Relational Mining in Spatial Domains: Accomplishments and Challenges

Donato Malerba, Michelangelo Ceci, and Annalisa Appice

Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro
via Orabona, 4 - 70126 Bari - Italy
{malerba,ceci,appice}@di.uniba.it

Abstract. The rapid growth in the amount of spatial data available in Geographical Information Systems has given rise to substantial demand of data mining tools which can help uncover interesting spatial patterns. We advocate the relational mining approach to spatial domains, due to both various forms of spatial correlation which characterize these domains and the need to handle spatial relationships in a systematic way. We present some major achievements in this research direction and point out some open problems.

1 Introduction

Several real world applications, such as fleet management, environmental and ecological modeling, remote sensing, are the source of a huge amount of spatial data, which are stored in spatial databases of Geographic Information Systems (GISs). A GIS is a software system that provides the infrastructure for editing, storing, analyzing and displaying spatial objects [10]. Popular GISs (e.g. ArcView, MapInfo and Open GIS) have been designed as a toolbox that allows planners to explore spatial data by zooming, overlaying, and thematic map coloring. They are provided with functionalities that make the spatial visualization of individual variables effective, but overlook complex multi-variate dependencies.

The solution to this limitation is to integrate GIS with *spatial data mining* tools [25]. Spatial data mining investigates how interesting, but not explicitly available, knowledge (or pattern) can be extracted from spatial data [34]. Several algorithms of spatial data mining have been reported in the literature for both predictive tasks (e.g., regression [24], [14], and localization [32]) and descriptive tasks (e.g., clustering [28,16] and discovery of association rules [19,3], co-location patterns [33], subgroups [18], emerging patterns [6], spatial trends [11] and outliers [31]).

Spatial data mining differs from traditional data mining in two important respects. First, spatial objects have a locational property which implicitly defines several spatial relationships between objects such as topological relationships (e.g., intersection, adjacency), distance relationships, directional relationships (e.g., north-of), and hybrid relationships (e.g., parallel-to). Second, attributes of spatially interacting (i.e., related) units tend to be statistically correlated. *Spatial*

M. Kryszkiewicz et al. (Eds.): ISMIS 2011, LNAI 6804, pp. 16–24, 2011.

cross-correlation refers to the correlation between two distinct attributes across space (e.g., the employment rate in a city depends on the business activities both in that city and in its neighborhood). *Autocorrelation* refers to the correlation of an an attribute with itself across space (e.g., the price level for a good at a retail outlet in a city depends on the price for the same good in the nearby). In geography, spatial autocorrelation is justified by Tobler's First law of geography, according to which "everything is related to everything else, but near things are more related than distant things" [35].

The network of data defined by implicit spatial relationships can sometime reveal the network of statistical dependencies in a spatial domain (e.g., region adjacency is a necessary condition for autocorrelation of air pollution). Nevertheless, the two concepts do not necessarily coincide. On the contrary, it is the identification of some form of spatial correlation which help to clarify what are the relevant spatial relationships among the infinitely many that are implicitly defined by locational properties of spatial objects.

The presence of spatial dependence is a clear indication of a violation of one of the fundamental assumptions of classic data mining algorithms, that is, the independent generation of data samples. As observed by LeSage and Pace [21], "anyone seriously interested in prediction when the sample data exhibit spatial dependence should consider a spatial model", since this can take into account different forms of spatial correlation. In addition to predictive data mining tasks, this consideration can also be applied to descriptive tasks, such as spatial clustering or spatial association rule discovery. The inappropriate treatment of sample data with spatial dependence could obfuscate important insights and observed patterns may even be inverted when spatial autocorrelation is ignored [20].

In order to accommodate several forms of spatial correlation, various models have been developed in the area of spatial statistics. The most renowned types of models are the spatial lag model, the spatial error model, and the spatial cross-regressive model [1], which consider autocorrelation, correlation of errors, and cross-correlation, respectively.

Despite the many successful applications of these models, there are still several limitations which prevent their wider usage in a spatial data mining context. First, they require the careful definition of a spatial weight matrix in order to specify to what extent a spatially close observation in a given location can affect the response observed in another location. Second, there is no clear method on how to express the contribution of different spatial relationships (e.g., topological and directional) in a spatial weight matrix. Third, spatial relationships are all extracted in a pre-processing step, which typically ignores the subsequent data mining step. In principle, a data mining method, which can check whether a spatial relationship contributes to defining a spatial dependency, presents the advantage of considering only those relationships that are really relevant to the task at hand. Fourth, all spatial objects involved in a spatial phenomena are uniformly represented by the same set of attributes. This can be a problem when spatial objects are heterogeneous (e.g., city and roads). Fifth, there is no clear distinction between the *reference* (or target) *objects*, which are the main subject of analysis, and the *task-relevant objects*, which are spatial objects "in the neighborhood" that can help to account for the spatial variation.

