International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 10, October 2014 **ISSN 2319 - 4847**

A Survey on Co-location and Segregation Patterns Discovery from Spatial Data

Miss Priya S. Shejwal[1], Prof.Mrs. J. R. Mankar[2]

[1] Savitribai Phule Pune University,Department of Computer Engineering,
K..K..Wagh College Nashik,India

[2] Savitribai Phule Pune University Department of Computer Engineering,
K.K..Wagh College Nashik,India

ABSTRACT

Spatial data mining is the extraction of spatial patterns, spatial relations and implicit knowledge, that are not explicitly stored in databases. In spatial domains, interaction between features generates two types of interaction patterns. A positive interaction (aggregation) brings a subset of features close to each other whereas a negative interaction (inhibition) results in subsets of features segregating from each other. Co-location patterns, intended to represents positive interactions, have been defined as subsets of Boolean spatial features whose instances are often seen to be located at close spatial proximity. Segregation patterns, representing negative interactions, can be defined as subsets of Boolean spatial features whose instances are infrequently seen to be located at close spatial proximity. In this paper, different approaches to discover co-location and segregation patterns like Rule based approach, event centric model approach and approaches for finding complex patterns are discussed. There are several measures which are used to compute spatial interaction.

Keywords:- Spatial data, co-location patterns, segregation patterns, spatial interaction, complex patterns.

1. INTRODUCTION

Spatial data mining is an application of data mining methods in spatial data. The main objective of spatial data mining is to find pattern in data with respect to geography. Spatial data mining is a process of finding interesting and useful patterns from spatial data that are generated from geographic space. It is very difficult to extract patterns from spatial data than transaction data because; it is due to the complexity of spatial data types, spatial relationship and autocorrelation. Advanced spatial data collecting system, such as NASA, Earth's Observing System (EOS) and Global Positioning System (GPS), has been accumulating increasingly large spatial data set. For insatance, since 1999, more than a terabyte of data has been produced by EOS every day. These spatial data set with explosive growth rate are considered nuggets of valuable information. The automatic discovery of interesting, potentially useful, and previously unknown pattern from large spatial data set is being widely investigated via various spatial data mining technique. Classical spatial pattern mining method includes spatial clustering, spatial characterization spatial outlier detection, spatial prediction, and spatial boundary shape matching.

2. INTERACTION PATTERN MINING

Interaction pattern mining can lead to important domain related insights in areas such as ecology, biology, epidemiology, earth science, and transportation. Pattern mining is data mining method that find existing pattern in data. A data set consist of instance of several Boolean spatial features, each represented by a distinct shape. Data input of spatial data mining is complex than input of transaction data mining. Spatial data mining include extended object such as point, lines and polygons. There are two different type of attributes used in spatial data mining as data input. They are non-spatial attributes and spatial attributes. In non-spatial attributes, they are used to characterize non spatial features of objects like name, population and unemployment rate of cities. In spatial attributes they are used to define the spatial location and spatial object. It includes longitude, latitude, elevation, slop etc. Boolean spatial features describe the presence and absence of geographic object types at different location in a 2 dimensional or 3 dimensional metric space , such as surface of the earth. Example of Boolean spatial features include plant and animal species, road type, cancer, crime, and business type. In spatial domains, interaction between features generates two types of interaction patterns: co-location and segregation patterns.

2.1 Co-location Patterns

A spatial co-location pattern is a pattern which represents a subset of spatial features whose instance frequently located in close geographic proximity.Mining of spatial co-location patterns problem can be related to various application domain. E.g. in location based services, various services are requested by service subscribers from their mobile PDA's with location devices such as GPS. Some type of services may be requested in proximate geographic area such as

finding the nearest Italian restaurant and nearest parking place. Location based service providers are very eager in finding what services are requested frequently together and located in spatial proximity. This information can help them improve the effectiveness of their location based recommendation system where users request a service in a nearby location. By knowing co-location pattern in location based services may enable the use of pre-fetching to speed up service delivery. In ecology, scientists are interested in finding frequent co-occurrences among spatial features such as substantial increase or drop in vegetation, drought and extremely high precipitation. The previous studies on co-location pattern mining emphasize frequent co-occurrences of all involved features.These marks off some valuable pattern involving rare spatial features. A spatial feature is rare if its instances are substantially less than those of the other features in a co-location. This definition of "rareness" is relative with respect to other features in a co-location. A feature could be rare in one co-location but not rare in another. For instance, the Nile crocodile and the Egyptian plover (a bird that has a symbiotic relationship with the Nile crocodile) are often seen together giving rise to a co-location pattern {Nile crocodile, Egyptial plover}. In urban areas, there are co-location patterns such as {shopping mall, restaurant}.

