2do Taller LAtino Iberoamericano de Investigación de Operaciones
"La IO aplicada a la solución de problemas regionales"

Mejora al algoritmo de agrupamiento *K-means* mediante un nuevo criterio de convergencia y su aplicación a bases de datos poblacionales de cáncer

J. Pérez[1], M. F. Henriques[2], R. Pazos[1], L. Cruz[3], G. Reyes[1], J. Salinas[1], A. Mexicano[1]

[1] Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), México,
[2]Secretaría de Saúde do Estado de Pernambuco, Brasil,
[3] Instituto Tecnológico de Ciudad Madero, México.

Resumen: En este trabajo se propone una mejora al algoritmo heurístico de agrupamiento (clustering) *K-means* y se muestran los resultados de su aplicación a bases de datos poblacionales reales de cáncer de México. Los problemas de agrupamiento surgen en varios tipos de aplicaciones: minería de datos, aprendizaje de máquina, descubrimiento de conocimiento, compresión de datos y reconocimiento de patrones, entre otros. Uno de los métodos más usados y estudiados es *K-means*, al cual se le han realizado varias mejoras, la mayoría de éstas relacionadas con la definición de los parámetros iniciales. En contraste, en este trabajo se propone una mejora usando un nuevo criterio de convergencia que consiste en parar la ejecución del algoritmo cuando se encuentra un óptimo local, o bien cuando ya no se dan intercambios de objetos entre los grupos. Dicha mejora surge del análisis experimental del algoritmo clásico, el cual en algunas corridas y ejemplares de problemas pasaba de largo en óptimos locales y convergía en soluciones peores. Experimentalmente, el algoritmo mejorado mostró importantes reducciones en el número de iteraciones y en la calidad de la solución. Debido a los resultados alentadores, se decidió usarlo en una aplicación real en el área de salud pública de México, en particular en el problema de encontrar patrones de agrupamiento de municipios con afinidad en los parámetros de localización y tasas de mortalidad por cáncer de pulmón y estómago. Como resultado de la aplicación del algoritmo mejorado, se encontraron, para el cáncer de pulmón, grupos de municipios con altas tasas de mortalidad en la región norte y noroeste. Para el cáncer de estómago se identificó un grupo muy conocido de alta mortalidad en la región sureste (altos de Chiapas), el cual sirvió para validar el enfoque de solución. Por otra parte, se identificó otro grupo con una mayor tasa de mortalidad en la región noroeste. Finalmente, por una parte, consideramos que la mejora al algoritmo puede ser de utilidad en muchos tipos de aplicaciones, y por otra, que los resultados obtenidos en la mencionada aplicación pueden servir como una herramienta de apoyo para investigaciones sobre el cáncer y para la toma de decisiones en cuanto a la asignación de recursos para su prevención y tratamiento.

Palabras clave: optimización, algoritmos heurísticos, agrupamiento, aplicaciones en salud, inteligencia artificial.

Introducción

El problema de agrupación de objetos de acuerdo a sus atributos ha sido ampliamente estudiado debido a sus aplicaciones en áreas como aprendizaje de máquina [4], minería de datos y descubrimiento de conocimiento [3, 11], reconocimiento y clasificación de patrones [2]. El objetivo del agrupamiento es particionar un grupo de objetos, los cuales tienen asociados vectores multidimensionales de atributos en grupos homogéneos, tales que los patrones en cada grupo sean similares.

Varios algoritmos de aprendizaje no supervisado han sido propuestos, los cuales particionan el conjunto de objetos en un determinado número de grupos de acuerdo a un criterio de optimización. Uno de los métodos de agrupación más popular y ampliamente utilizado es K-means [10]; particularmente, porque su implementación es relativamente simple. Sin embargo, el algoritmo K-means tiene los siguientes inconvenientes:

- El agrupamiento final depende de los centroides iniciales.
- La convergencia en el óptimo global no está garantizada, y para problemas con muchos ejemplares, requiere de un gran número de iteraciones para converger.

