

Spatial Data Mining

SHASHI SHEKHAR, JAMES KANG,
VIJAY GANDHI

5 University of Minnesota, Minneapolis, MN, USA

Definition

10 Spatial data mining is the process of discovering nontrivial, interesting, and useful patterns in large spatial datasets. The most common spatial pattern families are co-locations, spatial hotspots, spatial outliers, and location predictions.

15 Historical Background

Spatial data mining research began several decades ago when practitioners and researchers noticed that critical assumptions in classical data mining and statistics were violated by spatial datasets. First, whereas classical datasets often assume that data are discrete, spatial data were observed to reside in continuous space. For example, classical data mining and statistical methods may use market-basket datasets (e.g., history of Walmart's transactions), where each item-type in a transaction is discrete. However, "transactions" are not natural in continuous spatial datasets, and decomposing space across transactions leads to loss of information about neighbor relationships between items across transaction boundaries. In addition, spatial data often exhibits heterogeneity (i.e., no places on the Earth are identical), whereas classical data mining techniques often focus on spatially stationary global patterns (i.e., ignoring spatial variations across locations). Finally, one of the common assumptions in classical statistical analysis is that data samples are independently generated. However, this assumption is generally false when analyzing spatial data, because spatial data tends to be highly self-correlated. For example, people with similar characteristics, occupation and background tend to cluster together in the same neighborhoods. In spatial statistics this tendency is called spatial auto-correlation. Ignoring spatial auto-correlation when analyzing data with spatial characteristics may produce hypotheses or models that are inaccurate or inconsistent with the data set. Thus, classical data mining algorithms often perform poorly when applied to spatial data sets. Better methods are needed to analyze spatial data to detect spatial patterns.

Foundations

55 The spatial data mining literature has focused on four main types of spatial patterns: (i) spatial outliers, which are spatial locations showing a significant

difference from their neighbors; (ii) spatial co-locations, or subsets of event types that tend to be found more often together throughout space than other subsets of event types; (iii) location predictions, that is, information that is inferred about locations favored by an event type based on other explanatory spatial variables; and (iv) spatial hotspots, unusual spatial groupings of events. The remainder of this section presents a general overview of each of these pattern categories.

A *spatial outlier* is a spatially referenced object whose non-spatial attribute values differ significantly from those of other spatially referenced objects in its spatial neighborhood. For example, consider spatial outliers, detected in traffic measurements for sensors on highway I-35W (North bound) in the Minneapolis-St. Paul area, for a 24-h time period. Station 9 may be considered a spatial outlier as it exhibits inconsistent traffic flow compared with its neighboring stations. Once a spatial outlier is identified, one may proceed with diagnosis. For example, the sensor at Station 9 may be diagnosed as malfunctioning. Spatial attributes are used to characterize location, neighborhood, and distance. Non-spatial attribute dimensions are used to compare a spatially referenced object to its neighbors. Spatial statistics literature provides two kinds of bipartite multidimensional tests, namely graphical tests and quantitative tests. Graphical tests, such as Variogram clouds and Moran scatterplots, are based on the visualization of spatial data and highlight spatial outliers. Quantitative methods provide a precise test to distinguish spatial outliers from the remainder of data.

Spatial co-location pattern discovery finds frequently co-located subsets of spatial event types given a map of their locations. Spatial co-location is a generalization of a classical data mining pattern family called association rules, since transactions are not natural in spatial datasets, and partitioning space across transactions leads to loss of information about neighbor relationships between items near transaction boundaries. Additional details about co-location interest measures, e.g. participation index and K functions, and mining algorithms are described in [2].

Location prediction is concerned with the discovery of a model to infer preferred locations of a spatial phenomenon from the maps of other explanatory spatial features. For example, ecologists may build models to predict habitats for endangered species using maps of vegetation, water bodies, climate, and other related species. For example, consider an example of a dataset used in building a location prediction model for red-winged blackbirds in the Darr and Stubble wetlands on the shores of Lake Erie in Ohio, USA. This dataset consists of nest location, distance to open

water, vegetation durability and water depth maps. Classical prediction methods may be ineffective in this problem due to the presence of spatial auto-correlation. Spatial data mining techniques that capture the spatial autocorrelation of nest location such as the Spatial Autoregressive Model (SAR) [1] and Markov Random Fields based Bayesian Classifiers (MRF-BC) are used for location prediction modeling.

