Fuzzy Systems and Data Mining VII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210203

A Novel Method for Mining Fuzzy Co-Location Patterns

Jinyu GUO and Lizhen WANG[1]

School of Information Science and Engineering, Yunnan University, Yunnan, China

Abstract. As an important branch of the spatial data mining, spatial co-location pattern mining refers to discovering a subset of the set of features whose instances are often neighbor in space. In many practical scenes, the instances of spatial features include not only location information, but also attribute information. Some scholars use the type-1 fuzzy membership function to mine fuzzy co-location patterns from spatial instances with attribute information. However, the type-1 membership function itself is uncertain. Therefore, there is some deviation in describing the membership degree of attributes of spatial instances by using a type-1 membership function. To solve this problem, we propose a fuzzy co-location pattern mining method based on type-2 fuzzy membership function. Firstly, we collected interval evaluation values of interval data of attribute information from 1000 experts, and formed granular data. Then, on the basis of the original type-1 membership function, a type-2 fuzzy membership function based on elliptic curve is expanded, and the parameters of the type-2 fuzzy membership function are adjusted by using a gradual method, so that the footprint of uncertainty (FOU) in the function meets the connectivity and the threshold given by the user. After that, we design a fuzzy co-location pattern mining algorithm incorporating type-2 fuzzy membership function into the traditional Join-based algorithm. In which, we define the concepts of fuzzy feature, fuzzy co-location pattern, upper bound participation index, lower bound participation index. In order to improve the efficiency of our method, we also put forward a pruning strategy. We have done a lot of experiments on synthetic and real data sets, which proves the effectiveness and efficiency of our proposed algorithm.

Key words: Spatial data mining; Fuzzy co-location pattern; Type-2 fuzzy set; Pruning

1. Introduction

Spatial co-location pattern mining is a very important branch in the field of spatial data mining. A spatial co-location pattern is a set of spatial features, and the instances of these features are often neighbor in space. For example, {convenience store, pharmacy} may be a prevalent co-location pattern, because convenience stores and pharmacies are often neighbor to each other in a city central. Spatial co-location pattern mining has a wide range of applications, such as geographic information science, symbiotic plant distribution, urban facility distribution, and so on.

We find that spatial data often contains attribute information besides spatial location information, such as the content of heavy metals in topsoil at a certain spatial location, which is usually expressed by specific numerical values. However, people are usually

[1] Corresponding author: Lizhen Wang, School of Information Science and Engineering, Yunnan University, Dongwaihuan SouthRoad, Kunming, Yunnan 650500, China. E-mail: lzhwang@ynu.edu.cn

less sensitive to a specific value of attribute information, but when the attribute value changes fundamentally, that is, when it changes from a certain range to a new range, people are sensitive and attach importance to the result. In this paper, a method of mining fuzzy spatial co-location pattern based on type-2 fuzzy membership function and Join-based algorithm is proposed, which is used to mine fuzzy co-location pattern from spatial data sets with attribute information.

In recent years, spatial co-location pattern mining methods can be roughly divided into two categories, one is Apriori-like method [1,2], which first generates candidate co-location patterns, and then generate row-instances of candidate patterns and check their prevalence. The second method is to generate maximal cliques from spatial data to find prevalent co-location patterns [3,4]. However, these two kinds of mining methods almost only emphasize the spatial location information of spatial data, while ignoring their attribute information. In the traditional methods of mining co-location patterns, we can only find the spatial association of different functional areas, but not the spatial association of functional areas with attribute information. For example, whether there is a correlation between the content of heavy metals in industrial area and the content of heavy metals in residential area, etc. Therefore, we mine fuzzy co-location pattern based on the type-2 fuzzy membership function, which not only reduces the deviation caused by the uncertainty of the type-1 fuzzy membership function, but also accurately mines fuzzy co-location pattern.

In this paper, the heavy metal content in a surface soil is taken as an example to illustrate the application of the proposed method. Through questionnaire survey, 1000 geological researchers were counted to evaluate the membership of two heavy metals (Cu and Zn) in topsoil, that is, the evaluation data of particle type [5]. Aiming at mining more reasonable fuzzy co-location patterns from spatial data sets with attribute information, this paper makes the following contributions to the field of spatial co-location pattern mining:

(1). In the fuzzy processing of corresponding attribute information, on the basis of the original type-1 membership function, a type-2 membership function based on elliptic curve is expanded from the evaluation data of particle type. Among them, we use a gradual method to adjust the parameters of the function, so that the footprint of uncertainty (FOU) in the type-2 membership function not only satisfies the connectivity [6], but also makes its confidence reach the given threshold (the default is 80%).