Un factor que afecta en gran medida el costo computacional del algoritmo K-means es el número de

iteraciones que necesita realizar, ya que por cada iteración calcula la distancia de cada objeto a los centroides de los grupos. En este trabajo se propone una nueva condición de convergencia, la cual permite en muchos casos reducir el número de iteraciones y mejorar la calidad del agrupamiento. La mejora propuesta fue obtenida mediante una perspectiva de optimización; es decir, mediante el estudio del comportamiento de la función objetivo, que en este caso es el error al cuadrado.

En la siguiente sección se describe el algoritmo K-means tradicional.

Descripción del algoritmo K-means estándar

De acuerdo a la literatura especializada [1, 6, 7, 8, 14, 15] se pueden identificar cuatro pasos en el algoritmo:

- **Paso 1. Inicialización:** Se definen un conjunto de objetos a particionar, el número de grupos y un centroide por cada grupo. Algunas implementaciones del algoritmo estándar determinan los centroides iniciales de forma aleatoria; mientras que algunos otros procesan los datos y determinan los centroides mediante de cálculos.
- **Paso 2. Clasificación:** Para cada objeto de la base de datos, se calcula su distancia a cada centroide, se determina el centroide más cercano, y el objeto es incorporado al grupo relacionado con ese centroide.
- **Paso 3. Cálculo de centroides:** Para cada grupo generado en el paso anterior se vuelve a calcular su centroide.
- **Paso 4. Condición de convergencia:** Se han usado varias condiciones de convergencia, de las cuales las más utilizadas son las siguientes: converger cuando alcanza un número de iteraciones dado, converger cuando no existe un intercambio de objetos entre los grupos, o converger cuando la diferencia entre los centroides de dos iteraciones consecutivas es más pequeño que un umbral dado. Si la condición de convergencia no se satisface, se repiten los pasos dos, tres y cuatro del algoritmo.

Trabajos Relacionados

Se han realizado varias mejoras al algoritmo K-means estándar relacionadas con varios aspectos asociados a cada uno de los pasos del algoritmo. Básicamente el algoritmo consta de cuatro pasos:

1. Inicialización.
2. Clasificación.
3. Cálculo de centroides.
4. Condición de convergencia.

Con respecto a la mejora de los cuatro pasos, probablemente la que ha recibido mayor atención es la inicialización; vale la pena mencionar los trabajos de Meila [12] y Peña [15]. En lo que concierne al segundo paso, se han definido varias medidas de afinidad para los elementos de los grupos; en este caso las contribuciones en [6, 17] son destacables. Con respecto a los pasos tres y cuatro y de acuerdo a la literatura especializada, no existen trabajos reportados sobre la mejora del algoritmo. En contraste, este artículo se enfoca en el cuarto paso, proponiendo una nueva condición de convergencia.

Por otra parte, otros autores han hecho algunas contribuciones importantes en relación a la implementación computacional del algoritmo; v.gr., se han propuesto estructuras de datos especiales y optimización de código para los cálculos del algoritmo; con respecto a esto, los siguientes trabajos pueden ser mencionados [7, 9, 14].

Hasta este punto es conveniente destacar que, ya que este enfoque corresponde al cuarto paso, no se intenta competir con otros enfoques, sino que podría ser usado para complementarlos.

Análisis de resultados experimentales del algoritmo

El algoritmo K-means estándar fue implementado, incluyendo varias líneas de código para mostrar en cada iteración el error cuadrático para cada grupo y la suma de errores de los grupos. Para propósitos de prueba, se utilizó la base de datos *Diabetes* del repositorio de aprendizaje de máquina de la Universidad de California en Irvine (UCI) [13]. Al observar los errores cuadráticos de las iteraciones, se encontró que en muchos casos el algoritmo pasaba un óptimo local y continuaba realizando iteraciones que incrementaban el error. La Figura 1 ilustra el resultado de una corrida usando la base de datos *Diabetes*. Como se puede ver en la figura, en la iteración 2 se encuentra un óptimo local del error cuadrático, el cual difiere del error cuadrático final.

