UNA HERRAMIENTA COMPUTACIONAL PARA EL AGRUPAMIENTO DE DATOS BASADO EN EL COMPORTAMIENTO COLECTIVO DE LAS ABEJAS

Niriaska Perozo - Oscar Gutiérrez - Raúl Pérez

Unidad de Inteligencia Artificial, Decanato de Ciencias y Tecnología. - UCLA
nperozo@ucla.edu.ve; bgutierrez600@gmail.com; rauleduardops@gmail.com

Recibido: Mar [2016]
Aceptado: Sep. [2016]

Resumen

En el ámbito de la minería de datos y el aprendizaje de máquina no supervisado, la agrupación de datos es definido como la tarea de agrupar objetos de acuerdo con una medida de similitud o disimilitud. Esto es, los objetos que son similares entre sí, se reúnen en el mismo grupo, y los que son disímiles se organizan en diferentes grupos, a partir de lo cual, puede emerger una estructura descriptiva de los datos. En las ciencias sociales, la clasificación y agrupamiento de individuos respecto a patrones de comportamiento puede dar lugar a descripciones y predicciones cuantitativas que permitan un estudio más preciso de cómo funcionan las sociedades bajo ciertos parámetros como por ejemplo: la predicción de un comportamiento emergente de la delincuencia en algunos sectores sociales. En general, el problema de agrupamiento puede formularse con la optimización de multi-objetivos, la cual puede ser muy compleja en términos de tiempo y espacio, en términos computacionales. En este sentido, el algoritmo de colonia de abejas, perteneciente al área de la inteligencia de enjambre y basado en la optimización numérica, intenta obtener la mejor solución al problema, explotando y explorando el espacio de búsqueda. En este trabajo se propone una herramienta computacional implementada en Java para simular el comportamiento de los enjambres de abejas como un sistema multi-agentes, en el cual es posible observar la agrupación en los datos de prueba que se utilizan para ajustar los parámetros clave y comparar los resultados obtenidos con trabajos similares. A través de la experimentación realizada, se propone utilizar el algoritmo de optimización de enjambre de partículas, como una técnica heurística para obtener mejores soluciones iniciales en el agrupamiento de datos, de tal manera que el algoritmo de colonia de abejas pueda converger a un óptimo global, mejorando su velocidad de convergencia.

Palabras Clave: Agrupamiento de Datos, Colonia Artificial de Abejas, K- Medias, Inteligencia de Enjambre, Minería de Datos.

A COMPUTER TOOL FOR DATA GROUPING BASED ON BEHAVIOR OF BEE

Abstract

In the field of data mining and unsupervised machine learning, data clustering is defined as the task of grouping objects according to a similarity or dissimilarity measure. That means, objects that are similar among them are grouped in the same cluster, and objects that are dissimilar are grouped into different clusters so a data descriptive structure can emerge. In social sciences, the classification and the grouping regarding to behavior patterns can take place to quantitative descriptions and predictions which let more specific study about how societies work under some parameters such as prediction of a crime emergent behavior in some social sectors. In general, the clustering problem can be formulated as a multi-objective optimization problem, which can be very complex in time and space computationally speaking. In this sense, the Artificial Bee Colony Algorithm which is a swarm intelligence algorithm based on numeric optimization, tries to get the best solution to the problem, exploiting and exploring the search space. In this work, we propose a computationally tool implemented in java for simulating the behavior of the honey bee swarms as a multi-agent system, where it is possible to observe the data clustering in training data that is used to tune the key parameters and compare them with similar papers. Through this experimentation, it is proposed to use the particle swarm optimization algorithm as a heuristic technique to get better initial solutions to the problem, so that the ABC algorithm can converge to a global optimum improving its convergence rate.

Keywords: Clustering; Artificial Bee Colony; K-Means; Swarm Intelligence, Data Mining.

INTRODUCCIÓN

El algoritmo de la colonia artificial de abejas ("*Artificial Bee Colony*, ABC", en inglés) es una de las metas heurísticas bio-inspiradas más usadas en el ámbito de computación emergente, para resolver problemas computacionales complejos. Desde su aparición, su aplicabilidad se ha extendido a diversos campos, entre los cuales figuran: optimización numérica y combinatoria, ingeniería de software, sistemas expertos, entrenamiento de redes neuronales, ingeniería eléctrica, procesamiento y segmentación de imágenes, entre otras.