(2). According to the constructed type-2 membership function, this paper defines the concepts of upper bound participation ratio and lower bound participation ratio of fuzzy features, and upper bound participation index and lower bound participation index of fuzzy co-location patterns. Furthermore, because our fuzzy co-location patterns haves both upper and lower membership degrees, according to the given prevalence threshold, we divide the fuzzy co-location pattern into three kinds of patterns: absolute non-prevalent fuzzy co-location patterns, fuzzy co-location patterns with prevalent tendency degree and absolute prevalent fuzzy co-location patterns.

(3). We propose a method based on type-2 fuzzy membership function and traditional Join-based algorithm of mining prevalent co-locations, which is used to mine spatial fuzzy co-location patterns. The fuzzy co-location pattern composed of fuzzy features between different fuzzy sets of the same attribute in the same functional area has no practical significance. Therefore, we propose a bucket-based candidate fuzzy co-location generation method, which puts fuzzy features of the same attribute in the same bucket, and then selects fuzzy features from different buckets to form candidate fuzzy co-location patterns. Then, prevalent fuzzy co-location patterns are generated by filtering. In order to improve the efficiency of our algorithm, we propose a pruning strategy.

2. Related work

In recent years, there are some researches that apply fuzzy sets to attribute data mining [7,8]. We introduce the related work of fuzzy co-location pattern mining method based on type-2 membership function from two parts: constructing type-2 fuzzy membership function and discovering spatial co-location patterns.

2.1. Constructing type-2 fuzzy membership function

Type-2 membership functions are usually characterized by the shape of the footprint of uncertainty (FOU). Therefore, many type-2 fuzzy membership functions have been proposed, for example, ladder type, triangle type, Gaussian type [9,10], π type [11], etc. We find that these functions have a common shortcoming: there are many parameters that determine the confidence and uncertainty width, which makes it very difficult to select the appropriate parameter values. However, the parameters of the type-2 membership function based on elliptic curve are relatively simple [12] (there is only one parameter). Therefore, we symmetrically expand from the original type-1 membership function [13] to generate type-2 membership function based on elliptic curve.

2.2. Spatial co-location pattern mining algorithm

Spatial co-location pattern mining algorithms are mainly divided into two categories: one is Apriori-like algorithm, for example, Join-based algorithm [2], Join-less algorithm [1], etc., and another is prefix tree-based algorithm, for example, CPI-Tree algorithm [14], ordered clique-based algorithm [15], etc. As the fuzzy co-location pattern mining based on type-2 membership function is still in the preliminary exploration stage, we improve the Join-based algorithm and propose a fuzzy co-location pattern mining method based on type-2 fuzzy membership function and Join-based.

3. Related definitions and lemmas

We take the contents of heavy metals copper and zinc in the topsoil of an area as an example. We divided the area into agricultural area (The number of the functional area is A), residential area(The number of the functional area is B), industrial area(The number of the functional area is C), and then sampled the topsoil of each functional areas, and measured the heavy metal content at each sampling point. The attribute information of sampling points is shown in Table 1.

Table 1. Attribute information of sampling points

The number of the functional area	**The number of sampling points**	**The content of copper**	**The content of zinc**
A	1	62	226
A	2	31	105
A	3	46	136

B	4	86	183
B	5	29	112
C	6	63	210
C	7	51	155

Traditional co-location pattern mining algorithm can mine patterns like {A, B, C}, which means that the instances of functional areas A、B and C are often neighbor. This pattern can reflect the spatial relevance of instances, but ignores the influence of attribute information in instances on mining results. For example, the relationship of heavy metal content among functional areas (for example, what is the relationship between the copper content in functional area B and the zinc content in functional area C). Obviously, the traditional co-location pattern mining method cannot directly solve this kind of problem. Therefore, we give the following definitions.

Definition 1 (Data fuzzification) The most important step in fuzzifying attribute data is to determine the membership function. Assuming that the fuzzy set on the domain is F, the membership function is expressed as U_F, we use the membership function U_F to describe the fuzzy set F, that is $U_F: x \rightarrow [0,1]$.

