Una introducción al análisis topológico de datos [*]

Natalia García-Colín [**]

Resumen

Uno de las nuevas técnicas desarrolladas para el ánalisis de grandes cúmulos de información (*Big Data*) es el Análisis Topológico de Datos. Este se ha desarrollado con el propósito de inferir información de un sistema de datos a partir de muestras representadas como un espacio *topológico combinatorio*. En esta comunicación se presenta una introducción a algunas de las técnicas del análisis topológico de datos.

1. Introducción

La cantidad de datos recaudada por instituciones públicas y privadas ha explotado en los últimos 15 años gracias a la creciente cobertura de las redes de internet y la disminución del costo de almacenamiento de información [8]. En el año 2000 se almacenaron a nivel mundial 800 mil petabytes (PB) de datos y esta cantidad aumenta constantemente. En la actualidad, por ejemplo, Twitter genera siete terabytes (7 TB) de datos diariamente, Facebook 10 TB. Se calcula que la cantidad de datos almacenada anualmente alcance 35 zettabytes (ZB=un billon de terabytes) para el año 2020 [5].

Se denomina Big Data a un conjunto de información tan grande, complejo y, en la mayoría de casos, sin estructura, que resulta imposible estudiarlo con las herramientas usuales de manejo de base de datos. El estudio del manejo del Big Data incluye retos como la optimización de la captura, almacenamiento, búsqueda, trasferencia, análisis, visualización, etc.

Actualmente en muchísimas ramas de la ciencia y la industria se tiene acceso a bases de datos gigantescas con información cruda de la cual se pueden extraer patrones, relaciones y en un siguiente paso, teorías.

Uno de las nuevas técnicas desarrolladas es el Análisis Topológico de Datos (TDA, por sus siglas en inglés), éste se ha practicado con éxito en los últimos 15 años para estudiar como se puede inferir información de un sistema de datos a partir de muestras representadas como un espacio *topológico combinatorio*. En el TDA se construyen complejos simpliciales asociados a los datos y se infieren características cualitativas del conjunto a partir de la homología de dicho complejo [4].

El propósito de esta comunicación es presentar una introducción a algunas de las técnicas y resultados del análisis topológico de datos.

2. Preliminares

Usualmente, los datos recabados se pueden representar como nubes de puntos en $\mathbb{R}^d$, donde la dimensión es el número de parámetros que se está estudiando. Un modelo popular que ha surgido para representar dicho conjunto es el de las gráficas geométricas aleatorias y complejos geométricos aleatorios. [3]

Una gráfica geométrica aleatoria $\mathcal{G}(n, r)$ se construye escogiendo n puntos de manera independiente y distribuidos identicamente (i.i.d), de acuerdo con una medida de probabilidad en $\mathbb{R}^d$. Estos puntos corresponderán a los vértices de una gráfica. Dos vértices x y y se conectan por una arista si y sólo si la distancia entre x y y satisface $d(x, y) < r$. Normalmente se está interesado en las propiedades asintóticas de estas gráficas cuando $n \to \infty$; es por esto que se piensa a la distancia r como una función de n.

[*]Trabajo realizado con apoyo de los proyectos CONACyT y PAPIIT IN

[**]INFOTEC Centro de Investigación e Innovación en Tecnologias de la Información y Comunicación, natalia.garcia@infotec.com.mx

Algunas de las formas naturales de extender el concepto de una gráfica geométrica a un complejo simplicial son el complejo de Čech y el complejo de Vietoris-Rips.

Definición 1 *[Complejo de Čech] Sea $X = \{x_1, \dots x_n\}$ una colección de puntos en $\mathbb{R}^d$, y sea $r > 0$. El complejo de Čech de X, $\mathcal{C}(X, r)$, es aquel que tiene como vértices a los puntos de X y cuyos k-simplejos $\{x_{i_0} \dots x_{i_k}\}$ son aquellos tales que $\cap_{j=0}^{k} B_{\frac{r}{2}}(x_{i_j}) \neq \emptyset$.*

Definición 2 *[Complejo de Vietoris-Rips] Sea $X = \{x_1, \dots x_n\}$ una colección de puntos en $\mathbb{R}^d$, y sea $r > 0$. El complejo de Vietoris-Rips de X, $\mathcal{R}(X, r)$, es aquel que tiene como vértices a los puntos de X y cuyos k-simplejos $\{x_{i_0} \dots x_{i_k}\}$ son aquellos tales que $\|x_{i_j} - x_{i_l}\| \leq r$ para toda pareja $1 \leq j, l \leq k$.*

A continuación, se usarán $\mathcal{C}(n, r)$ y $\mathcal{R}(n, r)$ para denotar a los complejos de Čech y de Vietoris-Rips, respectivamente, generados aleatoriamente por conjuntos de puntos distribuidos independiente e idénticamente (i.i.d) en $\mathbb{R}^d$, con función de densidad f, medible y acotada.

