Universidad Nacional Autónoma de México

Instituto de Investigaciones en Matemáticas
Aplicadas y Sistemas

Detectando transiciones críticas en series de tiempo de criptomonedas con Análisis Topológico de Datos

Proyecto Minería de datos

Autores:

Barajas Cervantes Alfonso

Flores Tiburcio Luis Fernando

Yañez Espindola José Marcos

Vázquez Rojas José David

Diciembre 2021

Índice

1. Marco Teórico

1.1. Topología y Homotopía

Definición 1. *Sea X un conjunto, una topología τ_X sobre X es una familia de subconjuntos de X, es decir, $\tau_X \subseteq \mathcal{P}(X)$, tal que:*

(I) $\emptyset, X \in \tau_X$

(II) $\forall \mathcal{U} \subseteq \tau_X, \bigcup \mathcal{U} \in X$

(III) $\forall U_1, U_2 \in \tau_X, U_1 \bigcap U_1 \in X$

Decimos que (X, τ_X) es un espacio topológico.

Definición 2. *Sea $f : X \longrightarrow Y$ una función entre espacios topológicos, decimos que f es una función continua si $f^{-1}[V] \in \tau_X, \forall V \in \tau_Y$.*

Nota: esta definición coincide con la definición usual de funciones continua de $\mathbb{R}^n$ en $\mathbb{R}^m$

Definición 3. *Sea X un conjunto, una métrica d en X es una función $d : X \times X \longrightarrow [0, \infty)$ que satisface:*

(I) $d(x, y) = 0 \iff x = y$

(II) $d(x, y) = d(y, x), \forall x, y \in X$

(III) $d(x, y) \leq d(x, z) + d(z, y), \forall x, y, z \in X$

Decimos que (X, d) es un espacio métrico. Notamos que d induce una topología natural $\tau_d = \{U \subseteq X | \forall x \in U \exists \delta > 0$ tal que $B_\delta(x) \subset U\}$

Definición 4. *Decimos que dos espacios topológicos X y Y son homeomorfos si existe $f : X \longrightarrow Y$ una función continua y biyectiva y tal que $f^{-1} : Y \longrightarrow X$ es también una función continua.*

"Dos espacios homeomorfos son iguales".

1.1.1. Homotopías entre funciones y espacios homotópicos

Definición 5. *Sea $f, g : X \longrightarrow Y$ dos funciones continuas, decimos que f, g son homotópicas si existe $H : X \times I \longrightarrow Y$ continua tal que $H(x, 0) = f(x), H(x, 1) = g(y), \forall x \in X$.*

Definición 6. *Sean X, Y dos espacios topólogicos, decimos que X y Y son homotópicos si existen $f : X \longrightarrow Y, g : Y \longrightarrow X$ continuas tales que $f \circ g$ es homotópica a Id_Y y $g \circ f$ es homótopico a Id_X.*

"Dos espacios homótopicos son casi iguales".

1.2. Complejos simpliciales

Definición 7. *Un $n-$simplejo o $n-$simplex σ de $\mathbb{R}^{\mathbb{N}}$ es la envolvente convexa de un conjunto V de $n+1$ puntos $\{v_0, \ldots, v_n\}$ afines independientes (es decir $\{v_1 - v_0, \ldots, v_{n+1} - v_0\}$ son independientes como vectores). A V le llamamos los vértices del simplejo y a los simplejos generados por los subconjuntos de V le llamamos caras de σ.*

Definición 8. *Un complejo simplicial es una colección $\mathcal{K}$ de simplejos tal que cualquier cara de $\sigma \in \mathcal{K}$ es también un simplejo en $\mathcal{K}$ y la intersección de dos simplejos es vacía o es una cara de algún simplejo.*

Definición 9. *Un complejo simplicial abstracto con conjunto de vértices V (finito) es una colección $\tilde{\mathcal{K}}$ de subconjuntos de V tal que $\{v\} \in \tilde{\mathcal{K}}$ para cada $v \in V$ y $\forall \sigma \in \tilde{\mathcal{K}}$ todo subconjunto de σ también está en $\tilde{\mathcal{K}}$.*

Definición 10. *Una realización geométrica de un complejo simplicial abstracto $\tilde{\mathcal{K}}$ es un complejo simplicial $\mathcal{K}$ donde el conjunto de vértices de $\mathcal{K}$ se puede biyectar con el conjunto de vértices de $\tilde{\mathcal{K}}$ de tal forma que la imagen de todo simplejo esté en $\tilde{\mathcal{K}}$.*

