An efficient Framework of Spatial Fuzzy Association Rule Mining

Sanghamitra Panda, Dr. Laxman Sahoo

Abstract: Although there are many types of databases, in almost all application areas, however the basic functions of the database remains same but they require a vastly wider range of techniques integrating from classical to specialized approaches. The specialized data which has been handled today is spatial data, which has emerged as central to many applications which include geographical information systems (GIS), computer-aided design, robotics, image processing etc., all of which have at their core spatial objects. In a relatively very short period, research of spatial database has been enhanced by the use of data mining and fuzzy techniques in order to give better outcomes. Spatial database, stores a large number of space related data and has many different characteristics with relational databases, such as it contains topology and distance information, which is generally represented as multi relational data. So extracting knowledge from spatial database is more difficult than extracting knowledge from traditional relational database. The support and confidence of an association rule mining generally has been used to extract implicit knowledge from spatial database. However some attributes of spatial data does not allow to make the support more effective because of the crisp properties, of the harsh boundaries created by the partitioning of quantitative attributes into intervals generating unnatural division of the data and subsequently of the resultant rules. To overcome this difficulty the fuzzy association rule has been used, but it is found that the extraction of the knowledge from multiple spatial objects is not an easy task. In order to find multidimensional association rule, we have, developed a new framework, which is the main objective in this research work. For this purpose we have used a concept of neighbourhood relation, upon which spatial literals can be built, to find association rules involving spatial relations among spatial objects using fuzzy association rule mining. This new framework generates a multidimentional association rule based on spatial relations, which has better potential for extracting knowledge from large spatial database to enable decision makers to capture more complex pattern and spatial occurrences in a very efficient way.

Keywords: Spatial data mining, spatial association rule mining, fuzzy association rules, spatial fuzzy association rules.

---

1 INTRODUCTION

Spatial data mining, i.e. mining knowledge from large amount of spatial data, is a highly demanding field because huge amount of spatial data have been collected in various applications. These spatial data are the data related to objects that occupy space [10]. Spatial data carries topological and distance properties. A spatial database stores data about spatial objects (represented by spatial data types) and spatial relationships among such objects [5]. There is a certain interaction between adjacent objects, so the relationship between the spatial objects is more complex (It is not about merely adding topological and directional relations, but also to measure relationship referred to spatial location and the distance between spatial objects) [5][11].

When mining spatial data, spatial association rules are needed.The spatial association rule mining is used for extraction of implicit knowledge from spatial database [1].The intent is to discover relationship between spatial objects. For example one may find the rule as **" If a house is close to the highway, then it is expensive".** Although such rules are usually not 100% true, they carry some nontrivial knowledge about spatial associations and thus it is interesting to mine them from large spatial databases. When the rule uses the terms or properties like "close","expensive", these terms may not be concise and meaningful enough for human experts to easily obtain knowledge from those rules discovered. These properties are expressed earlier using some user defined threshold values as objects are close if their distance is less than some user defined threshold "d". Still some attributes of spatial data does not allows to make association rules more effective because of the crisp properties, of the harsh boundaries created by the partitioning of quantitative attributes into intervals which generate unnatural division of the data and subsequently the resultant rules. To overcome this difficulty fuzzy association rule has been introduced.

The use of fuzzy techniques has been considered as one of the key components of data mining system because of its affinity with the human knowledge representation [2]. Still extraction of knowledge from multiple spatial objects remains a challenge for us, because the dynamics and complexity of the spatial objects, captured by spatial predicates are often overlooked. In this paper a framework is proposed for mining multiple spatial objects using fuzzy association rules.

