Daena: International Journal of Good Conscience. 14(1)350-363. Junio 2019.
ISSN 1870-557X

Revisión integrada de agrupamiento de series temporales basado en densidad

Integrated review of density-based time series clustering

M.C. Kathia Gabriela Flores Rodríguez, Dra. Sara E. Garza, Dra. Leticia Neira-Tovar

Universidad Autónoma de Nuevo León, Facultad de Ingeniería Mecánica y Eléctrica

San Nicolás de los Garza, N.L., México. Email: kathia.floresrd@uanl.edu.mx

Palabras Clave: data clustering, data mining, data warehouse, time series.

Resumen

Los avances tecnológicos en el manejo de datos han generado nuevas técnicas para ser utilizadas en minería de datos. El propósito de este trabajo es presentar el estado del arte del agrupamiento de datos aplicado en series temporales, la cual es una de las tecnicas utilizadas en minería de datos, en donde datos son agrupados con otros de características similares, sin tener un conocimiento previo de las características de dichos grupos. El objetivo de esta investigación es proponer una variante del algoritmo de agrupamiento de datos basado en picos de densidad (*density peaks*) desarrollado por Rodríguez y Laio (2014), el cual sirva como base para el desarrollo de un estudio posterior.

Abstract

Technological advances in data management have generated new techniques to be used in data mining. The purpose of this paper is to review recent research on data clustering applied in time series, which is a method used in data mining, where the data are grouped with other data with similar characteristics without having a previous knowledge of the characteristics of these groups. The goal of this research is to explores the possibility of developing a variant of the data clustering algorithm based on density peaks developed by Rodríguez and Laio (2014), which will be applied in a later study case.

1. Introducción

El uso de la tecnología hoy en día permite que casi cada acción que realizamos genere datos digitales. Debido al uso cotidiano de sensores digitales, equipos de cómputo, dispositivos móviles, entre otros, los cuales permiten recolectar, almacenar y procesar información de manera automática y en tiempo real, es que existen una cantidad masiva de datos. Sin embargo, debido al volumen y a la velocidad con la que estos datos son adquiridos y almacenados, los métodos tradicionales de estadística pierden su efectividad al momento de procesarlos para generar información (Buyya y Vahid, 2016). Esto conlleva la necesidad de utilizar métodos que permitan analizar datos que están en constante flujo. Una de las técnicas utilizadas en la actualidad para el análisis de datos masivos es la minería de datos (del inglés *data mining*), la cual no solo consiste en extraer y procesar todo el conjunto de datos para su análisis, sino más bien tiene como objetivo el extraer patrones que puedan generar conocimiento relevante. Los métodos que conforman la minería de datos son el agrupamiento de datos (del inglés *data clustering*), clasificación, descubrimiento de patrones y análisis de valores atípicos. Mediante estas técnicas es posible analizar diferentes tipos de estructuras las cuales presentan dependencia entre sus datos, por ejemplo, las secuencias de valores recolectados continuamente por un sensor pueden estar relacionados entre sí, en donde el atributo tiempo permite analizar e interpretar la posible relación entre los datos (Aggarwal, 2015).

1.1 Teorías básicas y conceptualización de una serie temporal

Un dato puede referirse a cantidades numéricas o a los elementos de un proceso, los cuales son obtenidos de una población o muestra, pueden representar medidas o conteos. Los datos no son útiles o significativos hasta que no son procesados y convertidos en información, de tal manera que se otorga un valor real y perceptible para tomar decisiones y se presenta en forma de indicadores (Celis y Labrada, 2014) (ver Figura 1).

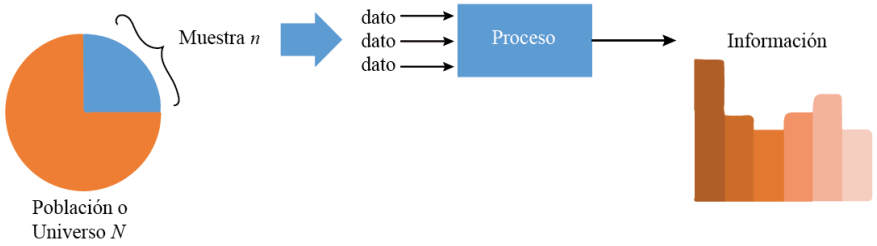


Figura 1: Generación de información a partir de los datos tomados de una muestra.

Minería de datos es la práctica de examinar bases de datos utilizando análisis matemático, estadístico y técnicas computacionales con el objetivo de generar nueva información, deducir patrones y tendencias que existen en los datos.

