Asociación Argentina
de Mecánica Computacional

Mecánica Computacional Vol XXXIV, págs. 1195-1202 (artículo completo)
Sebastián Giusti, Martín Pucheta y Mario Storti (Eds.)
Córdoba, 8-11 Noviembre 2016

SEGMENTACIÓN DE SECUENCIAS DE RANGO CONTINUO MEDIANTE INFORMACIÓN MUTUA

Miguel A. Ré[a,b] **y Guillermo G. Aguirre Varela**[b]

[a]*CIII, Facultad Regional Córdoba, Universidad Tecnológica Nacional, Maestro López y Cruz Roja Argentina, Ciudad Universitaria, 1016 Córdoba, Argentina, mgl.re33@gmail.com,*

[b]*Facultad de Matemática Astronomía, Física y Computación, Universidad Nacional de Córdoba, Haya de la Torre y Medina Allende, Ciudad Universitaria, 1016 Córdoba, Argentina, re@famaf.unc.edu.ar, guiava@gmail.com*

Palabras Clave: Divergencia Jensen-Shannon, distancias entrópicas, segmentación.

Resumen. La detección de bordes de dominio en secuencias de rango continuo encuentra aplicaciones en la detección del comienzo de una contracción muscular en electromiografía o del comienzo y propagación de una crisis epiléptica en el análisis de electroencefalogramas. La divergencia de Jensen Shannon (DJS), una versión simetrizada de la divergencia de Kullback Leibler, permite cuantificar la diferencia entre distribuciones de probabilidad. Esta propiedad ha sido ampliamente utilizada en el análisis de secuencias simbólicas o secuencias de rango discreto como cadenas de ADN. A pesar de estar bien definida para distribuciones continuas, la DJS no ha sido tan extensamente utilizada para la segmentación de secuencias de rango continuo. A partir de la identificación de la DJS con la Información Mutua entre una distribución continua y una discreta proponemos aquí un método para la segmentación de secuencias de rango continuo y evaluamos su desempeño a partir de secuencias generadas artificialmente.

1. INTRODUCCIÓN

El problema de la discriminación (o diferenciación) entre distribuciones de probabilidad tiene un papel relevante en la teoría de inferencia estadística. Ya en los trabajos de Fisher (1928) surge como relevante la noción de distancia entre distribuciones de probabilidad para un planteo formal de la inferencia estadística. La cuantificación de la distancia entre distribuciones de probabilidad presenta interés en el estudio de problemas de Física Estadística como la estabilidad de funcionales termodinámicos (Lesche, 1982), cuantificación del concepto de complejidad (López Ruiz et al., 1995), o distinción entre procesos caóticos y estocásticos (Rosso et al., 1995). También encontramos aplicación del concepto de distancia (Deza y Deza , 2006) en problemas de segmentación de secuencias como en la detección de segmentos estacionarios en cadenas de ADN (Grosse et al., 2002) o la detección de un cambio de régimen en secuencias de rango continuo como en la detección del comienzo de una contracción muscular en electromiografía (López et al., 2011) o la detección temprana del comienzo de una crisis epiléptica en el registro de un electroencefalograma (Pereyra et al., 2007).
Si bien existen diversas definiciones de distancia (Deza y Deza , 2006) entre distribuciones de probabilidad muchas de ellas presentan dificultades tanto formales como de aplicabilidad. Una medida de distancia que ha mostrado resultados satisfactorios es la denominada divergencia de Jensen-Shannon (DJS) (Lin , 1991), una versión simetrizada de la divergencia de Kullback-Leibler, de gran utilidad en diversos contextos. También puede verse como el contenido de información mutua (IM) entre una variable continua y otra discreta, definida como la divergencia de Kullback entre la distribución conjunta y el producto de las marginales de las mencionadas variables. La DJS entre las distribuciones P_1 y P_2 se expresa en términos de la entropía de Gibbs-Shannon como

$$D\left[P_1, P_2\right] = H\left[\pi_1 P_1 + \pi_2 P_2\right] - \pi_1 H\left[P_1\right] - \pi_2 H\left[P_2\right] \tag{1}$$

con π_1 y π_2 factores de peso que cumplen con la condición $\pi_1 + \pi_2 = 1$ y la entropía

$$H\left[P\right] = -\sum_j p_j \ln\left(p_j\right)$$

para distribuciones discretas o la versión continua

$$H\left[\phi\right] = -\int_{-\infty}^{\infty} dy\ \phi\left(y\right) \ln\left(\phi\left(y\right)\right)$$

