Universidad de Córdoba

Máster Universitario en Ingeniería Informática

Trabajo Fin de Máster

Algoritmo multi-criterio para la recomendación de asignaturas en los Grados Universitarios

Autora:
Aurora Esteban Toscano

Directores:
Prof. Dra. Amelia Zafra Gómez
Prof. Dr. Cristóbal Romero Morales

Córdoba, febrero de 2019

DECLARACIÓN DE AUTORÍA

Dña. Amelia Zafra Gómez y D. Cristóbal Romero Morales, profesores del Departamento de Informática y Análisis Numérico

INFORMAN:

Que el presente Trabajo Fin de Máster, titulado *"Algoritmo multi-criterio para la recomendación de asignaturas en los Grados Universitarios"*, ha sido desarrollado por Dña. Aurora Esteban Toscano, para aspirar al título de Máster Universitario en Ingeniería Informática, bajo nuestra dirección en el Departamento de Informática y Análisis Numérico de la Universidad de Córdoba, reuniendo, a nuestro juicio, las condiciones necesarias exigidas a este tipo de trabajos.

Y para que conste, se expide y firma el presente informe en Córdoba, a 12 de febrero de 2019.

Firmado:

Dña. Amelia Zafra Gómez D. Cristóbal Romero Morales

Mathematical science shows what is. It is the language of unseen relations between things. But to use and apply that language, we must be able fully to appreciate, to feel, to seize the unseen, the unconscious.

— Ada Lovelace

AGRADECIMIENTOS

El presente proyecto marca un punto de inflexión: el fin de una enseñanza más guiada que, a la vez, me ha hecho darme cuenta de que quiero seguir aprendiendo y ser útil a la sociedad por este medio. Un punto éste, al que no hubiera llegado sin la ayuda de muchas personas que me han rodeado en este poco más de año y medio que ha durado el Máster.

En primer lugar, he de agradecer a mis directores, Amelia y Cristóbal, por mostrarme su pasión hacia el mundo de la investigación y la enseñanza, por todo su trabajo para que yo pueda poco a poco abrirme camino en él y por su orientación y ayuda durante la realización de este trabajo.

También quisiera agradecer a todo el equipo de profesores que se han esforzado hasta el último día del Máster por transmitirme sus conocimientos e inspirarme a nunca querer dejar de aprender. Igualmente, agradecer a mis compañeros de clase la amistad y los buenos ratos brindados durante las tardes de clase.

Por último, muchas gracias a mis padres, Aurora y Paco, por todo su amor, entrega y apoyo. Sin ellos no hubiera podido llegar hasta aquí. A mis abuelos y mis hermanas, muchas gracias por acompañarme siempre y hacerme el camino más fácil. A Javi, por su cariño, por estar siempre conmigo y por su buen humor, muchas gracias porque sin ti este trabajo hubiera sido mucho más duro.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ALGORITMOS

ACRÓNIMOS

RS Sistema de Recomendación, *Recommendation System*

CF Filtrado Colaborativo, *Collaborative Filtering*

CBF Filtrado basado en Contenido, *Content-based Filtering*

EDM Minería de Datos Educativos, *Educational Data Mining*

GA Algoritmo Genético, *Genetic Algorithm*

SGEA Algoritmo Generacional Elitista Simple, *Simple Elitist Generational Algorithm*

CHC *Constrained Hill Climbing*

NNH Vecindario Más Cercano, *Nearest Neighborhood*

IR Recuperación de Información, *Information Retrieval*

nDCG Ganancia Acumulada Descontada Normalizada, *Normalized Discounted Cumulative Gain*

MAE Error Medio Absoluto, *Mean Absolute Error*

RMSE Raíz del Error Cuadrático Medio, *Root Mean Squared Error*

1

INTRODUCCIÓN

1.1 PLANTEAMIENTO

Vivimos en la era de la información, una era en la que el ritmo de uso de Internet crece imparablemente. Esto ha propiciado que se almacenen grandes cantidades de datos resultantes de la interacción de las personas con los diferentes portales web. En este contexto de aumento exponencial de datos, las técnicas tradicionales que desde hace siglos el ser humano ha ido mejorando para la extracción de conocimiento comienzan a quedarse pequeñas. Así, surgen técnicas de análisis automático y minería enfocadas a grandes cantidades datos cuyo propósito es obtener tendencias a gran escala y, en definitiva, información de temática tan amplia como los datos con los que se trabaje: procesos industriales, diagnóstico médico, aplicación comercial... Entre estas técnicas, un campo de investigación ya asentado son los Sistemas de Recomendación, *Recommendation Systems* (RS), que abordan el problema del filtrado de información para ofrecer aquella que pueda ser más relevante para el usuario. En una sociedad cada vez más saturada de información, los RS se encuentran en un momento único, siendo una necesidad real para los usuarios a los que ahorran tiempo y estrés utilizando información de su perfil para recomendarles ítems en multitud de dominios: desde páginas webs, libros, vídeos, canciones, aplicaciones hasta material educativo [1].

Desde hace un par de décadas, los RS se han popularizado gracias a su utilidad en los portales de comercio electrónico, donde se utilizan para recomendar productos específicos en base a las interacciones anteriores del usuario en el portal. En 2001, en un informe de investigación fundamentado en congresos y presentaciones relacionadas, el analista Doug Laney del META Group (ahora Gartner) definía el crecimiento constante de datos como una oportunidad y un reto para investigar en el volumen,

la velocidad y la variedad[1]. Entre los pioneros en el uso de técnicas de RS destaca la multinacional de comercio electrónico Amazon, que en torno al año 2003 ya tenía un potente RS basado en ítem que utilizaba información de compras anteriores y del historial de navegación en la tienda [2]. En torno a 2010, un informe de Microsoft Research estimó que el 30% de las páginas que se veían en el portal provenían del RS. Otras importantes compañías como YouTube o Netflix también han confirmado utilizar RS [3]. En el caso de Netflix, su Director de Producto, Neil Hunt confirmó que el 80% de las películas que se veían en su plataforma provenían de la recomendación.

Los RS hacen uso de diferentes fuentes de información para proporcionar a los usuarios predicciones y recomendaciones de ítems según cómo se aborden los diferentes aspectos que se quieran potenciar en la recomendación: precisión, novedad, diversidad... La técnica de Filtrado Colaborativo, *Collaborative Filtering* (CF) ha sido la más importante en el desarrollo de los RS. Está basada en cómo los seres humanos han tomado decisiones a lo largo de la historia: además de en nuestras propias experiencias, también basamos nuestras decisiones en el conocimiento y la experiencia que nos llega de un grupo más o menos grande de conocidos. Así, CF se basa en las valoraciones que los usuarios dejan sobre los ítems. Otra importante técnica es el Filtrado basado en Contenido, *Content-based Filtering* (CBF), que hace recomendaciones en base ítems que gustaron al usuario en el pasado, por tanto requiere establecer una medida de similitud entre ítems. Otra técnica menos explorada es el Filtrado Demográfico, *Demographic Filtering*, basada en que los individuos con ciertas características en común (género, país, edad...) también tendrán preferencias en común. El avance en RS ha puesto en evidencia la importancia de las técnicas híbridas, así como el multi-criterio, que ayudan a compensar los defectos que por separato tiene cada técnica.

El auge de las nuevas tecnologías no es ajeno al mundo educativo: desde hace algunos años cada vez son más los recursos *e-learning*, el software educativo instrumental, la creación de bases de datos de información de estudiantes y el uso de Internet en la educación. Toda esta información proporciona un interesante campo de investigación para comprender cómo aprenden los estudiantes [4]. Surge así la Minería de Datos Educativos, *Educational Data Mining* (EDM), que se ocupa del desarrollo, la investigación y la aplicación de métodos computarizados para detectar patrones en grandes colecciones de datos educativos. Éste es un área que comienza a tomar importancia hace una década escasa y, por

1 Douglas, Laney. "3D Data Management: Controlling Data Volume, Velocity and Variety"

tanto, con mucho por descubrir aún. Dentro de EDM existen multitud de campos de estudio, entre los que se encuentra la recomendación aplicada a los recursos de aprendizaje de toda índole. Pero si uno destaca, es la recomendación enfocada a cursos o asignaturas en un marco de estudios universitarios [5].

Es común encontrar que los estudiantes universitarios de todo el mundo deben escoger entre una serie de asignaturas o cursos optativos para alcanzar el número de créditos necesarios para obtener su grado universitario. Ésta no es una decisión trivial para los estudiantes, que pasan una gran cantidad de tiempo buscando información sobre estas asignaturas. La inclusión de los RS en este proceso resulta natural en la medida en que tratan de modelar la tendencia natural de los estudiantes de pedir consejo a compañeros con gustos similares. Así, existen varias propuestas capaces de recomendar cursos a partir de las valoraciones previas de los estudiantes a otros cursos [6], [7] o partir de sus calificaciones [8], [9]. También existen algunas propuestas que exploran la inclusión de múltiples criterios para hacer las recomendaciones [10]. Sin embargo, no existen propuestas que permitan identificar realmente qué criterios son los más relevantes a la hora de hacer recomendaciones de asignaturas. Otro problema existente en este campo de investigación es que se trabaja con conjuntos de datos sensibles que no se suelen difundir, por tanto cada RS es adaptado a la información con la que va a trabajar y resulta complicado extrapolarlo a otros conjuntos de datos y por tanto aplicarlo a otras universidades.

En este contexto, en este trabajo se propone un sistema de recomendación de cursos que combine un sistema CF, con información del estudiante, como son sus valoraciones sobre las asignaturas, sus calificaciones y la especialidad escogida, con un sistema CBF, basado en la información de la asignatura, concretamente sus profesores, sus competencias, sus contenidos tanto teóricos como prácticos y su área de conocimiento. El RS propuesto será probado con datos reales de alumnos del Grado en Ingeniería Informática de esta Universidad. Así, un objetivo importante del trabajo será determinar la relevancia de cada criterio en el proceso de recomendación. En esta línea, se propone la utilización de un Algoritmo Genético, *Genetic Algorithm* (GA) que automáticamente optimice tanto los pesos asignados a cada criterio como otros parámetros utilizados en la configuración del RS como las métricas de similaridad usadas o el tamaño de vecindario utilizado en el sistema CF. Así, primeramente el GA es entrenado para producir la configuración optimizada del RS para a continuación construir el modelo RS que produzca las recomendaciones con la garantía de que se ha utilizado la mejor configuración posible.

1.2 OBJETIVOS

Este trabajo continúa la línea que ya se abordó en el Trabajo Fin de Grado [11] en el que se estudiaba el desarrollo de varios RS para la recomendación de asignaturas. Así, en este proyecto se profundizará sobre todo en las técnicas de evaluación del mismo y en métodos de optimización automáticos basados en GA, que permitirán determinar qué criterios son los más relevantes a la hora de realizar recomendaciones de asignaturas.

Más concretamente, los objetivos científicos que se persiguen son los siguientes:

1. Revisión de los modelos propuestos en el ámbito de la recomendación de asignaturas o cursos.

2. Ampliación de la información relacionada con la asignatura utilizada en el RS del que se parte, incluyendo el área de conocimiento, las competencias y la descripción de los contenidos teóricos y prácticos de las mismas.

3. Propuestas de mejoras de las técnicas de evaluación del RS, incluyendo muestreo estratificado de los datos y validación cruzada.

4. Propuesta de optimización de la configuración completa del RS mediante GA empezando con un Algoritmo Generacional Elitista Simple, *Simple Elitist Generational Algorithm* (SGEA) [12], explorando la búsqueda adaptativa con CHC [13] y abordando técnicas de búsqueda local con algoritmos meméticos con *Clearing* [14].

5. Análisis experimental de los resultados obtenidos por el GA a fin de determinar qué criterios se consideran relevantes en la recomendación.

6. Análisis experimental del rendimiento del RS propuesto frente a otras propuestas más simples que utilicen menos criterios.

1.3 ESTRUCTURA

El resto de este Trabajo Fin de Máster se estructura como sigue. En el capítulo 2 se hace una revisión de las técnicas más relevantes que ya existen en el ámbito de la recomendación de cursos. En el capítulo 3

se describen los conocimientos teóricos que este trabajo tiene de base. En el capítulo 4 se definen en detalle tanto el RS propuesto como los datos con los que se trabajará y las técnicas de aprendizaje automático utilizadas para su optimización. En el capítulo 5 se realiza un estudio de rendimiento de las diferentes técnicas de optimización utilizadas, se exponen los resultados experimentales obtenidos con el RS propuesto y se lleva a cabo una comparativa con otros modelos de recomendación. Por último, en el capítulo 6 se presentan las conclusiones obtenidas, las publicaciones asociadas a este trabajo y las líneas de investigación futuras propuestas.

2

TRABAJO RELACIONADO

En este capítulo se presenta una revisión estructurada de las propuestas que han estudiado los RS para la recomendación de asignaturas. Con esta revisión se ha pretendido profundizar en las técnicas de recomendación existentes y sus métodos de evaluación y sobre todo obtener una visión general de las técnicas aplicadas a la recomendación de cursos y los criterios utilizados en el campo.

Las fuentes que se han utilizado para obtener los trabajos son principalmente dos, de las que se han priorizado los trabajos más actuales (de 2012 en adelante):

- Web of Science [1]: plataforma basada en tecnología Web que recoge las referencias de las principales publicaciones científicas de cualquier disciplina del conocimiento desde 1945. Maneja los índices de citas más importantes. Su acceso es vía licencia FECYT, con la que la Universidad de Córdoba tiene convenio.
- Google Scholar [2]: basado en el algoritmo de PageRank, abarca más publicaciones, pero sus resultados está más desestructurados, tendiendo a mostrar más contenido de peor calidad.

Los términos que se han utilizado para realizar la búsqueda son:

- *Educational Data Mining*
- *Academic Advising System*
- *Course Recommendation*
- *Curriculum Mining*

1 https://www.fecyt.es/es/recurso/web-science
2 https://scholar.google.es/

2.1 RECOMENDACIÓN DE CURSOS

El uso de la minería de datos en la educación, también llamada EDM, se ha consolidado como un importante campo de investigación debido al aumento de los sistemas de información que apoyan los procesos de aprendizaje en el ámbito educativo. Romero y Ventura [4] presentan una revisión actualizada de los congresos, revistas y técnicas más actualizadas en este campo. Igualmente, un completo análisis de las técnicas más importantes en los últimos años en EDM se puede encontrar en la revisión de Peña-Ayala [15]. Paralelamente, los RS han emergido en los últimos años como una técnica muy útil para guiar a los usuarios en diferentes dominios en los que hay una gran cantidad de información disponible, como el comercio electrónico, plataformas de música o películas. En este campo la técnica más importante es el CF, basado en usuarios similares, seguido del CBF, basado en ítems similares [1]. No obstante, en los últimos años la importancia de las técnicas híbridas, que toman las ventajas de cada técnica, ha aumentado [16]. En este contexto, los RS se han aplicado exhaustivamente al problema de la recomendación de asignaturas o cursos desde diferentes enfoques. En [5] Iatrellis y col. presentan una revisión sistemática, desde una perspectiva experimental, de las técnicas más recientes de RS aplicadas a la recomendación de cursos, todo ello enmarcado en la disciplina de los *Academic Advising Systems*.

Entre el trabajo reciente en la recomendación de asignaturas, CF es aún una técnica ampliamente utilizada. P.C. Chang, C.H. Lin y M.H. Chen [17] presentan un CF basado en usuario en dos etapas, usando sistemas inmune artificiales para la predicción de las calificaciones de los estudiantes. Con el fin de abordar el problema de la cantidad de retroalimentación requerida de los estudiantes para producir recomendaciones, los autores segregan la población de estudiantes con información demográfica e introducen un mecanismo de control de calidad basado en el filtrado de cursos cuyos profesores tienen bajas valoraciones. K. Taha [18] presenta un sistema colaborativo basado en XML que recomienda a los estudiantes elegir asignaturas que fueron superadas con éxito por otros estudiantes con intereses y rendimiento académico similares. La clasificación de estudiantes se basa en características respecto a las asignaturas como memorización o programación entre otras. B. Bakhshinategh y col. [19] exploran la inclusión de un sistema normalizado que describa las competencias que cada asignatura proporciona y la posibilidad de que los estudiantes valoren en qué medida cada asignatura les ayudó a alcanzarlas. K. Ganeshan y X. Li [20] diseñan un RS basado en web que usa *K-medias* para determinar la similaridad entre estudiantes.

Otra gran categoría en los enfoques de RS es el CBF. En esta línea, L. Mostafa y col. [21] presentan un sistema de razonamiento basado en casos que hace las recomendaciones basándose en asociar características relativas a las asignaturas. El desarrollo de software guiado por ontología [22] también es explorado como sistema CBF, en este caso modelando varios aspectos relacionados con el plan de estudios con el fin de recomendar asignaturas que ayuden a los estudiantes a completar los créditos requeridos. En [23] se explora la aplicación del análisis semántico a la descripción de los cursos para proporcionar las recomendaciones.

Los RS híbridos que combinan varias técnicas de recomendación están tomando cada vez más importancia. Unelsrød [16] explora la combinación de CF y CBF mediante la generación de recomendaciones independientes que luego son presentadas conjuntamente. Ese estudio muestra la importancia de contar con un conjunto de datos relativamente grande y accesible para evaluar el RS y sugiere que el factor principal en las recomendaciones es el rendimiento académico de los estudiantes. O. Daramola y col. [24] presenta un sistema CBF híbrido que combina reglas de asociación con razonamiento basado en casos utilizando información relativa a la asignatura. A. Al-Badarenah y J. Alsakran [25] combinan CF con reglas de asociación con el fin de predecir el rendimiento del estudiant. Z. Gulzar y col. [26] exploran el uso de una ontología conjuntamente con consultas basadas en N-gramas.

La combinación de los RS con GA no ha sido muy explorada en el campo de la recomendación de cursos. De acuerdo a nuestro conocimiento, sólo se puede mencionar el trabajo en [27], que combina un GA con árboles de decisión para recomendar asignaturas basándose en las restricciones sobre créditos y el rendimiento académico de los estudiantes. Más allá de la recomendación de asignaturas, C.S. Hwang [28] utiliza un GA en la última etapa de un sistema CF para ponderar los criterios tenidos en cuenta en la recomendación de películas.

En este trabajo se presenta un RS híbrido que combina CF y CBF usando múltiples criterios relativos tanto al curriculum del estudiante como a la información de la asignatura, con especial énfasis en la parametrización de la importancia de cada criterio considerado. Además, se propone que toda la configuración relativa al RS sea ajustada automáticamente por medio de un GA adaptado que garantice una solución óptima para el conjunto de datos usado. Las principales contribuciones de esta propuesta en relación al trabajo relacionado se resumen en la tabla 1.

La tabla 1 muestra un resumen de las principales características de cada propuesta. Se considera tanto los criterios específicos al estudiante

como a la asignatura, así como las métricas de similaridad utilizadas, que son un elemento clave en los RS para encontrar los estudiantes y asignaturas más similares. Es relevante mencionar que la mayoría de las propuestas utilizan o información del estudiante o de la asignatura, pero sin combinar. También se puede ver que la mayoría de trabajos utilizan uno o dos criterios y una o dos métricas de similaridad. Así, la principal contribución de esta propuesta en relación con el trabajo relacionado sería que utiliza un número representativo de criterios, así como de métricas de similitud. Concretamente, se utilizan 7 criterios diferentes combinando información del estudiante y de la asignatura, así como siete métricas de similitud diferentes.