Por otra parte, en varios experimentos el algoritmo convergió en el óptimo local como se muestra en la Figura 2. En estos casos el error cuadrático siguió un patrón de decremento continuo.

Con la finalidad de identificar la tasa de ocurrencia de este comportamiento, el algoritmo fue ejecutado 30 veces con la misma base de datos (pero con diferentes centroides iniciales), y se encontró que en un 73% de las corridas el algoritmo pasó del óptimo local y en un 27% de las corridas convergió en el mínimo local. Alentados por este resultado, se realizó un nuevo conjunto de experimentos con otras cinco bases de datos reales extraídas de UCI [13], las cuales se describen brevemente a continuación:

- **Vehicle silhouette.** Su propósito es clasificar una silueta dada en uno de cuatro tipos de vehículos, usando un conjunto de características extraídas de la silueta. El vehículo puede ser visto desde diferentes ángulos. La base de datos *Vehicle* contiene 846 ejemplares, 4 grupos y 18 atributos.
- **Glass identification.** Se usa para clasificar tipos de cristales y fue motivada por investigación criminológica. La base de datos *Glass* contiene 214 ejemplares, 7 grupos y 9 atributos.
- **Diabetes.** La base de datos *Diabetes* contiene el diagnóstico de pacientes de diabetes de acuerdo a la Organización Mundial de la Salud. La base de datos contiene 768 ejemplares, 2 grupos y 8 atributos.
- **Heart disease**. Esta base de datos almacena diagnósticos de enfermedades del corazón. La base de datos *Heart* contiene 270 ejemplares, 2 grupos y 13 atributos.
- **Wine recognition.** Estos datos consisten en los resultados de un análisis químico de vinos producidos un una región de Italia y derivados de tres diferentes variedades de vinos. La base de datos *Wine* contiene 178 ejemplares, 3 grupos y 13 atributos.
- **Liver disorders.** Estos datos son el resultado de una investigación médica de pacientes que pueden ser sensibles a trastornos de hígado, lo cual puede ser originado por abuso de alcohol. La base de datos *Liver* contiene 345 ejemplares, 2 grupos y 6 atributos.

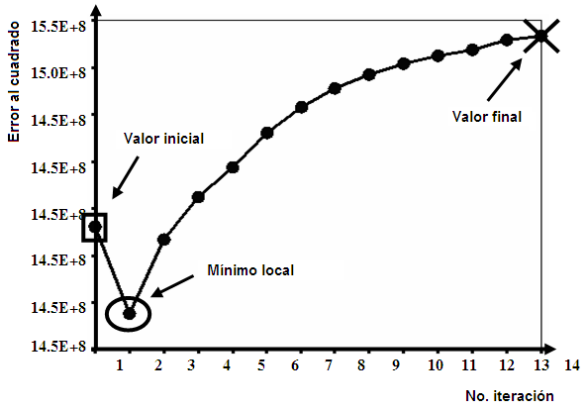

Figura 1: Convergencia pasando del óptimo local

Tabla 1: Información de las bases de datos usadas en los experimentos

Base de Datos	Número de grupos	Número de atributos	Número de ejemplares
Vehicle	4	18	846
Glass	7	9	214
Diabetes	2	8	768
Heart	2	13	270
Wine	3	13	178
Liver	2	6	345

La Tabla 2 muestra el resultado de los experimentos, en los cuales el algoritmo fue ejecutado 30 veces para cada base de datos.

Como se puede observar en la Tabla 2, el mayor porcentaje de ocurrencia del comportamiento en el que el algoritmo rebasó el óptimo local fue del 73% con la base de datos *Diabetes*; mientras que el porcentaje menor fue de 20% con la base de datos *Glass*. Este comportamiento sugiere la conveniencia de definir una nueva condición de convergencia para mejorar el desempeño del algoritmo en los casos donde el algoritmo pase más allá del mínimo.

