma la clase $((BELI, ROMA)$, $((DRAM, MUSI)$, $TERR))$, y por ello,

$$R\Pi_4 = \left\{ \begin{array}{l} (DRAM, MUSI); (BELI, ROMA); (DRAM, TERR); (MUSI, TERR); \\ (BELI, DRAM); (BELI, MUSI); (BELI, TERR); (ROMA, DRAM); \\ (ROMA, MUSI); (ROMA, TERR) \end{array} \right\},$$

$$S\Pi_4 = \left\{ \begin{array}{l} (TERR, BELI); (MUSI, BELI); (TERR, ROMA); (DRAM, BELI); \\ (DRAM, ROMA); (ROMA, BELI); (TERR, DRAM); (MUSI, ROMA); \\ (TERR, MUSI); (MUSI, DRAM) \end{array} \right\},$$

$r_4 = 10$, $s_4 = 10$, $I_4 = \{19, 11, 13, 17, 7, 3, 1, 10, 16, 5\}$, $f_1^4 = 0$, $f_2^4 = 0$, $f_3^4 = 0$, $f_4^4 = 0$, $f_5^4 = 0$, $f_6^4 = 0$, $f_7^4 = 0$, $f_8^4 = 1$, $f_9^4 = 1$, $f_{10}^4 = 0$ y $Card[G_s(\Omega) \cap [S\Pi_4 \times R\Pi_4] = 45$.

En la Tabla 3.5 se simplifican los cardinales de los conjuntos formados y se muestran los valores del índice centrado $S(\Omega, k)$ y de la función $v(\Omega, k)$, además en la Figura 3.2 aparece el gráfico de esta última columna:

Nivel	r_k	s_k	$Card[G_s(\Omega) \cap [S\Pi_k \times R\Pi_k]]$	$S(\Omega, k)$	$v(\Omega, k)$
1	1	19	18	1.474087	1.474087
2	2	18	27	1.133893	-0.340194
3	4	16	50	1.700840	0.566947
4	10	10	45	-0.377964	-2.078805

Tabla 3.5: *Valores del índice centrado y de la función* $v(\Omega, k)$.

Conforme la Definición 2, los nodos significativos serían los que surgen a los niveles 1 y 3, debido a que corresponden a máximos locales de $v(\Omega, k)$ (Figura 3.2).

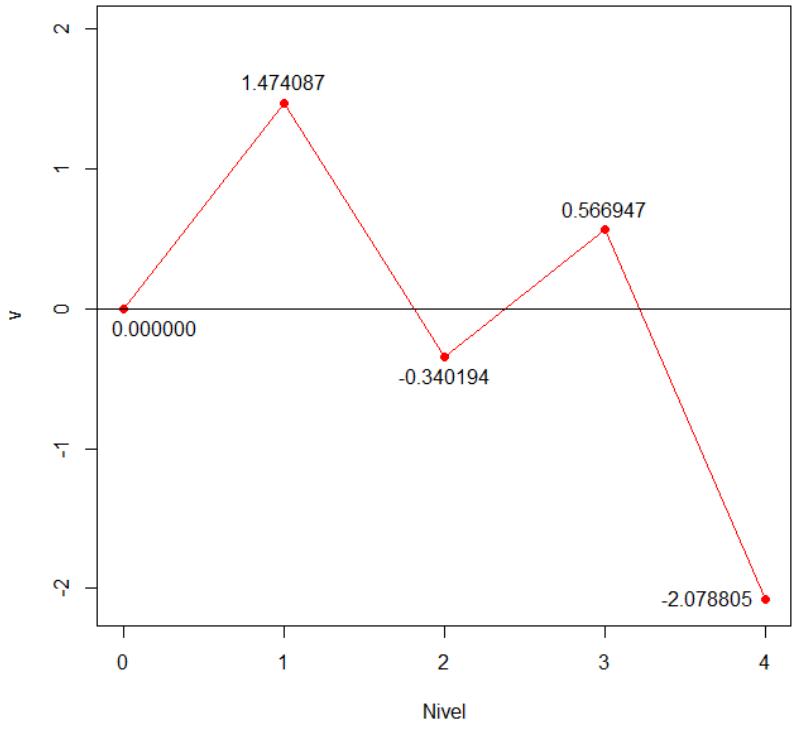

Figura 3.2: *Gráfica de* $v(\Omega, k)$.

Los cálculos realizados para obtener los niveles significativos han sido evaluados según el orden establecido en (3.3). Dentro de estar ordenados por los índices de similaridad, los pares siguen este orden en concreto: $i_1 = (BELI, TERR)$, $i_2 = (TERR, BELI)$, ..., $i_{19} = (DRAM, MUSI)$, $i_{20} = (MUSI, DRAM)$. Para cada bloque del preorden Ω, donde los pares comparten índice de similaridad, los tándems pueden ir ordenados según el usuario haya concretado, es por esta razón que los valores de $Card[G_s(\Omega) \cap [S\Pi_k \times R\Pi_k]]$, y en consecuencia de $S(\Omega, k)$ y $v(\Omega, k)$ difieren de los que CHIC ha calculado. No podemos saber qué orden ha seguido el programa dentro de cada pila de Ω, por lo que se ha optado por proceder por el preorden inicial que aparece en (3.3).

3.4. Análisis Estadístico Implicativo

Con el propóstio de modelar la extracción y representación de reglas deductivas no simétricas "si a entonces casi b", profundizaremos en los datos para poder presuponer una posible relación causal que describa y estructure una población con fines descriptivos y, de ser posible, predictivos. Al igual que en el análisis clasificatorio, en el ASI se contempla un conjunto I compuesto de n individuos y un conjunto A formado por p atributos $A = \{a_1, a_2, ..., a_p\}$, además se supone:

$$A_i = \{x \in I \mid a_i(x) = 1\}, \quad Card(I) = n, \quad Card(A_i) = n_{a_i} \quad \text{y} \quad Card(\overline{A}_i) = n_{\overline{a_i}}.$$

El análisis se desarrolla por medio de los índices de implicación y de cohesión presentados en los trabajos de R. Gras, estos índices se han definido en términos de una probabilidad . En este caso la variable aleatoria de interés es $Card(X_i \cap \bar{X}_j)$, que hace referencia al número de individuos que tienen el atributo a_i pero no el a_j.

