Artículo original

Clasificación no supervisada de imágenes médicas y minería de datos. Algoritmo S3 vs. K-medias

Unsupervised classification of medical images and data mining: S3 algorithm vs. k-means

Reinaldo Sánchez Álvarez[1*] https://orcid.org/0000-0001-6605-6590

[1]Universidad de Guantánamo. Cuba.

*Autor para la correspondencia: reinaldo.sa@cug.co.cu

RESUMEN

Uno de los desafíos que los programadores tienen que enfrentar es la alta dimensión de grupos de datos. El proceso de reconocimiento de patrones en imagen y la minería de datos para los volúmenes grandes de información son ejemplos de ellos, optimizar la cantidad de veces que se recorre el conjunto de datos, disminuye el tiempo de procesamiento. Éste documento tiene el objetivo de caracterizar el algoritmo de tres pasos (S3), paralelo a K-medias, como una alternativa para afrontar la alta dimensión del conjunto de datos, en la clasificación no supervisada de imagen. Para el análisis de la concurrencia, se escoge, flujo de datos y el esquema instrucción única con datos múltiples. El resultado obtenido confirma que la concurrencia en ambos es posible, S3 no depende de la selección inicial de los representantes y puede ser el proceso de escogimiento de los primeros vectores centrales en K-medias. S3 es una alternativa a ser tenida en cuenta en la clasificación no supervisada de imágenes médicas y procesos de minería de datos.

Palabras clave: medias; algoritmo; imágenes médicas; clasificación no supervisada; representantes; minería de datos.

ABSTRACT

One of the challenges to be faced by programmers is the large dimensions of data groups. The process of pattern recognition in images and data mining for great volumes of information is an example. Optimizing the number of times that the set of data is run saves processing time. The purpose of the study was to characterize the three-step (S3) algorithm, parallel to k-means, as an alternative to cope with the large dimension of the data set in unsupervised image classification. Concurrence analysis is based on data flow and the single instruction multiple data scheme. The result obtained confirms that concurrence of both is possible. S3 does not depend on initial selection of representatives, and may be the process for selection of the first central vectors in k-means. S3 is an alternative to be considered in the unsupervised classification of medical images and data mining processes.

Keywords: means; algorithm; medical images; unsupervised classification; representatives, data mining.

Recibido: 28/01/2021

Aceptado: 05/02/2021

Introducción

El aprendizaje computacional, es un grupo de métodos que pueden detectar automáticamente patrones ocultos, para predecir datos futuros.

Una imagen es una matriz de celdas, donde cada celda se denomina píxel. A cada píxel se le asigna un valor digital, que corresponden a la reflectividad discretizada, recogida por un sensor específico. Por lo tanto, una imagen multiespectral es un conjunto de matrices, con las mismas propiedades geométricas, donde cada matriz

almacena el valor de reflectancia de los píxeles en un intervalo de longitud de onda concreto del espectro electromagnético.[1]

Las técnicas clásicas de clasificación basada en píxeles pueden ser supervisadas, no supervisadas o mixtas. El método supervisado a pesar de ser el más preciso, al requerir una interpretación por medio de la delimitación de áreas de entrenamiento, requiere de un arduo trabajo de recolección de muestras en campo. Además, los resultados de una clasificación supervisada dependen no solo de la capacidad del algoritmo utilizado para discriminar las categorías; sino también de supuestos con respecto al comportamiento de las categorías.[2]

En el proceso de segmentación de una imagen, los algoritmos de segmentación consideran una imagen I(X) con N píxeles, donde X = (x, y), $X \in R^d$ representa las coordenadas de los píxeles en la imagen y cada pixel se denota como x_i, i = {1,2,.., N}, R^d es la representación de la imagen en el espacio de color RGB, donde d = 3.[3]

A pesar del aumento continuo de la velocidad de procesamiento y la capacidad de almacenamiento obtenida por la industria de equipos computacionales, existe una amplia gama de aplicaciones donde la computadora tradicional más rápida no puede operar en un tiempo razonable. Dichas tareas representan una demanda inmediata, siendo inviable esperar el aumento de la velocidad de los componentes hasta el punto de hacer posible su realización.

Una forma de aumentar la potencia computacional es el uso de múltiples procesadores que funcionan juntos en una misma tarea. El problema más grande se divide en partes; cada parte se resuelve mediante un procesador que funciona en paralelo. Una computadora paralela puede ser una computadora específica, que contiene múltiples procesadores interconectados, o incluso varias computadoras independientes conectadas a través de una red.