Dicha mejora puede ser benéfica de dos maneras:

- Reducción en el número de iteraciones (incremento de la eficiencia).
- Mejora en la calidad de la solución (incremento de la eficacia).

Las siguientes secciones presentan una breve descripción del algoritmo modificado, que incluye la nueva condición de convergencia.

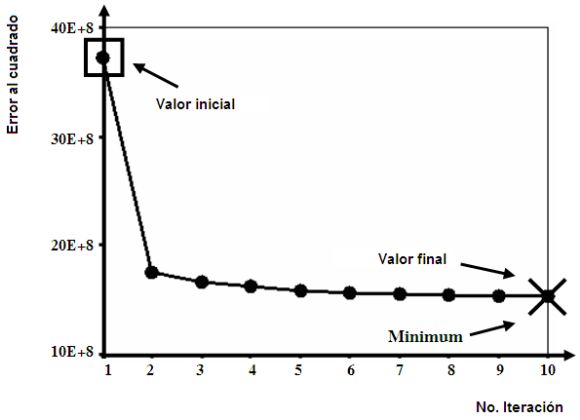

Figura 2 Convergencia al mínimo local

Tabla 2: Porcentaje de corridas que pasaron del mínimo local

Base de Datos	Porcentaje de Corridas
Diabetes	73
Liver	70
Vehicle	50
Wine	43
Heart	33
Glass	20

Mejora propuesta para el Algoritmo K-means Estándar

Tradicionalmente, el algoritmo K-means ha sido considerado como un algoritmo voraz; es decir, puede quedar atrapado en un mínimo; sin embargo, analizando las condiciones de convergencia y los experimentos realizados, se encontró que no existe una condición de convergencia que involucre el error cuadrático, y por tanto, el algoritmo estándar no puede garantizar la convergencia en el óptimo local, como lo revelan los experimentos efectuados (Figura 1). En este sentido la principal contribución de este trabajo consistió en asociar los valores del error cuadrático a una condición de convergencia, por tanto el algoritmo se detiene cuando se encuentra un óptimo local. La nueva condición de convergencia ocurre cuando en dos iteraciones consecutivas el error cuadrático de la última iteración excede al de la iteración precedente.

Nótese que la condición de convergencia propuesta fue concebida para encontrar una solución al menos tan buena como la del algoritmo K-means estándar con un número de iteraciones menor o igual.

Aplicación a Bases de Datos de Cáncer

En este trabajo se usaron varias bases de datos oficiales. Los datos de mortalidad fueron obtenidos del "Núcleo de acopio y análisis de información de salud" (NAAIS) del Instituto Nacional de Salud Pública (INSP), en particular se seleccionaron los datos de muerte por cáncer de pulmón y estomago del año 2000. Los datos fueron preprocesados y se obtuvieron para cada uno de los municipios de México el número de muertes por cáncer así como la tasa de muertes por 100,000 habitantes, ésta información mas la localización geográfica de los municipios se integró en un almacén de datos o *datawarehouse.* Se implementó un prototipo de sistema de minería de datos que constaba de dos módulos uno para la generación de patrones y otro de visualización de resultados. En el módulo de generación de patrones se codificó la mejora del algoritmo *k-means* propuesta en este trabajo. Se realizaron un conjunto de experimentos usando el prototipo sobre el almacén de datos de cáncer, seleccionando municipios con poblaciones mayores que 100,000 habitantes para el año 2000, y estableciendo el número de grupos *k* igual a 5, 10, 15, 20, y 30. El mejor resultado se obtuvo para *k* igual a 20 de acuerdo con los especialistas.