Gracias a su simplicidad en computación y potencia, el algoritmo ABC se ha convertido en una buena alternativa para resolver problemas de complejidad NP-duros, entre ellos el agrupamiento de datos ("*Clustering*", en inglés), técnica de aprendizaje automático no supervisado, enmarcada en el área de minería de datos y descubrimiento de conocimiento. Esta técnica encuentra sus aplicaciones en muchas ramas de la ciencia, como los negocios, la sociología, la medicina y la biología, quedando en evidencia su importancia y gran interés de investigación.

El agrupamiento de datos (y posteriormente el análisis de los grupos por parte de expertos) también es usado con frecuencia en el área de las ciencias sociales. Dymnicki & D. Henry (2011) describen la capacidad del agrupamiento y el análisis de los grupos

descubiertos, para resaltar ciertos procesos sociales. Especialmente, se explica cómo agrupar o clasificar familias de acuerdo con ciertas prácticas paternales, creencias familiares, estrategias de comunicación y disciplinarias, para descubrir los patrones de comportamiento que llevan a un mayor riesgo de delincuencia, tomando en cuenta el funcionamiento general de la familia. Adicionalmente, en Filho et al. (2014), usan simulaciones básicas para clasificar regímenes políticos en base a dos dimensiones de poliarquía: inclusividad y participación.

Por otra parte, el algoritmo K-medias ("k-means", en inglés), es el enfoque más simple de agrupamiento de datos, además, es sencillo de implementar, pero tiende a estancarse en óptimos locales. En este trabajo se implementa una herramienta computacional como sistema multi-agente que simula el comportamiento de las colonias de abejas hibridizado con el algoritmo k-medias, para realizar el agrupamiento en tres (3) distintos conjuntos de datos estándares (iris, wine, cmc) del repositorio de aprendizaje automático de la Universidad de California, Irvine (UCI). Los resultados obtenidos en este trabajo se comparan con los de otras contribuciones similares, a los fines de verificar la herramienta propuesta.

ASPECTOS TEÓRICOS RELEVANTES

Agrupamiento de datos. Aprendizaje automático no supervisado

El objetivo del agrupamiento de datos, es descubrir un nuevo conjunto de categorías, grupos o clases dado un conjunto de datos generalmente multidimensional, para describir dicho conjunto. Formalmente, la estructura del agrupamiento es representada como un conjunto de subconjuntos $C = C_1, \ldots, C_K$ de *S* tal que: $S = U_{i=1}^{k} C_i$ y $C_i \cap C_j = \emptyset$ para $i \neq j$. Consecuentemente, cualquier instancia de *S* pertenece a uno y sólo a un grupo.

Cios. Pedrycz, Swiniarski y Kurgan. (2007) muestran la complejidad computacional del problema de agrupamiento, considerando todas las posibles particiones, dados los parámetros del modelo:

$$Numero\ de\ posibles\ particiones = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-i}\binom{k}{i}i^{N} \quad (1)$$

Donde N es el tamaño del conjunto de datos y k es el número de grupos. Esto da lugar a un número bastante grande, cuando el tamaño del conjunto de datos aumenta. Así, queda en evidencia la importancia de una solución eficiente.

Algoritmo k-medias

La idea básica de este algoritmo es encontrar una estructura de agrupamiento que minimice cierto criterio de error, que mide la distancia de cada instancia a su valor representativo (centroides o centros de grupos). Los centroides de un gran número de datos deberían ser tales, que minimicen la dispersión de dichos datos a su alrededor K. (Cios, Pedrycz, Swiniarski y Kurgan. 2007). Teniendo N instancias en R^n y asumiendo que se quieren formar k grupos, se calcula la suma de los cuadrados del error entre las instancias y un conjunto de centroides $v_1, v_2, \ldots, v_k$:

$$Q= \sum_{i=1}^{k}\sum_{j=1}^{N} U_{ij} \left|\left|x_j - v_i\right|\right|^2 \quad (2)$$

Donde x_j es la j-esima instancia del conjunto de datos $j = 1,2,\dots,N$, v_i es el i-esimo centroide (valor representativo del grupo i) $i = 1,2,\dots,k$ y $U = [u_{ij}]$ es la matriz de partición (también llamada matriz de membresía), que organiza a las instancias en los grupos. Cabe destacar que esta matriz posee entradas binarias (1 si la instancia j pertenece al grupo i y 0 en caso contrario).

El algoritmo comienza con un conjunto inicial de centros de grupos (centroides) elegidos aleatoriamente o usando alguna heurística. En cada iteración, cada instancia es asignada a su centro de grupo más cercano, de acuerdo con la distancia euclidiana entre ambas. Después, se recalculan los centros de cada grupo.

