TDA

López Fernando
Maldonado Areli
Kosoi Nathan

January 2022

1. Objetivo y motivación

Nuestro objetivo es predecir colapsos financieros mediante señales tempranas de alarma. La motivación detrás de esto es que aplicando TDA al análisis de series de tiempo se ha logrado caracterizar y detectar patrones en los datos. Esto sería de gran ayuda para la planeación económica, pero también brindaría luz sobre el misterio del mecanismo de las crisis financieras.

Además, el método que proponen los autores es tan general que puede ser aplicado a cualquier tipo de fenómeno que implique series de tiempo.

2. Introducción

"El análisis topológico de datos (TDA) hace referencia a la combinación de métodos estadísticos, computacionales y topológicos que permiten encontrar estructuras geométricas en datos" (M. Gidea, Y. Katz, 2017, p. 821)

Para comenzar daremos algunos datos interesantes, en el año 2000 fueron almacenados 800,000 petabytes (PB) de datos y se esperaba que para el año 2020 esta cantidad aumentará a 35 zettabytes (ZB). Cada petabyte equivale a 10^{15} bytes mientras que un zettabyte equivale a 10^{21} bytes. Esta estimación se vio rebasada por la realidad ya que según Global Datasphere la cifra de datos que se crearon o replicaron durante el 2020 alcanzó unos 64,2 zetabytes. Además, estiman que para el 2025 esta cifra al menos alcance los 180 zettabytes.

La cantidad de datos es gigantesca. En un mar de datos crudos la implementación del análisis topológico de datos permite procesar bases de datos complejas y sin estructura que no pueden ser tratadas con métodos convencionales. Esto se logra a través de la construcción de complejos simpliciales asociados a los datos, obteniendo características cualitativas del conjunto a partir de la homología de dicho simplejo.

3. Métodos en abstracto

Empezaremos diciendo que para que podamos aplicar análisis topológico de datos, la base debe ser codificada como un conjunto finito de puntos en un espacio métrico. Además, es muy importante mencionar que el concepto subyacente a TDA es la persistencia homológica.

El proceso para computar la persistencia homológica asociada a una nube de puntos es, primero, la construcción de un complejo simplicial. Este se construye con respecto a un resolution scaling parameter que conforme vaya cambiando irá propiciando la aparición de algunas características topológicas, mientras que algunas otras desaparecerán. Cada característica topológica corresponde a un valor de nacimiento y muerte y la diferencia entre estos dos valores representa su persistencia. Una característica topológica con un rango de persistencia mayor es más significante que una con un rango menor, ya que estas pueden ser interpretadas como ruido. En realidad, todas las características que emergen de la base de datos son conservadas y se les asigna un peso acorde a su persistencia. Todo lo antes descrito es algo que en el artículo describen como una filtración, y el output de este proceso es el diagrama de persistencia.

Las dos coordenadas de cada punto corresponden a los valores de nacimiento (eje x) y muerte (eje y) de un hoyo k-dimensional. La información del diagrama de persistencia es condensada en el paisaje de persistencia, el cual consiste en una secuencia de funciones continuas y lineales definidas a trozos que se basan en las coordenadas de nacimiento-muerte. Los diagramas de persistencia tienen una estructura de espacio métrico mientras que los paisajes de persistencia tienen que ser encajados en un espacio de Banach, esto permite estudiar sus características estadísticas.

Una característica importante de la persistencia homológica es que tanto el diagrama de persistencia como el paisaje de persistencia son robustos a perturbaciones, es decir, si la base sufre cambios pequeños, el diagrama/paisaje de persistencia se modificara en esta proporción y esto constituye un punto clave para su estudio estadístico.

4. Series de tiempo caoticas con ruido

Consideraremos d series de tiempo $\{x_n^k\}_n$ donde $k = 1, ..., d$ y una ventana deslizante de tamaño w.

Consideremos las ecuaciones del mapa de Hénon:

$$\begin{aligned} x_{n+1} &= 1 - ax_n^2 + by_n \\ y_{n+1} &= x_{n+1}, \end{aligned} \tag{1}$$

donde a, b son parametros reales. Para algun rango de valores de a, b cada condición inicial (x_0, y_0) en alguna región apropiada del plano (la cuenca de atracción), la secuencia (x_n, y_n) se aproxima al mismo conjunto invariante, a esto se le conoce como el atractor de Hénon.

Para algunos valores de a, b el atractor es caotico.Se fija el parámetro $b = 0.3$ y $0 \leq a \leq 1.4$. Para valores fijos de a con $0 < a < 1.06$ existe un atractor que experimenta bifurcaciones que duplican el periodo, mientras que para $a \approx 1.06$ un atractor caótico aparece. Cuando a se encuentra entre 1.06 y 1.4 existen intervalos de valores de a para los cuales hay un atractor caótico, que se intercalan con intervalos de valores de a para los que existe una órbita periódica atractora.

