Proceso para la Obtención de Patrones de Co-localización con Relaciones Difusas

Giovanni Daián Rottoli[1,2] and Hernán Merlino[2]

[1] Programa de Doctorado en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata. La Plata. Argentina.
[2] Grupo de Investigación en Sistemas de Información (GISI). Universidad Nacional de Lanús. Lanús. Argentina.
rottolig@frcu.utn.edu.ar, hmerlino@gmail.com

Resumen La detección de patrones de co-localización es una de las actividades más demandadas de la ciencia de datos espaciales. Sin embargo, la gran variedad de algoritmos especializados en distintos aspectos de esta tarea y la falta de disponibilidad de los mismos en las herramientas de minería de datos disponibles hace que su utilización se dificulte. Por este motivo se diseña un framework que hace uso de teoría de grafos y lógica difusa para la extracción de estas regularidades considerando distintos aspectos fundamentales de la minería de datos espaciales utilizando herramientas fácilmente asequibles.

Keywords: Patrones de co-localización · Minería de datos espaciales · Lógica difusa · Relaciones espaciales

1. Introducción

En minería de datos espaciales, un patrón de co-localización es un tipo de regularidad de interés para la inteligencia de negocio que consiste en un subconjunto de objetos u eventos espaciales ubicados frecuentemente de manera próxima entre sí. Esta regularidad es utilizada comúnmente para determinar comportamientos delictivos posiblemente relacionados con características urbanas, para la búsqueda de simbiosis entre especies animales, o para la detección de minerales raros por asociación con particularidades del terreno [1–4].

Existe una amplia variedad de trabajos relacionados con la búsqueda de patrones de co-localización, entre los que se destacan los aportes y algoritmos de Shekhar y Huang (2001), Huang et al. (2004), Yoo y Shekhar (2006), Wang et al. (2008, 2009), Yao et al. (2017), entre otros [1, 5–9], incluyendo además acercamientos que tienen en cuenta la localidad de los patrones por el fenómeno de la heterogeneidad espacial [10], y trabajos que contemplan objetos y relaciones espaciales vagas mediante el uso de lógica difusa, disminuyendo así la rigidez de los modelos tradicionales [2, 11].

Esta variedad de métodos permite la búsqueda de patrones de co-localización ante una gran variedad de escenarios, pero dificulta la selección de las técnicas de modelado. Además, el gran costo computacional al que se incurre hace que sea

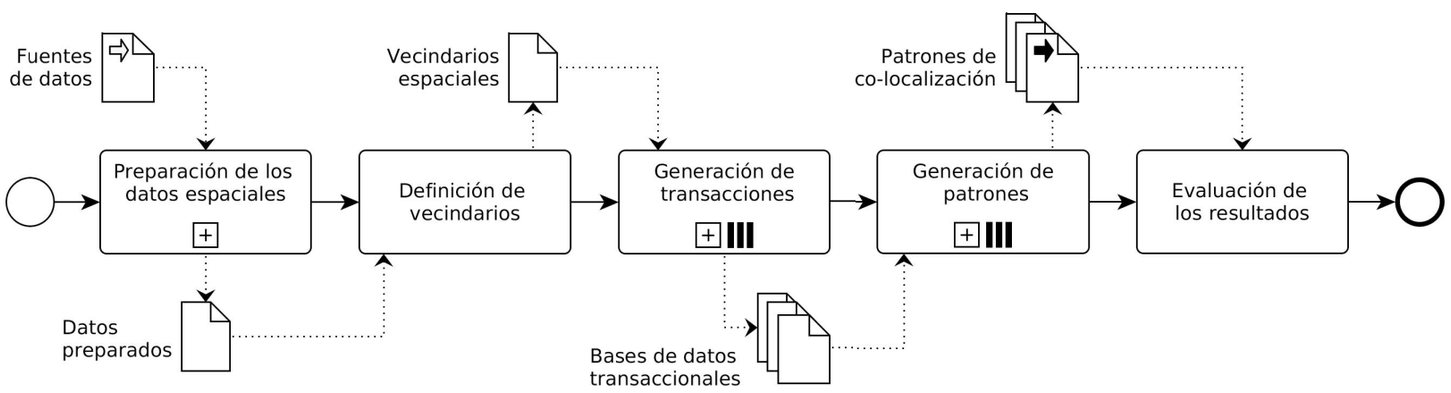

Figura 1. Proceso para el descubrimiento de patrones de co-localización

necesario un proceso que permita la generación de modelos prototipos de patrones de co-localización de una forma más dinámica lo cual a su vez también se ve limitado por la baja disponibilidad de herramientas software que implementen estas opciones.