El objetivo de este estudio es, caracterizar el algoritmo de tres pasos (S3), paralelo a K-medias, como una alternativa para afrontar la alta dimensión del conjunto de datos, en la clasificación no supervisada de imágenes médicas y a grandes volúmenes de datos.

Métodos

Para desarrollar la investigación, fue considerado una exploración orientada a la a caracterizar el algoritmo S3 en paralelo con k-medias, teniendo en cuenta: la programación concurrente, selección inicial de los representantes y la cantidad de veces que se recorre el conjunto de datos. Se utilizó Google, seleccionándose los documentos de interés referidos a estos 3 pilares. Para la caracterización de la concurrencia fue escogida la arquitectura SIMD (simple instrucción, múltiples datos), teniendo en cuenta los errores al fusionar los resultados de cada tarea en paralelo.

la implementación de S3, se utilizó el lenguaje de programación c++, el IDE de programación QT 5.14.1. Mediante reingeniería se realizó la reutilización de los módulos ya implementado en la plataforma adimg. Es una plataforma de aprendizaje computacional, en desarrollo, de la universidad de Guantánamo y está dirigida por el autor.

Para crear los patrones se utiliza como características, convolución de matrices de dimensión 3x3, con ventanas deslizantes. La plataforma adimg dispone de hasta 12 atributos para cada instancia. La métrica de similitud que se utiliza es la distancia euclidiana.

Alta dimensión del grupo de instancias

La alta dimensión del grupo de instancias está presente en el análisis de datos de expresión genética,[4] así como en el reconocimiento de patrones de imágenes,

debido al hecho de que el progreso tecnológico permite almacenar bandas espectrales de alta dimensión.

Modelo de programación recurrente

Es frecuente clasificar arquitecturas paralelas mediante dos conceptos: flujo de instrucciones y flujo de datos. Un flujo de instrucciones corresponde a un contador de programa, un sistema con n CPU que poseen n contadores de programa y n flujos de instrucciones.[5]

Descomposición funcional

La descomposición funcional tiene que ver con el hecho de hacer que cada procesador realice una determinada tarea, y cada tarea es responsable de una parte de todo el proceso.[5] Las arquitecturas paralelas se clasifican en:

- SIMD (simple instrucción, múltiples datos)
- MMD (múltiples instrucciones, múltiples datos)

Modelo SIMD

Esta arquitectura considera múltiples microprocesadores idénticos, donde cada uno posee una memoria local y ejecuta la misma secuencia de instrucciones a los diferentes datos, requiere menos memoria y el esquema de los programas es más complejo.[5]

Satyanarayana,[6] realiza una discusión sobre la complejidad del modelo K-medias, implementar la programación paralelo con esquema SIMD es factible.

Procesamiento recurrente de imágenes

Un enfoque directo, es dividir las imágenes en varias particiones, ejecutar una tarea paralela en cada partición al mismo tiempo y combinar los resultados de cada procesamiento. Un método de partición operacional es el método basado en áreas, o dominio, que divide una escena de imágenes en sub-rectángulos de igual dimensión de acuerdo con los valores de abscisa y ordenada.

Castillo Reyes realiza una discusión sobre éste enfoque como descomposición en dominios.[7]

En resumen, la aplicación del método de partición en áreas, representa un resultado parcial. Simplemente fusionar los resultados parciales conduce a graves errores de segmentación, ya que la agrupación necesita la información global de toda la imagen.

Un método para resolver este problema, es utilizar un proceso maestro encargado de almacenar todos los datos en la memoria del sistema, con el objetivo de que otros hilos de procesos accedan a los datos a través de la comunicación con el proceso maestro.[7] La descomposición en dominios usa un proceso iterativo. Primero ejecutan la etapa del agrupamiento en cada partición de datos en paralelo, para obtener resultados parciales, luego se calcula un resultado global conforme a una función específica, como votar, o promediar.[7]

La biblioteca GDAL es utilizada para el procesamiento paralelo de datos raster. *Castillo Reyes* describe la tendencia de utilización de esta biblioteca.[7]

Algoritmo K-medias

K- medias fue creado por *MacQueen* en 1967 y es reconocidos como uno de los algoritmos más simples.[8] También *Barba* y otros,[9] lo exponen como uno de los utilizados con más frecuencia. La idea del algoritmo K-medias (también llamado K-promedios) es proporcionar una clasificación de la información de acuerdo con los propios datos.