El centro de cada grupo se calcula como la suma de todas las instancias en ese grupo dividida entre la cantidad de instancias en ese grupo (media ponderada):

$$v_i = \frac{\sum_{j=1}^{N} u_{ij} x_j}{\sum_{j=1}^{N} u_{ij}} \quad (3)$$

Donde x_j es la j-esima instancia del conjunto de datos y u_{ij} corresponde a las entradas de la matriz de membresía de cada instancia j al grupo i.

La estructura algorítmica del algoritmo K-medias es mostrada en la tabla 1.

Tabla 1. Algoritmo K-medias.

Inicio
{
Fase de inicialización aleatoria de k centroides
Repetir
{
Construir la matriz de partición (a)
Actualizar los centroides $v_1, \dots, v_k$ ***(b)***
} Mientras (Q no cambie o los cambios sean imperceptibles y no se haya Completado el número máximo de iteraciones)
}
Fin

De acuerdo con la tabla 1, la sección (a) representa la matriz de partición con valores binarios (1 o 0) para cada instancia. Es decir, se indica su membresía a un grupo específico de acuerdo con la partición inicial de centroides. En la sección (b) se calculan los nuevos centros de grupos tomando en cuenta todas las instancias del mismo, a los fines de obtener mejores valores representativos para cada grupo. Para más detalle sobre este algoritmo, ver Cios, Pedrycz, Swiniarski y Kurgan (2007) y Armano y Farmani (2014).

Algoritmo de la colonia artificial de abejas para el agrupamiento de datos (ABCK)

La colonia de abejas artificiales en el algoritmo ABC contiene tres (3) grupos de abejas: **las empleadas, las observadoras y las exploradoras (**Karaboga, Gorkemli, Ozturk y Karaboga, 2014). Para cada fuente de comida sólo existe una abeja empleada. Inicialmente, todas las fuentes de comida son descubiertas por las abejas observadoras (es decir, se descubren aleatoriamente). Luego, la mitad de las abejas en la colonia está conformada por las abejas empleadas y la otra mitad serán las abejas observadoras. La abeja empleada de una fuente de comida abandonada, se convierte en una exploradora, posteriormente. Una fuente de comida puede ser abandonada debido a que se ha agotado su rentabilidad.

La estructura algorítmica general del enfoque de agrupamiento de datos de la colonia artificial de abejas basado en el K-medias es llamada ABCk y se describe en la tabla 2.

Tabla 2. Algoritmo ABCK.

Inicio
{
Fase de inicialización de fuentes de comida (conjunto de centroides)
Repetir
{
Fase de las abejas empleadas
Fase de las abejas observadoras
Fase de las abejas exploradoras
Memorizar la mejor posición obtenida hasta ahora
} Mientras (Ciclo = Máximo número de ciclos o Tiempo máximo de CPU)
}
Fin

En la fase de iniciación, toda la población de fuentes de comida, (donde cada una de estas fuentes es un conjunto de centroides), x_i es inicializada por las abejas exploradoras (es decir, aleatoriamente) y los parámetros de control son asignados (límites para visitar una fuente de comida antes de que sea abandonada). Luego, se aplica el algoritmo K-medias a cada una de estas fuentes de comida, se calcula la suma de los cuadrados del error para cada una (función objetivo a minimizar) y se calculan sus respectivos valores en la función de aptitud fit_i. La función de aptitud para una fuente i se representa mediante la siguiente expresión:

$$fit_i = \begin{cases} \frac{1}{1+f_i}; Si\ f_i \geq 0 \\ 1 + abs(f_i); Si\ f_i < 0 \end{cases} \quad (4)$$

En la fase de las abejas empleadas, éstas buscan nuevas fuentes de comida que tengan más néctar que aquellas actualmente en su memoria. Encuentran una fuente de comida vecina $v_{i,j}$ de $x_{i,j}$. Esto se realiza en base a la siguiente ecuación:

