Modern Management based on Big Data III
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220104

Mining Maximal Fuzzy Colocation Patterns

Meijiao WANG, Yu CHEN, Libo HE and Yunyun WU[1]
Yunnan Police College, Kunming, China

Abstract. Spatial colocation pattern mining is to discover the subsets of spatial objects frequently appearing together in adjacent geographic locations. In the existing research, several algorithms were proposed for excavating maximal prevalent colocation patterns. Furthermore, fuzzy neighborhood relationship(FNR) was employed to evaluate the proximity between spatial instances for improving the accuracy of the mining results. However, the approach for discovering the maximal prevalent colocation patterns based on FNR is not studied yet. This paper defines the maximal fuzzy prevalent colocation pattern (MFPCP). We propose a maximal fuzzy prevalent colocation pattern mining algorithm to generate the MFPCPs instead of all of the prevalent colocation patterns. We conduct experiments on the real datasets to evaluate the performance of the proposed algorithm.

Keywords. Spatial colocation pattern mining, maximal fuzzy prevalent colocation pattern, fuzzy neighborhood relationship, fuzzy instance tree

1. Introduction

Spatial colocation pattern mining which is an important branch of spatial data mining has attracted more and more attention in recent years. A spatial colocation pattern is a subset of spatial objects of which the prevalence index is no less than the prevalence threshold. The instances of its objects are frequently located together in adjacent space. Spatial colocation pattern mining is mainly applied in the following domains: Biology, Earth science, transportation, public health, etc. [1].

According to the downward closure property of prevalent colocation patterns, a prevalent colocation pattern is maximal when any of its subsets is prevalent while all of its supersets are not prevalent[2-4]. This means that all of the prevalent colocation patterns can be deduced from the maximal prevalent colocation patterns. Since the number of maximal prevalent colocations is much smaller than that of all prevalent colocations, maximal co-locations is more convenient for people to use.

Tobler's First Law demonstrates that the contributions of instances to their pattern's effect decrease along with the distance diminishes. To take into account the proximity level between instances in mining maximal prevalent colocations, Yao etc. proposed the SGCT-K algorithm[5]. SGCT-K employed a kernel density estimation (KDE) model for evaluating the proximity level between instances, and defined a KDE-based prevalence index (PI-K) as the prevalence measure of a colocation. Our previous research [6] defined the fuzzy neighborhood relationship(FNR) to evaluate the proximity level between instances, and proposed the CPFNR algorithm for mining

[1] Corresponding Author: Yunyun WU, Yunnan police college, 249 Jiaochang North Road, Wuhua District, Kunming City, China; E-mail:156799251@qq.com.

colocation patterns based on FNR for improving the accuracy of mining results. The PI-K in SGCT-K is so small that it is hard to set a prevalence index threshold to filter the prevalent colocations.

It can be seen from the above that it's very meaningful to mine maximal fuzzy prevalent colocation based on FNR. The major contributions are as follows:

(1) Based on the FNR and the downward closure of prevalent colocation patterns, we define the Maximal Fuzzy Prevalent colocation Pattern (MFPCP).

(2) Put forward a maximal Fuzzy Prevalent colocation Mining(MFPCM) algorithm for obtaining the MFPCPs.

(3) The efficiency of the MFPCM algorithm are evaluated by experiments.

The related works is stated in Section 2. Section 3 describes the relative definitions. Section 4 presents the algorithm. Section 5 performs the experiments to evaluate the presented algorithm. Finally, a summary is given in Section 6.