Tabla 1: Comparativa entre propuestas de recomendación de asignaturas en la literatura

Algoritmos	Criterios		Medidas
	Estudiante	Asignatura	
Biclustering con CF basado en XML [18]	• Valoraciones sobre competencias de cursos		• Similaridad del coseno
CF basado en usuario [19]	• Valoraciones sobre competencias de cursos		• Correlación Pearson
Clustering [20]	• Calificaciones medias • Información demográfica		• *K*-medias
AIS con clustering [17]	• Calificaciones	• Profesores	• Similaridad del coseno • Correlación Pearson
Razonamiento basado en casos [21]		• Palabras clave de cursos	• Similaridad del coseno
Agente basado en ontología [22]		• Departamento • Créditos	• Relaciones de ontología
Clustering [23]		• Descripción de cursos	• Análisis semántico • Doc2vec
CF basado en usuario y CF basado en ítem [16]	• Valoraciones sobre cursos	• Área de curso • Profesores	• Similaridad del coseno • Correlación Pearson
Reglas de asociación y razonamiento basado en casos [24]	• Calificaciones	• Requisitos de curso	• Soporte de los nuevos casos frente a los antiguos
Clustering y reglas de asociación [25]	• Calificaciones		• Distancia Manhattan • Soporte de las reglas y confianza
Expansión de consultas N-gramas y ontología [26]		• Palabras clave de cursos • Base de datos de sinónimos	• Frecuencia de términos • Relaciones de ontología
RS híbrido multi-criterio propuesto	• Valoraciones sobre cursos • Calificaciones • Especialidad de estudiantes	• Profesores • Contenidos de cursos • Área de conocimiento • Competencias	• Distancia euclídea • Distancia Manhattan • Correlación Pearson • Correl. Spearman • Índice Jaccard • Función de ver. log. • Similaridad semántica

3

ANTECEDENTES

En este capítulo se explican los conocimientos teóricos necesarios para abordar el diseño de la del RS propuesto. Para ello se introducen al principio los fundamentos de los RS y sus métodos de evaluación y a continuación las técnicas utilizadas relativas a GA.

3.1 SISTEMAS DE RECOMENDACIÓN

Los RS son una técnica de filtrado de información que trata de recomendar ítems (libros, películas, etc.) que son probables de ajustarse a los gustos de un usuario. La principal motivación para desarrollar un RS es la creciente sobrecarga de información en la sociedad actual. Un RS puede usarse para filtrar cualquier tipo de información y con cualquier clase de preferencia: los más conocidos quizá sean asociados a comercio electrónico, como el motor de recomendación del portal Amazon, aunque podrían ser también desde destinos turísticos a artículos de investigación. O, como se discute en este trabajo, recomendación de asignaturas universitarias a estudiantes. Hay varias funciones que el RS debe desempeñar adecuadamente para satisfacer a todos los usuarios. El objetivo principal es encontrar ítems que, con un alto valor de probabilidad, satisfagan los gustos de los usuarios. Pero también es importante darle a esa recomendación un contexto que permita a los usuarios entender con qué criterios esos ítems fueron seleccionados para ellos. También es importante tener en cuenta las variaciones subjetivas que cada usuario tendrá a la hora de valorar objetos, que producirán desviaciones a la alza o a la baja en los historiales de valoración de éstos.

Así, un RS debe ser capaz de descifrar qué características gustan a un usuario y a su vez categorizar cada ítem de la colección en base a dichas características. Hay dos grandes aproximaciones para resolver

este problema: CF y CBF [1]. En general, CF empareja a usuarios con un historial de valoraciones similar, asumiendo que si compartieron gustos similares en el pasado, lo volverán a hacer en el futuro. Por su parte, CBF se basa en construir un perfil del usuario y entonces recomendarle aquellos ítems que más se ajusten a dicho perfil. Ambas aproximaciones serán abordadas con más detalle en las próximas secciones. Por último se estudiarán las diferentes alternativas a la hora de hibridar distintas técnicas de recomendación.

3.1.1 *Filtrado Colaborativo*

CF es una de las aproximaciones más importantes y clásicas a los RS. Se basa en la evaluación de la información usando la opinión de los usuarios. Por lo general, las predicciones sobre los intereses del usuario se generan recolectando las opiniones de otros usuarios; por tanto, debe procesar grandes cantidades de información. En la última década el CF ha ido mejorando continuamente, convirtiéndose en una de las técnicas más notorias en el campo de los RS.

El enfoque no probabilístico más popular del CF es el basado en vecindario de usuarios [29]. Para esta técnica, dado un usuario activo a con un cierto historial de valoraciones, se buscarán los n usuarios que hayan valorado más parecido los ítems valorados por a. Esta medida de similitud entre usuarios puede abordarse mediante diferentes métricas basadas en distancia o correlación. A dicho usuario a se le recomendarán aquellos ítems no vistos por él que tengan valoraciones más altas en su vecindario. La estimación de la preferencia que a tendría por un cierto ítem i, $p_{a,i}$, vendrá dada como una media de las valoraciones que el vecindario ha hecho sobre i, ponderada con el valor de similitud que tengan con a.

$$p_{a,i} = \frac{\sum_{j=1}^{n} w(a,j) v_{j,i}}{\sum_{j=1}^{n} w(a,j)} \tag{1}$$

Donde $w(a,j)$ refleja un valor de similitud, que típicamente estará en el intervalo $[0,1]$, entre el usuario activo a y cada usuario del vecindario j.

3.1.2 *Filtrado basado en Contenido*

CBF es la otra gran categoría entre los enfoques de RS. Éste se basa en encontrar similitud entre los ítems a recomendar analizando sus

características. El objetivo es encontrar ítems que tengan características similares a aquellos por los que el usuario ha mostrado preferencia. Una diferencia importante respecto a CF es que esta técnica permite recomendar ítems que aún no han recibido ninguna valoración.

Un paso importante en CBF es crear un buen sistema de filtrado de contenido con características relevantes para el dominio que se trate de recomendar [30]. Una forma común de abordar esto es construir un perfil de los gustos del usuario activo a a partir de los ítems que ha valorado positivamente. A continuación se le recomendarán aquellos ítems no vistos que más se asemejen a los anteriores. Así, la preferencia por un ítem desconocido i, $p_{a,i}$ vendría dada como una media de las valoraciones del usuario sobre cada ítem visto j, ponderada con los valores de similitud que éstas tengan con i.

$$p_{a,i} = \frac{\sum_{j=1}^{n} w(i,j) v_{a,j}}{\sum_{j=1}^{n} w(i,j)} \tag{2}$$

Donde $w(i,j)$ refleja un valor de similitud, que típicamente estará en el intervalo $[0,1]$, entre el par de ítems i,j.

3.1.3 *Sistemas de Recomendación Híbridos*

Una forma eficaz de compensar las debilidades que CF y CBF tienen por separado es combinar ambas técnicas en un RS híbrido. En [31], Burke desglosa la siguiente clasificación de RS híbridos atendiendo a cómo se combinan las técnicas que aúna:

Ponderación	La estimación de un ítem para un usuario es calculada combinando las salidas ponderadas de varios RS independientes.
Intercambio	Depende de las condiciones, uno de los diferentes RS es seleccionado para generar la recomendación.
Combinado	Los diferentes RS generan las recomendaciones independientemente, las cuales son presentadas todas juntas.
Combinación de características	Usa información colaborativa como una característica adicional de los ítems en una aproximación basada en contenido.

Cascada	Una de las técnicas de recomendación filtra un conjunto de ítems candidatos que son refinados mediante una segunda técnica antes de ofrecer las recomendaciones al usuario.
Aumento de características	Usa una de las técnicas para crear características o valoraciones adicionales que pueden ser usadas por una segunda técnica para producir recomendaciones.
Meta-nivel	Uno de los RS construye un modelo que se usa como entrada a la segunda técnica, que producirá las recomendaciones.

3.2 EVALUACIÓN DE LOS SISTEMAS DE RECOMENDACIÓN

Evaluar el rendimiento de los RS no es trivial. En primer lugar, porque los objetivos del RS pueden ser diversos [32]: algunos sistemas se centran en encontrar todas las sugerencias viables, mientras que otros encuentran menos pero mejores resultados. Otros factores a tener en cuenta son el orden de las recomendaciones o el posible impacto que puedan tener éstas (si son para grandes decisiones o triviales).

Otro factor importante a la hora de evaluar el RS es la elección del conjunto de datos. Existe la posibilidad de generar los datos sintéticamente, pero las investigaciones muestran que este tipo de conjuntos de datos tienden a dar resultados poco realistas [32]. Esto se debe a que los investigadores tienden an sesgar los datos para que encajen mejor con el algoritmo que está siendo probado. Por tanto, suele ser una elección más fiable probar el modelo con un conjunto de datos real.

En cualquier caso, las métricas de evaluación pueden dar una primera idea de cómo de bueno es el modelo del RS. Existen dos aproximaciones: medición de la desviación entre las valoraciones reales y las estimadas, y atendiendo a la relevancia que tienen para el usuario las recomendaciones hechas. Ambos enfoques serán explicados con más detalle en las siguientes secciones.

3.2.1 *Evaluación basada en desviación*

Para evaluar la desviación entre las valoraciones reales y las estimadas por el RS, basta con entrenar el RS ocultando las valoraciones que se van a estimar y a continuación preguntar explícitamente por ellas [29]. Con las valoraciones reales y las estimadas se pueden calcular diversas métricas basadas en el error de la desviación:

- Error Medio Absoluto, *Mean Absolute Error* (MAE). Es la métrica más simple e intuitiva, basada en diferencias absolutas:

$$MAE = \frac{\sum_{(i,j)\in K} |p_{i,j} - v_{i,j}|}{\#K} \tag{3}$$

  Donde

$p_{i,j}$	es la estimación para un usuario i del ítem j.
$v_{i,j}$	es la valoración real del usuario i para el ítem j.
$K = \{(i,j)\}$	es el conjunto de valoraciones estudiante-curso a predecir.

- Raíz del Error Cuadrático Medio, *Root Mean Squared Error* (RMSE). Es más sofisticado que el MAE, buscando la penalización de los errores más grandes:

$$RMSE = \sqrt{\frac{\sum_{(i,j)\in K} (p_{i,j} - v_{i,j})^2}{\#K}} \tag{4}$$

  Donde

$p_{i,j}$	es la estimación para un usuario i del ítem j.
$v_{i,j}$	es la valoración real del usuario i para el ítem j.
$K = \{(i,j)\}$	es el conjunto de valoraciones estudiante-curso a predecir.

3.2.2 *Evaluación basada en relevancia*

Mediante técnicas de Recuperación de Información, *Information Retrieval* (IR), se puede evaluar cómo de efectivo es un RS efectuando recomendaciones relevantes para el usuario. Para ello se asume que el proceso de predicción es binario: o el elemento recomendado agrada al usuario o no. Será caso de estudio a partir de qué umbral dentro del rango de

posibles valoraciones se considera que una valoración es positiva. El proceso de evaluación en este caso consistiría en ocultar las n valoraciones más positivas de un usuario, entrenar el modelo con el resto de datos y pedir al RS que obtenga las recomendaciones para dicho usuario. Con la lista ordenada de los ítems más relevantes para un usuario y aquellos estimados se pueden computar medidas como las siguientes:

- *Precisión*. Se define como el ratio de ítems relevantes frente al total de recomendados:

$$Precision = \frac{R \cap P}{P} \tag{5}$$

  Donde

  R es el conjunto de ítems relevantes para el usuario.
  P es el conjunto de ítems recomendados para el usuario.

- *Recall*. Se define como la proporción de ítems relevantes que son recomendados frente al total de ítems recomendados:

$$Recall = \frac{R \cap P}{R} \tag{6}$$

  Donde

  R es el conjunto de ítems relevantes para el usuario.
  P es el conjunto de ítems recomendados para el usuario.

- *Métrica F1*. Considera conjuntamente las métricas *precision* y *recall*:

$$F1 = \frac{2Recall \cdot Precision}{Recall + Precision} \tag{7}$$

- *Fall-out*. Es la proporción de ítems no relevantes que son recomendados frente al total de ítems no relevantes para el usuario:

$$Fall - out = \frac{N \cap R}{N} \tag{8}$$

  Donde

  N es el conjunto de ítems no relevantes para el usuario.
  R es el conjunto de ítems relevantes para el usuario.

- Ganancia Acumulada Descontada Normalizada, *Normalized Discounted Cumulative Gain* (nDCG). Relacionada con el orden en que aparecen las recomendaciones, busca penalizar, proporcionalmente a la posición del resultado, aquellos ítems relevantes para el usuario que aparezcan al final de la recomendación. Esta métrica

se define acumulada para una posición particular del ranking p como:

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{9}$$

Donde

DCG_p es la ganancia descontada acumulada en la posición p.
$IDCG_p$ es la DCG ideal en la posición p.

La DCG para hasta la posición p se define como:

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{log_2(i+1)} \tag{10}$$

Donde rel_i es la relevancia graduada del resultado en la posición i.

La normalización, utilizada para poder comparar consistentemente rankings de distinta longitud, viene dada por la división del DCG Ideal en la posición p. Ésta se calcula ordenando todos los ítems relevantes por su relevancia relativa, produciendo el máximo DCG posible a través de la posición p:

$$IDCG_p = \sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{log_2(i+1)} \tag{11}$$

Donde $|REL|$ es la lista de ítems relevantes (ordenados por su relevancia) en el corpus hasta la posición p.

- *Alcance*. CF se basa en la similaridad entre estudiantes. Dependiendo de la información que se utilice para calcular esta similaridad, para algunos estudiantes *outliers* podría no haber valores de similaridad satisfactorios, con lo que no se les podría hacer recomendaciones. Este comportamiento se puede medir con el alcance o cobertura del RS, cuyo propósito es maximizarse y viene dado por la ecuación:

$$Alcance = \frac{\#K - \sum_{(i,j)\in K} p_{i,j}}{\#K} \forall p_{i,j} = \emptyset \tag{12}$$

Donde

$p_{i,j}$ es la estimación para un usuario i del ítem j.
$K = \{(i,j)\}$ es el conjunto de valoraciones estudiante-curso a predecir.

3.3 ANÁLISIS DE DOCUMENTOS

En el campo de la IR, el análisis e indexado de documentos es una técnica muy importante para tratar la sobrecarga de información, pues permite procesar automáticamente grandes textos para sintetizarlos en términos clave sobre los que se pueden realizar operaciones como búsqueda o cálculo de similitud entre documentos [33]. Así, dado un cierto documento de texto se tratará en dos fases:

1. Análisis léxico: transforma el documento a un conjunto de palabras candidatas a ser términos clave. Para ello se deberán eliminar aquellas palabras que no aporten significado (dependerá del contexto) y aplicar *steaming*, es decir, reducir las palabras a su raíz morfológica para evitar duplicidad de términos.

2. Indexación de términos: procesar las palabras y transformarlas en una estructura de términos que permitan luego aplicarle las operaciones previstas. El agente encargado de esta tarea es el Sistema de Recuperación de Información, que atendiendo a qué información guarde por documento podrá ser de naturaleza booleana, difusa, probabilista, etc. Aunque en el que se va a centrar este trabajo es el vectorial, que para cada documento construirá un vector documental con sus descriptores y un coeficiente que represente en qué grado ese descriptor, o término clave, está presente en el documento.

Las aproximaciones típica para calcular la frecuencia de los términos en el documento es mediante TF-IDF, que suelen usarse de manera combinada para mostrar tanto los términos que aparecen con frecuencias altas en documentos individuales, como aquellos que son poco comunes en la colección de documentos:

- TF. La frecuencia del término i en el documento j se define como:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{13}$$

Donde

$n_{i,j}$	es el número de ocurrencias de i en j.
$\sum_k n_{k,j}$	es la suma del número de ocurrencias de todos los términos en j.

- IDF. La frecuencia inversa de documento de un término i en un conjunto de documentos se define como:

$$IDF_i = log \frac{|D|}{|d : t_i \in d|} \tag{14}$$

Donde

$\|D\|$	es el número de documentos en el conjunto.
$\|d : t_i \in d\|$	es el número de documentos en el que aparece i.

3.4 ALGORITMOS GENÉTICOS

Los GA [12] son técnicas de búsqueda estocástica que guían a una población de soluciones utilizando los principios de la evolución y la genética natural. La amplia investigación que los respalda garantiza sus robustas propiedades y demuestra su potencial en un amplio espectro de problemas de optimización, incluida la selección de características y tareas de ponderación. Los GA se modelan tomando como base los principios de la evolución a través de la selección natural, empleando una población de individuos que se someten a la selección en presencia de operadores que inducen variaciones, como la mutación y el cruce. Se utiliza una función de aptitud, o *fitness*, para evaluar individuos, variando el éxito reproductivo con ella. Las técnicas que se van a abordar en este trabajo para la optimización automática del RS son las siguientes:

3.4.1 *Algoritmo Generacional Elitista Simple*

Este tipo de GA [12] se inicia con un conjunto de soluciones (representadas por cromosomas) llamado población, que será generado de forma aleatoria. En cada paso evolutivo (generación), las soluciones en la población actual se evalúan de acuerdo con un criterio de calidad predefinido, la función *fitness*. Las soluciones de una población serán utilizadas para formar una nueva población (próxima generación). Así, los individuos que producirán la próxima generación (los padres) se seleccionan de acuerdo con su estado: cuanto más adecuadas sean, más posibilidades tienen de reproducirse. Los individuos seleccionados serán cruzados entre sí para producir soluciones nuevas (descendencia). Adicionalmente, la próxima generación también puede verse modificada mediante mutaciones aleatoria. La parte elitista del GA se produce al seleccionar los individuos finales para la próxima generación: siempre se

conservará el mejor individuo encontrado hasta el momento. El proceso descrito se repetirá a lo largo de un alto número de generaciones, hasta que se encuentra una solución satisfactoria.

3.4.2 *Algoritmo CHC*

El algoritmo de *Constrained Hill Climbing* (CHC) [13] surge en 1991 como un GA no tradicional que combina una estrategia de selección conservadora que preserva los mejores individuos encontrados, con un operador de cruce altamente disruptivo, que produce descendencia lo más diferente posible de los padres. Así, se introducen algunas modificaciones en el algoritmo generacional clásico que permiten hablar de búsqueda adaptativa. Al igual que el algoritmo generacional, en este caso se parte de una población de individuos (representados por cromosomas) que se irán cruzando a través de las generaciones hasta que se encuentre una solución satisfactoria, que vendrá dada por la función *fitness*. Sin embargo, en CHC se introduce el factor de la distancia entre individuos, que debe superar un cierto umbral para permitir que éstos se crucen (prevención de incesto). El umbral se establecerá al inicio del algoritmo en un valor máximo. En cada generación se formarán parejas de individuos cuya distancia deberá superar ese umbral. Conforme la población vaya convergiendo, se espera que dejen de poder formarse parejas al umbral establecido, con lo que este descenderá en una unidad, adaptándose así a esta convergencia. Llegará un punto en que el umbral llegue a 0, entendiéndose que la población se ha estancado. A continuación el algoritmo aplicará el proceso divergente: conservará las mejores soluciones encontradas (los mejores individuos), reiniciará el resto de la población de forma aleatoria y volverá a establecer el umbral al máximo inicial, iniciándose de nuevo el proceso intensificador, ya con los mejores individuos en la población.

