ARTÍCULO DE REVISIÓN

La minería de datos espaciales y su aplicación en los estudios de salud y epidemiología

Spatial data mining and its application in health and epidemiology studies

Ing. Liset González Polanco, Ing. Yadian Guillermo Pérez Betancourt

Universidad de las Ciencias Informáticas. La Habana, Cuba.

RESUMEN

La acumulación de información espacial producto del desarrollo de los sistemas informáticos, y en especial de los sistemas de información geográfica, propicia la aplicación de técnicas de minería de datos espaciales para la extracción de nuevos conocimientos que asistan a la toma de decisiones. Las áreas de salud y epidemiologia no han estado ajenas al desarrollo y utilización de estos sistemas; han revalorizado la importancia de la componente espacial en sus investigaciones y en el diseño de estrategias diferenciadas de prevención y control por área de salud. En este trabajo se describen los aspectos metodológicos y los conceptos asociados a la minería de datos espaciales. Se describen los principales algoritmos y herramientas existentes para la minería de datos espaciales y se muestran algunos trabajos, tendencias de su aplicación y potencialidades en las áreas de salud y epidemiología.

Palabras clave: análisis espacial, epidemiología, salud, sistemas de información geográfica.

ABSTRACT

The accumulation of spatial data resulting from the development of information systems, especially geographic information systems, has paved the way for the application of spatial data mining techniques for the extraction of new knowledge which could in turn assist in decision making. Health and epidemiology areas have not been alien to the development and use of these systems, revalidating the importance of the spatial component both in research and in the design of differentiated prevention and control strategies for each health area. The paper presents the methodological aspects and concepts associated with spatial data mining. A description is provided of the main algorithms and tools used in spatial data mining, and some experiences are presented which illustrate the application trends and potential of this technique in health and epidemiology areas.

Key words: spatial analysis, epidemiology, health, geographic information systems.

INTRODUCCIÓN

La aplicación de estrategias diferenciadas para la prevención y control sanitario es una prioridad para el sector de la salud, y muestra de esto es el gran número de trabajos relacionados con la temática. La salud pública no ha estado alejada del creciente desarrollo tecnológico e incorpora a su quehacer los resultados de otras áreas, entre ellas la informática.

El desarrollo de la informática y de los sistemas informáticos ha propiciado una transformación importante en esta área. Se recogen datos de enfermedades, la distribución de los servicios y la utilización de la geoinformática en la vigilancia, y se desarrollan las bases cartográficas que facilitan los estudios.

Los estudios epidemiológicos tienen el objetivo de determinar las relaciones persona-espacio-tiempo.[1] Si bien el espacio es un componente importante en los estudios sobre la salud,[2] no siempre se le da la importancia requerida, motivado por: 1) el acceso limitado a los sistemas de información geográfica (GIS) por los costos que ellos implican, 2) el poco conocimiento de las herramientas y 3) el tiempo de formación en el área de los GIS, por ser elevado.

Determinar los riesgos existentes y presentar cartográficamente las áreas expuestas a estos es parte importante de la vigilancia epidemiológica,[3] que se favorece con el empleo de los GIS. Muchas son las posibilidades aportadas por los GIS al sector de la salud en el fortalecimiento de las capacidades de análisis, gestión, monitoreo y toma de decisiones.

Los GIS en salud pública son utilizados en el análisis de la situación de salud, la vigilancia de eventos, el estudio epidemiológico, la planeación y la evaluación de estrategias por zonas de salud, así como en la gestión y toma de decisiones.

El desarrollo de los GIS y su aplicación en diferentes áreas ha brindado la posibilidad de analizar grandes volúmenes de datos espaciales. Aunque los GIS están creados para manipular datos espaciales, se demanda el uso de técnicas que permitan extraer conocimiento de estos datos acumulados en bases de datos espaciales (SDBMS) y el descubrimiento de patrones que sean más fáciles de entender. Inicialmente se pudiera pensar en la extracción de conocimiento automatizada mediante la minería de datos, que permite encontrar conocimiento implícito[4] en grandes volúmenes de datos. Sin embargo, producto de la complejidad de los tipos de datos que se manejan en SDBMS, y los objetos que se almacenan (puntos, líneas, polígonos y las estructuras de datos utilizadas) se dificulta la utilización de aproximaciones tradicionales de la minería de datos. La minería de datos espaciales provee un grupo de técnicas y herramientas para la explotación de estos datos que permiten encontrar patrones potencialmente útiles.