El objetivo del algoritmo es minimizar una función de error cuadrático:[9,10]

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

El algoritmo K-medias se basa en la minimización de una medida, la distancia interna entre los patrones de una agrupación. La minimización del costo garantiza encontrar un mínimo local de la función, objetivo que dependerá del punto inicial del algoritmo.[11]

K-medias

m: *Amounts of instances.*

g: *Amounts of Cluster.*

S(i): *it Lists of labels of instances with m, where i = 1, 2, ..., m*

$$I_{ij} = \{I_j(C_i)\}_{i=1,..,n}^{j=1,..,m}$$ *Study Objects (instances)*

$$M_{ij} = X^0 = \{I_{h1}, I_{h2}, \dots, I_{hk}\}$$ *Initial Centroids*

$$MT_{ij} = X^0 = \{I_{h1}, I_{h2}, \dots, I_{hk}\}$$ *Initial Centroids*

$$MT_{ij} = X^0 = \{I_{h1}, I_{h2}, \dots, I_{hk}\}$$ *Centroids previous Step*

Cuadro 1

1	*begin*
2	*repeat*
3	*for i =1 until m do*
4	*begin*
5	$S(i) = IndexKOf(Min(I_i, M_k))$ *// it Labels of most similar Centroid // to instances i*
6	*MT(S(i))=Sum(I_i) // it Adds instances i to Centróide S(i)*
7	*end*
8	$MT1 = M$ *// it copies Centroids*
9	$M = means(MT)$ *// new Centroids*
10	*MT=MT1 // Centroids previous iterance*
11	*until (MT=M) // stop condition*
12	*end.*

Enfoque de la programación paralela de K-medias

En cada iteración, una instancia seleccionada se procesa con el mismo grupo de instrucciones, entonces, se puede definir un cierto número de procesos para que procesen cada subconjunto de instancias. Si elegimos 10 procesos para un grupo de 100 mil instancias, entonces, cada uno de estos procesos ejecuta la tarea de agrupamiento para 10 mil instancias. Se necesitan dos parámetros de entrada para cada proceso: Principio y fin, que definen los subconjuntos de datos a procesar para cada uno de los 10 procesos. Proceso1 (1, 10 000), Proceso 2 (10 001, 20 000) ..., Proceso 10 (90 001, 100 000).

Algoritmo S3

Para realizar la clasificación, se realizan tres exploraciones al conjunto de datos, como se observa en el cuadro 2.

1. En la primera exploración se crean los vectores I_{max} e I_{min} con los valores máximos y mínimos de las coordenadas de los patrones o instancias.

$$\boldsymbol{I_{max}}\begin{pmatrix} Max\{C_1\}, \\ Max\{C_2\}, \\ \dots, \\ Max\{C_n\} \end{pmatrix} \qquad \boldsymbol{I_{min}}\begin{pmatrix} Min\{C_1\}, \\ Min\ \{C_2\}, \\ \dots, \\ Min\{C_n\} \end{pmatrix}$$

Estos vectores permiten elaborar una bifurcación, que consiste en un segmento dividido en partes iguales, estas divisiones se forman de la siguiente manera: Se calculan las normas de ambos vectores.

$$\|I_{max}\| = \sqrt[2]{I_{max}(1)^2 + I_{max}(2)^2 + .. + I_{max}(k)^2}$$

$$\|I_{min}\| = \sqrt[2]{I_{min}(1)^2 + I_{min}(2)^2 + .. + I_{min}(k)^2}$$

$$\Delta\mu = \frac{|\|I_{max}\| - \|I_{min}\||}{\omega}$$

Donde ω es la cantidad de grupos a formar dentro del conjunto de datos, la bifurcación será un vector de dimensión ω y sus coordenadas se expresan con la siguiente expresión.

$$\begin{pmatrix} o_1 = \|I_{max}\|, \\ o_2 = o_1 - \Delta\mu, \\ ., \\ ., \\ ., \\ o_w = \|I_{min}\|, \end{pmatrix}$$

2. En la segunda exploración las instancias similares se juntan con la misma etiqueta que identifica a un grupo.

Se dice que una instancia pertenece a un grupo O_k de los posibles ω, cuando la norma de la instancia está más cerca de la coordenada O_k, que de cualquiera de las coordenadas restantes:

$$I_i \in O(h)\ if\ \ \forall\ k \neq h\ \ |\|I_i\| - O(k)| > |\|I_i\| - O(h)|\ onde\ (2 \leq k\ \leq\ \omega)$$