Las series temporales (del inglés *time series data*) son un tipo de estructura temporal utilizado en la minería de datos, en donde, los datos se presentan en secuencia y contienen el atributo de tiempo, es decir, representan un conjunto de datos los cuales varían con respecto al tiempo. En este tipo de conjunto de datos la extracción, almacenamiento y procesamiento se realizan de manera dinámica y sus valores cambian respecto al tiempo (Aggarwal, 2015; Falk y Frank, 2011). Para el análisis de los datos en series temporales se usan diversos métodos que permiten extraer información o patrones sobre la relación entre los datos la cual permite realizar pronósticos o extraer valores atípicos. Debido al incremento de las tecnologías de la información y la recolección de información en bases de datos en donde los registros presentan datos que varían con el tiempo, surge el interés de la investigación en este campo (Aggarwal, 2015; Fu, 2011; Zhang, Shen, Yao y Pei, 2016).

Para que un conjunto de datos pueda considerarse serie temporal debe contener al menos uno de los siguientes factores (Anderson y Senmelroth, 2015; Falk y Frank, 2011):

- **Tendencia:** dirección general en la que algo se desarrolla o cambia.
- **Temporal:** mostrar una relación o característica en un punto en particular.
- **Cíclico:** se presenta con recurrencia.
- **Atípico:** contener parámetros que son inusuales o anormales.

1.2 Minería de datos en series temporales

Las series temporales son estructuras de datos las cuales están conformados por grandes cantidades de datos y necesitan actualizarse constantemente, debido al tamaño de estas estructuras se les conoce como macrodatos (del inglés *Big Data*), es decir es un conjunto de datos en donde los métodos estadísticos e informáticos normales no son suficientes para procesar y extraer patrones (Zhang, Shen, Yao y Pei, 2016; Baoyan, 2014), para esto se utilizan diversas técnicas de minería de datos, las cuales combinan métodos matemáticos, computacionales y estadísticos para extraer información relevante previamente desconocida (Benmouiza y Cheknane, 2016).

Los datos en series temporales provienen de bases de datos en las cuales se almacena y maneja la información en forma de tablas; se les considera una colección de datos organizados, relacionados y estructurados; puede ser considerada como un depósito de datos que representan un modelo o proceso del mundo real (Jiménez, 2014), sin embargo, una base de datos es más que eso, es más bien una fuente de información y conocimiento, en donde un conjunto de datos genera información y con dicha información se puede obtener algún conocimiento (Collen, 2012). No obstante, para obtener información relevante es importante tener suficientes datos. Una de las problemáticas para poder aplicar técnicas de minería de datos es la poca cantidad de registros o datos que se tenga, y esto se debe a que los datos obtenidos de experimentos, registros web, etc., son de uso restringido o bien se mantienen en un ambiente local y no son reportados a la comunidad científica (Lemke, Budka y Gabrys, 2013). A su vez, una base de datos se puede integrar de información obtenida de diferentes fuentes, ya sea de repositorios web, datos de otra base de datos, reportes obtenidos de archivos físicos, resultados de cuestionarios, bases de datos no estructuradas, encuestas o registros en formato que no es electrónico, los cuales se integran en una sola estructura, la cual se conoce como almacén de datos (del inglés *data warehouse*). En la integración de datos la exactitud en los mismos es importante, ya que los datos incorrectos y corruptos generan un análisis impreciso. Entre las dificultades que esto representa está el extraer información de bases de datos normalizadas, inconsistencias en los datos, manejo de datos redundantes y pérdida de datos. Además, al ser obtenidos de diversas fuentes los datos se presentan de manera no estructurada y con formato diferente, por lo cual para que en una base de datos se pueda aplicar técnicas de minería de datos y algoritmos de agrupamiento es necesario pre procesar los datos no estructurados (Poenaru, Merezeanu, Dobrescu y Posdarascu, 2017). El preprocesamiento de los datos consiste en 1) extraer información de diversas fuentes, 2) transformar en datos coherentes, estructurados y des normalizados y 3) cargar en base de datos (almacén de datos). En la Figura 2 se muestra un diagrama de integración de los datos para obtener reportes estadísticos o análisis predictivo.

Los datos almacenados en una almacén de datos son temporales y no volátiles, es decir, proporcionan una perspectiva histórica, lo cual permite ver el comportamiento de dichos datos en el transcurso del tiempo, permitiendo así mediante algoritmos de agrupamiento de datos calcular tendencias, encontrar patrones y al ser variantes en el tiempo cuando la base de datos o el almacén

de datos tiene una actualización no se alteran o eliminan los datos existentes (Poenaru, Merezeanu, Dobrescu y Posdarascu, 2017).

