International Journal of Research
Available at https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 09
May 2016

Analysis of Spatial Association Rule Mining

Ranu Sahu[1]& Raghvendra Kumar[2]
[1,2]Dept. of Computer Scienece & Engineering, LNCT Group of College, Jabalpur, MP, India
Ranu_sahu_7182@yahoo.com, raghvendraagrawal7@gmail.com

Abstract:

The association rule mining technique emerged with the objective to find novel, useful, and previously unknown associations from transactional databases, and a large amount of association rule mining algorithms have been proposed in the last decade. Their main drawback, which is a well known problem, is the generation of large amounts of frequent patterns and association rules. In geographic databases the problem of mining spatial association rules increases significantly. Besides the large amount of generated patterns and rules, many patterns are well known geographic domain associations, normally explicitly represented in geographic database schemas. The majority of existing algorithms do not warrant the elimination of all well known geographic dependences

Keywords: Data Mining; Distributed Data Mining; Association Rule Mining; Spatial Data Mining; Spatial Association Rules; Weka Tool.

1. Introduction

Due to the increased demand for knowledge discovery in all industrial domains, it is necessary to store all the raw data and to provide useful patterns with respective to the user needs. Generally, the storage of all raw data will be done in a database maintained by respective organizations. Data mining techniques are available to retrieve useful information from large database. Prediction and description are the two fundamental goals of data mining. To full fill these goals many data mining techniques exist such as association rules, classification, clustering and so on. Among these, association rule has wide applications to discover interesting relationship among attributes in large databases. Association rule mining is used to find the rules which satisfy the user specified minimum support and minimum confidence. In the process of finding association rules, the set of frequent item sets are computed as the first step and then association rules are generated based on these frequent item sets.

2. Spatial Data Mining

Spatial describes how objects fit together in space, either among the planets or down here on earth. Spatial data refers to all types of data objects or elements that are present in a geographical space or horizon. It enables the global finding and locating of individuals or devices anywhere in the world. Spatial data base is a database that is enhanced to store and access spatial data or data that defines a geometric space. These data are often associated with geographic locations and features, or constructed features like cities. Data on spatial databases are stored as coordinates, points, lines, polygons and topology. Some spatial databases handle more complex data like three-dimensional objects, topological coverage and linear networks. Spatial data mining is the application of data mining to spatial models. In spatial data mining, analysts use geographical or spatial information to produce business intelligence or other results. This requires specific techniques and resources to get the geographical data into relevant and useful formats.

3. Spatial Association Rules

Spatial association rules consist of an implication of the form X Y, where X and Y are sets of predicates, and at least one element in X or Y is a spatial. While in transactional association rule mining every row in the dataset is usually a transaction and columns are items, in spatial association rule mining every row is an instance of a reference object type (e.g. city), called target feature type, and columns are predicates. Every predicate is related to a non-spatial attribute (e.g. population) of the target feature type or a spatial predicate. Spatial predicate is a relevant feature type that is spatially related to specific instances of the target feature type (e.g. contains_factory). In SAR mining the set F = {f1, f2, …,fk, ..., fn} is a set of non-spatial attributes and spatial predicates, and Ψ (dataset) is a set of instances of a reference feature type, where each instance is a row W such that W ⊆ F. There is exactly one tuple in the dataset Ψ for each instance of the reference feature type.

While the problem of mining non-spatial association rules is performed in two steps, the problem of mining spatial association rules is decomposed in at least three main steps, where the first one is usually performed as a data preprocessing method because of the high computational cost:

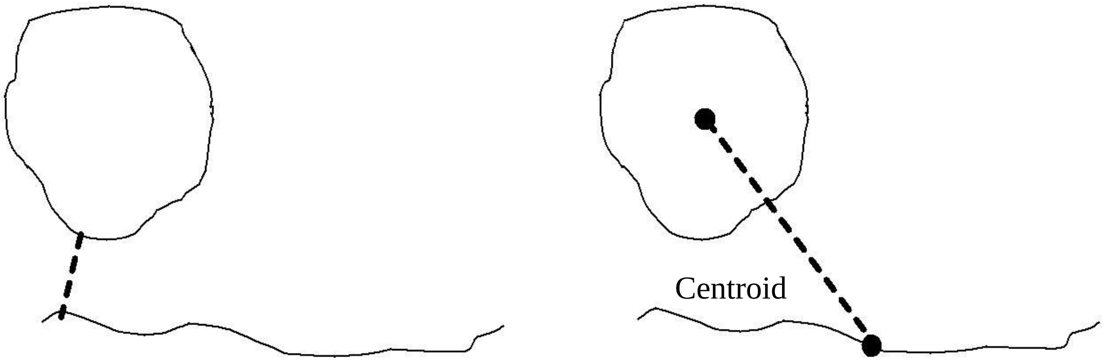
1. Extract spatial predicates: spatial predicate is a spatial relationship (e.g. distance, order, topological) between the reference feature type and a set of relevant feature types;
2. Find all frequent patterns/predicates/sets: a set of predicates is a frequent pattern if its support is at least equal to a certain threshold, called minsup;
3. Generate strong rules: a rule is strong if it reaches minimum support and the confidence is at least equal to a certain threshold, called min conf.