Un problema de particular interés dentro del contexto del TDA es estudiar la homología de los complejos de Čech y Vietoris-Rips.

Recuérdese que dado un espacio topológico T, su i-homología, denotada $H_i(T)$ es un espacio vectorial, que $\dim H_0(T)$ indica el número de componentes conexas del espacio, que cuando $i > 1$, $H_i(T)$ da información sobre los hoyos de tamaño i y que se le llama $\dim H_i(T) = \beta_i(T)$ al i–ésimo número de Betti.

Nótese que la conectividad de un complejo simplicial depende únicamente de su 1-esqueleto, es decir de su gráfica subyacente. En el caso de los complejos de Čech y Vietoris-Rips dicha gráfica es, en ambos casos, precisamente la gráfica geométrica aleatoria de sus vértices generadores.

3. Gráficas geométricas aleatorias

La conectividad de las gráficas aleatorias se ha estudiado ampliamente [9]. Aquí presentamos algunos resultados que conciernen precisamente a la conectividad de dichas gráficas, es decir, resultados sobre la 0-homología de los complejos de Čech y Vietoris-Rips.

Teorema 1 *Si $nr^d \to 0$ entonces $\mathbb{E}[\beta_0(n, r)] \approx n$*

Teorema 2 *Si $nr^d \to \lambda \in (0, \infty)$ entonces $\mathbb{E}[\beta_0(n, r)] \approx Cn$ para una constante $C(\lambda) < 1$.*

Teorema 3 *Sea $c \in \mathbb{R}$ un número fijo, y sea $r = (\frac{\log n + c}{\omega_d n})^{\frac{1}{d}}$. Entonces $\mathbb{P}(\mathcal{G}(n, r)$ sea conexa $) \to e^{-e^{-c}}$ cuando $n \to \infty$.*

Teorema 4 *Para $d = 2$ existe $C > 0$, tal que si $A \leq nr^2 \leq B \log n$, entonces asintóticamente casi seguramente (a. a. s.) $\beta_0(n, r) \leq \frac{1}{r^2} e^{-Cnr^2}$, donde las constantes A y B solamente dependen de la función densidad f.*

4. Complejos gráficos aleatorios

Los números de Betti de los complejos geométricos aleatorios fueron estudiados en sus inicios por Robins [10] y después por [1, 2, 6, 7]. A diferencia de la conectividad en gráficas, que corresponde a la homología cero, en general $H_i(\mathcal{C}(n, r))$, con $i \geq 1$ no es monótona en respecto a r. A continuación presentamos algunos resultados recientes en este respecto.

Teorema 5 *Sea $nr^d \to 0$, $i \geq 1$ y $d \geq 2$. Entonces $\mathbb{E}[\beta_i(n, r)] \sim n^{i+2} r^{(i+1)d}$.*

Teorema 6 *Sea $d \geq 2$ y $1 \leq i \leq d - 1$ fijo. Supóngase que $nr^d \to 0$.*

- *si $nr^d \ll n^{\frac{-1}{i+1}}$, entonces a.a.s. $H_i(\mathcal{C}(n, r)) = 0$ y*

- *si* $nr^d \gg n^{\frac{-1}{i+1}}$, *entonces* a.a.s. $H_i(\mathcal{C}(n,r)) \neq 0$.

Teorema 7 *Sea* $d \geq 2$ *y* $1 \leq i$ *fijo. Supóngase que* $nr^d \to 0$.

- *si* $nr^d \ll n^{\frac{-1}{2i+1}}$, *entonces* a.a.s. $H_i(\mathcal{R}(n,r)) = 0$ *y*
- *si* $nr^d \gg n^{\frac{-1}{2i+1}}$, *entonces* a.a.s. $H_i(\mathcal{R}(n,r)) \neq 0$.

El estudio de los números de Betti es mucho más complicado cuando $nr^d \to \lambda \in (0,\infty)$.

