INTERNATIONAL JOURNAL OF ADVANCE RESEARCH AND DEVELOPMENT

IJARnD

(Volume 5, Issue 3)

Available online at: www.ijarnd.com

Discovery and comparative study on spatial co-location and association rule mining of spatial data mining

Zaw Lin Oo[1], Mya Sandar Kyin[2]

[1,2]Lecturer, University of Computer Studies, Taungoo, Myanmar

ABSTRACT

Spatial data mining is the process of discovering interesting implicit knowledge in spatial databases that is an important task for understanding and use if spatial data-and knowledge-base and previously unknown, but potentially useful patterns from large spatial datasets; it is an important task for understanding and use the spatial data. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from conventional transaction based database due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation The purpose of in this paper is to do comparative study on spatial co-location rule mining and association rule mining of spatial data mining application based on classical papers, and refine some previous algorithms.

Keywords*— Data Mining, Discovery, Comparative, Spatial*

1. INTRODUCTION

The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns, hence efficient tools for extracting information from geo-spatial data are crucial to make decisions based on large spatial datasets. Spatial objects by definition are embedded in a continuous space that serves as a measurement framework for all other attributes, the framework generates a wide spectrum of implicit distance, directional, and topological relationships, particularly if the objects are greater than one dimension (such as lines, polygons and volumes). Although some objects in typical knowledge discovering applications can be reduced to points in some multidimensional space without information loss, many geographic entities in spatial datasets cannot be reduced to point objects without significant information loss. Characteristics such as the size and morphology of geographic entities can have non-trivial influences on geographic processes.

Spatial data mining can process discovery on interesting and previously unknown but useful patterns from large spatial datasets. Extracting the input data from the datasets is more complex than that of traditional data mining due to huge amount of spatial data, complexity and relationships. Pattern types such as classes, associations, rules, clusters, outliers and trends all have spatial expressions since these patterns can be conditioned by the morphology as well as spatial relationships among these objects. Spatial data mining attributes may have spatial or non-spatial data attributes.

1.1 Characteristics of Spatial Data for Rule Learning

Rule learning is a promising technique for mining patterns of correlations with spatial big data. Rule learning for spatial data will be referred to as spatial rule learning Spatial data includes spatial features that are georeferenced (i.e., their locations are determined within a geographic coordinate system). Thus, spatial data possesses spatial attributes embedded in feature locations on or near the Earth’s surface. The nature of the geographic space, the complexity of the spatial object relationships, and the heterogeneous and sometimes ill-structured nature of geographic data, brings uniqueness to spatial rule learning. At the same time, it renders the standard rule learning techniques inefficient. Special characteristics of spatial information to be considered for spatial rule learning include: 1. appearances of object locations, 2. functional semantic and spatial relationships among objects, 3. functional complexity posed by spatial dependency and spatial heterogeneity, 4. spatio-temporal changes of objects’ semantic and spatial characteristics, and thus, their interactions, and 5. the heterogeneous and sometimes ill-structured nature of geographic data. Without considering these spatial factors, classical rule learning approaches are a poor-fit to mining spatial data tasks. [1] [9] [10].

Evolving from classical rule learning techniques, the objective of spatial rule learning is then to extract the frequent occurrence of both semantic and spatial attributes of analyzed spatial features or of object locations (co-locations). Spatial predicates expressing spatial characteristics and relationships of the learning units are often used in addition to non-spatial predicates when applying the APRIORI algorithm. The process of materializing all possible spatial characteristics and relationships to generate a complete set of spatial predicates become crucial. This task is, however, non-trivial. The achievements and remaining challenges in spatial rule learning can be broadly discussed under two learning problems: 1) spatial association rules and 2) co-location rules.

1.2 Spatial Data Mining

1.2.1 Major Techniques in Spatial Data Mining: Pattern types such as classes, associations, rules, clusters, outliers and trends all have spatial expressions since these patterns can be conditioned by the morphology as well as spatial relationships among these objects. In this section, we briefly review the major techniques in spatial data mining.