De los 20 grupos generados, los dos grupos con las tasas promedio más altas fueron seleccionados, ya que generan mayor interés. La Figura 3 muestra un grupo que corresponde a los altos de Chiapas en el sureste de México. Este grupo sirvió para validar el método de minería de datos utilizado en este prototipo, puesto que investigaciones clínicas han reportado una alta tasa de mortalidad por cáncer gástrico en esa región. Tales investigaciones han afirmado que uno de los factores que contribuye al desarrollo de este tipo de cáncer en la región es una infección crónica causada por una bacteria llamada helicobacter pylori (HP) [18]. Los detalles de los municipios del grupo y las tasas de mortalidad se muestran en la Tabla 3, incluyendo los valores de la media y la desviación estándar.

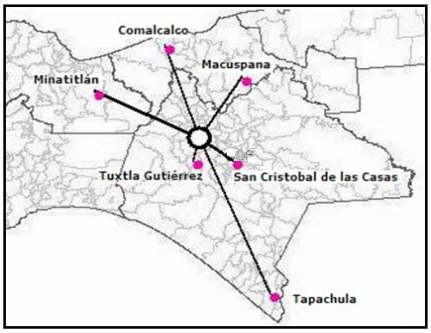

Figura 3: Grupo de los altos de Chiapas

Tabla 3: Valores del Grupo de los altos de Chiapas

Municipio	Muertes	Población	Tasa
Minatitlán	14	153001	9.15
Comalcalco	14	164637	8.50
Tapachula	21	271674	7.73
San Cristóbal	9	132421	6.80
Macuspana	9	133985	6.72
Tuxtla Gutiérrez	28	434143	6.45
Promedio de Tasa			7.56
Desviación estándar			0.99

Adicionalmente, se encontró otro patrón de interés y de gran utilidad en la región noroeste (Figura 4, Tabla 4), el cual tiene una tasa de mortalidad más alta que la del grupo 1. De acuerdo con la literatura especializada, no existen estudios que reporten una concentración de tasas de mortalidad altas para cáncer de estómago en esa región. Una posible explicación a esta situación es que las estadísticas de cáncer usualmente se efectúan por estado, y el grupo noroeste abarca dos estados.

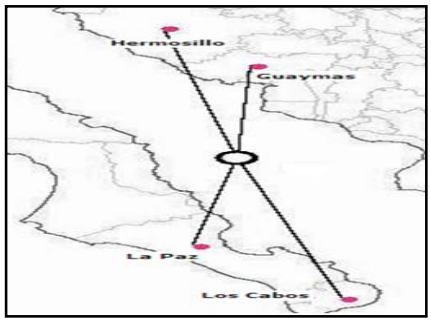

Figura 4: Grupo del noroeste

Tabla 4: Valores para grupo noroeste

Municipio	Muertes	Población	Tasa
Guaymas	15	130329	11.52
Hermosillo	48	609829	7.87
La Paz	14	196907	7.11
Los Cabos	7	105469	6.64
Promedio de Tasa			8.28
Desviación estándar			1.92

Para la base de datos de cáncer de pulmón, se encontró un patrón de utilidad potencial en la región noroeste con altas tasa de mortalidad, el cual abarca 13 municipios de la región, como se puede observar en la Figura 5 y cuyos valores se muestran en la Tabla 5.

