Clasificador no supervisado para series de tiempo

E.A. Santos-Camacho, J.G. Figueroa-Nazuno, J.C. Chimal Eguía

Instituto Politécnico Nacional, Centro de Investigación en Computación (CIC), México

araceli.libelula@gmail.com, {jfn,chimal}@cic.ipn.mx

Resumen. Una serie de tiempo es una secuencia de datos numéricos que describen fenómenos naturales o artificiales, por lo que su análisis es de gran interés en diversas áreas del conocimiento, siendo de gran utilidad en el desarrollo industrial, social y científico. Una alternativa para su análisis se presenta en la clasificación, que permite tener agrupaciones con características similares, sin embargo, existe una gran cantidad de fenómenos en la vida real que no cuentan con una clasificación previa o establecida, por lo que los clasificadores supervisados no pueden ser aplicados en este tipo de problemas, y es por ello que surge la necesidad de utilizar un enfoque de clasificación no supervisado. En este trabajo se propone una metodología, que permita obtener la clasificación no supervisada de un conjunto de series de tiempo utilizando un enfoque no supervisado, la cual consta de dos etapas: en la primera, se diseñaron doce técnicas diferentes que contemplan la búsqueda de series representativas y en la segunda, se propone un algoritmo de ensamble para obtener la agrupación final, mezclando las agrupaciones obtenidas en la etapa anterior.

Palabras clave: clasificación no supervisada, serie de tiempo, ensamble de algoritmos.

Unsupervised Classifier for Time Series

Abstract. A time series is a sequence of numerical data describing natural and/or artificial phenomena, therefore its analysis is of great interest in various areas of knowledge, being very useful in industrial, social and scientific development. An alternative for the analysis of time series is presented in classification techniques that allow the creation of groups of similar characteristics, however, there is a lot of phenomena in real life that do not have a prior or established classification, therefore supervised classifiers cannot be applied in such problems. This is why unsupervised classification approach arises, which allow us creating groups from a given set of data without any prior knowledge. In this work a methodology is proposed that allows the unsupervised classification of a set of time series using an unsupervised approach. This method consists of two stages: first, it was designed twelve different search techniques to find representative time series and second, an

assembly algorithm is proposed for the final grouping, mixing clusters obtained in the previous stage.

Keywords: unsupervised classification, time series, assembly algorithms.

1. Introducción

En la actualidad son innumerables las aplicaciones que se pueden citar en diferentes áreas de investigación y la industria, donde los datos están representados en forma de series de tiempo. En los últimos años se ha visto una dramática explosión en la cantidad de series de tiempo, por lo que su análisis desempeña un papel muy importante, ya que con él es posible entender los fenómenos que describen.

Se denomina "serie de tiempo" a un conjunto de datos de un cierto fenómeno o ecuación, registrados secuencialmente.

Una alternativa que permite conocer el comportamiento y dinámica de un conjunto de series de tiempo, se ha presentado en el problema de clasificación, sin embargo, es preciso mencionar que la mayoría de los fenómenos encontrados en la vida real, no tienen una clasificación como tal y es por eso que la clasificación no supervisada, ha traído gran interés.

La clasificación es organizar y categorizar los objetos en clases o grupos diferentes no etiquetados, los cuales deben ser coherentes u homogéneos [1, 2].

Es importante comentar que uno de los retos en la clasificación de series de tiempo, está dada por su estructura, generalmente al clasificar fenómenos descritos por Atributos, el orden de ellos no afecta el resultado, sin embargo, las series de tiempo conservan un orden o temporalidad, lo que no permite cambiar la posición de los datos, por lo que algoritmos que trabajan con atributos no pueden ser aplicados en este tipo de problemas.

Por lo anterior, en este trabajo se presenta una metodología de clasificación no supervisada y de forma libre para series de tiempo, basada en el desarrollo de distintos algoritmos que son ensamblados para obtener la agrupación final, con el propósito de que este proceso pueda revelar objetos/categorías desconocidas que ayuden a un mejor entendimiento de los datos, resaltando la estructura inherente al agrupar un conjunto de serie de tiempo.

