Agrupamiento de datos de series de tiempo. Estado del arte

Clustering of time series data. State of the art

Gustavo Cáceres Castellanos*
Jorge E. Rodríguez Rodríguez**

Fecha de recepción: 15 de enero del 2011
Fecha de aceptación: 16/ de junio del 2011

Palabras clave: datos de series de tiempo, agrupamiento de series de tiempo.

Abstract

Time series clustering has been an important research field in the last decade, providing useful and effective information in diverse domain. As outcome of the great existing interest for part of the scientific community of data mining area, innumerable research works have arisen that propose new algorithms and methodologies to identify cluster in the data time series. To provide an overview, this paper surveys and summarizes works that investigated the data time series clustering in diverse applications field. The basic concepts of time series clustering are presented and the surveyed works are organized into three groups: temporal-proximity-based, model-based and representation-based. The application areas are summarized with a brief description of the used data. The characteristics and particularities of some works are discussed.

Key words: Time series data, time seriesclustering

* Universidad Pedagógica y Tecnológica de Colombia. Teléfono: (57) 30056433462. Correo electrónico: gustavo.caceres@uptc.edu.co.
** Universidad Distrital Francisco José de Caldas. Teléfono: (57) 3203050462. Correo electrónico: jrodri@udistrital.edu.co.

Introducción

Debido al rápido desarrollo de las tecnologías de la información y la comunicación (TIC) y al avance de la globalización, las sociedades actuales se han enlazado conjuntamente de manera compleja en varios niveles y han surgido varios problemas en economía, medio ambiente, salud, seguridad, etc. Por consiguiente, la elucidación, predicción y control de estos sistemas complejos dinámicos son objetos de estudio muy importantes (Kitagawa, 2010, p. 252).

Como un hecho particular de dichos sistemas complejos, en algunos casos, se encuentran involucrados en ellos grandes volúmenes de datos que están representados en datos de series de tiempo. Una de las características de estas series de tiempo del mundo real es su ubicuidad (Yang y Chen, 2010, p.10); actualmente, se encuentran en muchas áreas de aplicación, como agricultura, finanzas, mercadeo, ingeniería, geofísica, medicina, economía, biología, bioquímica, meteorología, ciencias sociales, industria de procesos y producción, lenguaje natural, robótica, multimedia, entre otras (Wei, 2006, p. 634; Palit y Popovic, 2005, p. 381; Cowpertwait y Metcalfe, 2009, p. 262; Plant, Wohlschläger y Zherdin, 2009, pp. 914, 919; Pylvänen, Äyrämö y Kärkkäinen, 2009, p. 10; Savvides, Promponas y Fokianos, 2008, p. 15). Como consecuencia, en la última década, la administración de datos de series de tiempo se ha convertido en un área de investigación interesante e importante en la minería de datos (Ding et ál.,2008, p. 11; Kavitha y Punithavalli, 2010, p. 6; Guo, Jia y Zhang, 2008, p. 4; Luo, Liao y Zhan, 2010, p. 5), razón por la cual las técnicas y aplicaciones de minería de datos para el análisis de datos de series de tiempo han estado ganando amplia atención con temas de investigación interesantes sobre agrupamiento, búsqueda de similaridad, clasificación, predicción, etc. (Ding et ál.,2008, p. 11). Particularmente, el agrupamiento de series de tiempo ha centrado el interés de algunos investigadores, dando como resultado la introducción de gran cantidad de trabajos que desarrollan nuevas metodologías, como se evidencia en los artículos de estado del arte elaborados por Liao (2005, p. 18) y Kavitha y Punithavalli (2010, p. 6) y las diferentes técnicas revisadas en el presente artículo.

Debido al enorme interés existente, por parte de la comunidad científica, en relación con el área de minería de datos, enfocada a la investigación en agrupamiento de datos de series de tiempo, han surgido innumerables trabajos de investigación, en los últimos años, en los que se proponen nuevos algoritmos y metodologías para identificar grupos en los datos de series de tiempo. Las diversas metodologías desarrolladas han sido aplicadas en los más variados campos, como por ejemplo, la comparación de indicadores a través de países y regiones, investigación de datos financieros (Guo, Jia y Zhang, 2008, p. 4; Papanastassiou, 2009, p. 5; Olier y Vellido, 2008, p. 21); datos médicos, desde monitoreo, basado en sensores de pacientes afectados por patologías similares, datos granados en intervalos de tiempo regular en geología y climatología (Bandyopadhyay, Baragona y Maulik, 2010), en la bioquímica (Savvides, Promponas y Fokianos, 2008, p. 15; Kuenzel,2010, p. 14), meteorología (Horenko, 2010, p. 23), ecología (Debeljak et ál., 2010, p. 6), datos de expresión genética (Chiu, Hsu y Wang, 2010), entre otros.

Estos métodos están clasificados en tres importantes categorías: metodología basada en proximidad temporal, metodología basada en representación y metodología basada en modelos, en las cuales la idea básica de los dos últimos métodos es convertir las series

de tiempo en datos estáticos o parámetros de modelos, para luego aplicar directamente métodos de agrupamiento desarrollados para manipular los datos estáticos para completar las tareas de agrupamiento (Yang y Chen, 2010, p. 10; Guo, Jia, y Zhang, 2008; Wei y Jiang, 2010, pp. 135-138).

El presente artículo tiene la intención de introducir las bases del agrupamiento de series de tiempo y suministrar una revisión de trabajos realizados en años recientes. En el segundo aparte, se presentan conceptos fundamentales del agrupamiento datos de series de tiempo. En el tercero, se presenta la clasificación e informe de los trabajos de agrupamiento en series de tiempo que han sido presentados en la literatura abierta en los últimos años. Varios aspectos en relación con las técnicas presentadas son discutidos en el cuarto apartado y, finalmente, el artículo concluye con un apéndice, en el que se exponen las áreas de aplicación que son reportadas con links, en los cuales se encuentran datos de series de tiempo disponibles.

Agrupamiento de series de tiempo

El agrupamiento es una técnica poderosa reconocida en la minería de datos y ha sido estudiada exhaustivamente durante los últimos años. Este es un método para agrupar objetos de un conjunto de datos dentro de diferente "grupos", de acuerdo con las características encontradas en dichos datos. El mayor logro es la creación de particiones de objetos que tengan similitud entre ellos; teniendo un conjunto de datos X={x1; x2; … ; xn}, x se divide en k grupos {C1;C2; …;Ck}, donde Ci □ X(i = 1; 2;…; k). El resultado obtenido de este proceso puede revelar objetos/categorías desconocidas que pueden ayudar a un mejor entendimiento de los datos (Zhou, Li y Ma, 2009).

Similar al agrupamiento de puntos de datos estáticos, el propósito de agrupamiento de series de tiempo es resaltar la estructura inherente en un conjunto de datos de serie de tiempo agrupando los datos en un número de grupos homogéneos, de forma que la similaridad entre datos dentro de un grupo es máxima. Los datos de series de tiempo pueden ser de valor discreto o real, muestreados uniformemente o no, uni variables o multivariables y de igual longitud o longitud diferente (Chandrakala y Sekhar, 2008).

El agrupamiento de series de tiempo es una tarea importante en la minería de datos. En comparación con los problemas tradicionales de agrupamiento, en las series de tiempo se plantean algunas dificultades adicionales. La estructura única de series de tiempo hace que muchos métodos tradicionales de agrupamiento no se puedan aplicar a las series de tiempo de manera directa (Guo, Jia y Zhang, 2008). El objetivo final del análisis de agrupamiento de series de tiempo es dividir un conjunto de series de tiempo no etiquetadas, en las cuales las secuencias agrupadas deben ser coherentes u homogéneas (Yang y Chen, 2010, p. 10; Vilar, Alonso y Vilar, 2010, p. 16). Los componentes más importantes de un método de agrupamiento de series de tiempo son probablemente: la definición de una adecuada medida de similitud/distancia, y el algoritmo de agrupamiento (Piccardi y Calatroni, 2010). Sin importar el método de agrupamiento utilizado, siempre se requiere de una medida de distancia o similitud para la comparación de dos series de tiempo. La selección de esta medida de similitud determina el buen desempeño del agrupamiento (Zhang, Liu y Yan, 2010). Las medidas de similitud más ampliamente utilizadas para el agrupamiento de datos estáticos son la distancia euclidiana (*Euclidean Distance, EC*) y el coeficiente de correlación de Peason (*Peason's*

Correlation Coefficient, CC) (Zhang, Liu y Yan, 2010).