Intensidad de implicación

En lógica matemática, la regla "$a_i \to a_j$" es verdadera si para todo x en I, $a_j(x)$ sólo es nulo cuando $a_i(x)$ lo sea también. No obstante, en la realidad esta inclusión estricta se observa excepcionalmente. La formalización matemática de esta situación se expresa a través de la cuasi-implicación $a_i \to a_j$ [6].

La cuasi-implicación $a_i \to a_j$ significa que "en el momento que a_i está presente entonces generalmente a_j también lo está". El objetivo de la Intensidad Implicativa es expresar la inverosimilitud ("asombro") del número de contraejemplos que invalidad la regla $a_i \to a_j$, de manera que se compara el número de contraejemplos observados con el número de contraejemplos esperado bajo la hipótesis de ausencia de relación.

Los **nodos** internos que representan la jerarquía dirigida del grafo implicativo describen relaciones implicativas complejas, llamadas **R-reglas**, entre los atributos de A.

- Cuando $R \to a_i$, se concibe a a_i como una consecuencia de $\mathbf{R}$.
- La R-regla $a_i \to R$, indica que una R-regla $\mathbf{R}$ puede ser deducida de la observación de a_i.
- La R-regla $R' \to R''$, designa que la propiedad $\mathbf{R}''$ es el corolario de una propiedad previamente definida $\mathbf{R}'$.

Seleccionamos aleatoriamente dos subconjuntos U, $V \subseteq I$, con n_{a_i} y n_{a_j} elementos respectivamente. Sea $X_{a_i \wedge \overline{a}_j} = Card(U \cap \overline{V})$ la variable aleatoria asociada con el número de contraejemplos.

Definición 3 *La **regla** $a_i \to a_j$ se dice que es **admisible** al nivel de confianza $1 - \alpha$ si:*

$$P\left\lfloor X_{a_i \wedge \overline{a}_j} \leq n_{a_i \wedge \overline{a}_j} \right\rfloor \leq \alpha \tag{3.4}$$

donde $n_{a_i \wedge \overline{a}_j} = Card(A_i \cap \overline{A_j})$ es el número de contraejemplos a la regla $a_i \to a_j$ observados en la muestra [12].

Cuando una regla es admisible significa que el número de contraejemplos observados es pequeño. La distribución de probabilidad de $X_{a_i \wedge \overline{a}_j}$ puede ser la Binomial, Hipergeométrica o Poisson.

Definición 4 *Dados los subconjuntos $X, Y \subseteq I$ extraídos aleatoriamente con $Card(X) = n_{a_i}$ y $Card(Y) = n_{a_j}$ y sea $K = Card(X \cap \overline{Y})$, entonces la* ***Intensidad Implicativa*** *de la regla $a_i \to a_j$ se define como:*

$$\varphi(a_i, a_j) = 1 - P\left\lfloor K \leq n_{a_i \wedge \overline{a}_j} \right\rfloor, \quad si \;\; n_{a_j} \neq n,$$

en caso contrario, $\varphi(a_i, a_j) = 0$.

La regla es retenida para un α dado si: $\varphi(a_i, a_j) \geq 1 - \alpha$, lo cual es análogo a decir que la regla es admisible, según la Definición 3, con confianza $1 - \alpha$.

Definición 5 *El índice de Implicación de la regla $a_i \to a_j$ se define como:*

$$q(a_i, \overline{a}_j) = \frac{n_{a_i \wedge \overline{a}_j} - \frac{n_{a_i} n_{\overline{a}_j}}{n}}{\sqrt{\frac{n_{a_i} n_{\overline{a}_j}}{n}}}. \tag{3.5}$$

Proseguiremos nuestra ejemplificación con los datos de la Tabla 3.1. Consideramos que la variable aleatoria $K = Card(X \cap \overline{Y})$ sigue una distribución Binomial de parámetros n y p, con $p = \frac{n_{a_i \wedge \overline{a}_j}}{n}$, probabilidad que bajo la hipótesis de independencia se puede escribir como $p = \frac{n_{a_i}}{n} \frac{n_{\overline{a}_j}}{n}$ y utilizaremos $K_0 = Card(A_i \cap \overline{A}_j)$. Además trabajaremos con las variables BELI, DRAM, HIST, MUSI, POLI, ROMA y TERR.

En las tablas 3.6 y 3.7 se muestran, respectivamente, la sub-tabla con los datos para las variables seleccionadas y algunas de las implicaciones que se pueden formar, así como el cardinal, el valor de la probabilidad p y el número de contraejemplos de la regla $a_i \to a_j$.

	BELI	DRAM	HIST	MUSI	ROMA	TERR
p1	0	1	1	1	1	0
p2	1	1	1	1	0	1
p3	1	0	1	0	0	1
p4	0	0	1	0	0	1
p5	0	1	0	1	1	0
p6	0	0	0	1	1	1
p7	0	0	1	1	1	1
p8	0	1	1	1	0	1
p9	1	1	1	1	1	1
p10	0	0	1	0	0	0
p11	1	1	1	1	1	0
p12	0	1	1	1	0	1
p13	1	0	0	0	1	0
p14	1	1	1	1	0	0
p15	0	0	0	1	1	1
p16	1	0	1	1	1	0
p17	1	0	1	0	1	0
p18	0	1	0	1	1	1
p19	1	0	1	0	1	0
p20	0	1	0	1	1	1

Tabla 3.6: *Submatriz de la matriz de datos.*

a_i	a_j	n_{a_i}	$n_{\overline{a}_j}$	p	$K_0 = Card(A_i \cap \overline{A}_j)$
BELI	DRAM	9	10	0.225	5
BELI	HIST	9	6	0.135	1
DRAM	ROMA	10	7	0.175	4
DRAM	MUSI	10	6	0.150	0
HIST	ROMA	14	7	0.245	7
TERR	MUSI	11	6	0.165	2

Tabla 3.7: *Valores de cardinales y probabilidad de la ley Binomial.*

Empleando la Definición 4 sobre estos datos y teniendo en cuenta que bajo el modelo Binomial, $P(K = K_0) = \binom{20}{K_0} p^{K_0}(1-p)^{20-K_0}$, obtendremos la intesidad implicativa de las reglas formadas por estos pares de variables. Para ello, formamos la Tabla 3.8 con las probabilidades $P(K = K_0)$ para valores de K_0 variando desde 0 hasta 7, y para los diferentes valores de la probabilidad de éxito (p) en el modelo Binomial.