3.4.3 *Algoritmo Clearing*

Un SGEA es apto para encontrar el óptimo de una función unimodal en un espacio de búsqueda limitado, aunque está demostrado que esta capacidad falla cuando se trabaja con múltiples óptimos en funciones multimodales [34]. Esta limitación puede ser abordada mediante mecanismos que creen y mantengan varias subpoblaciones dentro del espacio de búsqueda de manera que cada máximo de la función multimodal

pueda atraer cada uno de los nichos [35]. Estas técnicas son conocidas como métodos meméticos. Un procedimiento ya asentado en este campo es el algoritmo *Clearing* [14]. Así, el procedimiento de *Clearing* para determinar qué individuos pertenecen a una misma subpoblación (nicho), en un GA se aplicará después de evaluar el *fitness* de una población y antes de aplicar el operador de selección, que se asume que será estocástico (guiado por el *fitness* de los individuos). Para ello utilizará una métrica de distancia entre individuos. Cada nicho contendrá un individuo dominante: aquel con mejor *fitness*. Para que otro individuo pertenezca al nicho dado, su distancia con el dominante debe ser inferior a un umbral dado (radio de limpieza). Si la distancia es mayor, este individuo formará otro nicho al que se irán añadiendo los individuos más cercanos. Así, el procedimiento de *Clearing* conservará el *fitness* del individuo dominante de cada nicho, mientras que *limpiará* o restablecerá el del resto de individuos a cero (si el problema es a maximizar). De esta forma, la posterior selección de padres partirá de una población en los que los mejores *fitness* (y por tanto las probabilidades más alta de ser seleccionados) corresponde a individuos tan separados entre sí que corresponden a distintos nichos.

4

METODOLOGÍA PROPUESTA

En este capítulo se aborda en detalle la metodología propuesta, que constará de varios pasos como se muestra en la figura 1. En primer lugar se realiza la descripción y el preprocesado de los datos con los que se va a trabajar. A continuación se lleva a cabo la descripción del RS híbrido multi-criterio que se propone, enfocado en recomendar asignaturas a estudiantes universitarios basándose en varios criterios relativos tanto al estudiante como a la asignatura. Por último, debido al alto número de criterios considerados y al resto de configuraciones posibles para el RS, se abordarán las técnicas de optimización automática basadas en GA.

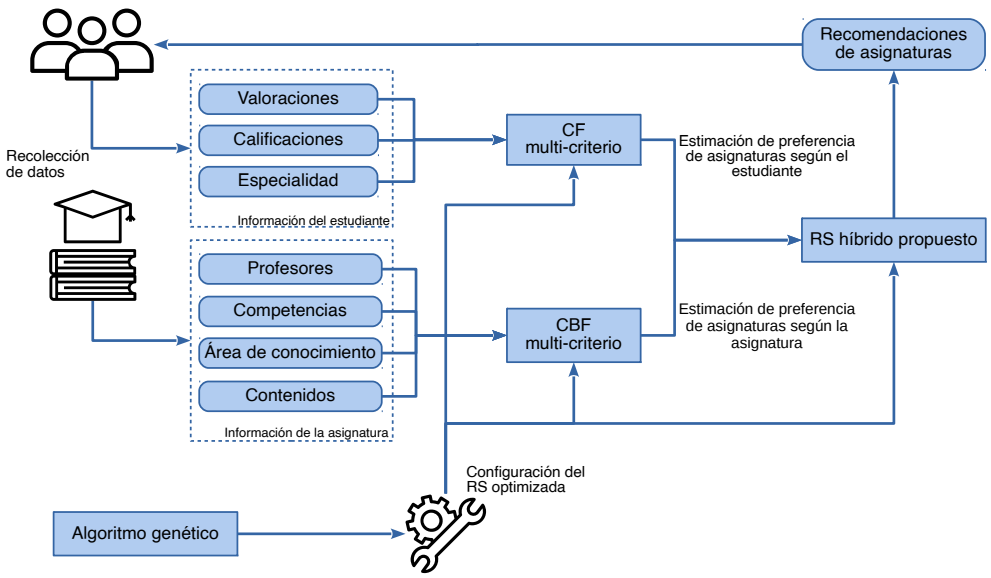


Figura 1: Pasos del RS híbrido multi-criterio propuesto

4.1 DESCRIPCIÓN Y PROCESADO DE LOS DATOS

Este trabajo ha sido desarrollado utilizando información real del Grado en Ingeniería Informática de la Universidad de Córdoba. Los datos obtenidos incluyen tanto información de los estudiantes como de las asignaturas.

4.1.1 *Información del estudiante*

La información del estudiante ha sido obtenida a través de encuestas realizadas durante tres cursos académicos (desde 2016 a 2018) a 95 estudiantes en cuarto curso de carrera. Así, se han conseguido 2500 valoraciones relativas a 63 asignaturas del plan de estudios del Grado en Ingeniería Informática. La información con la que se ha trabajado se muestra en la figura 2:

- Una valoración sobre la satisfacción general del estudiante respecto a cada asignatura cursada, en escala de *Likert* de 5 puntos. A las asignaturas no cursadas por el estudiante se le asigna un valor vacío.
- La calificación obtenida por el estudiante en cada asignatura cursada. En caso de haberse examinado más de una vez, la última calificación obtenida. Será un valor decimal en el rango $[0, 10]$ si la asignatura ha sido cursada o un valor vacío si no lo ha sido.
- La especialidad, mención o rama que el estudiante decide tomar dentro del área de la informática. Concretamente, el plan de estudios del Grado en Ingeniería Informática ofrece tres menciones a escoger en el 3º curso del plan de estudios: Computación, Ingeniería del Software o Ingeniería de Computadores. La elección del estudiante se representa mediante un identificador numérico de 1 a 3.

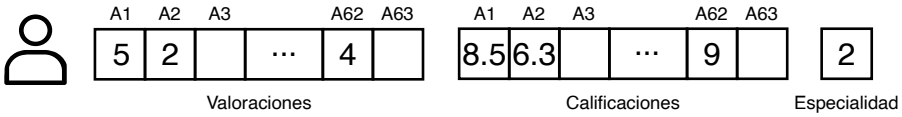


Figura 2: Información del estudiante

4.1.2 *Información de la asignatura*

La información de las 63 asignaturas consideradas ha sido obtenida de la web del Grado en Ingeniería Informática de la Universidad de Córdoba[1]. Los factores seleccionados para cada asignatura son representados en la figura 3:

- Los profesores que imparten docencia en cada asignatura, representados como un vector con un índice por cada profesor en el Grado. Si el profesor imparte la asignatura, se le asigna el valor 1 y si no lo está, 0.
- Las competencias que cada asignatura proporciona, representadas como un vector con un índice por cada competencia en el Grado. Si la competencia se relaciona con la asignatura, se le asigna el valor 1 y si no, se le asigna el valor 0.
- El área de conocimiento a la que la asignatura pertenece, representada mediante un identificador numérico. En el Grado considerado están presentes 8 áreas de conocimiento por tanto tomará un valor de 1 a 8.
- Los contenidos teóricos y prácticos que se detallan en la guía académica de la asignatura. Se representan mediante un vector de frecuencias de los términos relevantes, obtenido mediante minería de textos.

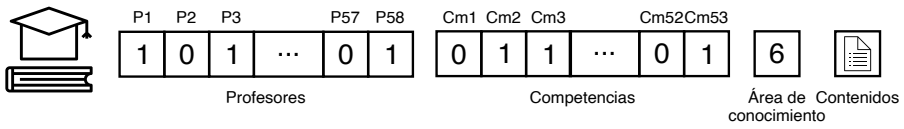


Figura 3: Información de la asignatura

4.2 SISTEMA DE RECOMENDACIÓN HÍBRIDO MULTI-CRITERIO

En esta sección se aborda en profundidad la propuesta de RS de este trabajo, cuyo propósito es recomendar asignaturas a estudiantes universitarios que les puedan ser relevantes para su curriculum. Para ello se utilizará información basada en múltiples criterios asociados tanto al estudiante como a la asignatura.

1 http://www.uco.es/eps/node/619

El RS propuesto es la combinación de dos subsistemas multi-criterio: un modelo basado en CF y otro en CBF. Así, la hibridación se implementa mediante la combinación lineal de las estimaciones de preferencia que estos subsistemas dan para cada estudiante y asignatura. El objetivo final del RS de recomendar asignaturas relevantes a estudiantes se alcanzaría seleccionando aquellas que hayan obtenido una estimación de la preferencia mayor. En el algoritmo 1 se detalla el proceso seguido.

Algoritmo 1: RS híbrido

```
inicio
    para cada estudiante i en el sistema hacer
        EstCF ← CF (i);
        EstCBF ← CBF (i);
        para cada asignatura k no valorada por i hacer
            Est[i, k] ← α · EstCF[k] + β · EstCBF[k];
        fin
        Recomendaciones ← N asignatura con Est[i] más altas;
    fin
fin
```

De acuerdo al algoritmo 1, los subsitemas CF y CBF proporcionan su estimación en el rango $[1, 5]$ y α y β son los parámetros que determinarán qué peso tiene cada subsistema en la estimación final. Por tanto, deben estar en el rango $[0, 1]$ y sumar 1. De este modo, las estimaciones finales para el estudiante i también estarán en el rango $[1, 5]$. En los siguientes apartados se aborda en detalle el diseño de los subsistemas CF y CBF.

4.2.1 *Filtrado Colaborativo*

El modelo CF encuentra relaciones entre estudiantes similares de forma que las recomendaciones se realizan en función de las asignaturas que más hayan gustado a los estudiantes más similares. El enfoque que se ha seguido para su desarrollo es el de Vecindario Más Cercano, *Nearest Neighborhood* (NNH). En este enfoque se construye para cada estudiante un vecindario con los estudiantes más similares para, a continuación, estimar la preferencia de una asignatura desconocida como la media de las valoraciones que recibió la asignatura en el vecindario, ponderada con los valores de similitud entre estudiantes. El aspecto multi-criterio se introduce en la forma de calcular la similaridad entre estudiantes, que determinará cómo de importante es la valoración de un estudiante sobre una asignatura. Así, la similaridad final entre estudiantes se calcula

como una combinación lineal de las similaridades individuales basadas en los criterios antes mencionados: las valoraciones que los estudiantes hacen sobre las asignaturas, las calificaciones obtenidas en cada una y la especialidad que cursan. El proceso se detalla en el algoritmo 2.

Algoritmo 2: Subsistema CF

```
inicio
    para cada estudiante i en el sistema hacer
        para cada estudiante j distinto de i hacer
            criterio1 ← simValoraciones (i,j);
            criterio2 ← simCalificaciones (i,j);
            criterio3 ← simEspecialidad (i,j);
            Similaridades[i, j] ← α · criterio1 + β · criterio2 + γ · criterio3;
        fin
        Nnh ← N estudiantes con Similaridades[i] más altas;
        para cada asignatura k no valorada por i hacer
            si en Nnh hay valoraciones para k entonces
                para cada estudiante j que ha valorado k hacer
                    Est[i, k] ←
                     Est[i, k] + Similaridades[i, j] · Valoraciones[j, k];
                fin
                Est[i, k] ← Est[i, k]/n;
            fin
        fin
    fin
fin
```

En el algoritmo 2 se puede ver que las funciones de similaridad entre estudiantes basadas en los diferentes criterios individuales proporcionan un valor en el rango $[0, 1]$ y que α, β y γ son los parámetros que determinarán qué peso tiene cada criterio en la similaridad final. Por tanto, deben estar en el rango $[0, 1]$ y sumar 1. De este modo, las similaridades finales entre estudiantes también estarán en el rango $[0, 1]$. Otro aspecto a configurar del algoritmo será el tamaño de vecindario utilizado para realizar las estimaciones.

Otra característica parametrizable de este modelo es qué métricas se utilizarán en el cálculo de las similitudes según los criterios considerados. Según la naturaleza de los datos, se pueden distinguir dos enfoques: las valoraciones y las calificaciones por un lado, y la especialidad por el otro. Dado que las valoraciones y las calificaciones son variables numéricas multidimensionales (tantas dimensiones como asignaturas) se les puede

aplicar métricas basadas en distancia (como la distancia euclidiana o la de Manhattan) o basadas en correlación lineal (como el coeficiente de correlación de Pearson o el de Spearman). Por otro lado, la especialidad, que es una única variable categórica, se mide atendiendo a si coincide o no entre los estudiantes. A continuación se detallan las medidas mencionadas.

- *Distancia euclidiana*. Se puede aplicar tanto a las valoraciones como a las calificaciones entre pares de estudiantes. Así, dados dos estudiantes X y Y, esta métrica se interpretará como una medida espacial n-dimensional donde X y Y son puntos colocados en las coordenadas que marquen sus valoraciones o calificaciones, como se ve en la ecuación 15.

$$d_E(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{15}$$

  Donde

n	es el número de asignaturas que X e Y tienen en común.
X_i, Y_i	son las valoraciones o calificaciones asociadas a una asignatura.

  Nótese que para ajustarse al estándar de similaridad $[0,1]$ deben subsanarse dos aspectos asociados a la interpretación de la métrica: por un lado a mayor asignaturas en común n, más posibilidades de que la distancia sea mayor, perjudicando una comparativa justa entre pares de usuarios. Por otro, menor distancia debe significar un valor más cercano a 1 en el valor final de similaridad. La regulación de estos dos aspectos se hará mediante la operación descrita en 16, acotándose sus resultados al intervalo $[0,1]$.

$$s = \frac{\sqrt{n}}{1 + d(X,Y)} \tag{16}$$

  Donde

n	es el número de asignaturas que X e Y tienen en común.
$d(X,Y)$	es el valor de distancia entre X e Y.

- *Distancia de Manhattan*. Se puede aplicar tanto a valoraciones como a calificaciones. Tendrá una interpretación análoga a la descrita en la distancia euclidiana, siendo en este caso su fórmula la descrita en la ecuación 17, a la que hay que aplicarle la misma regulación especificada en la ecuación 16.

$$d_M(X,Y) = \sum_{i=1}^{n} |X_i - Y_i| \tag{17}$$

Donde

n	es el número de asignaturas que X e Y tienen en común.
X_i, Y_i	son las valoraciones o calificaciones asociadas a una asignatura.

- *Coeficiente de correlación de Pearson*. Se puede aplicar tanto a valoraciones como a calificaciones. Así, dados dos estudiante X e Y, esta medida medirá la tendencia de sus valoraciones o calificaciones, emparejadas 1 a 1, a moverse en la misma dirección. Así, esta correlación se medirá como el producto de las covarianzas de sus asignaturas en común, normalizando el tamaño de los cambios mediante la división por las desviaciones típicas, como se muestra en la ecuación 18, cuyo resultado estará en el intervalo $[-1, 1]$.

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{18}$$

Donde

X, Y	son las series de valoraciones o calificaciones en común de los estudiantes.
cov	es la covarianza de las varianzas asociadas a X, Y.
σ_X, σ_Y	son las varianzas asociadas a X, Y.

- *Coeficiente de correlación de Spearman*. Se puede aplicar tanto a valoraciones como a calificaciones. Es una variación de la correlación de Pearson donde en vez de realizar la covarianza directamente sobre las dos series de valoraciones o calificaciones, a éstas se les asigna una posición en un ranking. Este proceso hace que se pierda información (cuánto gusta exactamente una asignatura), pero preserva el orden de las series y puede ser un interesante punto de estudio.

$$\rho(r_X, r_Y) = \frac{cov(r_X, r_Y)}{\sigma_{r_X} \sigma_{r_Y}} \tag{19}$$

Donde

r_X, r_Y	son los rankings asociados a las valoraciones o calificaciones en común de los estudiantes.
cov	es la covarianza de las varianzas asociadas a los rankings de X e Y.
σ_X, σ_Y	son las varianzas asociadas a los rankings de X e Y.

- *Similitud entre especialidades*. Se asume que cada estudiante sólo puede pertenecer a una especialidad, y que estas no tienen relación entre sí. Así, esta medida de similitud se limita a comprobar si un par de estudiantes comparten especialidad, devolviendo 1 si tienen la misma y 0 si es distinta.

4.2.2 *Filtrado basado en Contenido*

El modelo CBF encuentra relaciones entre asignaturas similares de forma que las recomendaciones se realizan en función de las asignaturas parecidas a las que más hayan gustado al estudiante en el pasado. El enfoque que se ha seguido para su desarrollo es el de *filtrado basado en ítem*. Para estimar el valor de preferencia que un estudiante tendría en una asignatura se realiza una media de las valoraciones que éste dio a asignaturas pasadas, ponderada con los valores de similitud entre la asignatura a estimar y las valoradas por él. El aspecto multi-criterio se introduce en la forma de calcular la similaridad entre asignaturas, que determinará cómo de importante es la valoración de un estudiante sobre una asignatura. Así, la similaridad final entre asignaturas se calcula como una combinación lineal de las similaridades individuales basadas en los criterios antes mencionados: profesores que imparten asignaturas, competencias, área de conocimiento y los contenidos, presentes en las guías docentes. El proceso se detalla en el algoritmo 3.

De acuerdo al algoritmo 3, las funciones de similaridad entre asignaturas basadas en los diferentes criterios individuales proporcionan un valor en el rango $[0, 1]$ y α, β, γ y δ son los parámetros que determinan qué peso tiene cada criterio en la similaridad final. Por tanto, deben estar en el rango $[0, 1]$ y sumar 1. De este modo, las similaridades finales entre asignaturas también estarán en el rango $[0, 1]$.

Otra característica parametrizable de este modelo es qué medidas de similaridad se usarán en el cálculo de las similitudes según los criterios considerados. Según la naturaleza de los datos se pueden distinguir tres enfoques: dado que los profesores y las competencias son variables booleanas multidimensionales (tantas dimensiones como profesores o competencias respectivamente) se les puede aplicar métricas basadas en teoría de conjuntos (como el índice Jaccard) o basadas en probabilidad (como el coeficiente de verosimilitud logarítmica). Por otro lado, el área de conocimiento, que es una única variable categórica, se mide atendiendo a si coincide o no entre dos asignaturas. Finalmente, la similaridad

Algoritmo 3: Subsistema CBF

inicio
  para cada *asignatura* k *en el sistema* **hacer**
    para cada *asignatura* l *distinta de* k **hacer**
      criterio1 ← `simProfesores` (k,l);
      criterio2 ← `simCompetencias` (k,l);
      criterio3 ← `simArea` (k,l);
      criterio4 ← `simContenidos` (k,l);
      Similaridades[k, l] ←
       $\alpha \cdot$ criterio1 + $\beta \cdot$ criterio2 + $\gamma \cdot$ criterio3 + $\delta \cdot$ criterio4;
    fin
  fin
  para cada *estudiante* i *en el sistema* **hacer**
    para cada *asignatura* k *no valorada por* i **hacer**
      para *cada asignatura* l *valorada por* i **hacer**
        Est[i, k] ←
         Est[i, k] + Similaridades[k, l] · Valoraciones[i, l];
      fin
      Est[i, k] ← Est[i, k]/n;
    fin
  fin
fin

según los contenidos sigue un enfoque basado en minería de textos y análisis semántico. A continuación se detallan las medidas mencionadas.