Esta exploración concluye con la selección de los centroides para cada grupo, estos representantes son aquellas instancias cuya norma es la más cercana de las coordenadas del segmento de bifurcación.

$$I_k^R\ \ is\ a\ representative\ of\ \ \ O(k)\ se\ \ \forall\ i \neq R\ \left|\left\|I_k^i\right\| - O(k)\right| > |\|I_k^R\| - O(k)|\ onde\ (1 \leq i \leq\ |\{I_i \in O(k)\}|)$$

3. La tercera exploración es la formación de los grupos, donde todas las instancias se agrupan de acuerdo con la proximidad a los centroides de cada grupo seleccionado en la segunda exploración.

$$I_h\ belongs\ to\ the\ group\ \ k\ if\ \ \forall i \neq k\ \ \left\|I_h - I_i^R\right\| > \|I_h - I_k^R\|\ onde\ (1 \leq k\ \leq\ \omega)$$

Algoritmo S3

m: Amounts of instances.

n: Amounts of characteristics.

g: Amounts of cluster.

$$I_{ij} = \left\{I_j(C_i)\right\}_{i=1,..,n}^{j=1,..,m}$$ Study Objects (instances)

S(i): it Lists of labels of instances with m, where i = 1; 2; ..., m.

Imax(i): maximum Vector of the characteristics i = 1; 2; ..., n.

Imin(i): minimum Vector of the characteristics i = 1; 2; ..., n.

Cuadro 2

1	*Begin*
2	*// Exploration One*
3	$Imax = I_1$
4	*for i =1 ate m do*
5	*For j=1 ate n*
6	*Begin*
7	$if(Imax(j) < I_i(j))\ Imax(j) = I_i(j)$ *// Vector maxiof all of the instances*
8	$if(Imin(j) > I_i(j))\ Imin(j) = I_i(j)$
9	*End*
10	$O = GetForquilha(Imax)$ *// creating the fork*
11	*// Exploration Two*
12	*for i =1ate m do // m- instancias*
13	$S(i) = IndexKOf(Min(O(k), I_i))$ *// it Labels in the most similar Fork*
14	$To\ upgrade\ Representatives\ (I_k^R, I_i)$
15	*// ExplorationThird*
16	*for i =1 ate m do*
17	$S(i) = IndexKOf(Min(I_k^R, I_i))$ *// it Labels in the most similar of Centroids*
18	*end.*

Aproximación de la programación paralela para S3

La primera exploración se encuentra entre las líneas 4 y 9. Las líneas 7 y 8 son ejecutadas en cada iteración, un vector de los valores máximos y mínimos de las características son calculados. Muchas tareas en paralelo pueden modificar el

vector de máximos o mínimos al mismo tiempo. Este es un elemento a tener en cuenta, ya que es un área de memoria crítica. En la segunda exploración en la línea 13, no se produce evento con memoria crítica, en la línea 14 se manifiesta un evento crítico. Mientras, en la tercera exploración no se produce evento crítico. El conjunto de datos se puede dividir en subconjuntos que son procesados por el mismo grupo de instrucciones en concurrencia.

Resultados

S3 al igual que K-medias tiene la desventaja de que depende de los valores iniciales del parámetro K (cantidad de cluster por descubrir), sin embargo, no depende de la selección de los centroides iniciales. En K-medias y S3, el parámetro K se puede calcular por algunos de los métodos que se utilizan para determinar la cantidad de grupos que predisponen el conjunto de datos. En este documento no se incluye dicho desafío (tendencia al agrupamiento).

El algoritmo S3 realiza solo tres exploraciones al conjunto de datos, mientras que K-medias depende del conjunto de datos (es muy difícil su convergencia en la segunda iteración). Esta ventaja, hace que sea una alternativa a tener en cuenta. Para evitar regiones críticas en la primera exploración con S3, se utilizan matrices dinámicas para cada uno de los hilos definidos. Cuando terminan todos los hilos se realiza la selección de los vectores máximos y mínimos de entre todos los seleccionados por cada tarea en paralelo. En la segunda exploración, por cada hilo se implementa una tabla hashing. Al concluir todos los hilos se seleccionan los centroides, teniendo en cuenta cada una de las tablas hashing creadas por cada hilo.

Algoritmo K-medias y S3

Para realizar una comparación entre K-medias y S3, usaremos los siguientes tópicos: Selección inicial de los centroides, cantidad de exploraciones al conjunto

de datos y concurrencia. La siguiente tabla muestra la ventaja de S3 con relación a K-medias.