A solution to above problems is offered by latest developments in *relational mining* or *relational learning*. Indeed, relational mining algorithms can be directly applied to various representations of networked data, i.e. collections of interconnected entities. By looking at spatial databases as a kind o networked data where entities are spatial objects and connections are spatial relationships, the application of relational mining techniques appears straightforward, at least in principle. Relational mining techniques can take into account the various forms of correlation which bias learning in spatial domains. Furthermore, discovered relational patterns reveal those spatial relationships which correspond to spatial dependencies.

This relational mining approach to spatial domains has been advocated in several research papers [18,23,12]. Major accomplishments in this direction have been performed, but there are still many open problems which challenges researchers. In the rest of the paper, we pinpoint the important issues that need to be addressed in spatial data mining, as well as the opportunities in this emerging research direction.

2 Integration with Spatial Databases

Spatial data are stored in a set of *layers*, that is, database relations each of which has a number of elementary attributes, called thematic data, and a geometry attribute represented by a vector of coordinates. The computation of spatial relationships, which are fundamental for querying spatial data, is based on spatial joins [30]. To support the efficient computation of spatial joins, special purpose indexes like Quadtrees and Kd-tree [27] are used.

Integration can be tight, as in SubgroupMiner [18] and Mrs-SMOTI [24], or loose as in ARES [2]. A tight integration:

- guarantees the applicability of spatial data mining algorithms to large spatial datasets;
- exploits useful knowledge of spatial data model available, free of charge, in the spatial database;
- avoids useless preprocessing to compute spatial relationships which do not express statistical dependencies.

A loose integration is less efficient, since it uses a middle layer module to extract both spatial attributes and relationships independently of the specific data mining step. On the other hand, this decoupling between the spatial database and the data mining algorithm allows researchers to focus on general aspects of the relational data mining task, and to exploit important theoretical and empirical results. A systematic study of these integration approaches should lead to valuable information on how a spatial data mining task should be methodologically dealt with.

Many relational mining methods take advantage of knowledge on the data model (e.g., foreign keys), which is obtained free of charge from the database schema, in order to guide the search process. However, this approach does not suit spatial databases, since the database navigation is also based on the spatial relationships, which are not explicitly modeled in the schema. The high number of spatial relationships is a further complication, since each individual relationship can become insignificant on its own, requiring the use of some form of spatial aggregation [12].

3 Dealing with Hierarchical Representations of Objects and Relationships

Both spatial objects and spatial relationships are often organized in taxonomies typically represented by hierarchies [37]. By descending/ascending through a hierarchy it is possible to view the same spatial object/relationship at different levels of abstraction (or granularity). Spatial patterns involving the most abstract spatial objects and relationships can be well supported but at the same time they are the less confident. Therefore, spatial data mining methods should be able to explore the search space at different granularity levels in order to find the most interesting patterns (e.g., the most supported and confident). In the case of granularity levels defined by a containment relationship (e.g., Bari → Apulia → Italy), this corresponds to exploring both global and local aspects of the underlying phenomenon. Geo-associator [19] and SPADA [22] are two prominent examples of spatial data mining systems which automatically support this multiple-level analysis. However, there is still no clear methodization for extracting, representing and exploiting hierarchies of spatial objects and relationships in knowledge discovery.

4 Dealing with Spatial Autocorrelation

Relational mining algorithms exploit two sources of correlation when they discover relational patterns: *local correlation*, i.e., correlation between attributes of each unit of analysis, and *within-network correlation*, i.e., correlation between attributes of the various units of analysis. In this sense, they are appropriate for spatial domains, which present both forms of correlation. For instance, the spatial subgroup mining system SubgroupMiner [18] is built on previous work on relational subgroup discovery [38], although it also allows numeric target variables, and aggregations based on (spatial) links. The learning system Mrs-SMOTI [24], which learns a tree-based regression model from spatial data, extends the relational system Mr-SMOTI [4] by associating spatial queries to nodes of model trees. UnMASC [12] is based on both the idea of the sequential covering algorithm developed in the relational data mining system CrossMine [39] and on aggregation-based methods originally proposed for relational classification [13].