	$BELI$, $DRAM$	$BELI$, $HIST$	$DRAM$, $ROMA$	$DRAM$, $MUSI$	$HIST$, $ROMA$	$TERR$, $MUSI$
p	0.225	0.135	0.175	0.150	0.245	0.165
K_0	$P(K = K_0)$					
0	0.006109899	0.054994866	0.021334328	0.038759531	0.003621921	0.027147479
1	0.035476834	0.171660276	0.090509271	0.136798345	0.023506510	0.107289439
2	0.097847398	0.254513646	0.182389895	0.229338402	0.072465434	0.201408617
3	0.170443855	0.238330697	0.232132593	0.242828896	0.141091640	0.238795846
4	0.210305724	0.158083512	0.209271050	0.182121672	0.194584994	0.200545613
5	0.195380802	0.078950378	0.142050652	0.102845180	0.202059119	0.126812076
6	0.141808647	0.030804338	0.075329891	0.045372873	0.163922133	0.062646684
7	0.082340504	0.009615227	0.031958136	0.016013955	0.106386550	0.024758570

Tabla 3.8: *Probabilidades* $P(K = K_0)$ *desde* $K_0 = 0$ *hasta* $K_0 = 7$.

Para calcular φ(BELI, DRAM) se debe restar a 1 la suma de las probabilidades que aparecen en la columna *BELI, DRAM* con $p = 0.225$, desde $K_0 = 0$ hasta $K_0 = 5$. Este resultado da $\varphi(\text{BELI, DRAM}) = 1 - P[K \leq 5] = 1 - (P[K = 0] + ... + P[K = 5]) = 0.284435487$. De forma análoga obtenemos las siguientes intensidades implicativas:

$$\varphi(BELI, HIST) = 1 - P[K \leq 1] = 0.773344858$$
$$\varphi(DRAM, ROMA) = 1 - P[K \leq 4] = 0.264362863$$
$$\varphi(DRAM, MUSI) = 1 - P[K \leq 0] = 0.961240469$$
$$\varphi(HIST, ROMA) = 1 - P[K \leq 7] = 0.092361697$$
$$\varphi(TERR, MUSI) = 1 - P[K \leq 2] = 0.664154465$$

Conforme a la Definición 3 y los valores obtenidos podemos asegurar que la regla $DRAM \rightarrow MUSI$ es admisible al 95 % de confiabilidad y el resto, que no llegan al 85 %, no se consideran admisibles.

	BELI	DRAM	HIST	MUSI	ROMA	TERR
BELI	00	28	77	12	39	09
DRAM	30	00	35	96	26	48
HIST	70	40	00	23	09	28
MUSI	20	88	23	00	56	64
ROMA	43	31	08	57	00	21
TERR	12	48	22	66	17	00

Tabla 3.9: *Intensidades de implicación calculadas por CHIC.*

CHIC muestra los índices de implicación entre variables en por ciento (Tabla 3.9), los índices de los ejemplos calculados previamente serían: 28, 77, 26, 96, 9 y 66 respectivamente. En la Figura 3.3 se muestra el grafo implicativo formado a partir de los valores de intensidad implicativa entre las variables de la Subtabla 3.6. En el grafo observamos sólo aquellas implicaciones mayores o igual al 50 % de intensidad, siendo la implicación en rojo la más sólida ($DRAM \rightarrow MUSI$ al 96 %).

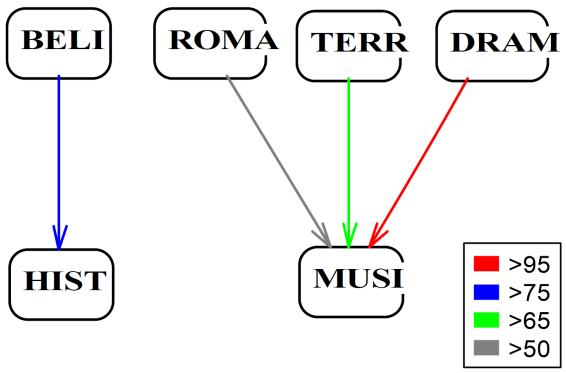

Figura 3.3: *Grafo Implicativo.*

Se sabe que en el caso de muestras grandes [1], el índice de intensidad implicativa no discrimina lo suficiente, de ahí que se estableciera un nuevo índice, de implicación-inclusión. No puntualizamos más en esta variación, puesto que el propósito de lo que se mide es idéntido.

3.5. Análisis Cohesitivo

La cohesión aparece como una medida de la calidad implicativa de la R-regla. El propósito es averiguar R-reglas $R' \to R''$ con una sólida relación implicativa entre los elementos de R' y los de R''. Por ejemplo, es inherente componer la regla $(a_1 \to a_2) \to (a_3 \to a_4)$ si las relaciones implicativas $a_1 \to a_3$, $a_1 \to a_4$, $a_2 \to a_3$ y $a_2 \to a_4$ son lo bastante significativas. Para medir esta calidad se plantea contrastar con el desorden de una experiencia aleatoria, para ello se propone la **Entropía**.

Se parte de considerar una regla $a_i \to a_j$ de orden 1 y de definir la variable aleatoria Y como indicativo del suceso $X_{a_i \wedge \overline{a}_j} \geq n_{a_i \wedge \overline{a}_j}$, luego:

$Pr[Y = 1] = Pr\lfloor X_{a_i \wedge \overline{a}_j} \geq n_{a_i \wedge \overline{a}_j} \rfloor = \varphi(a_i, a_j)$ y
$Pr[Y = 0] = 1 - \varphi(a_i, a_j)$.

Entonces se define la Entropía de este experimento como:

$$E = -p \log_2 p - (1-p) \log_2 (1-p), \text{ con } p = \varphi(a_i, a_j). \tag{3.6}$$

Dicho valor se puede interpretar como el valor medio de la incertidumbre de un observador previo a conocer la salida de una fuente binaria en la que puede o no acontecer el evento $a_i \to a_j$.

Se define el **grado de una R-regla** como la cantidad de variables comprendidas en la regla menos 1. Por ejemplo, la R-regla $R : a_i \to a_j$ es de orden 1, la R-regla $R' : R \to a_k$ es de orden 2, y así sucesivamente.

Definición 6 *La **Cohesión de una R-regla** $a_i \to a_j$ de grado 1 es:*

$$Coh(a_i, a_j) = \begin{cases} \sqrt{1 - E^2} & \text{si } p \geq 0,5 \\ 0 & \text{en caso contrario.} \end{cases} \tag{3.7}$$

Si $\varphi(a_i, a_j) = 0.5$, la regla $a_i \to a_j$ es imparcial, en vista de que la cantidad de contraejemplos es igual al esperado bajo la hipótesis de independencia. La entropía en este caso es máxima.