2. Related Work

Shekhar et al. first defined the concept of spatial colocation pattern[7]. They employed the participation index to measure the prevalent level of a colocation. Huang et al. present the join-based strategy for mining colocation patterns[1]. It was an Apriori-like algorithm which generated the prevalent colocations from short to long size. Because the table instance connection process consumed a lot of time, the papers [8,9] proposed the join-less algorithm and the partial join algorithm respectively. For efficiently pruning the candidates and reducing the memory usage for storing table instances, the CPI-tree algorithm[10] and the iCPI-tree algorithm[11] constructed the prefix-tree structure for reserving the table instances. Wang et al. studied the SPI-closed colocation discovery approach[12,13]. For massive spatial data, the work in [14,15] studied parallel colocation mining algorithms on map-reduce platform. The fuzzy set theory was adopted in the colocation discovery[6,16-19]. Especially, FNR was used to improve the accuracy of the prevalence index calculations in [6]. To reduce the number of prevalent colocations, mining maximal colocation patterns was disposed in [2-4]. But as far as we know, no work had been conducted on mining the maximal fuzzy prevalent colocation patterns based on the FNR, which will be addressed in this paper.

3. Related Definitions

Table 1 lists the abbreviations of the important concepts in this paper.

Table 1. the abbreviations of the important concepts.

Notations	Meaning	Notations	Meaning
O	spatial objects set	k	size of a colocation pattern
S	spatial instance set	FNR	fuzzy neighborhood relationship
s_u^i	an instance of o_u	FNR_α	α-cut set of FNR
c	a co-location pattern	FPR	fuzzy participation ratio
μ	membership function of FNR	FPI	fuzzy participation index
α	membership threshold of FNR	MFPCP	maximal fuzzy prevalent colocation pattern
d	the Euclidean distance	*min_fprev*	minimum fuzzy participation index threshold

Spatial objects (spatial features or attributes) represent different kinds of things in space. Let $O = \{o_1, o_2, ..., o_N\}$ be the spatial object set of N objects. Spatial instances are the appearance of spatial objects in different geographical locations. The spatial instance data sets is denoted as S, $S = \{s_1, s_2, ..., s_n\}$ is the set of n instances. For the objects o_u ($1 \le u \le N$), an instance of o_u is denoted as $s_u^i (1 \le i \le |S_u|)$.

3.1 Colocation Pattern mining Based on FNR

Definition 1(fuzzy neighborhood relationship(FNR)) . Let $D(D \to [0,\infty))$ be the Euclidean distance set between instances in S. The FNR of S is a fuzzy subset on D, which is formalized by the following mapping:

$$\text{FNR}: D \to [0,1], d \to \mu(d)$$

where, μ is the FNR's membership function, d($d \in D$) denotes the Euclidean distance between instances in S, $\mu(d)$ represents the membership value of d and it is exactly the probability of d pertinent to FNR.

Let $dist(s_i, s_j)$ be the Euclidean distance between two instances s_i and s_j. Then FNR can be expressed as:

$$\text{FNR} = \{<(s_i, s_j), \mu(dist(s_i, s_j))> | s_i, s_j \in S\} \quad (1)$$

The ***α*-cut set** of FNR is denoted as FNR_a which is defined as :

$$FNR_\alpha = \{<(s_u, s_v), \mu(dist(s_u, s_v))> | \mu(dist(s_u, s_v)) \ge \alpha, (s_u, s_v \in S)\} \quad (2)$$

where $\alpha \in [0,1]$ is a pre-defined **membership threshold**, and s_u and s_v are regarded as a **FNR_a neighbor pair** that will be connected by a solid line in the diagram of the datasets.

Example 1. Let α = 0.2. An example data sets is illustrated in Figure 1. It is convenient to obtain that $\mu(dist(\text{B.1,C.2}) = 0.75$ and $<(\text{B.1,C.2}), 0.75> \in FNR_\alpha$.

Let c be a **colocation pattern,** $c \in O$. Let I be a subset of S, $I \subseteq S$. I is called *a* **fuzzy row instance** of c, if I meets all of the following requirements:

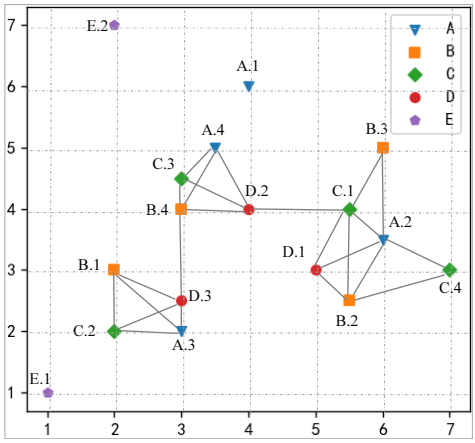
(1) The size of I is equal to that of c.