En este trabajo, entonces, se propone la utilización de un enfoque transaccional para la obtención de patrones de co-localización que utiliza tecnologías de minería de datos relacionales tradicionales para la obtención de patrones de co-localización regionales incorporando también la evaluación difusa de los mismos.

2. Solución propuesta

El proceso propuesto consiste en cinco actividades conceptuales dispuestas en forma secuencial, tal como se puede observar en la figura 1.

Primeramente, el proceso inicia con la preparación de los datos espaciales provistos de entrada, generalmente en formatos vectoriales o capas *shapefile* de forma tal que conste un único registro con los atributos espaciales necesarios para especificar la ubicación de los objetos (generalmente puntos o regiones) y un atributo nominal cuyo valor indique el tipo de objeto o evento espacial en cuestión.

La localidad de los patrones de co-localización, necesaria debido al fenómeno de heterogeneidad espacial que establece que las regularidades detectadas en datos espaciales pueden tener un alto soporte en ciertas regiones y no en otras, se ve asegurada por la definición de zonas de interés para los tomadores de decisiones en la actividad siguiente. Estas zonas o vecindarios son especificados de acuerdo a las características del problema, según las necesidades de información de los interesados. En caso de que no se posean requerimientos de este tipo, algoritmos de segmentación espacial basados en densidad como DBSCAN, debido a la primera ley de la geografía que establece que existe una correlación mayor entre objetos ubicados próximamente en el espacio que entre aquellos alejados [12].

Posteriormente, la generación de transacciones se realiza por cada uno de estos vecindarios. Para esto se calculan los grados de pertenencia de cada par

de instancias a la relación difusa de vecindad definida según el dominio del problema, se registran estos valores y se crea un grafo de vecindad a partir del soporte de la relación difusa, esto es, todas las relaciones con $\mu > 0$. Por ejemplo, la relación puede definirse como se muestra en la ecuación 1.

$$\mu_{vec}(x,y) = \begin{cases} 1 & \text{si } dist(x,y) < 10 \\ (\frac{-dist(x,y)}{5} + 3)^2 & \text{si } 10 \leq dist(x,y) \leq 15 \\ 0 & \text{si } dist(x,y) > 15 \end{cases} \quad (1)$$

A partir de este grafo se extraen los cliques máximos, esto es, subgrafos máximos completos K_n que representan subconjunto de instancias de datos espaciales que son todas vecinas entre sí en algún grado. Cada uno de estos cliques se traduce luego en una transacción consistente en una tupla $T = (t_1, t_2, ..., t_m)$ siendo m el número de tipos de objetos espaciales considerados, y donde $t_i = 1$ si el clique contiene una instancia del tipo de objeto espacial i y $t_i = 0$ en caso contrario.

Posteriormente, la extracción de patrones de co-localización puede realizarse de dos maneras: mediante algoritmos de búsqueda de asociaciones, tales como el algoritmo a-priori, disponibles en la mayoría de las suite de minería de datos como Rapidminer o lenguajes de programación como Python, o mediante el uso de algoritmos de clasificación basados en árboles de decisión utilizando uno de los atributos, el que resulta de interés para la inteligencia de negocio, como atributo objetivo. De esta forma se obtiene patrones de asociación que se corresponden a patrones de co-localización, o reglas de comportamiento que indican las condiciones que deben darse en un vecindario para que ocurra un tipo de objeto espacial dado.