La DJS ha sido utilizada ampliamente como herramienta para la detección de puntos de segmentación en secuencias simbólicas (o de rango discreto) incluyendo distintas generalizaciones (Lamberti y Majtey , 2003; Ré y Azad , 2014) para la definición de entropía.
El método consiste en proponer una partición de la secuencia y calcular la DJS entre la distribuciones de probabilidad de cada subsecuencia. Variando el punto de partición se determina el punto de segmentación para la partición que arroja el máximo valor de DJS (Grosse et al., 2002). Si bien la DJS está bien definida a partir de las distribuciones de probabilidad de cada subsecuencia, estas distribuciones no son conocidas y sólo pueden ser estimadas a partir de la secuencia misma. Para el caso de secuencias discretas las distribuciones de probabilidad p_j pueden estimarse a partir de las frecuencias de aparición relativa de cada símbolo en la secuencia $f_j^{(s)}$. Para secuencias continuas la estimación de las correspondientes densidades de probabilidad $\phi\left(y\right)$ no resulta tan simple o directa.

La forma pesada de la DJS en (1) puede identificarse con la IM entre una variable discreta (asociada a los pesos asignados a las distribuciones) y una variable continua (identificada con los valores registrados en la secuencia).

2. MÉTODO

En esta sección presentamos el uso de la DJS para la segmentación de una secuencia de números reales heterogénea, *i.e.* con propiedades estadísticas distintas en distintos segmentos. Consideremos una secuencia $\mathcal{S}$ de n números reales conformada por dos (o más) subsecuencias homogéneas con n_i elementos cada una ($\sum_i n_i = n$). Las propiedades estadísticas de cada subsecuencia pueden describirse por la densidad de probabilidad $\mu_i(x)$. Suponiendo que la única información disponible es la secuencia misma, el problema de la segmentación de la secuencia puede enunciarse como la determinación de los valores de n_i. Ilustramos esquemáticamente en la figura 1 el problema de segmentación para una secuencia constituída por dos subsecuencias.

$$\underbrace{v_1 v_2 \quad \dots \quad v_{n_1-1} v_{n_1}}_{n_1 \text{ valores}} \underbrace{v_{n_1+1} \quad \dots \quad v_{n_1+n_2}}_{n_2 \text{ valores}}$$

Figura 1: Esquema del problema de detección de segmentación: la secuencia $\mathcal{S}$ está conformada por n_1 números reales con densidad de probabilidad μ_1 seguidos de n_2 números reales con densidad de probabilidad μ_2 (ambos de rango continuo). El problema a resolver es encontrar el valor de n_1 sin el conocimiento previo de las densidades de probabilidad.

Para determinar el punto de segmentación se divide la secuencia en dos subsecuencias de longitud ν_1 y ν_2 respectivamente ($\nu_1 + \nu_2 = n$) y se calcula la DJS. Se repite la operación variando el valor de ν_1 y se toma como punto de segmentación el valor de ν_1 que hace máximo el valor de la DJS.

Dado que las probabilidades correspondientes a cada segmento de la secuencia son desconocidas, éstas deben estimarse para poder efectuar el cálculo de la DJS. Para el caso de variables de rango discreto el método de aproximación es relativamente directo: se aproximan las probabilidades por las frecuencias relativas de aparición de cada valor. Para secuencias de rango continuo la aproximación de las densidades de probabilidad presenta una dificultad mayor. Un método posible es el de *binning* pero requiere de un gran número de valores para tener una aproximación razonable.

Presentamos aquí un método alternativo de cálculo adaptando uno desarrollado para el cálculo de la IM entre una variable discreta y una continua. El cálculo se basa en la aproximación del *kernel* de densidad (Silverman B. W. , 2006) para las densidades de probabilidad continua como se describe a continuación.