(2) The object type of each instance in I is the same as that of c in the corresponding order.

(3) The instances in I form a clique under the FNR_a.

A fuzzy row instance of c is denoted as FR(c). All row instances of c compose the table instances of c denoted as FT(c).

Example 2. In Figure 1, let c = {A,B,C,D}, FT(c) = {{A.3,B.1,C2,D.3}, {A.2, B.2, C.1, D.1}, {A.4, B.4,C.3,D.2}}.



(a) distribution of data

$$\mu(d)=\begin{cases}1 & d<0.5\\ -\frac{1}{2}(d-0.5)+1 & 0.5\le d\le 2.5\\ 0 & d>2.5\end{cases}$$

(b) membership function

< (B.1,C.2), 0.75 >
< (B.1,C.3), 0.35 >
< (B.2,C.1), 0.5 >
< (B.2,C.4), 0.46 >
< (B.3,C.1), 0.69 >
< (B.4,C.3), 1 >

(c) FNR_a(B,C)

Figure 1. An example data sets.

Definition 2(the contribution of an instance). Given a fuzzy row instance FR(c), the instance $s_i \in$FR(c), the contribution of s_i is defined as the minimum membership value of all of the membership values between s_i and its fuzzy neighbors in FR(c), i.e.,

$$contri(FR(c),s_i)=min_{j=1}^{m}(\mu(dist(s_i,s_j))), i\neq j \tag{3}$$

Definition 2(fuzzy participation ratio(FPR), fuzzy participation index(FPI)). The FPR of $o_u (o_u \in c)$ is defined as the ratio of the sum of the contributions of non-repeating instances of o_u in FT(c) to the total number of o_u's instances, i.e.,

$$FPR(c,o_u)=\frac{\sum_{s_u^i\in FR(c),FR(c)\in FT(c)} \mathrm{Max}(Contri(FR(c),s_u^i))}{|o_u|} \tag{4}$$

where, $\mathrm{Max}(Contri(FR(c),s_u^i))$ refers to that the maximal contribution is added to the sum when $s^i{}_u$ is repeated in $FT(c)$.

The minimal FPR of the FPRs of all objects in c is regarded as the FPI of c :

$$FPI(c)=\min_{u=1}^{k}\{FPR(c,o_u)\} \tag{5}$$

Given a pre-defined FPI threshold *min_fprev*, if $FPI(c) \geq min\_fprev$ then c is fuzzy prevalent .

Example 3. In Figure 1, $FPR(\{B,C\},B)$ = (0.75+0.5+0.69+1)/4 = 0.735, $FPR(\{B,C\},C)$ = (0.75+1+0.69+0.46)/4 = 0.725, then $FPI(\{B,C\})$ = min(0.735,0.725) = 0.725. If *min_fprev* = 0.3, then {B,C} is prevalent.

3.2 Properties and Related Definitions

Lemma 1 (Monotonicity of FPR and FPI). Let c' and c be two colocation patterns, $c' \subseteq c$. For each object $o \in c'$, $FPR(c', o) \geq FPR(c, o)$. In addition, $FPI(c') \geq FPI(c)$.

Definition 3 (Maximal Fuzzy Prevalent Colocation Pattern(MFPCP)). For a colocation pattern *c,* if any subset of *c* is fuzzy prevalent while any superset of *c* is not fuzzy prevalent, *c* is called a Maximal Fuzzy Prevalent Colocation Pattern(MFPCP).

Example 4. In Figure 1, $FPI(\{A,B,C,D\}) = 0.434$, {A, B, C, D} is a MFPCP. Because its subsets {A, B, C}, {A, B, D},{A, C, D} and {B, C, D} are all fuzzy prevalent, while its superset {A, B, C, D, E} is not fuzzy prevalent.

