Revista de Matemática: Teoría y Aplicaciones
ISSN: 1409-2433
mta.cimpa@ucr.ac.cr
Universidad de Costa Rica
Costa Rica

FLORES–CRUZ, JORGE; LARA–VELÁZQUEZ, PEDRO; GUTIÉRREZ–ANDRADE, MIGUEL A.; DE-LOS-COBOS–SILVA, SERGIO G.; RINCÓN–GARCÍA, ERIC A.

UN SISTEMA CLASIFICADOR UTILIZANDO COLORACIÓN DE GRÁFICAS SUAVES

Revista de Matemática: Teoría y Aplicaciones, vol. 24, núm. 1, enero, 2017, pp. 129-156
Universidad de Costa Rica
San José, Costa Rica

REVISTA DE MATEMÁTICA: TEORÍA Y APLICACIONES 2017 **24**(1) : 129–156
CIMPA – UCR ISSN: 1409-2433 (PRINT), 2215-3373 (ONLINE)

UN SISTEMA CLASIFICADOR UTILIZANDO COLORACIÓN DE GRÁFICAS SUAVES

A CLASSIFIER SYSTEM USING SMOOTH GRAPH COLORING

JORGE FLORES–CRUZ* PEDRO LARA–VELÁZQUEZ†
MIGUEL A. GUTIÉRREZ–ANDRADE‡
SERGIO G. DE-LOS-COBOS–SILVA§ ERIC A. RINCÓN–GARCÍA¶

Received: 3/Mar/2016; Revised: 26/Aug/2016;
Accepted: 7/Oct/2016

*Universidad Autónoma Metropolitana-Iztapalapa, Posgrado en Ciencias y Tecnologías de la Información, Av. San Rafael Atlixco 186, Col. Vicentina, Del. Iztapalapa, México D.F., C.P. 09340, Mexico. E-Mail: jorge.floresc@xanum.uam.mx

†Misma dirección que/*Same address as*: J. Flores-Cruz. E-Mail: plara@xanum.uam.mx

‡Misma dirección que/*Same address as*: J. Flores-Cruz. E-Mail: gamma@xanum.uam.mx

§Misma dirección que/*Same address as*: J. Flores-Cruz. E-Mail: cobos@xanum.uam.mx

¶Universidad Autónoma Metropolitana-Azcapotzalco, Departamento de Sistemas, Av. San Pablo 180, Colonia Reynosa Tamaulipas, México D.F., C.P. 02200, Mexico. E-Mail: rigaeral@correo.azc.uam.mx

Resumen

Los clasificadores no supervisados permiten un método de agrupación de forma automatizada. Para ello es deseable agrupar los elementos con un menor procesamiento de datos. Este trabajo propone un sistema clasificador no supervisado que utiliza el modelo de coloración de gráficas suaves. El método se puso a prueba con algunas instancias clásicas de la literatura especializada y se comparan los resultados obtenidos con clasificaciones hechas con clasificadores supervisados, obteniéndose resultados tan buenos o mejores que con los clasificadores más aceptados y utilizados, proporcionando a veces clasificaciones alternativas que muestran información adicional que los humanos no consideraron.

Palabras clave: coloración suave; clasificación no supervisada; clasificación automática; agrupación; optimización.

Abstract

Unsupervised classifiers allow clustering methods with less or no human intervention. Therefore it is desirable to group the set of items with less data processing. This paper proposes an unsupervised classifier system using the model of soft graph coloring. This method was tested with some classic instances in the literature and the results obtained were compared with classifications made with human intervention, yielding as good or better results than supervised classifiers, sometimes providing alternative classifications that considers additional information that humans did not considered.

Keywords: soft coloring; unsupervised classification; clustering; optimization.

Mathematics Subject Classification: 90C90, 90C10, 05C15.

1 Introducción

Una de las aplicaciones de la Inteligencia Artificial consiste en el reconocimiento de patrones, que se puede entender como clasificar grandes cantidades de objetos físicos o abstractos con el propósito de extraer información útil, que permita establecer propiedades entre agrupaciones de dichos objetos.

Las razones para automatizar este proceso van desde hacer actividades autoadaptables, es decir, aplicaciones capaces de adaptarse a distintos tipos de problemas y que van aprendiendo a lo largo de su ejecución sin requerir de la intervención humana para su aprendizaje, hasta actividades tediosas, tardadas o hasta imposibles para una persona o un grupo de personas. El flujo de clasificación

consiste en la obtención de patrones mediante un proceso de segmentación, extracción de características y reseña de cada objeto como una colección de descriptores.

La ciencia y la tecnología se han valido del reconocimiento de patrones para diagnosticar enfermedades, detectar seres submarinos en un sonar, identificar clientes propensos a buró de crédito, entre otras aplicaciones. Los sistemas clasificadores son un tipo especial de reconocimiento de patrones, éstos colocan una etiqueta a un objeto de acuerdo a sus características; ya sea para personas, otros seres vivos o para objetos inanimados. Los seres humanos hacemos reconocimiento de patrones cotidianamente incluso de forma inconsciente.

El reto de un modelo de reconocimiento de patrones es enseñarle a una máquina, en este caso una computadora, a hacerlo de forma eficiente sin muchos parámetros. En un sistema clasificador, la etiqueta puede ser determinada previamente, luego se tiene que decidir para un objeto dado, en que clase particular es más adecuado colocarlo dadas las opciones existentes. Por ejemplo, un servidor debe decidir si un correo entrante es "spam" o se dirige a la bandeja de entrada. Hay veces donde las clases no están definidas con anterioridad y es necesario encontrar la forma de agrupar los elementos en alguna forma "óptima".

La coloración de gráficas suaves es una generalización del problema de coloración [5] en el que se busca encontrar una coloración que minimice la dureza en la gráfica, o dicho de otra forma, reducir la suma de distancias entre vértices con colores idénticos. Este modelo se utiliza en la programación de eventos susceptibles a cambios, en la asignación estable de frecuencias del espectro electromagnético, por ejemplo. Se ha demostrado que es un problema de tipo NP-difícil [11]. Para grafos de orden menor o igual a 20, se pueden utilizar algoritmos exactos que resuelven el problema; en caso contrario es necesario el uso de técnicas heurísticas con aproximaciones bastante aceptables [8].

El reconocimiento de patrones se refiere al problema de encontrar la estructura interna de los objetos etiquetados. Una de las formas para llevar a cabo la tarea es definir una distancia entre dos objetos, de esta manera si dos objetos están muy cerca respecto a una métrica dada, es muy probable que estén en la misma clase. Si los objetos están lejos uno del otro, van a estar en clases diferentes. La métrica más utilizada es la distancia euclidiana [5].

En este trabajo se usa el problema de coloración de gráficas suaves para clasificar diversas bases de datos clásicas de la literatura. En la Sección 2 se exponen las bases teóricas de un clasificador; en la Sección 3 se explica la metodología de la propuesta; la evaluación de algunas bases de datos se detalla en la Sección 4; mientras que el posicionamiento del clasificador propuesto respecto a otros trabajos se explica en la Sección 5; finalmente se presentan las conclusiones.

