International Conference on "Emerging Trends in Computer Engineering, Science and Information Technology "-2015
Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477

Mining Co-location and Segregation Patterns from Spatial Data

Miss Priya S. Shejwal **Prof. J. R. Mankar**

Abstract — **Interaction between features generates two types of interaction patterns in spatial domains. A positive interaction brings a subset of features close to each other and negative interaction results in subsets of features segregating from each other. Co-location patterns intended to represents positive interactions and Segregation patterns representing negative interactions. Existing approaches to finding co-location patterns have several limitations because they depend on user specified thresholds for prevalence measures, they may report co-locations even if the features are randomly distributed and they do not take spatial auto-correlation into account. Segregation patterns have not receive much attention yet. To overcome these limitations in a proposed approach, a model for the null distribution of features is developed so spatial auto-correlation is taken into account, and an algorithm for finding both co-location and segregation patterns is designed. Pruning strategy is developed to reduce the computational cost compared to existing approach and methods are developed to reduce the runtime of algorithm even further by using Grid based and Stratified sampling approaches. Proposed method is evaluated empirically using synthetic and real data sets and performance of the proposed system is compared with the existing method.**

Key Words — **Co-location; Sampling approach; Segregation; Spatial data; Spatial interaction**.

I. INTRODUCTION

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. Interaction pattern mining can lead to important domain related insights in areas such as ecology, biology, epidemiology, earth science, and transportation. 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. For instance, In urban areas, there are co-location patterns such as between shopping mall and restaurant. Examples of segregation patterns are common in ecology, where they arise from processes such as the competition between plants or the territorial behavior 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.

Existing approaches to finding Co-location patterns have several limitations: (1) They depend on user specified thresholds for prevalence measures which can lead to missing meaningful patterns or reporting meaningless patterns; (2) they do not take instances of a feature which have a tendency to form clusters (i.e. spatial auto correlation) into account; and (3) they may report co-locations even if the features are randomly distributed. Segregation patterns have not receive much attention yet. These limitations of existing system motivated the proposed work.

This paper is organized as follows: In Section I, a brief introduction of co-location and segregation patterns and motivation of the proposed system. Section II describes the related work in which we describe the motivational survey, efficiency and drawbacks of previous system. Section III describes the implementation details with Mathematical model, Process block diagram and Datasets. Section IV describes the Results Discussion. And finally in Section V, we conclude with the summary of this paper.

II. LITERATURE SURVEY

Most of the current algorithms [1], [3] to [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. In existing 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 approach in [3] uses the event centric model where a transaction is generated from a proximity neighborhood of feature instances. Feature instances present in such a neighborhood are neighbors of each other forming a clique. The proposed prevalence measure is called the Participation Index (PI).

The works in [7] and [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).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.

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 [9], new definition of co-location based on statistical significance test is introduced and a mining algorithm (SSCP) is proposed. 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.

In [10], above work is extended and a sampling approach is proposed to improve the runtime of the SSCP algorithm further. Proposed sampling approach uses a grid based technique to generate sample efficiently and can reduce the computational cost of SSCP approach further. Also an algorithm is designed for finding both co-location and
segregation patterns.

III. IMPLEMENTATION DETAILS

The proposed system takes a spatial data set with spatial features as an input and reports groups of features as co-location or as segregation patterns if the participating features have a positive or negative interaction and develops appropriate null models that take the possible spatial auto-correlation of individual features into account. Improve the runtime of the system by introducing pruning strategies and sampling based approaches like Grid based sampling and Stratified sampling.

A. Mathematical Model

The system aims to introduce a new definition of co-location and segregation pattern, we propose a model for the null distribution of features so spatial auto-correlation is taken into account, and we design an algorithm for finding both co-location and segregation patterns.

	Fn1	Fn2	Fn3	Fn4	Fn5
Fn1	0	1	0	0	0
Fn2	0	0	1	0	0
Fn3	0	0	0	1	0
Fn4	0	0	0	0	1
Fn5	0	0	0	0	0

The proposed system S is defined as follows,

S={D, N, IG, SP, DG, PC}

where S is a Proposed system,

$D = \{D_1, D_2, \dots, D_n\}$
D is a given dataset.