Teorema 1. *Si tenemos dos realizaciones geométricas de un complejo simplicial abstracto entonces son homeomorfas.*

1.2.1. Complejo de Vietoris-Rips y complejo de Cech

Definición 11. *Sea X un subconjunto finito de un espacio métrico $\mathcal{X}$ entonces el complejo de Vietoris-Rips* $\mathrm{Rips}_\alpha(X)$ *es el conjunto de simplejos $[x_0, x_1, \ldots, x_k]$ de puntos en X tales que $d(x_i, x_j) \leq \alpha, \forall i, j \in \{0, 1, \ldots, k\}$*

Se sigue inmediatamente que $\mathrm{Rips}_\alpha(X)$ es un complejo simplicial abstracto, el complejo de Cech es otro complejo simplicial relacionado a este.

Definición 12. *El complejo de Cech de X,* $\mathrm{Cech}_\alpha(X)$ *es el conjunto de simplejos $[x_0, x_1, \ldots, x_k]$ de elementos de X tales que*

$$\bigcap_{j=0}^{K} B_\alpha(x_j) \neq \emptyset$$

Observamos que

$$\mathrm{Rips}_\alpha(X) \subseteq \mathrm{Cech}_\alpha(X) \subseteq \mathrm{Rips}_{2\alpha}(X)$$

y si $X \subset \mathbb{R}^d$ entonces $\mathrm{Cech}_\alpha(X)$ y $\mathrm{Rips}_\alpha(X)$ tienen el mismo $1-$esqueleto.

Teorema 2. ***El teorema del nervio:*** *Sea $\mathcal{U} = \{U_\alpha\}_{\alpha \in I}$ una cubierta abierta de un espacio X tal que la intersección de una subfamilia finita siempre es vacía o contraíble, entonces X y el complejo simplicial generado por $\mathcal{U}$ son homotópicos.*

Corolario 1. *Si X es un conjunto de $\mathbb{R}^d$ entonces* $\mathrm{Cech}_\alpha(X)$ *es homotópico a $\bigcup_{x \in X} B_\alpha(x)$*

1.3. Homología simplicial

Definición 13. *Sea $\mathcal{K}$ un complejo simplicial y $\mathbb{F}$ un campo, el espacio de $k-$cadenas de $\mathcal{K}$, $C_k(\mathcal{K})$, es el conjunto de sumas de k-simplejos. Es decir si $\{\sigma_1, \sigma_2, \ldots, \sigma_p\}$ es el conjunto de k-simplejos de $\mathcal{K}$ entonces todos los elementos de $C_k(\mathcal{K})$ se pueden escribir como*

$$c = \sum_{j=1}^{p} \alpha_j \sigma_j, \quad \alpha_j \in \mathbb{F}$$

Si $c' = \sum_{j=1}^{p} \alpha'_j \sigma_j$ es otro elemento de $C_k(\mathcal{K})$ definimos

$$c + c' = \sum_{j=1}^{p} (\alpha_j + \alpha'_j)\sigma_j$$

y

$$\lambda c = \sum_{j=1}^{p} (\lambda \alpha_j)\sigma_j,$$

es fácil notar que bajo estas operaciones $C_k(\mathcal{K})$ es un espacio vectorial de dimensión p y que $\{\sigma_1, \sigma_2, \ldots, \sigma_p\}$ es una base de $C_k(\mathcal{K})$.

Si consideramos $\mathbb{F} = \mathbb{Z}_2$ podemos considerar una k-cadena como un conjunto de k-simplejos y la suma de k-cadenas como la diferencia simétrica de esos conjuntos.

Definición 14. *La frontera de un k-simplejo $\sigma = [v_0, v_1, \ldots, v_k]$ es la $(k-1)$-cadena*

$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k]$$

Donde $\hat{v}_i$ indica que omitimos ese elemento.

Como $\{\sigma_1, \ldots, \sigma_p\}$ forman una base de $C_k(\mathcal{K})$ entonces podemos extender ∂_k de forma lineal a todo $C_k(\mathcal{K})$ y entonces $\partial_k : C_k(\mathcal{K}) \longrightarrow C_{k-1}(\mathcal{K})$.

Definición 15. *El kernel de ∂_k,*

$$Z_k(\mathcal{K}) = \{c \in C_k(\mathcal{K}) : \partial_k(c) = 0\}$$

lo nombramos como el conjunto de k-ciclos de $\mathcal{K}$ y la imagen de ∂_{k+1},

$$B_k(\mathcal{K}) = \{c \in C_k(\mathcal{K}) : \exists c' \in C_{k+1}(\mathcal{K}) \text{ tal que } \partial_{k+1} c' = c\}$$

la nombramos el espacio de k- fronteras de $\mathcal{K}$.