Teorema 8 *Sea* $d \geq 2$ *y* $0 \leq i \leq d-1$ *fijo y* $nr^d \to \lambda \in (0,\infty)$. *Entonces* $\mathbb{E}[\beta_i(n,r)] \sim n$.

En el caso cuando $nr^d \to \infty$ el orden de magnitud correcto de los números de Betti no es conocido, pero existen cotas. En particular se tienen los siguientes resultados:

Teorema 9 *Sea* $\mathcal{R}(n,r)$ *complejo de Vietoris-Rips aleatorio generado por una distribución uniforme en un convexo de volumen unitario en* $\mathbb{R}^d$. *Entonces,* $\mathbb{E}[\beta_i(n,r)] = O(ne^{-c_d nr^d}(nr^d)^i)$, *para una constante independiente de* i, $c_d \geq 0$.

Corolario 10 *Sea* $C > \frac{1}{c_d}$ *una constante. Si* $nr^d \geq C \log n$ *entonces* a.a.s $H_k(\mathcal{R}(n,r)) = 0$.

Teorema 11 *Sea* $d \geq 2$ *fijo, y supóngase que se tiene una distribución subyacente uniforme sobre un convexo. Entonces existen* A, B *tales que:*

- *si* $nr^d \ll \frac{1}{n^{\frac{1}{2i+1}}}$, *entonces* a.a.s. $H_i(\mathcal{R}(n,r)) = 0$,
- *si* $\frac{1}{n^{\frac{1}{2i+1}}} \ll nr^d \leq A \log n$, *entonces* a.a.s. $H_i(\mathcal{R}(n,r)) \neq 0$,
- *y si* $nr^d \geq B \log n$, *entonces* a.a.s. $H_i(\mathcal{R}(n,r)) = 0$,

En el caso del complejo de Čech aleatorio no se conoce una cota superior para los números de Betti, pero sí se conoce el orden de magnitud para el desvanecimiento de la homología:

Teorema 12 *Sea* $d \geq 2$ *y* $1 \leq i \leq d-1$ *fijo, y supóngase que se tiene una distribución subyacente uniforme sobre un convexo. Entonces existen* A, B *tales que:*

- *si* $nr^d \ll \frac{1}{n^{\frac{1}{i+1}}}$, *entonces* a.a.s. $H_i(\mathcal{C}(n,r)) = 0$,
- *si* $\frac{1}{n^{\frac{1}{i+1}}} \ll nr^d \leq A \log n$, *entonces* a.a.s. $H_i(\mathcal{C}(n,r)) \neq 0$,
- *y si* $nr^d \geq B \log n$, *entonces* a.a.s. $H_i(\mathcal{C}(n,r)) = 0$,

5. Observaciones finales y conclusiones

Existen generalizaciones de los teoremas anteriores a casos donde los puntos se encuentran sobre variedades o superficies de Riemman.

Adicionalmente también dentro del TDA hay otras técnicas que se estudian, como la homología persistente, aplicaciones de la teoría de Morse, etc.

Entre las direcciones poco exploradas e interesantes se encuentra la teoría extremal para hipergráficas aplicada al análisis de datos, además de los aspectos algorítmicos derivados de la teoría existente.

Referencias

[1] Bobrowski, O., Adler, R.J.: Distance functions, critical points, and the topology of random Čech complexes, arXiv:1107.4775 (2011),

[2] Bobrowski, O., Mukherjee, S.: The topology of probability distributions on manifolds, to appear., Probability theory and related fields (2014).

[3] Bobrowski, O., Kahle, M.: Topology of random geometric complexes: a survey, arXiv:1409.4734v1 (2014).

[4] Carlsson, G.: Topology and Data., Bulletin of the American Mathematical Society no. 46 (2009) 255–308.

[5] 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.

[6] Kahle, M.: Random geometric complexes., Discrete and Computational Geometry 45 no. 3 (2011).

[7] Kahle, M. Meckes, E.: Linit theorems for Betti numbers of random simplicial complexes., Homology and homotopy Appl. 15 no. 1 (2013).

[8] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung-Byers, A.: Big data: The next frontier for innovation, competition, and productivity, `http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation`, 2011.

[9] Penrose, M.: Random Geometric Graphs. Oxford Studies in Probability, vol. 5, Oxford University Press, UK 2003.

[10] Robins,V.: Betti number signatures of homogeneous poisson point processes, Physical Review E, 74 no. 6 2006,