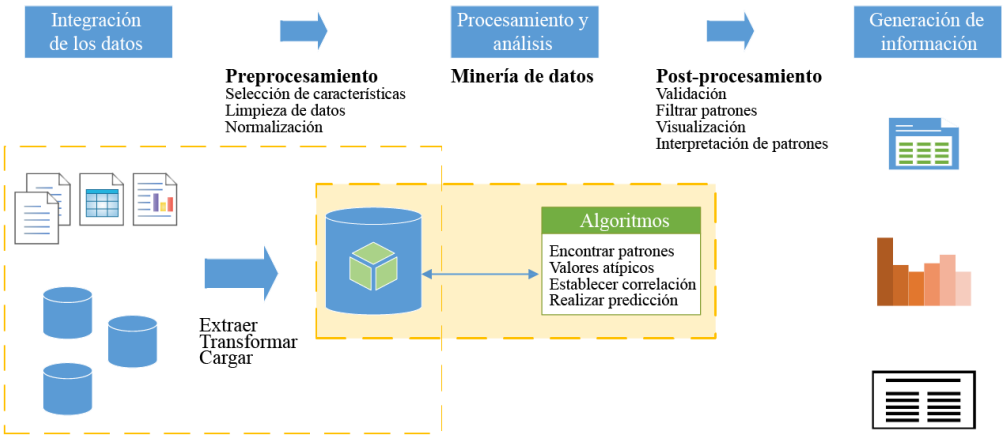


Figura 2: Proceso de preparación de los datos para su análisis.

2. Estado del arte

Aunque el análisis de datos no es algo nuevo, se puede notar que en los últimos años se ha incrementado el interés por la investigación en esta área, invirtiendo recursos científicos y financieros para generar, recolectar, analizar, comparar y condensar cantidades masivas de datos, para con esto obtener información valiosa, la cual permita obtener conclusiones y realizar predicciones que permitan entender y mejorar algún proceso (Pierson y Porway, 2017). Sin embargo, el encontrar patrones o valores extraños en series temporales representa un reto en la actualidad debido al volumen de datos, el cual conlleva a que se reduzca la velocidad de procesamiento. Los métodos estadísticos convencionales para identificar patrones en los datos incluyen el teorema de Bayes y el análisis de regresión y se basan en encontrar una correlación basada en muestras representativas de una población grande, sin embargo, los métodos estadísticos tradicionales requieren interacción con personas para validar la corrección de un modelo (Tanvi, Rekha y Kumar 2016; Croarkin y Tobias, 2012). A medida que los métodos estadísticos tradicionales se ven frustrados por la cantidad de datos, se vuelve más común el uso de los métodos de minería de datos, combinados con herramientas estadísticas y gestión de base de datos, los cuales, permiten procesar grandes cantidades de información y extraer patrones significativos (Urbach y Moore, 2011).

Las técnicas de minería de datos actuales se enfrentan a problemas en los cuales las bases de datos contienen ruido, es decir, valores sin relevancia o incorrectos. Algunos algoritmos de agrupamiento de datos que funcionan en muestras de datos pequeñas no se pueden implementar en series temporales debido a la dimensión y la variabilidad de los datos, lo cual se refleja en una alta complejidad computacional. Para resolver esta problemática las investigaciones se centran en utilizar métodos híbridos (del inglés *multi-step clustering*), en los cuales primero se utilizan técnicas para reducir para la dimensión del arreglo eliminando el ruido y convertir datos a un formato más conveniente para otro algoritmo de agrupamiento (Aghabozorgi, Seyed, Ying, 2015; Smith y Wunsch, 2015).

Debido a las características que presenta los datos en las series temporales, es difícil aplicar técnicas de agrupamiento de datos convencionales; muchas investigaciones se centran en reducir la dimensión de los algoritmos para que la serie de datos temporales sean compatibles con los algoritmos de agrupamiento, sin embargo, técnicas como *k*-medias, *k*-medoides o el agrupamiento jerárquico (del inglés *k-means, k-medoid* o *hierarchical clustering* respectivamente) no funcionan ya que estos trabajan con datos estáticos (Aggarwal, 2015). En la actualidad las investigaciones de la problemática en este método se enfocan en las siguientes áreas (Aghabozorgi, Seyed, Ying, 2015):