However, predictive modeling in spatial domains still challenges most relational mining algorithms when autocorrelation on the target (or response) variable is captured. Indeed, values of the target variable of unclassified units of analysis have to be inferred collectively, and not independently as most relational mining algorithm do. *Collective inference* refers to the simultaneous judgments regarding the values of response variables for multiple linked entities for which some attribute values are not known. Several collective inference methods (e.g., Gibbs sampling, relaxation labeling, and iterative classification) have been investigated in the context of relational learning. For the specific task of classification it has been proven that collective inference outperforms independent classification when the autocorrelation between linked instances in the data graph is high [17]. Collective inference in the context of spatial predictive modeling is still a largely unexplored area of research.

5 Dealing with Unlabeled Data

Learning algorithms designed for mining spatial data may require large sets of labeled data. However, the common situation is that only few labeled training data are available since manual annotation of the many objects in a map is very demanding. Therefore, it is important to exploit the large amount of information potentially conveyed by unlabeled data to better estimate the data distribution and to build more accurate classification models. To deal with this issue, two learning settings have been proposed in the literature: the semi-supervised setting and the transductive setting [29]. The former is a type of inductive learning, since the learned function is used to make predictions on any possible example. The latter asks for less - it is only interested in making predictions for the given set of unlabeled data.

Transduction [36] seems to be the most suitable setting for spatial classification tasks, for at least two reasons. First, in spatial domains observations to be classified are already known in advance: they are spatial objects on maps already available in a GIS. Second, transduction is based on a (semi-supervised) smoothness assumption according to which if two points x_1 and x_2 in a high-density region are close, then the corresponding outputs y_1 and y_2 should also be close [8]. In spatial domains, where closeness of points corresponds to some spatial distance measure, this assumption is implied by (positive) spatial autocorrelation. Therefore, we expect that a strong spatial autocorrelation should counterbalance the lack of labeled data, when transductive relational learners are applied to spatial domains. Recent results for spatial classification [7] and spatial regression tasks [5] give support to this expectation. Nevertheless, more experiments are needed to substantiate this claim.

6 Dealing with Dynamic Spatial Networks

Most of works on spatial data mining assume that the spatial structure is static. Nevertheless, changes may occur in many real-world applications (e.g., the public transport network can change). This causes the appearance and disappearance of spatial objects and spatial relationships over time, while properties of the spatial objects may evolve. By analyzing these changes, we can follow variations, adapt tools and services to new demands, as well as capture and delay undesirable alterations. Moreover, time associated to changes represent a valuable source of information which should be modeled to better understand both the whole dynamics and each change in the course of dynamics.

In the literature, the task of change mining has been mainly explored for time-series, transactional data and tabular data, by focusing on the detection of significant deviations in the values of the attributes describing the data. However, detecting and analyzing changes on spatially referenced data is critical for many applications. For instance, by taking snapshots of over time of the spatial distribution of plant species, it is possible to monitor significant changes, which may reveal important ecological phenomena. Pekerskaya et al. [26] address the problem of mining changing regions by directly comparing models (cluster-embedded decision trees) built on the original data snapshots. This approach is suitable when there are data access constraints such as privacy concerns and limited data online availability. Ciampi et al. [9] consider the case of distributed streams of unidimensional numeric data, where each data source is a geo-referenced remote sensor which periodically records measures for a specific numeric theme (e.g., temperature, humidity). A combination of stream and spatial data mining techniques is used to mine a new kind of spatio-temporal patterns, called trend clusters, which are spatial clusters of sources for which the same temporal variation is observed over a time window.

Spatial networks demand for attention not only on the attributes which may describe nodes and links but also on the structural and topological aspects of the network, namely the relationships among the nodes and the kind of links which connect the nodes. In this direction, research on network analysis has mainly investigated graph-theoretical approaches which oversimplify the representation of spatial networks. Indeed, graph-theory mainly investigates structural aspects, such as distance and connectivity, in homogeneous networks, while it almost ignores the data heterogeneity issue, which is typical of spatial networks, where nodes are of different types (e.g. in public transport networks, public services and private houses should be described by different feature sets), and relationships among nodes can be of different nature (e.g. connection by bus, railway or road). Methods for learning and inference with networks of heterogeneous data have been investigated in the context of statistical relational learning [15], however the scalability issue that characterizes many most statistical relational learning methods makes their application very challenging in the case of dynamic networks due to continuous changes in the network.