4. Algorithm

In this section, the algorithm for Maximal Fuzzy Colocation Pattern Mining(MFCPM) is designed by improving the SGCT algorithm given in [4]. It is described as follows:

Algorithm 1. the MFCPM algorithm

Input:
O, S, μ, α, *min_fprev*

Variables:
k: size of a colocation pattern
CP: candidate maximal colocation set
CP_k: size-k candidate set
TI_k: table instance of a size-k candidate
MP_k: size-k maximal fuzzy prevalent set
MP: maximal fuzzy prevalent colocation set
ITree: fuzzy instance tree of a candidate

Output:
MP with $fpi \geq min\_fprev$

Steps:
(1) FNR = get_FNR (S, μ);
(2) CP_2 = gen_candidate_colocations(O);
(3) TI_2 = get_table_instances(CP_2, FNR_α);
(4) P_2 = select_prevalent_colocations(CP_2, TI_2, *min_fprev*);
(5) CP = gen_maximal_candidates(P_2)
(6) k = longest_size (CP)
(7) while(not empty CP) do
(8) CP_k = gen_size-k_candidate_colocations(CP,k);
(9) for each c in CP_k
(10) *ITree* = constuct_instance_tree(c)
(11) TI_k = get_table_instances(c, *ITree*)
(12) *fpi* = calculate_fpi(c, *min_fprev*, TI_k)
(13) if *fpi* >= *min_fprev*
(14) $MP_k = MP_k \cup c$
(15) else
(16) $CP = CP \cup subset(c)$
(17) end for
(18) $MP = MP \cup P_k$
(19) $k = k - 1$;
(20) end do

The main steps of the MFCPM algorithm are as follows:

Step 1 (FNR_α-table construction): Based on the membership function μ and membership threshold α of FNR, the grid division technique is adopted to calculate the FNR_a of the spatial data set. We build the FNR_α-table which is a two-dimensional hash table for storing the FNR_a of the spatial data set. Two object types are used for indexing each cell in the FNR_α-table. They form a size-2 candidate maximal colocation pattern.

Example 4. Figure 2 illustrates the FNR_a-table of the data sets demonstrated in Figure 1. The cell FNR_a(B,C) in the FNR_a-table represents the FNR_a of the candidate size-2 colocation {B,C}.

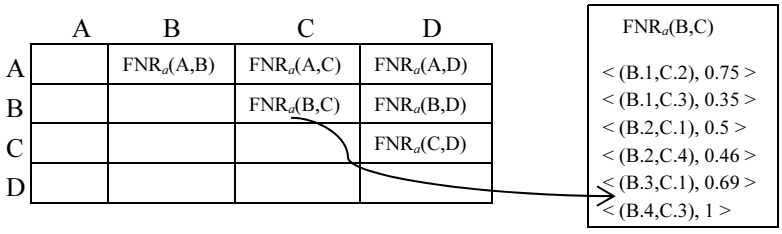


Figure 2. the FNR_a-table of the example data set.

Step 2 (size-2 fuzzy prevalent colocations generation): Generate size-2 candidate colocations from spatial objects and then obtain the table instance for each candidate from FNR_a-table. Filter the prevalent colocations whose FPI is not less than *min_fprev*;

Step 3 (candidate MFPCPs generation): The two objects in a size-2 prevalent colocations are connected by a solid line. Once this done, an undirected graph is constructed, of which each vertex is a spatial object and the two vertices connected by a straight line is just a size-2 prevalent colocation. We obtain all maximal cliques from the undirected graph based on the Bron-Kerbosch algrithm, and regard them as the candidate maximal colocations;

Step 4 (filtering prevalent maximal fuzzy colocations): Filter the final maximal fuzzy colocations from long to short by the size of candidates. The filtering process for each candidate is as follows: first, construct its fuzzy instance tree based on the FNR_a-table and the clique verification approach; second, obtain its fuzzy table instance from the fuzzy instance tree and calculate its fuzzy participation index; third, if its fuzzy participation index is no less than *min_fprev*, it will be reserved as a maximal fuzzy colocation; otherwise, it will be supplanted by its subsets.