Spatial Hotspots are unusual spatial groupings of events that tend to be much more closely related than other events. Examples of spatial hotspots can be incidents of crime in a city or outbreaks of a disease. Hotspot patterns have properties of clustering as well as anomalies from classical data mining. However, hotspot discovery [4] remains a challenging area of research due to variation in shape, size, density of hotspots and underlying space (e.g. Euclidean or spatial networks such as roadmaps). Additional challenges arise from the spatio-temporal semantics such as emerging hotspots, displacement etc.

135 Key Applications

Spatial data mining and the discovery of spatial patterns has applications in a number of areas. Detecting spatial outliers is useful in many applications of geographic information systems and spatial databases, including the domains of public safety, public health, climatology, and location-based services. As noted earlier, for example, spatial outlier applications may be used to identify defective or out of the ordinary (i.e., unusually behaving) sensors in a transportation system. Spatial co-location discovery is useful in ecology in the analysis of animal and plant habitats to identify co-locations of predator-prey species, symbiotic species, or fire events with fuel and ignition sources. Location prediction may provide applications toward predicting the climatic effects of El Nino on locations around the world. Finally, identification of spatial hotspots can be used in crime prevention and reduction, as well as in epidemiological tracking of disease.

(Abridged)

Recommended Reading

1. Cressie N.A. Statistics for Spatial Data (Revised Edition). Wiley, New York, NY, 1993.
2. Huang Y., Shekhar S., and Xiong H. Discovering co-location patterns from spatial datasets: a general approach. IEEE Trans. Knowl. Data Eng. (TKDE), 16(12):1472–1485, 2004.
3. Shekhar S., Schrater P., Vatsavai R., Wu W., and Chawla S. Spatial contextual classification and

- prediction models for mining geospatial data. IEEE Trans. Multimed. (special issue on Multimedia Databases), 4(2):174–188, 2002.
4. US Department of Justice - Mapping and Analysis for Public Safety report. Mapping Crime: Understanding Hot Spots, 2005
- (<http://www.ncjrs.gov/pdffiles1/nij/209393.pdf>).
5. Shekhar S. and Chawla S. A Tour of Spatial Databases. Prentice Hall, 2003.
6. Longley P.A., Goodchild M., Maquire D.J., and Rhind D.W. Geographic Information Systems and Science. Wiley, 2005.
7. Shekhar S., Zhang P., Huang Y., and Vatsavai R. Trend in spatial data mining. In Data Mining: Next Generation Challenges and Future Directions, H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha (eds.). AAAI/MIT Press, 2003.
8. Solberg A.H., Taxt T., and Jain A.K. A Markov random field model for classification of multisource satellite imagery. IEEE Trans. Geosci. Remote Sens., 34(1):100–113, 1996.
9. Kou Y., Lu C.T., and Chen D. Algorithms for spatial outlier detection. In Proc. 2003 IEEE Int. Conf. on Data Mining, 2003, pp. 597–600.
10. Shekhar S., Lu C.T., and Zhang P. A unified approach to detecting spatial outliers. GeoInformatica, 7(2):139–166, 2003.
11. Mamoulis N., Cao H., and Cheung D.W. Mining frequent spatio-temporal sequential patterns. In Proc. 2003 IEEE Int. Conf. on Data Mining, 2005, pp. 82–89.

Answer the following questions:

- 1) What is spatial data mining?
- 2) What are the basic differences between classical datasets and spatial datasets?
- 3) Why are classical data mining algorithms unsuitable for spatial datasets?
- 4) What does it mean that spatial data tends to be highly self-correlated? How would you explain spatial auto correlation?
- 5) Name the four main types of spatial patterns.
- 6) Explain a spatial outlier and give a simple example.
- 7) What are spatial and non-spatial attributes?
- 8) Which pairs are frequently co-located?
- 9) Give some examples of hotspots.
- 10) Name some areas where spatial data mining is applied.

Mark the following statements as *true* or *false*:

- 1) Many of the relationships on spatial data are implicit.
- 2) Spatial autocorrelation means that nearby things are more different than distant things.
- 3) In spatial data mining data samples are not independent.
- 4) Classical data mining techniques perform well when applied to spatial data sets.
- 5) Spatial data mining is based on the same assumptions as classical data mining.