Por último, es necesario evaluar los grados de pertenencia de las relaciones que ocurren en las instancias de cada uno de estos patrones, primero identificándolas y luego obteniendo el mínimo grado de pertenencia de la relación de vecindad dentro de cada vecindario. Este valor describe la distancia más lejana en el vecindario. Por último, se describen las medidas de tendencia central y medidas de dispersión de estos valores mínimos. Adicionalmente, se filtran las reglas que no satisfacen las métricas mínimas de calidad de los patrones como precisión o confianza.

3. Discusión de resultados

El proceso fue probado en diferentes escenarios sintéticos generados de forma tal que haya patrones de co-localización entre diferentes tipos de datos. El método propuesto ha sido eficaz en la detección de los mismos. La representación difusa de las relaciones espaciales ha modelado satisfactoriamente el escenario forzado en cada caso.

No obstante, aún queda pendiente la realización de dos experimentos: el primero con el propósito de determinar si el método propuesto presenta diferencia en cuanto a las respuestas obtenidas en relación a los métodos complejos del estado del arte. Para esto se plantea la implementación del método propuesto

y de los algoritmos seleccionados para la comparativa y la ejecución de los mismos sobre diferentes conjuntos de datos, apareando los resultados. Se buscará determinar si existen diferencias significativas en la cantidad de patrones descubiertos utilizando el una prueba de hipótesis no paramétrica como lo es la prueba de rangos con signo de Willcoxon. Asimismo, se busca validar la facilidad de implementación y la flexibilidad de esta alternativa mediante juicio de expertos.

Agradecimientos

La investigación desarrollada en este artículo se encuentra financiada parcialmente por el Programa de Becas Doctorales para el Fortalecimiento de Áreas I+D+i (2016-2020) de la Universidad Tecnológica Nacional y por el Proyecto de Investigación 80020160400001LA de la Universidad Nacional de Lanús.

Referencias

1. Shekhar, Shashi, and Yan Huang. "Discovering spatial co-location patterns: A summary of results." International symposium on spatial and temporal databases. Springer, Berlin, Heidelberg, 2001. https://doi.org/10.1007/3-540-47724-1_13
2. Ouyang, Zhiping, Lizhen Wang, and Pingping Wu. "Spatial co-location pattern discovery from fuzzy objects." International Journal on Artificial Intelligence Tools 26.02 (2017): 1750003. https://doi.org/10.1142/S0218213017500038
3. Yue, Han, et al. "The local colocation patterns of crime and land-use features in Wuhan, China." ISPRS International Journal of Geo-Information 6.10 (2017): 307.
4. Rottoli, Giovanni Daián, Hernán Merlino, and Ramón García-Martinez. "Co-location Rules Discovery Process Focused on Reference Spatial Features Using Decision Tree Learning." IEA-AIE. 2017. https://doi.org/10.1007/978-3-319-60042-0_25
5. Huang, Yan, Shashi Shekhar, and Hui Xiong. "Discovering colocation patterns from spatial data sets: a general approach." IEEE Transactions on Knowledge and data engineering 16.12 (2004): 1472-1485.
6. Yoo, Jin Soung, and Shashi Shekhar. "A joinless approach for mining spatial co-location patterns." IEEE Transactions on Knowledge and Data Engineering 18.10 (2006): 1323-1337.
7. Wang, Lizhen, et al. "A new join-less approach for co-location pattern mining." 2008 8th IEEE International Conference on Computer and Information Technology. IEEE, 2008.
8. Wang, Lizhen, et al. "An order-clique-based approach for mining maximal co-locations." Information Sciences 179.19 (2009): 3370-3382.
9. Yao, Xiaojing, et al. .A co-location pattern-mining algorithm with a density-weighted distance thresholding consideration.Ïnformation Sciences 396 (2017): 144-161.
10. Li, Yan, and Shashi Shekhar. "Local co-location pattern detection: a summary of results." 10th International Conference on Geographic Information Science (GIScience 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
11. Akbari, Mohammad, and Farhad Samadzadegan. "New regional co-location pattern mining method using fuzzy definition of neighborhood..Advances in Computer Science: an International Journal 3.3 (2014): 32-37.
12. Jiang, Zhe, and Shashi Shekhar. "Spatial big data science."Schweiz: Springer International Publishing AG (2017).