Teorema 3. *Se tiene*

$$\partial_k \circ \partial_{k+1} \equiv 0$$

Demostración. Basta con probar que $\partial_k \circ \partial_{k+1} c = 0$, para todo c $(k+1)$-simplejo.

Tenemos que si $c = [v_0, v_1, \ldots, v_{k+1}]$ entonces $\partial_{k+1} c = \sum_{i=0}^{k+1} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_{k+1}]$ y por tanto

$$\partial_k \circ \partial_{k+1}(c) = \sum_{i=0}^{k+1} (-1)^i \Big(\sum_{j<i} (-1)^j [v_0, \ldots, \hat{v}_j, \ldots, \hat{v}_i, \ldots, v_{k+1}] + \sum_{j>i} (-1)^{j-1} [v_0, \ldots, \hat{v}_i, \ldots \hat{v}_j, \ldots, v_{k+1}] \Big) = 0$$

□

Corolario 2. *Para todo $k \geq 0$ se tiene que*

$$B_k(\mathcal{K}) \subseteq Z_k(\mathcal{K}) \subseteq C_k(\mathcal{K})$$

Definición 16. *(Grupos de homología y números de Betti): El k-ésimo grupo de homología de $\mathcal{K}$ es el cociente de espacios vectoriales (o de grupos):*

$$H_k(\mathcal{K}) = Z_k(\mathcal{K}) \big/ B_k(\mathcal{K})$$

y el k-ésimo número de Betti es la dimensión

$$\beta_k(\mathcal{K}) = \dim(H_k(\mathcal{K}))$$

del espacio vectorial $H_k(\mathcal{K})$

1.4. Homología persistente

Definición 17. *Una filtración de un complejo simplicial $\mathcal{K}$ es una familia anidad de subcomplejos $(\mathcal{K}_r)_{r \in T}$, donde $T \subseteq \mathbb{R}$ tal que para todo r, r' con $r < r'$, $\mathcal{K}_r \subseteq \mathcal{K}_{r'}$ y $\mathcal{K} = \cup_{r \in T} \mathcal{K}_r$. El conjunto T puede ser finito o infinito. Esta definición se extiende de manera natural a cualquier espacio topológico $\mathbb{M}$*

Notemos que $f : \mathbb{M} \longrightarrow \mathbb{R}$ es una función entonces $M_r := f^{-1}((-\infty, r])$, define una filtración conocida como la filtración de conjuntos de subnivel de f.

Las funciones definidas en los vértices de un complejo simplicial dan lugar a otro ejemplo importante de filtración: sea $\mathcal{K}$ un complejo simplicial con un conjunto de vértices V y $f : V \longrightarrow \mathbb{R}$. Entonces f puede extenderse a todos los simplejos de $\mathcal{K}$ por $f([v_0, \ldots, v_k]) = \max\{f(v_i) : i = 1, \ldots, k\}$ para cualquier simplejo $\sigma = [v_0, \ldots, v_k] \in \mathcal{K}$ y la familia de subcomplejos $\mathcal{K}_r = \{\sigma \in \mathcal{K} : f(\sigma) \leq r\}$ define una filtración llamada la filtración de conjuntos de subnivel de f. Del mismo modo, se puede definir el filtración de conjuntos de nivel superior de f.

Para cada uno de estos complejos, podemos calcular su homología k -dimensional $H_k(\mathcal{K}_r)$ con coeficientes en algún campo.

Los generadores del grupo de homología 0-dimensional $H_0(\mathcal{K}_r)$ corresponden a los componentes conectados de $\mathcal{K}_r$, los generadores del grupo de homología unidimensional $H_1(\mathcal{K}_r)$, corresponden a los 'huecos', los generadores del grupo de homología bidimensional $H_2(\mathcal{K}_r)$, corresponden a 'vacíos' en $\mathcal{K}_r$ etc.