Conceptualmente, muchos de los criterios de disimilitud propuestos para agrupamiento de series de tiempo tratan con la noción de similaridad confiando en dos posibles criterios: proximidad entre datos de series en bruto y proximidad entre procesos de producción fundamentales. En ambos casos, la tarea de clasificación llega a ser inherentemente estática, dado que la búsqueda de la similitud es controlada solamente por el comportamiento de las series sobre sus periodos de observación (Vilar, Alonso y Vilar, 2010, p. 16).

Estudios empíricos en minería de datos de series de tiempo revelan que muchos de los algoritmos de agrupamiento existentes no trabajan bien, debido a su complejidad en la estructura fundamental y a su dependencia de los datos (Keogh y Kasetty, 2002, pp. 102-111), lo cual plantea un reto real en agrupamiento de series de tiempo de alta dimensionalidad, correlación temporal compleja y una cantidad sustancial de ruido (Yang y Chen, 2010, p. 10).

En el contexto del tratamiento de dependencia de datos, existen algoritmos de agrupamiento de series de tiempo que pueden ser clasificados en metodologías de agrupamiento *basadas en proximidad temporal, basadas en modelos y basadas en representación*. A continuación, se describen algunas de las características más relevantes de estas metodologías.

Agrupamiento de series de tiempo basado en proximidad temporal

Estos algoritmos trabajan directamente sobre las series de tiempo, en las cuales la correlación temporal es tratada directamente durante el análisis del agrupamiento por medio de medidas de similitud. Entre las ventajas presentadas por este método, se encuentra que previenen la pérdida de información y es una forma directa de capturar el comportamiento dinámico de una serie de tiempo; adicionalmente, es un medio flexible para tratar con longitudes de datos de series de tiempo variables.

Sin embargo, estas metodologías regularmente son sensibles a la inicialización, presentan problemas de selección del modelo y una complejidad computacional alta. Algunos trabajos desarrollados en esta metodología son: algoritmo de agrupamiento para video basado en series de tiempo de trayectoria de movimiento de transformadas wavelet de movimiento de objetos en video (Luo, Liao y Zhan, 2010), análisis de datos de series de tiempo sobre vegetación de agroecología utilizando árboles de agrupamiento predictivo (Debeljak et ál., 2010, p. 6), agrupamiento de datos de series de tiempo del gen basado en representaciones continuas y una mediada de similaridad basada en energía (Zhang, Liu y Yan, 2010), agrupamiento adaptativo para series de tiempo (Douzal-Chouakria, Diallo y Giroud, 2009, p. 13), un procedimiento de agrupamiento para minería exploratoria de series de tiempo vector (Liao, 2007, p. 13), un algoritmo de agrupamiento para datos de series de tiempo (Yin, Zhou y Xie, 2006).

Agrupamiento de series de tiempo basado en modelos

Trabajan directamente sobre las series de tiempo, en las cuales la correlación temporal es tratada directamente durante el análisis del agrupamiento por medio de medidas de similitud. Los grupos de las series de tiempo son especificados por una serie de modelos dinámicos, identifican la independencia de datos y la regularidad, más allá del comportamiento dinámico de las series de tiempo.

Esta metodología es adecuada para enfrentar dependencias de datos entre las series de tiempo; estas son caracterizadas con modelos generativos. Como desventaja, presentan una alta complejidad computacional y problemas en la selección del modelo. Algunos de los trabajos desarrollados utilizando esta metodología son: un algoritmo basado en interacción de series de tiempo multivariables (Plant, Wohlschläger y Zherdin, 2009), agrupamiento de series de tiempo meteorológicas no estacionarias (Horenko, 2010, p. 23), agrupamiento de series de tiempo no lineales basado en densidades de pronóstico no paramétricas (Vilar, Alonso y Vilar, 2010, pp. 2850-2865), algoritmo de agrupamiento k-means basado en modelo oculto de Markov (Wei y Jiang, 2010, pp. 135-138), un método de agrupamiento jerárquico basado en HMM para series de tiempo de expresión genética (Zhao y Deng, 2010).

Agrupamiento de series de tiempo basado en representaciones

Se extrae un conjunto de características de la serie de tiempo, convierte las series de tiempo en dimensionalidades más bajas con características de espacio, en las cuales cualquier algoritmo de agrupamiento de datos estático existente puede ser aplicado, lo cual, en especial, eficiente en computación. Presenta la ventaja de reducir significativamente el costo computacional y su compatibilidad con los algoritmos de agrupamiento para datos estáticos. No obstante, una representación tiende a codificar solamente aquellas características bien presentadas en su espacio de representación, lo cual inevitablemente causa pérdida de otra información útil llevada en la serie de tiempo original. Debido a la alta complejidad y variedad de las series de tiempo, no existe una representación universal que caracterice perfectamente diferentes tipos de series de tiempo (Ding et ál., 2008, pp. 1542–1552). Por consiguiente, una representación es simplemente aplicable a clases de series de tiempo, en las cuales sus características salientes puedan ser completamente capturadas en el espacio de representación; pero esta información es difícilmente disponible sin conocimiento previo y un análisis cuidadoso. Con esta metodología encontramos trabajos desarrollados como: agrupamiento de series de tiempo por medio de conjunto de redes RCPL(Yang y Chen, 2010. p. 10), agrupamiento de series de tiempo por análisis de comunidad de redes (Piccardi y Calatroni, 2010), agrupamiento difuso en series de tiempo en el dominio de frecuencia (Maharaj, E.-A. and D'urso, 2010, p. 25), agrupamiento basado en LLE (Locally Linear Embedding) para series de tiempo financieras multivariables [25]

Trabajos realizados

Basados en la clasificación para algoritmos de agrupamiento de series de tiempo, enunciada anteriormente, a continuación se presentan algunos trabajos que se han desarrollado en años recientes, que nos permiten ver el avance y las tendencias de la investigación en el campo de la minería de datos y particularmente del agrupamiento en series de tiempo.

Agrupamiento de series de tiempo basada en proximidad temporal

Estas metodologías centran su esfuerzo del proceso de agrupamiento en diseñar medidas de similitud o distancia entre secuencias, siendo una de las más efectivas la DTW (Wei y Jiang, 2010, pp. 135-138). A continuación, se enuncian algunas de las técnicas de agrupamiento de series de tiempo basadas en proximidad temporal, desarrolladas en los últimos años:

Cuando se trabaja con series de tiempo de expresión génica, a menudo, los puntos del tiempo no son muestreados uniformemente, lo cual representa un problema en el desempeño del agrupamiento. Con el propósito de mejorar dicho desempeño, Zhang, Liu y Yan (2010), presentan una nueva metodología de agrupamiento, que se basa en la representación continua y medidas de similaridad basadas en energía. La metodología propuesta modela cada perfil de la expresión génica como una expansión *B-spline*, para lo cual son estimados los coeficientes *spline* por medio del esquema del cuadrado mínimo regularizado sobre los datos observados. Luego de ajustar la representación continua del perfil de la expresividad del gen, utilizan la medida de similaridad basada en energía para tomar en cuenta la información temporal y cambios relativos de la serie de tiempo. El método propuesto está enfocado a mejorar el agrupamiento de series de tiempo del gen combinando la representación continua y la medida de similaridad, basada en energía, más que el algoritmo de agrupamiento mismo. Este método puede ser extendido a otros algoritmos de agrupamiento.

Luo, Liao y Zhan (2010) proponen un análisis de similaridad y un algoritmo de agrupamiento de videos, basado en la transformada *wavelet* de datos de series de tiempo de la trayectoria del movimiento de objetos en los videos. Este algoritmo detecta el movimiento de objetos desde la observación de escenas del video, calcula el centroide del movimiento del objeto y usa la serie centroide para caracterizar la trayectoria del movimiento del objeto. Luego, utiliza el método de análisis wavelet para lograr la reducción de la dimensionalidad y obtener el primer coeficiente wavelet k para sustituir los datos de serie de tiempo original. Basado en la distancia euclidiana, utiliza dos reglas de juicio para determinar la similaridad de los datos de series de tiempo y agruparlos, utiliza las reglas para ejecutar la búsqueda de similitud y agrupamiento del video. Para realizar el agrupamiento, se utiliza el algoritmo *K-means*.

Debeljak et ál. (2010, p. 6) describen una aplicación exitosa de árboles de agrupamiento predictivo en el análisis de series de tiempo en un conjunto de datos ecológico grande y complejo. Fueron combinados tres métodos en esta aplicación DTW para definir la distancia entre dos series de tiempo; el algoritmo *K-medoids* para hacer una partición de las series de tiempo y de los árboles de agrupamiento predictivo, a fin de asociar una variable destino, en este caso, la segunda serie de tiempo, para variables independientes (atributos de entrada), incluyendo pertenencias en los grupos definidos para la primera serie de tiempo.