- *Sanghamitra Panda, School of Computer Engineering, KIIT University, Bhubaneswar, Odisha,India, +919439468152*
- *Dr. Laxman Sahoo, Professor and Head of Database Group, KIIT University, Bhubaneswar, India, +919692259550*

2 BACKGROUND AND RELATED WORK

Data mining is more specifically defined as the application of specific algorithms for extracting patterns from data (Fayyad et al. 1996, p.39). As a sub-field of data mining, spatial data mining or knowledge discovery in spatial database refers to the discovery of interesting and useful but implicit knowledge or patterns in spatial databases.The challenges arising when mining spatial data include geographic measurement frameworks, spatial dependency, complexity of spatial objects and rules. From the spatial data mining literature survey we found, the output patterns are spatial clusters, spatial outliers (Zhang, Lu et al. 2003), movement patterns (Laube et al. 2005) and spatial association rules (koperski and Jiawei han 1995).

In an early work, Koperski and Han (1995) investigated spatial association rules (SAR) among a set of spatial and possibly some non spatial predicates. They present optimization techniques for association rules with a spatial antecedent (Ant) and a non spatial consequent (Cons), $Ant \rightarrow Cons$ as $(\text{is\_a}(x, \text{house}) \wedge \text{close\_to}(x, \text{river}) \rightarrow \text{is\_expensive}(x))$ and with a non spatial antecent and spatial consequent expressed as $(\text{is\_a}(x, \text{gasStation}) \rightarrow \text{close\_to}(x, \text{highway}))$. Kuok et al. (1998) introduced the fuzzy association rules of the form, 'If X is A then Y is B' to deal with quantitative attributes, where crisp intervals are replaced by fuzzy intervals. In other words, strict membership functions are replaced by fuzzy membership functions giving scores in the range [0.1]. More recently Patrick Laube and Berg et al. (2008) describe a theoretical approach for measuring the quality of an association rule by support and confidence using fuzzy association rule.

This study aims to improve the quality of an association rule by introducing a framework for multiple spatial objects using fuzzy association rule mining. The emphasis is on developing a methodology to incorporate spatial relations and spatial dependencies under the form of spatial associations.

3 Spatial Datamining

In many non spatial applications the numerical values of the attributes are given and it is rather straight forward to map these values to scores in the range [0, 1]. In the spatial domain, however, this is not the case. Spatial database stores a large number of space related data and has many different characteristics with relational database, such as it contains topology and distance information. The spatial data can be represented as multirelational data. But the multirelational approach generates huge amount of rules for spatial pattern. Figure 1 shows an example of spatial data stored in relational database, where the feature types are roads, houses and water resource are different relations (database tables) with spatial (shape) and nonspatial attributes.

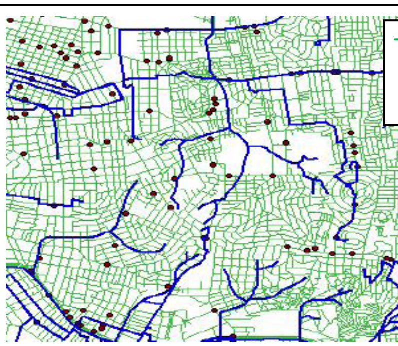

Fig. 1. Example of implicit spatial relationships

Spatial features for different spatial objects are represented in table 1, 2 and 3.

TABLE 1
Road

Gid	Type	Shape
1	Highway	Multiline[(X1,Y1),(X2, Y2),…]
2	Street	Multiline[(X1,Y1),(X2, Y2),…]

TABLE 2
House

Gid	Name	Shape
1	BR	Point[(X1,Y1)]
2	ELF	Point[(X1,Y1)]
3	ESSO	Point[(X1,Y1)]

TABLE 3
Water Resource

Gid	Name	Shape
1	Jauci	Multiline[(X1,Y1),(X2, Y2),…]
2	Uruguai	Multiline[(X1,Y1),(X2, Y2),…]

The spatial attributes of spatial object types, are represented by shape on above table. Because of these relationships real world entities can affect the behavior of other features in the neighbourhood. This makes spatial relationship be the main characteristic of geographic data to be considered for data mining and knowledge discovery.