7 Conclusions

In this paper, we have advocated a relational mining approach to spatial domains, and we have presented some major research achievements in this direction. Research results are encouraging but there are still many open problems which challenge current relational mining systems, namely:

1. a methodological support to the integration of spatial database technology with data mining tools;
2. the potentially infinitely many spatial relationships which are implicitly defined by spatial objects;
3. the efficient discovery of spatial patterns at various levels of granularity;
4. the demand for collective inference in predictive models which capture autocorrelation;
5. the exploitation of the many unlabeled spatial objects in a semi-supervised or transductive setting;
6. the need of new types of patters which capture the interactions between both spatial and temporal dimensions in spatially static structures;
7. the structural changes of dynamic networks with heterogeneous spatial objects.

Obviously, this list of challenges is not exhaustive, but rather it is indicative of the necessity for developing synergies between researchers interested in spatial data mining and relational learning. Some efforts have been made, but the two research communities still work in relative isolation from one another, with little methodological common ground. Nonetheless, there is good cause for optimism: there are many real-world applications which cry out for this collaboration.

References

1. Anselin, L., Bera, A.: Spatial dependence in linear regression models with an application to spatial econometrics. In: Ullah, A., Giles, D. (eds.) Handbook of Applied Economics Statistics, pp. 21–74. Springer, Heidelberg (1998)
2. Appice, A., Berardi, M., Ceci, M., Malerba, D.: Mining and filtering multi-level spatial association rules with ARES. In: Hacid, M.-S., Murray, N.V., Raś, Z.W., Tsumoto, S. (eds.) ISMIS 2005. LNCS (LNAI), vol. 3488, pp. 342–353. Springer, Heidelberg (2005)
3. Appice, A., Ceci, M., Lanza, A., Lisi, F.A., Malerba, D.: Discovery of spatial association rules in georeferenced census data: A relational mining approach. Intelligent Data Analysis 7(6), 541–566 (2003)
4. Apice, A., Ceci, M., Malerba, D.: Mining model trees: A multi-relational approach. In: Horváth, T., Yamamoto, A. (eds.) ILP 2003. LNCS (LNAI), vol. 2835, pp. 4–21. Springer, Heidelberg (2003)
5. Appice, A., Ceci, M., Malerba, D.: Transductive learning for spatial regression with co-training. In: Shin, S.Y., Ossowski, S., Schumacher, M., Palakal, M.J., Hung, C.-C. (eds.) SAC, pp. 1065–1070. ACM, New York (2010)
6. Ceci, M., Appice, A., Malerba, D.: Discovering emerging patterns in spatial databases: A multi-relational approach. In: Kok, J.N., Koronacki, J., de Mántaras, R.L., Matwin, S., Mladenic, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 390–397. Springer, Heidelberg (2007)
7. Ceci, M., Appice, A., Malerba, D.: Transductive learning for spatial data classification. In: Koronacki, J., Ras, Z.W., Wierzchon, S.T., Kacprzyk, J. (eds.) Advances in Machine Learning I. Studies in Computational Intelligence, vol. 262, pp. 189–207. Springer, Heidelberg (2010)

