Recibido, 11/05/2010 Aceptado, 18/08/2010 Artículo original

Segmentación de Imágenes Médicas Digitales mediante Técnicas de Clustering

Digital Medical Image Segmentation with Clustering Techniques

Gustavo Lorca T.[a], José Arzola R.[b], Osvaldo Pereira B.[c]

RESUMEN

En este trabajo se emplea técnicas de Clustering en la segmentación de imágenes médicas digitales para ser utilizadas en la reconstrucción de modelos anatómicos 3D a partir del estándar Digital Images and Comunication in Medicine (DICOM) con el fin de mejorar los resultados reportados en las fuentes bibliográficas. Son expuestos los algoritmos de Clustering Particional implementados y los resultados alcanzados con estos. Se compara entre sí los resultados alcanzados con ayuda de los métodos K-means y Fuzzy K-means y se recomienda procedimientos para la inicialización de los centroides.

Palabras clave: Segmentación de imágenes; Método de clustering K-means; Método de clustering Fuzzy K-means.

ABSTRACT

In this work Clustering techniques are employed in the segmentation of digital medical images to be used in the reconstruction of anatomical 3D models, starting from the standard DICOM format with the purpose of improving the results reported in the bibliographical sources. The implemented Partitional Clustering algorithms are exposed. The results reached helped by K-means and Fuzzy K-means methods are compared to each other and procedures are ecommended for the centroids initialization.

Key words: Image segmentation; K-means clustering method; Fuzzy K-means clustering method.

[1] Universidad de las Ciencias Informáticas

INTRODUCCIÓN

El veloz desarrollo de las tecnologías de adquisición de imágenes médicas digitales está revolucionando la medicina. La información contenida en las imágenes médicas es tratada por medios de cómputo que permiten mejorar la calidad. Con la utilización de estos se puede eliminar en cierta medida el ruido proveniente del equipo médico utilizado para la captura de la imagen, realizar resaltes en zonas y segmentar la imagen en diferentes partes constituyentes. Han sido creados innumerables métodos de segmentación de imágenes digitales basados en diferentes ramas de las matemáticas, entre ellos algunos de uso general y otros específicamente para un tipo de imagen, muchos juegan un papel relevante en numerosas aplicaciones. Ninguno de estos en la actualidad resuelve el problema de manera global, sino que presentan sus ventajas y desventajas según el uso que se les dé, aunque continuamente se crea nuevos métodos y se mejora los existentes obteniéndose cada vez mejores resultados y haciéndose más imperiosa su utilización.

"Si bien son muchas las opciones disponibles, aún no existen soluciones definitivas ni algoritmos generalmente aplicables, por lo que la segmentación de imágenes constituye un campo de continua investigación." (Del Fresno & J Vénere, 2002)

El Clustering de datos ayuda a discernir la estructura y simplifica la complejidad de cantidades masivas de datos. Es una técnica común y se utiliza en diversos campos, donde la distribución de la información puede ser de cualquier tamaño y forma. La eficiencia de los algoritmos de Clustering es extremadamente necesaria cuando se trabaja con enormes bases de datos y tipos de datos de grandes dimensiones. (Villagra, Guzmán, Pandolfi, & Leguizamón, 2008)

Este trabajo presenta una comparación entre diferentes métodos de segmentación de imágenes que justifica la aplicación de algoritmos de Clustering al problema de segmentación de imágenes médicas digitales, así como también son expuestos los resultados obtenidos con los algoritmos K-means y Fuzzy K-means sobre grandes volúmenes de datos como son las series de imágenes digitales pertenecientes al estándar DICOM.

MATERIALES Y MÉTODOS

Han surgido aplicaciones informáticas dedicadas al tratamiento de imágenes digitales y algunas de estas especializadas en la reconstrucción de modelos anatómicos 3D. La reconstrucción 3D es el proceso mediante el cual, objetos reales, son reproducidos en la memoria de una computadora manteniendo sus características físicas (dimensiones, volumen y forma). Estas técnicas posibilitan mejor visualización de la información obtenida por los equipos médicos especializados.