Por su parte, Douzal-Chouakria, Diallo y Giroud (2009, p. 13)aplican un algoritmo de agrupamiento adaptativo de series de tiempo para identificar el ciclo celular expresado en los genes. Este algoritmo primero se basa en un índice de disimilitud, cubriendo tanto la proximidad sobre los valores como los comportamientos. El método de agrupamiento debería ayudar a aprender la contribución apropiada sobre valores y comportamiento del índice de disimilitud. Finalmente, este permite extraer un conjunto de genes caracterizando, bien las fases del ciclo celular. Se utilizó el algoritmo *Partitioning Around Medoids* (PAM) para dividir el conjunto de genes estudiado en *n* grupos (siendo *n* el número de fases o interfaces de ciclo celular estudiado). Este algoritmo, siendo más robusto en relación al manejo de valores atípicos que el *K-means*, permite un mayor detalle en el análisis de la partición, suministrando características de agrupamiento; particularmente, indica, para cada gen, si este está bien clasificado o si se encuentra en el límite del grupo.

Liao (2007, p. 13) propone un procedimiento de dos pasos para minería exploratoria de series de tiempo multivariada valor continuo (*real-valued*) utilizando métodos de agrupamiento basados en partición. Esta metodología trabaja directamente sobre los datos en bruto y es capaz de manipular series de tiempo de longitud diferente. El primer paso de la metodología propuesta convierte la serie de tiempo continua en serie se tiempo univariable de valor discreto. Este primer paso se puede considerar un paso de reducción de dimensión. Se lleva acabo aplicando un algoritmo de agrupamiento a los datos multivariados trazados en el tiempo.

El segundo paso agrupa los *n* números de la serie de tiempo de valor discreto convertida dentro de un número predeterminado de grupos. Se tomaron dos diferentes metodologías para esto, la primera utiliza la distancia DTW y algoritmos de agrupamiento jerárquico, o basado en *medoid,* estos son necesarios si esta metodología es tomada. La segunda metodología primero expresa cada valor discreto de la serie de tiempo univariable como una matriz de transición de probabilidades *n*. Esta segunda alternativa puede hacer uso de todos los algoritmos de agrupamiento existentes.

Yin, Zhou y Xie (2006) proponen una mejora al método clásico de agrupamiento jerárquico, desarrollando un método de intercambio basado en la metodología de codificación Bitmap. Este método consta de seis pasos:

- Inicia asignando cada ítem a su propio grupo, así que si son *n* ítems. Hay *n* grupos, cada uno conteniendo solamente un ítem.
- Usa la relación grey como medida de similaridad de series de tiempo y deja las similitudes entre grupos igual a las similitudes entre los ítems que ellos contienen.
- Encuentra pares de grupos similares y los combina dentro de un grupo simple, así que un grupo puede ser reducido.
- Calcula el enlace promedio como similaridad entre los nuevos grupos y cada uno de los grupos viejos.
- Repite los pasos tres y cuatro hasta obtener *k* grupos.
- Adopta el intercambio basado en la metodología de codificación bitmap para refinar los *k* grupos del paso cinco y luego obtener los nuevos *k* grupos.

Chen (2007) resuelve algunos problemas encontrados en la popular técnica de agrupamiento STS, por consiguiente, se pudo concluir que esta técnica carece de sentido. Propone el algoritmo TF que produce agrupamiento de series de tiempo útiles. La metodología está basada en restringir el espacio de agrupamiento para extender solamente la región visitada por las series de tiempo en la subdivisión del vector espacio. Dicho algoritmo fue validado en doce conjuntos de datos sintéticos y de la vida real.

La tabla 1 resume los principales componentes utilizados en cada algoritmo de agrupamiento basado en proximidad temporal.

Agrupamiento de series de tiempo basada en modelos

La metodología basada en modelos como primer paso modela las series de tiempo y luego aplica el algoritmo de agrupamiento sobre el modelo obtenido. En algunos casos, la desventaja presentada en relación con otras metodologías es la complejidad computacional. Algunas de las metodologías propuestas en los últimos años se presentan a continuación.

Tabla 1. Resumen de los algoritmos de series de tiempo basados en proximidad temporal

Artículo	Variable	Longitud	Medida de distancia	Algortimo de agrupamiento	Aplicación
Zhang et ál. (2010)	Simple	Igual	SimilB	K-means	Expresión genética.
Luo et ál. (2010)	Simple	Igual	Euclidiana	K-means	Análisis de videos de deporte atlético.
Debeljak et ál. (2010)	Múlti-ples	Diferentes	DTW	K-medoids	Cobertura de cultivos y maleza Reino Unido.
Qu et ál. (2010)	Simple	Igual	Euclidiana	Fuzzy C-means	Conjuntos de da-tos sintéticos.
Douzal-Chouakria et ál. (2009, p. 13)	Simple	igual	Euclidiana	PAM	Expresión genética.
Liao (2007, p. 13)	Múltiple	Diferentes	DTW o	Jerárquico	Datos Sintéticosl gene-rados por procedimien-to *varmasim* de SAS.
			Kullback-Liebeler	Fuzzy C-means	
Yin (2006)	Simple	Igual	Relación gris	Jerárqui-co mejorado	Sistema de administración de tráfico en Washington.
Chen (2007)	Simple	Igual	Euclidiana	K-medoids	Doce series de tiempo del mundo real y sintéticas
Lytkin (2008, p. 33)	Múltiple	Igual	Euclidiana	Basado en gradiente	Fondos de inversión común.

Zhao y Deng (2010) proponen un novedoso método de agrupamiento para analizar los datos de series de tiempo de expresión génica, denominado agrupamiento jerárquico basado en el Modelo de Markov Oculto (*Hidden Markov Model-based hierarchical clustering*, HMM-HC). Convierten datos de puntos del tiempo en símbolos discretos sobre la base del hecho de que el logaritmo del dato obedece aproximadamente a una distribución normal y construyen los modelos de Markov ocultos con estos símbolos para secuencias del gen. En una serie de tiempo de expresión génica, el dato en el punto del tiempo es correlacionado con otros; el uso de HMM puede ayudar a tomar ventajas de su correlación especial. El algoritmo de agrupamiento HMM-HC está dividido en dos etapas. Primero, modelar los datos de la serie de tiempo de expresión génica, ya que cada modelo representa un grupo. Segundo, agrupar los modelos con la estrategia de jerarquía.

Wei y Jiang (2010, pp. 135-138) proponen un método que busca subsanar algunas insuficiencias presentadas en los algoritmos de agrupamiento basados en los modelos de Markov, como secuencias largas y longitudes iguales. El algoritmo de agrupamiento de series de tiempo *K-means*, basado en HMM; este algoritmo primero parte el conjunto de datos usando DWT, agrupa las secuencias similares en un grupo, a fin de evitar sobreajustes (*overfitting*) para el entrenamiento de muestras. Luego, entrena un modelo HMM para cada grupo, calcula las probabilidades que cada secuencia pertenece a cada modelo, asigna cada secuencia en una clase corres-

pondiente con el principio de máxima probabilidad. Por último, una mejora iterativa del modelo hasta que la función de probabilidad de unión converja a un umbral preajustado. Experimentos sobre datos artificiales han mostrado que la metodología propuesta funciona mucho mejor que las metodologías estándar de agrupamiento basadas en HMM.

Vilar, Alonso y Vilar (2010, p. 16) extienden el procedimiento de agrupamiento, propuesto por Alonso et ál (2006, p. 15), para cubrir el caso de modelos no paramétricos de autorregresión arbitraria. Esta metodología no asume ningún modelo paramétrico para la verdadera estructura autorregresiva de la serie, que es estimada usando técnicas de *kernel smoothing*. Como consecuencia, solamente aproximaciones no paramétricas para la verdadera función autorregresiva están disponibles en este nuevo ajuste y, por tanto, el filtro de *bootstrap* no es un proceso de producción válido. En este procedimiento, el mecanismo usado para obtener las predicciones *bootstrap* está basado en imitar el proceso de producción usando un estimado no paramétrico de la función autorregresiva y un remuestreo *bootstrap* de los residuos no paramétricos. De esta manera, suministran un dispositivo útil para clasificar series de tiempo autorregresivas no lineales, incluyendo modelos paramétricos estudiados ampliamente, como el autorregresivo de umbral (*Threshold Autoregressive*, TAR), el autorregresivo exponencial (*Exponential Autoregressive*, EXPAR), el autorregresivo de transición sin problemas (*Smooth-Transition Autoregressive*, STAR), y el bi lineal, entre otros.

