Co-Location Rules Discovery Process Focused on Reference Spatial Features Using Decision Tree Learning

Giovanni Daián Rottoli [1,2,3], Hernán Merlino[3], Ramón García-Martinez [3,4]

1 PhD Program on Computer Sciences. National University of La Plata. Argentina.
2 PhD Scholarship Program to Reinforce R+D+I Areas.
National Technological University. Argentina
3 Information Systems Research Group. National University of Lanús. Argentina
4 Scientific Researchs Comission - CIC Bs As. Argentina

gd.rottoli@gmail.com,hmerlino@gmail.com,rgm1960@yahoo.com

Abstract. The co-location discovery process serves to find subsets of spatial features frequently located together. Many algorithms and methods have been designed in recent years; however, finding this kind of patterns around specific spatial features is a task in which the existing solutions provide incorrect results. Throughout this paper we propose a knowledge discovery process to find co-location patterns focused on reference features using decision tree learning algorithms on transactional data generated using maximal cliques. A validation test of this process is provided.

Keywords: Co-Location Patterns, Spatial Data Mining, Decision Trees Algorithms, Maximal Cliques, Knowledge Discovery Process.

1 Introduction

Given a collection of boolean spatial features (also known as spatial events), the co-location pattern discovery process finds the subset of features frequently located together [1]. Some examples of this kind of relationships are symbiotic species, and public service buildings frequently built together, like hospitals and pharmacies [2]. Many algorithms and methods have been proposed for co-location pattern discovery based on association analysis. These algorithms generate transactional data from spatial objects neighborhoods and, based on that, they can be categorized into two classes: (i) transaction-free algorithms, which exploit the association analysis algorithm internally, e.g., the Aprori-like algorithms [3], but none of them generates or uses a transaction-type dataset externally; and (ii), transaction-based algorithms, which exploit association analysis methods after explicitly generating a transaction-type dataset. [1,4,5] In both options, it is necessary to choose a model to generate the transactional data. There are three different approaches: [1,5,6]

- **Window-Centric Model**: in a space discretized by a uniform grid, windows of size W can be enumerated. Each window corresponds to a transaction that contains a subset of spatial features related to the spatial instances found on the window.

- **Event-Centric Model**: used to find subsets of spatial features likely to occur in a neighborhood around instances of given subsets of event types.
- **Reference Feature-Centric Model**: the transactions are created by "materializing" the neighborhood of the instances of the reference spatial feature.

Nowadays, the spatial datasets are collected for a particular problem domain and, because of that, there is a spatial feature more relevant than the others. In this case, it is appropriate to select a Reference Feature-Centric Model for the generation of transactional data. However, two problems arise: all the applications and algorithms listed in previous works use an Event-Centric Model. Some examples of these publications are [1-14]. On the other hand, the transactional data generated using a Reference Feature-Centric Model may be incorrect and incomplete [5].

This paper is organized as follows: In Section 2, a problem derived from the analysis of the state-of-the-art is presented. In Section 3, we present a Knowledge Discovery Process to solve that problem. In Section 4, experimental results are presented. Finally, conclusions derived from the research are outlined in Section 5.

2 Problem Definition

When facing a co-location discovery problem, if there are many boolean spatial features to be considered for co-location pattern discovery, the Event-Centric Model may be expensive in terms of time and resources. With the presence of a spatial feature that is interesting in a particular problem domain, using a Reference Feature-Centric Model is a more suitable alternative. This model determines the neighborhoods in a special manner: first, a reference feature is selected and then, for each instance object of that feature, all spatial objects located within a pre-specified distance are selected, and transaction-type data generated [1,5,6]. This approach, however, cannot be used to generate correct or complete transactions, as it does not ensure that all objects in the transaction are neighbors; moreover, some neighborhoods may be lost [2,5,12]. For this reason, it is necessary to develop a solution that serves to discover correct and complete spatial co-location patterns around reference features. In this work, we develop a Knowledge Discovery Process [15,16] to give a solution to this problem using an Event-Centric Model to generate transaction-based data, and induction of decision trees to generate co-location rules.

3 Proposed Solution

As mentioned before, this paper proposes a knowledge discovery process to find co-location relations between spatial features. This process serves to find correct relations around reference features without using a reference feature-centric model to generate transactional data. This work is based on the work of Kim et al. [5], proposing a transactional framework that uses an event-centric model to find co-location patterns using maximal cliques as a way to generate complete and correct transactions. In this context, Spatial Maximal Clique (SMC) is defined as follows [5]: Given a spatial dataset consisting of spatial objects, a spatial clique is a subset of the

dataset whose elements have neighbor relationships with each other. A Spatial Maximal Clique is a Spatial Clique that is not part of the others. Using an SMC as a transaction generates two properties: all the elements in the SMC are neighbors with each other, ensuring the correctness of the method, and there is no neighbor relation that is excluded, ensuring the completeness in the transactional data. SMC seems to be a proper solution to solve the aforementioned problems, but finding a way of discovering co-location patterns around features relevant to the problem domain is necessary, because the classical association rules discovery algorithms used in the transactional-based approaches cannot be used to select a target feature. For this reason, a knowledge discovery process for co-location pattern discovery is proposed, focused on reference features, that uses an event-centric model for transaction-based data generation through spatial maximal cliques and using a Process of Discovery of Behavior Rules using Decision Tree Learning algorithms [15,16]. Figure 1 shows the proposed process using BPMN [17].