- **Medidas de similitud:** El problema con estas técnicas es que las series temporales son de dimensiones masivas y contienen ruido (datos sin relevancia) y valores atípicos.
- **Reducción de dimensión:** Transforman los datos crudos (del inglés *raw data*) en otra dimensión, mediante la transformación del conjunto en uno de menor dimensión o por la extracción de características. Este método es muy importante porque los datos en las series temporales se caracterizan por ser muy grandes y contener ruido, entonces, al reducir la dimensión de los datos, se reduce los requerimientos de almacenamiento y aumenta la velocidad de procesamiento. Existen diferentes algoritmos en esta área, sin embargo, para seleccionar uno se tienen que tomar en cuenta que tipo de datos se va a analizar, entre las desventajas se encuentra que no soportan cierto tipo de datos, no dan resultados estables, son costosos, su implementación es compleja, entre otros.
- **Algoritmos de agrupamiento de datos:** El agrupamiento de datos es una técnica en donde se colocan datos similares en grupos que están relacionados sin tener conocimiento de las

características de cada grupo. Se utiliza junto con el análisis de datos para descubrir patrones en conjuntos muy grandes de información. Los algoritmos se clasifican en métodos de particionamiento, jerárquicos, dirigido por modelos (del inglés *model based*), basados en densidad y técnicas de agrupamiento híbrido.

Actualmente las investigaciones se centran en la reducción de dimensión e intentan enfocarse en los conjuntos de datos temporales con diferente longitud. En cuanto a medir la similitud no se ha reportado que haya alguna mejoría en comparación con las técnicas más utilizadas, las cuales son los algoritmos de distancia euclidiana y DTW (del inglés *Dynamic Time Warping*). Debido a esto en donde existe mayor oportunidad para innovar es en los algoritmos de agrupamiento de datos, mediante algoritmos híbridos (Aghabozorgi, Seyed, Ying, 2015).

2.1 Algoritmo basado en densidad.

Una de las técnicas de agrupamiento de datos que no ha sido utilizada en conjunto de datos con el atributo de temporalidad son los algoritmos de agrupamiento basado en densidad, el cual, consiste en la identificación de clases, agrupando los datos disponibles en un número de grupos o clústeres mediante la estimación de la distribución de densidad de los nodos (Benmouiza y Cheknane, 2016). Estos algoritmos clasifican los datos de un conjunto en:

- **Centros:** se caracterizan por tener una densidad más alta que sus vecinos y por estar a una distancia relativamente alejada de otros puntos con densidad alta.
- **Puntos alcanzables:** se encuentran a un radio (máxima distancia permitida entre puntos de cada grupo) establecida de los puntos centros.
- **Ruido:** puntos que no son alcanzables desde algún centro.

Entre estas técnicas se encuentra el algoritmo basado en picos de densidad (*density peaks*) desarrollado por Rodríguez y Laio (2014), el cual utiliza la densidad de los datos y la distancia entre ellos para identificar los puntos núcleo y los miembros del grupo. Este algoritmo no necesita de un proceso iterativo, sin embargo, si el radio para estimar la distancia no se estima de manera adecuada entonces no toma en cuenta la diferencia en la distancia entre grupos o clústeres, dando como resultado que los grupos de pequeña densidad se combinen con otros (Hou y Cui, 2017). El agrupamiento basado en la búsqueda de picos de densidad se describe en el Algoritmo 1 y se basa

en la idea de que los puntos que se consideran centros (picos) se caracterizan por tener una densidad más alta que otros puntos vecinos y por una distancia grande en relación con otros puntos de densidad alta (Rodríguez y Laio, 2014). En este método el procedimiento de agrupamiento forma los grupos de clústeres independientemente de su forma y dimensión, además detecta valores atípicos. Para cada punto del conjunto de datos se calcula la densidad local (ρ) y la distancia entre puntos de alta densidad (δ). Ambas variables dependen de las distancias entre los puntos del conjunto de datos (d_{ij}).

Algoritmo 1. Algoritmo de agrupamiento basado en picos de densidad (Rodríguez y Laio, 2014).

1. Generar la matriz de proximidad (medida de similitud d_{ij}) utilizando distancia euclidiana.
2. Estimar el valor de la distancia d_c en la serie temporal
3. Encontrar los centros
   - Para cada punto y_i de la serie temporal $\bar{Y}$

$$\rho_i = \sum_{j \epsilon \bar{Y}, j \neq i} X(d_{ij} - d_c)$$

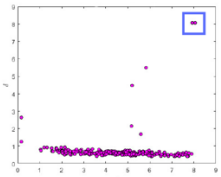
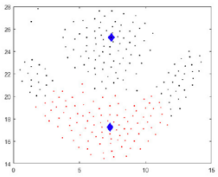
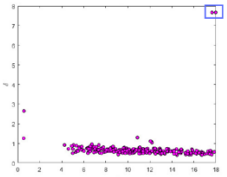
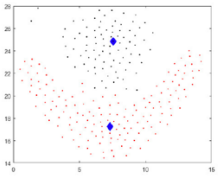
$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij})$$

$$\gamma_i = \delta_i \rho_i$$

   - Ordenar de manera descendente γ
4. Seleccionar los valores más altos de γ_i (centros) mediante un gráfico de selección
5. Asignar cada punto y_i con un centro