Tabla - Comparación de algoritmo K-medias y S3

Tópicos	K-medias	S3
Selección de los centroides iniciales.	Influye en la convergencia, existen variantes para la selección de los centroides iniciales.	No hay dependencia.
Cantidad de exploraciones del conjunto de datos.	Depende de la convergencia.	Realiza tres exploraciones.
Concurrencia, en relación a la forma de fusionar los resultados de tareas en paralelo.	En cada iteración se realiza ponderación y se actualizan para cada tarea en paralelo los valores de los parámetros globales.	En la primera exploración, se obtiene el mínimo y máximo global, de entre los mínimos y máximos por cada tarea en paralelo. En la segunda exploración se obtienen los representantes por votación.

El algoritmo K-medias sigue siendo objeto de estudio por parte de la comunidad científica. Desde su aparición, se han presentado muchos artículos relacionados con diferentes aspectos del algoritmo. De manera general se han identificado dos vertientes importantes. La primera está enfocada en artículos que analizan la aplicación del algoritmo K-medias para resolver un problema de un dominio particular y la segunda está enfocada en artículos que proponen una mejora de la etapa de inicialización de los centroides.[11]

Debido a que la elección de los centroides iniciales impacta en la solución del agrupamiento, no existe un método generalizado. Algunas alternativas están basadas en el uso de información de la media y la desviación estándar de los atributos (características) del conjunto de datos, o utilizando las dos variables. Asimismo, el uso de estructuras de datos representa una alternativa de mejora, por ejemplo, con el uso de información de densidad de regiones como en el caso de kd-trees. Otras alternativas han asignado peso a los grupos y optimizan la

función objetivo.[11] También *Pham* y otros[12] muestran una nueva variante de K-medias y los fundamentos teóricos. *Medina Veloz* y otros[13] exponen que el modelo K-medias puede ser calibrado con el lenguaje estadístico R, es decir se ejecutan diferentes variantes de selección de centroides iniciales, y luego realiza una votación.

Con el algoritmo S3, en la segunda exploración se determinan los representantes de cada grupo, y se puede combinar con el algoritmo k-medias, para seleccionar los centroides iniciales.

Entre los trabajos futuros para validar el algoritmo S3, se propone realizar experimentos mediante la utilización de los indicadores de evaluación de los resultados de la clasificación no supervisada, para una comparación más exhaustiva. K-medias se combina con otros algoritmos como es el caso de Enjambre de Partículas (*Particle Swarm Optimization:* PSO),[14] donde se hace una exposición sobre el tema. Sin embargo, sería conveniente realizar el experimento cuando se utilice la etapa de K-medias, en este ejemplo de aplicación, referido a la selección de centroides iniciales que utilizan S3.

Cuando los conjuntos de datos son a gran escala, dígase gran número de instancias, gran número de variables de entrada, o gran número de variables de salida, entonces se presentan serias limitaciones en cuanto a la eficiencia de los algoritmos que los utilizan. Hay muchos trabajos de investigación que se centran en resolver los problemas de escalabilidad causados por un gran número de instancias de datos, como son los métodos de selección de instancias de *Brighton* y *Mellish.*[15] *Rodríguez Álvarez* y otros, describen un método basado en prototipos.[16] Una variante de usar las dos primeras exploraciones de S3, es adicionar matrices dinámicas para n vecinos más cercanos, similar al enfoque basado en prototipos, lo cual es una alternativa a estudiar para la selección de instancias.

Para extraer los patrones de la imagen, la plataforma adimg utiliza convolución de matrices, similar a como se aplica en el proceso de filtrado utilizado por *Giménez P* y otros.[17] Al igual que K-medias, S3 es aplicable a conjuntos de datos cuantitativos.[18]

Aplicación en el procesamiento de imágenes

Mostramos el procesamiento de una imagen médica, con la intención de revelar el uso del algoritmo para ayudar en la interpretación de imágenes médicas. Una vez que el proceso de clasificación es realizado por el método S3, el modelo obtenido puede ser guardado y usado en un futuro como ayuda a otras interpretaciones. El análisis del médico no está incluido porque esta discusión es una demostración de la aplicación del algoritmo S3 que funciona en un entorno paralelo.

En la figura 1 se muestra la imagen de una biopsia cargada en la aplicación *adimg* (Análisis de datos de imágenes).