Pamminger y Frühwirth-Schnatter (2010, p. 24) plantean dos metodologías para agrupamiento basado en modelos de series de tiempo categóricas basadas en cadenas de Markov de primer orden de tiempo homogéneo. Para el agrupamiento de cadenas de Markov, las probabilidades de transición individual son fijadas a una matriz de transición de grupo específico. En la nueva metodología, llamada agrupamiento multinomial de *Dirichelt,* las filas de las matrices de transición individual se derivan del grupo medio y siguen una distribución Dirichlet con híper parámetros de grupo específico desconocidos. La estimación es llevada mediante Monte Carlo Cadenas de Markov (*Markov Chain Monte Carlo*, MCMC). Varios criterios de agrupamiento bien conocidos son aplicados para seleccionar el número de grupos.

Horenko (2010, p. 23) presenta un método para agrupamiento de series de tiempo meteorológicas no estacionarias multidimensionales. Esta metodología está basada en la optimización del agrupamiento promediado regularizado funcional, describiendo la calidad de la representación de datos en términos de *k* modelos de regresión y un proceso oculto meta estable intercambiado entre estos. El algoritmo de agrupamiento numérico propuesto está basado en aplicación del método de elementos finitos (*finite element method*, FEM) para el problema del análisis de series de tiempo no estacionarias. La principal ventaja del algoritmo presentado en comparación con HMM y con modelos de mezcla finita es que ninguna suposición a priori acerca del modelo de probabilidad para los procesos ocultos y observados es necesaria. Otra característica numérica atractiva de este algoritmo es la posibilidad para seleccionar el número óptimo de grupos metaestables y una oportunidad para controlar la ambigüedad de la descomposición resultante posteriormente, basado en la distinguibilidad estadística de los estados de grupos persistentes resultantes.

Plant, Wohlschläger y Zherdin (2009) proponen una noción nueva para series de tiempo multivariables. Definen un grupo como un conjunto de objetos compartiendo un patrón de interacción específica entre dimen

siones: proponen un algoritmo eficiente para agrupamiento basado en interacciones denominado interacción *K-Means* (IKM). Este algoritmo demostró que el grupo basado en interacción es un complemento valioso para agrupamiento en series de tiempo multivariado. La tabla 2 resume los principales componentes usados en cada algoritmo basado en modelos.

Tabla 2. Resumen de algoritmos de agrupamiento para series de tiempo basado en modelos

Artículo	Variable	Modelo	Medida de distancia	Algoritmo de agrupamiento	Aplicación
Zhao y Deng (2010)	Simple	HMM	Distancia de Ward	Hierarchical	Datos de expresión genética.
Wei y Jiang (2010)	Simple	HMM	DTW	K-means	Datos artificiales generados.
Vilar et ál. (2010)	Múltiple	Autorregresión bootstrap	L^1 y L^2	Jerárquico agregativo	Índice de producción industrial.
Horenko (2010)	Multiple	Modelos de regresión	Euclidiana	FEM-K_Trends (Finite element method-K-Trends)	Temperaturas diarias desde 1958-2002.
Piccardi y Calatroni (2010)	Múltiple	Modelo de red	Euclidiana	Jerárquico agregativo	Series de tiempo financieras.
Pamminger (2010)	Múltiple	Multinomial Dirichlet	N/A	multinomial Dirichlet	Panel de movilidad salarial Australiana.
Plant y Frühwirth-Schnatter (2009)	Múltiple	Modelo lineal	Euclidiana	K-means	Datos sintéticos fMRI, EEG, CAD.
Alonso et ál. (2006)	Múltiple	Procedimiento de bootstrap	L^2	Jerárquico aglomerativo	Emisión de CO_2.

Agrupamiento de series de tiempo basada en representaciones

Este método consiste primero en extraer características de la serie de tiempo y luego aplicar el algoritmo de agrupamiento sobre la representación de las características de la serie de tiempo. La desventaja que plantean algunos autores al utilizar esta metodología es que se pierde alguna información de la serie de tiempo al realizar la extracción. A continuación, se presentan algunas metodologías propuestas en años recientes.

Yang y Cheng (2010. p. 10) proponen una nueva metodología práctica aún para agrupamiento de series de tiempo por medio de una combinación de redes de aprendizaje competitivo rival-penalized (*rival-penalized competitive learning*, RPCL) con diferentes representaciones, las cuales direccionan tanto el agrupamiento y los problemas de selección del modelo en el análisis de agrupamiento de manera general. Esta metodología está motivada por el éxito previo en el uso de diferentes representaciones para construir un modelo combinado para tratar con difíciles tareas de aprendizaje, supervisado y semisupervisado donde el uso de diferentes representaciones explota mejor la información llevada en los datos en bruto y, por consiguiente, conduce a un mejor desempeño.

Por cada representación individual, primero emplean una red RPCL para análisis

de agrupamiento de la selección del modelo automático; por otra parte, la naturaleza de la red RPCL a menudo induce a análisis de agrupamiento rápidos. Esta combinación de redes RPCL hace frente a la diversidad de grupos generados por las redes RPCL sobre las diferentes representaciones reconciliándolas en una forma óptima. Como resultado del conjunto de red, RPCL reduce considerablemente ambigüedades resultantes del uso de diferentes inicializaciones, porcentajes de aprendizaje y condiciones de terminación en una red RPCL individual y, además, aumenta su capacidad de selección de modelo automático sobre diferentes representaciones.

La arquitectura del modelo de ensamble RPCL consta de tres módulos, es decir, extracción de la representación, aprendizaje competitivo RPCL y ensamble de agrupamiento. En el módulo de extracción, varias representaciones de naturaleza complementaria son utilizadas (*piecewise local statistics*, PLS; *piecewise discrete wavelet transform*, PDWT; *polynimial curve fitting*, PCF y *discrete fourier transforms*, DFT). Así, las series de tiempo son transformadas en diferentes representaciones para ser la entrada de las redes RPCL. En el módulo de aprendizaje competitivo, una red RPCL sobre una representación individual sería entrenada con sus reglas de aprendizaje para análisis de agrupamiento.

Piccardi y Calatroni (2010) proponen un algoritmo de agrupamiento no convencional, que es, de hecho, una aplicación de un desarrollo reciente de la teoría de redes complejas, llamado análisis de comunidad. Una red con *n* nodos es asociada a el conjunto de *n* series de tiempo, con el peso del enlace (i,j) cuantificando la similitud entre los dos componentes de la serie. Luego, buscando para comunidades de redes, se permite identificar grupos de nodos (por ejemplo, series de tiempo) con fuerte similaridad.

Maharaj y D'urso (2010, p. 25) presentaron una metodología de agrupamiento Fuzzy para series de tiempo basado en coeficientes cepstral, esto es, basado en lógica difusa; asimismo, clasificaron las series de tiempo en el dominio de frecuencia considerando su representación cepstral. En este enfoque, a diferencia del enfoque tradicional (no fuzzy), los elementos de datos pertenecen a más de un grupo y asociados con cada elemento está un conjunto de niveles de pertenencia. Esto indica la fuerza de la asociación entre ese elemento de dato y un grupo en particular. El agrupamiento fuzzy es un proceso de asignación de esos niveles de pertenencia y, luego, usando estos para asignar los elementos de datos a uno o más grupos.

Lai, Chung y Tseng (2010, p. 8) proponen un método de agrupamiento de dos niveles llamado 2LTSC (2 [*level time series clustering*]), el cual puede suministrar un profundo entendimiento para agrupamiento de series de tiempo por medio de la consideración de diferentes granulaciones de tiempo. El método considera tanto la serie de tiempo completa, denominada nivel 1, en el primer nivel, como la información subdividida de la serie de tiemi 2 debería ser diferente y así es también considerada en el segundo nivel este método.

Chiu, Hsu y Wang (2010) presentan un algoritmo de agrupamiento no supervisado para analizar series de tiempo de datos de expresión genética, el cual no requiere conocimiento previo. Este algoritmo combina la propagación de afinidad y el agrupamiento de consenso con varios intervalos de series de tiempo suministrando robustez y precisión progresiva. Este algoritmo suministra un apropiado y efectivo análisis sobre expe-

Zhou, Li y Ma (2009) proponen un enfoque basado en la agrupación local incrustación lineal (*Local Linear Embedding*, LLE) para la base de datos financieros, en este enfoque, primero se convierten los datos de series de tiempo *raw* en dimensiones menores mediante el algoritmo LLE, y luego se aplica un algoritmo modificado de *k-means* a las características de vectores extraídas. Primer paso: se realiza una reducción de la dimensionalidad por medio del LLE. Segundo paso: luego se aplica el algoritmo de agrupamiento *K-means* ajustado para agrupar la matriz de mezcla W obtenida por LLE. El algoritmo intenta dividir *n* cantidad de objetos en *k* grupos, donde cada uno tendría un objeto como el centro del grupo, lo que representa que todos los objetos de datos están asociados a un grupo. Luego de esto asignan cada uno de los objetos al grupo adecuado de acuerdo con el centro del grupo definido, una vez todos los objetos están asignados a algún grupo se recalculan los centros promediando los miembros de cada grupo, hasta estabilizar el centro. Con esto, se logra, después de cada iteración, una mejor calidad en los grupos y los centros de estos.