- *Índice Jaccard*. Se puede aplicar tanto a los profesores como a las competencias de una asignatura. Se basa en aplicar teoría de conjuntos a un par de asignaturas X e Y como el ratio entre los profesores o las competencias que ambas tienen en común entre el total de profesores o competencias que tienen entre las dos, como se describe en la ecuación 20. Así, la medida se acercará más a 0 cuanto menos profesores o competencias en común tengan X e Y, y más a 1 cuanto más, en relación a cuantos tengan por separado. función

$$J(X, Y) = \frac{X \cap Y}{X \cup Y} \tag{20}$$

- *Coeficiente de verosimilitud logarítmica*. Se puede aplicar tanto a los profesores como a las competencias de una asignatura. Es el logaritmo natural de la función de verosimilitud, que, a grandes rasgos, es una medida de cómo de improbable es que dos asignaturas se

superpongan fruto del azar, dado el número total de profesores y el número de profesores (o competencias) que cada una tiene por separado. Cuanto más improbable, más similares deberían ser las asignaturas en cuestión.

- *Similaridad entre áreas de conocimiento*. Se asume que cada asignatura sólo puede pertenecer a un área de conocimiento, y que éstas no tienen relación entre sí. Así, esta medida de similitud comprueba si un par de asignaturas comparten área de conocimiento, devolviendo 1 si tienen la misma y 0 si es distinta.

- *Similaridad entre contenidos*. La similaridad entre contenidos aplica minería de textos a los contenidos teóricos y prácticos presentes en las guías docentes de las asignaturas con el fin de obtener un coeficiente de similitud basado en términos en común. Para ello se siguen los siguientes pasos:

  1. Indexado del apartado *Contenidos* especificado en la guía de la asignatura: se ha implementado un *parseador* de texto personalizado y basado en el idioma del texto (español), usado junto a un conjunto de *stop words* (palabras a ignorar por no aportar significado) adaptado al dominio. Como resultado, para cada documento se genera una lista de *tokens* o lexemas con significado, además de la frecuencia y el número de apariciones que tienen en el documento.

  2. Para cada par de asignaturas X e Y, se genera un conjunto B con la unión de los *tokens* que tiene cada una. Para cada asignatura, se construye un vector, $\vec{X}$ o $\vec{Y}$, con tantos elementos como hay en B. Estos vectores almacenarán la frecuencia de cada *token*. Finalmente se normaliza cada vector usando la norma *L1*. Así, se obtienen las frecuencias relativas a cada par de asignaturas.

  3. Para obtener el valor de similaridad final entre cada par de asignaturas X e Y se aplica la similaridad del coseno a los vectores de frecuencias como se detalla en la ecuación 21, que estará acotada en el intervalo $[0, 1]$.

$$cos(\theta) = \frac{\vec{X} \cdot \vec{Y}}{\|\vec{X}\| \cdot \|\vec{Y}\|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \tag{21}$$

4.3 EVALUACIÓN DEL SISTEMA DE RECOMENDACIÓN

Dado que uno de los principales objetivos de este trabajo es la evaluación del RS y su comparativa con otros modelos no-híbridos o monocriterio, es muy importante desarrollar un sistema que permita evaluar de forma justa el RS adaptándose a la naturaleza de los datos con los que se trabaja. Así, el proceso de evaluación basada en desviación se ajustaría a lo expuesto en la sección 3.2.1, de forma que para cada estudiante del conjunto de datos se ocultaría un porcentaje de sus valoraciones sobre las asignaturas para que el RS construya el modelo con el resto de valoraciones y a continuación se tratará de estimar el valor de preferencia que los estudiantes mostraron sobre sus valoraciones ocultadas. Con las valoraciones reales y las estimadas se obtendrían las métricas de desviación presentadas en los antecedentes.

La primera mejora que se propone es transformar el proceso de evaluación, que esencialmente sería de tipo *hold-out*, a una validación cruzada de forma que las medidas de error sean estadísticamente significativas. Así, las asignaturas valoradas por cada estudiante se dividirían en n particiones y el proceso de evaluación se repetiría n veces tratando de predicir cada vez las valoraciones de una partición. El proceso se detalla en el diagrama de flujo de la figura 4. El número de particiones se establecerá para este trabajo en $n = 5$.

Como se muestra en la figura 4, la siguiente mejora que se propone afecta a cómo son construidas las particiones de asignaturas. Concretamente, se presenta un método de muestreo estratificado que trate de compensar el desbalanceo que existe en el número de valoraciones que recibe cada asignatura: al trabajar con unos datos que han sido recogidos de estudiantes reales, inevitablemente las asignaturas de primeros cursos y las que son comunes a todas las especialidades han recibido más valoraciones (se puede ver esto en el resumen de los datos presentado en el apéndice A). Si las particiones se construyen aleatoriamente, es posible que en una partición caigan muchas valoraciones sobre asignaturas poco valoradas, con lo que la medida del error estaría distorsionada. Para solucionar esto se propone el método que se detalla en el algoritmo 4, que esencialmente ordena todas las asignaturas de más a menos valoraciones y en base a ese orden va recorriendo las valoraciones de cada estudiante y asignando cada una a una partición cíclicamente. De esta manera, en cada partición se asegura que habrá equilibrio entre asignaturas con muchas valoraciones (fáciles de estimar) y con pocas (difíciles de estimar).

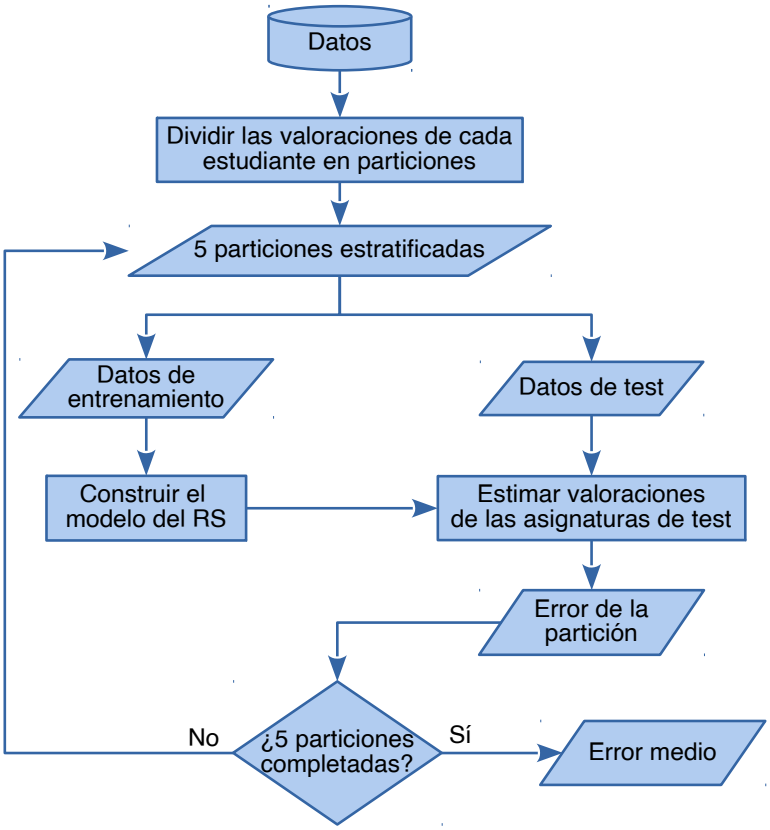


Figura 4: Proceso de evaluación del RS

Algoritmo 4: Partición estratificada de las valoraciones

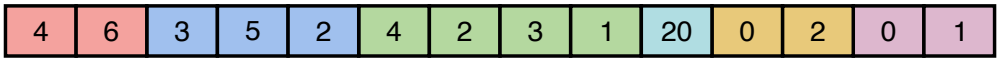
```
inicio
    A ← ordenarAsignaturas () ;
    para cada estudiante i en el sistema hacer
        p ← enteroAleatorio ([0, 4]) ;
        para k =0 a len(A)-1 hacer
            si existe valoración Valoraciones[i, A[k]] entonces
                añadir Valoraciones[i, A[k]] a Particiones[p] ;
                p ← (p + 1) % 5 ;
            fin
        fin
    fin
fin
```

4.4 OPTIMIZACIÓN AUTOMÁTICA DEL SISTEMA MEDIANTE BÚSQUEDA GENÉTICA

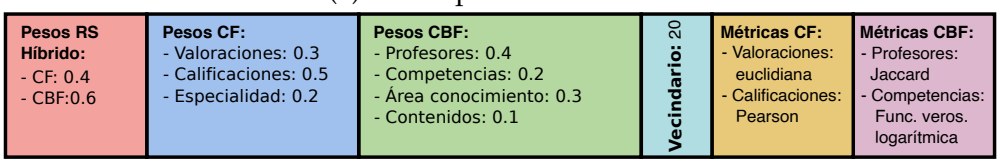
El RS propuesto tiene múltiples criterios que son ponderados con pesos para indicar su relevancia en las recomendaciones generadas. Además, hay otros parámetros que se deben configurar en un RS como son el tamaño de vecindario y las métricas de similaridad asociadas a valoraciones, calificaciones, profesores y competencias. Así, se han diseñado varios GA cuyo propósito es encontrar la configuración que mejor se adapte a los datos con los que se trabaja. Concretamente se va a probar con tres métodos distintos: un SGEA, una búsqueda adaptativa con el algoritmo CHC y un enfoque de búsqueda local con algoritmos meméticos con el algoritmo *Clearing*. Aunque los operadores de cada uno sean distintos, tienen elementos comunes dados por las características del problema, como son la representación de individuos y la función de aptitud o *fitness*. Se explican ambos a continuación.

4.4.1 *Representación de individuos*

El objetivo de los GA implementados es la búsqueda de una configuración óptima para el RS híbrido multi-criterio, así que los individuos que utilicen deberán ser posibles configuraciones para el sistema propuesto. Estas configuraciones se representarán mediante 14 genes codificados como enteros y agrupados en 5 categorías. En la figura 5 se muestra un posible individuo para los GA: en 5a se muestra cómo se codificaría su genotipo, es decir, los diferentes valores enteros que deberán ser optimizados por el GA. Igualmente, en la figura 5b se muestra el correspondiente fenotipo, es decir, el significado específico en la configuración del RS que representaría este individuo. A continuación se describe cada grupo de genes:



(a) Genotipo del individuo



(b) Fenotipo del individuo

Figura 5: Representación de individuos en los GA

- Los 2 primeros genes representan los pesos usados para combinar las estimaciones dadas por los subsistemas CF y CBF respectivamente. Concretamente, serían los valores α y β que muestra el algoritmo 1. En el momento de evaluar al individuo, a este grupo se le aplica una normalización de tipo min-max que escala los valores dados por los genes en el rango $[0, 1]$.

- Los siguientes 3 genes representan los pesos dados a cada criterio considerado en cuanto a la información del estudiante. Concretamente, serían los valores α, β y γ representados en el algoritmo 2. Se pueden definir en cualquier rango de valores enteros, y en el momento de la evaluación se le aplica al grupo una normalización de tipo min-max que escala los valores dados por los genes en el rango $[0, 1]$.

- Los siguientes 4 genes representan los pesos dados a cada criterio considerado en cuanto a la información de la asignatura. Concretamente serían los valores α, β, γ y δ en el algoritmo 3. Se pueden definir en cualquier rango de valores enteros, y en el momento de la evaluación se le aplica al grupo una normalización de tipo min-max que escala los valores dados por los genes en el rango $[0, 1]$.

- El siguiente gen se corresponde con el tamaño de vecindario usado por el subsistema de CF.

- Los siguientes 2 genes representan las métricas usadas para calcular las similaridades en cuanto a valoraciones y calificaciones en el subsistema CF. Estos genes siguen un enfoque categórico en el que toman un identificador numérico para cada una de las 4 métricas relativas a estos criterios que se han descrito en el apartado 4.2.1.

- Los últimos 2 genes representan las métricas usadas para calcular las similaridades en el subsistema CBF, concretamente las relativas a los profesores y las competencias. Estos genes siguen un enfoque categórico en el que se toma un identificador numérico para cada métrica mencionada en el apartado 4.2.2.

4.4.2 *Función fitness*

El *fitness* que se utilizará para medir la bondad de cada individuo será el RMSE entre las estimaciones de preferencia dadas por el RS, configurado

con la solución dada por el individuo, y las valoraciones reales. Esta medida tiende a penalizar los errores mayores más duramente que otras. Así, el RMSE, cuyo propósito es minimizarse, se define como:

$$RMSE = \sqrt{\frac{\sum_{(i,j)\in K}(p_{i,j} - v_{i,j})^2}{\#K}} \tag{22}$$

Donde

$p_{i,j}$	es la preferencia estimada del estudiante i sobre la asignatura j.
$v_{i,j}$	es la valoración real del estudiante i sobre la asignatura j.
$K = \{(i,j)\}$	es el conjunto estudiante-asignatura de las valoraciones a predecir.

En el cálculo del *fitness* se utilizará un proceso de tipo *hold-out* donde el 80% de los datos será utilizado para entrenamiento y el 20% restante para test.

4.4.3 *Algoritmo Generacional Elitista Simple*

La primera aproximación que se hace a la utilización de GA en la optimización del RS propuesto es a través de un SGEA. Así, se contará con una población de individuos que irá evolucionando hasta un cierto número de generaciones, siempre conservando al mejor individuo encontrado. Adicionalmente, se ha introducido un optimizador local similar al empleado en los algoritmos meméticos [36], que al final de cada generación tratará de mejorar el mejor elemento de la población mediante cambios locales durante n intentos. El proceso general del algoritmo se especifica en el diagrama de flujo de la figura 6. Además, los operadores genéticos se detallan en los siguientes apartados.

SELECCIÓN DE PADRES

El primer paso en una generación es seleccionar qué individuos de la población serán los padres que generarán la descendencia. El proceso de selección pasa por un procedimiento estocástico de torneo, donde 2 individuos son enfrentados para acabar seleccionando el de mejor *fitness*, repetido tantas veces como marque una probabilidad de cruce dada.

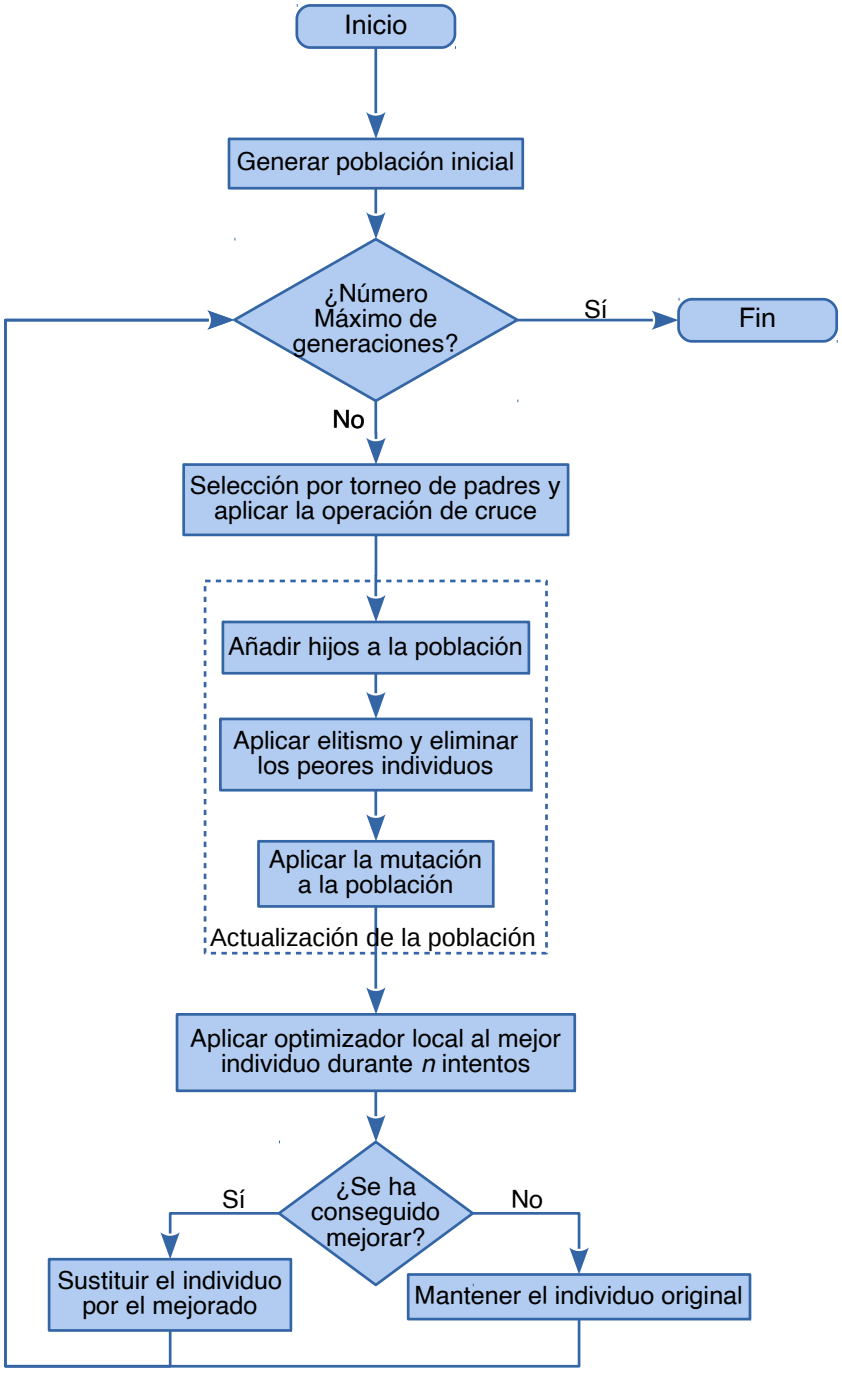


Figura 6: Diagrama de flujo del SGEA propuesto

CRUCE DE INDIVIDUOS

Se utiliza un operador de cruce de tipo uniforme que toma 2 padres para generar 2 hijos. Este operador evalúa cada posición del cromosoma y aleatoriamente asigna el gen de uno de los padres a uno de los hijos, y el del otro padre al otro hijo. Adicionalmente, en este trabajo se propone una modificación del operador con el fin de abordar aquellos genes que están relacionados entre sí (ver figura 5): $(1, 2)$, $(3, 4, 5)$ y $(6, 7, 8, 9)$. Estos grupos son tratados cada uno como un único bloque, de forma que cada uno de éstos será asignado del mismo padre para el mismo hijo. Los últimos 5 genes correspondientes a las configuraciones de vecindario y métricas de similaridad son tratados de forma individual: cada uno será asignado indistintamente a un hijo o al otro. En la figura 7 se muestra un ejemplo de cruce.