En consecuencia, en la Universidad de las Ciencias Informáticas el proyecto Visualización Médica fue creado con el objetivo de desarrollar aplicaciones que permitan la visualización 3D de modelos anatómicos obtenidos a partir del procesamiento de imágenes médicas digitales, basada en la reconstrucción tridimensional. Este proceso tiene una etapa que consiste en la segmentación de la imagen y que actualmente se realiza con la utilización de técnicas muy básicas que conllevan a que los modelos 3D no representen de forma exacta los órganos de la anatomía humana, esto trae como consecuencia que la visualización no sea realista, haciendo su interpretación más difícil para los especialistas.

Dada la situación expuesta anteriormente se plantea como problema científico: ¿Cómo lograr que los órganos de la anatomía humana representados en imágenes médicas digitales sean segmentados de forma precisa durante el proceso de reconstrucción tridimensional llevado a cabo por el proyecto Visualización Médica? Siendo así, se toma como objeto de investigación la segmentación de imágenes y es propuesto como campo de acción la segmentación de imágenes médicas digitales. Se plantea como objetivo de esta investigación la elaboración de un módulo para la segmentación de imágenes médicas digitales mediante técnicas de Clustering.

Destacan entre los métodos científicos de investigación, los siguientes:

- Métodos Teóricos:
  - Analítico – Sintético: Para concretar y resumir el conocimiento reflejado en los materiales consultados y utilizarlo en el desarrollo de esta investigación.
  - Modelación: Para predecir acontecimientos que no han sido observados aún.
- Métodos Empíricos:
  - Observación: Es el método empírico contemplativo que permite obtener información necesaria en cualquiera de las fases de la investigación.
  - Medición: Para obtener información numérica acertada de magnitudes medibles

que permitan realizar comparaciones en los resultados.

- o Experimentación Científica: Para llegar a conclusiones a través de la alteración controlada de las condiciones que permiten crear modelos, reproducir condiciones y extraer rasgos distintivos.

K-means

K-means es un método particional que intenta encontrar un número específico de grupos, los cuales están representados por sus centroides, aplicable a un grupo de objetos en un espacio continuo n-dimensional. Es uno de los algoritmos de Clustering más antiguos y ampliamente usados.

Es denominado centroide representativo de un cluster el vector formado por las medias de cada una de las componentes de los elementos pertenecientes al cluster.
La técnica general de Clustering K-means es muy simple. A continuación se presenta la descripción del algoritmo básico. (Tan, Steinbach, & Kumar, 2006)

1. Seleccionar *K* centroides, donde *K* es el número de clusters deseado.

2. Asignar cada punto al centroide más cercano y cada colección de puntos asignados a un centroide es un cluster (Región de Voronoi).

3. Actualizar los centroides de cada cluster, basados en los puntos asignados al cluster.

4. Repetir el proceso de asignación y actualización hasta que ningún punto cambie de cluster, o lo que es lo mismo, hasta que los centroides permanezcan iguales.

5. Fin.

Sea el conjunto de datos $X = \{x_1, x_2, ..., x_n\}$ Dado un centroide (representante de una agrupación) y_i, el conjunto de puntos de *X* que está más cercano a y_i que a cualquier otro centroide según la medida *d* (*x, y*) se denomina región de Voronoi de y_i y se denota por (Tan, Steinbach, & Kumar, 2006):

$$V_i = \{x \in X : d(x, y_i) < d(x, y_j), i \neq j\}$$

El número de vectores en una región de Voronoi se representa por $|V_i|$. El centroide de los vectores de una región de Voronoi viene dado por:

$$\frac{1}{|V_i|} \sum_{x \in V_i} x$$

El algoritmo descrito busca minimizar la siguiente función objetivo donde *SSE* es la suma del cuadrado de los errores, C_i es el i-esimo cluster de la partición, $d(x, C_i)$ es la medida de disimilitud o distancia entre el elemento *x* y el cluster C_i:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, C_i)^2$$

Fuzzy K-means

Fuzzy K-means (también llamado Fuzzy C-means) es una extensión del K-means. Mientras K-means encuentra particiones para las que un punto pertenece a un solo cluster, Fuzzy K-means es un método estadísticamente formalizado que encuentra *K* clusters donde un punto puede pertenecer a más de un cluster con cierto valor de pertenencia (Jain, Murty, & Flynn, 1999). Tiene su basamento en la teoría de conjuntos imprecisos o poco definidos (Fuzzy set) propuesta por Zadeh (1965) y fue creado por Ruspini (1969), Bezdek (1964) y Dunn (1974). La teoría Fuzzy set es una generalización del álgebra Booleana, por lo que la función de pertenencia de un elemento a los grupos se encuentra en el intervalo [0,1]; es decir un elemento puede pertenecer totalmente a una clase, a todas o a ninguna. (Ortega, Foster, & Ortega, 2002)