2.2 Segregation Patterns

Segregation patterns have yet to receive much attention. Segregation patterns, representing negative interactions, can be defined as subsets of Boolean spatial features whose instances are infrequently seen to be located at close spatial proximity (i.e., whose co-locations are "unusually" rare). Examples of segregation patterns are common in ecology, where they arise from processes such as the competition between plants or the territorial behaviour of animals. For instance, in a forest, some tree species are less likely found closer than a particular distance from each other due to their competition for resources.

3. RELATED WORK

In this section, an overview of mining co-location and segregation patterns from spatial data is provided. The objective of this survey clearly understands the limitations of existing approaches. Extraction of knowledge from huge amount of spatial data is a big challenge. Many technologies are available for knowledge discovery in large spatial database. Different type of data mining approaches are available, mainly extraction of implicit knowledge, spatial relation and other patterns which are not explicitly stored in spatial databases.

R. Agrawal and R. Srikant in [2] developed Co-location mining algorithms which are motivated by association rule mining (ARM). Most of the algorithms [1], [3]–[6] adopt an approach similar to the Apriori algorithm proposed for ARM in [2], by introducing some notion of transaction over the space, and a suitable prevalence measure. Shekhar *et al*. in [3] discuss three models (reference feature centric model, window centric model, and event centric model) that can be used to materialize "transactions" in a continuous spatial domain so that a frequent itemset mining approach can be used. Using the notion of an event centric model a mining algorithm is developed which utilizes the anti-monotonic property of a proposed prevalence measure, called participation index (*PI*), to find all possible co-location patterns. Many of the follow-up work on the co-location mining approach in [3] have focused on improving the runtime.

a. The Participation Index (PI)

The PI of an interaction pattern C is defined as $PI(C) = \min_k\{pr(C, f_k)\}$. For example, let an interaction pattern C = {P,Q,R} where the participation ratios of P, Q, and R are 2/4, 2/7 , and 1/8 respectively.The PI-value of C is 1/8. As an extension of the work in [3], [1] proposes a multi-resolution pruning technique to filter false candidates. To improve the runtime of the algorithm in [1],Yo et al. in [5] and [6] propose two instance look-up schemes where a transaction is materialized in two different ways. In [5], a transaction is materialized from a star neighborhood and in [6], a transaction is materialized from a clique neighborhood. The works by R. Munro, S. Chawla, and P. Sun in [7] and B. Arunasalam, S. Chawla, and P. Sun in [8] look for "complex patterns" that occur due to a mixed type of interaction (a combination of positive and negative), using a proposed prevalence measure called Maximum Participation Index (maxPI). In co-location mining algorithms [1], [3], [5], and [6] a co-location pattern is reported as prevalent, if its PI-value is greater than a user specified threshold. The complex pattern mining algorithm proposed in [8] also reports a pattern as prevalent if its maxPI-value is greater than a user defined threshold. Finding pattern defined in this way, is reasonably efficient since the PI is anti-monotonic and the maxPI is weakly anti-monotonic. However, using such an approach may not be meaningful from an application point of view. Munro et al. in [7] first discuss more complex spatial relationships and spatial patterns that occur due to such a relationship. A combination of positive and negative correlation behavior among a group of features gives rise to a complex type of collocation pattern.