Definición 7 *La **Cohesión de la clase** de variables $R = \{a_1, ..., a_k\}$ es la media geométrica de las cohesiones de los pares de variables que la conforman [9]:*

$$Coh(R) = \left\{ \prod_{\substack{i,j \\ i<j}} Coh(a_i, a_j) \right\}^{\frac{2}{k(k-1)}} \tag{3.8}$$

La implicación estadística de una clase de variables sobre otra, se modeliza mediante el siguiente índice:

Definición 8 *La **intensidad de implicación** de una clase A sobre una clase B es:*

$$\Psi(A, B) = \left\{ \sup_{\substack{a_i \in A \\ a_j \in B}} \varphi(a_i, a_j) \right\}^{card(A) \times card(B)} \cdot [Coh(A) Coh(B)]^{\frac{1}{2}} \tag{3.9}$$

Seguidamente, se muestran los índices de cohesión al nivel cero (reglas de grado 1) de las variables presentadas en la Tabla 3.6 utilizando la fórmula de la Definición 6.

Nivel 0	BELI	DRAM	HIST	MUSI	ROMA	TERR
BELI	0.00	0.00	0.64[*1]	0.00	0.00	0.00
DRAM	0.00	0.00	0.00	**0.97**[*2]	0.00	0.00
HIST	0.48	0.00	0.00	0.00	0.00[*3]	0.00
MUSI	0.00	0.85	0.00	0.00	0.16	0.33
ROMA	0.00	0.00	0.00	0.16	0.00	0.00
TERR	0.00	0.00	0.00	0.39	0.00	0.00

Tabla 3.10: *Índices de cohesión al nivel cero de la jerarquía obtenidos con CHIC.*

[*1] Para el par de variables $(BELI, HIST)$ se tiene del análisis de la Intensidad Implicativa entre variables que $\varphi(BELI, HIST) = 0.773344858$, luego aplicando la Fórmula 3.6 obtenemos un valor de entropía de $E = 0.77213474$, por lo que $Coh(BELI, HIST) = 0.63545884 \approx 0.64$.

[*2] Para el par de variables $(DRAM, MUSI)$ se tiene del análisis de la Intensidad Implicativa entre variables que $\varphi(DRAM, MUSI) = 0.961240469$, luego aplicando la Fórmula 3.6 obtenemos un valor de entropía de $E = 0.23657549$, por lo que $Coh(DRAM, MUSI) = 0.97161311 \approx 0.97$.

[*3] Para el par de variables $(HIST, ROMA)$ se tiene del análisis de la Intensidad Implicativa entre variables que $\varphi(HIST, ROMA) = 0.092361697$, lo cual es menor que 0.5, por lo que según (3.7) $Coh(HIST, ROMA) = 0$.

Se contempla que la matriz de las cohesiones no es necesariamente simétrica, por ejemplo $Coh(ROMA, MUSI) = Coh(MUSI, ROMA) = 0.16$, pero, $Coh(DRAM, MUSI) = 0.97 \neq 0.85 = Coh(MUSI, DRAM)$.

Inmediatamente después de obtener las cohesiones al nivel cero, se unen las variables aisladas cuyas cohesiones sean máximas para comenzar el proceso de aglomeración. De la Tabla 3.10 se observa que el mayor índice se da entre las variables "DRAM y MUSI" con valor de 0.97161311, en consecuencia en el nivel 1 de la jerarquía se han unido. Para continuar, se deben determinar las intensidades implicativas de esta clase con las variables aisladas.

Nivel 1	(DRAM, MUSI)	BELI	HIST	ROMA	TERR
(DRAM, MUSI)	0.00	0.09[*4]	0.12	0.31	0.40
BELI	0.08	0.00	**0.64**	0.00	0.00
HIST	0.16	0.48	0.00	0.00	0.00
ROMA	0.32	0.00	0.00	0.00	0.00
TERR	0.43[*5]	0.00	0.00	0.00	0.00

Tabla 3.11: *Intensidades de implicación entre clases de variables al nivel 1 de la jerarquía.*

[*4] Para el par $((DRAM, MUSI), BELI)$ formado por una regla $R : DRAM \rightarrow MUSI$ y una variable $BELI$, se aplica primero la Fórmula (3.8) para determinar la cohesión de una clase

variables. Por tanto:

$$Coh(R) = \{Coh(DRAM, MUSI)\}^1 = 0.97161311$$

Debemos tener en cuenta que la cohesión de una variable es 1, $Coh(BELI) = 1$. Luego, aplicamos la Fórmula (3.9) para determinar el índice:

$$\begin{aligned}
&\Psi((DRAM, MUSI), BELI) = \\
&= \{\sup\{\varphi(DRAM, BELI), \varphi(MUSI, BELI)\}\}^{2\times 1} \cdot [Coh(DRAM, MUSI)Coh(BELI)]^{\frac{1}{2}} = \\
&= \{\sup\{0.3, 0.2\}\}^2 \cdot [0.97161311 \cdot 1]^{\frac{1}{2}} = 0.3^2 \cdot 0.97161311^{\frac{1}{2}} = 0.0887134 \approx 0.09.
\end{aligned}$$

*[5] Para el par $(TERR, (DRAM, MUSI))$ formado por una variable aislada $TERR$ y por la regla $R : DRAM \rightarrow MUSI$, se procede la misma forma: $Coh(TERR) = 1$ y $Coh(R) = \{Coh(DRAM, MUSI)\}^1 = 0.97161311$.

$$\begin{aligned}
&\Psi(TERR, (DRAM, MUSI)) = \\
&= \{\sup\{\varphi(TERR, DRAM), \varphi(TERR, MUSI)\}\}^{2\times 1} \cdot [Coh(TERR)Coh(DRAM, MUSI)]^{\frac{1}{2}} = \\
&= \{\sup\{0.48, 0.66\}\}^2 \cdot [0.97161311 \cdot 1]^{\frac{1}{2}} = 0.66^2 \cdot 0.97161311^{\frac{1}{2}} = 0.429373 \approx 0.43.
\end{aligned}$$

En consecuencia de los datos de la Tabla 3.11 se determina que al nivel 2 de la jerarquía se juntan las variables BELI e HIST con un índice de intensidad implicativa máximo e igual a 0.64.