Suppose there is any $x \in X$, when $U_F(x)=0$, x does not belong to fuzzy set F at all; when $U_F(x)=0.4$, 40% of the tendency of x belongs to fuzzy set F; when $U_F(x)=1$, x completely belongs to fuzzy set F.

In previous studies, there were the following problems in obtaining type-1 membership function by expert experience: Different experts' evaluation of membership degree of specific value is likely not completely consistent; Experts may give specific membership degrees to specific values, which may cause some deviations (for example, The experts gave the evaluation of low Cu content with membership degree of 0.3 for the Cu content of 25ug/g, which may have some deviations). Therefore, we use the upper and lower bounds of the footprint of uncertainty (FOU) to deal with the problem in this paper, that is, we introduce two membership functions (one is called the upper bound membership function $\overline{U}_F(x)$, which is used to represent the upper bound of the membership degree of an instance's attribute; the another is called lower bound membership function $\underline{U}_F(x)$, which is used to represent the lower bound of the membership degree of an instance's attribute).

Definition 2 (Granular evaluation data) We divide the attribute of heavy metal content into three fuzzy sets: low、middle and high. The divided attributes are called fuzzy attributes, such as Zn(L)、Zn(M) and Zn(H). Then, in the form of questionnaire, we invited 1000 geologists to evaluate the interval membership of the interval content of heavy metals copper and zinc in surface soil, to form granular evaluation data. For example, for the interval value of Cu content [15, 20], 21 people think that the degree of middle membership is [0, 0.1], and 573 people think that the degree of middle membership is [0.1, 0.2], 391 people thought that the degree of low membership was [0.2, 0.3], and 15 people thought that the degree of low membership was [0.3, 0.4].

Definition 3 (Fuzzy feature) Fuzzy features refer to different type of things with fuzzy attributes. A set of fuzzy feature is represented as $\boldsymbol{Fuz\_Fea}$ $=\{Fuz\_f_1,Fuz\_f_2...,Fuz\_f_n\}$, a fuzzy feature is represented as $Fuz\_f_i(1\leq i\leq n)$.

For example, the fuzzy feature B.Zn(L) represents the functional area B with low zinc content. The instance of fuzzy feature refers to the sampling point with the fuzzy feature property in the space.

Definition 4 (Neighborhood Relationship) For any two instances s_1 and s_2 in space, if the Euclidean distance between the two instances is less than the given threshold *min_dist*, that is the distance(s_1, s_2)≤ *min_dist*, it indicates that the two instances are neighbor each other, denoted by $R(s_1, s_2)$.

Definition 5 (Fuzzy co-location pattern) The fuzzy co-location pattern ***Fuz_c*** is a set of fuzzy features, and the size of the fuzzy co-location pattern is the number of fuzzy features contained in the fuzzy pattern, denoted by |***Fuz_c***|. For example, The size of fuzzy co-location pattern {A.Zn(M), B.Zn(M), C.Cu(M), E.Cu(H)} is 4.

Definition 6 (Row-instance and Table-instance) Suppose there is a set of spatial instances $\boldsymbol{S}=\{s_1,s_2,\ldots,s_n\}$, where any two instances are neighbor, if ***S*** contains all fuzzy features of a fuzzy co-location pattern, and any subset of ***S*** cannot contain all the fuzzy features of ***Fuz_c***, then ***S*** is called a row-instance of ***Fuz_c***. All row-instances of ***Fuz_c*** constitute the table-instance of ***Fuz_c***.

Definition 7 (upper bound participation ratio and lower bound participation ratio) For a fuzzy pattern ***Fuz_c*** ={*Fuz_f*$_1$, *Fuz_f*$_2$,…, *Fuz_f*$_n$}, The lower bound participation ratio of the fuzzy feature *Fuz_f*$_i$(1≤i≤n) in this fuzzy pattern is expressed as $\underline{PR}$(***Fuz_c***, *Fuz_f*$_i$), defined as ratio of the sum of the lower bound membership degrees of the non-repeated instances of fuzzy feature *Fuz_f*$_i$ in the table-instance of fuzzy co-location pattern ***Fuz_c*** to the sum of the lower bound membership degrees of all the instances of *Fuz_f*$_i$. The upper bound participation ratio of the fuzzy feature *Fuz_f*$_i$(1≤i≤n) is expressed as $\overline{PR}$(***Fuz_c***, *Fuz_f*$_i$), defined as the ratio of the sum of the upper bound membership degrees of the non-repeated instances of fuzzy feature *Fuz_f*$_i$ in the table-instance of fuzzy co-location pattern ***Fuz_c*** to the sum of the upper bound membership degrees of all the instances in *Fuz_f*$_i$.