B. Arunasalam et al. in [8] develop a method to mine positive, negative, and complex (mixed). co-location patterns. For mining such patterns, their method uses a user specified threshold on prevalence measure called maximum participation index (maxPI) which was first introduced in [27] to detect a co-location pattern where at least one feature has a high participation ratio. maxPI of a co-location C is the maximal participation ratio of all the features of C. In mining such complex patterns, the total number of candidate patterns of size 1 is doubled for a given number of features. The exponential growth of candidate space with an increased number of features makes the pattern mining computationally expensive. Arunasalam et al. propose a pruning technique which works only when a threshold value of

0.5 or greater is selected. In their method, selection of the right threshold is important for capturing a pattern occurring due to a true correlation behavior. This method lacks a validation of the significance of a pattern when the pattern size is greater than two. Spatial auto-correlation behavior is also not considered in their approach In spatial statistics, the spatial co-location or segregation pattern mining problem is different from the data mining community. There, co-location or segregation pattern mining is similar to the problem of finding associations or interactions in multi-type spatial point processes. There are several measures used to compute spatial interaction such as Ripley's K-function [12], distance based measures (e.g., F function, G function) [11], and co-variogram function [21]. With a large collection of Boolean spatial features, computation of the above measures becomes expensive as the number of candidate subsets increases exponentially in the number of different features. This research develops an algorithm for discovering both types of interaction patterns, which is based on a statistical test and develops a model which takes spatial clustered features into account. It also develops strategies to reduce the computational cost and to reduce the runtime of the algorithm. Evaluate method empirically using real and synthetic data sets and show its advantages over an existing co-location mining algorithm. Mane *et al*. in [22] use bivariate *K*-function [11] as a spatial statistical measure with a data mining tool to find the clusters of female chimpanzees' locations and investigate the dynamics of spatial interaction of a female chimpanzee with other male chimpanzees in the community. There, each female chimpanzee represents a unique mark. Two clustering methods (SPACE- 1 and SPACE-2) are proposed which use Ripley's cross-*K* function to find clusters among different marked point processes. In the data mining community, co-location pattern mining approaches are mainly based on spatial relationship such as "close to" proposed by Han and Koperski in [23]**,**which presents a method to mine spatial association rules indicating a strong relationship among a set of spatial and some non-spatial predicates. A spatial association rule of the form $X \rightarrow Y$ states that in a spatial database if a set of features X is present, another set of features Y is more likely to be present. Rules are built in an apriori-like fashion and a rule is defined as strong if it has enough support and confidence [2]. Morimoto in [4] proposes a method to find groups of different service types originated from nearby locations and report a group if its occurrence frequency is above a given threshold. Xiao *et al*. in [24] improve the runtime of frequent itemset based methods [1], [4], and [6] by starting from the most dense region of objects and then proceeding to less dense regions. From a dense region, the method counts the number of instances of a feature of a candidate co-location. Assuming that all the remaining instances are in co-locations, the method then estimates an upper bound of the *PI*-value and if it is below the threshold the candidate co-location is pruned. In [26] Brin *et al*. use $\chi 2$ as a test statistic to measure the significance of an association of a group of items. Besides finding patterns occurring due to a positive association among spatial features, researchers often look for patterns that can also occur due to the effect of an inhibition or a negative association among spatial features. In association rule mining, several works are found which look for patterns occurring due to the correlation among items from the market basket data. Among those, the work of [26] can be mentioned. There are not many works in the spatial domain can be found that look for patterns occurring due to a negative interaction. All the above mentioned co-location pattern discovery methods use a predefined threshold to report a prevalent co-location. Therefore, if thresholds are not selected properly, meaningless co-location patterns could be reported in the presence of spatial auto-correlation and feature abundance, or meaningful co-location patterns could be missed when the threshold is too high. In [25] introduce a new definition of co-location based on statistical significance test and propose a mining algorithm (SSCP). The SSCP algorithm relies on randomization tests to estimate the distribution of a test statistic under a null hypothesis. To reduce the computational cost of the simulations conducted during the randomization tests, SSCP algorithm adapts two strategies – one in data generation and the other in prevalence measure computation steps.
Sajib Barua and Jörg Sander in [28] overcome above limitations. They propose a mining algorithm that reports groups of features as co-location or as segregation patterns if the participating features have a positive or negative interaction, respectively, among themselves. To determine the type of interaction they do not compare its frequency against a user specified threshold. They also develop appropriate models that take the possible spatial auto-correlation of individual features into account. They introduce pruning strategies to improve the runtime of proposed method.To improve runtime further, they propose a different approaches to identify pattern instances efficiently and compare the effectiveness, and trade-off between runtime and accuracy. They demonstrate the effectiveness of their approach to finding both co-location patterns as well as segregation patterns using a variety of real and synthetic data sets. Following is the block diagram for their proposed work