1.2.2 Spatial Classification: The classification techniques map spatial objects into meaningful categories that consider the distance, direction or connectivity relationships and the morphology of these objects. The spatial buffer can be used to classify objects based on attribute similarity and distance-based proximity. Ester et al. (1997) generalize this approach through a spatial classification-learning algorithm that considers spatial relationships defined as path relationships among objects in a defined neighbourhood of a target object. These paths are highly general and can be defined using any spatial relationship.[11]

1.2.3 Spatial Association: Spatial association rules are association rules that defined above different spatial predicates. Koperski and Han (1995) proposed this pioneering this concept, providing detailed descriptions of their formal properties as well as a top-down tree search technique that exploits background knowledge in the form of a geographic concept hierarchy. [9]

1.2.4 Spatial Clustering: Spatial clustering algorithms exploit spatial relationships among data objects in determining inherent groupings of the input data. Since finding the optimal set of k clusters is intractable (where k is some integer much smaller than the cardinality of the database), a large number of heuristic methods for clustering exist in the literature.

1.2.5 Spatial Outlier Analysis: The spatial outlier is a spatially-referenced object whose non-spatial attributes appear inconsistent with other objects within some spatial neighbourhood. Unlike a spatial outlier, this definition does not imply that the object is significantly different from the overall database as a whole: it is possible for a spatial object to appear consistent with the other objects in the entire database but nevertheless appear unusual with a local neighbourhood.

1.3 Categorization of spatial rules

Various kinds of rules can be discovered from spatial databases in general, such as co-location rules, spatial association rules, spatial characteristic rule, and spatial discriminant rule.

1.3.1 Spatial association rules: A spatial association rule is a rule which describes the implication of one or a set of features by another set of features in spatial databases. Spatial association rule is of the form X□Y, where X and Y are sets of spatial or non-spatial predicates and can be defined using the minimum support and minimum confidence.
Eg. is_a(X, Park) ^ close_to(X, Police_station) □ close_to(Public_Facility) [92.59%]
The spatial association rules have at least one of the predicates is spatial. Co-location pattern can be seen as a special type of association rule, subsets of spatial objects that are frequently located together.

1.3.2 Co-location rules: The co-location patterns represent the subsets of the Boolean spatial features whose instances are often located in a close geographic proximity, for example, frontage roads and highways in metropolitan road maps. The co-location rules can be a model to analyze and infer the presence of certain Boolean spatial features in the neighborhood of instances of other Boolean spatial features.

A co-location rule is of the form: $C_1 \rightarrow C_2(p, cp)$, $C_1 \cap C_2 = \varnothing$, where C_1 and C_2 are subsets of Boolean spatial features, p is a number representing the prevalence measure and cp is a number representing the conditional probability.
Some major types of co-location rule are as follows:

- *Reference-centric co-location:* In reference-centric co-location, there is a spatial predicate *sp* such that $sp(X, Y)$ is true for specific reference feature X, and for all $Y \in I$, $Y \neq X$. For example, if the spatial predicate is *close_to*, for every item Y in I, *close_to*(X, Y) must be true, but *close to*(Y_1, Y_2) is not necessary to be true if Y_1 and Y_2 are not reference features. Instances of this pattern exhibit *star* pattern with reference feature X as the center of the star.
- *Event-centric co-location:* In event-centric co-location, there is a spatial predicate *sp* such that $sp(X, Y)$ is true for all $X \in I$, $Y \in I$, $Y \neq X$. For example, if the spatial predicate is *close to*, for every item X, Y in I, *close_to*(X, Y) must be true. Instances of this pattern exhibit *clique* (fully connected subgraph) pattern where all pairs in the set satisfy the spatial predicate.
- *Complex co-location;* Complex co-location rule may include negative co-location (absence), and self co-location [2]. Example of complex co-location rule is that crime C is more likely to occur near a subway station S, with no lighting L: $S, –L \rightarrow C+$.