En donde d_c es un valor aleatorio de la serie temporal, el cual representa el radio mínimo que debe existir entre puntos para ser considerado dentro del grupo, el cual Rodríguez y Laio (2014) recomiendan que este entre un valor de 1-2% (%d_c) de los valores obtenidos de la matriz de proximidad, $\mathrm{X(x)} = 1$ sí $(d_{ij} - d_c) < 0$ y $\mathrm{X(x)} = 0$ en caso contrario, ρ representa la cantidad de puntos que tienen entre si una distancia menor a d_c, es decir, la densidad local entre puntos, δ es la distancia mínima a un punto de alta densidad, d_{ij} representa la matriz de proximidad y γ se utiliza para saber qué puntos se encuentran más alejados de otros y a su vez presentan mayor densidad local con otros puntos. Para encontrar los centros Rodríguez y Laio (2014) proponen en el Algoritmo 1 el parámetro γ ordenándolo de manera descendente y seleccionando los puntos que tengan mayor valor de γ, sin embargo, esto lo realizan de manera visual mediante una gráfica de decisión. La Tabla 1 se puede observar el resultado de la implementación del Algoritmo 1.

Tabla 1: Gráfico de decisión y resultados del agrupamiento de datos basado en picos de densidad (Rodríguez y Laio, 2014) con diferentes valores para calcular el radio d_c.

Grafico de selección de centros	parámetros	Resultado del agrupamiento de datos
	*dc* = 2% centros: 2	
	*dc* = 5% centros: 2	

Entre las ventajas del algoritmo basado en búsqueda de picos de densidad se encuentra que no requiere especificar el número de grupos que se quiere encontrar, detecta el ruido y valores atípicos, no es susceptible al orden en que se encuentren los datos en la base de datos. Sin embargo, tiene algunas limitaciones entre las cuales se encuentra que algunos métodos solo pueden agrupar los datos basándose en un umbral único (distancia euclidiana) por lo que en ocasiones no se puede caracterizar conjuntos de datos o bien no puede extraer los grupos más significativos (Campello, Moulavi y Sander, 2013). Otra problemática en este tipo de algoritmos es la integración en situaciones reales, debido a que las estructuras de datos contienen ruido o mediciones inexactas podría representar algoritmos con baja calidad. Para evitar esto es necesario modificar los métodos tradicionales mediante el uso del algoritmo híbrido que permitan lidiar con el volumen del conjunto datos, reducir la dimensión de este y excluir características irrelevantes (Aggarwal, 2015).

3. Método propuesto

Como parte de la investigación sobre agrupamiento de datos en series temporales en este trabajo se propone una variante del algoritmo basado en picos de densidad (Rodríguez y Laio, 2014) en el cual se seleccionen de manera automática los puntos que forman los centros de los grupos, los cuales se caracterizan por tener una densidad local alta en comparación con sus puntos

vecinos y estar alejados en distancia de otros puntos con densidad alta, lo cual puede ser representado de la siguiente manera:

$$\delta i \ > \ d_c \ \& \ \rho i \ >> \ \mu(\rho)$$

En donde $\mu(\rho)$ es el la media o promedio de la densidad local (ρ). Para realizar la selección de los centros de forma automática por el algoritmo, se propone la siguiente modificación al algoritmo original (Algoritmo 2):

Algoritmo 2. Variante del algoritmo de agrupamiento basado en picos de densidad (Rodríguez y Laio, 2014).

1. Generar la matriz de proximidad (medida de similitud d_{ij}) utilizando distancia euclidiana
2. Estimar el valor de la distancia d_c en la serie temporal
3. Encontrar los centros
   - Para cada punto y_i de la serie temporal $\bar{Y}$

     $$\rho_i = \sum_{j \epsilon \bar{Y}, j \neq i} X(d_{ij} - d_c)$$

     $$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij})$$

     $$\gamma_i = \delta_i \rho_i$$
   - Ordenar de manera descendente γ
4. Obtener posibles centros
   - Para cada punto y_i de la serie temporal $\bar{Y}$ (ordenada en base a γ)