Reescribimos la DJS como

$$D = \frac{\nu_1}{n} \int_{-\infty}^{\infty} dy\, \mu_1(y) \ln[\mu_1(y)] + \frac{\nu_2}{n} \int_{-\infty}^{\infty} dy\, \mu_2(y) \ln[\mu_2(y)] - \int_{-\infty}^{\infty} dy\, \phi(y) \ln[\phi(y)] \tag{2}$$

donde los factores de peso se han elegido como la fracción de elementos en cada subsecuencia: $\pi_i = \nu_i/n$. En la expresión anterior

$$\phi(y) = \frac{\nu_1}{n}\mu_1(y) + \frac{\nu_2}{n}\mu_2(y) \tag{3}$$

corresponde a la probabilidad marginal de la variable continua marginada sobre la subsecuencia de dependencia. Como fuera mencionado, el cálculo de la DJS requiere de la aproximación de las densidades de probabilidad. Usamos aquí la aproximción del *kernel* de densidad: si para una variable aleatoria Y con densidad de probabilidad $f\left(y\right)$ tenemos una realización de n_i repeticiones podemos aproximar la densidad de probabilidad por

$$\hat{f}\left(y\right) = \frac{1}{n_i h} \sum_{i=1}^{n_i} K\left(\frac{y - y_i}{h}\right) \tag{4}$$

con y_i los valores registrados en la realización. En la aproximación, h es un parámetro de suavizado y K es el *kernel* de densidad: una función normalizada. En nuestro trabajo hemos usado como *kernel* una densidad gaussiana y para h el valor informado como óptimo en la literatura (Steuer et al., 2002)

$$h_{opt} = 1{,}06\sigma n_i^{-1/5}$$

siendo σ^2 la varianza calculada desde el registro de valores.
Queda por calcular las integrales. Usamos aquí una segunda aproximación: cada integral corresponde a un valor de expectación

$$\left\langle \ln\left[f\left(y\right)\right]\right\rangle = \int_{-\infty}^{\infty} dy\ f\left(y\right) \ln\left[f\left(y\right)\right]$$

por lo que si el conjunto de valores registrados es una realización aceptable de la variable aleatoria podemos aproximar

$$\int_{-\infty}^{\infty} dy\ f\left(y\right) \ln\left[f\left(y\right)\right] \simeq \frac{1}{n_i} \sum_{j=1}^{n_i} \ln\left[\hat{f}\left(y_j\right)\right] \tag{5}$$

3. RESULTADOS

Para la verificación del método propuesto para la segmentación se generaron secuencias heterogéneas de 2000 valores conformadas por dos subsecuencias homogéneas (de largo 1500 y 500 respectivamente) como a continuación se detallan. Se calculó el promedio de la DJS para cada posición del punto de segmentación sobre 200 secuencias y se presentan los resultados obtenidos en forma gráfica.

3.1. Caso 1

Como primer caso de análisis elegimos la densidad de probabilidad

$$\mu_i\left(y\right) = 1/\eta_i \exp\left(-y/\eta_i\right) \tag{6}$$

para ambas subsecuencias, cambiando el valor medio en cada una siendo

$$\eta_1 = 1, \quad \eta_2 = 2$$

respectivamente en cada segmento.
En la figura 2 ilustramos los valores promedio de DJS promediados sobre 200 secuencias. Notamos que la curva es suave aún para el número relativamente bajo de secuencias consideradas. El método ha detectado correctamente la posición del punto de segmentación en $n_1 = 1500$ de acuerdo con la construcción de las secuencias.