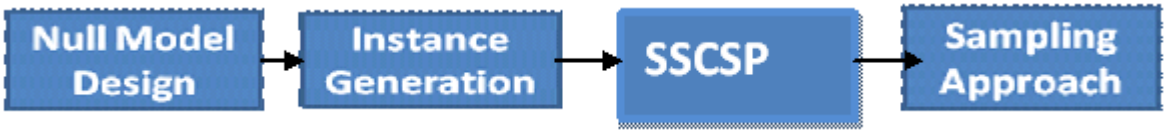

Figure 1 Proposed System Block Diagram

4. CONCLUSION AND FUTURE WORK

In this paper, we discussed different approaches to discover co-location and segregation patterns. Rule based approach showed the similarities and difference between collocation rules problem and classic association rules. Most of the

algorithms adopt an approach similar to the Apriori algorithm proposed for ARM in [2], by introducing some notion of transaction over the space, and a suitable prevalence measure. In event centric model a transaction is generated from a proximity neighborhood of feature instances. Some works look for "complex patterns" that occur due to a mixed type of interaction (a combination of positive and negative), using a proposed prevalence measure. There are several measures used to compute spatial interaction such as Ripley's *K*-function [12], distance based measures [11], and co-variogram function [21]. Some approaches improve the runtime of frequent itemset based methods by starting from the most dense region of objects and then proceeding to less dense regions, the method counts the number of instances of a feature of a candidate co-location. Existing algorithms to finding collocation patterns have several limitations: They depend on user specified thresholds which can lead to missing meaningful patterns or reporting meaningless patterns, they do not take clustered features into consideration, and they may show co-locations for randomly distributed features. Segregation patterns have not receive much attention yet. To overcome such limitations in future we will develop an algorithm for discovering both types of interaction patterns, which is based on a statistical test and develops a model which takes spatial clustered features into account. We will also develop strategies to reduce the computational cost and to reduce the runtime of the algorithm and evaluate method empirically using real and synthetic data sets and show its advantages over an existing co-location mining algorithm.