En el artículo modificaron las ecuaciones del atractor de Hénon haciendo que uno de los parámetros cambie lentamente en el tiempo, además, le metieron ruido. Para valores fijos del parámetro b ($b = 0.27, b = 0.28, b = 0.29, b = 0.3$) tenemos que el parámetro crece lentamente en el tiampo, de $a = 0$ a $a = 1.4$, también agregaron ruido gaussiano.
Esto se traduce al siguiente sistema de ecuaciones:

$$x_{n+1} = 1 - a_n x_n^2 + b y_n + \sigma W_n \sqrt{\Delta t} \tag{2}$$

$$y_{n+1} = x_{n+1} + \sigma W_n \sqrt{\Delta t} \tag{3}$$

$$a_{n+1} = a_n + \Delta t \tag{4}$$

$$\tag{5}$$

donde W_n es una variable normal aleatoria, $\Delta t > 0$ es un paso pequeño y $\sigma > 0$ es la intensidad del ruido.
Los términos estocásticos en las ecuaciones anteriores corresponden a un proceso difuso. El tiempo cambia acorde a $t_{n+1} = t_n + \Delta t$, entonces a_n puede ser visto como un equivalente de la variable tiempo.
Primero generaron $d = 4$ realizaciones de x series de tiempo $\{x_n^k\}_n$, donde $k = 1, ..., d$, uno para cada valor del parámetro $b = 0.27, b = 0.28, b = 0.29, b = 0.3$, por lo tanto, por cada tiempo t_n, tenemos un punto $x_n = (x_n^1, ..., x_n^d) \in \mathbb{R}^d$. Utilizaron ua ventana deslizante $w = 50$, generaron una secuencia de conjuntos de nuebes de puntos $X_n = (x_n, x_{n+1}, ..., x_{n+w-1})$, así que cada nube consiste en w puntos en $\mathbb{R}^d$.
Después, para cada nube de puntos X_n se les asocia el correspondiente complejo simplicial de Vietoris - Rips $R(X_n, \epsilon, \epsilon > 0$ y se computa el $1D$ diagrama de persistencia $P_1(X_n, \lambda(X_n)$ y las normas $L^1 = \| \lambda(X_n) \|_1$ y $L^2 = \| \lambda(X_n) \|_2$.
La conclusión es que las normas L^P de los paisajes de persistencia tienen la capacidad de detectar transiciones de comportamientos regulares a caóticos, en sistemas con parámetros de evolución lenta. Cuando los comportamientos pasan de regular a caóticos la dinámica determina cambios significativos en la topología del atractor, que son recogidas por la serie temporal de las normas L^P.

5. Análisis empírico de datos financieros

En el artículo se analizaron diariamente las series de tiempo de SP 500, DJIA, NASDAQ y Russell 2000 de entre el 23 de diciembre de 1987 y 8 de diciembre de 2016, esta información fue obtenida de Yahoo Finance.

Por cada índice y trading day se calcula el log - returns, que se define como los cambios diarios adelantados en el logaritmo del radio $r_{ij} = ln(P_{ij}/P_{i-1})$ donde P_{ij} representa el valor de cierre ajustado del índice j en el día i. Tanto nosotros como los autores del artículo deseamos investigar las propiedades topológicas de esta serie de tiempo multi-dimensional.

Cada nube de puntos está formada por w puntos en $\mathbb{R}^d$. Las coordenadas de cada punto en $\mathbb{R}^d$ representan la función log-returns diaria. Esto significa que cada nube de puntos está dada por una matriz $w \times d$, donde el número de columnas $d = 4$ es el número de series de tiempo $1D$ que están involucradas en el análisis. El número de filas w representa el tamaño de la ventana deslizante. En este trabajo solo se considerara $w = 50$.

Como fue descrito en la parte 3, una vez construido el complejo simplicial y habiendo asignado pesos a las características topológicas, se construye el diagrama de persistencia con su correspondiente función landscape, dada por:

$$\lambda_k(x) = k - max\{f_{(b_\alpha, d_\alpha)}(x) \mid (b_\alpha, d_\alpha) \in P_k\} \tag{6}$$

Y que se calcula con la ventana deslizante de 50 trading days.

Después se calculan las normas L^P del paisaje de persistencia λ para cada rolling window. Este conjunto de valores forman las series de tiempo diarias,después estas se normalizan.

6. Conclusión

El objetivo de este proyecto era demostrar que las series de tiempo formadas por las normas de los paisajes de persistencia pueden ser usadas como señales tempranas de alarma de colapsos financieros, lo cual concluimos cierto gracias al análisis de las series de tiempo de las normas L^P, ya que estas muestran un rápido crecimiento previo a un colapso.

Además, comprobamos que el análisis topológico de datos es muy útil para extraer características de la base datos que después de un proceso como el señalado en la parte 3, se puede llegar a conclusiones más específicas, por ejemplo, que es lo que sucede detrás de una crisis económica.

Al principio del artículo en que se basa este trabajo leímos que no hay consenso del mecanismo de la crisis financiera, y si bien este proyecto no buscaba esclarecer esto, el TDA es un buen método econométrico.

7. Bibliogrfía

Marian Gidea, Yuri Katz. (09 de octubre de 2017). Topological data analysis of financial time series: Landscapes of crashes. Physica A, 491, 820-834.

Eaton, C., Deroos, D., Deutsch, T., Lapis, G., Zikopoulos, P.: Understanding Big Data. Analytics for enterprise class Hadoop and Streaming Data, The McGraw Hill Companies, 2012.

Carlsson, G.: Topology and Data., Bulletin of the American Mathematical Society no. 46 (2009).