2 Conceptos y definiciones

2.1 Clasificador no supervisado (agrupador)

Un sistema clasificador es la abstracción informática de la habilidad humana de reconocer clases o categorías de varios objetos, ya sean caras, voces, letras, seres vivos, etc. Para la tarea que implica el reconocimiento de patrones, se busca clasificar ciertos objetos en una de las k categorías preestablecidas. Resultando la clase a la que pertenece el objeto o en algunos casos no poder determinar la clase a la que se llega.

Al buscar problemas cuya solución óptima resulta muy complicada, se recurre a métodos heurísticos, tratando de encontrar soluciones parciales o aproximadas pero que en la práctica obtengan buenos resultados. Para alcanzar una buena solución, los programadores deben introducir un conocimiento previo del problema para facilitar la solución, reduciendo el costo computacional de la búsqueda de solución [14].

El Aprendizaje Computacional es una rama de la Inteligencia Artificial que busca construir un modelo matemático para modelar el proceso cognitivo [12], el cual es descrito como un macro-teoría que provee el entorno para estudiar el comportamiento de los algoritmos bajo distintos tipos de entrenamiento o de datos de entrada.

En el aprendizaje no supervisado no hay clases predeterminadas. El objetivo es encontrar la clase más adecuada para el conjunto sin etiquetar. Utilizando el grado de similitud en sus atributos podemos determinar si los objetos van a una misma clase o en caso de encontrarse diferencias, colocarlos en distintas clases.

2.2 Tipos de clasificadores

Se han construido varios clasificadores aplicando diversos tipos de aprendizaje, entre los cuales destacan:

Clasificador Bayes Ingenuo: Aplica el teorema de Bayes de probabilidad condicional como modelo de clasificación con algoritmo simplificado pero que contiene las características más importantes. Se le llama clasificador ingenuo porque asume que hay independencia probabilística entre cada atributo o característica de los objetos; un valor de característica no influye en el valor de otra característica o atributo. Aún con esa limitante es un clasificador muy efectivo que en la práctica da resultados bastante aceptables y no calcula probabilidades cero siempre y cuando los valores de cada clase se presenten en el conjunto de entrenamiento [10].

Clasificador Parzen: Mientras algunos clasificadores buscan establecer distancias entre los objetos mediante métricas —distancias entre los objetos— de similitud o disimilitud, hay clasificadores que buscan encontrar las distancias entre objetos a las líneas de características. El concepto de líneas de características permite generalizar el espacio de similitudes o disimilitudes y una vez obtenido el espacio estimar densidades de probabilidad mediante el método Parzen [15] y hacer uso de ellas para la clasificación de patrones.

Clasificador Red Neuronal: Tratando de simular el comportamiento de una neurona biológica, se sintetizan los componentes principales que son las dendritas o entradas, el cuerpo de la neurona y los axones o salidas. En la parte final de cada axón se encuentra la conexión con dendritas de otra neurona, a este proceso de comunicación se llama sinapsis. Utilizando el Modelo de Perceptrón Multicapa (PMC por sus siglas en inglés) para clasificación tenemos un conjunto de señales de entrada multiplicadas por sus pesos correspondientes y luego sumadas a cada una de las nuevas entradas en el proceso de sinapsis o nivel de activación de la neurona [9].

2.3 Tipos de agrupadores [24]

A forma de mención, se mostrarán los tipos de algoritmos agrupadores de mayor relevancia, posteriormente se explicarán a detalle:

1. Por medidas de similitud y distancia.
2. Jerárquicos, (a) Aglomerativos, (b) Divisivos.
3. Basados en error cuadrático.
4. Estimación por mezcla de densidades.
5. Basados en teoría de grafos.
6. Basados en técnicas de búsqueda combinatoria.
7. De lógica borrosa.
8. Basados en redes neuronales.
9. Agrupamiento basado en núcleo.
10. Agrupamiento de datos secuenciales.
11. Agrupamiento de gran cantidad de datos.
12. Exploración de datos multidimensionales.

2.4 Coloración de gráficas suaves

Los algoritmos de agrupamiento o clústering son un tipo de clasificadores no supervisados, y consisten en métodos que dividen un conjunto de n datos dentro de k grupos de tal modo que los miembros de un mismo grupo sean lo más parecidos entre ellos y lo más diferentes a los miembros de otros grupos [16]. El número de grupos k puede estar predeterminado o puede ser decidido por el algoritmo. Formalmente un algoritmo de agrupamiento produce un mapeo $C : \{1, \ldots, n\} \longrightarrow \{i, \ldots, k\}$ asociando cada objeto con un grupo.

Dentro de los algoritmos de agrupamiento, la coloración de gráficas suaves es una generalización del problema de coloración o de etiquetado de grafos. Consiste en colocar una etiqueta o color a un vértice del grafo de tal modo que ningún vértice adyacente contenga el mismo color [5].

Al ser una generalización del problema de coloración robusta, se sabe que este problema es del tipo NP-difícil y se necesita de técnicas metaheurísticas para gráficas completas.

En ella se busca reducir la dureza de la coloración, es decir, una coloración que minimice la suma de las penalizaciones entre las aristas adyacentes con vértices del mismo color. Para lo cual consideraremos lo siguiente. Sea un grafo completo no dirigido $G = (V, E)$ con un conjunto de vértices $V = \{1, 2, \ldots, n\}$ y un conjunto de aristas (i, j) comportándose de la siguiente manera[11].

$$G = (V, E); |V| = n; |E| = n(n-1)/2.$$

Existe una penalización por arista (i, j), denotada por p_{ij} tal que:

$$p_{ij} \geq 0, \forall (i, j) \in E.$$

Una función de coloración de vértices del grafo con k colores que sirven como etiquetado del vértice i se define como:

$$C^k : 1, 2, \ldots, n \longrightarrow 1, 2, \ldots, k.$$

Para una coloración C^k en el grafo, la función de dureza está dada por la suma de las penalizaciones con el mismo color en los extremos, es decir:

$$H(C^k) = \sum_{(i,j)\in E, C^k(i)=C^k(j)} p_{ij}.$$

El objetivo es encontrar el óptimo C^k_{op} que minimice la dureza $H(C^k_{op})$.

2.5 Solidez de una coloración

Dada una coloración de k colores sobre un grafo completo de orden n, el promedio de los vértices etiquetados m viene dado por $m = n/k$ y el número de aristas que comparten los m vértices viene dado por $C(m, 2) = m(m-1)/2$ con un número promedio de penalizaciones proporcional al promedio de vértices pintados con el mismo color multiplicado por el número de colores. Entonces la función de solidez de una coloración se define como la dureza de un grafo dividido por el número medio de aristas que contribuyen a la dureza, es decir:

$$S(C_{op}^k) = \frac{H(C_{op}^k)}{km(m-1)/2} = \frac{2H(C_{op}^k)}{k(n/k(n/k-1))} = \frac{2kH(C_{op}^k)}{n(n-k)}.$$

2.6 Resiliencia de una coloración

La resiliencia de una coloración C^k consiste en el porcentaje en que disminuye la solidez de una coloración con $k-1$ colores respecto a una con k colores, expresado como:

$$R(C_{op}^k) = \frac{S(C_{op}^{k-1}) - S(C_{op}^k)}{S(C_{op}^k)}.$$

Si obtenemos la resiliencia de todas las coloraciones posibles desde 1 hasta k, los valores más grandes permitirán las mejores opciones de número de clases que se pueden utilizar para clasificar un conjunto [11].