2. Serie de tiempo

Una serie de tiempo, es un conjunto de datos numéricos, obtenidos a partir de una observación experimental o mediante el cálculo numérico de ecuaciones, es decir, una serie de tiempo es un conjunto de la forma $ST = \{x_1, x_2, \ldots\ldots, x_t, \ldots\ldots x_N\}$.

De lo anterior algunas características sobresalientes deben ser consideradas en el análisis de las series de tiempo, las cuales se describen a continuación:

- Dimensionalidad: son los grados de libertad en las series de tiempo, es decir, si se trabaja en el espacio binario, en los reales u otro.
- Tamaño: es la cantidad de datos que conforman una serie de tiempo.
- Representación: una serie de tiempo puede ser representada en un plano cartesiano, donde el eje 'y' representa el valor (o magnitud) y el eje 'x' es un índice consecutivo que corresponde a cada valor (el cual puede ser tiempo u otra variable) en la serie de tiempo, por lo tanto, las series de tiempo se encuentran en 1½ dimensión.
- Estructura: las series de tiempo contienen picos, los cuales no son derivables ni integrables.

3. Comparación entre series de tiempo

La comparación de series de tiempo se centra en la búsqueda de similitud de patrones semejantes, por lo que para realizar el análisis de semejanza entre dos series de tiempo es necesario que éstas sean cuantificables, es decir, asignarles un valor numérico. De acuerdo a Krantz [3, 4, 5] "medir es el acto o proceso de asignar un número a un fenómeno, con base en alguna regla".

En [6] se presentan las siguientes definiciones:

- "La similitud es una cantidad que refleja la fuerza o intensidad de relación, entre dos objetos".
- "La distancia es la medida de disimilitud entre dos objetos y se refiere a la discrepancia entre dos objetos, basada en varias características analizadas. También puede interpretarse como una medida de desorden, entre dos objetos".

En este trabajo se emplean técnicas de distancia para dar un valor de semejanza entre un par de series de tiempo, por lo que la relación con la similitud queda dada por: "Poca distancia equivale a Poca diferencia, lo que equivale a Gran similitud".

4. Ensamble de algoritmos de agrupación

Por el teorema de *No Free Lunch* [7], que nos dice que si un algoritmo funciona bien para un problema dado, este no va a tener los mismos resultados para otro problema, en este trabajo se presentan diferentes técnicas de agrupación de series de tiempo y para obtener la agrupación final, se aplica un algoritmo de ensamble.

El ensamble de algoritmos de agrupación [8], es generado por un conjunto de algoritmos de agrupación denominado "agrupación base", el cual combina las salidas de los algoritmos de la "agrupación base" tal que la información útil codificada en

cada algoritmo de agrupación, es aprovechada al máximo para obtener el agrupamiento final.

Comúnmente los métodos de ensamble son aplicados principalmente porque son capaces de impulsar algoritmos débiles y mejorar la estimación aleatoria.

En términos generales obtener una agrupación es relativamente fácil, ya que cualquier algoritmo de partición genera una agrupación, mientras que la mayor dificultad radica en la combinación de los algoritmos, por lo que para obtener éxito en el ensamble de algoritmos, la clave está en cómo se expresa la información dada por los "agrupación base" y como está es ensamblada.

5. Metodología

La metodología propuesta para la agrupación de series de tiempo considerando un enfoque no supervisado y de forma libre, se describe de la siguiente manera:

1. Se selecciona un conjunto de series de tiempo.
2. Se selecciona una medida de distancia.
3. Se aplican las técnicas de agrupación propuestas (sección 6.3).
4. Se aplica el algoritmo de ensamble modificado (sección 6.4), a las agrupaciones obtenidas en el paso 3.
5. Se evalúa la agrupación final.

5.1. Conjunto de datos

En este trabajo se tomaron 3 conjuntos de datos en los que se contemplan datos Sintéticos, Aleatorios y un conjunto de datos con que contempla la transformación de objetos 3D a 1D, donde 1D corresponde a una serie de tiempo.

5.1.1. Conjunto de datos sintético

Para ejemplificar los problemas más comunes de las series de tiempo se utiliza un conjunto de datos sintético con 21 series de tiempo, el cual contempla 3 estructuras diferentes (tres grupos), que son: Cuadros, Senos y Cuadros; a los cuales se les hicieron diferentes modificaciones para representar los problemas de escala, desfase, ruido y combinación de ellas.