$N = \{N_1, N_2, \dots, N_n\}$
N is Null Model Design

$IG = \{IG_1, IG_2, \dots, IG_n\}$
IG is Instance Generation

$SP = \{SP_1, SP_2, \dots, SP_n\}$
SP is Sampling method

$DG = \{DG_1, DG_2, \dots, DG_n\}$
DG is Data Generation

$PC = \{PC_1, PC_2, \dots, PC_n\}$
PC is PI-Value Computation

Y={N, IG, SP, DG,PC}
Y is a set of techniques used for Significant Co-location and Segregation Patterns.

O={O1, O2, ... , On}
Output is given as a set of co-location or segregation patterns

The system design includes following main functions:

1. Null model design (Fn1):
   The function inputs given dataset and generate null model for auto-correlated features
   Fn1(N) → IG

2. Instance generation of auto-correlated features (Fn2):
   The function inputs features from null model and generates instances .
   Fn2(IG) → SP

3. Sampling method (Fn3):
   This function uses sampling method and generates data.
   Fn3(SP) → DG

4. Data generation for computation (Fn4):
   Function Fn4 takes data generated from sampling method for PI-Value Computation.
   Fn4(DG) → PC

5. PI value computation (Fn5):
   Function Fn5 inputs PI-values and outputs patterns as colocation or segregation.
   Fn5(PC) → O

TABLE I: FUNCTIONAL DEPENDENCY

B. Process Block Diagram

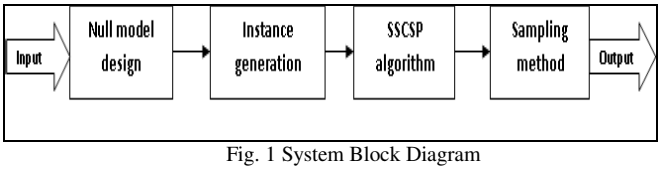


Fig. 1 System Block Diagram

As shown in the diagram below, Proposed system takes a spatial data set with spatial features as an input and reports groups of features as co-location or as segregation patterns if the participating features have a positive or negative interaction, respectively, among themselves. First null model detects auto-correlated features then by using pruning strategies instances of features are generated. Sampling based approaches identifies the instances of a pattern for the significance test at a reduced computational cost. As a result, the speedup is further improved. Proposed methods and approaches consist of following different stages:

1. Null Model Design

Null hypothesis must model the assumption that different features are distributed in the space independently of each other. A spatial feature could be either spatially auto-correlated or not spatially auto-correlated. A feature which is spatially auto-correlated in the given data is modeled as a cluster process. To determine if a feature is spatially auto-correlated or not, compute the PCF value (g(d)). Values of g(d) > 1 suggest clustering or attraction at distance d. A feature has a regular distribution (inhibition) if g(d) < 1, and a feature shows CSR if g(d) = 1. Hence for g(d) ≤ 1, the feature is considered to be not spatially auto-correlated.

2. Instance generation

The features which are auto-correlated their instances are generated by using pruning strategy. This will reduce the runtime by generating a reduced number of instances for an auto-correlated feature in a simulated data generation step and by pruning unnecessary candidate patterns in the PI-value computation step.

3. SSCSP Algorithm

It takes spatial dataset as input then data is generated for the simulation runs after that p - value computation is performed for detecting co-location and segregation patterns accurately as output. Algorithm for SSCSP contains following steps:

First determine the PI-value of each interaction pattern C, in each simulation run of the randomization tests. This requires identifying all instances of C, which naively amounts to checking the neighborhoods of each participating feature in C. In the following, both the data generation step and the p-value computation are described, including strategies for reducing the overall computational cost of this approach.