Savvides, Promponas y Fokianos (2008, p. 15) proponen un algoritmo de agrupamiento aplicado a series de tiempo biológicas. Para esto, se plantea una nueva medida de distancia basada en el coeficiente cepstral, el cual transporta información acerca del registro de espectros de una serie de tiempo estacionaria. Una vez que estos coeficientes son estimados, esta medida de distancia es dada como entrada a un método de agrupamiento para producir grupos disyuntos de datos. Para el agrupamiento se utiliza el algoritmo Diana (*Divisive Analysis*), el cual calcula una jerarquía divisionista, mientras que otros procedimientos para agrupamiento jerárquico es aglomerativo. El algoritmo Diana crea una jerarquía de grupos, iniciando con un grupo grande que incluye todos los puntos de datos. Luego, los grupos son divididos hasta que cada grupo contiene una observación única.

Otranto (2008, p. 14) propone un procedimiento de agrupamiento basado en herramientas estadísticas simples. En particular, considera la interferencia cuadrática del retorno de una serie de tiempo financiera como la volatilidad de las series. Luego, utilizan la representación GARCH (*Generalized Autoregressive Conditional Heteroskedasticity*) de una varianza condicional para derivar el modelo fundamental de la inferencia cuadrática. De este modelo se separa la volatilidad dentro de una parte constante y parte variante en el tiempo; esta subdivisión puede tener una interpretación atractiva, en particular, cuando se usa la volatilidad para representar el riesgo del activo. La parte constante de la volatilidad es medida en forma natural, mientras que miden la parte variante en el tiempo extendiendo la idea de distancia entre modelos AR para la familia GARCH. Se aplica un algoritmo aglomerativo y como característica de este procedimiento, a diferencia de los principales algoritmos aglomerativos, el número de grupos es detectado automáticamente y no es determinado por el usuario.

Hsu y Chen (2008) proponen una nueva metodología para estimar el índice de equilibrio utilizando mapas de autoorganización (self organizing map, SOM), que sirve como una red de dos capas no supervisadas que puede organizar un mapeo topológico. El mapeo resultante muestra las relaciones naturales entre los patrones que están dados en la red. Por su parte, SOM es adecuado para análisis de agrupamiento y ha sido aplicado para predicción de series de tiempo y en el proceso de investigación desarrollado primero, las series de tiempo son agrupadas con SOM. Segundo, el índice de equilibrio es calcula-

do basado en el grupo de la similaridad de los patrones de series de tiempo. Se asumió que los patrones de similaridad de las series de tiempo tendrán el mismo comportamiento y serán adecuados para la estimación del índice de equilibrio. Finalmente, varios modelos basados en SOM propuestos son investigados y comparados con los modelos tradicionales.

Bandyopadhyay, Baragona y Maulik (2010) proponen técnicas para agrupamiento de series de tiempo univariables y multivariables. Se utilizan dos pasos básicos, extracción de características y asignación de series de tiempo a grupos, de acuerdo con un criterio de optimización. Se ha introducido la optimización de Pareto como un criterio valioso para resolver problemas de optimización en los cuales varios objetivos conflictivos a menudo tienen que ser tomados en cuenta simultáneamente. Con el propósito de subir la velocidad de computación hasta que los pasos de optimización estén involucrados, se aplican algoritmos genéticos, debido a que estos son meta heurística más popular en cuanto a problemas de agrupamiento y buena cantidad de conocimiento está disponible. Una metodología de agrupamiento fuzzy ha sido considerada para agrupamiento de datos series de tiempo univariable y multivariable.

Chandrakala y Sekhar (2008) proponen un método de agrupamiento basado en densidad dentro del espacio de características de núcleo para agrupamiento de datos de serie de tiempo multivarible de longitud variable. Este método puede ser usado para agrupamiento de cualquier tipo de estructura de datos, suministrando un núcleo que puede manipular la clase de datos es usado. Presentan métodos heurísticos para encontrar los valores iniciales de los parámetros usados en el algoritmo propuesto.

Alonso et ál. (2006, p. 15) proponen una nueva metodología basada en modelos que crean las observaciones, pero con respecto a la predicción en un tiempo futuro específico. Este procedimiento está basado en la completa predicción de densidades para cada una de las series observadas en la muestra, en lugar de enfocarlas en el punto de pronóstico. Se aplica un procedimiento *bootstrap* de tamiz suavizado, combinado con estimación de densidades kernel no paramétricas ideado para aproximar la distribución de predicciones. Esto es hecho en un contexto general, sin restricciones a la habitual hipótesis de Gaussinidad. Las diferencias entre cada par de densidades bootstrap suministran una matriz de disimilitud, que será usada para examinar posibles estructuras de agrupamiento. El método propuesto reduce las dimensiones altas del problema de 3D, convirtiendo la estructura del cubo de datos de diferentes series de tiempo p, medidos en m individuos sobre T momentos de tiempo, en una o más estructuras 2D de p predicciones, obtenidas para m individuos en un tiempo fijo $T+h$. La metodología desarrollada consta de tres pasos. Paso 1: cálculo de predicciones; paso 2: cálculo de la matriz de disimilitud; paso 3: aplicación de un método clásico de agrupamiento a la matriz de disimilitud.

Guo et ál. (2008) tienen en cuenta que los resultados del agrupamiento no pueden reflejar apropiadamente la similaridad de las series de tiempo, debido a la distorsión del ruido y detalles en las series de tiempo, proponen una nueva metodología basada en la descomposición y eliminación de ruido wavelet. La aplicación de esta metodología primero realizan una descomposición Wavelet de la serie de tiempo, luego se realiza una reducción del ruido wavelet y una reconstrucción Wavelet. Y finalmente, se lleva a cabo un agrupamiento con el método *K-means*.

Wang en [46] presenta un nuevo método para agrupamiento multivariado de series de tiempo basado en la estructura global de datos. Una serie de tiempo de una sola variable puede ser representada por un vector de longitud fija cuyos componentes son características estadísticas de la serie de tiempo, capturando la estructura global. Estos vectores descriptivos, uno para cada componente de la serie de tiempo multivariante, son concatenados, antes de ser agrupados usando un algoritmo estándar rápido de agrupamiento, como agrupamiento k-means o jerárquico. Tal extracción de características estadísticas sirve como un procedimiento de reducción de dimensión para series de tiempo multivariadas. El método propuesto basado en estructuras de series de tiempo de una sola variable y métricas estadísticas suministra un novedoso, y aún simple y flexible forma de agrupar datos de series de tiempo multivariados eficientemente con precisión prometedora.

Toshniwal y Joshi (2005, p. 12) proponen una nueva metodología para agrupamiento de datos de series de tiempo basado en agrupamiento completo de secuencias. En este método, la extracción de características de los datos de serie de tiempo es hecha usando la tendencia acumulativa ponderada (*cumulative weighted slopes*). La tendencia acumulativa ponderada puede ser definida como la suma de tendencias ponderadas de la secuencia de tiempo computada sobre una base punto a punto. Los parámetros representan la tendencia acumulativa ponderada para varias secuencias de tiempo y son agrupadas dentro de grupos, y utilizan el método de agrupamiento *K-means* para identificar patrones similares.

Fujimaki, Hirose y Nakata (2008, p. 12) presentan una análisis teórico de agrupamiento de subdivisión de series de tiempo (Subsecuence Time series, STS) desde un punto de vista de análisis de frecuencias e identifica unos antecedentes matemáticos sobre los cuales el agrupamiento STS genera patrones de onda sinusoidal. También presenta una novedosa metodología de análisis teórico para descubrimiento de patrones desde datos de series de tiempo. Adicionalmente, proponen un algoritmo de agrupamiento usando una fase de preprocesamiento de ajustes para evitar patrones de onda sinusoidal y referirse a estos como agrupamiento fase de ajuste STS (Phase Alignment STS, PA-STS). El PA-STS es un algoritmo basado en análisis teórico, que permite obtener resultados de agrupamiento significativos.

Lytkin Kulikowski y Muchnik (2008, p. 33) proponen dos métodos de agrupamiento basados en la teoría del muestreo estadístico: generalización del criterio de Neyman para muestreo estratificado y generalización del método de selección de tipos representativos. Dichos algoritmos trabajan sobre datos de n dimensiones. Se realiza un estudio de los algoritmos planteados y el K-means aplicado a datos con series de tiempo diario. Los resultados experimentales obtenidos sobre series de tiempo de retorno diario del mundo real demostraron la credibilidad de las metodologías presentadas para la clasificación. La tabla 3 presenta el resumen de los algoritmos de agrupamiento basado en representaciones tratados en este estudio.