3 Modelo propuesto

Para generar la matriz de distancias de una cierta base de datos, se diseñó el siguiente procedimiento.

1. Limpieza de datos faltantes.

   Para implementar el modelo de agrupamiento, las bases de datos de prueba fueron sometidas primero a un proceso de limpieza de datos faltantes, el modelo es sensible al ruido por lo que se excluyeron todos las instancias con al menos un dato faltante de cualquier columna.

2. Ponderación de columnas alfanuméricas.

   En algunos casos había columnas con valores alfanuméricos, por ejemplo escalas de “malo, bueno y excelente” o “paciente vivo y paciente muerto”. Como el modelo trabaja exclusivamente con datos numéricos, se otorgó un valor numérico a cada uno de los datos, por ejemplo, malo se sustituyó

por un 0 y se siguió con la sucesión hasta excelente con un 2, vivo por un 1 y muerto por un 0 etc.

3. Normalización de los datos.

   Algunas columnas cuentan con escalas mayores a otras, por ejemplo hay columnas con métrica en miligramos (dosis) y otras columnas con métrica en kilogramos (peso del paciente), el modelo considera que todas las columnas tienen la misma influencia para la clasificación, así que, para evitar que la magnitud de cualquier atributo influya más que los otros, todas las bases de datos se sometieron a un proceso previo de normalización.

   La normalización consiste en tomar los valores mínimos y máximos de cada columna, una vez obtenido los límites, los demás valores se sustituirán con la siguiente fórmula, garantizando que todos los valores del atributo tengan una distribución entre 0 y 1:

$$x_i = \frac{x_i - x_{min}}{x_{\max} - x_{\min}}.$$

   Si los valores máximos y mínimos son los mismos, se dará valor de 1 a todos los valores de la columna, evitando así la división entre cero.

4. Matriz de distancias.

   La última fase de la estandarización de datos consiste en obtener la matriz de distancias. Existen muchas métricas que se pueden utilizar, por ejemplo Euclidiana, Euclidiana cuadrática, Manhattan, Jaccard, Sorensen-Dice, Pearson, para una referencia más completa consultar [23]. En nuestro caso los mejores resultados se obtuvieron con distancia Euclidiana cuadrática, cuya fórmula se describe a continuación.

$$D_{ij} = \sum_{l=1}^{d} (x_{il} - x_{jl})^2.$$

   El resultado es una matriz n x n simétrica con diagonal principal compuesta por ceros, donde n es el número de instancias trabajadas por el número de clases distintas del conjunto de prueba, por ejemplo si una base de datos es de 10 instancias y tiene 2 clases, se obtiene una matriz de distancias de 20×20.

   Con la matriz de distancias representando a la gráfica completa de los datos y sus correlaciones, se puede poner en marcha la técnica de coloración de gráficas suaves.

4 Instancias de prueba

Como se mencionó anteriormente para un número de instancias de prueba mayores a 20 es necesario utilizar una técnica metaheurística, en este caso se optó por recocido simulado [3], para los casos en el que el número de instancias es menor o igual, se utilizó un software licenciado llamado General Algebraic Modeling System (GAMS) [8] que ejecuta el algoritmo de coloración que garantiza el óptimo. Todas las bases de datos fueron extraídas del Repositorio de Aprendizaje Automático de la Universidad de Irvine California (UCI por sus siglas en inglés).

Aunque el sistema es no supervisado, se decidió comparar los resultados con instancias clásicas de la literatura, las cuales ya tienen una clasificación previa. En nuestros resultados se presenta la columna de porcentaje de eficiencia la cual se define como el total de elementos correctamente etiquetados por nuestro modelo contra el total de elementos utilizados.

Todas las pruebas se realizaron en una PC con Windows 7 Professional de 64 bits Service Pack 1 con procesador Intel Xeon a 2.27 GHz y 4 GB de memoria RAM.

4.1 Bases de datos utilizadas [19]

4.1.1 Hepatitis

Reúne los datos del tratamiento y los síntomas de 155 pacientes de hepatitis, cuenta con 19 características tomadas como atributos que son [2]:

1. Edad del paciente, los datos oscilan entre los 7 y 78 años.
2. Sexo del paciente: hombre (1), mujer (2).
3. Tratamiento con esteroides: no (1), sí (2).
4. Tratamiento con antivirales: no (1), sí (2).
5. Síntoma de fatiga: no (1), sí (2).
6. Malestar general en el paciente: no (1), sí (2).
7. El paciente presentaba un cuadro de anorexia: no (1), sí (2).
8. Hígado grande: no (1), sí (2).
9. Hígado duro: no (1), sí (2).
10. Bazo fácilmente palpable: no (1), sí (2).

11. Araña vascular o aparición de manchas rojizas en la piel: no (1), sí (2).

12. Ascitis, es decir, acumulación de líquido seroso alrededor del hígado: no (1), sí (2).

13. Várices en el paciente, entiéndanse como hinchazón de las venas: no (1), sí (2).

14. Cantidad de bilirrubina como consecuencia de una inflamación aguda del hígado: los datos se distribuyen entre 0.3 mg/dl y 8 mg/dl de forma continua.

15. Cantidad anormal de fosfatasa alcalina, una enzima muy sensible a problemas del hígado o los huesos: fluctuando entre 26 y 295 en números enteros.

16. Presencia de transaminasa sérica de glutamato-oxalacetato (SGOT por sus siglas en inglés): cuando esta enzima se encuentra en la sangre hay alerta de daños en el corazón o en el hígado, desde 14 hasta 648 todos números enteros.

17. Albúmina: una baja cantidad de esta proteína en la sangre es señal de células hepáticas dañadas, ubicados los datos entre 2.1 y 6.4.

18. Tiempo de protrombina: un monitoreo de la coagulación sanguínea comúnmente usado, distribuido entre 0 y 100 en cantidades enteras.

19. Histología o si el paciente cuenta con expediente clínico: no (1), sí (2).

Hay datos faltantes, tiene valores alfanuméricos por lo que se tuvo que efectuar una previa ponderación, los datos están divididos en 2 clases y están distribuidos de la siguiente manera:

- Vivo: 123 instancias.
- Muerto: 32 instancias.

Debido a que se trata de un clasificador no supervisado, no se conoce a priori la cantidad de clases, para ello es necesario apoyarse en la propiedad de resiliencia. Suponiendo que se desconoce la cantidad de colores de la base de datos, el proceso de coloración sugirió una cifra óptima de colores para después compararlo con las clases predefinidas en la base de datos.