Cada inclusión $\mathcal{K}_r \hookrightarrow \mathcal{K}_{r'}$, en la filtración de complejos simpliciales, induce un homomorfismo de grupo canónico $F_k^{r,r'} : H_k(\mathcal{K}_r) \to H_k(\mathcal{K}_{r'})$ entre los grupos de homología correspondientes. La imagen $\operatorname{Im} F_r^{r,r'} \subseteq H_k(\mathcal{K}_{r'})$ consta de generadores de homología que están presentes en el valor del parámetro r (por lo tanto, nacen en algún valor de parámetro $r_1 \leq r$), y permanecen presentes en el valor de parámetro r', y se denomina grupo de homología persistente. Para cada clase de homología k-dimensional distinta de cero α existe un par de valores r_1, r_2, tales que:

- $\alpha \in H_k(\mathcal{K}_r)$ pero no está en la imagen de ningún $H_k(\mathcal{K}_\delta), r_1 > \delta$ debajo del homomorfismo correspondiente.
- la imagen de $\alpha \in H_k(\mathcal{K}_r)$, no es cero para todos $r_1 < r < r_2$, pero la imagen de $\alpha \in H_k(\mathcal{K}_{r_2})$, es cero.

En este caso, se dice que la clase α "nace.$^{\text{en}}$ el valor del parámetro $b_\alpha := r_1$, y "muere.$^{\text{en}}$ el valor del parámetro $d_\alpha := r_2$; el par (b_α, d_α) representa los índices de "nacimientoz de "muerte"de α. La multiplicidad $\mu_{(b_\alpha,d_\alpha)}$ del punto (b_α, d_α) es igual a el número de clases α que nacen en b_α y mueren en d_α. Esta multiplicidad es finita ya que el complejo simplicial es finito . La información sobre los generadores de homología k-dimensional en todas las escalas puede codificarse en un *diagrama de persistencia* P_k. Dicho diagrama consta de:

- Para cada clase de homología k-dimensional α se asigna un punto $p_\alpha = (b_\alpha, d_\alpha) \in \mathbb{R}^2$ junto con su multiplicidad μ_α
- Además, P_k contiene todos los puntos en la diagonal positiva de $\mathbb{R}^2$; estos puntos representan todos los generadores de homología trivial que nacen y mueren instantáneamente en cada nivel; cada punto de la diagonal tiene una multiplicidad infinita.

Los ejes de un diagrama de persistencia son índices de nacimiento en el eje horizontal e índices de muerte en el eje vertical.

El espacio de los diagramas de persistencia se puede incrustar en un espacio de Banach, cuya norma se puede utilizar para derivar una métrica. Una de esas incrustaciones se basa en Paisajes de persistencia, que consisten en secuencias de funciones en el espacio de Banach. $L^p(\mathbb{N} \times \mathbb{R})$. Para cada punto de nacimiento-muerte $(b_\alpha, d_\alpha) \in P_k$, primero definimos una función lineal

$$f_{(b_\alpha,d_\alpha)}(x) = \begin{cases} x - b_\alpha & \text{si } x \in (b_\alpha, \frac{b_\alpha+d_\alpha}{2}]) \\ d_\alpha - x & \text{si } x \in (\frac{b_\alpha+d_\alpha}{2}, d_\alpha) \\ 0, & \text{si } x \in (b_\alpha, d_\alpha) \end{cases}$$

Para un diagrama de persistencia P_k que consta de un número finito de puntos fuera de la diagonal, asociamos una secuencia de funciones $\lambda_i : \mathbb{R} \longrightarrow [0, \infty), i \in \mathbb{N}$ dado por $\lambda_i(x) = i-\text{máx}\{(f_{(b_\alpha,d_\alpha)}\}|(b_\alpha, d_\alpha)\}$. donde $i-$máx denota el i-ésimo valor más grande de una función. Establecemos $\lambda_i = 0$ si el i-ésimo valor más grande no existe. Por supuesto, λ también depende de la dimensión n correspondiente al diagrama de persistencia P_k.

A través del encaje anterior, los paisajes de persistencia forman un subconjunto del Espacio Banach $L^p(\mathbb{N} \times \mathbb{R})$ donde la norma de λ es la usual de L^p.

1.5. Sistemas dinámicos y encajes

Definición 18. *Un sistema dinámico global de tiempo continuo es un par* $(\mathbb{M}, \Phi)$, *donde* M *es un espacio topológico y* $\Phi : \mathbb{R} \times \mathbb{M} \longrightarrow \mathbb{M}$ *es un mapeo continuo de modo que:*

- $\Phi(0, p) = p$,
- *y* $\Phi(s, \Phi(t, p)) = \Phi(s + t, p)$ *para todo* $p \in \mathbb{M}$ *y* $s, t \in \mathbb{R}$

Algunos subconjuntos de M son especialmente importantes ya que atraen la evolución de los estados cercanos.