Como el K-means, Fuzzy K-means trabaja con aquellos objetos que pueden ser representados en un espacio n-dimensional con una medida de distancia definida (Jain, Murty, & Flynn, 1999). El procedimiento Fuzzy K-means minimiza la siguiente función objetivo:

$$SSE(M, C) = \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^{\phi} d_{ij}^2$$

Sujeto a: $\sum_{j=1}^{k} m_{ij} = 1 \qquad i = 1,2,...,n$

$$\sum_{i=1}^{n} m_{ij} > 0 \quad j = 1,2,...,k \quad m_{ij} \in [0,1]$$

Donde *SSE* (*M, C*) es la suma del cuadrado de los

errores dentro de las clases, M es la matriz n x k de pertenencia a los grupos (m_{ij} =1 si el elemento i pertenece totalmente al cluster j y m_{ij} = 0 es lo contrario), C es la matriz k x p de centro de las clases siendo p el número de componentes del espacio, ϕ es el grado de imprecisión de la solución (fuzziness exponent) y d_{ij}^2 es el cuadrado de la distancia entre el elemento i y el centro representativo del cluster j. El algoritmo de solución de la función objetivo consta de las siguientes etapas iterativas (Ortega, Foster, & Ortega, 2002):

Seleccionar el número de clases k, con $1 < k < n$. Si k es 1 o n el análisis no es necesario.

Seleccionar el valor de fuzziness exponent ϕ, con $\phi > 1$ Los valores comúnmente usados están en el rango 1.1 a 2.

Seleccionar la definición de distancia en el espacio variable. Las distancias más usadas son la Euclídea y Mahalanobis. Seleccionar un valor del criterio de detención, $e = 0.001$ da una convergencia razonable.

Iniciar con $M = M_0$, por ejemplo con una agrupación aleatoria o con una agrupación de partición rígida (K-means).

En las iteraciones i = 1, 2, 3,... re-calcular $C = C_i$ usando M_{i-1} con la ecuación:

$$C_j = \frac{\sum_{i=1}^{n} m_{ij}^{\phi} x_i}{\sum_{i=1}^{n} m_{ij}^{\phi}}$$

Re-calcular $M = M_1$ usando C_i y la ecuación:

$$m_{ij} = \frac{d_{ij}^{2/(\phi-1)}}{\sum_{r=1}^{k} d_{ir}^{2/(\phi-1)}}$$

Si $\|M_i - M_{i-1}\| < e$ entonces parar, sino retornar al paso 5.

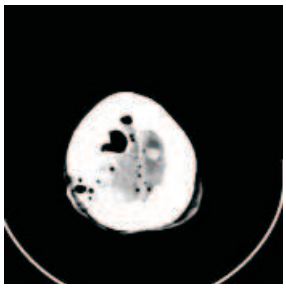
Donde $\|M_i - M_{i-1}\|$ es el mayor valor absoluto de la diferencia entre los elementos de la matriz M_i y sus correspondientes elementos de la matriz M_{i-1}.

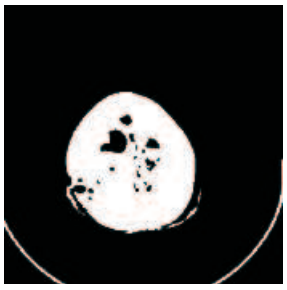
Según el criterio de detención e, el algoritmo converge con mayor o menor número de iteraciones. Sin embargo, la solución no es siempre óptima pues puede converger hacia mínimos locales en función de la estimación inicial (Oliva i Cuyás, de Cáceres Ainsa, Font Castell, & Cuadras Avellana, 2001), esta misma desventaja fue vista en el algoritmo K-means aunque este método mejora el problema de convergencia del K-means (Jain, Murty, & Flynn, 1999).

RESULTADOS

Como resultado de todo el proceso de investigación fue posible el desarrollo de un módulo de segmentación de imágenes médicas digitales mediante técnicas de Clustering que valida en buena medida la calidad de las conclusiones obtenidas. Los resultados que aquí se muestran fueron obtenidos en una PC Intel Pentium 4 a 3.00GHz y 1GB de RAM con sistema operativo Ubuntu Desktop 9.10 i386.