Nivel 2	(DRAM, MUSI)	(BELI, HIST)	ROMA	TERR
(DRAM, MUSI)	0.00	0.01	0.31	0.40
(BELI, HIST)	0.02	0.00	0.12	0.06
ROMA	0.32	0.15	0.00	0.00
TERR	0.43	0.02	0.00	0.00

Tabla 3.12: *Intensidades de implicación entre clases de variables al nivel 2 de la jerarquía.*

El procedimiento de seguir buscando R-reglas consistentes se detiene al nivel tres debido a que ningún valor de intensidad implicativa es mayor a 0.5 y no se podrían sacar conclusiones firmes. El árbol jerárquico dirigido que se alcanza para este conjunto de variables se muestra en la Figura 3.4.

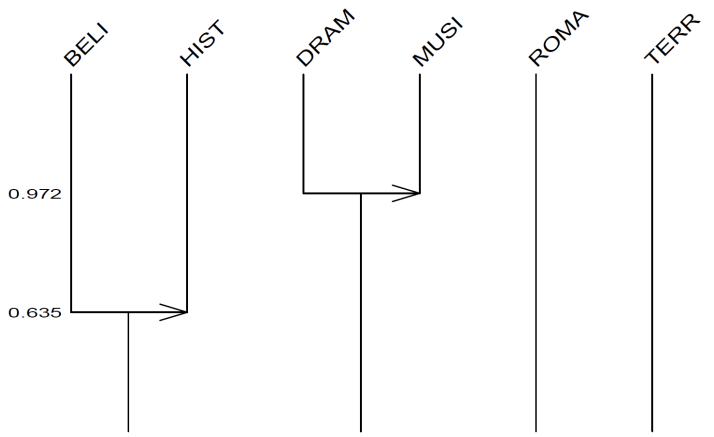

Figura 3.4: *Árbol jerárquico cohesitivo.*

Nodos significativos

Los nodos significativos del árbol jerárquico cohesitivo, construido mediante los índices de intensidad implicativa entre clases, son los correspondientes a una clasificación lo más acorde posible a los valores y la calidad de los valores de implicación y de cohesión [8].

El procedimiento empleado para el cálculo de los niveles y nodos significativos es idéntico al explicado en el apartado 3.3, pero esta vez empleando los índices de cohesión. Para la ejemplificación emplearemos el conjunto de variables $A^* = \{BELI, DRAM, HIST, MUSI\}$.

	BELI	DRAM	HIST	MUSI
BELI	0.00	0.00	0.64	0.00
DRAM	0.00	0.00	0.00	0.97
HIST	0.48	0.00	0.00	0.00
MUSI	0.00	0.85	0.00	0.00

Tabla 3.13: *Índices de cohesión de A^* obtenidos con CHIC.*

Los índices de cohesión de A^* al nivel cero de la jerarquía se muestran en la Tabla 3.13, y sobre ellos se forma el preorden inicial Ω:

$$\Omega = \left\{ \begin{bmatrix} Coh(BELI, DRAM) \\ Coh(BELI, MUSI) \\ Coh(DRAM, BELI) \\ Coh(DRAM, HIST) \\ Coh(HIST, DRAM) \\ Coh(HIST, MUSI) \\ Coh(MUSI, BELI) \\ Coh(MUSI, HIST) \end{bmatrix} < [Coh(HIST, BELI)] < [Coh(BELI, HIST)] < \right.$$

$$\left. < [Coh(MUSI, DRAM)] < [Coh(DRAM, MUSI)] \right\}.$$

Luego:

$$G_{Coh}(\Omega) = \left\{ \begin{array}{l} ((BELI, DRAM), (HIST, BELI)); ...((BELI, DRAM), (DRAM, MUSI)); \\ ((BELI, MUSI), (HIST, BELI)); ...((BELI, MUSI), (DRAM, MUSI)); ...; \\ ((MUSI, DRAM), (DRAM, MUSI)) \end{array} \right\},$$

$$Card(G_{Coh}(\Omega)) = \frac{m(m-1)}{2} - \sum_{l=1}^{d} \frac{Card(P_l)(Card(P_l)-1)}{2} = 66 - 28 = 38,$$

donde $t = Card(A^*) = 4$, $m = Card(\Omega) = 2\binom{t}{2} = 2\binom{4}{2} = 12$ y $d = 5$.

Nivel 1 de la jerarquía: se reúnen DRAM y MUSI (Figura 3.5) y obtenemos $R\Pi_1 = \{(DRAM, MUSI)\}$,

$$S\Pi_1 = \left\{ \begin{array}{l} (BELI, DRAM); (BELI, MUSI); (DRAM, BELI); (DRAM, HIST); \\ (HIST, DRAM); (HIST, MUSI); (MUSI, BELI); (MUSI, HIST); \\ (HIST, BELI); (BELI, HIST); (MUSI, DRAM) \end{array} \right\},$$

$r_1 = Card(R\Pi_1) = 1$, $s_1 = Card(S\Pi_1) = 11$, $I_1 = \{12\}$, $f_1^1 = 0$ y
$Card[G_{Coh}(\Omega) \cap [S\Pi_1 \times R\Pi_1] = \sum_{j=1}^{r_1} i_j - \frac{r_1(r_1+1)}{2} - \sum_{j=1}^{r_1} f_j^1 = 12 - 1 - 0 = 11.$

Nivel 2 de la jerarquía: a esta altura se reúnen BELI e HIST y por consiguiente $R\Pi_2 = \{(DRAM, MUSI), (BELI, HIST)\}$,

$$S\Pi_2 = \left\{ \begin{array}{l} (BELI, DRAM); (BELI, MUSI); (DRAM, BELI); (DRAM, HIST); \\ (HIST, DRAM); (HIST, MUSI); (MUSI, BELI); (MUSI, HIST); \\ (HIST, BELI); (MUSI, DRAM) \end{array} \right\},$$

$r_2 = Card(R\Pi_2) = 2$, $s_2 = Card(S\Pi_2) = 10$, $I_2 = \{12, 10\}$, $f_1^2 = 0$, $f_2^2 = 0$ y
$Card[G_{Coh}(\Omega) \cap [S\Pi_2 \times R\Pi_2] = 19.$

En la Tabla 3.14 se sintetizan los cardinales de los conjuntos reunidos y se muestran los valores obtenidos para el índice centrado y la función $v(\Omega, k)$.

Nivel	r_k	s_k	$Card[G_{Coh}(\Omega) \cap [S\Pi_k \times R\Pi_k]]$	$S(\Omega, k)$	$v(\Omega, k)$
1	1	11	11	1.59326	1.59326
2	2	10	19	1.93351	0.34025

Tabla 3.14: *Valores del índice centrado y de la función* $v(\Omega, k)$.