Por tiempos de ejecución se utilizó una cantidad pequeña de instancias, y así garantizar el óptimo de la función objetivo con el software GAMS.

Nodos	Colores	Dureza	Solidez	Resiliencia
80	1	116881	36.987596	
80	2	6851.37	4.3919014	**7.421772811**
80	3	4088.54	3.9823393	0.1028446
80	4	2742.92	3.6091105	0.1034129
80	5	2028.22	3.3803722	0.0676666
80	6	1584.05	3.2109081	0.0527776
80	7	1274.8	3.0560195	0.0506831
80	8	1059.84	2.9440103	0.0380465
80	9	898.269	2.8466274	0.0342099
80	10	758.11	2.7075357	0.0513721

Tabla 1: Resultados de resiliencia utilizando toda la BD Hepatitis.

Bajo las condiciones descritas en la Tabla 1 se puede observar que el proceso sugiere una cantidad de 2 colores como óptima para agrupar el conjunto de datos, tal y como lo sugiere la clasificación supervisada.

Instancias Por Clase	Tiempo de Ejecución (segs.)	Porcentaje de Eficiencia
2 vivos, 2 muertos	0.9165	100%
4 vivos, 4 muertos	1.952	87.50%
5 vivos, 5 muertos	2.3223	90%
10 vivos, 10 muertos	9.4864	85%
20 vivos, 13 muertos	23.6332	84.80%
30 vivos, 13 muertos	39.3526	81.30%

Tabla 2: Resultados de clasificación para la BD Hepatitis.

Para observar mejor el comportamiento del proceso de coloración se fue incrementando gradualmente el número de instancias hasta abarcar toda la base de datos.

Como puede verse en la Tabla 2 sólo 13 instancias de una clase cumplieron con todos los filtros previos, la eficiencia de clasificación es buena incluso para una cantidad dispar de instancias por clase, en cuanto al tiempo de ejecución, se pueden observar resultados buenos. En el apartado V se comparan los resultados obtenidos con otras técnicas similares para poder hacer un análisis a profundidad.

Figura 1: Dendrograma de clasificación para la base de datos Hepatitis.

4.1.2 Wines

Un conjunto de resultados del análisis químico de 178 instancias de vinos de una misma región, cuenta con 13 características tomadas como atributos que son ([6], [21]):

1. Grado de alcohol: cuyos valores oscilan entre 11.03 y 14.83.

2. Cantidad de ácido málico, responsable del sabor ácido de la fermentación de las frutas: el grado de acidez se encuentra entre 0.74 y 5.8.

3. Cenizas: como resultado de la calcinación del vino a $500°$C, los valores se dispersan entre 1.36 y 3.23.

4. Alcalinidad de las cenizas: tratándose de la suma de los cationes de amonio mezclados en los ácidos del vino cuyos valores están entre 10.6 y 30.

5. Magnesio: importante para determinar las condiciones de almacenamiento del vino, las concentraciones varían entre 70 y 162 en números enteros.

6. Fenoles totales, o pigmentación del vino: distribuidos entre 0.98 y 3.88.

7. Flavonoides o pigmentos amarillos que aumentan a envejecer el vino blanco: loos valores se colocan entre 0.34 y 5.08.

8. Fenoles no flavanoides: otro tipo de elemento que influye en la coloración del vino, con valor mínimo de 0.13 y máximo de 0.66.

9. Proantocianinas, sustancia más importante de la uva para este caso, otorga las propiedades beneficiosas del vino a la salud humana: los valores oscilan entre 0.41 y 3.58.

10. Intensidad del color sin importar si el vino es blanco o tinto: con datos desde 1.28 hasta 13. 11. Matiz del color del vino: de igual forma éste no distingue si el vino es blanco o tinto, su información de distribuye entre 0.48 y 1.71.

11. OD280/OD315: concentración de esas proteínas en vinos diluidos, los datos fluctúan entre 1.27 y 4.

12. Prolina, un aminoácido importante del metabolismo del nitrógeno de las levaduras: distribuido entre 278 y 1680 en número entero.

No hay datos faltantes, todos los valores son numéricos y los datos están distribuidos en 3 clases de la siguiente manera:

- Vino de mesa: 59 instancias.
- Vino de crianza: 71 instancias.
- Vino de reserva: 48 instancias.

En la Tabla 3 se muestra la cantidad óptima de colores que sugiere el algoritmo, para ello se utilizó la propiedad de resiliencia.

Nodos	Colores	Dureza	Solidez	Resiliencia
12	1	163.374	2.475363636	
12	2	49.4852	1.649506667	0.500669071
12	3	17.2711	0.959505556	**0.719121538**
12	4	10.8214	0.901783333	0.06400897
12	5	7.1476	0.850904762	0.059793497
12	6	4.5769	0.762816667	0.115477413
12	7	3.1643	0.738336667	0.033155607
12	8	2.1423	0.7141	0.033940158
12	9	1.3831	0.69155	0.03260791
12	10	0.715	0.595833333	0.160643357

Tabla 3: Resultados de resiliencia para la BD Wines.

Por tiempos de ejecución se utilizó una cantidad pequeña de instancias, y así garantizar el óptimo de la función objetivo con el software GAMS.

El modelo sugiere 3 colores como óptimo para agrupar el conjunto de datos, tal y como ya se sabía con anterioridad. Se estandarizaron los datos y se usó el algoritmo de coloración a varios casos obteniendo los siguientes resultados:

Instancias totales	Tiempo de ejecución (segundos)	Porcentaje de eficiencia
6	1.6493	100%
9	2.5924	100%
12	3.9579	100%
15	5.8813	100%
30	17.4964	96.6%
60	59.6774	96.6%
90	126.4094	95.5%
Toda la BD	471.2716	93.25%

Tabla 4: Resultados de clasificación para la BD Wines.

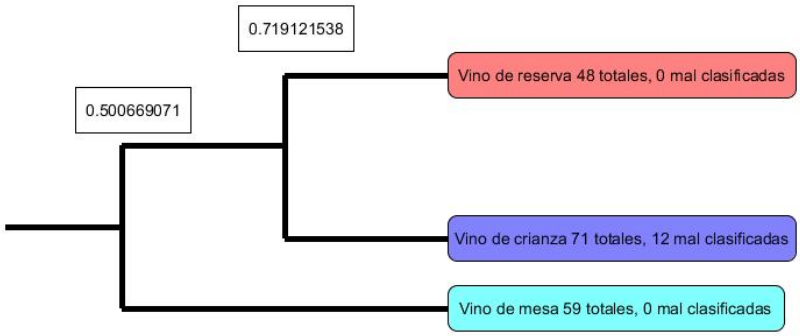

Figura 2: Dendrograma de clasificación para la base de datos Wine.

Se toma una parte proporcional de cada clase para las pruebas preliminares, sólo se utilizó la proporción original en la última prueba.