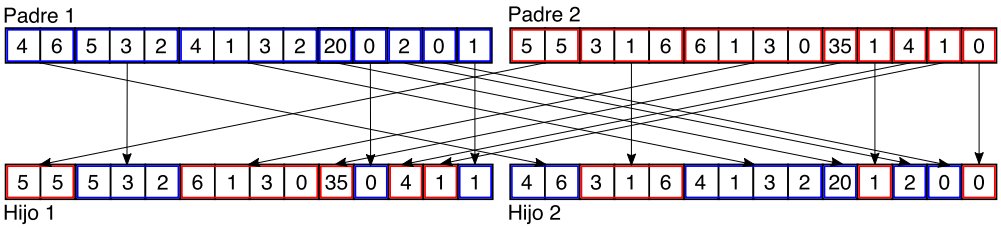


Figura 7: Ejemplo de cruce uniforme propuesto

OPERACIÓN DE MUTACIÓN

Tras haber aplicado el operador de cruce en una generación, se introduce una mutación con el fin de favorecer la diversidad en la población. Esta operación se aplica aleatoriamente (según una probabilidad dada) a los individuos actuales de la población, es decir, al conjunto de los individuos de la generación anterior que no se cruzaron y a los hijos de los que sí lo hicieron. El operador elegido para aplicar esta mutación es de tipo uniforme, esto es: evalúa cada gen del cromosoma y aleatoriamente (según la probabilidad dada) modifica su valor dentro del rango permitido. En la figura 8 se muestra un ejemplo de mutación. Nótese que el valor efectivo de la mutación sobre los genes que representan a los pesos se producirá una vez que estos hayan sido normalizados.

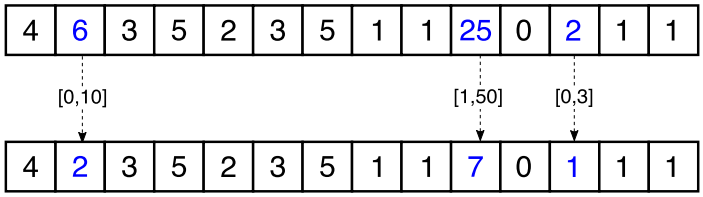


Figura 8: Ejemplo de mutación uniforme propuesta

OPTIMIZADOR LOCAL

La modificación principal de este SGEA se introduce como último paso de una generación. Se trata de un modificador que afecta aleatoriamente a alguno de los últimos genes correspondientes a vecindario y métricas de similaridad, cambiando su valor aleatoriamente por otro dentro de su rango permitido. Este operador se aplica durante n intentos a un individuo dado: al final en la población se introducirá aquella modificación que mejor *fitness* haya obtenido, descartando al individuo original, a no ser que ninguna haya conseguido mejorar a éste. Debido al coste computacional de este operador, solamente es aplicado al mejor individuo de la población.

ACTUALIZACIÓN ELITISTA DE LA POBLACIÓN

La población que pasará a la siguiente generación será la unión de aquellos individuos *estériles* que no se cruzaron junto con los nuevos hijos fruto de la operación de cruce. Así, en principio ningún padre pasa a la siguiente generación. No obstante, al tratarse de una aproximación elitista, el mejor individuo encontrado permanece en cualquier caso en la población. Así, si éste estaba entre los padres, volverá a ser introducido en la población mientras que el peor individuo de la misma es eliminado.

4.4.4 *Algoritmo CHC*

La segunda aproximación que se estudia para optimizar la configuración del RS es la búsqueda adaptativa de CHC [13]. Como se ha documentado ampliamente, [37], [38], el algoritmo CHC ha demostrado un excelente rendimiento en escenarios similares de selección de características. Su combinación de alta diversidad, dada por la prevención de incesto y el reinicio de la población, combinada con la alta convergencia, dada por la selección elitista, presentan a este GA como apropiado para las características del problema que se aborda. El proceso general del algoritmo se especifica en el diagrama de flujo de la figura 9. Además, los operadores genéticos se detallan en los siguientes apartados.

PREVENCIÓN DE INCESTO

Este operador promueve la exploración y reduce la deriva generacional. Así, establece una distancia de umbral que permite obtener descendencia sólo de aquellos individuos suficientemente distantes (prevención de incesto). Igualmente, el umbral, que es adaptativo, permite controlar

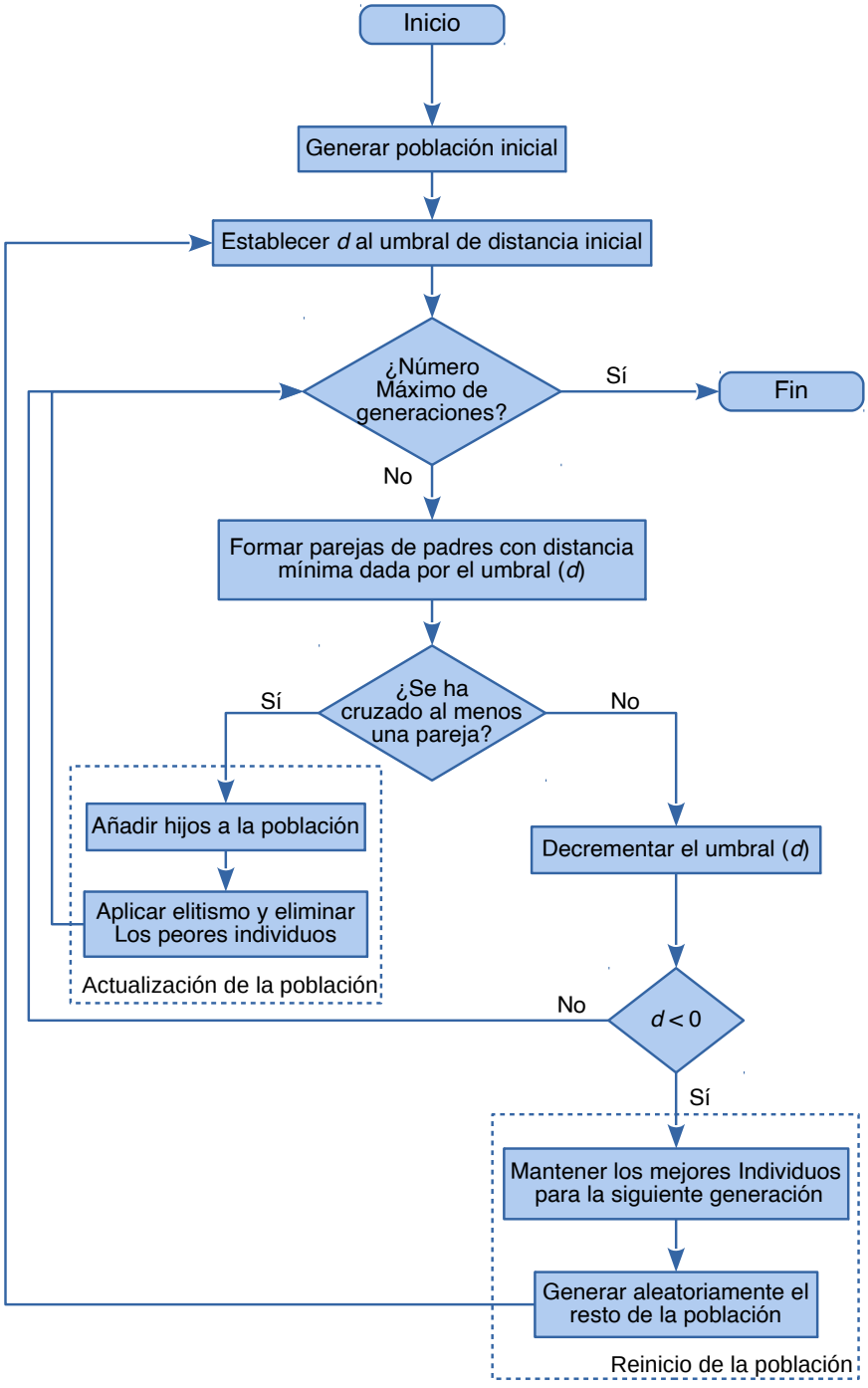


Figura 9: Diagrama de flujo del algoritmo CHC propuesto

cuándo reiniciar la población porque los individuos se hayan vuelto demasiado similares entre sí.

Este umbral (d en el diagrama de flujo de la figura 9) se establece a un valor inicial al inicio del algoritmo. En cada generación todos los individuos de la población se emparejan de forma aleatoria y se intentan cruzar sólo si la pareja excede el umbral d. Si no se puede producir ningún cruce, el umbral se decrementa en una unidad para la siguiente generación, es decir, en la siguiente generación el operador de cruce será menos restrictivo. Si la población se vuelve demasiado similar, el umbral alcanza el valor de 0. En este punto la población es reiniciada y el umbral establecido de nuevo a su valor inicial.

DISTANCIA ENTRE INDIVIDUOS

La prevención de incesto antes comentada viene dada por un nivel de similaridad variable aceptado entre padres. Con el fin de maximizar las posibilidades de diversificación de la población, es necesario usar una métrica de distancia apropiada que identifique sin lugar a dudas la similaridad entre individuos.

En este trabajo, el cromosoma propuesto para representar a los individuos tiene grupos de genes cuyo significado está relacionado (ver figura 5). Así, es necesario usar una métrica de distancia personalizada que evite introducir ruido en la medida entre individuos. Se propone utilizar la distancia de Hamming: dados 2 individuos, la distancia entre ellos viene dada por el número de genes en los que difieren. Esta métrica será aplicada con algunas modificaciones como se muestra en la figura 10: dados 2 individuos i, j, en primer lugar, sus genes asociados a pesos son normalizados y tratados por grupos como se especifica en la definición del cromosoma. Así, la suma de los genes de cada grupo será 1. A continuación la evaluación de la métrica de distancia será:

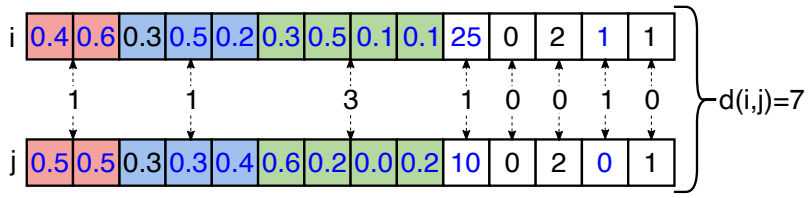


Figura 10: Ejemplo de distancia de Hamming propuesta

- Cuando se compara el primer grupo, compuesto de 2 genes, sólo dos estados son posibles: o los dos genes son iguales o los dos difieren. Así, este grupo contribuirá con 0 o 1 al valor final de

distancia $d(i,j)$. De acuerdo al ejemplo de la figura 10, los dos genes son distintos, así que en este ejemplo se suma 1 a $d(i,j)$.

- Al comparar el segundo grupo, compuesto de 3 genes, hay 3 estados posibles: cero, dos o tres genes difieren, así que este grupo contribuirá con 0, 1 o 2 a $d(i,j)$. De acuerdo a la figura 10, hay sólo un gen igual mientras que los otros dos son diferentes, así que el grupo sumará 1 a la distancia final $d(i,j)$. Nótese que, debido a la restricción de que los 3 genes suman 1, si dos de los genes son iguales el tercero también lo sera.

- Cuando se compara el tercer grupo, que se compone de 4 genes, se sigue el mismo principio contribuyendo con 0, 1, 2 o 3 a la distancia $d(i,j)$ según los 4 posibles estados. En la figura 10 todos los genes son distintos, así que el grupo suma 3 a $d(i,j)$.

- Finalmente, el resto de genes se computan individualmente: si son iguales sumarán 0 a la distancia final $d(i,j)$ y 1 en otro caso. En la figura 10, entre estos genes hay 2 distintos, contribuyendo con esta cantidad a $d(i,j)$.

CRUCE DE INDIVIDUOS

En este algoritmo todos los individuos de la población se seleccionan como posibles padres para cruzar, formándose las parejas de forma aleatoria. Si el umbral de incesto lo permite, las parejas serán cruzadas mediante el mismo operador de cruce de tipo uniforme y adaptado al problema que se ha expuesto para el caso del SGEA (ver sección 4.4.3).

ACTUALIZACIÓN DE LA POBLACIÓN

La población se actualiza de generación en generación por medio de una estrategia de elitismo. Concretamente, los individuos que pasan a la siguiente generación serán los mejores de la actual, que fusiona a todos los padres y la descendencia resultante.

REINICIO DE LA POBLACIÓN

El proceso elegido para actualizar la población puede introducir una alta presión de selección. Con el propósito de evitar una convergencia prematura, CHC utiliza el reinicio de la población para introducir diversidad en la búsqueda automática. El proceso de reinicio trata, por un lado, de mantener la generación elitista y, por el otro, conservar la diversidad en la población. Así, cuando el umbral de prevención de

incesto alcanza el valor mínimo, es decir, la población se ha vuelto tan similar que el umbral se ha tenido que reducir hasta 0 para permitir cruces entre individuos, la población debe ser reiniciada.

En la propuesta que se ha diseñado el reinicio funciona como sigue: los n mejores individuos de la población se conservan para ser introducidos sin modificaciones en la siguiente generación. El resto de individuos hasta que se completa el tamaño de la población son generados de forma aleatoria.

4.4.5 *Algoritmo Clearing*

La última aproximación que este trabajo estudia para la optimización del RS híbrido multi-criterio utiliza el algoritmo *Clearing* [14]. Este procedimiento no se diferencia en el flujo general de un GA simple donde se parte de una población que va evolucionando a lo largo de las generaciones. El procedimiento de *Clearing* se aplica entre la evaluación del *fitness* y la selección estocástica de individuos. Así, se realiza el agrupamiento de los individuos en nichos de manera que se limpie el *fitness* de todos aquellos individuos excepto el dominante de cada nicho. De esta manera, la selección de padres parte de un grupo más diverso entre sí. Nótese que en el diseño original del procedimiento de *Clearing* se asume que se trabajará con problemas donde el *fitness* sea a maximizar, con lo que la limpieza se limita a establecerlo a un valor de 0. En el caso que ocupa a este trabajo, el *fitness* es el RMSE, a minimizar por tanto. Por lo que para ignorar a los individuos *limpiados* durante la selección de padres se le debe asignar un *fitness* muy alto. Para el dominio del problema bastará con un valor de RMSE igual a 5. El proceso general del algoritmo se especifica en el diagrama de flujo de la figura 11.

RADIO DE NICHO

El radio de nicho, σ, determina hasta qué umbral de similaridad se considera que 2 individuos pertenecen al mismo nicho. Aunque podría fijarse manualmente, en este trabajo se propone calcularlo dinámicamente para cada generación como el 20% de la distancia entre los dos individuos más lejanos de la población.

$$\sigma = max(distanacias\_individuos) \cdot 0.2 \tag{23}$$

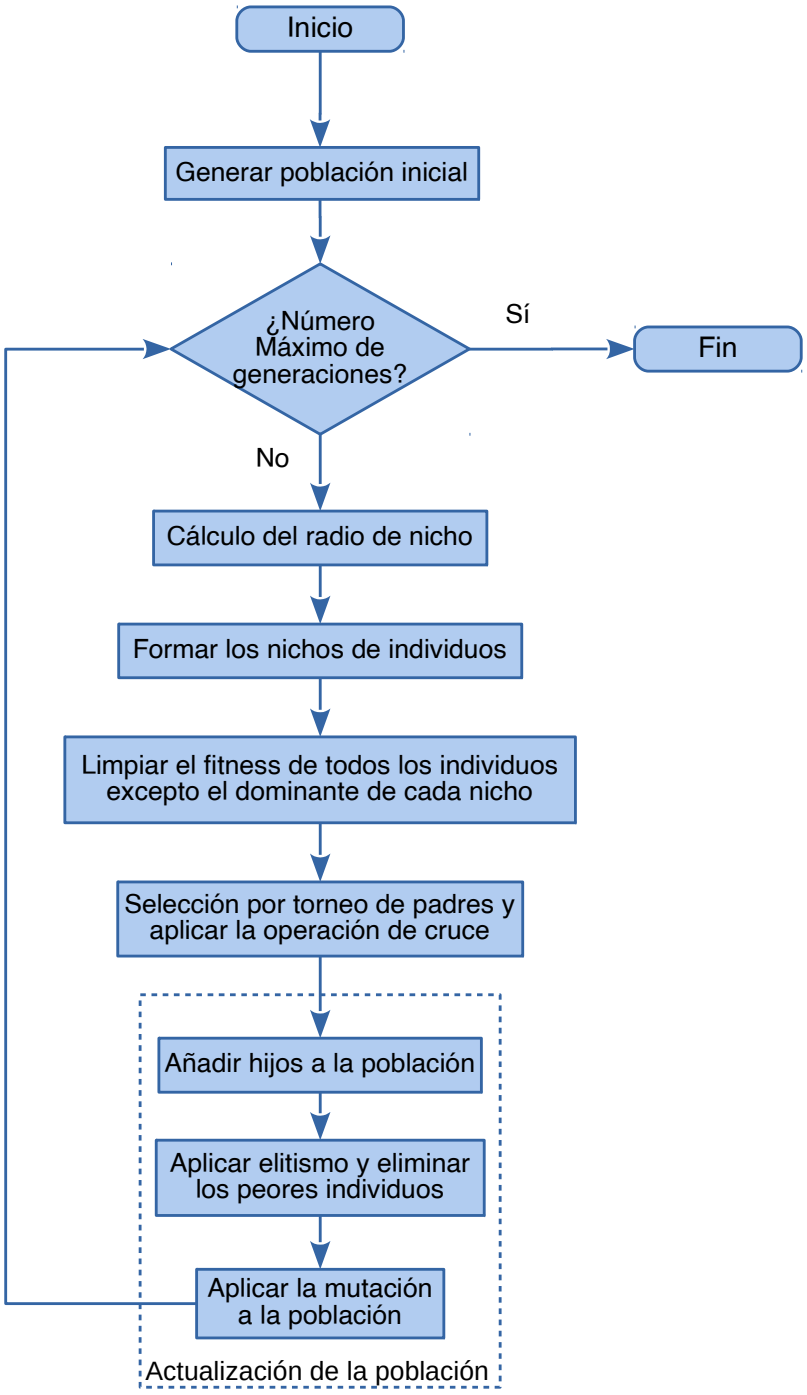


Figura 11: Diagrama de flujo del algoritmo de *Clearing* propuesto

Algoritmo 5: Creación de nichos en *Clearing*

```
inicio
    P ← ordenarPoblacionMejorAPeor ();
    n ← len(P);
    para i = 1 a n hacer
        si fitness (P[i]) < 5.0 entonces
            nGanadores ← 1;
            para j = i + 1 a n hacer
                si fitness (P[j]) < 5.0 and distancia (P[i],P[j]) < σ
                 entonces
                    si nGanadores < κ entonces
                        nGanadores ← nGanadores + 1;
                    fin
                    en otro caso
                        fitness (P[j]) ← 5.0;
                    fin
                fin
            fin
        fin
    fin
fin
```

CREACIÓN DE NICHOS

El núcleo del procedimiento de *Clearing* es la creación de los nichos, que vendrá determinada tanto por el radio del nicho σ como por el parámetro κ, que fija la capacidad máxima que tendrá cada nicho. De esta forma, aunque σ aún admitiera más individuos en un nicho dado, si se ha alcanzado el máximo κ se procede a crear un nuevo nicho. El procedimiento de creación de nichos se especifica en el algoritmo 5, en el que se muestra que es necesario partir de los individuos de la población ordenados de mejor a peor, para ir asignándolos a los nichos. También se muestran las adaptaciones propias al problema relativas al valor con el que se va a limpiar el *fitness*.