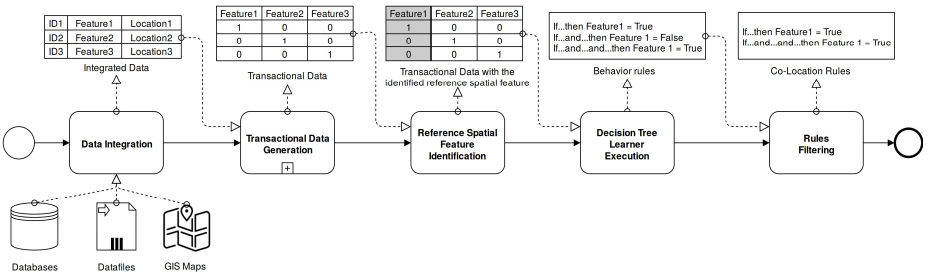

Fig. 1. Proposed co-location rules discovery process

The process takes a set of spatially referenced data as input, represented in different formats such as *inter alia*, plain text, databases, tables and geographic information system maps. These data are integrated to a table comprised of the object identifier, the spatial feature and the object location. Then, the integrated data are used to generate the transactional dataset. In this sub-process, as shown in Figure 2, all the neighbor relationships are calculated by evaluating the distance between the spatial objects.

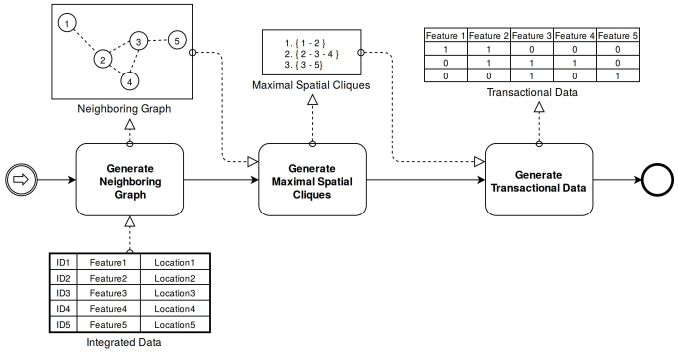

Fig. 2. Sub-process to generate transactional data

The distance function and the threshold will depend on the problem domain. Afterwards, finding all the maximal spatial cliques inside then neighboring graph is required to generate a transaction for each, in which the spatial features of each spatial object from that clique are presented. Once the transactional data are obtained, the reference spatial feature must be specified to find the co-location relations around it. That spatial feature will be used as the target attribute of a Decision Tree Learning algorithm, such as C4.5 [18] or Random Forest [19], using the rest of the attributes as input. A set of rules will be obtained from the generated decision tree in the last step as output. Due to the fact that the transactions have boolean values that show the presence or absence of the spatial features in the neighborhoods, it is necessary to filter the rules that show the presence of the reference spatial feature in the consequent.

4 Concept Proof

To proof the concept, we create 10 synthetic sets of 500 points automatically generated and classified in 7 types, with random location in a small 2D space, and then used as input for our proposed process and for the selected algorithm: Co-Location Miner with a Reference Feature-Centric Model [1]. On the other hand, the euclidean distance function has been used to calculate the neighbouring graph using a constant threesome. The algorithm CLIQUES has been used for the generation of maximal spatial cliques because of its superior efficiency over other methods [20]. The software Tanagra [21] was used to run the selected TDIDT Learning Algorithm C4.5 [18]. After the execution of both methods, the co-location relationships obtained were evaluated to corroborate their correctness in order to determine how many correct relationships were found with each method.

To show that the proposed process can find a greater number of co-location relationships, the statistical Wilcoxon signed-rank test was used [22] considering the hypotheses shown in Table 1, obtaining a W-Value equals to 0 (see Table 2) that allows to reject the null hypothesis, confirming that the knowledge discovery process proposed in this paper serves to find a greater number of correct co-location relationships that the method using a reference feature-centric model.

5 Conclusion

This paper has described a knowledge discovery process that can be used to find correct and complete co-location patterns around reference spatial features. This process uses an event-centric model through maximal spatial cliques in order to generate transactional data from neighboring relationship between spatial data, and a decision tree learning algorithm, to obtain behavior rules that describe the neighborhoods that contain the spatial reference feature, an innovative method to achieve this goal. The proof of concept, by means of a non-parametrical statistical test, shows that the proposed process finds a greater number of correct co-location patterns than the methods that use a reference feature-centric model to generate transactional data.

The next planned step is to conduct validation proofs in the fields of accident prevention, civil defense and environmental determinants of diseases.