Analizando la Tabla 4 tanto la eficiencia de clasificación como el tiempo de ejecución son bastante buenos incluso para una cantidad dispar de instancias por clase, aunque para analizar la eficiencia del algoritmo a mayor detalle fue necesario compararlo con otros clasificadores similares (ver en el apartado 5).

4.1.3 Iris

Esta es posiblemente la base de datos más conocida en la literatura de reconocimiento de patrones. Contiene 3 clases de 50 instancias cada una y cada clase corresponde a un tipo de planta. Los atributos de la base de datos son:

1. Longitud del sépalo (hoja que envuelve a la flor desde sus fases tempranas de desarrollo) en cm.

2. Ancho del sépalo en cm.

3. Longitud del pétalo (antófilo que forma parte de la corola de una flor) en cm.

4. Ancho del pétalo en cm.

Los tipos de planta se distribuyen entre:

- Iris setosa.

- Iris versicolor.

- Iris virginica.

Se especifica de antemano que una clase puede ser separada linealmente de las otras dos, mientras que las últimas clases no pueden ser separadas de forma lineal, tomando en cuenta esa información, el primer color será compuesto por la Iris setosa, mientras que el segundo color corresponderá a la Iris versicolor y la Iris virginica.

Se hizo un análisis de regresión [13], para decidir que columnas influyen más en el tipo de planta. Se encontró que la columna de ancho de sépalo tiene una alta probabilidad de valer cero, mientras que la columna de longitud de sépalo tiene valor negativo en su estadístico t, una vez teniendo esta información se tomaron las siguientes medidas:

- Invertir el signo de la característica longitud de sépalo.

- Eliminar la columna ancho de sépalo.

Una vez tomadas las medidas correctivas se volvió a ejecutar el modelo de coloración.

Nodos	Colores	Dureza	Solidez	Resiliencia
150	1	4504.2856	0.40306806	
150	2	517.0014	0.09315341	**3.32692783**
150	3	113.9274	0.03100065	**2.00488526**
150	4	60.2439	0.0220069	0.40867852
150	5	51.1956	0.02353821	-0.06505605
150	6	44.6389	0.02479939	-0.05085537
150	7	40.9661	0.02673778	-0.07249637
150	8	35.8347	0.02691808	-0.00669823
150	9	32.8099	0.02792332	-0.03599983
150	10	31.8684	0.03035086	-0.07998252

Tabla 5: Resultados de resiliencia para la BD Iris.

La resiliencia se calculó con el proceso de recocido simulado y así poder incluir todos los elementos de la BD.

El proceso sugiere 2 colores como óptimos, sin embargo ya ve más claramente que 3 colores es una buena clasificación. A continuación se ejecutó el proceso de clasificación ya con todos los nodos de la BD para los colores que obtuvieron los mejores valores de resiliencia.

Instancias totales	Colores	Tiempo de ejecución (segundos)	Porcentaje de eficiencia
150	2	383.8048	95.3%
150	3	350.753	96%

Tabla 6: Resultados de clasificación para la BD Iris con 2 y 3 colores.

La disposición de los datos en la clasificación con 2 colores no se alteró para que ambos colores tuvieran la misma cantidad de datos, por lo que la proporción se mantuvo original ya que al grupo 2 lo componen 2 tipos de planta.

En el proceso de clasificación es donde más se puede apreciar la mejora, el modelo ya es capaz de ver los 3 tipos de planta y las clasifica a un porcentaje de eficiencia muy alto, en el siguiente apartado se podrá comparar el modelo con otros clasificadores ya publicados y probados.

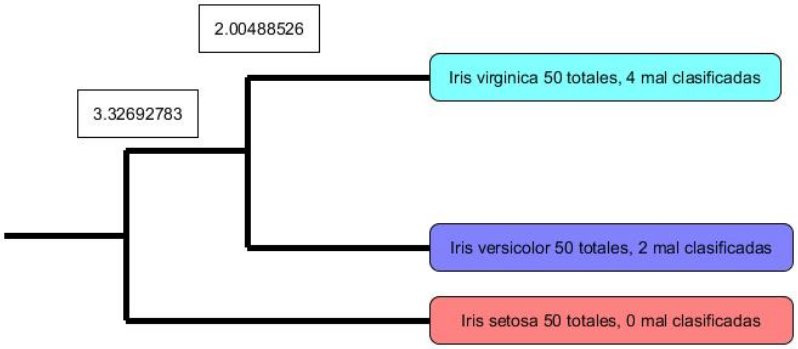

Figura 3: Dendrograma de clasificación para la base de datos Iris.

4.1.4 Car evaluation

1728 modelos distintos de automóviles evaluados entre atributos subjetivos y atributos técnicos, no contiene datos faltantes y las características se describen a continuación:

1. Precio: evaluado como muy alto, alto, medio y bajo, se reemplazaron por 4, 3, 2, 1 respectivamente.
2. Mantenimiento, (muy alto, alto, medio y bajo, se reemplazaron por 4, 3, 2, 1 respectivamente).
3. Puertas: 2, 3, 4, 5 o más.
4. Capacidad: 2, 4 o más personas.
5. Cajuela: pequeña, mediana y grande, se sustituyeron esos valores por 1, 2 y 3 respectivamente.
6. Seguridad: baja, media y alta, reemplazándose por 1, 2 y 3 respectivamente.

Revisando los valores de precio, mantenimiento, cajuela y seguridad, además de ser subjetivos no son numéricos, por lo que fue necesario establecer la ponderación de los valores descrita en la propuesta, todos los datos se encuentran distribuidos entre las siguientes categorías:

- Inaceptable: 1210 instancias.
- Aceptable: 384 instancias.

- Bueno: 69 instancias

- Muy bueno: 65 instancias.

Suponiendo que no sabemos que la agencia ha dividido los autos en estas cuatro clases, se ejecutó el proceso de resiliencia para conocer el número de clases en los que trabajará el agrupador.

Nodos	Colores	Dureza	Solidez	Resiliencia
20	1	367.2775	1.93303947	
20	2	96.4999	1.07222111	**0.80283661**
20	3	48.8055	0.86127353	0.24492519
20	4	22.1111	0.5527775	0.55808355
20	5	13.2778	0.44259333	0.24895126
20	6	8.3055	0.35595	0.24341434
20	7	5.5555	0.29914231	0.1899019
20	8	3.8889	0.25926	0.15383132
20	9	3.1389	0.25681909	0.00950439
20	10	2.5278	0.25278	0.01597868

Tabla 7: Resultados de resiliencia para la BD Car Evaluation.

Por tiempos de ejecución se utilizó una cantidad pequeña de instancias, y así garantizar el óptimo de la función objetivo con el software GAMS.

En la Tabla 7 se sugieren 2 colores como óptimos, es decir, para el proceso la mejor clasificación es sólo entre carros buenos y malos, hay que resaltar que el proceso ve como buena clasificación el uso de 4 colores, por lo que el algoritmo de coloración ve adecuada también la clasificación original. Estos son los resultados luego de estandarizar y ejecutar la coloración suave óptima:

Debido al tiempo de ejecución no fue posible utilizar toda la base de datos, por lo que las pruebas se acotaron hasta 50 instancias por clase.