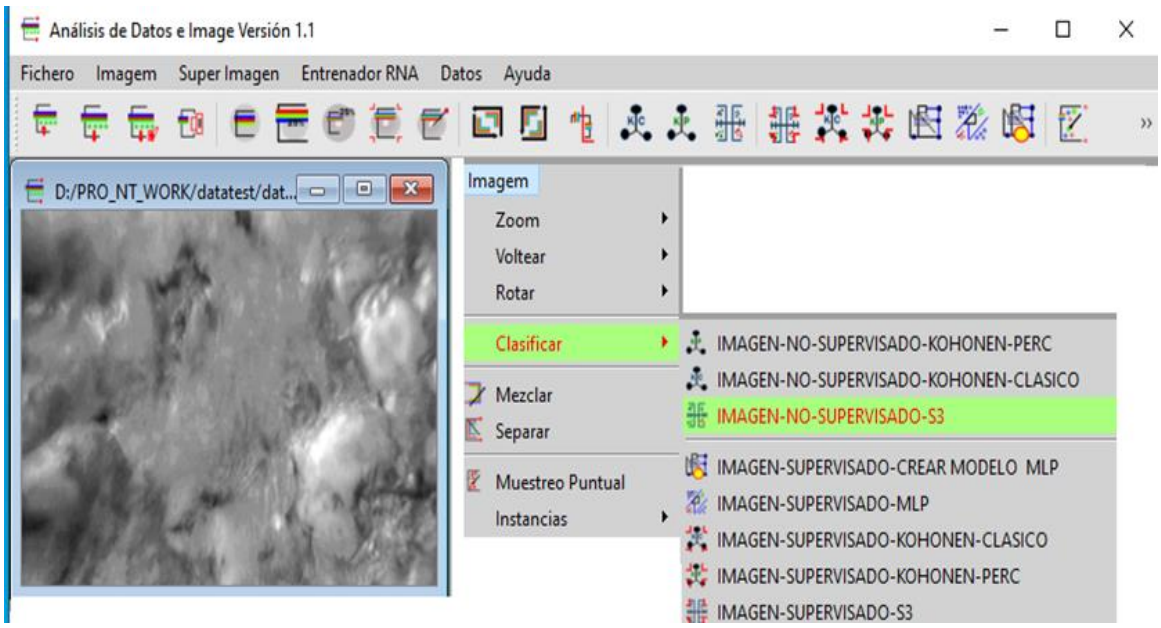

Fig. 1 - Aplicación del algoritmo S3.

En la figura 2 se observa la aplicación del algoritmo S3 con los resultados estadísticos.

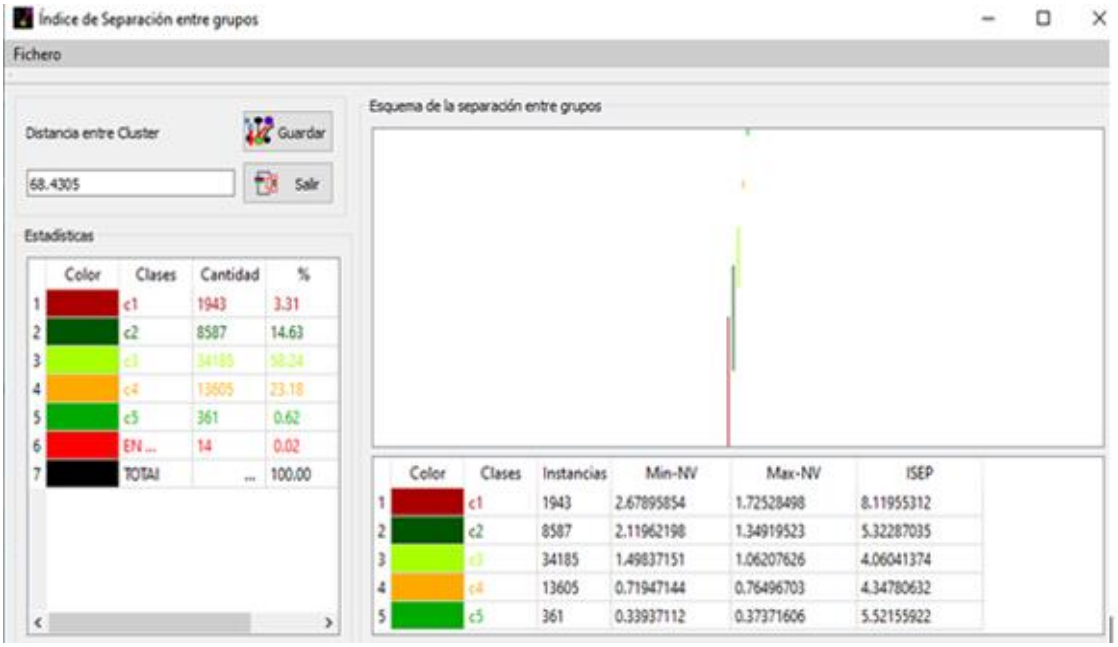

Fig. 2 - Resultado estadístico algoritmo S3.