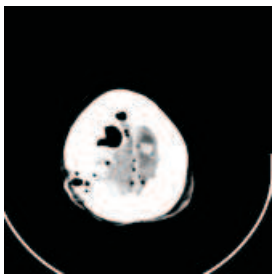
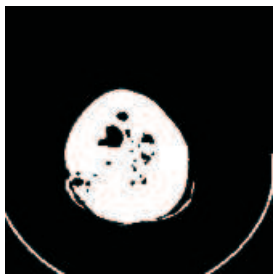
Las siguientes tablas muestran la configuración de los parámetros de inicialización del algoritmo K-means y algunos datos que permiten evaluar su funcionamiento. En las imágenes resultados del proceso de Clustering aparecen en las posiciones de los píxeles un color representativo del cluster al que pertenece:





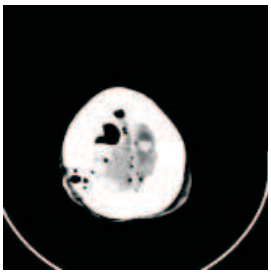
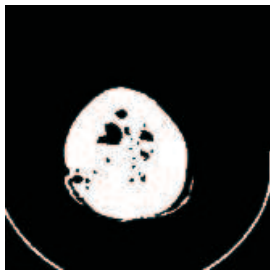
Cantidad de imágenes	Cantidad de clusters	Demora (segundos)	Inicialización de los centroides	Medida de distancia
1	2	1.43	Por frecuencia	Color

Figura 1. Experimento 1 del algoritmo K-means

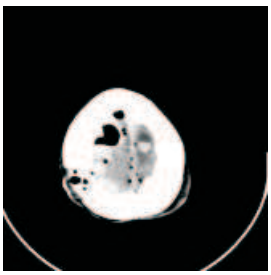
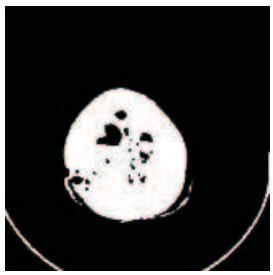
Cantidad de imágenes	Cantidad de clusters	Demora (segundos)	Inicialización de los centroides	Medida de distancia
1	2	2.45	Por frecuencia	Mahalanobis

Figura 2. Experimento 2 del algoritmo K-means

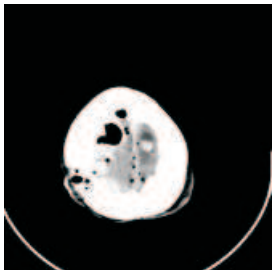
Cantidad de imágenes	Cantidad de clusters	Demora (segundos)	Inicialización de los centroides	Medida de distancia
1	2	1.819	Por frecuencia	Manhattan

Figura 3. Experimento 3 del algoritmo K-means

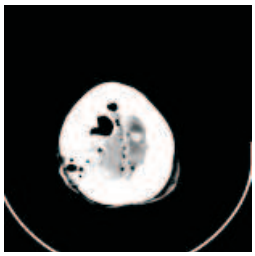
Cantidad de imágenes	Cantidad de clusters	Demora (segundos)	Inicialización de los centroides	Medida de distancia
1	2	1.605	Dispersos	Color

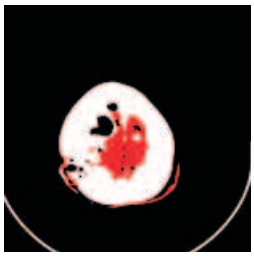
Figura 4. Experimento 4 del algoritmo K-means




Cantidad de imágenes	Cantidad de clusters	Demora (segundos)	Inicialización de los centroides	Medida de distancia
1	2	1.61	Aleatorio	Color