Discusión

A diferencia de lo observado por Liao (2005, p. 18), donde la mayoría de estudios de agrupamiento de series de tiempo estaban enfocados a series de tiempo de una sola variable, en los artículos revisados en el presente estudio se observa un crecimiento significativo en cuanto a la cantidad de propuestas que abordan las series de tiempo multivaluada,

llegando a corresponder aproximadamente al 65% de las técnicas revisadas.

Entre las metodologías estudiadas Debeljak et ál. (2010, p. 6), Wei y Jang (2010, pp. 135-138), Liao (2007, p. 13), se utilizan DTW para establecer las medidas de similitud, mientras que el 50% de las técnicas evaluadas utilizan la medida de distancia Euclidiana.

En los estudios realizados por Lytkin, Kulikowski, C.A., and Muchnik (2008, p. 33), Horenko (2010, p. 23), Pamminger y Frühwirth-Schnatter (2010, p. 24), Chiu, Hsu y Wang (2010), Otranto (2008, p. 14), desarrollan sus propias técnicas para realizar el agrupamiento.

Tabla 3. Resumen de algoritmos de agrupamiento de series de tiempo basado en representaciones

Artículo	Variable	Característica	Medida de distancia	Algoritmo de agrupamiento	Aplicación
Yang y Chen (2010, p. 10)	Simple	Picewise Local Statistics (PLS), Piecewise Discrete Wavelet Transfor (PDWT), Polynomial Curve fitting,(PCF) y Discrete Fourier Transformation (DFT)	Euclidiana	RPCL Network	16 series de tiempo de minería de datos de prueba
Maharaj y D'Urso [36]	Multiple	Dominio de frecuencia, coeficiente cepstral	Euclidiana	Fuzzy	Datos generados y serie de tiempo de 200 electroencefalogramas.
Lai, Chung y Tseng (2010, p. 8)	Múltiple	Representación de aproximación agregada simbólica (SAX)	Euclidiana	CAST	Datos sintéticos, mercado de acciones de Taiwan.
Chiu et ál. (2010)	Múltiple	Propagación de afinidad		Agrupamiento de consenso	Expresión génetica
Zhou et ál. (2009) [25]	Múltiple	Vectores	Mahalanobis	K-means	Series de tiempo financieras
Sawidesa et ál. (2008, p. 15)	Simple	Dominio de espectro, Coeficiente cepstral	Distancia cepstral	Diana	Series de tiempo biológicas
Otranto [43]	Múltiple Volatil	GRACH	Prueba de Wald y métricas autoregresivas	Algoritmo desarrollado	Índices del mercado italiano.
Hsu y Chen (2008)	Múltiple	Espacio	Euclidiana	SOM	Datos financieros Bolsa de Taiwan
Bandyopadhyay et ál. (2010)	Múltiple y sencilla	Pronosticabilidad, Interpolabilidad	Euclidiana	Fuzzy	Conjunto de datos artificiales.
Chandrakala y Shekar (2006)	Múltiple	Matriz de densidades o matriz distancia	Euclidiana	DBSCAN	Conjuntos de datos de carácter manuscrito.

Guo et ál. (2008)	Múltiples	Wavelet	Euclidiana	K-means	Datos sintéticos series de tiempo de control de flujo SCCTS.
Wang (2008)	Múltiples	Estructura global	Euclidiana	K-means o jerárquico	Secuencias de movimiento humano.
Toshniwal y Joshi (2005, p. 12)[47]	Múltiple	Tendencias ponderadas	Euclidiana	K-means	Datos sintéticos y datos de ventas al por menor de cadenas de almacenes de los Estados Unidos.
Fujimaki et ál. (2008, p. 12)	Simple	Patrones de onda senosidal	Euclidiana	K-means o Jerárquico	Datos sintéticos CBF

Solamente el artículo presentado por Bandyopadhyay, Baragona y Maulik (2010) utiliza técnicas de computación evolucionaria como algoritmos genéticos dentro del proceso de agrupamiento de series de tiempo. Entre los trabajos de metodologías basadas en modelos, Zhao y Deng (2010), Wie y Jiang (2010, pp. 135-138) y Pamminger y Frühwirth-Schnatter utilizan un método de agrupamiento basado en HMM. En el caso de metodologías basadas en representaciones Maharaj y D'ruso (2010, p. 25) y Savvides, Promponas y Fokianos (2008, p. 15) realizan el trabajo aplicado a dominio de frecuencias extrayendo el coeficiente cepstral.
Conclusiones (t1)

En este artículo, se han examinado algunos de los más recientes estudios sobre agrupamiento de series de tiempo, desde el año 2006, teniendo en cuenta que Liao (2005, p. 18) desarrolló un estudio de la misma naturaleza. Estos estudios están organizados en tres categorías principales si trabajan directamente sobre los datos originales, indirectamente con modelos construidos desde los datos en bruto, o indirectamente con extracción de características desde los datos en bruto. Las áreas de aplicación son resumidas con una breve descripción de los datos usados y algunas referencias de su obtención cuando son públicos. Las características y particularidades de algunos trabajos son discutidas.
Se aprecia que no se puede encontrar un algoritmo único que se ajuste a todas las series de tiempo que se puedan encontrar y que, en muchos casos, los algoritmos se desarrollan para un conjunto de datos en particular y considerando las características de estos.
Como trabajos futuros, es importante continuar la extensión de algunos algoritmos de agrupamiento de datos estáticos que presentan buen rendimiento y pueden ser una buena alternativa para lograr algoritmos de agrupamiento de series de tiempo que ofrezcan eficiencia y efectividad al determinar los grupos. Adicionalmente, es importante considerar el uso de técnicas de computación evolutiva para proponer nuevos algoritmos de agrupamiento de datos de series de tiempo en los cuales su complejidad computacional ofrezca mejores condiciones que los actualmente existentes.

Apéndice. Aplicaciones y datos utilizados

En algunos casos, cuando se desarrollan nuevos métodos no se utilizan datos específicos definidos previamente. Sin embargo, para probar dichos métodos y compararlos con los ya existentes, normalmente los investigadores generan datos simulados o se basan en

depósitos de datos de series de tiempo de acceso público, como la página UCR time series clasification/clustering [http://www.cs.ucr.edu/~eamonn/time_series_data/].

Para otros casos, las investigaciones están enfocadas en aspectos particulares y utilizan datos específicos del área de interés. Como se aprecia en el siguiente resumen, el agrupamiento de datos series de tiempo es necesario en aplicaciones ampliamente diferentes.

Negocios y socioeconómicos

Agrupamiento aplicado al análisis de estilos de administración de fondos de inversión común basado en series de tiempo de retorno diario (Lytkin, Kulikowski y Muchnik, 2008, p. 33). Se hicieron pruebas sobre datos sintéticos generados por una distribución Gaussiana y como datos reales las series de tiempo de retorno diario de dos fondos de inversión común (mayo del 2005 a mayo del 2006).
Agrupamiento de series de tiempo no lineales basado en densidades no paramétricas proyectadas (Vilar, Alonso y Vilar, 2010, p. 16). El conjunto de datos sobre el cual se aplicó el algoritmo consta de una colección de series de tiempo que representan el índice de producción industrial mensual mundial para veintiún países, desde enero de 1990 hasta noviembre del 2007. Todos los países considerados son miembros de la Organización para la Cooperación y Desarrollo Económico (Organization for Economic Cooperation and Development, OECD). (http://stats.oecd.org/index.aspx)

Agrupamiento basado en modelos de series de tiempo categóricas (Pamminger y Frühwirth-Schnatter, 2010, p. 24). El algoritmo se aplicó a los datos de movilidad salarial austriaco, los datos fueron tomados de las bases de datos de la Seguridad Social Austriaca (Austrian Social Security Data Base, ASSD). (http://ideas.repec.org/p/jkun nwps/2009_03.html)

Agrupamiento de series de tiempo mediante análisis de comunidad de redes (Piccardi y Calatroni, 2010). Esta técnica se aplicó sobre conjunto de datos de series financieras derivadas de los valores de bolsa diarios de las compañías incluidas en el índice promedio industrial Down Jones (*Down Jones Industrial Average*, DJIA).

Un método novedoso de agrupamiento de dos niveles para el análisis de datos de series de tiempo (Lai, C Chung y Tseng, 2010, p. 8). Para evaluar el algoritmo, datos del mundo real del mercado de acciones de Taiwan son probadas.