     Si $\delta i \ > \ dc \ \& \ \rho i \ >> \ \mu(\rho)$ entonces

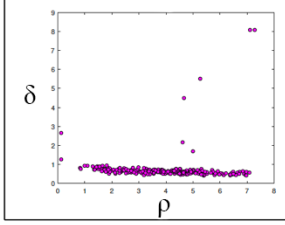
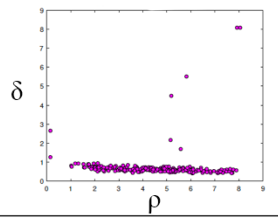
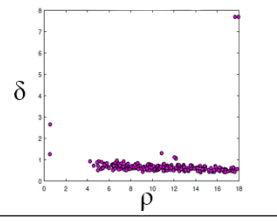
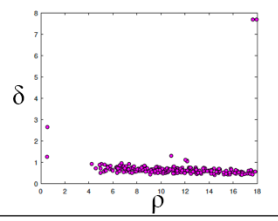
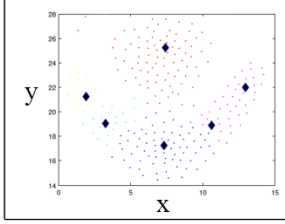
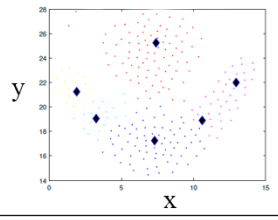
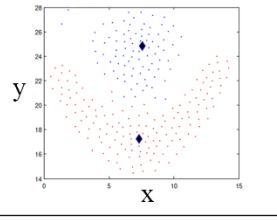
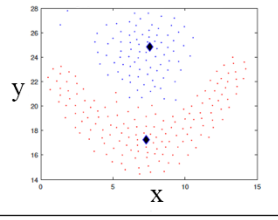
     $centro = centro + 1$
5. Asignar cada punto y_i con un centro