$$v_{i,j} = x_{i,j} + \Phi_{ij}(x_{i,j} - x_{k,j}); \quad (5)$$

Donde k es una solución en el vecindario de i, *j* es el parámetro a actualizar (centroide elegido aleatoriamente) y Φ_{ij} es un número aleatorio en el rango [-1,1].
Después de producir una nueva fuente, la suma de los cuadrados del error es calculada por medio del algoritmo K-medias, y posteriormente, se evalúa su calidad (rentabilidad) en base a la expresión (4). Se aplica una selección avara entre esta nueva fuente $v_{i,j}$ y la fuente padre $x_{i,j}$. Después de esto, las empleadas comparten su información con las abejas observadoras, esperando en la colmena mediante la danza de las abejas.
En la fase de las abejas observadoras, éstas eligen probabilísticamente su fuente de comida, dependiendo de la información provista por las abejas empleadas. Para este propósito, una técnica de selección basada en calidad puede ser usada, como el método de selección de la rueda de roulette. Este método indica que la probabilidad de una solución de ser escogida para su modificación, es proporcional a la aptitud de dicha solución, y aquellas fuentes con mayor calidad tienen más probabilidades de ser elegidas. Funciona de forma similar a una torta dividida en proporciones (porcentajes). El método de la rueda de roulette es el algoritmo más simple de selección en algoritmos genéticos, y la ecuación usada viene dada por:

$$p_i = \frac{fit_i}{\sum_{i=1}^{SN} fit_i} \quad (6)$$

Donde SN es el total de soluciones (fuentes de comida) siendo explotadas. Después de que una fuente de comida es probabilísticamente elegida por una observadora, una fuente vecina es producida (la fuente padre es modificada en base a las fuentes de su vecindario) por medio de la ecuación (5) y se calcula su calidad o aptitud por medio de (4). Así como en la fase de las abejas empleadas, se aplica una selección avara entre estas dos fuentes.
En la fase de las abejas observadoras, las fuentes de comida que no pudieron ser mejoradas a lo largo de un número determinado de pruebas llamado “límite”, se dice que han sido agotadas y su solución es abandonada. Después, las exploradoras comienzan a buscar por nuevas soluciones aleatoriamente (diversificación del espacio de búsqueda), por medio de:

$$x_{ij} = min_j + r(max_j - min_j); \quad (7)$$

Donde r es un número aleatorio que se distribuye uniformemente en el intervalo [0,1], max_j es la cantidad de elementos en el conjunto de datos, min_j es 0 y j corresponde a cada parámetro de la fuente de comida (centroides). Así, aquellas fuentes de comida que eran inicialmente pobres o se hicieron pobres mediante el proceso de explotación, son abandonadas y surge el comportamiento de retroalimentación negativa para compensar la creación de estructuras y explorar el espacio de soluciones (intensificación y diversificación).
Estos pasos se repiten hasta que se satisface un criterio de parada, como tiempo de CPU o número de ciclos. Para más detalle sobre el algoritmo ABC original, ver Karaboga, Gorkemli, Ozturk y Karaboga (2014) y Armano y Farmani (2014).

DESCRIPCIÓN DE LA PROPUESTA

Diseño de la herramienta computacional

A continuación, se describe el diseño y las funcionalidades de la herramienta computacional desarrollada para el agrupamiento de datos basado en el comportamiento colectivo de las abejas (ver figura 1 y tabla 3).