Esta base de datos es la que cuenta con el mayor éxito de clasificación, teniendo resultados de 100 por ciento en la mayoría de las pruebas, por lo que podemos concluir que mientras se use la cantidad de colores sugerida por el proceso, se obtienen excelentes resultados

Instancias Totales	Tiempo de Ejecución (segundos)	Porcentaje de Eficiencia
8	1.8176	100%
12	4.07	100%
16	5.7669	100%
20	8.7322	100%
40	27.7042	100%
80	102.1054	100%
120	228.0418	95.83%
200	618.1844	98.5%

Tabla 8: Resultados de clasificación para la BD Car Evaluation.

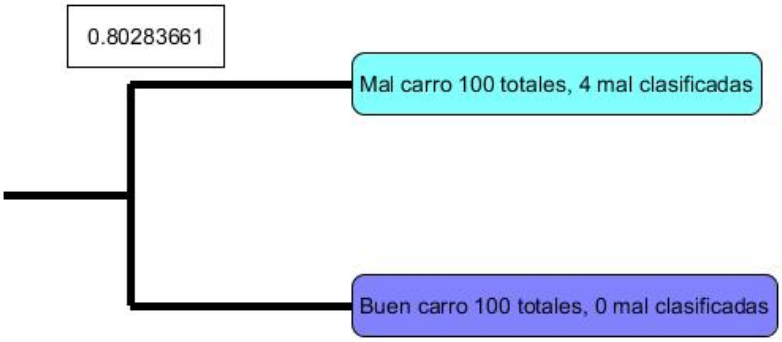

Figura 4: Dendrograma de clasificación para la base de datos Car Evaluation.

4.1.5 Stone flakes

Se trata de una base de datos publicada en 2014 y se compone de la recopilación de 79 objetos que sirvieron como auxiliares para crear puntas de flechas, lanzas o cuchillos en la Europa Occidental durante el periodo Paleolítico. Se busca encontrar una relación de progreso tecnológico con estas herramientas de trabajo o algún cambio en la técnica de creación de armas de la Prehistoria [22]. Hay presentes 8 atributos de la base de datos que se describen a continuación:

1. Ancho del objeto: distribuido en un intervalo desde 1.02 hasta 1.69 cm.

2. Grosor del objeto: entre 16.5 y 43.7 cm.

3. Profundidad del objeto: oscilando entre 1.66 y 4.9 cm.

4. Ángulo de golpeo entre el objeto y la superficie en la que se talló: desde los $105°$ hasta los $131°$.

5. Frecuencia de uso como herramienta primaria: los datos se encuentran entre 0 y 67.2.

6. Frecuencia de uso como herramienta multiusos: con valores desde 0 hasta 55.3.

7. Área de mayor desgaste: distribuyéndose los valores entre 5 y 94.1 cm^2.

8. Proporción de desgaste respecto a la superficie total: varía entre 30 y 98%.

La base de datos no cuenta con clases predefinidas, los arqueólogos sugieren la siguiente clasificación: Edad y tipo de homínido, desde el *Homo ergaster* en el Bajo Paleolítico entre 600,000 y 300,000 años de antigüedad, pasando por la técnica Levallois de hace 200,000 años, seguido del grupo de neandertales del Medio Paleolítico entre 130,000 y 60,000 años de antigüedad, llegando finalmente al *Homo sapiens* del Alto Paleolítico, hace unos 40,000 años.

Una vez estandarizados los datos y aplicando el proceso de coloración, se obtuvieron los resultados de la tabla 9.

Nodos	Colores	Dureza	Solidez	Resiliencia
73	1	1909.4213	0.72656823	
73	2	564.74	0.43584025	**0.6670517**
73	3	309.4432	0.3633384	**0.1995436**
73	4	201.9049	0.32067485	0.13304302
73	5	149.0708	0.30030379	0.06783484
73	6	120.4644	0.29555772	0.01605801
73	7	101.0152	0.29352694	0.00691854
73	8	84.661	0.28547439	0.0282076
73	9	76.99	0.29662243	-0.03758326
73	10	64.3584	0.27987997	0.05982013

Tabla 9: Resultados de resiliencia para la BD Stone Flakes.

Se utilizó el método de recocido simulado dado que se utilizaron todos los nodos de la base de datos.

En la Tabla 9 se sugieren 2 colores como óptimos, sin embargo no descarta la posibilidad de que se utilicen 3 colores. En la siguiente tabla a diferencia de las anteriores se ejecutó el modelo de clasificación tanto con el software GAMS como con el proceso de recocido simulado. La eficiencia se tomó respecto a la clasificación sugerida de edad y tipo de homínido (ver Tabla 10).

Modelo de ejecución	Tiempo de ejecución (segundos)	Valor de la función objetivo	Porcentaje de eficiencia
GAMS	8042.54	564.74	89.04%
Recocido Simulado	100.68	564.74	89.04%

Tabla 10: Clasificación para la BD Stone Flakes con 2 colores.

GAMS garantiza el óptimo, pero como se puede observar en la Tabla 10 el tiempo de ejecución es mucho mayor que mediante el proceso de recocido simulado; el proceso de recocido simulado es más rápido aunque no garantiza encontrar el óptimo. En este caso alcanzó la misma solución que GAMS, pero no se puede tener la certeza de que siempre encuentre el óptimo.

Respecto a la clasificación, el modelo es capaz de distinguir las piezas entre los homínidos más antiguos de los más recientes, lo que podría confirmar de manera matemática la hipótesis de que existe una evolución en las técnicas de tallado de piedras en el Paleolítico [22]. En la Tabla 11 se muestran los resultados del proceso de clasificación ahora con 3 colores.

Modelo de ejecución	Tiempo de ejecución (segundos)	Valor de la función objetivo	Porcentaje de eficiencia
GAMS	50704.21	–	Desconocido
Recocido Simulado	102.9938	309.4432	82.19%

Tabla 11: Clasificación para la BD Stone Flakes con 3 colores.

Se utilizaron las 73 instancias que cumplían con los criterios de limpieza y preparación de los datos.

En esta ocasión GAMS fue incapaz de encontrar una solución tras 14 horas aproximadas en ejecución por lo que nos apoyamos en los resultados del recocido simulado encontrando un porcentaje de eficiencia bastante aceptable tomando en cuenta que no es el óptimo de colores que sugiere el modelo.

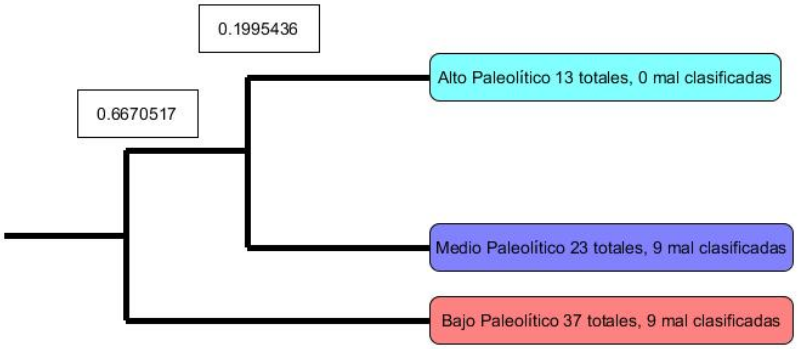

Figura 5: Dendrograma de clasificación para la base de datos Stone Flakes.