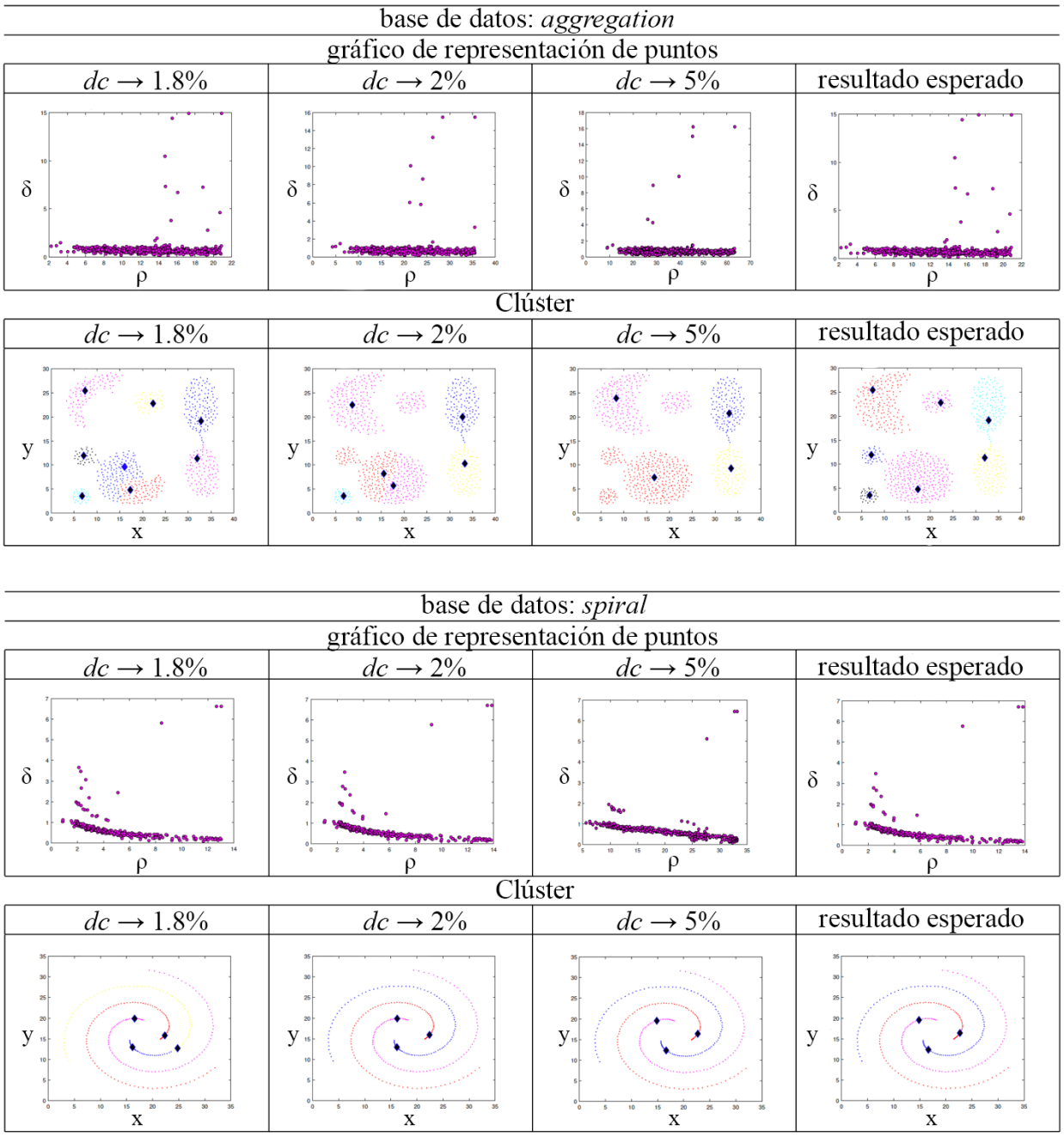
4. Experimentos y Resultados

El algoritmo fue implementado teniendo como base el método propuesto por Rodríguez y Laio (2014). En este estudio previo para evaluar la modificación al algoritmo de agrupamiento basado en picos de densidad no se utilizaron conjuntos de datos de series temporales, ya que es necesario un preprocesamiento. Para estos experimentos se utilizaron cuatro bases de datos obtenidas del repositorio de UCI (Dheeru and Karra Taniskidou, 2017), las cuales se describen en la Tabla 2, a su vez se muestra el grafico de decisión para visualizar los posibles centros para formar los grupos y el resultado del agrupamiento de datos utilizando como parámetro de entrada diferentes valores para calcular el d_c, el cual Rodríguez y Laio (2014) recomiendan que este entre un valor de 1-2%, sin embargo en estos experimentos tomamos un rango de 1-5%, ya que en

algunos conjuntos de datos el resultado esperado se encontró en 5%. En estos resultados se puede ver que la selección del valor para encontrar la variable d_c influye de manera directa en la formación de grupos, sin embargo, al ser un valor de entrada propuesto, es difícil estimar un valor que se considere fijo para diferentes conjuntos de datos. También se observa que los puntos que se consideran candidatos a ser los centros para formar clústeres son aquellos que están por encima de la media de la densidad, es decir, aquellos puntos que tienen densidad alta y están alejados de otros puntos con alta densidad, aquellos puntos que presenten una medida de distancia alta, pero no tengan densidad se consideran ruido.

Tabla 2: Resultados de la modificación al agrupamiento de datos basado en picos de densidad.

base de datos: *flame*			
gráfico de representación de puntos			
$dc \rightarrow 1.8\%$	$dc \rightarrow 2\%$	$dc \rightarrow 5\%$	resultado esperado
			
Clúster			
$dc \rightarrow 1.8\%$	$dc \rightarrow 2\%$	$dc \rightarrow 5\%$	resultado esperado
			



5. Discusión

Este estudio sirve como base para una investigación posterior y se enfoca en el agrupamiento de datos. Revisando la literatura se encontró que debido a las características que presentan los datos en las series temporales, es difícil aplicar técnicas de minería de datos

convencionales, a su vez se encontró que existía un área de oportunidad para innovar en los algoritmos de agrupamiento de datos basado en densidad.

Los resultados de esta primera aproximación muestran que, aunque se implementó un método para seleccionar de manera automática la cantidad de centros para formar los grupos (basado en la idea de que los centros son puntos que tienen alta densidad y están separados por una distancia relativamente grande de otros puntos considerados centros), se observa que los resultados de los experimentos dependen a su vez de la selección del valor aleatorio para encontrar el radio o distancia mínima entre puntos (d_c) para formar parte del clúster, el cual influye en la cantidad de centros y la formación de clústeres. Se pretende continuar desarrollando una variante del algoritmo basado en picos de densidad desarrollado por Rodríguez y Laio (2014), con enfoque en estimar de manera automática el valor del radio d_c e implementarlo en conjunto de datos con atributo de temporalidad, por lo cual a su vez, debido a las características que presentan los datos en las series temporales es necesario realizar un preprocesamiento de los datos, ya que, las series temporales son el resultado de recolección continua de datos y en ocasiones estos valores no solo provienen de base de datos estructuradas, sino también de diversas fuentes, por lo cual la integración de los datos puede ser redundante, incoherente, presentar ruido o pérdida o duplicidad de los mismos. Sin embargo, ni la recolección de datos, preparación de datos, ni la interpretación y evaluación de los resultados y la información no son parte de la minería de datos, pero forman parte del proceso de descubrir información en las bases de datos.