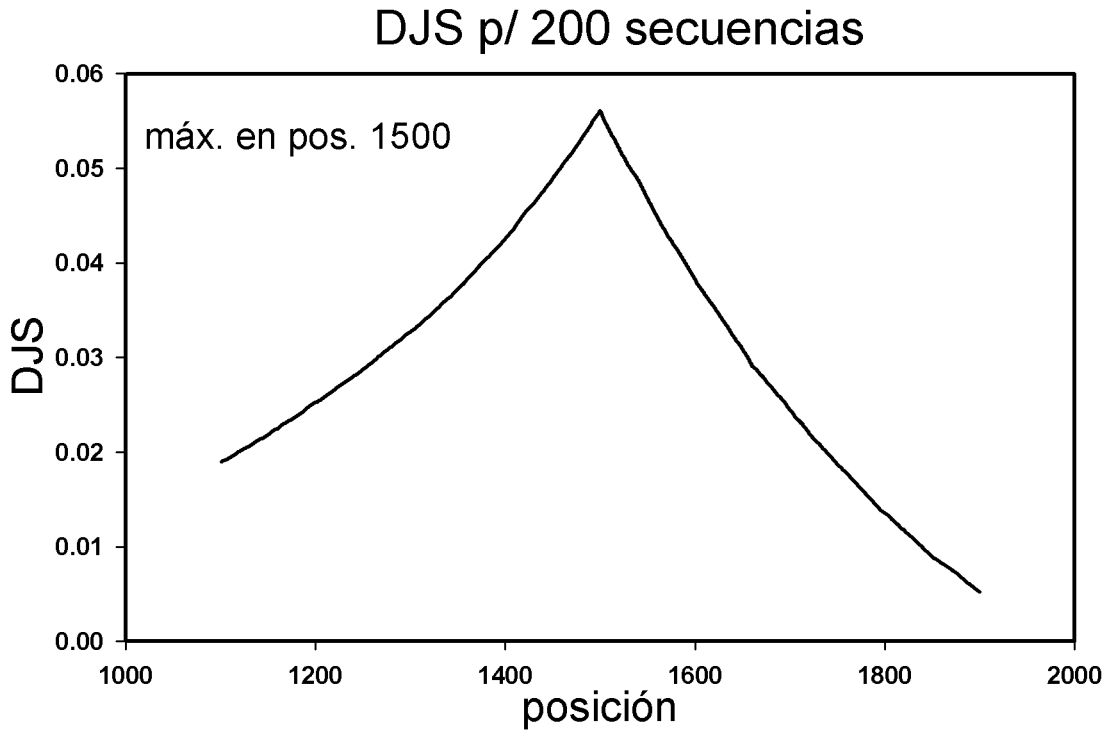

Figura 2: Resultado del cálculo de JSD para el caso 1. La secuencia $\mathcal{S}$ está conformada por 1500 números generados según μ_1 seguidos de 500 números generados con μ_2 (ver texto). El máximo del promedio sobre 200 secuencias marca el punto de segmentación ($n_1 = 1500$).

