Análisis de conglomerados y clasificación de series de tiempo. Utilizando Kullback-Liebler sobre la densidad espectral. Un ejemplo aplicado a audio de gatos y perros

Campos Martínez Joaquín Nicolás
Alumno de Actuaría
Facultad de Ciencias Actuariales
Universidad Anáhuac Campus Norte

Introducción

En este practicum queremos mostrar el uso de la densidad espectral para clasificar series de tiempo, algunas de las razones que impulsaron está idea fueron: primero algunos ejemplos, particularmente la clasificación de temblores basada en esta misma metodología(Shumway, 2003)(Yoshihide Kakizawa, 1998), en el caso de Shumway como un medio para monitorear que se respeten las reglas internacionales de pruebas nucleares. También tenemos el caso de la clasificación de flamas según el sonido generado en la combustión.(J. Yuan, 2000)

La primera parte del texto consiste en un ejemplo de la clasificación de audios de perros y gatos que forma parte de una base de datos de 48 clases acústicas adicionales, utilizada en un artículo sobre redes neuronales (Naoya Takahashi y Gool, 2016).

La base utilizada consta de 164 archivos WAV con sonidos de gatos, maullidos, ronroneos, o similares y 113 sonidos de perros, ladridos principalmente. Decidimos seguir el ejercio propuesto en el sitio Kaggle(marc moreaux, 2017), donde se propone dividir la muestra en conjunto de entranamiento con 115 archivos de gatos y 64 de perros, los 49 restantes de cada categoría se asignan al conjunto de prueba.

Los resultados preliminares son bastante satisfactorios, y notamos que el espectro de maullidos es muy distinto de ladridos en términos de duración y harmonía.

La segunda parte del texto utiliza los mismos 277 archivos de audio utilizados en el ejemplo anterior, en este caso presentamos los resultaod s

Al final proponemos como ejercicio el posible uso en la clasificación de series de tiempo económicas, particularmente de precios de acciones o activos, de modo que podemos categorizar por estructuras de covarianza con un análisis de conglomerados, así podríamos decir que activos en un mismo grupo tienen un riesgo similar.

1. Objetivo

Nosotros queremos presentar una metodología de clasificación y conglomeración de series de tiempo basada en la estructura de correlación. Nuestra base es utilizar la medida de discrepancia de Kullback-Liebler sobre la densidad espectral para clasificar en un grupo, según el espectro medio de cada grupo.

2. Relevancia

Las series de tiempo pueden ser observadas en diferentes ámbitos, información de ventas, precios de acciones, tasas cambiarias, información sobre el clima, datos biomédicos, etc. (Saeed Aghabozorgi, 2015)

Las aplicaciones de análisis de conglomerados en estos ámbitos son variadas, por ejemplo, en medicina es importante reconocer la diferencia entre señales normales en un ECG, EEG, EMG de aquellas producidas por enfermedades. En sismología, para discriminar las ondas sísmicas, de los moviemientos naturales de la tierra (Corduas Marcella, 2008) o para monitorear violaciones al CTBT (Tratado de prohibición completa de los ensayos nucleares, en español) (Shumway, 2003).

Otras aplicaciones son, en astronomía las curvas de luz muestran el brillo de una estrella en un periodo de tiempo, en medicina la actividad cerebral, en el medio ambiente y urbanización los niveles de la marea, niveles de contaminantes en el aire ($PM_{2.5}, PM_{10}$) (Saeed Aghabozorgi, 2015).

Otra razón es la gestión de portafolios de inversión, donde se requiere diversificar el riesgo, y no tener muchas acciones de empresas con giros similares, o que esten altamente correlacionadas.

3. Antecedentes

Hay múltiples artículos sobre análisis de conglomerados, y algunos de ellos se enfocan en series de tiempo. *Time-series clustering - A decade review* (Saeed Aghabozorgi, 2015) tiene como objetivo comparar los enfoques más populares.

«This review will expose four main components of time-series clustering and is aimed to represent an updated investigation on the trend of improvements in efficiency, quality and complexity of clustering time-series approaches during the last decade and enlighten new paths for future works» (Saeed Aghabozorgi, 2015)

Clustering of time series data - a survey busca resumir investigación realizada en el tema y sus aplicaciones en varios campos. Incluye aspectos basicos de los algoritmos más generales que se utilizan en estudios de conglomerados, medidas de similitud y disimilitud, evaluación de conglomerados. (Liao, 2005)

Ambos artículos distinguen 3 principales categorizaciones de los acercamientos a conglomeración de series de tiempo, trabajar con los datos brutos, extraer «características» de los datos, basarse en un modelo al que se ajusten los datos.