5.1.2. Conjunto de datos aleatorios

La distribución normal o distribución de Gauss es sin duda la más importante y la demás aplicación en todas las distribuciones continuas, ya que es bastante adecuada para describir la distribución de muchos conjuntos de datos que ocurren en la naturaleza, la industria y la navegación, entre otros. Por esta razón se generó un conjunto de datos conformado por 180 series de tiempo, con distribución normal de

forma aleatoria; con el objetivo es contar con un conjunto de datos controlado, se generaron 9 grupos.

5.2. Conjunto de datos objetos 3D

En Computación la representación de datos espaciales de una figura 3D está dada por la definición de malla de polígonos, la cual es muy popular para modelos tridimensionales debido a su simplicidad. Se tomó un conjunto de datos correspondiente al trabajo realizado en [9], donde realizan la transformación de objetos 3D a 1D bajo la siguiente idea "Se coloca el objeto 3D en un cubo y dado un orden predeterminado, se registra la distancia de forma consecutiva".

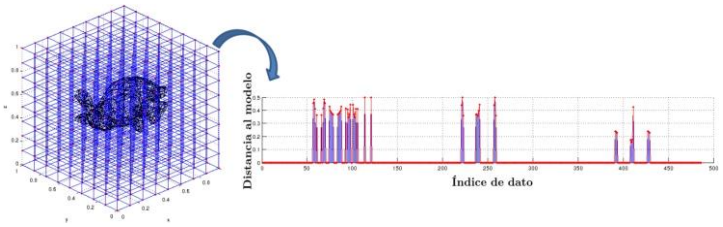

Fig. 2. Conjunto de datos de objetos 3D, extraída de [9].

El conjunto de datos está conformado por 5 clases que incluye a: delfines, perro, rostros, copas y pistolas, con un total de 40 series de tiempo, donde cada serie de tiempo tiene 1014 datos.

5.3. Medidas de distancia

Las medidas de distancia utilizadas e implementadas en este trabajo, contemplan la comparación directa utilizando la distancia de *Minkowski* (ecuación 1) e indirecta *Fast Dynamic Time Warping* o Fréchet evolucionado [10].

$$d = \sqrt[\lambda]{\sum_{k=1}^{n} |x_k - y_k|^{\lambda}}, \tag{1}$$

donde x, y corresponden a la serie 1 y 2 respectivamente, k es el valor contenido en esa posición en la serie de tiempo, n es la cardinalidad de la serie de tiempo y λ es el orden de distancia a calcular, si es 1 corresponde a distancia *City Block*, 2 es *Ecuclidiana* y >= 3 *Minkowski*.

5.4. Técnicas de agrupación

Los algoritmos de agrupación propuestos, están basados en la búsqueda de series de tiempo representativas dado un conjunto de datos, a través de la relación que existe entre la serie de tiempo representativa y una medida de distancia hacia todas las

demás. Para formar cada una de los grupos encontradas en el conjunto de datos, se emplea un criterio de corte estadístico [11].

5.5. Agrupación mediante serie representativa

Para la agrupación de series de tiempo, en este trabajo nos centramos en la búsqueda de series representativas, para lo cual se consideran los siguientes cuatro criterios:

- Aleatorio: se escoge de forma aleatoria una serie de tiempo.
- Mínima distancia: se obtiene la distancia de cada serie de tiempo contra todas las demás y aquella que tenga la mínima distancia con alguna otra.
- Máxima distancia: se obtiene la distancia de cada serie de tiempo contra todas las demás y aquella que tenga la máxima distancia con alguna otra.
- Centroide: por cada serie de tiempo se obtiene la distancia a todas las demás y se promedian, se toma como serie representativa a la serie con el menor promedio.

Por cada criterio propuesto se presentan a continuación tres métodos para la agrupación de series de tiempo, con un enfoque no supervisado y de forma libre:

- **Método de agrupación 1**

Este método considera la agrupación de un conjunto de datos dado, empleando solo una serie de tiempo denominada serie representativa.