References

[1] Y. Huang, S. Shekhar, and H. Xiong, "Discovering co-location patterns from spatial data sets: A general approach," IEEE Trans. Knowl. Data Eng., vol. 16, no. 12, pp. 1472–1485, Dec. 2004.
[2] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in Proc. 20th Int. Conf. VLDB, Santiago, Chile, 1994, pp. 487–499.
[3] S. Shekhar and Y. Huang, "Discovering spatial co-location patterns: A summary of results," in Proc. 7th Int. SSTD, Redondo Beach, CA, USA, 2001, pp. 236–256.
[4] Y. Morimoto, "Mining frequent neighboring class sets in spatial databases," in Proc. 7th ACM SIGKDD Int. Conf. KDD, New York, NY, USA, 2001, pp. 353– 358.
[5] J. S. Yoo and S. Shekhar, "A joinless approach for mining spatial collocation patterns," IEEE Trans. Knowl. Data Eng., vol. 18, no. 10, pp. 1323–1337, Oct. 2006.
[6] J. S. Yoo and S. Shekhar, "A partial join approach for mining co-location patterns," in Proc. 12th ACM Int. Workshop GIS, Washington, DC, USA, 2004, pp. 241–249.
[7] R. Munro, S. Chawla, and P. Sun, "Complex spatial relationships," in Proc. 3rd IEEE ICDM, 2003, pp. 227–234.
[8] B. Arunasalam, S. Chawla, and P. Sun, "Striking two birds with one stone: Simultaneous mining of positive and negative spatial patterns," in Proc. 5th SIAM ICDM, 2005, pp. 173–182.
[9] P. J. Diggle, "Displaced amacrine cells in the retina of a rabbit : Analysis of a bivariate spatial point pattern," J. Neurosci. Meth., vol. 18, no. 1–2, pp. 115–25, 1986.
[10] P. J. Diggle, Statistical Analysis of Spatial Point Patterns, 2nd ed. London, U.K.: Arnold, 2003.
[11] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, Statistical Analysis and Modelling of Spatial Point Patterns. Hoboken, NJ, USA: Wiley, 2008.
[12] B. D. Ripley, "The second-order analysis of stationary point processes," J. Appl. Probab., vol. 13, no. 2, pp. 255–266, 1976.
[13] J. Neyman and E. L. Scott, "Statistical approach to problems of cosmology," J. Roy. Statist. Soc. Ser. B, vol. 20, no. 1, pp. 1–43, 1958.
[14] P. J. Diggle and R. J. Gratton, "Monte Carlo methods of inference for implicit statistical models," J. Roy. Statist. Soc. Ser. B, vol. 46, no. 2, pp. 193–227
[15] J. Besag and P. J. Diggle, "Simple Monte Carlo tests for spatial patterns," Appl. Statist., vol. 26, no. 3, pp. 327–333, 1977.
[16] C. J. Geyer, "Likelihood inference for spatial point processes," in Stochastic Geometry: Likelihood and Computation, O. E. Barndorff- Nielsen, W. S. Kendall, and M. V. Lieshout, Eds. Boca Raton, FL, USA: Chapman and Hall / CRC, 1999, no. 80, ch. 3, pp. 79–140.
[17] K. Schladitz and A. Baddeley, "A third order point process characteristic,"Scandinavian J. Statist., vol. 27, no. 4, pp. 657–671, 2000.
[18] R. D. Harkness and V. Isham, "A bivariate spatial point pattern of ants' nests," J. Roy. Statist. Soc. Ser. C (Appl. Statist.), vol. 32, no. 3, pp. 293–303
[19] M. J. Hutchings, "Standing crop and pattern in pure stands of Mercurialis Perennis and Rubus Fruticosus in mixed deciduous woodland," Nordic Society Oikos, vol. 31, no. 3, pp. 351–357, 1979.
[20] G. L. W. Perry, B. P. Miller, and N. J. Enright, "A comparison of methods for the statistical analysis of spatial point patterns in plant ecology," Plant Ecol., vol. 187, no. 1, pp. 59–82, 2006.
[21] N. A. C. Cressie, Statistics for Spatial Data. New York, NY, USA: Wiley, 1993.
[22] S. Mane, C. Murray, S. Shekhar, J. Srivastava, and A. Pusey, "Spatial clustering of chimpanzee locations for

Neighborhood identification," in Proc. 5th IEEE ICDM, Washington, DC, USA, 2005, pp. 737–740.
[23] K. Koperski and J. Han, "Discovery of spatial association rules in geographic information databases," in Proc. 4th Int. SSD, Portland, ME, USA, 1995, pp. 47–66.
[24] X. Xiao, X. Xie, Q. Luo, and W.-Y. Ma, "Density based co-location pattern discovery," in Proc. 16th ACM Int. Symp. Adv. GIS, Irvine, CA, USA, 2008, pp. 250–259.
[25] S. Barua and J. Sander, "SSCP: Mining statistically significant collocation patterns," in Proc. 12th Int. SSTD,Minneapolis,MN, USA,2011, pp. 2–20.
[26] S.Brin, R. Motwani, and C. Silverstein, "Beyond market baskets:Generalizingassociation rules to correlations," in Proc. SIGMOD, New York, NY, USA, 1997, pp. 265–276.
[27] Y. Huang, J. Pei, and H. Xiong, "Mining co-location patterns with rare events from spatial data sets," GeoInformatica, vol. 10, no. 3, pp. 239–260, 2006.
[28] Sajib Barua and Jörg Sander, " Mining Statistically Significant Co-location and Segregation Patterns" ,IEEE Trans. Knowledge and Data Eng., Vol. 26, no. 5, may 2014.

AUTHOR

Priya Shejwal received the B.E. degree in Information Technology Engineering from Brahma Valley Collage of Engg. And Research Institute , Nashik, Savitribai Phule Pune University in 2013. Now pursuing M.E. from K. K. Wagh Institute of Engineering Education & Research, Nashik, India.

Prof. Jyoti Mankar, Assistant Professor, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik, India.