Tal como se especifica en la Definición 2 habría un solo nodo significativo, el 1, puesto que se trata de un máximo local de $v(\Omega, k)$ (Figura 3.6).

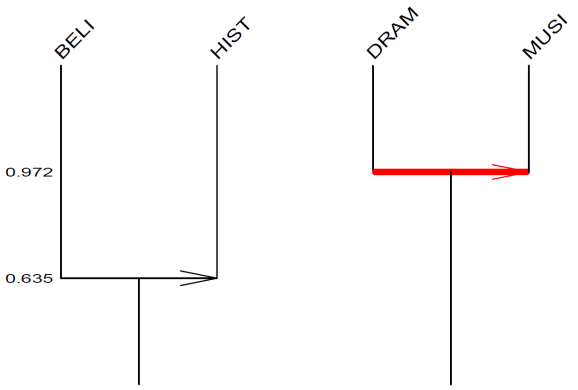

Figura 3.5: *Árbol cohesitivo.*

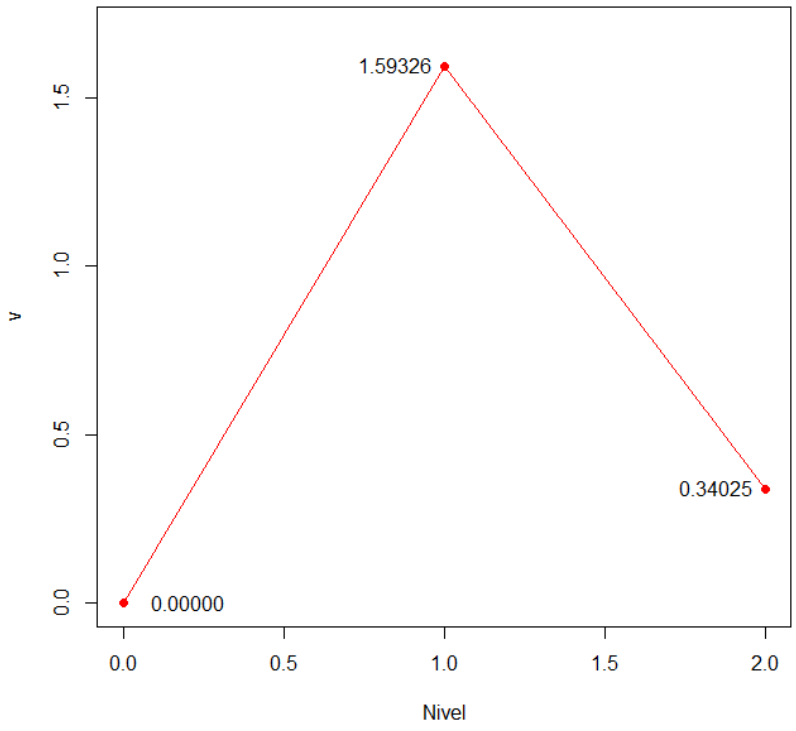

Figura 3.6: *Gráfica de* $v(\Omega, k)$.

3.6. Resultados

Una vez dados a conocer los fundamentos de la teoría del Análisis Estadístico Implicativo, vamos a aplicar las tres medidas presentadas sobre los datos de la investigación (Tabla 3.1) para obtener así grupos de variables y reglas de asociación sólidas, y ver las similitudes y disimilitudes entre los distintos métodos. Las figuras presentadas en este apartado se han extraído del programa informático CHIC al procesar los datos.

Análisis Clasificatorio

Siguiendo el mismo orden de la teoría presentada, empezaremos con el análisis clasificatorio. Con el fin de obtener un álbol de similaridad, CHIC calcula todos los índices de similaridad y a partir de éstos forma agrupaciones que organiza en forma de árbol. El resultado de estas operaciones se ve reflejado en la Figura 3.7.

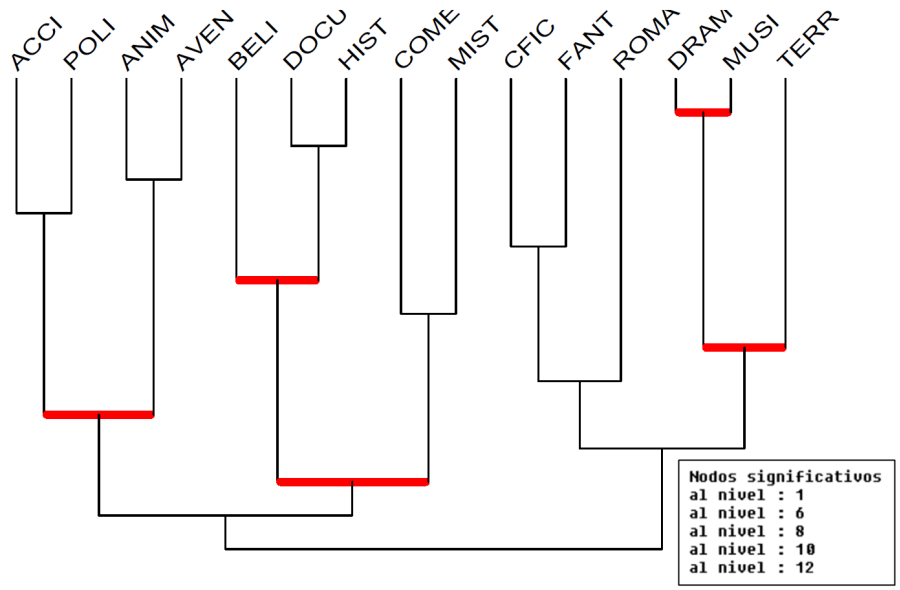

Figura 3.7: *Árbol de Similaridad.*

Tal y como se contempla, podemos diferenciar 3 clusters que contienen variables más semejantes entre sí:

$$C_1 = \{ACCI, POLI, ANIM, AVEN\},$$
$$C_2 = \{BELI, DOCU, HIST, COME, MIST\},$$
$$C_3 = \{CFIC, FANT, ROMA, DRAM, MUSI, TERR\}.$$

Sin embargo, se debe prestar más atención a los valores de los índices de similaridad y los niveles significativos que aparecen en rojo. Las variables reunidas hasta el nivel 5 de la clasificación (nivel donde se unen CFIC y FANT), tienen un índice superior al 70 % de semejanza y el nodo significativo hasta esa altura es la unión entre DRAM y MUSI. Así pues, la deducción que podemos sacar a partir de estas relaciones es que "*Normalmente, las personas a las que les gustan las películas dramáticas, suelen sentirse atraídas por los musicales y viceversa*". Al tratarse de un método de clasificación simétrico no podemos deducir una regla hacia un único

sentido, es en las siguientes metodologías donde sí podremos formar reglas orientadas, debido a que el ASI es un método no simétrico.