Other rules like spatial characteristic rule and spatial discriminant rule can also be mined from spatial databases, the spatial characteristic rule is a general description of spatial data, for e.g. rule describing general price range of the houses in various geographic regions in a city. Whereas the spatial discriminant rule is a general description of the discriminating or contrasting features of a class of spatial data from other classes, for e.g. comparison of price range of houses in different geographical regions.

1.4 Algorithms for spatial rule mining

1.4.1 Spatial association rule mining: To find spatial association rules hidden inside the spatial datasets, a top-down, progressive deepening search technique for spatial association rule mining has been proposed by Han and Koperski.K [1]. This technique is iterative in nature and is essentially a Apriori, it first searches at the highest concept level for large patterns and strong implication relationships at a coarse resolution scale, where several approximate spatial computation algorithms.

such as R* tree and Minimum Bounding Rectangle method (MBR) can be employed. This will generate the large 1-predicates. Then, the algorithm deepens the search to lower levels for only the candidate spatial predicates generated at the coarse level. For example, general_close_to predicate is replaced by the detailed predicates such as intersect, adjacent_to or close_to, etc. Rows with support count less than the minimum support threshold can be removed from the predicate table and there is no need to be considered in the next step. This process can be deepening further by predicates with certain geographical feature, such as Adjacent_to_East Lake, here the east lake is a specific spatial object. This process continues to find the large k-predicates and it will stop when no large patterns can be found.

When all large predicates are found, we can generate spatial association rules from the large predicates table. For any large predicates A and B from the predicate table, if A is not a subset of B, the support for the predicate A^B is computed, we can add the rule A→B to the result set as long as the sup(A^B) / sup(A) > minimum confidence. By scanning the large predicate table and computing the confidence ratio, we can find all possible association rules.

1.4.2 Spatial co-location rule mining: Several algorithms have been proposed for spatial co-location rule mining. Two of them are the Co-location Miner Algorithm [2] proposed by Sekhar and Huang, and Synch Sweep algorithm [8] proposed by Zhang et. al.

1.4.2.1 Co-location Miner Algorithm: The co-location miner algorithm is a breadth first search algorithm for mining event centric co-location rules. The algorithm keeps tracking all prevalent candidates and their instances to generate prevalence co-location patterns. From all prevalence co-location patterns, the co-location rules can then be generated.

The item set generation is very similar to Apriori algorithm. It starts from 1 co-location set C_1, and generate k co-location candidates C_k from $k – 1$ co-location candidates $C_{k–1}$. If two candidates in $C_{k–1}$ differ only at the last feature, they can be joined to form k co-location pattern. For example, if A,B,C and A,B,D are in $C_{k–1}$, then A,B,C,D is inserted into C_k. All instances in $C_{k–1}$ are prevalence co-location patterns, and therefore if certain feature combinations are not found in $C_{k–1}$, they are not prevalent. Using this fact, C_k can be pruned further by removing candidates if their subsets are not prevalent enough. For example, if B,C,D is not found in $C_{k–1}$, then A,B,C,D should not be in C_k, even though A,B,C and A,B,D are in $C_{k–1}$.

The generation of co-location instances is also similar to the candidate generation. It starts from 1 co-location instances T_1, and generate k co-location candidates T_k from $k – 1$ co-location candidates $T_{k–1}$. If two instances in $T_{k–1}$ differ only at the last point, they can be joined to form k co-location instance if the distance between the last points is within the distance threshold. For example, if P1,P2,P3 and P1,P2,P4 are in $T_{k–1}$, then P1,P2,P3,P4 is inserted into T_k. All instances in $T_{k–1}$ are prevalence co-location instances, and therefore if P1,P2,P3 is in $T_{k–1}$, the distances among them are already within the distance threshold. Using this fact, only the distance between P3 and P4 need to be computed to construct instance P1,P2,P3,P4. The co-location rules are then generated for each prevalent candidate co-location patterns by enumerating all possible subset of the patterns, and pruning them using conditional probability threshold.