It is well known that spatial joins to extract spatial predicates are the processing bottleneck in spatial data mining, but only little attention has been devoted to this problem. In a top-down progressive refinement method is proposed and spatial approximations are calculated in a first step, and in a second step, more precise spatial relationships are computed to the outcome of the first step. The method has been implemented in the module Geo-Associator of the Geo Miner system, which is no longer available. Ester proposed new operations such as graphs and paths to compute spatial neighborhoods. However, these operations are not implemented by most GIS, and to compute all relationships between all objects in the database in order to obtain the graphs and paths is computationally expensive for real databases. Appice proposed an upgrade of Geo-Associator to first-order logic, and all spatial relationships are extracted. This process is computationally expensive and many spatial relationships might be unnecessarily computed. In, we proposed to use geo-ontologies as prior knowledge to compute only topological relationships semantically consistent, and only among a target feature type and relevant feature types specified by the user. While the above approaches consider different spatial relationships and any geometric object type, a few approaches such as compute only distance relationships for point object types.

Spatial relationships are computed with spatial joins between all instances t (e.g. Porto Alegre) of a target feature type T (e.g. city) and all instances o (e.g. rio de la Plata) of every relevant feature type O (e.g. river) in a set of relevant feature types S (e.g. river, port, street, factory) that have any spatial relationship (e.g. touches, contains, close, far) with T. Being T a set of instances T={t1, t2,…,tn}, S = { O1, Oi,…, Om}, and Oi = { o1, o2,…, oq}, the extraction of spatial predicates implies the comparison of every instance of T with every instance of O, for all O in S.

Existing spatial association rule mining algorithms are in general Apriori-like approaches, i.e., generate candidates and frequent sets, and then extract association or co-location rules. In SAR mining the candidate generation is not a problem as it is in transactional databases. According to the number of predicates is much smaller than the number of items in transactional databases. Therefore, the computational cost relies on the spatial predicate extraction (step a), and depends on the number of instances of the target feature type and the relevant feature types, as well as their respective geometric representation.

The number of spatial association rule mining algorithms is much smaller than transactional rule mining algorithms, and can be classified in two main types. The first is based on quantitative reasoning, which mainly computes distance relationships during the frequent set generation. These approaches deal with geographic data (coordinates x,y) directly. Although they have the advantage of not requiring the definition of a reference object, they have some general drawbacks: usually deal only with points, consider only quantitative relationships, and do not consider non-spatial attributes of geographic data, which may be of fundamental importance for knowledge discovery. For spatial objects/features represented by lines or polygons, their centroid1 is extracted. Indeed, geographic coordinates are transformed into integer values, which reduce precision still further. This process loses significant information, and generates non-real patterns. Figure 3.3 shows an example of how the distance relationship can vary between two spatial features A and B when considering their original geometry (Figure) and their centroid (figure).



distance

distance

centroid

(a) Distance between polygon and line

(b) Distance between the centroid of polygon and the centroid of line

Figure 3.3: Distance relationship for real geometry (left) and for the centroid (right)

The second category is based on qualitative reasoning, which usually considers distance and topological relationships between a reference geographic object type and a set of relevant feature types represented by any geometric primitive (e.g. points, lines, and polygons). Relationships are normally extracted in a first step, in data preprocessing tasks, while frequent sets are generated in another step.

In both qualitative and quantitative reasoning approaches prior knowledge has rarely been used to eliminate irrelevant geographic domain patterns and to produce more interesting rules. Presented an approach which exploits taxonomies of both spatial feature types and spatial relationships only for mining spatial association rules at different granularity levels. Only minimum support is used to prune frequent sets and spatial association rules. A similar method has still been used by Mennis (2005). Clementini extended this method for mining multi-level spatial association rules from geographic objects with broad boundaries.

In both frequent sets and rules are pruned a posteriori. The user can define a pattern constraint and specify how many times a predicate should appear in the frequent sets or in association rules. For example, a pattern constraint such as pattern_constraint removes from the frequent sets the specified predicate when it appears in less than 5 sets. This step is performed after all frequent sets have already been generated. A rule constraint such as body_constraint only shows rules with the specified predicates if they appear in at least 10 association rules, otherwise they are removed. In (APPICE, 2005) this method has been extended with the possibility to specify one more cardinality for a constraint. This method can be used to remove well known rules. For example, a constraint such as pattern_constraint[[intersects(gas station),(intersects(road)],0,0)) would remove from the set of frequent sets all combinations having this pair of predicates. The minimum and maximum cardinality 0 defines that the pair of predicates should not appear in the resultant set of rules.