1. Obtención de serie representativa: se toma la serie representativa dependiendo del criterio deseado (aleatorio, mínima distancia, máxima distancia o centroide), del conjunto de datos y su distancia con respecto a todas las demás.
2. Criterio de corte: para generar las agrupaciones se ordena de menor a mayor las distancias de la serie representativa contra las demás y se aplica el criterio de corte empleando el umbral establecido en 2.

- **Método de agrupación 2**

Este método modifica el criterio de corte, ya que se podrían tener agrupaciones en los extremos (un grupo o tantos grupos como series de tiempo existentes en el conjunto de datos), dependiendo de la distribución de los datos. Dado el método 1 se modifica el paso 2 y se agrega el paso 3.

2. Criterio de corte: para generar las agrupaciones se ordena de menor a mayor las distancias de la serie representativa contra las demás y se aplica el criterio de corte empleando el umbral 0.5 (en cada iteración se aumentara el umbral en 0.5, hasta llegar a 3.5).
3. Evaluación: dada la agrupación con un umbral de corte, se evalúa mediante el índice I (la agrupación final será la que tenga el Índice I mayor).

- **Método de agrupación 3**

Este método examina la posibilidad de que exista, más de una serie representativa. Dado el método 2, se modifica el paso 2 y 3, y se agrega el paso 4.

2. Criterio de corte: para generar las agrupaciones se ordena de menor a mayor las distancias de la serie representativa contra las demás y se aplica el criterio de corte empleando el umbral en 2.
3. Agrupación final: del paso 2 se obtiene el primer grupo formado y se agrega a la agrupación final, después estas series se eliminan del conjunto de datos.
4. Posteriormente se repite el paso 1, 2 y 3, hasta que no existen series de tiempo en el conjunto de datos.

5.6. Método de ensamble

Dado que en este trabajo se proponen varios algoritmos para la agrupación de series de tiempo, se presenta el inconveniente de tener diversas agrupaciones, por lo que se hace necesario implementar un método que permita obtener una sola agrupación. En este trabajo se propone la modificación del método de ensamble por re-etiquetado [12], las modificaciones realizadas son:

- No se re-etiquetan los grupos.
- No se fija la cardinalidad de grupos finales como lo hace el método, basado en votación.
- Se examinan grupos con cardinalidad >3.
- Se considera un grupo en la agrupación, siempre y cuando este aparezca en la solución en más de 6 de las técnicas propuestas.

6. Análisis experimental

Se aplicaron las 12 técnicas de agrupación propuestas y posteriormente el algoritmo de ensamble en los tres conjuntos de datos, los resultados obtenidos se presentan en la tabla 1. Así mismo se examinan los conjuntos de datos utilizando el algoritmo k-medoids, con el objetivo de tener un referente de comparación de resultados.
En los resultados obtenidos se puede observar que el método de agrupación propuesto obtuvo mejor precisión que el algoritmo más *k-medoids*, independientemente de la técnica de distancia empleada.

Para ejemplificar una de las agrupaciones obtenidas en la figura 3, se presenta la agrupación del conjunto de datos de objetos 3D, en donde se presenta al objeto 3D que representa cada serie de tiempo, donde se puede apreciar en los resultados, que los objetos se agrupan correctamente, teniendo solo seis objetos mal agrupados; si bien no se logran obtener los 5 grupos existentes en el conjunto de datos como lo haría el algoritmo *k-medoids*, esto no afecta a la propuesta, debido a que los grupos formados son consistentes, es decir, los grupos contienen elementos que pertenecen a

la misma clase. Cabe mencionar que la agrupación empleada fue utilizando la transformación de estos objetos 3D a una dimensión.

Tabla 1. Precisión obtenida utilizando el ensamble de algoritmos.