En la figura 3 se puede apreciar la imagen segmentada.

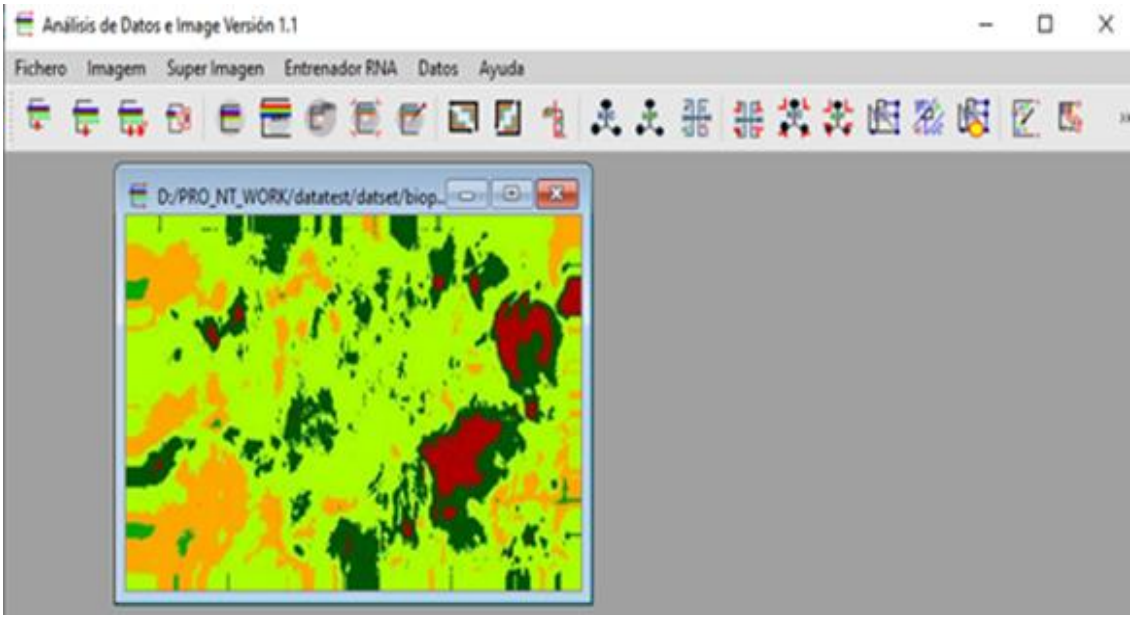

Fig. 3 - Imagen clasificada

En la clasificación realizada con el algoritmo S3, se utiliza la distancia euclidiana, sin embargo, se pueden utilizar otras distancias. *Soto* realiza un análisis de esta temática.[19] Una tendencia actual es el aprendizaje de métricas, cuyo objetivo es adaptar una función de distancia basada en pares y evaluada en los reales a un problema específico mediante la utilización de la información proporcionada por ejemplos de entrenamientos.[20]

Conclusiones

Al utilizar los algoritmos K-medias y S3 con una arquitectura SIMD, se pueden realizar procesos de clasificación no supervisada en imágenes médicas, en un tiempo relativamente corto.

El algoritmo S3 es una alternativa a tener en cuenta en los procesos de clasificación no supervisados, que no depende de la selección inicial de centroides, y solo realiza tres exploraciones de todo el conjunto de datos.

Agradecimientos

A la Dra. Zulma Luisa Barrera Jay, del Hospital Clínico Quirúrgico Dr. Agosthino Neto de Guantánamo, Departamento de Medios de Diagnóstico, por realizar la validación de las interfaces de la implementación del algoritmo S3.

A la Dra. Nora Luisa Mendoza Fonseca de la Universidad de Ciencias Médicas de Guantánamo, Departamento de Ciencia y Tecnología, por realizar la coordinación de la Universidad de Guantánamo con el hospital.