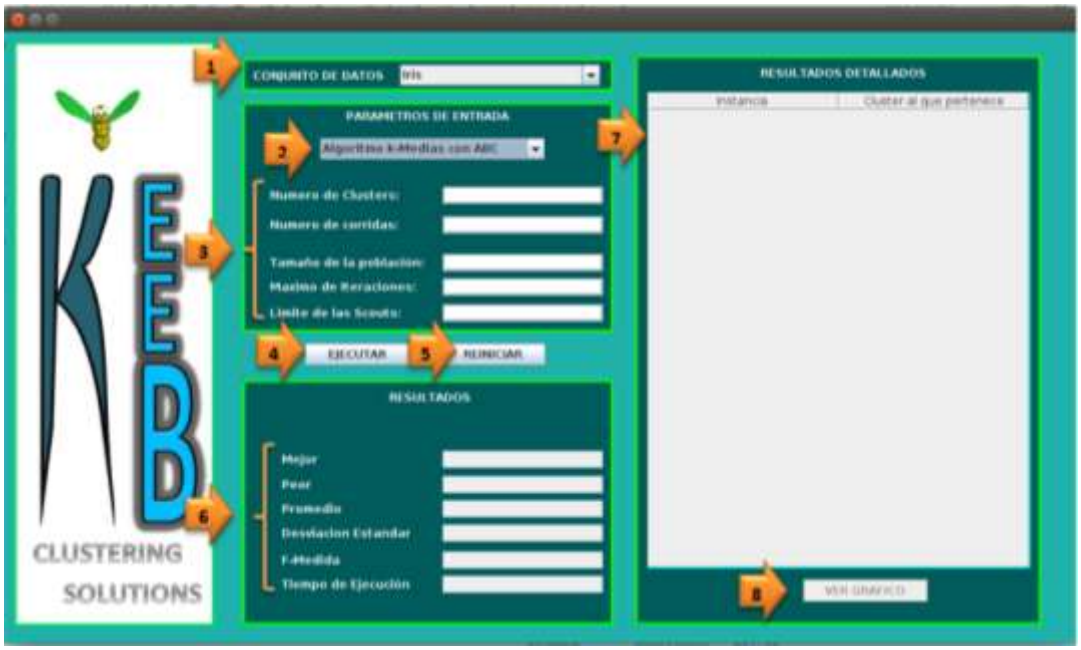

Figura 1. Interfaz gráfica de la herramienta computacional desarrollada (Perozo, Gutiérrez y Pérez, 2016).

Tabla 3. Descripción de cada una de las opciones establecidas en la herramienta computacional (Perozo, Gutiérrez y Pérez, 2016).

N. Ítem	Descripción
1	En este ítem se selecciona el conjunto de datos en el que se aplicarán los algoritmos de agrupamiento, teniendo las siguientes tres (3) opciones: Iris, Wine y CMC.
2	En este ítem se selecciona el algoritmo a utilizar en el conjunto de datos previamente seleccionado, teniendo actualmente las siguientes opciones: K-medias, Algoritmo K-medias con ABC.
3	En estos campos de texto se ingresan los parámetros necesarios para ejecutar el algoritmo previamente seleccionado. En caso de ser el algoritmo K-medias, sólo se permite ingresar el número de grupos (clusters) y el número de ejecuciones o corridas. Si el algoritmo seleccionado es el ABCk, se debe, además, ingresar el tamaño de la población de abejas (agentes del sistema), el máximo número de iteraciones (criterio de parada) y el límite para las abejas exploradoras.
4	Este botón sirve para ejecutar el algoritmo seleccionado en un conjunto de datos determinado.

5	Esta opción sirve para limpiar o borrar el texto ingresado en los campos de texto de los parámetros, de modo que puedan ingresarse parámetros nuevos.
6	En estos campos de texto se muestran los resultados arrojados por el algoritmo seleccionado. Estos son: mejor solución, peor solución, promedio, desviación estándar, medida F y tiempo de ejecución.
7	En esta tabla se muestran todas las instancias con su número y el grupo al que pertenecen, luego de haber aplicado el algoritmo. Esto, para ofrecer de forma más detallada el resultado final de la tarea de agrupamiento.
8	Esta opción permite mostrar los resultados gráficos arrojados por el algoritmo, que consisten en un plano cartesiano, donde el eje horizontal lo conforman las instancias del conjunto de datos y el eje vertical se compone de los grupos ingresados por el usuario. Cada grupo está representado por un color distinto (ver figura 2).

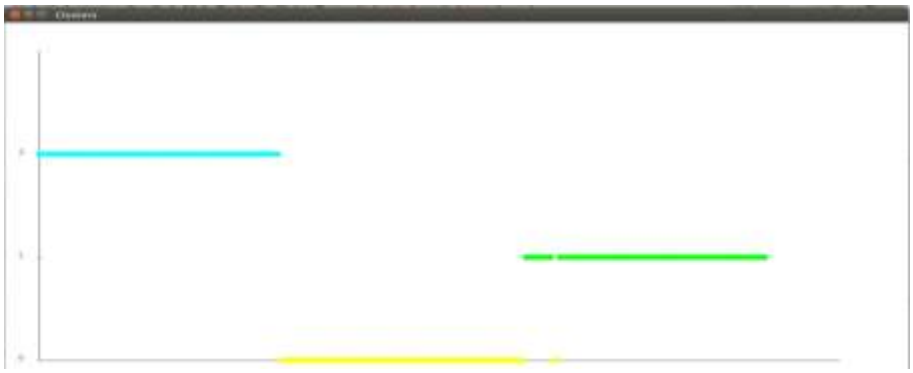

Figura 2. Pantalla para visualizar el agrupamiento de datos obtenidos (Perozo, Gutiérrez y Pérez, 2016).