Agrupamiento basado en empotramiento localmente lineal (Locally Linear Embedding, LLE) (Zhou, Li y Ma, 2009). Las pruebas fueron realizadas sobre datos de la bolsa de valores de Shangai.

Agrupamiento de series de tiempo de variables aleatorias (*heteroskedatic*) mediante procedimientos basados en modelos (Otranto,2008, p. 14). El procedimiento fue aplicado al sector de índice del mercado italiano.
Agrupamiento de datos de series de tiempo mediante SOM para la estimación del índice de equilibrio óptimo (Hsu y Chen, 2008). Se realizan pruebas del algoritmo en datos financieros de la bolsa de Taiwan.

Agrupamiento difuso de series de tiempo univariable y multivariable, mediante la optimización genética multiobjetivo (Bandyopadhyay, Baragona y Maulik, 2010). Las pruebas del algoritmo se realizaron utilizando datos sintéticos y como datos reales, los índices mensuales de la producción industrial en Italia.

Agrupamiento de datos de series de tiempo usando inclinaciones acumulativas ponderadas (Toshniwal y Joshi, 2005, p. 12). Los datos reales utilizados para aplicar el algoritmo representan el conjunto de datos de ventas al por menor de cadenas de almacenes de los Estados Unidos.

Ingeniería

Agrupamiento de series de tiempo multivariado (vector) (Liao, 2007, p. 13). Se aplica en cuatro conjuntos de datos. El primero es un conjunto de datos sintéticos de series de tiempo de valor real generado por el procedimiento *varmasin* de SAS. El segundo conjunto de datos es el conformado por las señales de fuerza de tres componentes en un esmeril. El tercer conjunto de datos está conformado por señales basadas por multi sensores en un esmeril. Y por último, un conjunto de datos de series de tiempo multivariados que consta de una muestra del lenguaje de señas autraliano (*Australia Sign Language, Auslan*). (http://kdd.ics.uci.edu/summary.data.type.html)

Agrupamiento para las series de tiempo del flujo de tráfico (Yin, Zhou y Xie, 2006). Los datos utilizados fueron tomados de datos reales de tráfico del *sistema* de administración de tráfico en el estado de Washington en los Estados Unidos. http://www.wsdot.wa.gov/traffic/seattle/products/webflow.htm

Agrupamiento para subdivisiones de series de tiempo (subsequence time series, STS) mejorado (Chen, 2007). Para validar la metodología se utilizaron doce conjuntos de datos sintéticos y de la vida real tomados de: (http://www.cs.ucr.edu/~eamonn/time_series_data/)

Agrupamiento *K-means* de series de tiempo basado en el modelo oculto de Markov (Wei y Jiang, 2010, pp. 135-138). Para probar el algoritmo propuesto se utilizó un conjunto de datos artificiales generados desde dos modelos ocultos de Markov.

Agrupamiento de series de tiempo por medio de ensamble de redes RPCL con diferentes representaciones(Yang y Chen, 2010, p. 10). Los experimentos y los resultados de simulación de esta técnica se realizaron sobre una colección de conjuntos de datos de prueba estándar para minería de datos ya mencionado anteriormente. (http://www.cs.ucr.edu/~eamonn/time_series_data/)

Identificación automática basada en agrupamiento de modelos de inferencia difusos para series de tiempo (Montesino-Pouzols y Barriga-Barros, 2010, p. 13). Los análisis fueron realizados sobre cinco diversos conjuntos de datos: 1) conjunto de datos de muestras semanales de temperatura del fenómeno de oscilación sureño del niño; 2) serie del número de promedio de mancha solar mensual desde enero de 1749 hasta diciembre del 2009, suministrado por el Centro de Datos Geográfico Nacional de los Estados Unidos; 3) serie de tiempo que representa la demanda de electricidad diaria promedio normalizada en Polonia en la década de los noventa; 4) series de tiempo multidimensionales del promedio mensual de descriptores químicos diferentes de cierta área del mar Báltico; 5) series de tiempo univariable de la cantidad promedio diaria del tráfico en una red de datos.

Agrupamiento de series de tiempo basado en descomposición y reducción de ruido Wavelet (Guo et ál., 2008). Datos de cartas de control generados sintéticamente. (http://kdd.ics.uci.edu/databases/synthetic_control/synthetic_control.html)

Agrupamiento de subdivisión de series de tiempo desde un punto de vista de análisis

de frecuencia (Fujimaki, Hirose y Nakata, 2008, p. 12).

Ciencia

Agrupamiento en series de tiempo de expresión genética basado en representaciones continuas y medida de similaridad basada en energía (Zhang, Liu y Yan, 2010). Se utilizó un conjunto de datos de la proteína de control de división celular cdc15. (http://www.wikigenes.org/e/gene/e/2541975.html)

Agrupamiento de series de tiempo de expresión genética (Qu, Ng y Chen, 2010), plicado a diversos datos artificiales. Primero se generó un conjunto de datos sintéticos que incluían tres valores constantes bicluster y luego para medir la eficiencia computacional generaron un gran conjunto de datos de expresión genética con 10.000 genes y 100 puntos de tiempo.

Agrupamiento adaptativo para series de tiempo aplicado para identificar el ciclo celular expresado en genes (Douzal-Chouakria, Diallo y Giroud, 2009, p. 13). Utilizaron un conjunto de datos de acceso público denominado datos transcriptómicos de la línea celular del cáncer cervical humano HeLa. (http://genome-www.stanford.edu/Human-CellCycle/Hela/).

Agrupamiento jerárquico basado en HMM para analizar datos de series de tiempo de expresión genética (Zhao y Deng, 2010). Utilizan dos conjuntos de datos de series de tiempo de expresión genética ampliamente utilizados para validar agrupamientos: ciclo celular del hongo y respuesta del fibroblasto humano al suero. (http://genome-www.stanford.edu/cellcycle/data/rawdata/ y http://genome-www.stanford.edu/serum/).

Agrupamiento en consenso basado en propagación de afinidad (AP) (Chiu, Hsu y Wang, 2010). Para la prueba de estos datos, además de utilizar datos sintéticos, se utilizaron datos reales conformados por un conjunto de datos de expresión genética del hongo de galactosa y ciclo celular del honho. Análisis de datos de series de tiempo sobre vegetación de agroecología utilizando árboles predictivos de agrupamiento (Debeljak et ál., 2010, p. 6). El conjunto de datos utilizado consta de las parejas de series de tiempo del porcentaje cubierto de cultivos y maleza de 128 sitios experimentales del Reino Unido.

Agrupamiento de series de tiempo meteorológicas no estacionarias [18]. Se utilizan datos históricos de temperatura multidimensionales de Europa y un conjunto de datos de temperatura global.Agrupamiento de series de tiempo basada en pronóstico de densidades(Alonso et ál., 2006, p. 15). Los datos utilizados para probar el algoritmo fueron datos relacionados con la emisión de CO_2 en países industrializados.

Agrupamiento de series de tiempo biológicas mediante distancia basada en coeficiente cepstral (Savvides, Promponas y Fokianos, 2008, p. 15). Esta metodología fue aplicada para clasificar secuencias de aminoácidos.
Agrupamiento de series de tiempo multivariable basado en estructuras (Wang, 2008). Se aplicó esta técnica a secuencias de movimiento humano.

Medicina

Agrupamiento basado en interacción de series de tiempo multi variable (Plant, Wohlschläger y Zherdin, 2009). Para probar este método se generaron un conjunto de datos sintéticos con seiscientos objetos y trece dimensiones. Adicionalmente, se utilizaron

veintiséis imágenes funcionales de resonancia magnética.

Agrupamiento difuso de series de tiempo en el domino de frecuencia (Maharaj y D'urso, 2010, p. 25). Para aplicar esta metodología desarrollada se utilizó una serie de tiempo de 200 electroencefalogramas (EEG), la cual está dividida en dos conjuntos denotados por A y E, cada uno contiene 100 EEG de 23,6 s de duración (4096 observaciones): el conjunto A EEG tiene registros de voluntarios saludables mientras que el conjunto E registra pacientes de epilepsia durante la actividad de ataque epiléptico.

Arte y entretenimiento

Agrupamiento para videos basado en transformadas Wavelet de series de tiempo de trayectoria de movimiento de objetos en movimiento en video (Luo, Liao y Zhan, 2010). Los datos utilizados en el estudio son videos jugando baloncesto.

Un método basado en densidades para agrupamiento de series de tiempo en *kernel feature space* (Chandrakala y Sekhar, 2008). Se utilizaron dos conjuntos diferentes de datos caracteres manuscritos en línea: el conjunto de datos 1 contiene tres caracteres en escritura de lenguaje indio, telugu; el conjunto de datos 2 contiene los datos de tres caracteres en ingles.