6. Referencias

Aggarwal, C. (2015). Data Mining. The textbook. New York, EE. UU. Springer.

Aghabozorgi, S., Seyed, A., Ying, W. (2015). Time-series clustering – A decade review. Elsevier, 53, 16-38. doi: 10.1007/978-3-319-14142-8.

Anderson, A., Senmelroth, D. (2015). Statics for big data. New Jersey, EE.UU. John Wiley & Sons, Inc.

Baoyan, L. (2014). Utilizing big data to build personalized technology and system of diagnosis and treatment in traditional Chinese medicine. Frontiers of Medicine, 8(3), 272-278. doi: 10.1007/s11684-014-0364-9.

Benmouiza, K., Cheknane, A. (2016). Density-Based Spatial Clustering of Application with noise Algorithm for the classificaction of Solar Radiation Time Series. International Conference on Modelling, Identification and Control. Algiers, Algeria. IEEE. doi: 10.1109/ICMIC.2016.7804123.

Buyya, R., Vahid, R. (2016). Big data: Principles and paradigms, 3-38. Cambridge, EE.UU. Elsevier.

Campello R., Moulavi D., Sander J. (2013) Density-Based Clustering Based on Hierarchical Density Estimates. Springer, 7819, 160-172. doi: 10.1007/978-3-642-37456-2_14.

Celis, A., Labrada, B. (2014). Bioestadística, 3-5. D.F., México. Manual Moderno.

Charu, A., Chandan, k. (2014). Data Clustering: Algorithms and Applications. New York, EE.UU. Taylor and Francis Group.

Collen, M. (2012). Computer Medical Databases: The first six decades (1950-2010), pp. 1-31,. London, Inglaterra. Springer. doi: 10.1007/978-0-85729-962-8.

Croarkin, C., Tobias, P. (2012). Engineering statics hand book. e-handbook.

Dua, D. y Karra, E. (2017). UCI machine learning repository. URL http://archive.ics.uci.edu/ml.

Falk, M., Frank, M. (2011). A First Course on Time Series Analysis with SAS, University of Würzburg.

Fu, T. (2011). A review on time series data mining. Elsevier, 24(1), 164-181. doi: 10.1016/j.engappai.2010.09.007.

Hou, J., Cui, H. (2017). Density Normalization in Density Peak Based Clustering. Springer, 103(10), 187-196. doi: 10.1007/978-3-319-58961-9_17.

Jiménez, M. (2014). Bases de datos relacionales y modelado de datos. D. F., México. IC, editorial.

Lemke, C., Budka, M., Gabrys, B. (2013). Metalearning: a survey of trends and technologies. Springer, 44(1), 117–130. doi: 10.1007/s10462-013-9406-y.

Pierson, L., Porway, J. (2017). Data Science for dummies. New Jersey, EE.UU. John Wiley & Sons, Inc.

Poenaru, C., Merezeanu, D., Dobrescu, R., Posdarascu, E. (2017). Advanced solutions for medical information storing: Clinical data warehouse. E-Health and Bioengineering Conference, 37-40. IEEE. Sinaia, Romania. doi: 10.1109/EHB.2017.7995355.

Rodríguez, A., Laio, A. (2014). Clustering by fast search and find of density peaks. Science, 344(6191):1492–1496. doi: DOI:10.1126/science.1242072.

Smith, C., Wunsch D. (2015) Time series prediction via two-step clustering, International Joint Conference on Neural Networks, 12-16. Killarney, Irland. IEEE. doi: 10.1109/IJCNN.2015.7280586.

Tanvi, A., Rekha, P. Kumar, S. (2016). Data Mining in Healthcare Informatics: techniques and Applications. 2016 3rd International Conference on Computing for Sustainable Global Development, 4023-4029. New Delhi, India. IEEE.

Thangarasu, G., Dominic, D. (2014). Prediction of Hidden Knowledge from Clinical Database using Data Mining Techniques. International Conference on Computer and Information Sciences, 1-5. IEEE. Kuala Lumpur, Malasia. doi: 10.1109/ICCOINS.2014.6868414.

Urbach, D., Moore, J.H. (2011). The spatial dimension in biological data mining. BioMed Central, 4(6). doi:10.1186/1756-0381-4-6.

Zhang, C., Shen X., Pei, X., Yao, Y. (2016). Applying Big Data Analytics into Network Security: Challenges, Techniques and Outlooks. International Conference on Smart Cloud (SmartCloud), 325-329. New York, EE.UU. IEEE. doi: 10.1109/SmartCloud.2016.62.