El interés de la salud pública y la epidemiologia en el estudio y análisis de la distribución geográfica de las enfermedades, su relación con los riesgos potenciales y el desarrollo de herramientas que permiten el manejo de datos epidemiológicos con su componente espacial, ha impulsado considerablemente el desarrollo de métodos estadísticos para ambas disciplinas, y como resultado un mejor desarrollo de planes preventivos.[1] Bajo estas circunstancias, la utilización de la minería de datos espaciales presenta muchas potencialidades para estudios epidemiológicos y de salud.

El descubrimiento de patrones relacionados con desigualdades socioeconómicas, los eventos epidemiológicos, los focos de contaminación ambiental y los factores de

riesgos permiten identificar las regiones donde hay que prestar especial atención en la vigilancia epidemiológica y adecuar más los planes de prevención.

En este trabajo se describen los aspectos metodológicos y los conceptos asociados a la minería de datos espaciales. Se describen los principales algoritmos y herramientas existentes y se muestran algunos trabajos, tendencias de su aplicación y potencialidades en las áreas de salud y epidemiologia.

La minería de datos es un proceso de búsqueda de información relevante en grandes volúmenes de datos, semejante a la que podría realizar un experto humano.[4,5] La amplia difusión de información espacial producto del desarrollo de los GIS ha favorecido la explotación de los datos con el objetivo de encontrar conocimiento de manera automatizada. La complejidad de los tipos de datos existentes en SDBMS[6] y las estructuras de datos que las soportan limitan la utilización de técnicas tradicionales de minería de datos, lo que propicia la aparición de nuevas técnicas que de conjunto forman la minería de datos espaciales.

MINERÍA DE DATOS ESPACIALES

La minería de datos espaciales posee la base teórica y metodológica para la identificación de patrones sobre los datos[7] y tiene como objetivo descubrir de forma automatizada patrones inesperados potencialmente útiles en SDBMS y que serán validados por expertos del área en cuestión. Se puede definir como el proceso automático o semiautomático[8] de seleccionar, explorar, modificar, visualizar y valorar grandes volúmenes de datos espaciales con el objetivo de descubrir conocimientos. La minería de datos espaciales es considerada una rama de la minería de datos[9] con la característica de extraer conocimiento referente a la naturaleza espacial de los datos.

El descubrimiento de conocimiento o patrones en bases de datos espaciales a través de la minería de datos espaciales es más complejo, pues no solo se encarga de los datos no espaciales, sino que además tiene en cuenta la localización de los objetos y sus relaciones topológicas. En este proceso se utilizan métodos basados en la generalización, en el reconocimiento de patrones, de agrupamiento: de exploración de asociaciones espaciales y mediante el uso de aproximación y agregación. A continuación se describen los métodos utilizados en la minería de datos espaciales:

- *Basados en la generalización:* requieren de la implementación de jerarquías de conceptos, bien temática o espacial. Dentro de las temáticas se incluyen los datos no espaciales; de ellos se colectan sus características más importantes para la búsqueda, se caracterizan por regiones y se agrupan como datos no espaciales generalizados. Para el caso de los espaciales esta generalización puede ser presentada como la partición en regiones y su posterior fusión dependiendo de los atributos espaciales de los datos.

- *Basados en el reconocimiento de patrones:* son utilizados en la clasificación de información que pueden ser imágenes de satélites, fotografías, textos o cualquier fuente de datos:

- *De agrupamiento:* permiten agrupar los objetos de una base de datos en grupos llamados conglomerados, conformados por elementos tan similares como sea posible.

- *De exploración de asociaciones espaciales:* permiten descubrir reglas de asociación espacial que relacionen a uno o más objetos espaciales.

- *Mediante el uso de aproximación y agregación:* permiten descubrir conocimiento a partir de las características representativas de los objetos.

ALGORITMOS DE MINERÍA DE DATOS ESPACIALES

Los algoritmos de minería de datos espaciales deben operar sobre conjuntos de datos de tamaño considerable,[10] por lo que se debe trabajar en propuestas donde el conjunto de datos completo no resida en la memoria principal. Deben hacer un correcto uso de las técnicas de optimización de búsquedas espaciales y del razonamiento espacial y realizar su tarea de forma eficiente y rápida. A continuación se describen algunos de los algoritmos más utilizados de la minería de datos espaciales:

- *CLARANS:* consiste en la búsqueda aleatoria de un grupo de datos. Tiene complejidad temporal de $O(n^2)$. Producto de la importancia de los datos espaciales, este algoritmo se deriva del SD CLARANS, que busca descubrir características no espaciales en grupos espaciales, y del NSD CLARANS para descubrir conglomerados espaciales en grupos de datos no espaciales.[11]

- *DBSCAN:* este algoritmo pertenece a la familia de algoritmos de conglomeración espacial.[12] Aborda la integración entre la minería de datos espaciales y la interfaz con el sistema de bases de datos espaciales. No todos los datos deben permanecer en memoria principal y tienen un orden de ejecución de $O(logn)$. Este algoritmo se basa en los conceptos de conglomerado, alcance directo por densidad, alcance por densidad y conexión por densidad.[13]

- *ST-DBSCAN:* es un algoritmo de agrupamiento por densidad y basa su funcionamiento en el DBSCAN. Tiene la característica de descubrir grupos de acuerdo con valores no espaciales y espacio-temporales de los objetos. Tiene complejidad temporal $O(n^3)$.

- *PDBSCAN:* es un algoritmo de conglomeración paralelizable,[14,15] que se basa en DBSCAN; utiliza una estructura de datos distribuida, basada en árboles de tipo R(dR*-tree) y curvas de Hilbert para encontrar puntos pertenecientes a los diferentes conglomerados en el momento de la partición del problema, pues permite que puntos espaciales cercanos se encuentren en la misma partición siempre que sea posible. Logra disminuir los tiempos de ejecución de algoritmos, como CLARANS, que tienen orden de ejecución cuadrático. Permite que el problema de la búsqueda de conglomerados en un conjunto de datos de gran tamaño sea paralerizable y tenga un tiempo de ejecución de $O(logn)$.[16]

SISTEMAS INFORMÁTICOS CON SOPORTE PARA LA MINERÍA DE DATOS ESPACIALES

La aplicación de la minería de datos espaciales busca resolver diversos problemas mediante el descubrimiento de conocimiento, aplicando diversas técnicas donde los objetos espaciales cuentan además con características no espaciales y sirven de entrada a algoritmos de minería. En aplicaciones desarrolladas para este fin se ha incluido también el uso de otras técnicas de la inteligencia artificial como las redes bayesianas y árboles de decisión. A continuación se describen algunos sistemas

informáticos con soporte para realizar la minería de datos espaciales o con base en dichos procesos para obtener resultados específicos:

- *GeoDMA:* es un complemento para la minería de datos del sistema informático TerraView. Soporta el uso de datos espaciales para la comparación de imágenes y regiones obtenidas en los procesos de segmentación y análisis de imágenes. Optimiza la segmentación de imágenes y la extracción de características, incluyendo atributos de selección, clasificación, validación y visualización.[17] Utiliza árboles de decisión y algoritmos para mapas autoorganizados.[18] Se encuentra liberado bajo la licencia de software libre GNU General Public License (GPL), lo que facilita el desarrollo de nuevas funcionalidades. Su implementación es en lenguaje C++ e interfaz en QT.

- *SDMiner:* tiene soporte para las principales técnicas de minería de datos espaciales, como agrupamiento, clasificación espacial, caracterización espacial y espacio-temporal y reglas de asociación espacial. Posee gran facilidad para el uso de datos espaciales y no espaciales con la característica de determinar su naturaleza. Los parámetros de entrada para la minería son en forma de tablas de la base de datos, lo que aumenta su flexibilidad. Implementa sus algoritmos en una librería que permite que sean utilizados por otros sistemas y, a su vez, incorporar nuevas funcionalidades.[19]

- *SaTScan:* es gratuito y se utiliza para realizar análisis estadístico espacial, implementa la técnica de detección de conglomerados de Kulldorff[20] para la detención de conglomerados espaciales, temporales, espacio-temporales y prospectivos.[21] Se pueden utilizar varios modelos estadísticos, entre ellos la distribución de Poisson o Bernoulli. Fue concebido inicialmente para su uso en el área de la salud en estudios sobre la distribución espacial de las enfermedades y epidemias. Si bien su mayor utilización radica en los estudios de sanidad, puede ser utilizado en problemas similares de otros campos para: 1) efectuar vigilancia geográfica de una variable, detectar patrones espaciales o espacio-temporales estadísticamente significativos, 2) probar si una variable tiene distribución aleatoria en el espacio, tiempo o espacio-tiempo, y 3) evaluar valores umbrales alcanzados por agrupaciones espaciales de la variable.[21]