«Computational Models of Music Similarity and their Application in Music Information Retrieval» es una tesis que busca mostrar el desarrollo de las técnicas de descubrimiento de música basado en medidas de similitud y disimilitud. Uno de los enfoques consiste en utilizar de función de densidad espectral. Sin embargo también utiliza técnicas aplicables a series de tiempo. (Pampalk, 2006).

4. Como perros y gatos

Como dijimos nos centramos en 2 ejemplos, el primero consiste en crear una regla de clasificación entre la clase de perros y gatos. El segundo ejemplo es para mostrar el uso la divergencia K-L en series de tiempo.'

4.1. Perros vs gatos

Nuestro ejemplo está basado en (marc moreaux, 2017), utilizamos 115 audios de gatos y 64 de perros para entrenar el algoritmo y 69 de cada clase para validar nuestro modelo. Primero tenemos un problema de desbalanceo en nuestro conjunto de entrenamiento, para tener la misma cantidad de observaciones de perros como de gatos utilizamos la técnica de bootstraping con reemplazo para sobremuestrear los 64 perros de entrenamiento.

4.1.1. Análisis exploratorio

Observemos algunas series de tiempo de gatos y perros. La figura 1 es la gráfica de series de tiempo de un maullido, es bastante sostenido, y notamos que la variación aumenta del inicio del maullido a la mitad para despues bajar. Por otro lado, 2 nos muestra un ladrido, en comparación es más explosivo, en el sentido de que la varación más alta es al inicio, y luego disminuye. Las gráficas de series anteriores, así

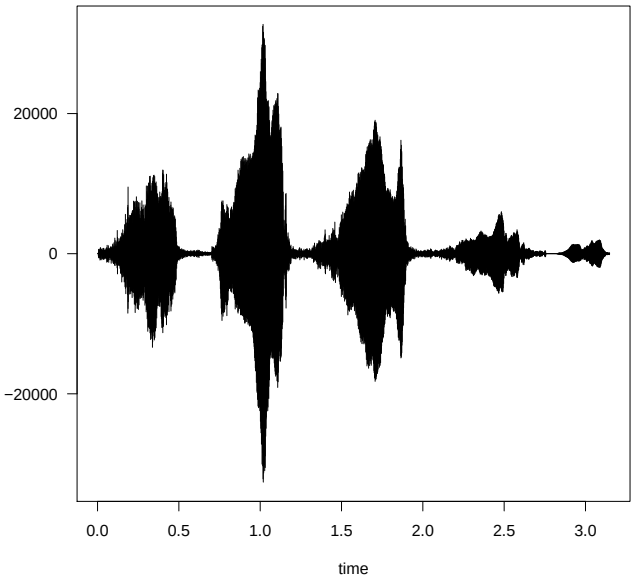

Figura 1: La representación gráfica de un maullido bastante nítido

como los espectrogramas siguientes corresponden a los archivos *cat_75.wav* y *dog_barking_101.wav* tomados al azar de la muestra.

Por un lado el maullido 3 muestra un comportamiento harmónico y con mayor duración que el ladrido 4. Esto nos lleva a pensar que podemos utilizar la densidad espectral como representación de la serie de tiempo en un análisis discriminante.

5. Densidad espectral

El estimador que utilizamos de la función de densidad espectral es la correción al periodograma con la ventana de Hann definida por

$$w(n) = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right)$$

Donde N es el número de observaciones en la serie. De modo que el espectro queda representado por