Análisis Implicativo

En la clasificación que hace el análisis implicativo destacan las reglas más significativas que se forman al calcular las intensidades implicativas entre variables. Con la finalidad de mostrar aquellas reglas más relevantes, CHIC forma un grafo implicativo (Figura 3.8) con 4 colores/niveles escogidos en función del valor de la intensidad de implicación, además se uiliza una flecha para simbolizar la implicación entre dos variables ($A \rightarrow B$). En este caso, al tratarse de una muestra pequeña, la intensidad implicativa es más que suficiente para discriminar las reglas más importantes de otras que no lo son.

La Figura 3.8 muestra un grafo implicativo obtenido en base a los datos de la investigación inicial. Según la leyenda, se muestran aquellas reglas cuya intensidad sea mayor al 50 % y están distribuidas de modo que: la flecha roja engloba a las implicaciones con una intensidad superior al 95 %, la azul abarca a las reglas con intensidad entre el 85 % y el 95 %, la verde a aquellas entre el 75 % y el 85 %, y la gris a las comprendidas entre el 50 % y el 75 %. En este grafo se pueden observar algunas propiedades interesantes, tomaremos por norma significativas a aquellas cuya intensidad sea mayor de 85 (rojas y azules), como es el caso de las reglas: DRAM → MUSI, POLI → ACCI, HIST → DOCU y ANIM → AVEN. En consecuencia podemos suponer las siguientes afirmaciones:

- "*Por lo general, los aficionados al género dramático lo son también al género musical*",
- "*Usualmente, a los simpatizantes del género cinematográfico policíaco les gusta ver acción en las películas*",
- "*Los partidiarios del género histórico disfrutan, habitualmente, de los documentales*",
- Por último, "*Quienes se divienten con las películas de animación, suelen inclinarse por los filmes de aventuras*".

A diferencia del análisis clasificatorio, la inversa de las implicaciones no tiene por que ser tan sólida, esto se debe a que el ASI toma en consideración el número de contraejemplos que que anulan la regla. Además, cabe remarcar que en el grafo no aparece la variable MIST, esto se debe a que el cardinal del conjunto $A_{MIST} = \{x \in I \mid MIST(x) = 1\}$ es igual al número de individuos en la muestra, $Card(A_{MIST}) = n = 20$ (ver Tabla 3.1).

Análisis Cohesitivo

Una vez calculadas las intensidades implicativas, el experto debe evaluar la calidad de estas implicaciones, para lo cual se aplica el análisis cohesitivo. CHIC computa la cohesión entre variables y clases y las intensidades de implicación sobre las clases para obtener un árbol jerárquico cohesitivo orientado (Figura 3.9).

Como contemplamos en el árbol cohesitivo (Figura 3.9), se han estructurado las 15 variables en distintas clases que establecen R-reglas a partir de la cohesión entre ellas. Además, marcado en rojo tenemos los nodos significativos que nos indican qué factores influyen más a la hora de formar las clases.

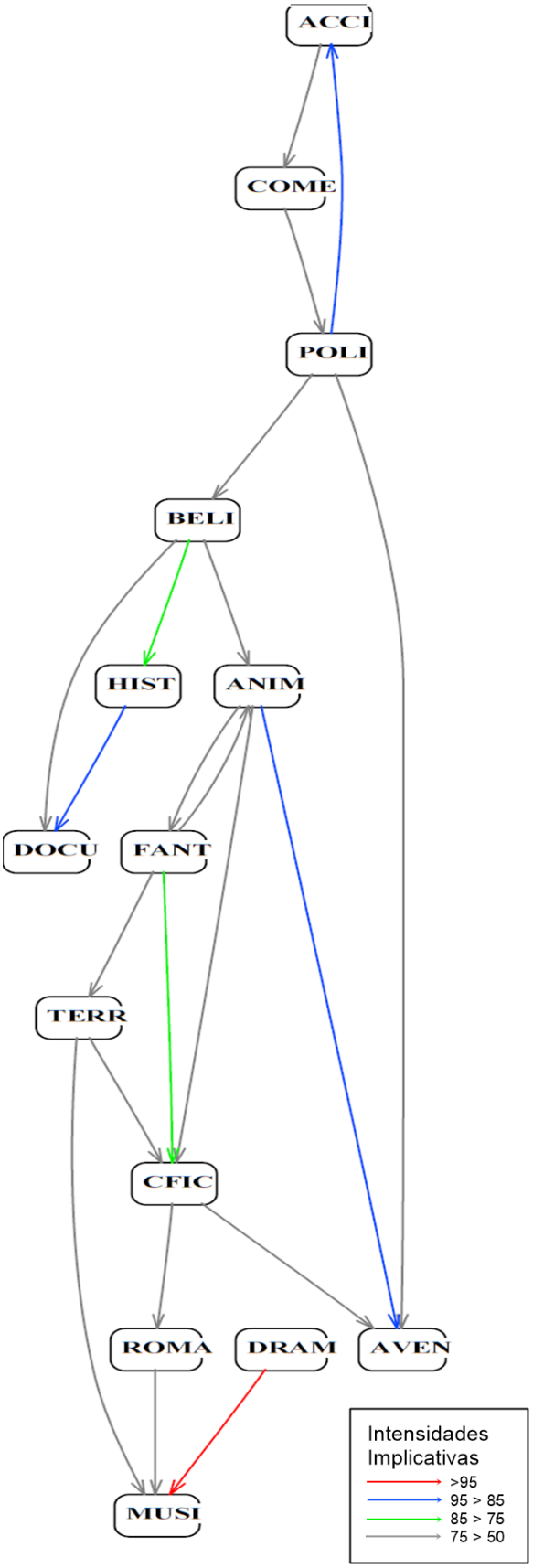

Figura 3.8: *Grafo Implicativo.*

Esta clasificación no simétrica de las clases nos acerca más a la realidad de los datos, sólo se han unido aquellas variables con una intensidad implicativa mayor a 0.5 y después se han organizado según la cohesión. Obtenemos 5 agrupaciones:

$$ANIM \to AVEN,$$
$$BELI \to (HIST \to DOCU),$$
$$COME \to (POLI \to ACCI),$$
$$DRAM \to MUSI \text{ y}$$
$$TERR \to (FANT \to CFIC).$$

Una vez más, se forman las clases ANIM → AVEN y DRAM → MUSI, cuyo sentido ya hemos comentado. Por otra parte, destaca la clase BELI → (HIST → DOCU) establecida a un nivel superior al 0.5 de cohesión y en rojo en el árbol, visto que se trata de un nodo significativo. El significado de la clase podría ser: "*Por lo general, la audiencia de las películas bélicas siente interés por los géneros histórico y documental*".