DISTANCIA ENTRE INDIVIDUOS

La distancia entre individuos se resolverá mediante la misma métrica de Hamming adaptada que se presentaba para el algoritmo CHC (ver sección 4.4.4) en la que se tienen en cuenta el significado de los genes y sus relaciones para el problema que se estudia.

SELECCIÓN Y CRUCE DE INDIVIDUOS

Tras el limpiado de *fitness* de los individuos no relevantes del nicho, se procede a aplicar la operación de cruce entre individuos. Se va a seguir un procedimiento similar al utilizado en el SGEA (ver sección 4.4.3): selección estocástica de cruce por torneo de tamaño 2, seleccionando tantos individuos como marque la probabilidad de cruce. Los cruces se forman mediante el operador de cruce uniforme adaptado al problema. Nótese que ahora un alto número de individuos con una aptitud aceptable pero poco diversos entre sí han sido *limpiados*, es decir, su *fitness* se ha falseado a un valor malo. Esto implica que tendrán una baja probabilidad de ser cruzados, favoreciendo la diversidad a través de los nichos que han quedado tras el *Clearing*.

OPERACIÓN DE MUTACIÓN

La mutación es el último paso en la generación, con el fin de aumentar la diversidad de la población. Se aplicará aleatoriamente sobre todos los individuos actuales. Su operador será de tipo uniforme, tal y como se ha diseñado para el SGEA (ver sección 4.4.3).

ACTUALIZACIÓN DE LA POBLACIÓN

La población se actualiza de generación en generación por medio de una estrategia de elitismo combinada con la información de cada nicho. Concretamente, se transmitirán los hijos generados por el cruce (que principalmente incluirá descendientes de los individuos dominantes de cada nicho) y se completará el tamaño de la población con los mejores individuos ya considerando sus *fitness* reales. Además, el mejor individuo de la población se transmitirá siempre de generación en generación.

5

ESTUDIO EXPERIMENTAL

En este capítulo se describirá la experimentación llevada a cabo. Como se comenta en la metodología (sección 4) y se puede comprobar en el apéndice A, se ha trabajado con un conjunto de datos obtenido a partir de alumnos reales del Grado en Ingeniería Informática, contando con 2500 valoraciones procedentes de 3 años académicos seguidos (2016 a 2018). Así mismo, la implementación del RS se ha llevado a cabo dentro del *framework* en Java de Apache Mahout [29] para la computación distribuida y el álgebra lineal. Por otro lado, se ha utilizado la librería, también en Java, de computación evolutiva JCLEC [39] para la implementación y adaptación de los GA utilizados. Todos los experimentos realizados se han ejecutado en un equipo con procesador AMD Ryzen 5 1600, 8 GiB de RAM y Ubuntu 16.04 de 64 bits como sistema operativo.

La experimentación llevada a cabo consta de 3 fases: en la primera se lleva a cabo un estudio comparativo entre los 3 GA propuestos, analizando cuál puede ser la mejor técnica para optimizar el RS propuesto. En la segunda fase, se analiza la relevancia de los criterios y los valores para el resto de parámetros que el GA escogido obtiene para el RS. Por último, la tercera fase estudia el rendimiento del RS propuesto comparándolo con los modelos CF o CBF basados ambos en múltiples criterios. Además, en la comparativa se incluyen los RS mono-objetivos, que emplean solamente cada uno de los criterios por separado. Así, se muestra la importancia de ponderar y utilizar técnicas híbridas para obtener un mejor rendimiento en la recomendación.

5.1 CONFIGURACIÓN Y SELECCIÓN DEL ALGORITMO GENÉTICO

Como en la mayoría de los problemas, en este trabajo se parte de un espacio de búsqueda de soluciones con unas características particulares, lo

que implica que es posible que un GA se adapte mejor que otro al problema. Así pues, se van a realizar una serie de experimentos exploratorios para identificar cuál de las propuestas se ajusta mejor. Las pruebas consistirán en primera instancia en encontrar una configuración adecuada para cada uno de los GA considerados, para lo cual se estudian diferentes valores de los parámetros. A continuación, la mejor configuración de cada GA se comparará con las del resto de GA para finalmente decidir cuál será la técnica de optimización que se considerará en la posterior optimización del RS propuesto.

Comparar distintos GA entre sí no es trivial [12], tanto por qué criterio considerar para decidir cuál es mejor, como por cómo fijar los parámetros para conseguir unas condiciones de ejecución igualitarias. En este caso, se va a considerar como criterio de selección principal encontrar un individuo con mejor *fitness*, aunque también se tendrá en cuenta el *fitness* medio de la población y que se mantenga una varianza en ésta relativamente alta. En cuanto a fijar unas condiciones de ejecución igualitarias entre algoritmos, se aplicarán las siguientes condiciones:

- Ya que se trata de un elevado número de pruebas exploratorias donde cada individuo equivale a entrenar un RS con 2500 valoraciones, éstas deben ser asumibles en tiempo de cómputo. Por ello se van a fijar tanto un número corto de generaciones (100), como una población pequeña (50). En esta línea, se han llevado a cabo varios estudios con un número mayor de generaciones, y mayor población, para estudiar la convergencia, demostrando que los resultados finales se pueden extrapolar con los conseguidos con dicha configuración.

- Un parámetro común a todos los GA considerados es la probabilidad de cruce. Está ampliamente asentado que debe ser un valor alto [12]. Las pruebas realizadas determinan que un valor de 0.9 es el más apropiado para los algoritmos estudiados.

- Un parámetro que puede determinar tanto la varianza potencial de la población como cuánto afectan los cambios del GA a la configuración del RS es el rango en que podrán variar los genes asociados a los pesos antes de que se aplique la normalización. Las pruebas realizadas muestran que un rango apropiado es $[0, 50]$. De igual forma, un rango suficiente para explorar el número de vecinos óptimo sería $[1, 50]$.

Fijados estos parámetros comunes a los 3 GA propuestos, los parámetros de cada GA se variarán entre unos valores intermedios para

Tabla 2: Resumen de pruebas de parámetros en los GA

Parámetro	Probados	Seleccionado	Coef. corr.
	SGEA		
Prob. mutación	0.1 a 0.9 con $\Delta = 0.1$	0.5	−0.510 31
Nº aplic. opt.	1 a 9 con $\Delta = 1$	7	−0.740 59
	CHC		
d inicial	2 a 16 con $\Delta = 2$	4	0.837 45
Nº superviv.	3 a 20 con $\Delta = 3$	5	0.344 80
	Clearing		
Prob. mutación	0.1 a 0.9 con $\Delta = 0.1$	0.5	−0.045 64
Cap. máx. nicho	5 a 35 con $\Delta = 5$	25	−0.534 52

determinar cuál podría ser una buena configuración para éstos. En la tabla 2 se muetra un resumen de la experimentación realizada: por cada fila se muestra qué valores se estudiarán para los parámetros de cada GA: la probabilidad de mutación y el número de veces que se aplicará el optimizador local en SGEA, el umbral de distancia inicial entre individuos y el número de supervivientes a conservar en el reinicio de la población en el algoritmo CHC y la probabilidad de mutación y la capacidad máxima por nicho en el algoritmo *Clearing*. La 2ª columna muestra los distintos valores que se han probado. Para cada GA se han probado todas las combinaciones de los valores estudiados de cada parámetro. Finalmente la mejor configuración obtenida se muestra en la 3ª columna. En la 4ª columna se muestra el coeficiente de correlación entre el parámetro y el mejor *fitness* obtenido, calculado a partir de todos los valores probados para ese parámetro y dejando fijo el otro parámetro en el mejor valor encontrado. Este coeficiente puede dar una idea de cuánto influye el parámetro en el problema estudiado. En el apéndice B se pueden ver los resultados obtenidos para cada combinación probada.

En la figura 12 se muestra qué resultados han obtenido las configuraciones óptimas de cada GA atendiendo al mejor *fitness*, al *fitness* medio de la población y al del peor individuo en la última generación.

Los resultados muestran en todos los casos que la población tiene una alta tendencia a estancarse y no tener grandes diferencias en su *fitness*. Por eso las tácticas diversificadoras parecen ser las más deseables. En esta línea, CHC obtiene los resultados más prometedores en cuanto a un mejor *fitness* a la par que una varianza en la población relativamente

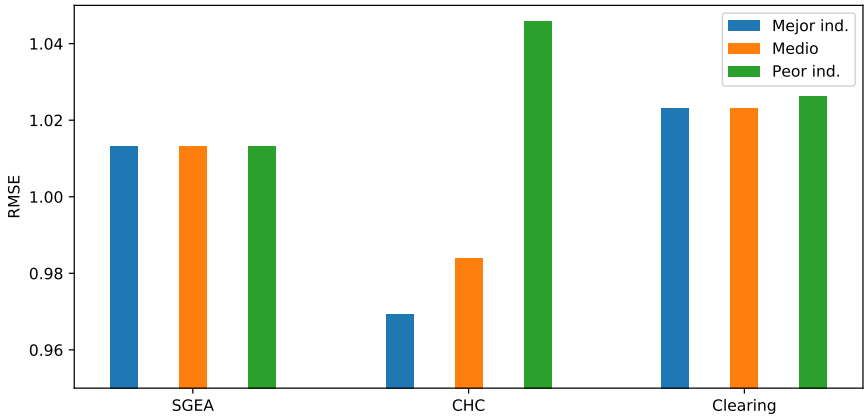


Figura 12: Resultados obtenidos por las mejores configuraciones de los GA

moderada. Así, para este algoritmo es deseable un umbral de distancia entre individuos bajo que permita que pronto se llegue al reinicio de la población y conservar pocos supervivientes para favorecer más diversidad. En SGEA, a pesar de tratar de combinar una probabilidad de mutación moderadamente alta sin llegar a demasiada aleatoriedad (más diversificación), la población ha convergido a un óptimo local prematuramente y está totalmente estancada. Por otro lado, el optimizador local sí que parece efectivo, siendo beneficioso tratar de mejorar al mejor individuo durante tantos intentos como sea posible. Por último, *Clearing* converge en todos los casos al peor resultado de los obtenidos en cuanto a mejor *fitness*, si bien mantiene una varianza relativamente alta de la población. Esto sugiere que, si bien las técnicas meméticas podrían tener potencial, haría falta probar con poblaciones mayores y más capacidad por nicho. Sin embargo, para las características de este problema consumiría altos recursos computacionales, mientras que se ha demostrado que el algoritmo CHC, consigue introducir mayor diversidad con menos recursos computacionales.

5.2 INFLUENCIA DE LOS CRITERIOS EN EL SISTEMA DE RECOMENDACIÓN

El peso asignado automáticamente a cada criterio por el GA da una idea de la influencia de éste en la construcción del modelo del RS. En esta sección se presenta un estudio de la evolución de los pesos y el resto de parámetros del RS durante una experimentación más en profundidad del algoritmo CHC, que es la técnica que mejor se adaptaba a este problema. La configuración final que se ha utilizado para la búsqueda automática se detalla en la tabla 3.

Tabla 3: Configuración de los parámetros de CHC

PARÁMETRO	VALOR
Número de generaciones	1000
Tamaño de la población	50
Probabilidad de cruce	0.9
Umbral de distancia inicial	4
Rango para genes de pesos	$[0, 50]$
Rango para gen de vecindario	$[1, 50]$
Rango para genes de métricas	$[0, 4]$ o $[0, 1]$

En la tabla 4 se muestran los pesos de cada criterio, así como las métricas de similaridad y el tamaño de vecindario, asignados por el GA al RS propuesto. Los resultados muestran la importancia de usar sístemas híbridos que combinen información tanto del estudiante como de la asignatura. Así, la relevancia de estos modelos para obtener buenas recomendaciones está balanceada, asignando una importancia del 54 % al modelo CF y del 46 % al modelo CBF. Atendiendo al modelo CF basado en información del estudiante, se aprecia que el criterio más determinante son las valoraciones (0.60), seguido por las calificaciones (0.30). Finalmente, el criterio de la especialidad es el menos relevante (0.10). También es destacable que el número de vecinos óptimo a tener en cuenta es relativamente pequeño (15 estudiantes). Así, una asignatura será recomendada a un estudiante principalmente si los estudiantes con valoraciones y calificaciones parecidas la valoraron positivamente. Atendiendo al modelo CBF basado en información de la asignatura, los criterios más relevantes son los profesores en común (0.65) y los contenidos similares (0.35), mientras que las competencias y el área de conocimiento parecen ser irrelevantes. Estos resultados revelan la importancia del profesor en la percepción que el estudiante tiene de la asignatura. Así, una asignatura será recomendada a un estudiante principalmente si los profesores que imparten en ella son los mismos a otras asignaturas que interesasen al estudiante en el pasado, aunque también influye si tienen contenidos en común.

Atendiendo a la evolución del peso de los criterios en la búsqueda automática, así como del resto de parámetros, se analizará la mejor configuración obtenida a lo largo de las generaciones por CHC. La figura 13 muestra la optimización del *fitness*, es decir, la minimización del RMSE entre las estimaciones y las valoraciones reales. Aquí se puede ver, por un lado, los efectos del reinicio de la población en los picos de incremento del *fitness* medio en la población. Por el otro lado, la búsqueda elitista

Tabla 4: Configuración del RS obtenida por el GA

RS Híbrido	
Peso de CF	0.54
Peso de CBF	0.46
CF (basado en estudiante)	
Valoraciones (mét. sim.)	Correlación de Pearson
Calificaciones (mét. sim.)	Correlación de Pearson
Valoraciones (peso)	0.60
Calificaciones (peso)	0.30
Especialidad (peso)	0.10
Tam. vecindario	15
CBF (basado en asignatura)	
Profesores (mét. sim)	Índice Jaccard
Competencias (mét. sim)	Índice Jaccard
Profesores (peso)	0.65
Contenidos (peso)	0.35
Competencias (peso)	0.00
Área conocimiento (peso)	0.00

y el reinicio siempre mantendrán al mejor individuo encontrado, por eso el *fitness* relativo al mejor individuo nunca se incrementa, aunque se mejora lentamente. Finalmente se obtiene un RMSE en torno a 0.96, corroborando que es durante las 100 primeras generaciones que se usaron para las pruebas exploratorias donde se producen los cambios más relevantes. También es importante notar que el conjunto de datos con el que se trabaja es complejo: no todas la asignaturas tienen un número de valoraciones suficientemente elevado para poder recomendarse con garantías, la mayoría de los estudiantes encuestados pertenecen a una misma especialidad... Se ha optado por mantener esta información y poder contar con al menos algo de información de la mayoría de las asignaturas, pero resulta evidente que algunas perjudican a la precisión de las recomendaciones que se realicen sobre ellas. Más información del conjunto de datos con el que se trabaja puede verse en el apéndice A. La evolución de las configuraciones del RS que producen este mejor individuo estudiado se representa en las figuras 14 y 15.

Atendiendo a la importancia dada a cada criterio, en la figura 14 se muestra la evolución de la mejor configuración del RS en función de la hibridación que se realiza y de la ponderación de los diferentes criterios utilizados. Relativo a los pesos que ponderan la relevancia que

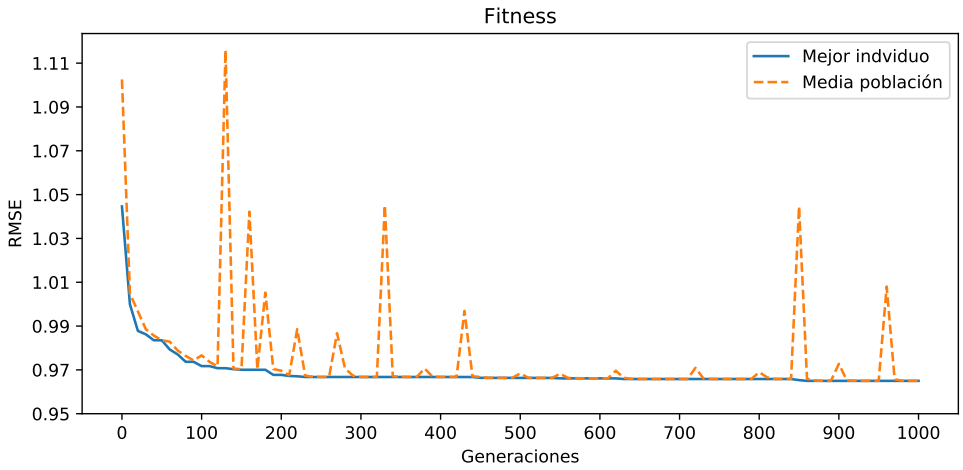


Figura 13: Evolución del *fitness* en el GA

se asigna a cada modelo utilizado en el RS híbrido, se puede ver una evolución muy estable, que tiende a balancear siempre la información del estudiante y la de la asignatura, aunque la primera es ligeramente más importante en las recomendaciones. Relativo a la evolución de los criterios específicos al estudiante, se pueden apreciar dos fases u óptimos locales: en la primera mitad de la experimentación las valoraciones tienden a monopolizar toda la importancia, aunque finalmente son compensadas con el criterio de las calificaciones. Por otro lado, el criterio de la especialidad no parece contribuir mucho a las recomendaciones, tal vez debido a las características de los estudiantes encuestados, que principalmente pertenecen a la especialidad en Computación por lo que este factor no aporta información relevante. Los criterios relativos al curso son los que presentan más inestabilidad. Aún así, hay una clara tendencia en la que los profesores y los contenidos tienen la importancia principal, mientras que las competencias y el área de conocimiento son prácticamente irrelevantes. Esta baja relevancia del área de conocimiento puede deberse a las características de los datos con los que se trabaja: la mayoría de las asignaturas consideradas pertenecen al mismo área, resultando en que este factor no proporciona información relevante. Por otro lado, la baja relevancia de las competencias puede deberse a que éstas son muy genéricas y muchas asignaturas comparten las mismas.