1.4.2.2 Synch Sweep Algorithm: The synch sweep algorithm is a depth first search algorithm for mining reference centric co-location rules. Even though the original synch sweep algorithm was designed for mining reference centric, several extensions for mining event centric or more general pattern were also proposed [8].

The basic synch sweep algorithm works by sorting all the feature instances by x-coordinate. For each feature, it then sweeps the instances by its x-coordinate, and tries to find other feature instances that are close in their x-coordinate. In this way, all instances that are not close in x-coordinate will be pruned, and will not be considered at all. For the instances that are close in x-coordinate, their actual distances are then computed. If they are close, then it will be considered as an instance of the co-location pattern. At the end of synch sweep algorithm, all prevalent co-location patterns will be collected. The co-location rules are then generated for each prevalent co-location patterns by enumerating all possible subset of the patterns, and pruning them using conditional probability threshold.

2. CONCLUSION

In this paper studied and discovered the classical algorithms in spatial data mining, such as spatial association rule mining algorithm, co-location rule miner algorithm and synch sweep algorithm. The spatial mining algorithms can be improved by building complex spatial index on the spatial datasets, so that when we try to filter out most irrelevant instances for certain feature, we can directly retrieve the relevant instance according to the spatial index, no scanning of the whole datasets is needed, hence we can significantly decrease the processing time. Fast algorithms for computing distances between different geometries should be proposed to reduce processing time. In association rule mining, for large datasets where transactions for predicates can be huge, we can store these transactions in a relational database instead of in memory.

3. ACKNOWLEDGEMENT

I would like to express my sincere thanks to Dr. Ei Ei Hlaing, Rector, University of Computer Studies, Taungoo, Pegu Regional Division, Myanmar, for her kind permission to this paper. I would like to express my profound gratitude and appreciation to Dr. Thin Thin Htwe, Head of Faculty of Information Science, University of Computer Studies, (Taungoo).

4. REFERENCES

[1] Koperski, K., and Han, J., Discovery of Spatial Association Rules in Geographic Information Databases (1995)
[2] Shekhar, S. and Huang, Y., Discovering Spatial Co-Location Patterns: A Summary of Results (2001)

[3] Malerba, D., Esposito, F. and Lisi, F., Mining Spatial Association Rules in Census Data (2001)

[4] Rakesh, A., and Ramakrishnan. S., Fast Algorithms for Mining Association Rules

[5] Hipp, J., Güntzer, U. and Nakhaeizadeh, G. (2000) "Algorithms for association rule mining: A general survey and comparison," SIGKDD Explorations, 2, 58-64.

[6] Shekhar, S. and Chawla, S. (2003) Spatial Databases: A Tour, Upper Saddle River, N. J.: Prentice-Hall.

[7] Harvey J. Miller. "Geographic Data Mining and Knowledge Discovery" Handbook of Geographic Information Science, in press

[8] Zhang, X, Mamoulis, N., Cheung, D.W., Shou, Y., "Fast Mining of Spatial Collocations." (2004).

[9] Mennis, J., & Guo, D. (2009). Spatial Data Mining and Geographic Knowledge Discovery - An Introduction. Computers, Environment and Urban Systems, 33(6), 403-408.

[10] Miller, H., & Han, J. (Eds.). (2009). Geographic Data Mining and Knowledge Discovery: An Overview: CRC Press, Taylor and Francis Group.

[11] Ester M., Kriegel H.-P., and Sander J. 1997 "Spatial Data Mining: A Database Approach", Proc. 5th Int. Symp. on Large Spatial Databases, Berlin, Germany, pp. 47-66.

BIOGRAPHY

Zaw Lin Oo
Faculty of Information Science,
University of Computer Studies (Taungoo), Pegu Regional Division, Myanmar

Mya Sandar Kyin
Faculty of Computer Science,
University of Computer Studies (Taungoo), Pegu Regional Division, Myanmar