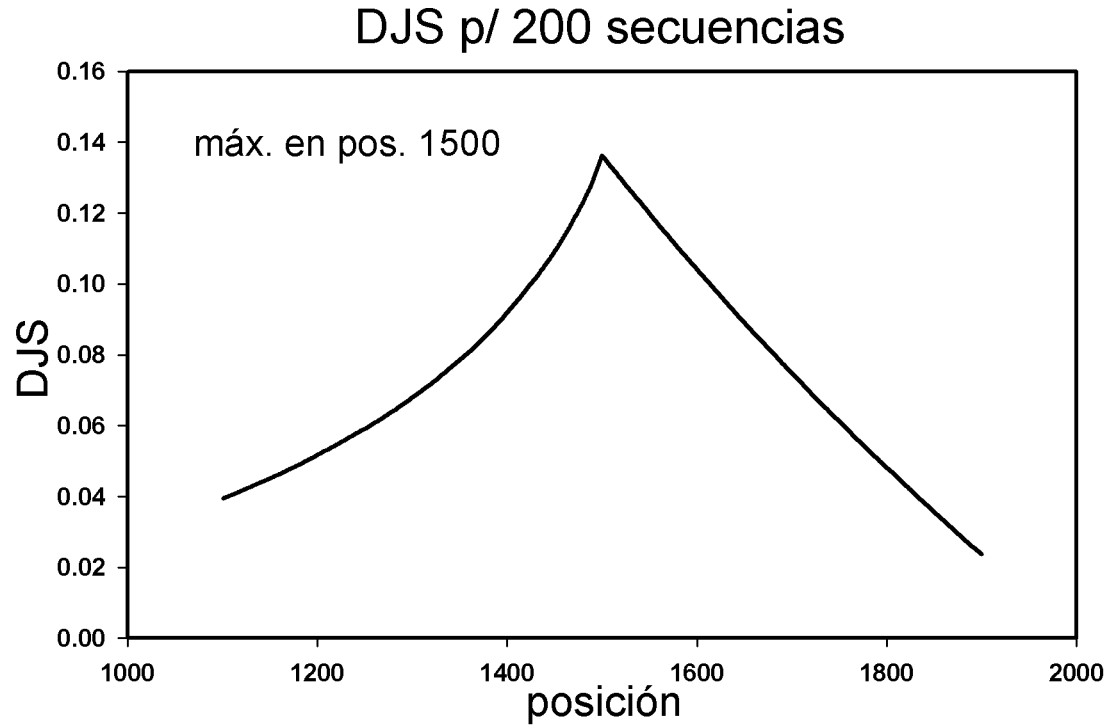

Figura 3: Resultado del cálculo de JSD para el caso 2. La secuencia $\mathcal{S}$ está conformada por 1500 números generados según μ_1 seguidos de 500 números generados con μ_2 (ver texto). El máximo del promedio sobre 200 secuencias marca el punto de segmentación ($n_1 = 1500$).

3.2. Caso 2

En este caso las densidades de probabilidad elegidas son

$$\begin{aligned} \mu_1(y) &= \frac{1}{\eta}\exp(-y/\eta) \\ \mu_2(y) &= \frac{\lambda}{\eta}\left(\frac{\lambda y}{\eta}\right)^{\lambda-1} \frac{1}{\Gamma(\lambda)}\exp(-\lambda y/\eta) \end{aligned} \tag{7}$$

para cada subsecuencia respectivamente. Ilustramos los valores del promedio de la DJS sobre 200 secuencias. Nuevamente encontramos una buena respuesta del método detectando el punto de segmentación correctamente.

3.3. Caso 3

Finalmente como último caso de análisis en esta presentación consideramos las densidades de probabilidad

$$\begin{aligned} \mu_1(x) &= \frac{1}{\pi}\frac{1}{1+x^2} \\ \mu_2(x) &\frac{1}{\sqrt{2\pi}}\exp(-x^2) \end{aligned} \tag{8}$$

para la primer y segunda subsecuencia respectivamente. Se eligió esta opción por ser la distri-

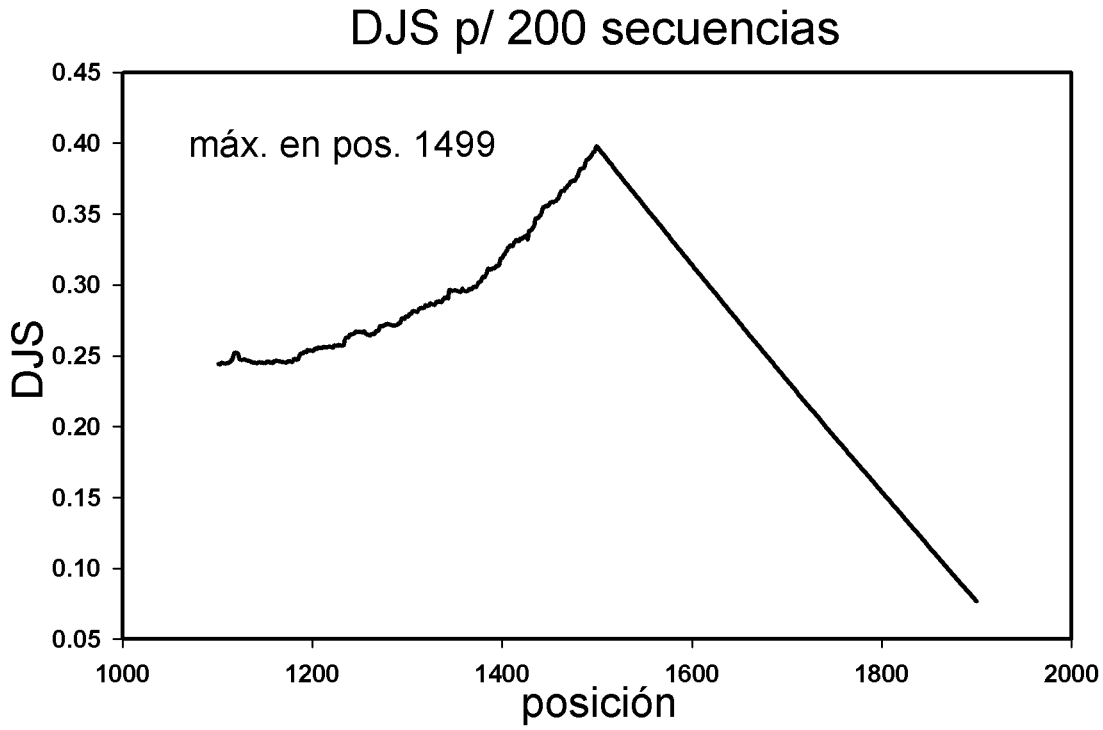

Figura 4: Resultado del cálculo de JSD para el caso 3. La secuencia $\mathcal{S}$ está conformada por 1500 números generados según μ_1 seguidos de 500 números generados con μ_2 (ver texto). El máximo del promedio sobre 200 secuencias marca el punto de segmentación ($n_1 = 1500$).