Atendiendo al resto de configuración del RS, la figura 15 muestra la evolución del mejor individuo relativa al vecindario y las métricas usadas para calcular la similaridad en los criterios que permiten su configuración, divididos entre relativos al estudiante y relativos a la asignatura. Se aprecia que el tamaño de vecindario se va reduciendo a medida que se optimiza la solución, señal de que son necesarios menos estudiantes

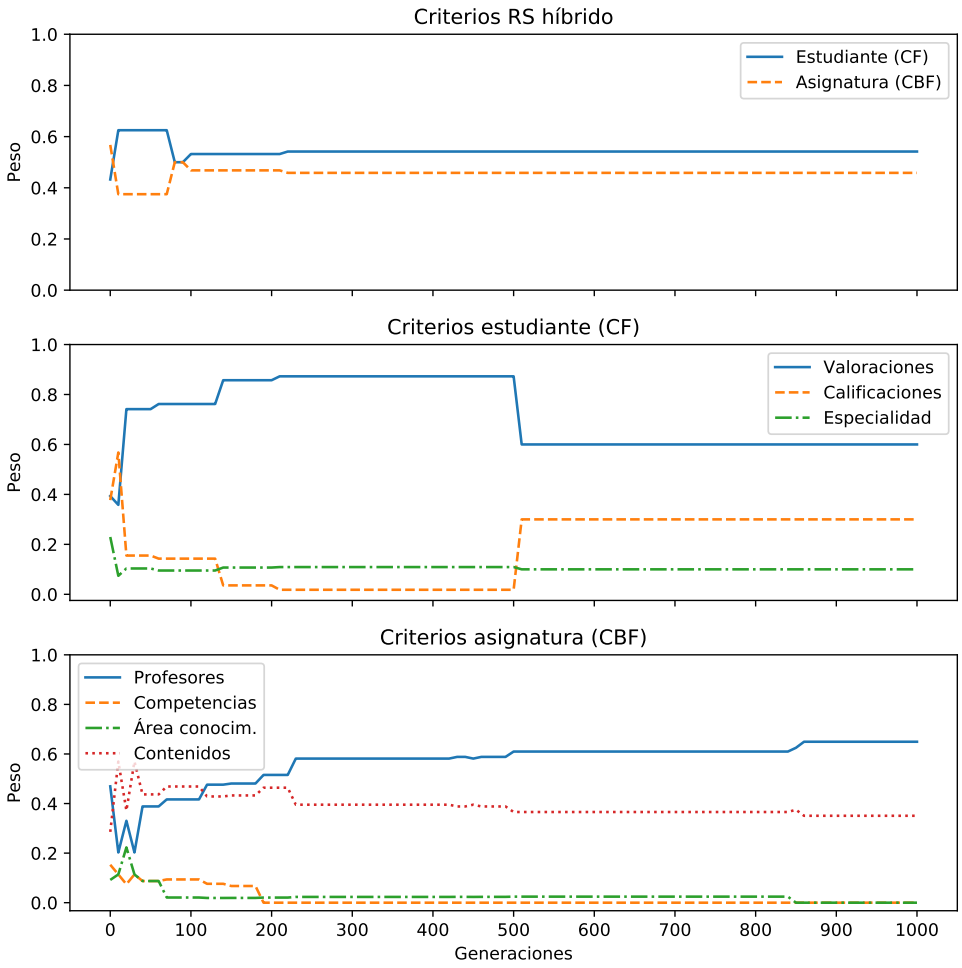


Figura 14: Evolución del peso de los criterios en el GA

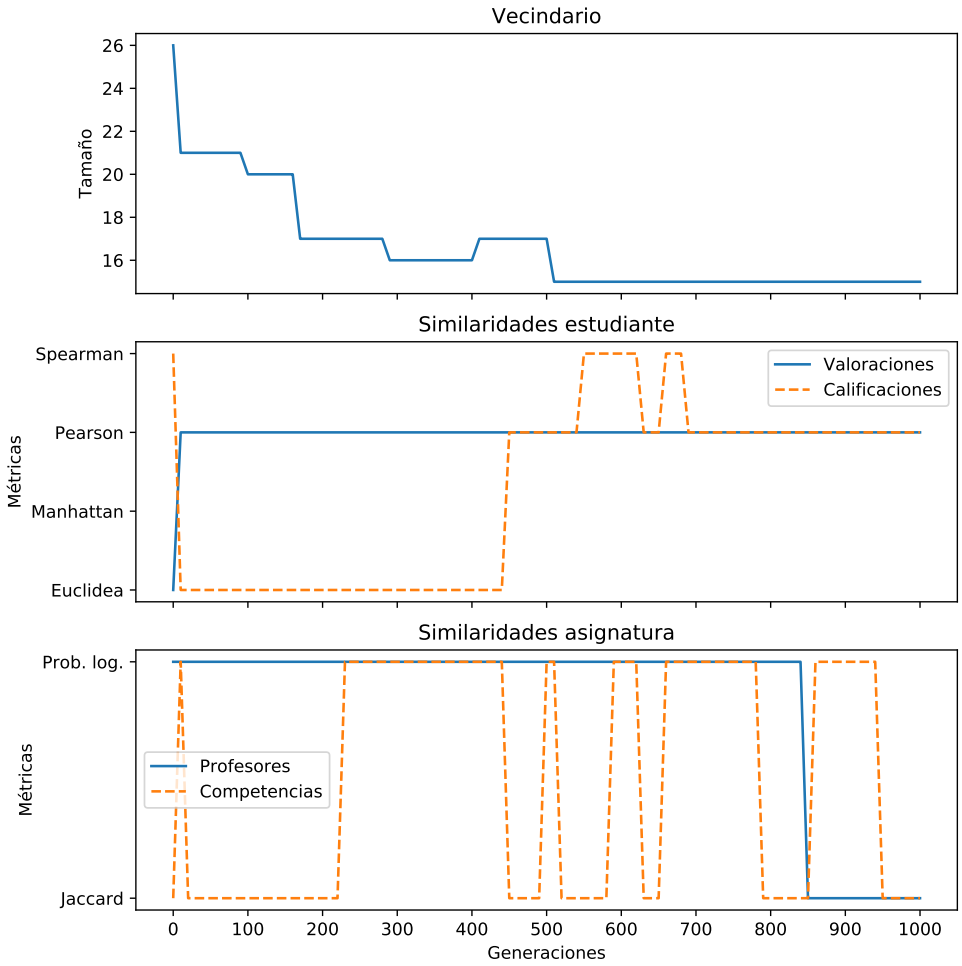


Figura 15: Evolución de vecindario y métricas de similaridad en el GA

para obtener buenas recomendaciones. Sobre las medidas de similaridad consideradas en el modelo CF del estudiante, parece que el coeficiente de correlación de Pearson es el enfoque que mejor funciona tanto para las valoraciones como para las calificaciones. Aún así, en las calificaciones se puede apreciar más inestabilidad, destacando la distancia euclidiana, que es un óptimo local durante la primera mitad de la experimentación. Atendiendo a las métricas de similaridad consideradas en la información de la asignatura: para los profesores se puede ver cómo cambia entre el coeficiente de verosimilitud y el índice Jaccard acorde con la optimización del peso de este criterio. En el caso de la similaridad de las competencias, se puede ver que cambia frecuentemente pero no parece que utilizando una u otra métrica haya mejora y por tanto la relevancia de este criterio tampoco aumenta.

5.3 COMPARATIVA DE RENDIMIENTO RESPECTO A OTROS MODELOS

Para estudiar la relevancia de la propuesta de este trabajo, se va a comparar con las versiones multi-criterio del modelo CF y del CBF con la misma configuración establecida por el GA, pero sin combinar sus resultados en un RS híbrido. Además, con la finalidad de mostrar la importancia de utilizar criterios apropiados, también se realizará un estudio comparativo con versiones mono-criterio de toda la información considerada. El proceso de evaluación de cada RS se basará en validación cruzada estratificada, como se describe en la sección 4.3 y se obtendrán las métricas basadas en desviación y en relevancia más relevantes que se definen en la sección 3.2. Adicionalmente también se calculará el tiempo de ejecución que cada RS necesita para construir el modelo y proporcionar las recomendaciones a un estudiante.

Los resultados finales se pueden ver en la tabla 5, donde se muestra en la primera fila el rendimiento del RS híbrido propuesto, en las 2 siguientes los RS multi-criterio no híbridos, en las 3 siguientes filas se muestran los resultados de los RS basado en CF y un único criterio (relativo al estudiante) en cada caso y en las 4 últimas filas se muestran los resultados de los RS basados en CBF y un único criterio (relativo a la asignatura) en cada caso.

Tabla 5: Comparativa de rendimiento entre RS

Modelo	RMSE	F1	nDCG	Alcance (%)	Tiempo (s)
RS híbrido	0.971	0.661	0.682	100.00	3.022
CF multi-criterio	1.123	0.695	0.714	79.30	1.582
CBF multi-criterio	1.206	0.215	0.199	100.00	1.324
CF de valoraciones	1.198	0.625	0.635	95.09	1.020
CF de calificaciones	1.347	0.518	0.534	96.14	1.014
CF de especialidad	1.221	0.642	0.644	88.42	0.250
CBF de profesores	2.608	0.234	0.291	100.00	0.785
CBF de contenido	1.224	0.234	0.234	100.00	1.874
CBF de competencias	1.229	0.187	0.184	100.00	0.833
CBF de área de conoc.	1.564	0.278	0.237	100.00	0.370

Los resultados demuestran la relevancia de usar un enfoque híbrido con múltiples criterios, cuyas estimaciones son significativamente menores que las del resto de modelos (esto es, un RMSE más bajo). Es más, el uso de múltiples criterios funciona mejor que las versiones mono-

criterio de cada uno. En general, se puede ver que la información de los estudiantes es la más útil para hacer recomendaciones, como demuestra el RMSE más bajo para CF que para CBF. En cuanto a la relevancia de las asignaturas finalmente recomendadas, tanto F1 como nDCG muestran que CBF tienen mal rendimiento lo que puede significar que aunque su estimación para valores de preferencia de muy bajos o muy altos no sea mala, sí que fallaría en valores intermedios que son los que determinarían si una asignatura es apta o no para recomendar. Por otro lado, CF tiene problemas para obtener recomendaciones para todos los estudiantes (alcance menor que 100 %) ya que algunos son demasiado diferentes al resto y no se puede construir un vecindario para ellos. En el RS híbrido se combinan las ventajas de CF y CBF, obteniendo un alcance del 100 % y recomendaciones más precisas. Por otro lado, hay que tener en cuenta que cuanto más complejo sea el modelo, más tiempo llevará que éste se construya y por tanto pueda ofrecer recomendaciones, algo que sería necesario optimizar en una eventual fase de producción.

También es importante notar que, aunque un criterio sea el más relevante en el enfoque multi-criterio, no quiere decir que vaya a obtener mejores resultados por separado. Por ejemplo: los profesores es un criterio relevante, pero hay pocas asignaturas impartidas por el mismo profesor. Por tanto, es necesario contar con otra información para ser capaz de recomendar nuevas asignaturas interesantes para el estudiante. De aquí la importancia de combinar y ponderar adecuadamente toda la información disponible en este sistema.

6

COMENTARIOS FINALES

6.1 CONCLUSIONES

En este Trabajo Fin de Máster se ha desarrollado un RS híbrido multicriterio para la recomendación de asignaturas en el ámbito de una carrera universitaria. El modelo propuesto combina tanto información del estudiante como de la asignatura utilizando varias herramientas como son el CF, el CBF y el análisis semántico. Se hace especial énfasis en cómo esta información se combina por medio de pesos configurables para determinar la relevancia de cada criterio. En esta línea, se han adaptado diversas técnicas de optimización genética con el fin de encontrar la que mejor se ajuste a las características del problema y, en última instancia, automatizar la búsqueda de una configuración óptima gracias a modelos interpretables en los que se puede controlar claramente la relevancia de cada criterio así como el resto de parámetros utilizados en el RS.

Los resultados experimentales muestran que, para las características del problema, el GA más adecuado es CHC gracias a su búsqueda adaptativa y la alta diversidad que introduce. En cuanto al rendimiento del RS, se ha demostrado que considerar varios criterios proporciona mejores resultados, pero es necesario estudiar la relevancia de cada uno de ellos ya que no todos los factores son igualmente relevantes. Además, el uso de sistemas híbridos como el presentado, que combina CF y CBF también optimiza los resultados alcanzados no sólo en una estimación más precisa de la recomendación de asignaturas a los estudiantes, sino relativa a otros aspectos tan importantes como puede ser la capacidad de obtener recomendaciones para todos los usuarios.

6.2 PUBLICACIONES ASOCIADAS

Este trabajo, como ya se ha mencionado, es la continuación del Trabajo de Fin de Grado *Algoritmos Multi-criterio y basados en Matriz de Factorización para la Recomendación de Asignaturas* [11], sobre el que se ha aumentado la información utilizada para las recomendaciones, se ha diseñado un sistema de evaluación del RS más preciso y se han presentado diversos métodos de optimización automática basada en GA. Así mismo, durante la investigación y desarrollo concernientes a este trabajo se han realizado las siguientes publicaciones: un congreso internacional [1], en el que se empieza a explorar la idea de la optimización de los pesos del RS mediante un SGEA, y un congreso nacional [2], en el que se introduce la optimización completa del RS mediante CHC y se amplía considerablemente el tamaño del conjunto de datos con el que se trabaja. Así mismo, se ha enviado un artículo para revista [3] en el que se profundiza en el estudio del trabajo relacionado, la definición del modelo y su optimización y se ha ampliado la experimentación y el estudio de pesos llevados a cabo.

[1] A. Esteban, A. Zafra y C. Romero, «A Hybrid Multi-Criteria approach using a Genetic Algorithm for Recommending Courses to University Students,» en *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)*, Buffalo, NY, 2018, págs. 273-279.

[2] A. Esteban, A. Zafra y C. Romero, «Un Sistema de Recomendación de Asignaturas Multi-Criterio con Optimización Genética,» en *Actas de la XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 18)*, Granada, Spain, 2018, págs. 802-807.

[3] A. Esteban, A. Zafra y C. Romero, «Helping university students to choose elective courses by using a Hybrid Multi-criteria Recommendation System with Genetic Optimization,» págs. 1-34, 2019.

6.3 TRABAJO FUTURO

La principal línea de trabajo futuro que se propone a esta investigación sería la aplicabilidad a un entorno real. Así, se propone la inclusión de restricciones a las recomendaciones que ayuden a los estudiantes a filtrar las asignaturas por curso, cuatrimestre u otros parámetros. También se propone la puesta en producción de una aplicación móvil que permita

a los estudiantes interactuar con el RS. Por medio de esta aplicación podría proveerse a los estudiantes no sólo de las recomendaciones de las asignaturas, sino también información de cómo se han generado estas recomendaciones, algo que sería inmediato gracias a la parametrización de criterios en base a la que se ha desarrollado este RS.

También se puede estudiar la extensión de los criterios considerados a más titulaciones, de forma que se pueda llegar a más estudiantes. El fin último de esta propuesta sería conseguir más datos de estudiantes reales para poder ampliar la experimentación y, en última instancia, poder extender las conclusiones a otras áreas educativas.

Otra interesante línea de investigación futura podría ser la inclusión del análisis de redes sociales para contemplar la confianza entre sus miembros a través de un mecanismo de reputación que capte las conexiones implícitas y explícitas entre éstos para mejorar las recomendaciones.

BIBLIOGRAFÍA

[1] J. Bobadilla, F. Ortega, A. Hernando y A. Gutiérrez, «Recommender systems survey,» *Knowledge-Based Systems*, vol. 46, págs. 109-132, 2013. DOI: 10.1016/j.knosys.2013.03.012. arXiv: arXiv:1011.1669v3.

[2] G. Linden, B. Smith y J. York, «Amazon. com recommendations: Item-to-item collaborative filtering,» *IEEE Internet computing*, nº 1, págs. 76-80, 2003.

[3] A. Sharma, J. M. Hofman y D. J. Watts, «Estimating the Causal Impact of Recommendation Systems from Observational Data,» en *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, ACM, 2015, págs. 453-470. DOI: 10.1145/2764468.2764488. arXiv: 1510.05569.

[4] C. Romero y S. Ventura, «Data mining in education,» *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, nº 1, págs. 12-27, 2013. DOI: 10.1002/widm.1075.

[5] O. Iatrellis, A. Kameas y P. Fitsilis, «Academic Advising Systems: A Systematic Literature Review of Empirical Evidence,» *Education Sciences*, vol. 7, nº 4, págs. 1-17, 2017. DOI: 10.3390/educsci7040090.

[6] N. Bendakir y E. Aïmeur, «Using association rules for course recommendation,» en *Proceedings of the AAAI Workshop on Educational Data Mining*, vol. 3, The AAAI Press, 2006, págs. 31-40.

[7] J. Han, J. Jo, H. Ji y H. Lim, «A collaborative recommender system for learning courses considering the relevance of a learner's learning skills,» *Cluster Computing*, vol. 19, nº 4, págs. 2273-2284, 2016. DOI: 10.1007/s10586-016-0670-x.

[8] D. Shah, P. Shah y A. Banerjee, «Similarity based regularization for online matrix-factorization problem: An application to course recommender systems,» en *Region 10 Annual International Conference TENCON*, Penang, Malaysia: IEEE, 2017, págs. 1874-1879. DOI: 10.1109/TENCON.2017.8228164.

[9] F. Carballo, «Masters ' Courses Recommendation : Exploring Collaborative Filtering and Singular Value Decomposition with Student Profiling,» Master Thesis, Técnico Lisboa, 2014, págs. 1-71.

[10] F. Le Roux, E. Ranjeet, V. Ghai, Y. Gao y J. Lu, «A Course Recommender System Using Multiple Criteria Decision Making Method,» en *International Conference on Intelligent Systems and Knowledge Engineering*, ép. Advances in Intelligent Systems Research, Atlantis Press, 2007, págs. 346-350. DOI: 10.2991/iske.2007.238LB-leroux2007.

[11] A. Esteban, «Algoritmos Multi-criterio y basados en Matriz de Factorización para la Recomendación de Asignaturas,» TFG, Universidad de Córdoba, 2017, págs. 1-144.

[12] M. Srinivas y L. M. Patnaik, «Genetic algorithms: A survey,» *Computer*, vol. 27, nº 6, págs. 17-26, 1994. DOI: 10.1109/2.294849.

[13] L. J. Eshelman, «The CHC Adaptive Search Algorithm: How to Have Safe Search When Engaging in Nontraditional Genetic Recombination,» *Foundations of genetic algorithms*, vol. 1, págs. 265-283, 1991. DOI: 10.1016/B978-0-08-050684-5.50020-3.

[14] A. Petrowski, «A clearing procedure as a niching method for genetic algorithms,» *Proceedings of IEEE International Conference on Evolutionary Computation ICEC-96*, págs. 798-803, 1996. DOI: 10.1109/ICEC.1996.542703.

[15] A. Peña-Ayala, «Educational data mining: A survey and a data mining-based analysis of recent works,» *Expert Systems with Applications*, vol. 41, nº 4, págs. 1432-1462, 2014. DOI: 10.1016/j.eswa.2013.08.042. arXiv: 1503.05296.

[16] H. F. Unelsrød, «Design and Evaluation of a Recommender System for Course Selection,» Master Thesis, Norwegian University of Science y Technology, 2011, págs. 1-56. DOI: 10.1016/j.iheduc.2011.01.001.

[17] P. C. Chang, C. H. Lin y M. H. Chen, «A hybrid course recommendation system by integrating collaborative filtering and artificial immune systems,» *Algorithms*, vol. 9, nº 3, págs. 1-18, 2016. DOI: 10.3390/a9030047.