En este caso el proceso de clasificación ve claramente la separación de los homínidos más antiguos, pero le cuesta trabajo diferenciar las piedras talladas entre Neandertales y *Homo sapiens*.

4.2 Comparación de los resultados obtenidos con técnicas similares

A continuación se presentan trabajos de otros clasificadores sobre las mismas bases de datos trabajadas, ahí se puede contrastar la eficiencia de la técnica de coloración de gráficas suaves respecto a los demás clasificadores [20].

4.2.1 Hepatitis

Método y métrica	Eficiencia	Tipo de prueba
21 NN, Manhattan k=1	90.3	leave-one-out
FSM	90	leave-one-out
14-NN, Euclidiana k=1	89	leave-one-out
LDA	86.4	leave-one-out
Coloración GS Euclidiana k=2	**84.8**	**Subconjuntos sin . entrenamiento.**
CART (árbol de decisión)	82.7	leave-one-out
MLP+backprop	82.1	leave-one-out

Tabla 12: Comparación de clasificadores para la base de datos Hepatitis en pruebas *leave-one-out*.

Método y métrica	Eficiencia	Tipo de prueba
9NN con pesos	92.9	10 x cross validation
18NN Manhattan	90.2 ± 0.7	10 x cross validation
FSM con rotaciones	89.7	10 x cross validation
15 NN Euclidiana k=1	89 ± 0.5	10 x cross validation
VSS 4 neuronas, 5 it	86.5 ± 8.8	10 x cross validation
FSM sin rotaciones	88.5	10 x cross validation
LDA,análisis lineal discriminante	86.4	10 x cross validation
Bayes ingenuo y semi NB	86.3	10 x cross validation
IncNet	86	10 x cross validation
QDA, Análisis cuadrático discrimiante	85.8	10 x cross validation
1-NN	85.3 ± 5.4	10 x cross validation
VSS 2 neuronas, 5 it	85.1 ± 7.4	10 x cross validation
ASR	85	10 x cross validation
Coloración GS, Euclideana k=2	**84.8**	**Subconjuntos sin entrenamiento**
Análisis discrimiante de Fisher	84.5	10 x cross validation
LVQ	83.2	10 x cross validation
CART (árbol de decisión)	82.7	10 x cross validation
MLP con BP	82.1	10 x cross validation
ASI	82	10 x cross validation
LFC	81.9	10 x cross validation
RBF	79	10 x cross validation
MLP+BP	77.4	10 x cross validation

Tabla 13: Comparación de clasificadores para la BD Hepatitis en pruebas por validación cruzada de tamaño 10.

Como se puede observar en las tablas 12 y 13 el clasificador se encuentra bien posicionado respecto a los demás, lo que significa que los resultados de eficiencia son bastante buenos.

4.2.2 Wine

Para este caso también se obtuvieron buenos resultados, cabe mencionar que la fuente no especifica las características del modelo de clasificación que utilizaron (ver Tablas 14 y 15).

Método y métrica	Eficiencia	Tipo de prueba
RDA	100	leave-one-out
QDA	99.4	leave-one-out
LDA	98.9	leave-one-out
kNN, Manhattan k=1	98.7	leave-one-out
1NN	96.1	leave-one-out
kNN, Euclidiana k=1	95.5	leave-one-out
kNN, Chebyshev k=1	93.3	leave-one-out
Coloración GS, Euclidiana k=2	**93.2**	**Subconjuntos sin entrenamiento**

Tabla 14: Comparación de clasificadores para la BD Wine en pruebas *leave-one-out*.

Método y métrica	Eficiencia	Tipo de prueba
kNN, Manhattan, k de 1-10	98.9 ± 2.3	10 x cross validation
IncNet, 10CV, Gauss	98.9 ± 2.4	10 x cross validation
10 CV SSV, con podado	98.3 ± 2.7	10 x cross validation
10 CV SSV, 7 nodos	98.3 ± 2.7	10 x cross validation
kNN, Euclidiana, k=1	97.8 ± 2.8	10 x cross validation
kNN, Manhattan, k=1	97.8 ± 2.9	10 x cross validation
kNN, Manhattan, k de 1-10	97.8 ± 3.9	10 x cross validation
kNN, Eucliana, k=3, pesos alterados	97.8 ± 4.7	10 x cross validation
IncNet, 10CV, bicentral	97.2 ± 2.9	10 x cross validation
kNN, Euclidiana k de 1-10	97.2 ± 4.0	10 x cross validation
10 CV SSV nodos optmizados	97.2 ± 5.4	10 x cross validation
FSM a=0.99	96.1 ± 3.7	10 x cross validation
FSM 10CV, Gauss, a=0.999	96.1 ± 4.7	10 x cross validation
FSM 10CV triangular, a=0.99	96.1 ± 5.9	10 x cross validation
kNN, Euclidiana k=1	95.5 ± 4.4	10 x cross validation
Coloración GS, Euclidiana k=1	**93.2**	**Subconjuntos sin entrenamiento**
10 CV SSV nodos optimizados, BFS	92.8 ± 3.7	10 x cross validation
10 CV SSV nodos optimizados, BS	91.6 ± 6.5	10 x cross validation
10 CV SSV con podado, BFS	90.4 ± 6.1	10 x cross validation

Tabla 15: Comparación de Clasificadores para la BD Wine en pruebas por validación cruzada de tamaño 10.

4.2.3 Iris

Como se mencionó anteriormente, Iris es una de las bases de datos más utilizadas, los autores también han tenido que enfrentarse al problema de separar los otros 2 tipos de planta con varias técnicas como son algoritmos evolutivos, redes neuronales o árboles de decisión [18], la comparación de resultados se hará sobre la clasificación de los 3 tipos de plantas.

Método y métrica	Eficiencia	Tipo de prueba
Coloración GS, Euclidiana k=2	**96**	**Sin entrenamiento**
C4.5	93.6	Algoritmo0 k-medias, CV
C4.5+m	93.1	Algoritmo0 k-medias, CV
C4.5+m+cf	93.1	Algoritmo0 k-medias, CV
C4.5+cf	91.6	Algoritmo0 k-medias, CV

Tabla 16: Comparación de clasificadores para la BD Iris; CV indica validación cruzada.

Tal y como se muestra en la Tabla 16 [4] el clasificador está muy bien colocado respecto a otros trabajos y aunque la muestra de trabajos es muy pequeña, se obtuvo la comparación de un artículo dedicado a la comparación de distintos clasificadores.

4.2.4 Car evaluation

Con lo mostrado en la Tabla 17 el agrupador se muestra bien posicionado respecto a clasificadores supervisados, incluso contra el Bayes Ingenuo [1], lo que parece indicar que el clasificador es bastante eficiente para distintos tipos de bases de datos.