Data generation for the simulation runs: In a simulation, instances of each feature are generated. For an auto-correlated feature, only generate instances of those clusters which are close enough to different features (auto-correlated or not) to be potentially involved in interactions. Fig. 2 shows an example with two auto-correlated features. Fig. 2(b) shows the partial amount of instances generated to compute the same PI-value that would be computed from all instances as in Fig. 2(a).

p-value computation: First, compute the PI-value, $PI_{obs}(C)$, of each possible interaction pattern C in the observed data. To calculate the p-values p_{pos} and p_{neg}, maintain two counters for the PI-value of C: $R^{\geq PIobs(c)}$ and $R^{\leq PIobs(c)}$. To compute p_{pos} and p_{neg}, do randomization tests, generating a large number of simulated data sets that conform to the null hypothesis. Then compute the *PI*-value of a pattern *C*, $PI_0(C)$, in each simulation run and compute p_{pos} and p_{neg} respectively as:

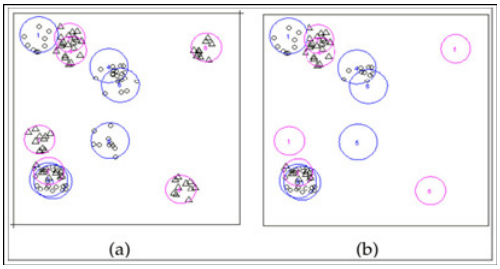


Fig. 2: (a) Instances of all clusters. (b) Generated instances.

$$p_{pos} = \frac{R^{\geq PI_{obs}} + 1}{R + 1} \quad (1)$$

$$p_{neg} = \frac{R^{\leq PI_{obs}} + 1}{R + 1} \quad (2)$$

Here $R^{\geq PIobs}$ of equation 1 represents the number of simulations where the computed $PI_0(C)$ is not smaller than the PI_{obs}- value. $R^{\leq PIobs}$ of equation 2 is the number of simulations where the computed $PI_0(C)$ is not greater than the PI_{obs}-value. R represents the total number of simulations.

In both the numerator and the denominator one is added to account for the observed data.

4. Sampling Methods

If a true co-location or segregation relationship exists among a group of features C, this should be reflected even in a subset of the total instances of C, and a statistical test should be able to capture this dependency from such a subset. Instead of looking at the full neighborhood S_0 of a feature instance I, consider only a sub-region S of S_0. By considering a larger sub-region which covers more area of S_0, the computed PI^*-value will be more similar to the original PI-value.

A neighborhood sampling approach using a grid based space partitioning: To select sub-regions of actual neighborhoods, a grid is placed over the whole study area. Each grid cell is a square with a diagonal length l being equal to Rd/ w , where Rd is the interaction neighborhood radius and w ≥ 1 is an integer. If l = Rd, the selected sub-region represents a sampled

neighborhood for a feature instance I that consists of a single cell X that contains I. If l = Rd /2 , the sampled neighborhood consists of the cell X that contains I, plus the 8 cells surrounding X. In general, if l = Rd/w , the sampled neighborhood of I consists of $(2w-1)^2$ cells including
X. Corresponding neighborhood is denoted by $S_{(2w-1)}^2$. Fig. 2 illustrates the sampled neighborhoods for w equal to 1, 2, and 3, i.e. S_1, S_9, and S_{25}. Note that any other feature instance located in a sampled neighborhood of I is necessarily co-located with I by construction.

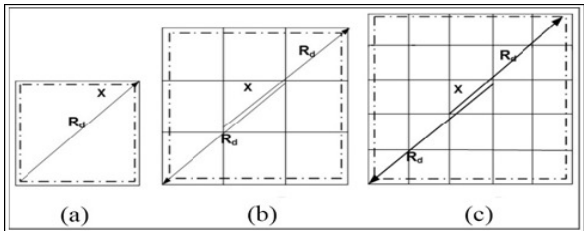


Fig. 3: Dashed bordered region is a sampled neighborhood for a feature instance present anywhere of cell X.(a) l = Rd (b) l = Rd/2 (c) l = Rd/3

Stratified Sampling approach: This is a Latin hypercube method and three dimensional approach. In this approach sampling frame is divided into subsets of known size. It is very flexible method which achieves better precision and recall values than existing methods. It can work on larger pattern sizes. It will give more accurate and fast results than the existing methods.