Figura 5. Experimento 5 del algoritmo K means

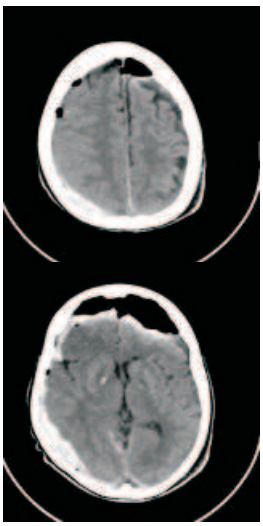


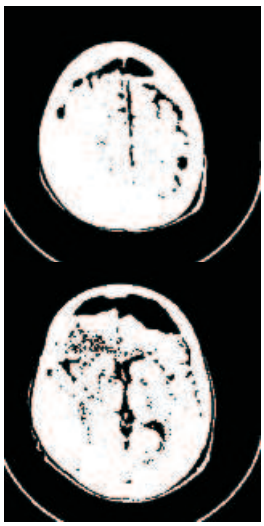


Cantidad de imágenes	Cantidad de clusters	Demora (segundos)	Inicialización de los centroides	Medida de distancia
1	3	2.391	Por frecuencia	Color

Figura 6. Experimento 6 del algoritmo K- means

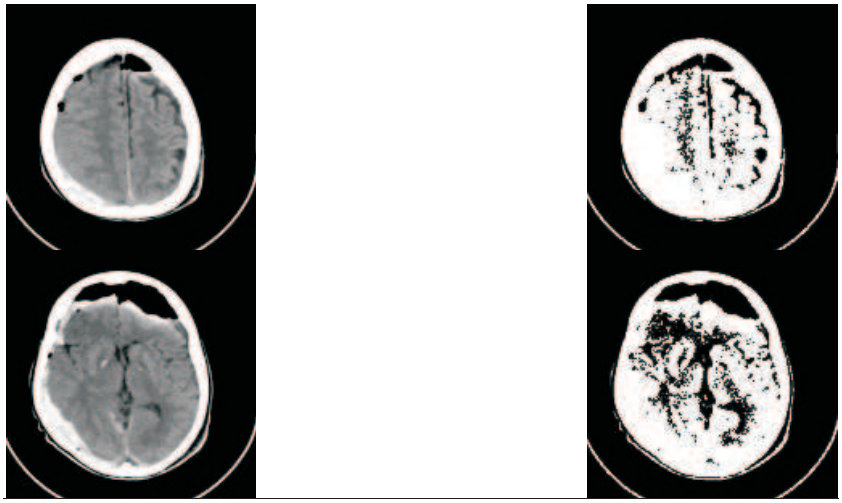
Para el algoritmo K-means la medida de distancia que mostró los mejores resultados en cuanto a tiempo y calidad es la diferencia de color. El procedimiento de inicialización de los centroides con el cual se hizo notable la reducción del tiempo de ejecución del algoritmo fue la aproximación mediante la frecuencia de color. Las imágenes DICOM utilizadas en los ejemplos son de 512x512 píxeles, lo que representa un total de 262144 píxeles, para los cuales las demoras en la obtención de los resultados de una sola imagen nunca excedió los tres segundos. Para 24 imágenes de 512x512 la demora estuvo sobre los 20 segundos, pero cabe destacar que en casos como estos se habla de 6029312 píxeles como entrada del algoritmo. Los experimentos aquí mostrados son solo algunos de los muchos que se realizaron. El algoritmo Fuzzy K-means aplicado a la segmentación de imágenes médicas digitales presentó las mismas características que el K-means en cuanto a inicialización de los centroides y medida de distancia se refiere, por tanto los experimentos que se mostrarán estarán basados por motivos de simplificación en inicialización de los centroides por el método de mayor frecuencia de color y la medida de distancia será la diferencia de color. La elección del exponente difuso se realizó en el intervalo [1.1, 2] y los resultados de los valores próximos a 1.1 superan los resultados de los valores próximos al otro extremo del intervalo, sin embargo, la duración real del algoritmo en la mayoría de los casos disminuyó para los exponentes difusos cercanos a dos. Las siguientes dos tablas ejemplifican lo antes expuesto:





Cantidad de imágenes	Cantidad de clusters	Demora (segundos)	Inicialización de los centroides	Medida de distancia	Exponente difuso	Error máximo
32	2	283.108	Por frecuencia	Color	1.1	0.001

Figura 7. Experimento 1 del algoritmo Fuzzy K -means



Cantidad de imágenes	Cantidad de clusters	Demora (segundos)	Inicialización de los centroides	Medida de distancia	Exponente difuso	Error máximo
32	2	181.66	Por frecuencia	Color	2.0	0.001