5. Experiments

This section conducts experiments of the MFCPM algorithm on real datasets, which concludes 31 species plants with 336 instances in the Three Parallel Rivers of Yunnan Protected Area. The 31 features were denoted by A to Z and a to e in our experiments. Besides MFCPM, the SGCT-K algorithm is the only one that mine maximal prevalent colocations with the proximity level between instances consideration [5]. Both of the two algorithms adopt the SGCT frame given in [4]. The difference of them is that the prevalence measure of the former is FPI while of the latter is the KDE-based prevalence index (PI-K). As the PI-K in SGCT-K is much less than the FPI in MFCPM, we cannot set a definite prevalence index threshold for the two algorithms in the experiments. The algorithms are coded and compiled by Python in the experiments and execute the programs on the Windows 7 operating system with 8 GB memory, 3.4 GHz main frequency and a Intel core i7-6700 processor.

We normalize the real datasets into a 2000 × 2000 space, and define the membership function of the MFCPM algorithm as following:

$$\mu(d) = \begin{cases} 1 & d \le a \\ -\dfrac{(d-a)^2}{(b-a)^2} + 1 & a < d \le b \\ 0 & d > b \end{cases} \tag{6}$$

where, a and b are the arguments to the membership function, d represents the Euclidean distance between instances. The parameter b in MFCPM and the distance threshold in SGCT-K have similar meanings.

We set $a = 20$, $b = 230$, $\alpha = 0$, min_*fprev* = 0.3 in MFCPM, and distance threshold is 230, prevalence threshold is 0.025 in SGCT-K. Table 2 lists partial mining results of the two algorithms. It can be seen that is the PI_K is much less than FPI, which is very different from the classic co-location mining with a prevalence index interval [0,1]. Figure 3 show the running time of MFCPM when min_*fprev* take 0.2 and 0.3 respectively. We can observe that the program executes very fast when min_*fprev* is

0.3, because it can only produce no more than 150 maximal prevalent co-locations with the size no less than 3 and no more than 5, while it can generate no more than 560 maximal prevalent co-locations with the size no less than 3 and no more than 7 when min_*fprev* is 0.2. The larger the number of results, the higher the size, the more time the algorithm consumes.

Table 2. partial results of the two algorithms.

Size	Maximal prevalent co-locations	FPI (MFCPM)	PI_K (SGCT-K)
Size-5	AJLZc	0.3467	0.0327
	AHJLZ	0.304	0.0267
Size-4	AKLc	0.3759	0.0401
	ALbc	0.3073	0.028
Size-3	ABX	0.3607	0.0257
	BSX	0.3795	0.0624

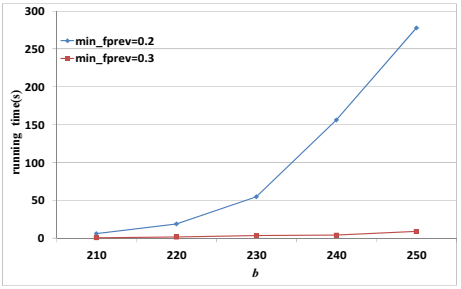


Figure 3. the efficiency of MFCPM

6. Conclusions

We proposed the MFCPM algorithm for mining maximal fuzzy prevalent colocation patterns. MFCPM produces the maximal prevalent colocations from long to short size candidates based on the fuzzy neighborhood relationship(FNR). The candidate maximal colocations are obtained from the undirected graph consisting of the size-2 maximal prevalent colocations, and the table instances of a candidate colocation is generated by its instance tree. Experiments show that the MFCPM algorithm performs good performance.

Acknowledgments

This work is supported by the major project of Science and Technology Department of Yunnan Province(No.2019BC003), the project of Science and Technology Department of Yunnan Province(No.202101AU070158), the projects of Yunnan Education Department (No.2021J0695, No.2020J0484) and the projects of Yunnan Police College (No.19A021, No.19A010).