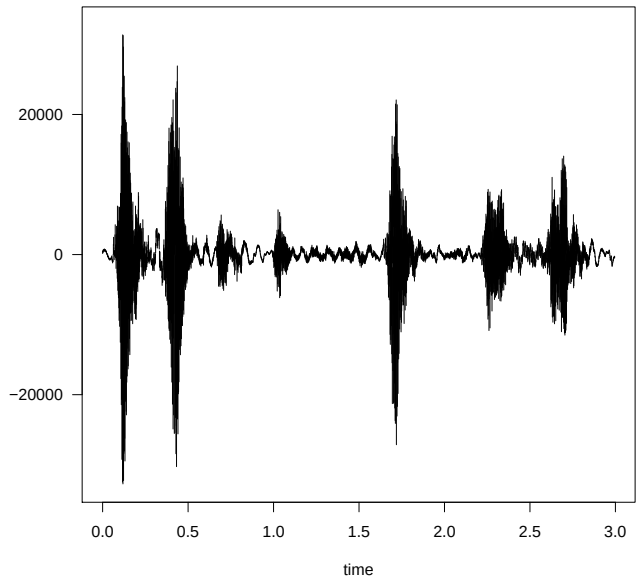

Figura 2: La representación gráfica de un ladrido bastante nítido

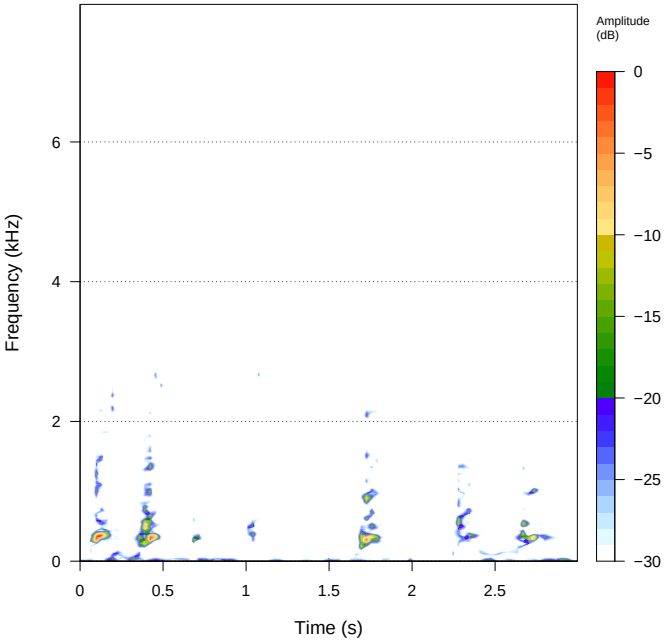

Figura 4: Espectograma ladrido

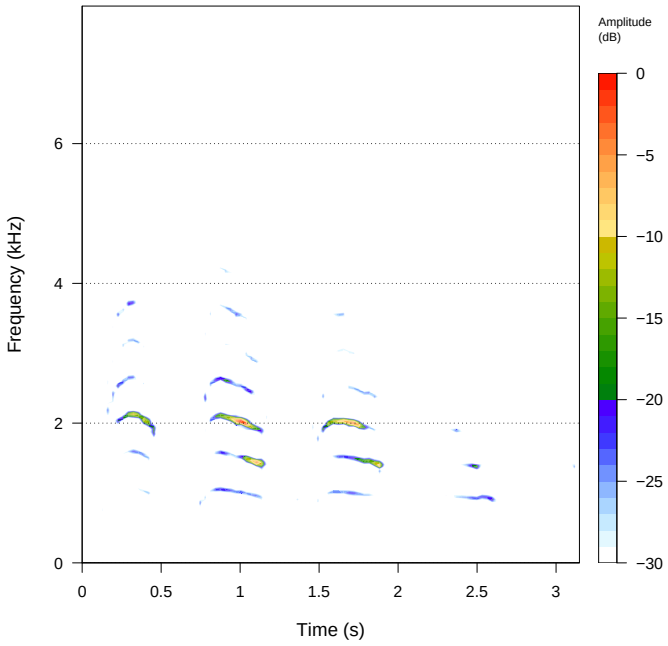

Figura 3: Espectograma maullido

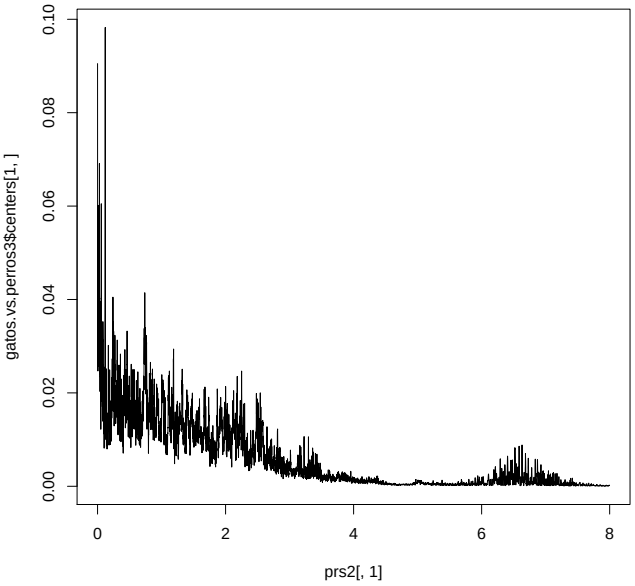

Figura 5: Densidad gatos 1

6. Clasifiación cortando las series

Los audios son de distinta duración, de modo que los espectros estimados serán es distintas frecuencias, para solucionar este problema cortamos las series con la duración más corta (*cat_55.wav*) de poco menos de un segundo. Partimos del punto medio de las mismas y tomamos aproximadamente medio segundo a la izquierda y medio a la derecha.