Definition 8 (upper bound participation index and lower bound participation index) The upper bound participation index of fuzzy co-location pattern ***Fuz_c*** is expressed as $\overline{PI}$ (***Fuz_c***), which is defined as the minimum of the upper bound participation ratio of all fuzzy features *Fuzz_f*$_i$ in ***Fuz_c***. the lower bound participation index of fuzzy co-location pattern ***Fuz_c*** is expressed as $\underline{PI}$(***Fuz_c***), which is defined as the minimum of the lower bound participation ratio of all fuzzy features *Fuzz_f*$_i$ in ***Fuz_c***.

Definition 9 (Prevalent fuzzy co-location patterns) According to Definition 8, a fuzzy co-location pattern ***Fuz_c*** has upper bound participation index and lower bound participation index. When the user gives a threshold *min_prev*, when the upper bound participation index $\overline{PI}$(***Fuz_c***) <*min_prev*, ***Fuz_c*** is called an absolute non-prevalent fuzzy pattern; when the lower bound participation index $\underline{PI}$(***Fuz_c***) ≥*min_prev*, ***Fuz_c*** is called an absolute prevalent fuzzy pattern; when the lower bound participation index $\underline{PI}$(***Fuz_c***) <*min_prev*≤upper bound participation index $\overline{PI}$(***Fuz_c***), ***Fuz_c*** is called a pattern with prevalent tendency degree, and the prevalent tendency degree of the pattern is:

$$\sigma = (\overline{PI}(\boldsymbol{Fuz\_c}) - min\_prev)/(\overline{PI}(\boldsymbol{Fuz\_c}) - \underline{\mathrm{PI}}(\boldsymbol{Fuz\_c}))$$

We call absolute prevalent fuzzy co-location patterns and fuzzy co-location patterns with prevalent tendency degree as prevalent fuzzy co-location patterns.

Lemma 1 In a candidate fuzzy co-location pattern ***Fuz_c***, if the upper bound participation index of its any fuzzy sub-patterns is less than the prevalence threshold *min_prev*, the ***Fuz_c*** is absolute non-prevalent.

Proof: For a n size pattern ***Fuz_c*** , and a n+1 size pattern contain it is ***Fuz_c'***, according to the downward closure, $\overline{PI}$(***Fuz_c'***)≤$\overline{PI}$(***Fuz_c***), if the n size pattern ***Fuz_c*** is absolute non-prevalent, then $\overline{PI}$ (***Fuz_c***) < *min_prev*. Therefore, $\overline{PI}$(***Fuz_c'***)≤$\overline{PI}$(***Fuz_c***)< *min_prev*, then *n*+1 size pattern ***Fuz_c'*** is also absolute non-prevalent.

This lemma can be used as our pruning strategy.

4. A fuzzy Co-location mining algorithm based on type-2 fuzzy membership function and Join-based

4.1. Construct type-2 fuzzy membership function

Based on the original type-1 membership function, we adopt the equal expansion method to generate the type-2 membership function based on elliptic curve, and use a gradual method to adjust the parameters of the function, so as to ensure that the footprint of uncertainty (FOU) in the type-2 membership function meets the connectivity and given threshold. We take the interval membership degree of fuzzy set Cu(M) as an example when the copper content is 10 to 50 (we choose the length of interval content as 5 and the length of interval membership degree as 0.1). This example is used to illustrate the steps of constructing type-2 fuzzy membership function.

We have counted the data of 1000 geologists who evaluated the interval value of copper content under the fuzzy set Cu.(M). The darker the color is, the more people are evaluated in this interval. Figure 1 shows the number of evaluators for the interval membership of the interval data.

According to the construction method of type-1 fuzzy membership function of the content of heavy metal copper in [7], we get the type-1 fuzzy membership function of heavy metal copper. We draw the type-1 membership function, which is expressed as y=$\frac{x-a}{b-a}=\frac{x-10}{40}$, As shown in Figure 2. Among them, *a* is the lowest value that experts believe the heavy metal content tends to belong to middle content, and *b* is the value that experts believe the copper content must belong to middle content.