Definición 19. *Un conjunto* $A \subset M$ *se llama atractor si cumple tres condiciones:*

1. *A es compacto,*
2. *A es un conjunto invariante. es decir, si* $p \in A$ *entonces* $\Phi(t, p) \in A, \forall p \in A, t \geq 0$ *y*
3. *Tiene una cuenca de atracción abierta. En otras palabras, hay vecindad abierta* $A \subset V$ *tal que*

$$\bigcap_{t \geq 0} \{\Phi(t, p) : p \in V\} = A$$

Un atractor A *se llama extraño, y esto es solo un nombre, si hay puntos cercanos arbitrariamente* p, p' *en una cuenca de atracción de* A *para la cual la distancia entre* $\Phi(t, p)$ *y* $\Phi(t, s')$ *crece exponencialmente rápido con* t *y* A *tiene una dimensión de Hausdorff no entera.*

En la práctica es extremadamente raro tener una descripción explicita de un sistema dinámico de interes. En cambio, a menudo se pueden recopilar mediciones de cantidades relevantes para cada estado $p \in \mathbb{M}$, por ejemplo, en la predicción del clima, se puede estimar la temperatura, la presión, etc. Se puede pensar en una forma de medir como un mapeo continuo $F : \mathbb{M} \longrightarrow \mathbb{R}$ y dado un estado inicial $p \in \mathbb{M}$ podemos obtener la serie de tiempo

$$\begin{aligned} \varphi_p : \mathbb{R} &\to \mathbb{R} \\ t &\mapsto F(\Phi(t, p)) \end{aligned}$$

Una sola serie de tiempo puede parecer una completa simplificación de la dinámica subyacente. Sin embargo, el teorema de encaje de Takens, implica que pueden ser muy utiles. Para formalizar esto, sean $\mathbb{M}, \mathbb{N}$ variedades diferenciales y $k \in \mathbb{N}$, definimos $C^k(\mathbb{M}, \mathbb{N})$ como las funciones con k derivadas continuas, entonces $C^k(\mathbb{M}, \mathbb{N})$ puede ser dotado de una topología de tal forma que dos funciones están cerca si y solo si las funciones y sus derivadas están cerca en subconjuntos compactos.

El teorema de taken dice lo siguiente:

Teorema 4. *Sea $\mathbb{M}$ una variedad sub-riemanniana suave y compacta, sea $\tau > 0$ y $d \geq \dim(\mathbb{M})$. Entonces para funciones 'genéricas' $\Phi \in C^2(\mathbb{R} \times \mathbb{M}, \mathbb{M})$ y $F \in C^2(\mathbb{M}, \mathbb{R})$ y para ϕ_p definida como antes, el siguiente mapeo:*

$$\begin{aligned} \phi : \mathbb{M} &\longrightarrow \mathbb{R}_{d+1} \\ p &\mapsto (\varphi_p(0), \varphi_p(\tau), \ldots, \varphi_p(d\tau)) \end{aligned}$$

es un encaje.

Más aún, si $\mathbb{A} \subset \mathbb{M}$ es un atractor extraño, entonces ϕ restringido a A será (genéricamente) un encaje siempre que d es al menos el doble de la dimensión de Hausdorff de A.

Definición 20. *Sea $f : \mathbb{R} \longrightarrow \mathbb{R}$ una función y $\tau > 0$ y $d \in \mathbb{N}$. La ventana deslizante de encaje de f, con parámetros d y τ es la función:*

$$\begin{aligned} \mathbb{SW}_{d,\tau} f : \mathbb{R} &\longrightarrow \mathbb{R}^{d+1} \\ t &\mapsto (f(t), f(t+\tau), \ldots, f(t+d\tau)) \end{aligned}$$

El entero $d+1$ es la dimensión, τ es el retraso y el producto $d\tau$ s el tamaño de la ventana. Para $T \subset \mathbb{R}$, el conjunto

$$\mathbb{SW}_{d,\tau} f = \{\mathbb{SW}_{d,\tau} f(t) : t \in T\}$$

es la nube de puntos de la ventana deslizante asociada al muestreo del conjunto T

Por lo tanto, dados los datos de series de tiempo $f(t) = \phi_p(t)$ observado de un sistema dinámico potencialmente desconocido $(\mathbb{M}, \Phi)$, el teorema de Takens implica que (genéricamente) la nube de puntos de la ventana deslizante $\mathbb{SW}_{d,\tau} f$ proporciona una copia topológica de

$$\{\Phi(t, p) : t \in T\} \subseteq \mathbb{M}.$$

En particular, esto reconstruirá atractores. La forma subyacente de $\mathbb{SW}_{d,\tau} f$ puede entonces ser cuantificado con homología persistente, y el asociado códigos de barras y se pueden utilizar como características en tareas de inferencia, clasificación y aprendizaje.