Técnicas de distancia	Sintético	Aleatoria	Imágenes 3D
City Block	80.95	100	82.5
Euclidiana	76.19	100	85
Minkowski	80.95	100	57.7
FDTW	95.23	100	77.5
K-medoids	47.61	80	57.5

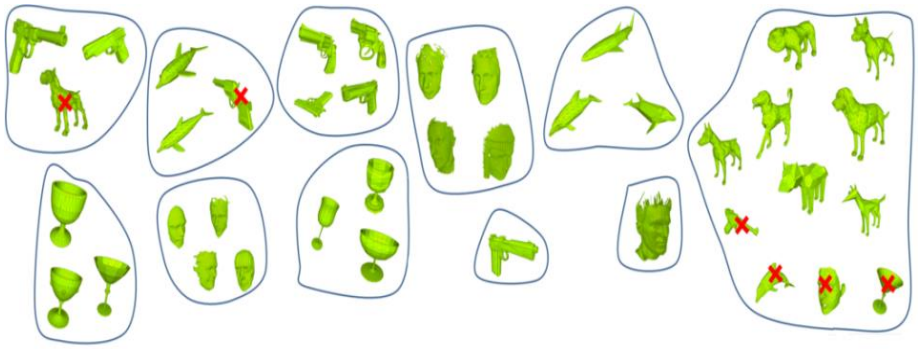

Fig. 1. Agrupación del conjunto de datos 3D empleando el método de ensamble y técnicas de agrupación propuestas y la medida de distancia *City Block*.

7. Conclusiones

El problema de clasificación no supervisada de series de tiempo, consiste en organizar las series de tiempo que son similares y distinguir entre aquellas que no lo son. Considerando el teorema de *No Free Lunch*, en este trabajo se presenta la agrupación de series de tiempo bajo el enfoque no supervisado y de forma libre, usando doce técnicas que involucran 4 criterios diferentes y finalmente para obtener una agrupación final, se combinan las salidas de las técnicas propuestas mediante la modificación del método de ensamble por re-etiquetado. Los resultados obtenidos por el ensamble de algoritmos supera en todos los casos probados al algoritmo *k-medoids*, lo que indica el potencial del método propuesto; además una de las aplicaciones de este método sobresaliente es la agrupación de objetos 3D con transformación a 1D, en donde además de tener semejanza en 1D en los grupos formados, al identificar el objeto 3D correspondiente, estos pertenecen a la misma categoría.

Referencias

1. Yang, Y., Chen, K.: Time series clustering via RPCL network ensemble with different representations. IEEE Transactions on Systems, Man, and Cyberneticspart C. Applications and Reviews (2010)
2. Vilar, J.A., Alonso, A.M., Vilar, J.M.: Non-linear time series clustering based on non-parametric forecast densities. Computational Statistics & Data Analysis (2010)
3. Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A.: Foundations of measurement. Vol. 1. Additive and polynomial representations. New York: Academic (1971)
4. Suppes, P., Krantz, D.H., Luce, R.D., Tversky, A.: Foundations of measurement. Vol. 2. Geometrical, threshold, and probabilistic representations. San Diego, CA: Academic, (1989)
5. Suppes, P., Krantz, D.H., Luce, R.D., Tversky, A.: Foundations of measurement. Vol. 3. Representation, axiomatization, and invariance. San Diego, CA: Academic (1990)
6. Teknomo, K.: Similarity measurement. [http://people.revoledu.com/kardi/tutorial/Similarity/index.html] (Agosto, 2015)
7. Wolpert, D.H., Macready, W.G.: No Free Lunch Theorems for Optimization. IEEE Trans. Evolutionary Computation, Vol. 1, No. 1, pp. 67–82 (April 1997)
8. Zhou, H.: Ensemble Methods Foundations and Algorithms. Machine Learning & Pattern Recognition Series, Chapman & Hall/CRC, Microsoft Research Ltd. Cambridge, UK, pp. 135–156 (2012)
9. Valle Chávez, Figueroa Nazuno, J.: Recuperación y Comparación de Figuras en 3D. Tesis de la Maestría en Ciencias de la Computación del Centro de Investigación en Computación del Instituto Politécnico Nacional (2012)
10. Santos Camacho, E.A.: Clasificación de series de tiempo mediante una comparación elástica. Tesis de licenciatura, Centro Universitario UAEM Valle de México (2012)
11. Renteria Agualimpia, W., Figueroa Nazuno, J.: Análisis e Identificación de Cambios Abruptos en Arreglos Unidimensionales. IV Congreso Internacional en Tecnologías Inteligentes y de la Información CITII, pp. 1–11 (2008)