Referencias bibliográficas

1. Espínola AM. Clasificación de imágenes de satélite mediante autómatas celulares. España: Edit. Antonio Moisés Espínola; 2014.

2. Hernández Reséndiz JD, Marín Castro HM, Tello Leal E. A Comparative Study of Clustering Validation Indices and Maximum Entropy for Sintonization of Automatic Segmentation Techniques. IEEE LATIN AMERICA TRANSACTIONS. 2019 august;17(8).

3. Chuvieco E. Teledetección Ambiental. La observación de la tierra desde el espacio. Barcelona: Editorial Planeta S.A.; 2010.

4. Gómez Sánchez JA. Análisis comparativo de diferentes métodos de agrupamiento de datos de expresión genética. Tonos Digitales. 2018 Julio.

5. Pacheco P. Parallel programing with MPI. USA: Elsevier; 2011.

6. Satyanarayana A, Davidson I. Speeding up K-Means clustering using bootstrap averaging. In International Conference on Data Mining Workshop on Clustering Large Data Sets. Melbourne: IEEE; 2003. p. 19-22.

7. Castillo Reyes G. Técnicas de programación paralela aplicadas al procesamiento de datos ráster mediante la biblioteca GDAL. Revista Cubana de Ciencias Informáticas. 2016 enero-marzo;10(1).

8. Ochoa RA. Componente Web para el análisis de información clínica usando la técnica de minería de datos. Informática Médica. 2014 enero-junio;6(1).

9. Barba LAR, Guerrero HH, Salazar GJ. Análisis de Clúster para la clasificación de datos económicos. Publicando. 2016;3(7).

10. Leyva Vázquez M, González Benítez N, Hechavarría Hernández J, Rivero Peña Y, Daher Nader JE. El diagnóstico de enfermedades desde el Análisis Inteligente de los Datos. Espacios. 2018;39(28):16.

11. Pérez Ortega J, Hidalgo Reyes M, Castro Sánchez NA, Pazos Rangel R, Díaz Parra O, Olivares Peregrino V, *et al*. Una heurística eficiente aplicada al algoritmo K-means para el agrupamiento de grandes instancias altamente agrupadas. Computación y Sistemas, 2018; 22(2).

12. Pham T, Lobos AG, Vidal Silva CL. Innovación en Minería de Datos para el Tratamiento de Imágenes: Agrupamiento K-media para Conjuntos de Datos de

Forma Alargada y su Aplicación en la Agroindustria. Innovación Tecnológica. 2019;30(2):135-42.

13. Medina Veloz G, Luna Rosas FJ, Tavarez Avendaño JF, Narvaez Murillo RU. Calibración y selección del modelo de aprendizaje no supervisado K-Medias, de una encuesta sobre factores de riesgo en el consumo de drogas entre estudiantes. Revista de Análisis Cuantitativo y Estadístico. 2016 junio;3(7):1-9.

14. Chavarría MJ, Fallas MJ. El algoritmo PSO aplicado al problema de particionamiento de datos cuantitativos. Matemática, Educación e Internet. 2019 marzo;19(1).

15. Camejo Corona J, González Diez H, Morell C. Los principales algoritmos para regresión con salidas múltiples. Una revisión para Big Data. Revista Cubana de Ciencias Informáticas. 2019 diciembre;13(4).

16. Rodríguez Álvarez Y, Bello Pérez R, Caballero Mota Y, Filiberto Cabrera Y, Fernández Hernández Y, Frías Hernández M. Estudio del comportamiento de métodos basados prototipos y en relaciones de similitud ante "hubness". Revista Cubana de Ciencias Informáticas. 2017 enero;11(2).

17. Giménez Palomares F, Monsoriu Serrá JA, Alemany Martínez E. Aplicación de la convolución de matrices al filtrado de imágenes. Modelling in Science Education and Learning. 2016;9(2).

18. López D, Fernández. Aplicación en los medios de prensa de un Agrupamiento K-Means. Economía y Sociedad. 2018;12(12).

19. Soto AJ, Ponzonia I, Vázquez GE. Análisis numérico de diferentes criterios de similitud en algoritmos de clustering. Mecánica Computacional. 2006 noviembre; XXV: 993-1011.

20. Pérez Verona IC, Arco García L. Una revisión sobre aprendizaje no supervisado de métricas de distancia. Revista Cubana de Ciencias Informáticas. 2016 octubre-diciembre; 10(4).

Conflicto de intereses

El autor declara que no presenta conflicto de intereses.