Conjunto de datos reales utilizados en la experimentación

Los resultados experimentales para comparar el enfoque de agrupamiento basado en el comportamiento de las colonias de abejas propuesto por Armano y Farmani (2014), en base al algoritmo K-medias, son provistos para tres (3) conjuntos de datos reales tomados del repositorio de aprendizaje automático de la Universidad de California. Los conjuntos de datos utilizados son: Iris, Wine y CMC. La elección de este conjunto de datos obedece a que constituyen bases de datos etiquetadas, lo que permite contar con agrupamientos modelo, contra los cuales es posible evaluar la calidad de los grupos obtenidos por los algoritmos. El conjunto de datos Iris es, tal vez, el más conocido en la literatura de reconocimiento de patrones. El conjunto de datos contiene tres (3) clases de cincuenta (50) instancias cada una (se tienen 150 instancias en total), donde cada clase se refiere a un tipo de la planta Iris. El problema consiste en predecir el tipo o la clase de planta Iris. El conjunto de datos Wine, es el resultado del análisis químico de vinos crecidos en la misma región de Italia, pero derivados de tres (3) distintos cultivadores. Se tienen 178 instancias y la tarea de clasificación consiste en identificar el origen del vino. El conjunto de datos

CMC (elección de método anticonceptivo) es un conjunto de muestras de mujeres casadas que no estaban embarazadas o no sabían si lo estaban al momento de la entrevista. El problema consiste en determinar el método anticonceptivo actual (ninguno, métodos a largo plazo, métodos a corto plazo) que usa una mujer basándose en sus características demográficas y socio económicas. Se tienen 1473 instancias y 9 atributos en total (entre numéricos y categóricos o nominales).

EXPERIMENTACIÓN Y RESULTADOS OBTENIDOS

Los parámetros clave del algoritmo ABCk: "tamaño de la colonia", "límite" y "número de iteraciones" fueron asignados a 10, 100 y 20 respectivamente. Los algoritmos son implementados en una computadora de procesador AMD E-350 de 1.60 GHz y 2.00 GB de RAM. Se toman en cuenta los resultados de la función objetivo y el tiempo de ejecución para cada uno de los algoritmos en 100 distintas corridas. En este sentido, la calidad del agrupamiento para cada algoritmo se basa en:

a. El criterio de distorsión o suma de los cuadrados del error (SSE) definido anteriormente en la ecuación (2). Es claro que a menor valor de la función objetivo, mejor es la calidad del agrupamiento.

b. La medida F (F-score) que usa las ideas de precisión y sensibilidad (o exhaustividad), los cuales son conceptos de recuperación de información. Cada clase *i* es considerada como el conjunto de n_i ítems deseados para una consulta. Cada grupo *j* (generado por el algoritmo) es considerado como el conjunto de n_j ítems recuperados por una consulta; n_{ij} da el número de elementos de la clase *i* en el grupo *j.* Entonces, para cada clase *i* y grupo *j* se definen la precisión y la sensibilidad como $p(i,j) = \frac{n_{ij}}{n_j}$ y $r(i,j) = \frac{n_{ij}}{n_i}$ y el valor correspondiente bajo la medida F es:

$$F(i,j) = \frac{(b^2+1) * p(i,j) * r(i,j)}{b^2 * p(i,j) + r(i,j)} \quad (8)$$

Donde *b=1* es elegido para obtener el mismo peso para la precisión y la sensibilidad. La F como medida general para todo el conjunto de datos viene dada por:

$$F^* = \sum_I \frac{n_i}{n} MAX\{F(i,j)\} \quad (9)$$

Donde a mayor valor, mejor calidad de agrupamiento.

A continuación, se muestran los resultados obtenidos por la herramienta computacional para los 3 distintos conjuntos de datos considerados en dos escenarios diferentes: aplicando el algoritmo K medias y el algoritmo ABCk.

RESULTADOS OBTENIDOS CON NUESTRA HERRAMIENTA COMPUTACIONAL

A nivel gráfico, se puede ver que al aplicar el algoritmo K-medias y ABCk (ver figuras 3 y 4), se logra un agrupamiento de datos similar en cada base de datos empleada. Ahora bien, veamos la evaluación realizada en relación a la mejor solución, peor solución, valor

promedio, desviación estándar y F-medida, a fin de verificar la calidad de los grupos obtenidos por los algoritmos en cada base de datos.