C. Data Sets

Proposed method is evaluated empirically using synthetic and real data sets. They are as follows:

1. Synthetic Data Sets

- Inhibition

A set of negatively associated features can be wrongly reported as a prevalent collocation pattern by the existing co-location mining algorithms, using typical threshold values.

- Auto-correlation

This show that even though participating features of a pattern are independent of each other, their spatial auto-correlation properties can generate a PI-value higher than a typical threshold.

- Mixed Spatial Interactions

Synthetic data set with 5 different feature types. Among these features, we impose different spatial relationships such as positive association, auto-correlation, inhibition, and randomness.

2. Real Data Sets

- Ants Data

The nesting behavior of two species of ants (Cataglyphis bicolor and Messor wasman) is investigated to check if they have any dependency on biological grounds. The Messor ants live on seeds while the Cataglyphis ants collect dead insects for foods which are for the most part dead Messor ants. Zodarium frenatum, a hunting spider, kills Messor ants. The full data set gives the spatial locations of nests. It comprises 97 nests (68 Messor and 29 Cataglyphis) inside an irregular convex polygon.

- Bramble Canes Data

The blackberry bush is known as Bramble. Bramble canes data records the locations (x,y) and ages of bramble canes in a field of a 9m square plot. The canes were classified according to age as either winter buds breaking the soil surface, unbranched and nonflowering first year stems, or branched and flower bearing second year stems. These three classes are encoded as marks 1, 2, and 3 respectively in the data set. There are 359 canes with mark 1, 385 with mark 2, and 79 with mark 3.

- Lansing Woods Data

This is famous multi-type point data set from a plot of 19.6 acre in Lansing Woods, Clinton County, Michigan, USA. This data set records the location of 2251 trees of 6 different species (135 black oaks, 703 hickories, 514 maples, 105 red
oaks, 346 white oaks, and 448 miscellaneous trees).

RESULTS AND DISCUSSION

The sampling based approach find exactly the same patterns as the all-instances-based approach. Precsion, Recall, and F-measure (harmonic mean of precision and recall), for the standard co-location algorithm, as well as for proposed method are calculated.

The computational time of the existing algorithms depends on the selection of the PI_{thre}-value. A low PI_{thre}-value is computationally more expensive than a high PI_{thre}-value. A low PI_{thre}-value allows fewer pruning and thus results in more candidate patterns as being prevalent. Hence there is no fair way to compare our algorithm with the existing algorithms. Sampling based methods are faster than all-instances-based approach and existing algorithms.

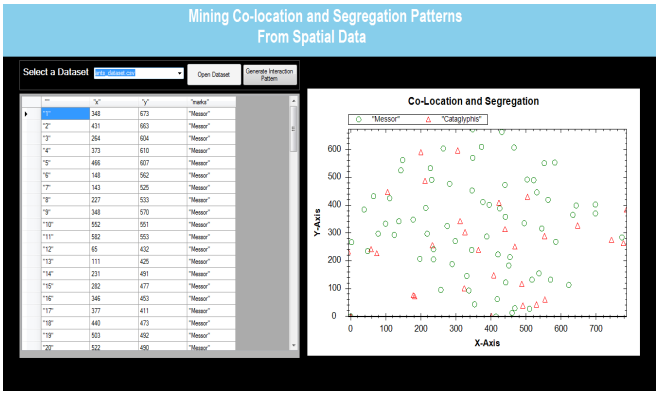
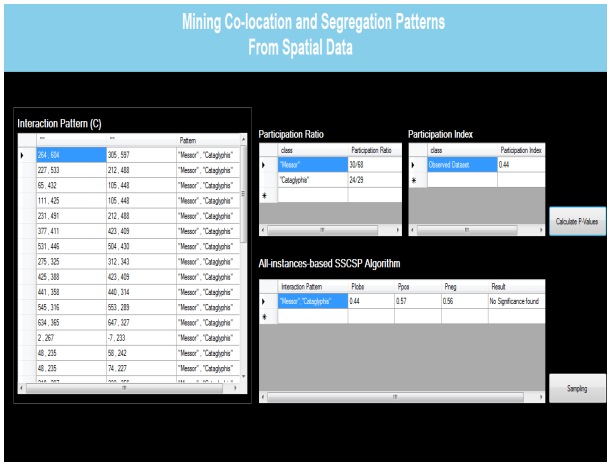