bución de Cauchy particularmente difícil de aproximar y permite verificar el comportamiento del método.

En este caso el máximo de la DJS detecta el punto de segmentación en 1499, que podemos considerar una respuesta razonable.

4. CONCLUSIONES

Hemos presentado un método para el análisis y segmentación de secuencias heterogéneas de valores reales, basado en el cálculo de la Divergencia de Jensen-Sahnnon (DJS) a partir de la aproximación de las distribuciones de probabilidad asociadas a los valores en los segmentos homogéneos.

El algoritmo utilizado se basa en la estimación de las densidades de probabilidad por el método del *kernel* de densidad, sin necesidad de dar una forma funcional explícita a estas densidades. Esta aproximación es una altenativa al método de *binning* que presenta dificultades, en particular para números pequeños de datos.

La evaluación del método mostró un desempeño que consideramos satisfactorio al detectar los puntos de segmentación en secuencias generadas a partir de distribuciones de probabilidad conocidas. En particular en los casos de análisis 1 y 2 hemos recurrido a distribuciones de probabilidad con momentos definidos para ambos segmentos. Por otra parte, en el caso 3 hemos usado una distribución de Cauchy que no tiene momentos definidos exponiendo al método a una situación más exigente. En todos los casos se pudo detectar el punto de segmentación.

Para el método de aproximación de las densidades de probabilidades hemos recurrido a un *kernel* gaussiano. Queda pendiente la evaluación del desempeño con *kernels* alternativos, que será mótivo de futuras comunicaciones.

Agradecimientos

Los autores agradecen el apoyo de UTN a través del proyecto UTN3559.

REFERENCIAS

Deza E. y Deza M. M. *Dictionary of Distances, First Edition* Elsevier B. B. 2006.

Fisher R. The general sampling distribution of the multiple correlation coefficient. *Proc. Roy. Soc.*, 121:654–673, 1928.

Grosse I., Bernaola-Galván P., Carpena P., Román-Roldán R., Oliver J. y Stanley H. Analysis of symbolic sequences using Jensen-Shannon divergence. *Phys. Rev. E*, 65:041905-1/16, 2002.

Lamberti P. y Majety A. Non-logarithmic JensenŰShannon divergence. *Phys. A*, 329:81–90, 2003.

Lesche B. Instabilities of Rényi entropies. *J. Stat. Phys*, 27:419–422, 1982.

Lin J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory*, 37:145–151, 1991.

López N., Orosco E., di Sciacio F. Surface Electromyographic Onset Detection Based on Statistics and Information Content. *J. Phys.: Conference Series*, 332:012043, 2011.

López Ruiz R., Mancini H., Calbet X. A statistical measure of complexity. *Phys. Lett. A*, 209:321–326, 1995.

Pereyra M., Lamberti P., Rosso O. Wavelet Jensen-Shannon divergence as a tool for studying the dynamics of frequency band components in EEG epileptic seizures. *Phys. A*, 379:122–132, 2007.

Ré M. y Azad R. Generalization of Entropy Based Divergence Measures for Symbolic Sequence Analysis. *PLoS ONE*, 9:e93532, 2014.

Rosso R., Larrondo H., Martin M. T., Plastino A. Distinguishing Noise from Chaos. *Phys. Rev. Lett.*, 99:154102, 2007.

Silverman B. W. *Monographs on Statistics and Applied Probability 26: Density Estimation for Satatistics and Data Analysis* Chapman and Hall/CRC 1986.

Steuer R., Kurths J., Daub C., Weise J., Selbig J. The Mutual Information: Detecting and Evaluating dependencies between variables. *Bioinformatics*, 18 S2:S231–S240, 2011.