Figura 8. Experimento 2 del algoritmo Fuzzy K -means

El error máximo permitido para la mayoría de los experimentos fue fijado a 0.001 con el objetivo de tener una medida confiable de evaluación y ajustarse a las recomendaciones de la documentación consultada, no obstante valores menos extremistas como 0.005 garantizan resultados similares. Las próximas tres imágenes exhiben una comparación entre los resultados de los algoritmos K-means y Fuzzy K-means utilizando las mismas 32 imágenes de entrada, con inicialización de los centroides por máxima frecuencia de color y como medida de disimilitud la diferencia de color. Además el exponente difuso y el error máximo permitido para el Fuzzy K-means fueron de 1.1 y 0.001 respectivamente. El algoritmo Fuzzy K-means consumió 283.108 segundos en ejecución mientras que el K-means solo consumió 97.591 segundos.

Imagen original	Resultado del algoritmo K-means	Resultado del algoritmo Fuzzy K-means

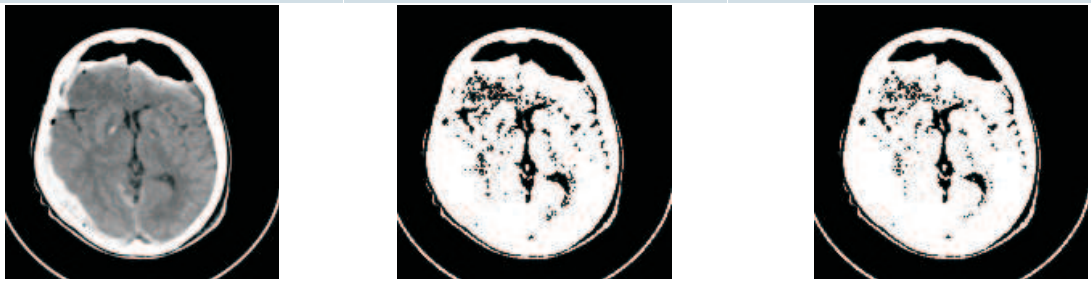


Figura 9. Comparación K -means - Fuzzy K -means

DISCUSIÓN

Clustering puede ser considerado el más importante problema de aprendizaje no supervisado. Un cluster es una colección de objetos que son similares entre sí según un determinado criterio de similitud y distintos a los objetos que pertenecen a otros clusters. Los algoritmos basados en el agrupamiento particional tienen como objetivo minimizar la varianza intracluster y maximizar la varianza intercluster. Estos métodos descomponen el conjunto de objetos en un conjunto de clusters disjuntos, minimizando una función criterio que enfatiza la estructura local de los objetos, asignando clusters a máximos locales en la estructura global. (Miguel Jiménez, 2008)

La determinación de la inicialización de los centroides juega un papel crucial debido a que cuanto mejor sea la partición inicial más rápido convergerá el algoritmo. Aunque se asegura la convergencia, ésta no tiene por qué ser un mínimo global. Una forma de realizar esta elección es calcular el histograma n-dimensional de la imagen y determinar los picos dominantes del mismo. Los *K* picos dominantes se hacen corresponder con los *K* centroides inicializados. Las agrupaciones al comenzar se pueden seleccionar aleatoriamente, con lo que disminuye la complejidad pero aumenta el tiempo de convergencia del algoritmo (Acha Piñero, 2002). Una opción recomendable y que suele ofrecer buenos resultados es la de realizar un análisis cluster jerárquico y elegir como partición inicial la obtenida con un nivel de similitud que aplicado al árbol ultra-métrico conduzca al número de grupos deseado. (Oliva i Cuyás, de Cáceres Ainsa, Font Castell, & Cuadras Avellana, 2001)
Así, una de las desventajas de los algoritmo K-means y Fuzzy K- means es que el cluster resultante es sensible a la elección inicial de los centroides y puede converger en un mínimo local. Por ambos métodos se realiza una búsqueda local en la vecindad de la solución inicial y va refinando la partición resultante, por esta razón se puede utilizar algún algoritmo de búsqueda global para inicializar los centroides. Los resultados de Fuzzy K-means fueron superiores para valores del exponente difuso próximos a 1.1 en cuanto a la calidad misma de la segmentación avalada por expertos. Sin embargo, la duración real del algoritmo en la mayoría de los casos disminuyó para los exponentes difusos cercanos a dos.