3.1 Spatial Fuzzy Association Rules

Association rules have been used for the detection of spatial patterns from spatial objects. Association analysis is one of the most widely research topics in data mining. The main focus of association rule mining is to generate hypothesis rather than to test them as is commonly achieved using statiatical techniques. We have seen that in order to apply the theory of fuzzy ARM (FARM), we need to define score functions in the range [0, 1] for the antecedent and for the consequent of the association rule. The task of defining score functions for attributes can be split into two subtasks. Consider as an example the condition "close to the road". Then the first task is to determine how to quantify distance to the road, and the second task is to convert this distance to a score in the range [0, 1]. Hence, the concepts of support and confidence are redefined. The support and confidence of the rule $Ant \rightarrow Cons$ can be expressed using t-norm (Dubois et al. 2006). The support of $Ant \rightarrow Cons$ is given by

$$\text{support} = \sum_{X} \text{SAnt}(x) \otimes \text{SCons}(x) \qquad (1)$$

and the confidence is given by

$$\text{confidence} = \frac{\sum_X \text{SAnt}(x) \otimes \text{SCons}(x)}{\sum_X \text{SAnt}(x)} \qquad (2)$$

A t-norm that is used often is the product t-norm, which multiplies its arguments.

4 DATA PROCESSING FRAMEWORK

Spatial data processing is one of the main process in spatial data mining. The framework represented here has the capability to handle the spatial component, as well as solves the purpose of knowledge discovery process from spatial database using spatial FAR mining.

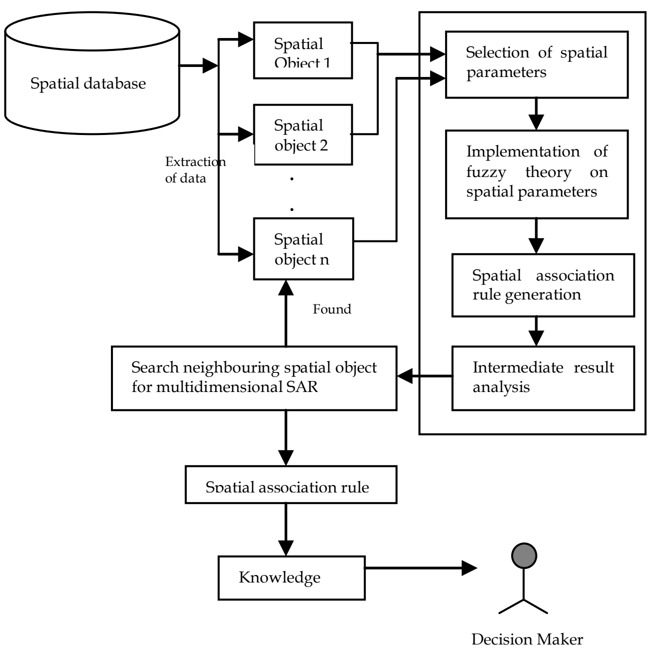

Fig. 2 Framework for spatial data processıng to generate SFARM(Spatial fuzzy association rule mining)

5 A METHOD FOR MINING SPATIAL ASSOCIATION RULE

5.1 An example of Mining Spatial Fuzzy Association Rules

Let the spatial database to be studied adopt an extended-relational data model and a SFARM (spatial fuzzy associateon rule mining) framework. The study of spatial association relationship is confined to a city with the following database relations for organizing and representing spatial objects.

1. Mall (M_id, name, spatial, …)
2. Houses (H_id,name, spatial, …)
3. Road (R_id, name, spatial, …)

In the above relational schema "spatial" represents a spatial object (a point, multiline, etc.). Suppose a user is interested in finding within the map of a city the strong spatial association relationship between large malls and other near by objects including houses, hotels, and major highways. The Geo Miner query is used to find relevant attributes. Moreover "close_to" is a condition dependent attribute and is defined by a set of knowledge rules. For example, if X is a town and Y is a country, then X is close to Y if their distance is within 80kms. For our example we have taken the relation as "If a mall has neighbouring houses with distance less than 20km then mall is profitable".