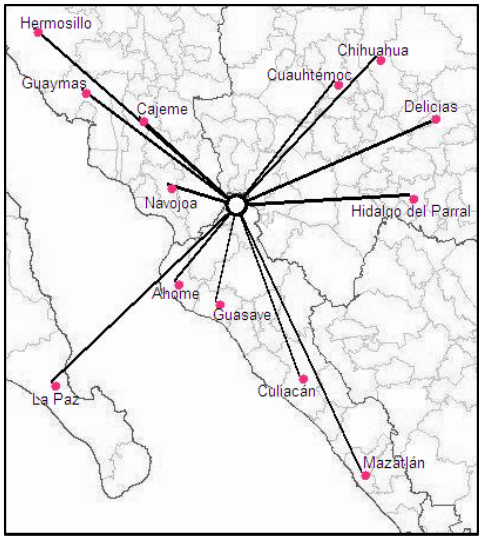

Figura 5: Grupo noroeste de cáncer de pulmón

Tabla 5: Valores para el grupo noroeste de cáncer de pulmón

Nombre	Muertes	Población	Tasa
Cajeme	67	356290	18.80
Hermosillo	104	609829	17.05
Hidalgo del Parral	16	100821	15.86
Culiacán	113	745537	15.15
Navojoa	21	140650	14.93
Ahome	52	359146	14.47
Guasave	39	277402	14.05
Delicias	16	116426	13.74
La Paz	27	196907	13.71
Mazatlán	51	380509	13.40
Guaymas	17	130329	13.04
Cuauhtémoc	14	124378	11.25
Chihuahua	75	671790	11.16
Promedio de Tasa			14.3546154
Desviación estándar			2.03128461

Otro patrón de interés y con alta tasa de mortalidad se detectó en la región norte (Figura 6, Tabla 6).

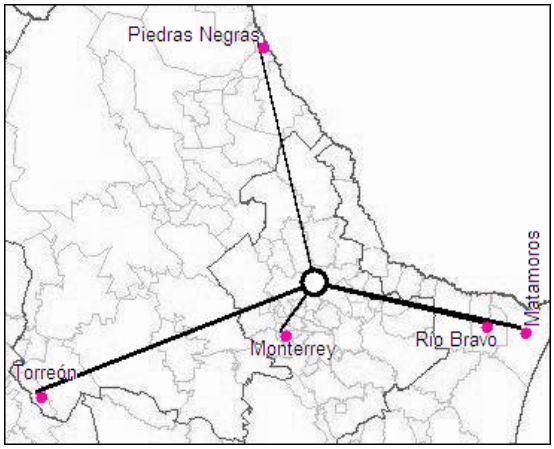

Figura 6: Grupo norte de cáncer de pulmón

Tabla 6: Valores para el grupo cáncer de pulmón

Nombre	Muertes	Población	Tasa
Río Bravo	14	104229	13.43
Matamoros	54	4148141	12.91
Torreón	65	529512	12.27
Monterrey	113	1110997	11.97
Piedras Negras	15	128130	11.70
Promedio de Tasa			12.456
Desviación estándar			0.7065

Conclusiones

Varias mejoras del algoritmo K-means se han enfocado en los parámetros de inicialización del algoritmo [12]; sin embargo, uno de los factores que más afecta su costo computacional es el número de iteraciones realizadas hasta converger.

Este trabajo muestra que es posible mejorar el algoritmo K-means estándar mediante una nueva condición de convergencia. Durante el estudio de la implementación del algoritmo K-means estándar; específicamente, analizando la evolución de la suma de los errores cuadráticos para cada iteración, se encontró que en muchos casos el algoritmo no paraba en un óptimo local. A pesar de haber alcanzado un óptimo local, el algoritmo continúo ejecutándose y generó una solución final peor que el óptimo local.

Un análisis detallado del algoritmo estándar reveló que pasó del óptimo local porque la condición de convergencia no toma en cuenta el error cuadrático. En contraste, este trabajo propone una nueva condición de convergencia que incorpora el error cuadrático, la cual garantiza que el algoritmo parará en un óptimo local, reduciendo el número de iteraciones y mejorando la calidad de la solución, en numerosos casos.

Es importante destacar que la técnica de mejora propuesta no es incompatible con otras técnicas para mejorar el algoritmo K-means, ya que éstas son aplicables a los pasos de inicialización y clasificación del algoritmo. Por lo tanto, la técnica propuesta puede ser combinada con algunas otras variantes del algoritmo K-means, contribuyendo así a mejoras adicionales en su desempeño.