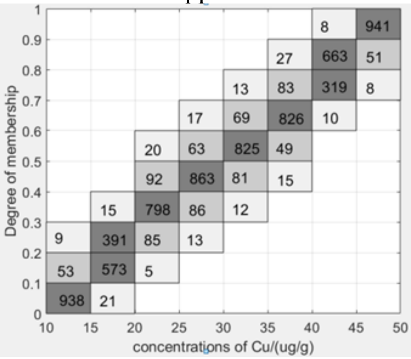


Figure 1. the number of evaluators for the interval membership of the interval data

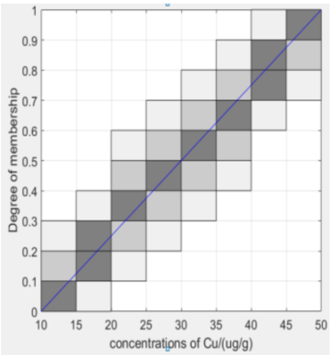


Figure 2. type-1 fuzzy membership function

We construct a type-2 membership function based on the original type-1 membership function. The type-2 membership function we proposed is represented by the upper bound membership function and the lower bound membership function. When the original type-1 membership function is a decreasing function, the formula is expressed as:

$$\text{Upper bound function } \bar{u}(x)=\sqrt[\eta]{1-\left(\frac{x-a}{b-a}\right)^{\eta}}$$

Lower bound function $\underline{u}(x)$ = 1-$\sqrt[\eta]{1-(\frac{b-x}{b-a})^{\eta}}$ $(a \leq x \leq b)$

When the original type-1 membership function is an increasing function, the formula is expressed as:

Upper bound function $\bar{u}(x)$ = $\sqrt[\eta]{1-(\frac{b-x}{b-a})^{\eta}}$

Lower bound function $\underline{u}(x)$ = 1-$\sqrt[\eta]{1-(\frac{x-a}{b-a})^{\eta}}$ $(a \leq x \leq b)$

Where, a is the lowest value that experts start to believe it has a tendency to belong to the fuzzy set, and b is the value that experts believe it must belong to the fuzzy set.

The area covered by the upper bound function and the lower bound function is expressed as the footprint of uncertainty (FOU). Among them, η determines the degree of expansion and contraction of the elliptic curve. The larger the η, the larger the FOU; the smaller the η, the smaller the FOU. For example, when η=1.1, our type-2 membership function is shown in Figure 3.

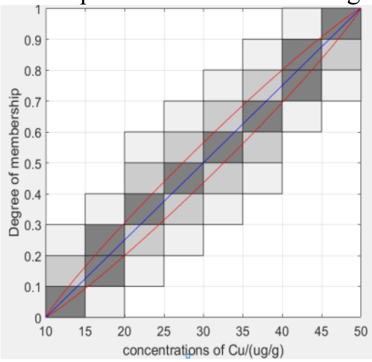


Figure 3. Type-2 membership function when η=1.1

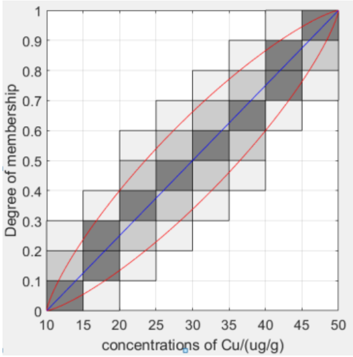


Figure 4. Type-2 membership function when η=1.283

The granular data indicates that for any specific value in the interval data, the possibility of taking any specific membership value in the interval membership is the same. Therefore, for an interval data, the confidence of the type-2 membership function = $\sum_{i=1}^{m} Area\ ratio\ S_i\ of\ particles\ in\ FOU *\ The\ number\ x\ of\ evaluators\ in\ this\ particle$, Among them, m is the number of particles in each interval data. Within a range, the average of the confidence of all interval data is the confidence of the type-2 membership function in this range.