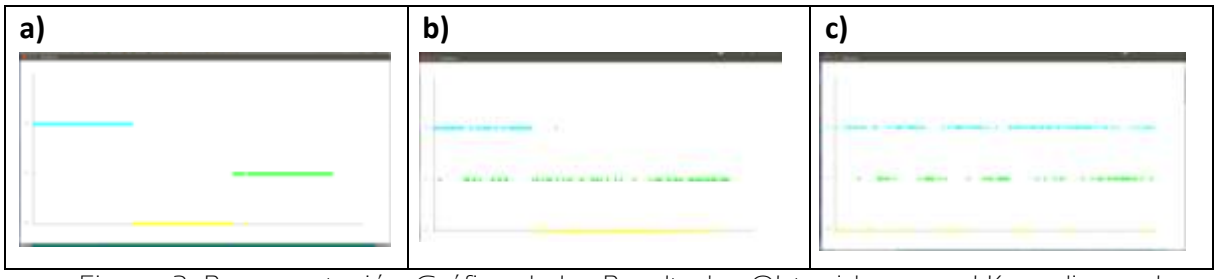

Figura 3. Representación Gráfica de los Resultados Obtenidos con el K-medias en la Base de Datos: a) Iris; b) Wine; c) CMC. (Perozo, Gutiérrez y Pérez, 2016).

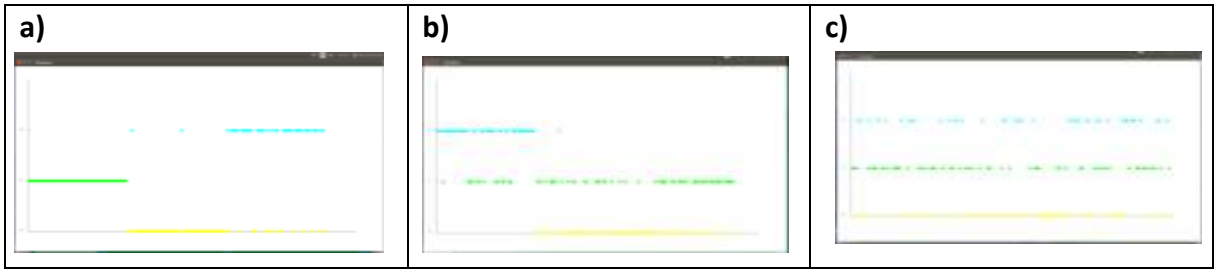

Figura 4. Representación Gráfica de los Resultados Obtenidos con el ABCk en la Base de Datos: a) Iris; b) Wine; c) CMC. (Perozo, Gutiérrez y Pérez, 2016).

Tabla 4. Resultados Obtenidos con el K-medias y ABCk en la Base de Datos: Iris. (Perozo, Gutiérrez y Pérez, 2016).

Resultados Iris.arff	Mejor	Peor	Media	Desviación estándar	F-Medida
K-Medias	97,33	123,84	103,30	10,91	84,30%
ABCk	97,33	97,33	97,33	0	86,88%

Tabla 5. Resultados Obtenidos con el K-medias y ABCk en la Base de Datos: Wine. (Perozo, Gutiérrez y Pérez, 2016).

Resultados wine.arff	Mejor	Peor	Media	Desviación estándar	F-Medida
K-Medias	16555,68	18436,95	16896,06	705,63	70,04%
ABCk	16555,68	16555,68	16555,68	0	70,94%

Tabla 6. Resultados Obtenidos con el K-medias y ABCk en la Base de Datos: CMC. (Perozo, Gutiérrez y Pérez, 2016).

Resultados cmc.arff	Mejor	Peor	Media	Desviación estándar	F-Medida
K-Medias	5542,18	5545,33	5543,62	1,55	40,38%
ABCk	5542,18	5542,18	5542,18	0	40,33%

Se puede apreciar una disminución total de la desviación estándar en todas las corridas para el algoritmo ABCk, al igual que se mejora la precisión del algoritmo (F-Medida), es decir, la calidad de los grupos obtenidos (ver tablas 4 a 6).

RESULTADOS OBTENIDOS EN TRABAJO SIMILAR

Ahora bien, al comparar los resultados obtenidos por Armano y Farmani (2014), con los obtenidos al aplicar la herramienta computacional propuesta (ver tablas 7 a la 9), se logra ver que se obtienen resultados similares después del proceso de calibración realizado con los parámetros clave: tamaño de la colonia, límite y número de iteraciones.

Tabla 7. Resultados Obtenidos por Armano y Farmani (2014), con el K-medias y ABCk en la Base de Datos: Iris.

Resultados Iris.arff	Mejor	Peor	Media	Desviación estándar	F-Medida
K-Medias	97,33	123,97	102,73	10,52	87,33%
ABCK	97,33	97,33	97,33	0	89,25%

Tabla 8. Resultados Obtenidos por Armano y Farmani (2014), con el K-medias y ABCk en la Base de Datos: Wine.