Rule 1: $(\text{is\_a}(x, \text{mall}) \wedge \text{close\_to}(y, \text{house}) \rightarrow \text{is\_profitable})$

Then the first task is to quantify the distance using Euclidean measure. For this rule the "spatial" parameters are points, from their X and Y coordinates distance can be measured between mall and houses. Otherwise when "spatial" parameters are represented by lines or polygons their centroids are taken, and distance is measured between the centroid. Then the second task is to convert the distance (dist) to a score in the range [0, 1]. Here we have used two threshold parameters d_{min} (taken for minimum distance) and d_{max} (taken for maximum distance), with $d_{min} < d_{max}$ and define the score as

$$\text{Score (d)} = \begin{cases} 1 & \text{if dist} <= d_{min} \\ \dfrac{d_{max} - \text{dist}}{d_{max} - d_{min}} & \text{if } d_{min} < \text{dist} <= d_{max} \\ 0 & \text{if dist} > d_{max} \end{cases} \tag{3}$$

The best values for the parameters d_{min} and d_{max} depends on the application, and can be determined by the user. Suppose we have a score $S_{Ant}(x)$ in the range [0, 1] that captures to what extent a house X is close to the mall. Then we have to map the profit (p) of mall to a score $S_{Cons}(x)$ in the range [0, 1] to capture what extent a mall is profitable. For example if the houses distance to the mall are within 5km that gives more profit to the mall which is $10000 annualy, gets a score of 1 or the distance is more than 20km then profit is less than $500, gets a scoreof 0, and in between we interpolate the score as

$$\text{Score (p)} = \begin{cases} 1 & \text{if } p >= p_{max} \\ \dfrac{p - p_{min}}{p_{max} - p_{min}} & \text{if } p_{min} <= p <= p_{max} \\ 0 & \text{if } p < p_{min} \end{cases} \tag{4}$$

Here we have taken two threshold values p_{min} (for minimum profit), p_{max} (for maximum profit). The best values for the parameters p_{min} and p_{max} depends on the application, and can be determined by the user. Using product t-norm in the definition of support gives us the spatial support and cofidence.

Then the turning point is that if the antecedent and/or the consequent of a spatial association rule which is composed of several predicates for multiple spatial objects, then it generates a multidimensional association rule. For example "If a mall has neighbouring houses with distance less than 20km and mall is close to highways then the mall is more profitable". From this we found the rule as

Rule 2:

$(\text{is\_a}(x, \text{mall}) \wedge \text{close\_to}(y, \text{house}) \wedge \text{close\_to}(z, \text{road}) \rightarrow \text{is\_profitable})$

Here we can follow the same procedure to find the score, for

ISSN 2229-5518

distance between road and mall and also for profit generated due to the road. Then product t-norm is used to find spatial support and confidence for the rule. When two rules are compared with their support and confidence, it can be found that second rule gives the better result than first one.

5.2 Data Processing for Mining Spatial Fuzzy Association Rule

In general, spatial data preprocessing can be formulated as follows

Input:

1. Spatial data(base)
2. Spatial association rule mining algorithm
3. Set of thresholds

Process:

Find spatial relation based on parameters determined of input .

This process consists of:

1. Feature selection based on spatial parameters.
2. Implementation of fuzzy theory on spatial parameters.
3. Spatial association rule generation.
4. Result analysis from support and confidence.
5. Searching neighbouring spatial object to generate multidimensional spatial association rule.
6. Transform into output form.

Output:

A multi dimensional spatial association rule for the relevant set of objects and relations.

Feature selection is performed to find spatial attributes from spatial objects using any query language. Fuzzy theory is implemented on spatial parameters. This process generates fuzzy association rule in spatial context, which can be used for further analysis.

6 Conclusion

This paper presents a framework for extracting knowledge from large spatial database using spatial fuzzy association rule mining. This framework will help decision makers for business analysis process. Main steps in this spatial data processing is that, multiple spatial objects are associated in neighbouring relation to form multidimensional spatial association rule. This approach can be further enhanced for implementation on various application areas of spatial data mining.