8. Chapelle, O., Schölkopf, B.B., Zien, A.: Semi-supervised learning. MIT Press, Cambridge (2006)
9. Ciampi, A., Appice, A., Malerba, D.: Summarization for geographically distributed data streams. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS, vol. 6278, pp. 339–348. Springer, Heidelberg (2010)
10. Densham, P.: Spatial decision support systems. Geographical Information Systems: Principles and Applications, 403–412 (1991)
11. Ester, M., Gundlach, S., Kriegel, H., Sander, J.: Database primitives for spatial data mining. In: Proceedings of the International Conference on Database in Office, Engineering and Science, BTW 1999, Freiburg, Germany (1999)
12. Frank, R., Ester, M., Knobbe, A.J.: A multi-relational approach to spatial classification. In: Elder IV, J.F., Fogelman-Soulié, F., Flach, P.A., Zaki, M.J. (eds.) KDD, pp. 309–318. ACM, New York (2009)
13. Frank, R., Moser, F., Ester, M.: A method for multi-relational classification using single and multi-feature aggregation functions. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 430–437. Springer, Heidelberg (2007)
14. Gao, X., Asami, Y., Chung, C.: An empirical evaluation of spatial regression models. Computers & Geosciences 32(8), 1040–1051 (2006)
15. Getoor, L., Taskar, B. (eds.): Introduction to Statistical Relational Learning. MIT Press, Cambridge (2007)
16. Han, J., Kamber, M., Tung, A.K.H.: Spatial Clustering Methods in Data Mining: A Survey. In: Geographic Data Mining and Knowledge Discovery, pp. 1–29. Taylor and Francis, Abington (2001)
17. Jensen, D., Neville, J., Gallagher, B.: Why collective inference improves relational classification. In: Kim, W., Kohavi, R., Gehrke, J., DuMouchel, W. (eds.) KDD, pp. 593–598. ACM, New York (2004)
18. Klösgen, W., May, M.: Spatial subgroup mining integrated in an object-relational spatial database. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) PKDD 2002. LNCS (LNAI), vol. 2431, pp. 275–286. Springer, Heidelberg (2002)
19. Koperski, K., Han, J.: Discovery of spatial association rules in geographic information databases. In: Egenhofer, M.J., Herring, J.R. (eds.) SSD 1995. LNCS, vol. 951, pp. 47–66. Springer, Heidelberg (1995)
20. Kühn, I.: Incorporating spatial autocorrelation invert observed patterns. Diversity and Distributions 13(1), 66–69 (2007)
21. LeSage, J.P., Pace, K.: Spatial dependence in data mining. In: Grossman, R., Kamath, C., Kegelmeyer, P., Kumar, V., Namburu, R. (eds.) Data Mining for Scientific and Engineering Applications, pp. 439–460. Kluwer Academic Publishing, Dordrecht (2001)
22. Lisi, F.A., Malerba, D.: Inducing multi-level association rules from multiple relations. Machine Learning 55, 175–210 (2004)
23. Malerba, D.: A relational perspective on spatial data mining. IJDMMM 1(1), 103–118 (2008)
24. Malerba, D., Ceci, M., Appice, A.: Mining model trees from spatial data. In: Jorge, A., Torgo, L., Brazdil, P., Camacho, R., Gama, J. (eds.) PKDD 2005. LNCS (LNAI), vol. 3721, pp. 169–180. Springer, Heidelberg (2005)
25. Malerba, D., Esposito, F., Lanza, A., Lisi, F.A., Appice, A.: Empowering a GIS with inductive learning capabilities: The case of INGENS. Journal of Computers, Environment and Urban Systems 27, 265–281 (2003)

26. Pekerskaya, I., Pei, J., Wang, K.: Mining changing regions from access-constrained snapshots: a cluster-embedded decision tree approach. Journal of Intelligent Information Systems 27(3), 215–242 (2006)
27. Samet, H.: Applications of spatial data structures. Addison-Wesley, Longman (1990)
28. Sander, J., Ester, M., Kriegel, H., Xu, X.: Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. Data Mining and Knowledge Discovery 2(2), 169–194 (1998)
29. Seeger, M.: Learning with labeled and unlabeled data. Technical report, University of Edinburgh (2001)
30. Shekhar, S., Chawla, S.: Spatial databases: A tour. Prentice Hall, Upper Saddle River (2003)
31. Shekhar, S., Huang, Y., Wu, W., Lu, C.: What's spatial about spatial data mining: Three case studies. In: Grossman, R., Kamath, C., Kegelmeyer, P., Kumar, V., Namburu, R. (eds.) Data Mining for Scientific and Engineering Applications. Massive Computing, vol. 2, pp. 357–380. Springer, Heidelberg (2001)
32. Shekhar, S., Schrater, P.R., Vatsavai, R.R., Wu, W., Chawla, S.: Spatial contextual classification and prediction models for mining geospatial data. IEEE Transactions on Multimedia 4(2), 174–188 (2002)
33. Shekhar, S., Vatsavai, R., Chawla, S.: Spatial classification and prediction models for geospatial data mining. In: Miller, H., Han, J. (eds.) Geographic Data Mining and Knowledge Discovery, 2nd edn., pp. 117–147. Taylor & Francis, Abington (2009)
34. Shekhar, S., Zhang, P., Huang, Y.: Spatial data mining. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 837–854. Springer, Heidelberg (2010)
35. Tobler, W.: A computer movie simulating urban growth in the detroit region. Economic Geography 46, 234–240 (1970)
36. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
37. Vert, G., Alkhaldi, R., Nasser, S., Harris Jr., F.C., Dascalu, S.M.: A taxonomic model supporting high performance spatial-temporal queries in spatial databases. In: Proceedings of High Performance Computing Systems (HPCS 2007), pp. 810–816 (2007)
38. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Komorowski, J., Żytkow, J.M. (eds.) PKDD 1997. LNCS, vol. 1263, pp. 78–87. Springer, Heidelberg (1997)
39. Yin, X., Han, J., Yang, J., Yu, P.S.: CrossMine: Efficient classification across multiple database relations. In: ICDE, pp. 399–411. IEEE Computer Society, Los Alamitos (2004)