References

[1] Y. Huang, S. Shekhar, H. Xiong, Discovering colocation patterns from spatial data sets: a general approach, IEEE Educational Activities Department , 2004 , 16 (12) :1472-1485.
[2] L. Wang, L. Zhou, J. Lu, et al. An order-clique-based approach for mining maximal co-locations[J]. Information Sciences, 2009, 179(19):3370-3382.
[3] J. Yoo, M. Bow. Mining Maximal Co-located Event Sets[C]. Pacific-asia Conference on Advances in Knowledge Discovery & Data Mining. Springer-Verlag, 2011:351-362.
[4] X. Yao, L. Peng, L. Yang, et al. A fast space-saving algorithm for maximal co-location pattern mining[J]. Expert Systems with Applications, 2016, 63: 310-323.
[5] X. Yao, L. Peng, L. Yang, et al. A co-location pattern-mining algorithm with a density-weighted distance thresholding consideration[J]. Information Sciences, 2017, 396:144-161.

[6] M. Wang, L. Wang, L. Zhao. Spatial Co-location Pattern Mining Based on Fuzzy neighbor relationship. Journal of Information Science & Engineering, 2019, 35(6).
[7] S. Shekhar,Y. Huang. Discovering Spatial Co-location Patterns: A Summary of Results[C]. Proc. The 7th International Symposium on Advances in Spatial and Temporal Databases(SSTD), Heidelberg, Springer, 2001: 236-256.
[8] J. Yoo, S. Shekhar, M. Celik. A join-less approach for colocation pattern mining: a summary of results, IEEE International Conference on Data Mining. IEEE, 2005:813-816.
[9] J. Yoo, S. Shekhar, J. Smith, et al. A partial join approach for mining co-location patterns[C]. Proc. the 12th annual ACP international workshop on Geographic information systems, New York, Washington DC, USA, ACP Press, 2004: 241-249.
[10] L. Wang, Y. Bao, J. Lu, J. Yip, A new join-less approach for co-location pattern mining, IEEE International Conference on Computer and Information Technology. IEEE, 2008:197-202.
[11] L. Wang, Y. Bao, Z. Lu, Efficient discovery of spatial co-location patterns using the iCPI-tree, The Open Information Systems Journal, 2009, 3(1):69-80.
[12] L. Wang, X. Bao, L. Zhou, Redundancy reduction for prevalent co-location patterns, IEEE Transactions on Knowledge & Data Engineering, 2018, 30(1):142-155.
[13] L. Wang, X. Bao, H. Chen, Effective lossless condensed representation and discovery of spatial co-location patterns, Information Sciences, 2018,436: 197-213.
[14] J. Yoo, Boulware, D. and Kimmey, A Parallel Spatial Co-location Mining Algorithm Based on MapReduce. IEEE International Congress on Big Data, 2014:25-31.
[15] P. Yang, L. Wang , X. Wang, A Parallel Spatial Colocation Pattern Mining Approach Based on Ordered Clique Growth, International Conference on Database Systems for Advanced Applications, 2018:734-742.
[16] Z. Ouyang, L. Wang, P. Wu, Spatial co-location pattern discovery from fuzzy objects, International Journal on Artificial Intelligence Tools, International Journal on Artificial Intelligence Tools, 2017, 26(02): 1750003.
[17] M. Wang, Y. Chen, L.He, Y. Wu. Spatial Colocation Pattern Mining based on improved density peak clustering and fuzzy neighbor relationship. Mathematical Biosciences and Engineering, 2021, 18(6): 8223-8244.
[18] M. Wang, L. Wang, L. Zhou. Spatial Colocation Pattern Mining with the Maximum Membership Threshold[M]//Fuzzy Systems and Data Mining V. IOS Press, 2019: 1092-1100.
[19] M. Wang, L. Wang, Y. Qian, et al. Incremental mining of spatial co-location Patterns based on the fuzzy neighborhood relationship[M]//Fuzzy Systems and Data Mining V. IOS Press, 2019: 652-660.