Resultados wine.arff	Mejor	Peor	Media	Desviación estándar	F-Medida
K-Medias	16555,68	16923,11	16890,16	718,65	70,22%
ABCK	16436,95	16678,52	16574,49	188,13	71,47%

Tabla 9. Resultados Obtenidos por Armano y Farmani (2014), con el K-medias y ABCk en la Base de Datos: CMC.

Resultados cmc.arff	Mejor	Peor	Media	Desviación estándar	F-Medida
K-Medias	5840,44	5934,50	5864,22	51,32	39,71%
ABCK	5700,82	5764,27	5711,27	3,41	42,31%

Cabe señalar, que la fase de agrupamiento inmersa en el algoritmo ABCk, realizada mediante el K-Medias, no tiene especificado un valor para el máximo número de iteraciones usadas en el trabajo de Armano y Farmani (2014). En la implementación propuesta, este parámetro es inicializado en 100.

También es importante resaltar que Armano y Farmani (2014), no especifican una forma para trabajar los atributos nominales del conjunto de datos CMC. En nuestra implementación, estos atributos se tratan como números enteros. Esto podría explicar la diferencia entre los resultados obtenidos en este trabajo y los presentados por los autores referidos.

CONCLUSIONES Y TRABAJO FUTURO

En este trabajo se implementó una herramienta computacional como un sistema multiagente, que simula el comportamiento colectivo de las colonias de abejas para realizar la tarea de agrupamiento con el algoritmo K-medias. Además, al comparar los resultados obtenidos en la experimentación realizada con trabajos similares como Armano y Farmani (2014), se pudo verificar tanto el diseño como implementación de la herramienta, ya que pudieron reproducirse los resultados en las mismas condiciones.

La debilidad más importante del algoritmo K-medias se encuentra en su inicialización, ya que al ser aleatoria, tiende a quedarse estancado normalmente en un óptimo local, debido a que depende netamente de los centroides iniciales. Este inconveniente del k-medias es superado con el ABCk, ya que usa las múltiples interacciones entre los agentes del sistema para encontrar mejores soluciones, logrando, así, una variación menor entre la mejor solución, la peor solución y la solución media para las tres (3) bases de datos consideradas en 100 corridas distintas.

Esta herramienta computacional para el agrupamiento de datos, con sus funcionalidades, favorece su uso en la minería de datos en diversas aplicaciones, tanto a nivel académico como comercial. Adicionalmente, la generalidad de la herramienta computacional permite que la misma sea flexible ante propuestas y modificaciones planteadas por investigadores y estudiantes, que en un futuro deseen añadir más algoritmos, parámetros y resultados.

Existen diversos enfoques para inicializar los algoritmos de agrupamiento, de modo que se obtengan mejores centroides iniciales. Algunas de estas propuestas involucran el uso de heurísticas y otras técnicas inteligentes, como, por ejemplo, el algoritmo de optimización por enjambre de partículas PSO (“Particle Swarm Optimization”, en inglés), cuyo funcionamiento se inspira en el comportamiento de las bandadas de pájaros. Debido a las bondades del PSO, se propone como trabajo futuro incorporar el PSO al algoritmo ABC, de manera que se pueda optimizar la etapa de explotación de las abejas y, así, posiblemente, disminuir la velocidad de convergencia.

REFERENCIAS

K. Cios, W. Pedrycz, R. Swiniarski & L. Kurgan. (2007). Data Mining: A Knowledge*. Discovery Approach.* Springer.

D. Karaboga, B.Gorkemli, C.Ozturk & N. Karaboga (2014). A comprehensive survey: artificial bee colony (ABC) algorithm and applications. *Artificial Intelligence Review: an International Science and Engineering Journal,* vol. 42*,* No. 1, pp.21-57.

G. Armano & M. Farmani (2014). Clustering Analysis with Combination of Artificial Bee Colony. Algorithm and K-Means Technique. *International Journal of Computer Theory and Engineering*, Vol. 6, No. 2, pp.141-145.

A. Dymnicki & D. Henry (2011). Use of Clustering methods to understand more about the case. *Methodological Innovations Online*, vol. 6, No. 2, pp.6-26,

D. Filho, E. Da Rocha, J. Da silva, R. Paranhos, M. Da Silva & B. Felix (2014). Cluster Analysis for Political Scientist. *AppliedMathematics*, vol. 5, pp.2408-2415.