El algoritmo de Cúmulo de Partículas (PSO Particle Swarm Optimization) es una técnica de optimización estocástica que puede utilizarse para encontrar una solución óptima o cercana, ha sido aplicado a Clustering de datos y de textos con muy buenos resultados. (Villagra, Guzmán, Pandolfi, & Leguizamón, 2008)

CONCLUSIONES

El análisis minucioso de los algoritmos de Clustering más significativos y la experimentación fueron los precedentes para la construcción de un módulo de segmentación de imágenes médicas digitales que dio cumplimiento a los objetivos planteados, para ello se determinaron los siguientes aspectos:

1. Las técnicas de Clustering presentan ventajas con respecto a las otras técnicas de segmentación de imágenes.

2. La diferencia de color es el indicador de disimilitud más adecuado para la ejecución de los algoritmos tratados.

3. La heurística de análisis de frecuencia de color para la inicialización de los centroides de los algoritmos vistos es la que proporciona los resultados más acertados.

4. Para el perfeccionamiento ulterior de las técnicas de segmentación de imágenes médicas mediante técnicas de Clustering se ha detectado las siguientes vías:
   - Búsqueda de procedimientos más avanzados de inicialización de los centroides.
   - Utilización de algoritmos que permitan alcanzar el óptimo global o al menos aproximarse al mismo, tanto para el método K-means como para el Fuzzy K-means.

5. El módulo de segmentación de imágenes médicas digitales presenta un diseño extensible y refinado que no lo mantiene atado a alguna biblioteca, fue desarrollado sobre estándares y brinda la posibilidad de ser adaptado fácilmente a diferentes sistemas. Sobre los sistemas operativos Ubuntu 9.10 y Windows XP se realizaron las pruebas que alegan alta fiabilidad y robustez.

REFERENCIAS BIBLIOGRÁFICAS

- Acha Piñero, B. (Abril de 2002). Segmentación y clasificación de imágenes en color. Aplicación al diagnóstico de quemaduras. Tesis Doctoral. Sevilla, Sevilla, España: Universidad de Sevilla.
- del Fresno, M., & J Vénere, M. (2002). Segmentación de Imágenes Médicas por Crecimiento de Regiones con Conocimiento Adicional. Tandil, Buenos Aires, Argentina: PLADEMA-ISISTAN, Universidad Nacional del Centro.
- Jain, A. J., Murty, M. N., & Flynn, P. J. (Septiembre de 1999). Data Clustering: A Review. ACM Computing Surveys, Vol. 31, No. 3.
- Miguel Jiménez, J. M. (Marzo de 2008). Introducción al Tratamiento Digital y Clustering de Imágenes. Alcalá, Madrid,

España: Departamento de Electrónica, Universidad de Alcalá.

- Oliva i Cuyás, F., de Cáceres Ainsa, M., Font Castell, X., & Cuadras Avellana, C. M. (Noviembre de 2001). Contribuciones desde una perspectiva basada en proximidades al Fuzzy K-means Clustering. Jaén, Andalucía, España: XXVI Congreso Nacional de Estadística e Investigación Operativa: Úbedabeda, Universidad de La Rioja, ISBN 84-8439-080-2.
- Ortega, J. A., Foster, W., & Ortega, R. (2002). Definición de Sub-Rodales para una Silvicultura de Precisión: Una aplicación del método Fuzzy K-means. Santiago, Chile: Facultad de Agronomía e Ingeniería Forestal, Pontificia Universidad Católica de Chile, Casilla 306-22.
- Tan, P. N., Steinbach, M., & Kumar, V. (Marzo de 2006). *Introduction to Data Mining*. Recuperado el 10 de Febrero de 2010, de Addison-Wesley Companion Book Site: http://www-users.cs.umn.edu/~kumar/dmbook/index.php
- Villagra, A., Guzmán, A., Pandolfi, D., & Leguizamón, G. (2008). Análisis de medidas no-supervisadas de calidad en clusters obtenidos por K-means y Particle Swarm Optimization. Argentina: Universidad Nacional de la Patagonia Austral, Unidad Académica Caleta Olivia, Universidad Nacional de San Luis, Laboratorio de Tecnologías Emergentes, Laboratorio de Investigación y Desarrollo en Inteligencia Computacional.

Correspondencia:
Gustavo Lorca
Carretera de San Antonio de los Baños Km. 2 ½
La Habana - Cuba
9toronzo@estudiantes.uci.cu