[18] K. Taha, «Automatic Academic Advisor,» en *8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, Pittsburgh, PA, USA: IEEE, 2012, págs. 262-268. DOI: 10.4108/icst.collaboratecom.2012.250338.

[19] B. Bakhshinategh, G. Spanakis, O. Zaiane y S. ElAtia, «A Course Recommender System based on Graduating Attributes,» en *Proceedings of the 9th International Conference on Computer Supported Education (CSEDU)*, vol. 1, SCITEPRESS, 2017, págs. 347-354. DOI: 10.5220/0006318803470354.

[20] K. Ganeshan y X. Li, «An intelligent student advising system using collaborative filtering,» en *Proceedings - Frontiers in Education Conference (FIE)*, El Paso, TX, USA: IEEE, 2015, págs. 1-8. DOI: 10.1109/FIE.2015.7344381.

[21] L. Mostafa, G. Oately, N. Khalifa y W. Rabie, «A Case based Reasoning System for Academic Advising in Egyptian Educational Institutions,» en *2nd International Conference on Research in Science, Engineering and Technology (ICRSET'2014)*, Dubai, UAE, 2014, págs. 5-10. DOI: 10.15242/IIE.E0314513.

[22] C. Y. Huang, R. C. Chen y L. S. Chen, «Course-recommendation system based on ontology,» en *International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 3, Tianjin, China: IEEE, 2013, págs. 1168-1173. DOI: 10.1109/ICMLC.2013.6890767.

[23] H. Ma, X. Wang, J. Hou e Y. Lu, «Course recommendation based on semantic similarity analysis,» en *3rd International Conference on Control Science and Systems Engineering (ICCSSE)*, Beijing, China: IEEE, 2017, págs. 638-641. DOI: 10.1109/CCSSE.2017.8088011.

[24] O. Daramola, O. Emebo, I. Afolabi y C. Ayo, «Implementation of an Intelligent Course Advisory Expert System,» *International Journal of Advanced Research in Artificial Intelligence*, vol. 3, nº 5, págs. 6-12, 2014. DOI: 10.14569/IJARAI.2014.030502.

[25] A. Al-Badarenah y J. Alsakran, «An Automated Recommender System for Course Selection,» *International Journal of Advanced Computer Science and Applications*, vol. 7, nº 3, págs. 166-175, 2016. DOI: 10.14569/IJACSA.2016.070323.

[26] Z. Gulzar, A. A. Leema y G. Deepak, «PCRS: Personalized Course Recommender System Based on Hybrid Approach,» *Procedia Computer Science*, vol. 125, nº The 6th International Conference on Smart Computing and Communications, págs. 518-524, 2018. DOI: 10.1016/j.procs.2017.12.067.

[27] R. S. Abdulwahhab, H. S. Al Makhmari y S. N. Al Battashi, «An educational web application for academic advising,» en *8th GCC Conference and Exhibition (GCCCE 2015)*, IEEE, 2015, págs. 1-6. DOI: 10.1109/IEEEGCC.2015.7060084.

[28] C.-S. Hwang, «Genetic Algorithms for Feature Weighting in Multi-criteria Recommender Systems,» *Journal of Convergence Information Technology*, vol. 5, nº 8, págs. 126-136, 2010. DOI: 10.4156/jcit.vol5.issue8.13.

[29] S. Owen, R. Anil, T. Dunning y E. Friedman, «Recommendations,» en *Mahout in Action*. Manning Publications Co., 2012, cap. Part 2, págs. 11-114.

[30] M. Balabanović e Y. Shoham, «Fab: content-based, collaborative recommendation,» *Communications of the ACM*, vol. 40, nº 3, págs. 66-72, 1997.

[31] R. Burke, «Hybrid Recommender Systems: Survey and Experiments,» *User Modeling and User-Adapted Interaction*, vol. 12, págs. 331-370, 2002. DOI: 10.1023/A:1021240730564.

[32] J. L. Herlocker, J. A. Konstan, L. G. Terveen y J. T. Riedl, «Evaluating collaborative filtering recommender systems,» *ACM Transactions on Information Systems (TOIS)*, vol. 22, nº 1, págs. 5-53, 2004.

[33] M. McCandless, E. Hatcher y O. Gospodnetic, «Core Lucene,» en *Lucene in action: covers Apache Lucene 3.0*, Manning Publications Co., 2010, cap. Part 1, págs. 36-255.

[34] D. E. Goldberg, «Genetic algorithms in search, optimization, and machine learning,» *Choice Reviews Online*, vol. 27, nº 02, págs. 27–0936–27-0936, 1989. DOI: 10.5860/CHOICE.27-0936. arXiv: 1304.4595 [quant-ph].

[35] S. Mahfoud y E. Bay, «Niching methods for genetic algorithms,» *Urbana*, nº 95001, pág. 251, 1995. DOI: http://dx.doi.org/10.1016/j.fertnstert.2011.05.079.

[36] E. Özcan y C. Başaran, «A case study of memetic algorithms for constraint optimization,» *Soft Computing*, vol. 13, nº 8-9, págs. 871-882, 2009. DOI: 10.1007/s00500-008-0354-4.

[37] J. Derrac, S. García y F. Herrera, «A First Study on the Use of Coevolutionary Algorithms for Instance and Feature Selection,» en *International Conference on Hybrid Artificial Intelligence Systems*, E. Corchado, X. Wu, E. Oja, Á. Herrero y B. Baruque, eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, págs. 557-564.

[38] D. Peralta, S. del Río, S. Ramírez-Gallego, I. Triguero, J. M. Benitez y F. Herrera, «Evolutionary feature selection for big data classification: A mapreduce approach,» *Mathematical Problems in Engineering*, vol. 2015, págs. 1-12, 2015. DOI: 10.1155/2015/246139.

[39] S. Ventura, C. Romero, A. Zafra, J. A. Delgado y C. Hervás, «JCLEC: A Java framework for evolutionary computation,» *Soft Computing*, vol. 12, nº 4, págs. 381-392, 2008. DOI: 10.1007/s00500-007-0172-0.

A

RESUMEN DE LOS DATOS RECOLECTADOS

En este apéndice se muestra un resumen de los datos relativos a las valoraciones que se han recolectados a través de encuestas a estudiantes del Grado en Ingeniería Informática de la Universidad de Córdoba durante 3 años académicos seguidos (2016 a 2018). En la tabla 6 se lista cada asignatura de la que se tienen valoraciones (anonimizada para evitar mostrar datos sensibles), junto al curso al que pertenece, el número de valoraciones que ha recibido, la valoración media y la calificación media que los estudiantes encuestados obtuvieron, el área de conocimiento al que pertenece, el número de profesores que imparten docencia en ella, el número de competencias que se le relacionan y, por último, la longitud del texto que describe sus contenidos teóricos y prácticos. En la tabla 7 se muestra información relativa a los estudiantes encuestados: el año de realización de la encuesta, las especialidades de la titulación, el número de estudiantes pertenecientes a cada especialidad por año y el número de valoraciones recogidas por cada especialidad y año.

Tabla 6: Resumen de los datos relativos a las asignaturas

ID	Curso	N Val.	Val. Media	Cal. Media	Área	N Prof	N Comp	Long. Cont
asig02	1	95	4.18	7.23	1	3	3	112
asig09	1	95	4.21	6.72	1	3	3	1481
asig01	1	94	4.50	8.04	5	5	1	667
asig04	1	94	3.10	6.59	1	3	5	821
asig11	2	94	3.77	7.84	1	4	1	621
asig03	1	93	2.73	5.65	2	5	4	1301
asig05	1	93	2.35	6.30	2	4	4	320
asig06	1	93	4.40	7.91	5	4	1	1060
asig18	2	93	1.96	7.13	1	8	1	1393
asig07	1	92	3.34	7.16	4	4	2	939
asig13	2	92	3.51	6.66	1	3	1	1061

asig14	2	92	2.51	6.00	1	5	1	885
asig08	1	91	2.79	6.81	2	3	2	2015
asig10	1	91	3.69	6.94	1	3	3	1453
asig17	2	90	3.78	7.44	3	3	1	820
asig19	2	90	3.88	6.96	5	3	1	1351
asig12	2	86	2.00	4.38	2	5	1	1069
asig15	2	86	3.86	6.48	3	3	2	1169
asig20	2	83	2.94	5.43	2	3	2	1267
asig21	3	82	2.72	5.87	2	4	1	433
asig22	3	80	4.25	7.49	1	6	1	1737
asig16	2	79	4.06	6.84	2	2	1	698
asig23	3	76	3.87	7.03	2	7	1	2683
asig24	3	58	3.81	8.32	1	5	6	3181
asig42	3	47	3.81	7.25	1	3	1	1125
asig41	3	46	3.39	7.27	1	3	1	1554
asig43	3	30	3.30	7.79	2	6	1	533
asig26	3	26	4.12	8.00	1	13	1	563
asig44	3	26	3.58	7.54	2	4	1	928
asig45	3	25	3.28	7.14	1	2	1	300
asig46	3	25	4.16	7.18	1	4	1	1145
asig27	3	23	3.70	6.30	1	5	1	627
asig29	3	23	2.43	7.32	1	9	1	547
asig25	3	19	3.16	5.98	1	3	1	1514
asig28	3	19	1.95	6.46	1	6	1	872
asig30	3	17	3.53	6.61	1	13	1	452
asig32	4	7	3.29	8.29	1	10	1	452
asig47	4	5	3.20	7.78	1	4	1	313
asig48	4	5	4.00	5.85	2	4	1	638
asig50	4	5	3.80	7.80	1	9	6	616
asig52	4	5	4.40	8.36	1	3	3	1568
asig54	4	4	3.25	7.73	1	5	1	1821
asig55	4	4	4.50	8.43	1	2	8	774
asig59	4	4	3.75	7.08	1	3	1	505
asig31	4	3	4.00	7.67	1	3	1	760
asig57	4	3	3.67	7.33	2	3	1	588
asig33	3	2	4.00	8.30	1	6	2	457
asig34	3	2	3.50	5.45	2	11	2	667
asig62	4	2	4.00	8.00	1	6	1	431
asig63	4	2	3.50	3.75	2	7	2	602
asig35	3	1	4.00	7.80	1	3	2	1920
asig36	3	1	5.00	7.40	2	6	2	2722
asig38	3	1	3.00	7.30	2	8	1	783
asig39	4	1	2.00	6.00	1	10	2	997

asig40	4	1	5.00	7.10	2	5	2	492
asig51	4	1	5.00	9.50	1	6	1	334
asig53	4	1	5.00	10.00	2	4	7	1074
asig56	4	1	5.00	10.00	1	3	3	487

Tabla 7: Resumen de los datos relativos a los estudiantes

Año	Especialidad	N Estudiantes	N Valoraciones
2016	Vacío	1	17
2016	Computación	21	610
2016	Software	3	72
2016	Computadores	3	61
2017	Computación	12	296
2017	Software	23	644
2018	Computación	20	519
2018	Software	6	145
2018	Computadores	6	135

B

RESULTADOS DE EXPERIMENTACIÓN

En este apéndice se muestran los resultados completos de las pruebas realizadas para determinar cómo influye cada parametro en cada uno de los GA estudiados.

Tabla 8: Análisis de sensibilidad para SGEA

N GEN.	TAM. POB.	P(CRU)	P(MUT)	N MEJ.	BEST F.	AVG. F.	VAR. F.
100	50	0.9	0.1	3	1.013 62	1.013 62	-3.109×10^{-15}
100	50	0.9	0.1	5	1.013 62	1.013 62	1.776×10^{-15}
100	50	0.9	0.1	7	1.013 26	1.013 26	2.442×10^{-15}
100	50	0.9	0.1	9	1.013 26	1.013 26	-1.554×10^{-15}
100	50	0.9	0.3	3	1.013 62	1.013 62	2.220×10^{-16}
100	50	0.9	0.3	5	1.013 62	1.013 62	6.661×10^{-16}
100	50	0.9	0.3	7	1.013 26	1.013 26	2.442×10^{-15}
100	50	0.9	0.3	9	1.013 26	1.013 26	1.998×10^{-15}
100	50	0.9	0.5	3	1.013 62	1.013 62	2.220×10^{-16}
100	50	0.9	0.5	5	1.013 26	1.013 26	2.442×10^{-15}
100	50	0.9	0.5	7	1.013 26	1.013 26	1.998×10^{-15}
100	50	0.9	0.5	9	1.013 26	1.013 26	2.442×10^{-15}
100	50	0.9	0.7	3	1.013 62	1.013 62	-4.441×10^{-16}
100	50	0.9	0.7	5	1.013 62	1.013 62	6.661×10^{-16}
100	50	0.9	0.7	7	1.013 26	1.013 26	1.998×10^{-15}
100	50	0.9	0.7	9	1.013 26	1.013 26	-3.109×10^{-15}
100	50	0.9	0.2	7	1.013 26	1.013 26	1.554×10^{-15}
100	50	0.9	0.4	7	1.013 26	1.013 26	2.442×10^{-15}
100	50	0.9	0.6	7	1.013 26	1.013 26	1.998×10^{-15}
100	50	0.9	0.8	7	1.013 26	1.013 26	1.998×10^{-15}
100	50	0.9	0.9	7	1.013 26	1.013 26	1.332×10^{-15}
100	50	0.9	0.5	1	1.013 97	1.013 97	-1.554×10^{-15}
100	50	0.9	0.5	2	1.013 62	1.013 62	1.776×10^{-15}
100	50	0.9	0.5	4	1.013 62	1.013 62	-3.109×10^{-15}

100	50	0.9	0.5	6	1.013 26	1.013 26	-3.109×10^{-15}
100	50	0.9	0.5	8	1.013 62	1.013 62	6.661×10^{-16}
100	50	0.9	0.5	9	1.013 26	1.013 26	2.442×10^{-15}

Tabla 9: Análisis de sensibilidad para CHC

N GEN.	TAM. POB.	P(CRU)	D	N SUP.	BEST F.	AVG. F.	VAR. F.
100	50	0.9	2	5	0.969 44	1.045 85	0.001
100	50	0.9	2	10	0.969 42	0.969 61	2.908×10^{-9}
100	50	0.9	2	15	0.985 18	0.985 18	-1.443×10^{-15}
100	50	0.9	2	20	0.983 22	0.983 22	2.331×10^{-15}
100	50	0.9	4	5	0.971 72	0.976 62	8.525×10^{-6}
100	50	0.9	4	10	0.977 42	0.982 20	6.954×10^{-5}
100	50	0.9	4	15	0.970 18	0.971 07	1.154×10^{-7}
100	50	0.9	4	20	0.981 08	0.983 49	3.050×10^{-6}
100	50	0.9	6	5	0.971 10	1.039 47	0.001
100	50	0.9	6	10	0.971 10	1.038 43	0.002
100	50	0.9	6	15	0.971 10	0.971 17	3.259×10^{-8}
100	50	0.9	6	20	0.971 10	1.005 33	0.002
100	50	0.9	8	5	0.973 87	0.980 78	2.557×10^{-6}
100	50	0.9	8	10	0.977 63	0.982 54	3.766×10^{-6}
100	50	0.9	8	15	0.973 27	0.980 81	3.058×10^{-6}
100	50	0.9	8	20	0.979 95	0.981 46	2.477×10^{-7}
100	50	0.9	4	3	0.975 40	0.976 79	1.306×10^{-6}
100	50	0.9	4	8	0.975 58	0.981 58	9.957×10^{-5}
100	50	0.9	4	13	0.969 54	0.977 00	0.000
100	50	0.9	4	17	0.978 75	0.979 09	1.462×10^{-8}
100	50	0.9	10	5	0.999 76	1.001 83	1.134×10^{-6}
100	50	0.9	12	5	0.997 92	0.998 18	3.444×10^{-8}
100	50	0.9	14	5	0.985 55	0.990 94	4.921×10^{-6}
100	50	0.9	16	5	1.003 73	1.005 29	3.664×10^{-7}

Tabla 10: Análisis de sensibilidad para *Clearing*

N GEN.	TAM. POB.	P(CRU)	P(MUT)	KAPPA	BEST F.	AVG. F.	VAR. F.
100	50	0.9	0.1	5	1.023 15	1.070 43	4.563×10^{-5}
100	50	0.9	0.3	5	1.023 15	1.070 43	4.563×10^{-5}
100	50	0.9	0.5	5	1.023 15	1.081 13	6.861×10^{-5}
100	50	0.9	0.7	5	1.023 15	1.066 72	3.873×10^{-5}
100	50	0.9	0.1	15	1.023 15	1.038 09	9.295×10^{-6}
100	50	0.9	0.3	15	1.023 15	1.026 28	4.070×10^{-7}

100	50	0.9	0.5	15	1.023 15	1.038 09	9.295×10^{-6}
100	50	0.9	0.7	15	1.023 15	1.026 28	4.070×10^{-7}
100	50	0.9	0.1	25	1.023 15	1.027 51	4.880×10^{-5}
100	50	0.9	0.3	25	1.023 15	1.024 26	2.379×10^{-6}
100	50	0.9	0.5	25	1.023 15	1.024 13	2.226×10^{-6}
100	50	0.9	0.7	25	1.023 15	1.023 15	8.882×10^{-16}
100	50	0.9	0.1	35	1.023 15	1.023 15	8.882×10^{-16}
100	50	0.9	0.3	35	1.023 15	1.023 15	8.882×10^{-16}
100	50	0.9	0.5	35	1.023 15	1.023 15	1.110×10^{-15}
100	50	0.9	0.7	35	1.023 15	1.023 15	1.110×10^{-15}
100	50	0.9	0.2	25	1.023 15	1.023 15	8.882×10^{-16}
100	50	0.9	0.4	25	1.023 15	1.023 15	1.332×10^{-15}
100	50	0.9	0.6	25	1.023 15	1.023 15	1.332×10^{-15}
100	50	0.9	0.8	25	1.023 15	1.023 15	8.882×10^{-16}
100	50	0.9	0.9	25	1.023 15	1.023 15	8.882×10^{-16}
100	50	0.9	0.5	10	1.023 15	1.023 15	8.882×10^{-16}
100	50	0.9	0.5	20	1.023 15	1.023 15	1.332×10^{-15}
100	50	0.9	0.5	30	1.023 15	1.023 15	1.332×10^{-15}

Tabla 11: Estudio del rango permitido para los genes de los pesos

Rango	N gen.	Tam. pob.	GA	Best f.	Avg. f.	Var. f.
0-10	100	50	CHC	0.974 01	0.975 46	1.950×10^{-6}
0-20	100	50	CHC	0.984 25	0.987 40	1.806×10^{-6}
0-30	100	50	CHC	0.971 07	0.971 07	2.220×10^{-16}
0-40	100	50	CHC	0.976 47	0.977 65	1.964×10^{-6}
0-50	100	50	CHC	0.965 34	0.966 19	2.697×10^{-7}
0-60	100	50	CHC	0.984 68	0.984 95	1.099×10^{-7}
0-70	100	50	CHC	0.971 72	0.976 62	8.525×10^{-6}
0-80	100	50	CHC	0.967 55	0.967 62	1.661×10^{-9}