La diferencia más significativa en comparación con el análisis de similaridad es la construcción de las clases, el árbol de similaridad al final conduce a una única clase final, mientras que en el cohesitivo no.

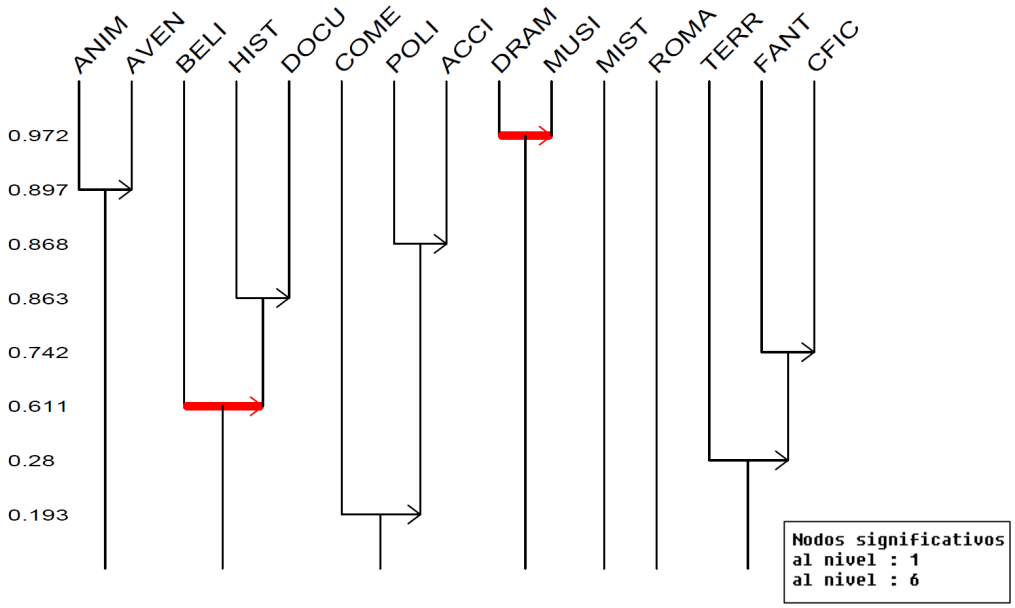

Figura 3.9: *Árbol jerárquico cohesitivo.*

3.7. Conclusiones

El presente trabajo conforma una recopilación y sumario de los conceptos fundamentales del Análisis Clasificatorio y del Análisis Estadístico Implicativo, así como su apoyo computacional en el programa informático CHIC. La teoría plasmada va seguida de una ejemplificación de los cálculos que deben realizarse, por lo que se utilizaron los datos de una pequeña investigación propia, con 20 individuos. Comenzando por el análisis de similaridad, éste permite revelar relaciones entre las variables con suficiente éxito. Sin embargo, la naturaleza no simétrica del ASI complementa favorablemente la información proporcionanda por la similaridad.

La visión general del desarrollo del análisis estadístico implicativo presenta cómo una teoría de proceso de datos se construye paso a paso en respuesta a las situaciones humanas, donde los teoremas (si a entonces b), en sentido racional del término, no pueden establecerse a causa de las excepciones que los niegan. La presente metodología nos permite descubrir cuasi-implicaciones entre atributos y clases de atributos, mostrando estructuras llamativas debido a su perspectiva no simétrica, fundamento principal de la teoría. Sus correspondientes funciones, reveladoras y analizadoras, parecen llevar a cabo las operaciones exitósamente en diversos ámbitos aplicativos.

Bibliografía

[1] Bodin, A. (1997). Analyse implicative : modèles sous-jacents à l'analyse implicative et outils complémentaires. Prépublication IRMAR. No. 97-32, 1997.

[2] Couturier, R. (2008). CHIC: Cohesive Hierarchical Implicative Classification, En Statistical Implicative Analysis, Studies in Computational Intelligence, vol. 127, p.41–52.

[3] Diday, E. (1971). La méthode des nuées dynamiques. Revue de statistique appliquée 19(2), 19–34.

[4] Diday, E. (1972). Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes, Thèse d'Etat, Université de Paris VI, 1972.

[5] Gras, R. (1979). Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Thèse d'Etat, Université de Rennes I, 1979.

[6] Gras R. (1999). Les fondements de l'analyse statistique implicative. Quaderni di Ricerca in Didattica, Vol. 9.

[7] Gras, R.; Ag Almouloud, S.; Bailleul, M.; Larher, A.; Polo M.; Ratsimba-Rajohn H.; Totohasina, A. (1996). L'implication Statistique, Collection Associée à Recherches en Didactique des Mathématiques", La Pensée Sauvage, Grenoble, 1996.

[8] Gras, R.; Kuntz P.; Régnier J.C. (2004). Significativité des niveaux d'une hiérarchie orientée. Classification et fouille de données, RNTI-C-1, Cépaduès- Editions, p 39-50, ISBN 2.85428.667.7.

[9] Gras, R.; Kuntz, P. (2005). Discovering R-rules with a directed hierarchy, Soft Computing, A Fusion of Foundations, Methodologies and Applications, Volume 1, p. 46-58, ISSN 1432-7643, Springer Verlag, 2005.

[10] Lerman, I.C. (1970). Sur l'analyse des données préalable à une classification automatique (proposition d'une nouvelle mesure de similarité). Mathématiques et Sciences Humaines, 32, p. 5-15.

[11] Lerman, I.-C. (1981). Classification et analyse ordinale des données, Dunod, Paris, 1981.

[12] Lerman, I.-C.; Gras, R.; Rostam H. (1981). Elaboration et évaluation d'un indice d'implication pour des données binaires, I et II, Mathématiques et Sciences Humaines, 1981, n. 74, 5-35 et n. 75, 5-47.

[13] Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of abilities, Psychological Monographs, 61, n. 4.