Table 1. Null hypotesis and alternative hypotesis considered in the Wilcoxon signed-rank test

H_0:	The number of correct co-location relationships discovered using the Co-Location Miner Algorithm is greater than or equal to the number of correct co-location relationships discovered using the proposed process.
H_A:	The number of correct co-location relationships discovered using the proposed process is greater than the number of correct co-location relationships discovered using the Co-location Miner Algorithm.

Table 2. Wilcoxon signed-rank test execution, sorted by the absolute value of the differences.

Set	Proposed process	Co-location miner algorithm.	Absolute value of the differences	Rank
7	3	3	0	-
8	2	1	1	2
9	2	1	1	2
10	5	4	1	2
1	3	1	2	5
2	3	1	2	5
5	4	2	2	5
4	4	1	3	7
3	6	2	4	8.5
6	7	3	4	8.5
Sum				45

Acknowledgements

The research presented in this paper was partially funded by the PhD Scholarship Program to reinforce R+D+I areas (2016-2020) of the Technological National University, Research Project 80020160400001LA of National University of Lanús, and PIO CONICET-UNLa 22420160100032CO of National Research Council of Science and Technology (CONICET), Argentina. The authors also want to extend their gratitude to Kevin-Mark Bozell Poudereux for proofreading the translation.

References

1. Shekhar, S., & Huang, Y. (2001). Discovering spatial co-location patterns: A summary of results. In Advances in Spatial and Temporal Databases (pp. 236-256). Springer Berlin Heidelberg.
2. Yu, W. (2016). Spatial co-location pattern mining for location-based services in road networks. Expert Systems with Applications, 46, 324-335.
3. Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).
4. Shekhar, S., Evans, M. R., Kang, J. M., & Mohan, P. (2011). Identifying patterns in spatial information: A survey of methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3), 193-214.

5. Kim, S. K., Lee, J. H., Ryu, K. H., & Kim, U. (2014). A framework of spatial co-location pattern mining for ubiquitous GIS. Multimedia tools and applications, 71(1), 199-218.
6. Xiong, H., Shekhar, S., Huang, Y., Kumar, V., Ma, X., & Yoo, J. S. (2004, April). A Framework for Discovering Co-Location Patterns in Data Sets with Extended Spatial Objects. In SDM (pp. 78-89).
7. Huang, Y., Xiong, H., Shekhar, S., & Pei, J. (2003, March). Mining confident co-location rules without a support threshold. In Proceedings of the 2003 ACM symposium on Applied computing (pp. 497-501). ACM.
8. Huang, Y., Pei, J., & Xiong, H. (2006). Mining co-location patterns with rare events from spatial data sets. Geoinformatica, 10(3), 239-260.
9. Yoo, J. S., & Shekhar, S. (2006). A joinless approach for mining spatial colocation patterns. Knowledge and Data Engineering, IEEE Transactions on, 18(10), 1323-1337.
10. Celik, M., Kang, J. M., & Shekhar, S. (2007, October). Zonal co-location pattern discovery with dynamic parameters. In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on (pp. 433-438). IEEE.
11. Eick, C. F., Parmar, R., Ding, W., Stepinski, T. F., & Nicot, J. P. (2008, November). Finding regional co-location patterns for sets of continuous variables in spatial datasets. In Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems (p. 30). ACM.
12. Adilmagambetov, A., Zaiane, O. R., & Osornio-Vargas, A. (2013). Discovering co-location patterns in datasets with extended spatial objects. In Data Warehousing and Knowledge Discovery (pp. 84-96). Springer Berlin Heidelberg.
13. Venkatesan, M., Thangavelu, A., & Prabhavathy, P. (2011). Event Centric Modeling Approach in Colocation Pattern Snalysis from Spatial Data. arXiv preprint arXiv:1109.1144.
14. Yoo, J. S., Shekhar, S., & Celik, M. (2005, November). A join-less approach for co-location pattern mining: A summary of results. In Data Mining, Fifth IEEE International Conference on (pp. 4-pp). IEEE.
15. García-Martínez, R., Britos, P., Rodríguez, D. 2013. Information Mining Processes Based on Intelligent Systems. Lecture Notes on Artificial Intelligence, 7906: 402-410. ISBN 978-3-642-38576-6.
16. Martins, S., Pesado, P., & García-Martínez, R. (2016, August). Intelligent Systems in Modeling Phase of Information Mining Development Process. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 3-15). Springer International Publishing.
17. Silver, B. (2011). BPMN Method and Style, with BPMN Implementer's Guide: A structured approach for business process modeling and implementation using BPMN 2.0. Cody-Cassidy Press, Aptos, CA, 450.
18. Quinlan, J. R. (1993). C4. 5: programs for machine learning.
19. Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8), 832-844.
20. Tomita, E., Tanaka, A., & Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. Theoretical Computer Science, 363(1), 28-42.
21. Rakotomalala, R. (2005). TANAGRA: a free software for research and academic purposes. In Proceedings of EGC (Vol. 2, pp. 697-702).
22. Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics bulletin, 1(6), 80-83.