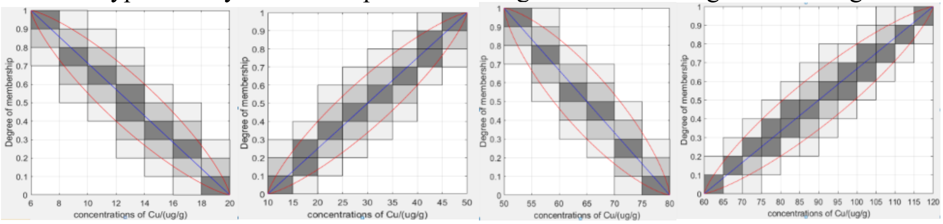
We use the gradual method to determine the unique parameter η. First, we set an increment α= 0.1, when η=1+$k_1\alpha$ (k_1=1,2,…,n), Observe whether the FOU is continuous. If there is φ ($\varphi \in \{1,2,\ldots,n\}$), when η=1+$\varphi\alpha$, FOU is continuous, when η=1+(φ+1)α, FOU is not continuous, then set the gradual value to one-tenth of the original (i.e. β = 0.1α), η=1+$\varphi\alpha$+$k_2\beta$ ($k_2 \in \{1,2,\ldots,10\}$), and so on, until the accuracy reaches the user's requirement and the FOU is connected. This process is shown in Table 2.

Table 2. connectivity of FOU with different values

η	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Is it connected?	Yes	Yes	No	No	No	No	No	No	No	No
η	1.21	1.22	1.23	1.24	1.25	1.26	1.27	1.28	1.29	1.3
Is it connected?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No
η	1.281	1.282	1.283	1.284	1.285	1.286	1.287	1.288	1.289	1.29
Is it connected?	Yes	Yes	Yes	No	No	No	No	No	No	No

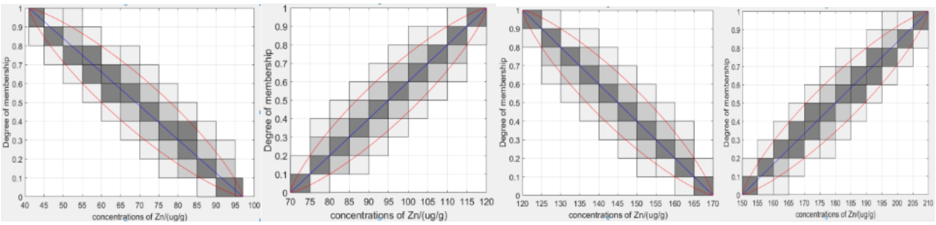
When η=1.283, the FOU of the type-2 membership function is connected. At this time, the data in the FOU accounts for 85.13% of the total evaluation data, that is, the confidence of FOU is 0.8513, which reaches the default threshold of 0.8. As shown in Figure 4.

The type-2 fuzzy membership function we get is shown in Figure 5 and Figure 6.



(a) Low content (b) Middle content (c) Middle content (d) High content

Figure 5. type-2 membership function of Cu



(a) Low content (b) Middle content (c) Middle content (d) High content

Figure 6. type-2 membership function of Zn

As shown in Figure 5 and Figure 6, the red lines represent the upper bound membership function and the lower bound membership function of the type-2 membership function. The parameters of Cu's type-2 membership function are η_1=1.296, η_2=1.283, η_3=1.272, η_4=1.272; the parameters of Zn's type-2 membership function are η_5=1.28, η_6=1.284, η_7= 1.284, η_8=1.272. The confidence of the type-2 membership function of Cu are 0.8502, 0.8513, 0.8172, 0.8953, and the confidence of the type-2 membership function of Zn are 0.8605, 0.8904, 0.8910, and 0.9022, respectively.

4.2. Prevalent fuzzy Co-location pattern mining method based on Join-based

In this section, we propose a fuzzy co-location pattern mining method based on type-2 fuzzy membership function and Join-based.

A. Fuzzy co-location pattern mining algorithm

Input: (1) Fuzzy feature set ***Fuz_Fea*** $=\{Fuz\_f_1, Fuz\_f_2,\ldots, Fuz\_f_n\}$