ALGUNOS TRABAJOS RELACIONADOS CON LA MINERÍA DE DATOS ESPACIALES APLICADA A LOS ESTUDIOS DE SALUD

Si bien los enfoques tradicionales de la minería de datos son los más utilizados en los estudios de salud, con la reevaluación de los datos espaciales en estas áreas existe una tendencia a incorporar enfoques de la minería de datos espaciales. Por ejemplo, la aplicación de reglas de asociación espacial[22] para encontrar relaciones entre determinadas variables socioeconómicas y las cuatro causas principales de muerte por cáncer (colorrectal, pulmón, mama y próstata) en los Estados Unidos permitió conocer qué áreas de salud con índices de educación bajo, alto desempleo y trabajos mal pagados tenian las mayores tasas de mortalidad por esta enfermedad.

La utilización de SaTScan en los programas de control de la malaria en la provincia de Mpumalanga, Sudáfrica, para la detención de grupos locales de malaria[21] permitió identificar las zonas de riesgo y facilitó la planificación para el control de la enfermedad. Se detectaron cinco conglomerados espaciales y dos

espaciotemporales con una gran coincidencia entre los conglomerados identificados y las zonas que reportaron casos de malaria. Esto permitió destinar los recursos a áreas específicas a nivel local. En estudios sobre esquistosomiasis[23] también se muestran resultados positivos en la caracterización espacial.

Un trabajo[1] de aplicación de la técnica de conglomerados de Kulldorff, permitió mostrar la diferenciación de los determinantes del síndrome de Down en zonas rurales y urbanas en la provincia de Villa Clara, Cuba. Este resultado permite la implementación de estrategias diferenciadas para la prevención de esta cromosomopatía. En estudio relacionado con el cáncer de mama y el de cérvix[24] también se mostró la presencia de conglomerados en las áreas de mayor riesgo.

PRINCIPALES TENDENCIAS Y PERSPECTIVAS EN LOS ESTUDIOS DE SALUD

La determinación de agrupaciones espaciales de las enfermedades tiene un renovado interés sobre todo en las zonas con menos recursos. La aplicación de diversas técnicas informáticas en diferentes estudios de salud sirve como soporte a la toma de decisiones. En la bibliografía consultada se encuentran trabajos de investigaciones relacionados con la minería de datos espaciales con resultados interesantes; sin embargo, su aplicación en las aéreas de salud y la epidemiología aún es insuficiente. El aumento del uso de los GIS en las áreas de salud facilita la incorporación de estas técnicas a los análisis diarios por parte del personal de la salud.

Los trabajos publicados que utilizan técnicas de minería de datos espaciales en la salud aplican métodos muy específicos para lograr determinados objetivos. Se evidencia un mayor uso de los conglomerados, sobre todo en la caracterización de territorios y la determinación de zonas de riesgos. La incorporación de otras técnicas a estos estudios puede aportar más información, aunque para esto se necesita el desarrollo de herramientas que lo faciliten.

CONCLUSIONES

Con la reevaluación de la información espacial en áreas como la salud pública y la epidemiologia producto del desarrollo y la aplicación de los GIS, la minería de datos espaciales se convierte en una herramienta con muchas potencialidades para sus estudios y proporciona la base teórica y metodológica para la búsqueda de patrones potencialmente útiles que faciliten el diseños de estrategias específicas para cada área de salud.

La minería de datos espaciales provee los mecanismos y herramientas como respuesta a la dificultad de resolver problemas de descubrimiento de conocimiento en bases de datos espaciales con los enfoques tradicionales de la minería de datos.

El descubrimiento de patrones mediante las técnicas de minería de datos espaciales relacionados con desigualdades socioeconómicas, eventos epidemiológicos, focos de contaminación ambiental y factores de riesgos permite identificar las regiones donde hay que prestar especial atención en la vigilancia epidemiológica y adecuar más los planes de prevención.

La mayoría de los trabajos publicados sobre la aplicación de métodos de la minería de datos espaciales en estudios de salud utilizan los basados en agrupamiento. La utilización de otras técnicas o la fusión de varias de ellas pueden arrojar más información para tomar decisiones.

REFERENCIAS BIBLIOGRÁFICAS

1. Alegret Rodríguez M, Herrera M, Grau Abalo R. Las técnicas de estadística espacial en la investigación salubrista: caso síndrome de Down. Rev Cubana Sal Públ [revista en la Internet]. 2008 [citado 13 de agosto de 2013];34(4). Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-34662008000400003&lng=es

2. Barcellos C, Buzai GD. La dimensión espacial de las desigualdades sociales en salud: aspectos de su evolución conceptual y metodológica. Departamento de Ciencias Sociales. Universidad Nacional de Luján: Anuario de la División Geografía; 2006:275-92.