TABLE II: SPATIAL INTERACTION OF ANTS BY USING SSCSP ALGORITHM

Interaction Pattern	PI_{obs}	P_{pos}, P_{neg}	Pattern Reported
{Cataglyphis,Messor}	0.44	0.53,0.52	No

TABLE III: SPATIAL INTERACTION OF ANTS BY USING SAMPLING APPROACH WITH DIFFERENT CELL RESOLUTION

Inte-raction Pattern	PI_{obs} (For l=Rd)	P_{pos}, P_{neg}	PI_{obs} (For l=Rd/ 2)	P_{pos}, P_{neg}	PI_{obs} (For l=Rd /3)	P_{pos}, P_{neg}	Patter n Repor td
{C, M}	0.08	0.38, 0.67	0.16	0.60, 0.41	0.23	0.25, 0.78	No



CONCLUSION

Existing approaches in the literature find prevalent patterns based on a predefined threshold value which can lead to missing meaningful patterns or reporting meaningless patterns. Proposed method uses a statistical test. Such statistical test is computationally expensive and have two approaches to improve the runtime. In first approach, it reduces the runtime by generating a reduced number of instances for an auto-correlated feature in a data generation step and by pruning unnecessary candidate patterns in the PI-value computation step. In the second approach, a PI-value of a pattern computed from a subset of the total instances is, in general, sufficient to test the significance of a pattern. Sampling approaches like Stratified sampling and Grid based sampling are introduced to identify the instances of a pattern for the significance test at a reduced computational cost. As a result, the speedup is further improved compared to the first approach.

Proposed methods are evaluated using synthetic and real data sets. Sampling approach never misses any true patterns when the number of feature instances is not extremely low. Both the all-instance-based and sampling algorithms find all the true patterns from the synthetic data sets. Using real data sets, algorithms do not miss any pattern of size 2. The pattern finding approach proposed in ecology cannot detect patterns of size greater than 2. Proposed methods also find meaningful patterns of larger sizes.

ACKNOWLEDGEMENT

I would like to express my sentiments of gratitude to all who rendered their valuable help for the successful completion of this project work. I am thankful to my guide Prof. J. R. Mankar , for her guidance and encouragement in this work. Her expert suggestions and scholarly feedback had greatly enhanced the effectiveness of this work. I am also thankful to Prof. S. M. Kamalapur and Prof. N. M. Shahane for the interest shown in this project by timely suggestions and helpful guidance.

REFERENCES

[1] 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.

[2] 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.

[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] R. Munro, S. Chawla, and P. Sun, "Complex spatial relationships," in Proc. 3rd IEEE ICDM, 2003, pp. 227–234.

[6] 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.

[7] 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.

[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] J. S. Yoo and S. Shekhar, "A joinless approach for mining spatial colocation patterns," IEEE Trans. Knowl. Data Eng., vol. 18, no. 10, pp. 1323–1337, Oct. 2006.

[10] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, Statistical Analysis and Modelling of Spatial Point Patterns. Hoboken, NJ, USA: Wiley, 2008.

[11] S. Barua and J. Sander, "SSCP: Mining statistically significant Collocation patterns," in Proc. 12th Int. SSTD,Minneapolis,MN, USA, 2011, pp. 2–20.

[12] Sajib Barua and Jorg Sander, " Mining Statistically Significant Co- location and Segregation Patterns" ,IEEE Trans. Knowledge and Data Eng., Vol. 26, no. 5, may 2014.

AUTHOR'S PROFILE

Miss. Priya Shejwal
Received the B.E. degree in Information Technology from Brahma Valley Collage of Engg. & 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.