Método y métrica	Eficiencia	Tipo de prueba
GBN	86.11 ± 1.46	Entrenamiento previo, CS
BAN	94.04 ± 0.44	Entrenamiento previo, CS
TAN	94.10 ± 0.48	Entrenamiento previo, CS
Coloración GS, Eucl. k=2	**98.5**	**Entrenamiento previo, CS**
Bayes Ingenuo	86.58 ± 1.78	Entrenamiento previo, CS

Tabla 17: Comparación de clasificadores para la BD Car Evaluation; CS indica con subconjunto.

4.2.5 Stone flakes

El trabajo encontrado [17] también realiza el agrupamiento de los datos haciendo uso de la clasificación geo-cronológica propuesta.

Método y métrica	Eficiencia	Tipo de prueba
Algoritmo de Agrupación recortada	84.93%	Probabilidad aplicada en análisis de agrupamiento y selección de variables
Coloración GS Euclidiana k=2	**82.19%**	**Sin Entrenamiento**

Tabla 18: Comparación de clasificadores para la BD Stone Flakes.

Los resultados son muy similares en la ejecución a 3 colores. Se puede percibir una ventaja importante del modelo de coloración de gráficas suaves respecto al modelo de Ritter, la cual consiste en que el método de coloración no conlleva un análisis tan exhaustivo de los datos ni consideró previamente varios tipos de separación (lineal, Fisher, p-valor, etc.).

5 Conclusiones

El agrupador propuesto, es bastante eficiente bajo condiciones de limpieza de ruido, de normalización de datos y de clases sugeridas por él mismo, esto lo posiciona muy bien entre clasificadores ya publicados y usados frecuentemente.

Consideramos que la solución como primer proceso de filtrado debe ser rápida y fácil de utilizar; un manejo complejo de los datos va en contra del paradigma original de la propuesta. Nuestro algoritmo se desempeña bien, por ejemplo en las bases de datos complicadas para los clasificadores más utilizados el clasificador propuesto, ha obtenido resultados bastante buenos.

Otro punto a resaltar es que, a diferencia de un clasificador supervisado, donde mediante el entrenamiento, se le dice qué debe encontrar, un agrupador no supervisado tiene su propia inteligencia y sus propios criterios sin necesidad de ser entrenado.

Si los resultados de clasificación arrojados por el agrupador no coinciden con los que se conocen a priori, no necesariamente se trata de un error, sino otro enfoque del problema que debería ser tomado en cuenta, sería el caso de la evaluación de automóviles, donde nuestro clasificador distingue con gran claridad un auto bueno de uno malo.

Referencias

[1] Cheng, J.; Greiner, R. (1999) "Learning theory and language modeling", in: N. Friedman, M. Goldszmidt & A. Wyner (Eds.) *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, Alberta: 101–108.

[2] Chow, J.H.; Chow, C. (2006) *The Encyclopedia of Hepatitis and Other Liver Diseases.* Facts On File, New York.

[3] De los Cobos, S.G.; Goddard, J.; Gutiérrez, M.A.; Martínez, A.E. (2010) *Búsqueda y Exploración Estocástica.* Universidad Autónoma Metropolitana, Ciudad de México.

[4] Demšar, J. (2006) "Statistical comparisons of classifiers over multiple data sets", *Journal of Machine Learning Research* **7**(Jan): 1–30.

[5] Diestel, R. (2000) *Graph Therory.* Springer-Verlag, New York.

[6] Fernández, V.; Berradre, M.; Sulbarán, B.; Ojeda, G.; Peña, J. (2009) "Caracterización química y contenido mineral en vinos comerciales venezolanos", *Revista de la Facultad de Agronomía* **26**(3): 392–396.

[7] GAMS Development Corporation. (2015) "General Algebraic Modeling System", en: `http://www.gams.com`, consultado el 18/08/2015, 17:30.

[8] Gutiérrez, M.A.; Lara, P.; Lopez, R.; Ramírez, J. (2011) "Heuristics for the robust coloring problem", *Revista de Matemática: Teoría y Aplicaciones* **18**(1): 137–147.

[9] Jain, A.K.; Murty, M.N.; Flynn, P.J. (1999) "Data clustering: a review", *ACM Computing Surveys (CSUR)* **31**(3): 264–323.

[10] Joachims, T. (1996) "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization", *Dept. of Computer Science, Carnegie-Mellon Univ., Pittsburgh PA* CMU(CS): 96–118.

[11] Lara, P.; Gutiérrez, M.A.; De los Cobos, S.G.; Rincón, E. (2015) "Coloración de gráficas suaves", *Revista de Matemática: Teoría y Aplicaciones* **22**(2): 1–13.

[12] McAllester, D.; Schapire, R.E. (2002) "Learning theory and language modeling", in: G. Lakemeyer & B. Nebel (Eds.) *Exploring Artificial Intelligence in the New Millenium*, Morgan Kaufmann, San Francisco: 271–285.

[13] Montgomery, D.C.; Runger, G.C. (2003) *Applied Statistics and Probability for Engineers.* John Wiley & Sons, New York.

[14] Moreno, B. (2009) *Minería sobre Grandes Cantidades de Datos*. Tesis de Maestría, Departamento de Ingeniería Eléctrica, Posgrado en Ciencias y Tecnologías de la Información, Universidad Autónoma Metropolitana, México D.F.

[15] Parzen E. (1962) "On estimation of a probability density function and mode", *The Annals of Mathematical Statistics* **33**(3): 1065–1076.

[16] Ripley, B.D. (1996) *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.

[17] Ritter, G. (2014) *Robust Cluster Analysis and Variable Selection.* CRC Press, Boca Raton.

[18] Universidad de California en Irvine (2015) "Iris related papers", en: `http://archive.ics.uci.edu/ml/datasets/Iris`, consultado el 25/11/2015, 12:53.

[19] Universidad de California en Irvine (2015) "UCI machine learning repository", en: `http://archive.ics.uci.edu/ml/index.html`, consultado el 28/05/2015, 18:18.

[20] Universidad Nicolás Copérnico de Polonia (2010) "Datasets classifier", en: `http://www.is.umk.pl/projects/datasets.html`, consultado el 23/11/2015, 14:22.

[21] Waterhouse, A. L.; Ebeler, S. E. (1998) *Chemistry of Wine Flavor.* American Chemical Society, Washington DC.

[22] Weber, T. (2009) "The lower/middle palaeolithic transition. Is there a lower/middle palaeolithic transition?", *Preistoria Alpina* **44**: 17–24.

[23] Xing, E.P.; Ng, A.Y.; Jordan, M.I.; Russell, S. (2003) "Distance metric learning with application to clustering with side-information", *Advances in Neural Information Processing Systems* **15**(1): 505–512.

[24] Xu, R.; Wunsch, D. (2005) "Survey of clustering algorithms", *IEEE Transactions on Neural Networks* **16**(3): 645–678.