Referencias

Alonso, A.M., et al. (2006). Time series clustering based on forecast densities. *ScienceDirect,* 15.

Bandyopadhyay, S., Baragona, R. y Maulik, M. (2010). Fuzzy clustering of univariate and multivariate time series by genetic multiobjective optimization. *Comisef Working Papers Series*. Computational Optimization Methods in Statistics, Econometrics and Finance.

Chandrakala, S. and Sekhar, C.C. (2008). A density based method for multivariate time series clustering in kernel feature space. *International Joint Conference on Neural Networks*, *6*.

Chen, J.R. (2007). Useful clustering outcomes from meaningful time series clustering. En Proc. 6th Australasian Data Mining Conference (AusDM'07). Gold Coast (Australia).

Chiu, T.Y., Hsu, T.C. y Wang, J.S. (2010). Ap-based consensus clustering for gene expression time series. En 20th IAPR International Conference on Pattern Recognition. Istanbul, Turkey, *IEEE, 2512-2515.*

Cowpertwait, P.S.P. y Metcalfe, A.V. (2009). *Introductory time series with r*. New York, NY: Springer.

Debeljak, M. et ál. (2010). Analysis of time series data on agroecosystem vegetation using predictive clustering trees. *Ecologicall Modeling*, Volume 222, Issue 14 6.

Ding, H. et ál. (2008) Querying and mining of time series data: Experimental comparison of representations and distance measures. En *Proceedings of the VLDB Endowment*. Auckland, New Zealanda: ACM.

Douzal-Chouakria, A., Diallo, A. y Giroud, F. (2009). Adaptive clustering for time series: Application for identifying cell cycle expressed genes. *Computational Statistics and Data Analysis, 53*, 13.

Fujimaki, R., Hirose, S. y Nakata, T. (2008). *Theoretical analysis of subsequence time-series clustering from a frequency-analysis viewpoint*. Society fro Industrial and Applied Mathematics, SIAM.

Guo, C., Jia, H. y Zhang, N. (2008). Time series clustering based on ica for stock data analysis. en Proceedings of the fourth international conference wire-

less communications, networking and mobile computing, WiCOM '08. *IEEE, 4*

Guo, H. et ál. (2008). An application on time series clustering based on wavelet decomposition and denoising. En Fourth International Conference on Natural Computation. Jinan, Shandong, China: IEEE, 4.

Horenko, I. (2010). On clustering of non-stationary meteorological time series. *Dynamics of Atmospheres and Oceans, 49* (2-3), 23.

Hsu, Y.C. y Chen, A.P. (2008). Clustering time series data by som for the optimal hedge ratio estimation. En Third 2008 International Conference on Convergence and Hybrid Information Technology. Daejeon, (Korea): IEEE, 6

Kavitha, V. y Punithavalli, M. (2010). Clustering time series data stream a literature survey. I*nternational Journal of Computer Science and Information Security, 8* (1), 6.

Keogh, E. y Kasetty, S. (2002). On the need for time series data mining benchmarks: Asurvey and empirical Knowl. *Data Discov, 6*, 102-111.

Kitagawa, G. (2010). *Introduction to time series modeling. Monographs on satatistics and applied probability*. Boca Raton, (FL): Chapman & Hall/CRC.

Kuenzel, L. (2010). *Gene clustering methods for time series microarray data*.

Lai, C.P., Chung, P.C. y Tseng, V.S. (2010). A novel two-level clustering method for time series data analysis. *Expert Systems with Applications, 37*, 8.

Liao, T.W. (2007). A clustering procedure for exploratory mining of vector time series. *Pattern Recognition, 40*, 13.

Liao, T.W. (2005). Clustering of time series data a survey. *The Journal of the pattern Recognition Society, 38*, 18.

Luo, Y., Liao, M. y Zhan, Z. A similarity analysis and clustering algorithm for video based on moving trajectory time series wavelet transform of moving object in video. En 2nd International Conference on Image Analysis and Signals Processing, IASP 2010. XiaMen, (China): IEEE, 5

Lytkin, N.I., Kulikowski, C.A. y Muchnik, I.B. (2008). *Variance-based criteria for clustering and their application to the analysis of management styles of mutual funds based on time series of daily returns*. New Jersey (USA): New Brunswick.

Maharaj, E.A. y D'urso, P. (2010). Fuzzy clustering of time series in the frequency domain. *Information Sciences,* 25.

Montesino-Pouzols, F. y Barriga-Barros, A. (2010). Automatic clustering-based identification of autoregressive fuzzy inference models for time series. *Neurocomputing, 73*, 13.

Olier, I. y Vellido, A. (2008). Advances in clustering and visualization of time series using gtm through time. *Neural Networks, 21*, 10.

Otranto, E. (2008). Clustering heteroskedastic time series by model-based procedures. *Computational Statistics & Data Analysis, 52*, 14.

Palit, A.K. y Popovic, D. (2005). Computational intelligence in time series forecasting: Theory and engineering applications. En M.J. Grimble y M.A. Johnson (ed.). *Advances in industrial control* (381). Glasgw (Scotland, UK): Springer.

Pamminger, C. y Frühwirth-Schnatter, S. (2010). Model-based clustering of categorical time series. Bayesian Analysis articulo.

Papanastassiou, D. Classification and clustering of garch time series. En XIII International Conference Applied Stochastic Models and Data Analysis ASMDA 2009. 2009. Vilnius, Lithuania. 5

Piccardi, C. y Calatroni, L. Clustering time series by network community analysis.

En *COMPENG 2010 Complexity in Engineering*. Roma (Italy): IEEE, 94-96.

Plant, C., Wohlschläger, A.M. y Zherdin, A. (2009). Interaction-based clustering of multivariate time series. En Ninth International Conference on Data Mining. Venice, (Italy): IEEE, 914-919.

Pylvänen, M., Äyrämö, S. y Kärkkäinen, T. (2009). *Visualizing time series state changes with prototype based clustering*.

Qu, J., Ng, M. y Chen, L. (2010). Constrained subspace clustering for time series gene expresion data. En The Fourth International Conference on Computational Systems Biology (ISB2010). Suzhou, (China): ORSC & APORC, 323-330.

Savvides, A., Promponas, V.J. y Fokianos, K. (2008) Clustering of biological time series by cepstral coefficients based distances. *Pattern Recognition, 41*, 15.

Toshniwal, D. y Joshi, R.C. (2005). Using cumulative weighted slopes for clustering time series data. *GESTS Int'l Trans. Computer Science and Engr, 20*, 12.

Tsiporkova, E. y Boeva, V. (2008). A novel gene-centric clustering algorithm for standardization of time series expression data. En 4th International IEEE Conference "Intelligent Systems". Varna, (Bulgaria): IEEE.

Vilar, J.A., Alonso, A.M. y Vilar, J.M. (2010). Non-linear time series clustering based on non-parametric forecast densities. *Computational Statistics & Data Analysis,* 2010, *54*, 16.

Vilar, J.A., Alonso, A.M. y Vilar, J.M. (2010). Non-linear time series clustering based on non-parametric forecast densities. *Computational Statistics & Data Analysis, 54*, 2850-2865.

Wang, X. (2008). Structure-based multivariate time series clustering. En *Computer Science Colloquium*. Hong Kong, China.

Wei, L.L. y Jiang, J.Q. (2010). A hidden markov model-based k-means time series clustering algorithm. In International Conference on Information Systems (ICIS) 2010 (135-138). Saint Louis, Missouri, (USA): IEEE.

Wei, W.W.S. (2006). *Time Series Analysis:Univariate and Multivariate Methods* (2ª ed.). New York (NY, USA): Pearson Education, Inc.

Yang, Y. y Chen, K. (2010). Time series clustering via RPCL network ensemble with different representations. *IEEE Transactions on Systems, Man, and Cybernetics-part C: Applications and Reviews*, 10.

Yin, J., Zhou, D. y Xie, Q.Q. (2006). A clustering algorithm for time series data. En Proceedings of the Seventh International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'06). IEEE, 4.

Zhang, W.F., Liu, C.C. y Yan, H. (2010). Gene time series data clustering based on continuos representations and an energy based similarity measure. En Proceedings of the Ninth International Conference on Machine Learning and Cybernetics. Qingdao, Shandong (China): IEEE. 2079-2083.

Zhao, G. y Deng, W. (2010). An hmm-based hierarchical clustering method for gene expression time series data. En The Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2010). Liverpool (United Kingdom): IEEE, 219-222.

Zhou, D., Li, J. y Ma, W. (2009). Clustering based on lle for financial multivariate time series. En International Conference on Management and Service Science (MASS 2009). Wuhan/Beijing (China): IEEE, 4.