3. Más Bermejo P. Desarrollo, tendencia actual y retos de la Epidemiología en Cuba. Rev Cubana Med Trop. 2011;63:5-6.

4. Orallo JH, Quintana MJR, Ramírez CF. Introducción a la minería de datos: Pearson Prentice Hall; 2004.

5. McDonnell R, de la Fuente Aragón M, McDonnell R, editors. Minería de datos aplicada a la gestión de la información urbanística. Data mining applied to urban information management. 6th International Conference on Industrial Engineering and Industrial Management; 2012.

6. Rigol-Sánchez JP, Chica-Olmo M, Pardo-Igúzquiza E, Rodríguez-Galiano V, Chica-Rivas M. Análisis e integración de datos espaciales en investigación de recursos geológicos mediante sistemas de información geográfica. Bol Soc Geol Mex. 2011;63(1):61-70.

7. Cangrejo Aljure D, Agudelo JG. Minería de datos espaciales Spatial data miningAn overview. Rev Avanc Sist Informát. 2011;8(3):71-7.

8. Dueñas Reyes MX. Minería de datos espaciales en búsqueda de la verdadera información. Ing Univ. 2009:137-56.

9. Han J, Kamber M. Data mining: concepts and techniques. Morgan Kaufmann; 2006.

10. Ester M, Kriegel HP, Sander J. Knowledge discovery in spatial databases. KI-99. Advanc Artif Intellig. 1999:696.

11. Ng RT, Han J. Clarans: a method for clustering objects for spatial data mining. Knowledge and Data Engineering. IEEE Transactions on. 2002;14(5):1003-16.

12. Celik M, Dadaser Celik F, Dokuz A, editors. Anomaly detection in temperature data using DBSCAN algorithm. Innovations in Intelligent Systems and Applications (INISTA). International Symposium. IEEE; 2011.

13. Ester M, Kriegel HP, Sander J. Algorithms and applications for spatial data mining. Geographic Data Mining and Knowledge Discovery. 2001.

14. Olman V, Mao F, Wu H, Xu Y. Parallel clustering algorithm for large data sets with applications in bioinformatics. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2009;6(2):344-52.

15. Danalis A, McCurdy C, Vetter JS. Efficient Quality Threshold Clustering for Parallel Architectures. Parallel & Distributed Processing Symposium (IPDPS) IEEE 26th International; 2012: 1068-79.

16. Xu X, Jäger J, Kriegel HP. A fast parallel clustering algorithm for large spatial databases. High Perform Dat Min. 2002:263-90.

17. Kux HJ, Souza UD. Object-based image analysis of WorldView-2 satellite data for the classification of mangrove areas in city of São Luís. Brazil: An Photogr, Rem Sens Spat Inform Sc. 2012.

18. Korting TS, Fonseca LMG, Escada MIS, da Silva FC, dos Santos Silva MP. GeoDMA - A novel system for spatial data mining. Data Mining Workshops. IEEE International Conference; 2008.

19. Bae DH, Baek JH, Oh HK, Song JW, Kim SW. SD-Miner: A spatial data mining system. Network Infrastructure and Digital Content. IEEE International Conference; 2009.

20. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Statistics in medicine. 1995;14(8):799-810.

21. Coleman M, Mabuza AM, Kok G, Coetzee M, Durrheim DN. Using the SaTScan method to detect local malaria clusters for guiding malaria control programmes. Malar J. 2009;8:68.

22. Vinnakota S, Lam NS. Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach. Internat J Heal Geogr. 2006;5(1):9.

23. Zhao F, Zhu R, Zhang L, Zhang Z, Li Y, He M, et al. Application of satscan in detection of schistosomiasis clusters in marshland and lake areas. Zhongguo xue xi chong bing fang zhi za zhi. Chin J Schistosom Contr. 2011;23(1):28.

24. Hernández NEB, Rodríguez MA, Fleites OA. Análisis espacial de la morbimortalidad del cáncer de mama y cérvix. Villa Clara. Cuba. 2004-2009. Rev Esp Sal Públ. 2013;87:49-57.

Recibido: 19 de marzo de 2013.
Aprobado: 25 de julio de 2013.

Ing. *Liset González Polanco.* Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 1/2, municipio Boyeros, La Habana, Cuba. Correo electrónico: lgpolanco@uci.cu