Finalmente, un trabajo futuro se enfocará en analizar entre parar en el primer mínimo local y la realización de más iteraciones hasta encontrar otro mínimo local o global. Experimentos preliminares del algoritmo K-means con 30 corridas con un *datawarehouse* de base poblacional de mortalidad por cáncer en México, muestra que en 24 casos la función objetivo convergió en un valor final sin pasar por un mínimo, y en 6 casos la función objetivo convergió pasando uno o más mínimos. De los 6 casos anteriores, en 3 casos el valor final fue mejor que el del primer mínimo intermedio, y la diferencia en valores no fue mayor que 0.6%.

Referencias

[1] Bottou, L., Bengio, Y.: Convergence Properties of the K-means Algorithms. Advances in Neural Information Processing. Systems. MTI Press, 1995.

[2] Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. John Wiley & Sons, New York, NY. 1973.

[3] Fayyad,U. M., Piatetsky-Shapiro, G., Smyth P., Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[4] Fisher, D.: Knowlwdge Acquisition via Incremental Conceptual Clustering. Machine Learning, Vol. 2, No. 2 (1987) 139-172.

[5] Garner, S.: Weka: The Waikato Environment for Knowledge Analysis. Proc. New Zealand Computer Science Research Students Conference (1995) 57-64.

[6] Hamerly, G., Elkan, C.: Alternatives to the K-means Algorithm that Find Better Clusterings. Proc. 11th International Conf. on Information and Knowledge Management CIKM'02. ACM. Virginia, USA (2002).

[7] Kanungo, T., Netanyahu, N.S., Wu, A.Y.: An Efficient K-means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7 (2002).

[8] Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: A Local Search Approximation Algorithm for *k*-Means Clustering. Proc. 18th Annual ACM Symposium on Computational Geometry (SoCG'02). Barcelona, Spain (2002) 10-18.

[9] Likas, A., Vlassis, N., Verbeek, J.J.: The Global K-means Clustering Algorithm. Pattern Recognition, 451-461. 2003.

[10] MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. Proc. 5th Berkeley Symp. Math. Statistics and Probability, 1:281-297, 1967.

[11] Mehmed, K.: Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. 2003.

[12] Meila, M., Heckerman, D.: An Experimental Comparison of Several Clustering and Initialization Methods. Proc. 14th Conf. on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers, Inc., San Francisco, CA.

[13] Merz, C., Murphy, P., Aha, D.: UCI Repository of Machine Learning Databases. Department of Information and Computer Science, University of California. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[14] Pelleg, D., Moore, A.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters. Proc. 17th International Conf. on Machine Learning (2000).

[15] Peña, J.M., Lozano, J.A., Larrañaga, P.: An Empirical Comparison of Four Initialization Methods for the K-Means Algorithm. Dept. of Computer Science and Artificial Intelligence, University of the Basque Country, San Sebastian, España.

[16] SPSS, Inc. Headquarters, Chicago, Illinois. http://www.spss.com/es/

[17] Su, M.C., Chou, C.H.: A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 6 (2001) 674-680.

[18] Mohar, A., Ley, C., Guarner, J., Herrera-Goepfert, R., Sánchez L., Halperin D., Parsonnet J.: Alta Frecuencia de Lesiones Precursoras de Cáncer Gástrico Asociadas a Helicobacter Pylori y Respuesta al Tratamiento, en Chiapas, México. Gaceta Médica de México, Vol. 138, No.5 (2000) 405-410

Nombre Completo del Autor

Joaquín Pérez O., Rodolfo Pazos R., Gerardo Reyes S., Jesús Salinas C., Adriana Mexicano S.

Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), México, {jperez, pazos, greyes, jsalinas-05c, iscmexs05c} @cenidet.edu.mx.

M. Fátima C. Henriques, Secretaría de Saúde do Estado de Pernambuco, Brasil, fhenriques@saude.pe.gov.br.

Laura Cruz R., Instituto Tecnológico de Ciudad Madero, México, lcruz @prodigy.net.mx.