The pruning method proposed in has some general disadvantages that make the method hard to be used with real databases. First, in data preprocessing all spatial relationships must be computed from geographic databases and transformed to first-order logic. In large geographic databases the extraction of all relationships is non-trivial, and many relationships can be unnecessarily computed. Second, the pruning step is very hard for the data mining user, since for every different relationship or geographic element, a different pattern constraint must be specified to remove non - interesting rules. Moreover, as concluded by the authors about their proposed method a lot of knowledge is required from the data mining user. Third, it is hard for the data mining user to a priori know all possible frequent sets and rules that might have a non -interesting pattern or rule. At lower granularity levels, which will be explained latter in this chapter, for example, such difficulty increases since a different constraint must be specified for every different relationship and feature at a different concept level. For example, to eliminate dependence between gas station and road, some of the constraints that the user has to specify include:

pattern_constraint([contains(X,GasStation),crossed_by(X,Road)],0,0),
pattern_constraint([contains(X,GasStation),contains(X,Road)],0,0),
pattern_constraint([contains(X,GasStation),touches(X,Road)],0,0)),
pattern_constraint([contains(X,Large_GasStation),crossed_by(X,StateHighWay)],0,0),
pattern_constraint([contains(X,Large_GasStation),contains(X,NationalHighWay)],0,0),
pattern_constraint([contains(X,Large_GasStation),contains(X,NationalHighWayBR-116)],0,0)

Conclusion

One of the main limitations of SPADA, which is also a problem of many other relational data mining algorithms, is the requirement of some expertise in data and knowledge engineering. Indeed, the user should know how data are organized in the spatial database (e.g., layers and physical representation of objects), the semantics of spatial relations that can be extracted from digital maps, the meaning of some parameters used in the discretization process and in the generation of spatial association rules, as well as the correct and most efficient way to specify the domain knowledge and declarative bias. Finally, in future work, we will investigate some “interestingness measures” of rules for presentation purposes, so that the user can browse the output XML file of spatial association rules as simply as possible.

References

[1]. Adriaans, P.; Zantinge, D. “Data Mining”. Harlow, England: Addison Wesley Longman, 1997.

[2]. Agrawal, R.; Imielinski, T.; Swami, A. “Mining Association Rules Between Sets of Items In Large Databases”. In: Acm Sigmod International Conference on Management of Data, Sigmod, 1993, Washington, D.C. Proceedings. New York: Acm Press, 1993. P. 207-216.

[3]. Agrawal, R.; Srikant, R. “Fast Algorithms for Mining Association Rules In Large Databases.” In: International Conference On Very Large Databases, Vldb, 20. 1994, San Francisco.

Proceeding. California: Morgan Kaufmann, 1994. P.487 – 499.

[4]. Appice, A. Et Al. "Discovery of Spatial Association Rules in Geo-Referenced Census Data": A Relational Mining Approach. Intelligent Data Analysis, V.7, N.6, P. 542-566, 2003.

[5]. Appice, A. Et Al. "Mining And Filtering Multi-Level Spatial Association Rules With Ares". In: International Symposium On Methodologies For Intelligent Systems, Ismis, 15.,2005, New York. Proceedings: Springer, 2005. P.342-353.

[6]. Bigolin, N. M.; Marsala, C. "Fuzzy Spatial Oql for Fuzzy Knowledge Discovery in Databases". In: European Conference On Principles And Practice Of Knowledge Discovery In Databases, Pkdd, 3., 1998, Nantes, France. Proceedings: Springer-Verlag, 1998. P.246-254.

[7]. Bogorny, V.; Engel, P. M.; Alvares, L.O.: A "Reuse-Based Spatial Data Preparation Framework For Data Mining". In International Conference On Software Engineering And Knowledge Engineering, Seke, 17. 2005, Taipei, Taiwan. Proceedings: Knowledge Systems Institute, 2005a. P.649-652.

[8]. Bogorny, V.; Engel, P. M.; Alvares, L.O. "Towards The Reduction Of Spatial Join For Knowledge Discovery In Geographic Databases Using Geo-Ontologies And Spatial Integrity Constraints". In: Workshop On Knowledge Discovery And Ontologies Of The Ecml/Pkdd, Kdo, 2., 2005, Porto. Proceedings, 2005b. P.51-58.

[9] Bogorny, V.; Engel, P. M.; Alvares, L.O. Geoarm – "An Interoperable Framework To Improve Geographic Data Preprocessing And Spatial Association Rule Mining". In: International Conference On Software Engineering And Knowledge Engineering, Seke, 18., 2006, San Francisco. Proceedings: Knowledge Systems Institute, 2006a. P. 70-84.

[10] Bogorny, V.; Camargo, S.; Engel, P. M.; Alvares, L.O. "Towards Elimination of Well Known Geographic Domain Patterns In Spatial Association Rule Mining". In: Ieee International Conference On Intelligent Systems, Ieee-Is, 3., 2006, London. Proceedings: Ieee Computer Society, 2006a. P. 532-537.

[11] Bogorny, V.; Camargo, S.; Engel, P.; Alvares, L. O. "Mining Frequent Geographic Patterns With Knowledge Constraints". In: Acm International Symposium on Advances In Geographic Information Systems, Acm-Gis, 14, 2006, Arlington. Proceedings 2006c. Accepted For Publication.