Una vez obtenida la densidad espectral, obtuvimos el espectro promedio por clase, la figura 5 nos muestra la densidad espectral de los audios de gatos, podemos notar que tiene su moda en frecuencias bajas, y luego decae lentamente a 0. En cambio, la densidad espectral para ladridos, 6, tienen su moda en frecuencias medias, y decar rápidamente a 0. Ahora que tenemos nuestros espectros promedio, el siguiente paso es la regla de clasificación

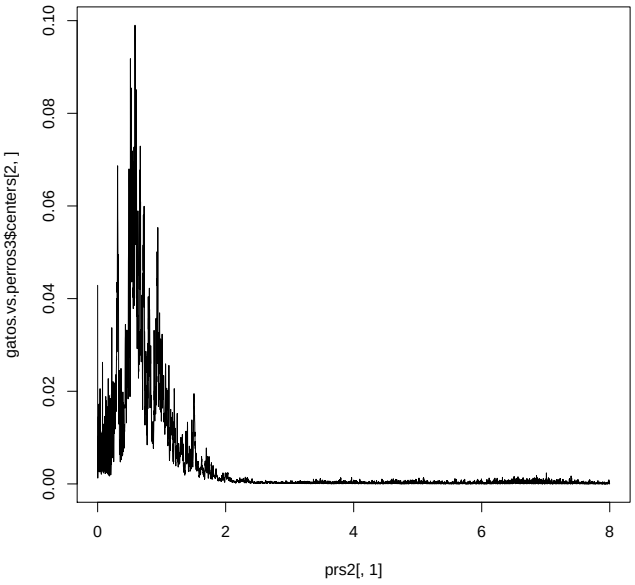

Figura 6: Densidad perro 1

$$\text{clasifica en} \left\{ \begin{array}{l} G \text{ si } I(f_g, f) \leq I(f_p, f) \\ P \text{ si } I(f_g, f) > (f_p, f) \end{array} \right\}$$

Donde $I(f,g) = \int_x f(x) \log(\frac{f(x)}{g(x)}) dx$ es la información de Kullback. La interpretación de I es la información media por observación en favor de f contra g. En este caso nuestras x son las frecuencias con las que estamos trabajando.

Al aplicar está regla obtenemos la matriz de confusión para el conjunto de entrenamiento 6, donde nuestra tasa de error es del 17%. Para el conjunto de prueba nuestra matriz de confusión nos indica una tasa de error del 26%

	Gato	Perro
Gatos	77	37
Perro	6	58

	Gato	Perro
Gatos	36	13
Perro	3	46

Referencias

Corduas Marcella, P. D. (2008). *Time series clustering and classification by the autoregressive metric.*

J. Yuan, Y. Z. (2000). *Spectral analysis of combustion noise and flame pattern recognition.*

Liao, T. W. (2005). *Clustering of time series data - a survey.* ([Online; Obtenido 09-Octubre-2017])

marc moreaux. (2017). *Audio cats and dogs, classify raw sound events.* https://www.kaggle.com/mmoreaux/audio-cats-and-dogs/home. ([Online; Obtenido 04-Noviembre-2018])

Naoya Takahashi, B. P., Michael Gygli, y Gool, L. V. (2016). *Deep convolutional neural networks and data augmentation for acoustic event recognition.*

Pampalk, E. (2006). *Computational models of music similarity and their application in music information retrieval.* ([Online; Obtenido 09-Octubre-2017])

Saeed Aghabozorgi, T. Y. W., Ali Seyed Shirkhorshidi. (2015). *Time-series clustering – a decade review.*

Shumway, R. H. (2003). *Time-frequency clustering and discrimination analysis.*

Yoshihide Kakizawa, M. T., Robert H. Shumway. (1998). *Discrimiantion and clustering for multivariate time series.*