References

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann, pp. 600-607, 2003.

[2] Keith C.C Chan, Wai-Ho Au, "Mining Fuzzy Association Rules", Department of computing, The Hong Kong Polytecnic University, Hung Hom, Kowloon, Hong Kong, pp. 209-215, CIKM 1997.

[3] Chan Man Kuok, Ada Fu, Man Hon Wong, "Mining Fuzzy Association Rules in Databases", ACM SIGMOD, Record Vol. 27, pp. 41-46, 1998.

[4] Harvey J. Miller and Jiawei Han, "Geographic data mining and knowledge discovery: An overview", in H. J. Miller and J. Han (eds.) Geographic Data Mining and Knowledge Discovery, London, pp. 3-32, 2001.

[5] Roy Ladner, Frederick E Petry, Maria A Cobb, "Fuzzy Set Approaches to Spatial Data Mining of Association Rules", Transactions in GIS, pp. 123-138, 2003.

[6] Patrick Laube, Mark de Berg, Marc van Kreveld, "Spatial Support and Confidence for Spatial Association Rules", pp. 575-593, SDH 2008.

[7] S. Shekhar, C.T. Lu, P. Zang, "A Unified Approach to Detecting Spatial Outliers", Geoinformatica, An International Journal on Advances of Computer Science for Geographic Information System, Vol. 7, No.2, pp. 139-166, June 2003.

[8] Krzysztof Koperski, Jiawei Han, "Discovery of Spatial Association Rules in Geographic Information Databases", SSD, pp. 47-66, 1995.

[9] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From Data mining to Knowledge Discovery in Databases", AAAI/MIT Press, AI Magazine, Vol. 17, pp. 37-54, 1996.

[10] P. Laube, S. Imfeld and R. Weibel, "Discovering Relative motion Patterns in Groups of moving point objects", International Journal of Geographic Information Science, pp. 639-668, 2005.

[11] Richard Frank, Martin Ester, Arno Knobbe, "A Multi-Relational Approach to Spatial Classification", SIGKDD, pp. 309-318, 2009.

[12] Imam Mukhlash, Benhard Sitohang, "Spatial Data Preprocessing for Mining Spatial Association Rule with Conventional Association Mining Algorithms", International Conference on Electrical Engineering and Informatics, Indonesia June 17-19, 2007.

[13] Vania Bogorny, Paulo Martins Engel, Luis Otavio Alvares, "GEOARM: An Interoperable Framework to Improve Geographic Data Preprocessing and Spatial Association Rule Mining", SEKE, pp. 79-84, 2006.

[14] Theresa Beaubouef, Frederick E. Petry, "Fuzzy and Rough Set Approaches for Uncertainty in Spatial Data", Methods for Handling Imperfect Spatial Info. , Springer, STUDFUZZ 256, pp. 103–129, 2010.

Sanghamitra Panda received her MCA degree in the year 2009 from CEB, Bhubaneswar affiliated to BPUT University, Odisha and currently pursuing M.Tech in Computer Science Engineering from KIIT University, Bhubaneswar. She was involved in C# and .NET application based project for CRRI from STPI, Bhubaneswar during her MCA. Her research area includes Database systems, Spatial Data Mining, Cloud Computing and soft computing.

Dr. Laxman Sahoo received his Ph.D in 1987 from IIT, Kharagpur. Presently, he is professor and head of Database Engg. Group at KIIT University, Odisha. He served as Director/Coordinator/Head of Department in BITS Pilani, BITS Ranchi and Lucknow Indian Engg. College. He has guided over 400 Master Degree students. He is associated with many professional bodies as a member/chair person of technical committee and conference. He has published and presented many Research papers in national and international Journals. He is also author of a good number of computer related books. His research area includes VLSI design, DBMS, Spatial Data mining, AI and Fuzzy Expert Systems.