Match the following terms with their definitions:

- 1) Spatial data mining (SDM)
 - 2) Spatial outlier
 - 3) Location prediction
 - 4) Spatial co-location
 - 5) Hotspot analysis
 - 6) Spatial autocorrelation
 - 7) Variogram clouds and Moran scatterplots
-
- a) Presence of two or more spatial objects at the same location or at significantly close distances from each other
 - b) Process of finding unusually dense event clusters across time and space
 - c) Process of discovering interesting, useful, non-trivial patterns from large spatial datasets
 - d) Spatial outlier detection algorithms based on visualization
 - e) Observation which appears to be inconsistent with its neighborhood
 - f) Prediction of events occurring at particular geographic locations
 - g) Spatial data values are influenced by values in their immediate vicinity

Vocabulary

anomaly [ə'nɒm.ə.li] [-'naɪ.mə-] - odchylka

assumption [ə'sʌmp.ʃən] - předpoklad, domněnka

boundary ['baʊn.dər.i] [-dri] [-də-]

- hranice, mez

defective [dɪ'fek.tɪv] - vadný, chybný

discrete [dɪ'skri:t] - nespojitý, oddělený, samostatný

displacement [dɪ'spleɪs.mənt] - přemístění, posun

endangered [ɪn'deɪn.dʒəd] [-dʒəd]

- ohrožený

grouping ['gruː.pɪŋ] - seskupení

habitat ['hæb.ɪ.tæt] - přirozené prostředí, domov

hypothesis, hypotheses (pl.) [haɪ'pɒθ.ə.sɪs] [-'paɪ.θə-] - hypotéza, předpoklad, domněnka

ignition [ɪg'niʃ.ən] - vznícení, vzplanutí

incident ['ɪn.t.sɪ.dənt] - případ, incident, příhoda, událost

inconsistent [ɪn.kən'sɪs.tənt] - neslučitelný, jsoucí v rozporu

namely ['neɪm.li] - a to, jmenovitě

nontrivial [nɒn'trɪv.i.əl] - podstatný, důležitý

out of the ordinary [aʊt] ['ɔː.di.nə.ri]

['ɔːr.dən.ər-] - neobvyklý, zvláštní

outbreak ['aʊt.breɪk] - vypuknutí

outlier ['aʊt.laɪ.ə] - co leží mimo

pattern ['pæt.ən] ['pæʧ.ən] - vzorec, způsob, průběh

phenomenon, phenomena (pl.) [fə'nɒm.i.nən]

[-'naɪ.mə.nə:n] - jev

predator ['pred.ə.tər] [-tər] - dravec, predátor

prey [preɪ] - kořist

remainder [rɪ'meɪn.dər] [-dər] - zbytek

species ['spiː.ʃiːz] - druh

stationary ['steɪ.ʃən.ər.i] [-fə.nər-]

- neměnný, nehybný

throughout space [θruː'aʊt] [speɪs] - po celém prostoru

to be concerned with st [kən'sɜːnd]

[-'sɜːnd] - zabývat se čím

to cluster ['klʌs.tər] [-tər] - shlukovat se, seskupit

to co-locate [kəʊ.ləʊ'keɪt] [kəʊ.ləʊ'keɪt]

- vyskytovat se společně

to decompose [ˌdiː.kəm'pəʊz] [-'pəʊz]

- rozložit

to favor ['feɪ.vər] [-vər] - podporovat

to highlight ['haɪ.laɪt] - zvýraznit, poukázat

to infer [ɪn'fɜːr] [-'fɜːr] - dedukovat,

vyvozovat

to note [nəʊt] [nəʊt] - upozornit, poukázat

to partition [pɑː'tɪʃ.ən] [pɑːr-] - rozdělit

to proceed [prə'ʊsiːd] [prəʊ-] - pokračovat, postupovat

to reference ['ref.ər.ənts] [-ə-] - odkazovat

to tend [tend] - mít sklon, tendenci, být

náchylný

to track [træk] - sledovat

to violate ['vaɪə.leɪt] - porušit, nedodržet

Phrases

due to - kvůli

either - or - buď - nebo

from the standpoint - z hlediska

in infant stages - na počátku vývoje