(2) Instance set of Fuzzy features $\boldsymbol{S} = \{s_1, s_2, \ldots, s_n\}$

(3) Minimum distance threshold *min_dist*

(4) Minimum prevalence threshold *min_prev*

Output: prevalent fuzzy co-location pattern set *Fre_P*

Variable：

C_n: *n*-size candidate fuzzy co-location patterns

T_n: Table instance of C_n

P_n: Prevalent *n*-size fuzzy co-location patterns

Step:

```
1.    $n = 1$; $Fre\_P = \emptyset$; $C_1 = Fuzz\_Fea$, $P_1 = Fuzz\_Fea$;
```

```
2.  while($P_n \neq \Phi$){
3.      $C_{n+1}$ = Gen_Candidate($P_n$,$n$);
4.      $T_{n+1}$ = Gen_table_instance($C_{n+1}$, $T_n$,$min\_dist$);
5.      $P_{n+1}$ = select_prev_pattern($C_{n+1}$, $T_{n+1}$,$min\_prev$);
6.      $Fre\_P = Fre\_P \cup P_{n+1}$ ;
7.      $n = n + 1$;}
8.      Return $Fre\_P$
```

Step 1 initializes size-1 candidate fuzzy pattern and size-1 prevalent fuzzy pattern(all fuzzy features are considered as the first size prevalent pattern). Then, the following steps are carried out: In step 3, Apriori-like method is used to generate *n*+1 size candidate fuzzy patterns from *n* size prevalent fuzzy patterns; Step 4 generates the table-instance of *n*+1 size candidate fuzzy pattern from the table-instance of *n* size prevalent fuzzy pattern by connecting; In Step 5, we calculate the upper bound participation index and the lower bound participation index of *n*+1 size candidate fuzzy patterns according to the table-instances of *n*+1 size candidate fuzzy patterns. By comparing the prevalence threshold given by the user, we filter out *n*+1 size prevalent fuzzy patterns.

B. Generate candidate fuzzy patterns

When generating candidate fuzzy patterns, some of them have no practical significance, such as {A.Cu(L), A.Cu(M), A.Cu(H)}, that is, the fuzzy co-location pattern composed of fuzzy features of the same functional area between different fuzzy sets of the same attribute has no practical significance. In order to avoid producing such meaningless candidate fuzzy patterns, we propose a bucket-based candidate fuzzy pattern generation algorithm, which puts the fuzzy features of the same attribute in the same functional area into the same bucket, and then selects the fuzzy features from different buckets to form the candidate fuzzy patterns. The steps of this algorithm are as follows:

Input: (1) Fuzzy feature set ***Fuzz_Fea*** ={*Fuz_f*$_1$, *Fuz_f*$_2$,…,*Fuz_f*$_n$}
(2) Fuzzy feature instances' set of ***S*** = {s_1,s_2,…,s_n}
(3) Minimum distance threshold *min_dist*
(4) Minimum prevalence threshold *min_prev*

Output: C_{n+1}: *n*+1 size candidate fuzzy co-location patterns

Step:

```
if($n$ == 1)
  for each Fuz_f in Fuz_Fea do
    Initialize_bucket(Fuz_f)
    SetId(Fuz_f)
else if($n$ == 2)
  for each Fuz_$c_1$ in bucket
    for each Fuz_$c_2$ in other bucket
    Generate a candidate pattern from Fuz_$c_1$ and Fuz_$c_2$
else
  for($i$=1; $i$<bucket_count; i++)
    for each fuzzy pattern Fuz_$c_1$ in bucket[i]
      for each fuzzy pattern Fuz_$c_2$ in bucket[i]
        Generate a pattern Fuz_$c_{n+1}$
        if(check($n$-1, Fuz_$c_{n+1}$, bucket))
                Fuz_$c_{n+1}$ is a new candidate pattern
```

Steps 2 to 4 first initialize each bucket, and then number all fuzzy features in the bucket. Among them, for any two fuzzy features in the same bucket, there are Id(Fuz_f_i)%*bucket_count* == Id(Fuz_f_j)%*bucket_count*, that is, put all fuzzy features with the same features and the same attributes into the same bucket.

In Step 6 to Step 8, we select two fuzzy features from different buckets in order, to form a size-2 fuzzy pattern. Step 10 to 15 generate *n*+1 size candidates from *n* size prevalent fuzzy patterns: first, for any two prevalent fuzzy patterns, check whether the fuzzy features of the first *n*-1 of the two fuzzy patterns are same, if they are same, then connect them to generate a *n*+1 size candidate. At the same time, check whether all its *n* size subsets are prevalent.

C. Generate table-instance

As we know, for a fuzzy co-location pattern, all of its row-instances constitute its table-instance. The process of obtaining the row-instances of fuzzy co-location candidates can be divided into the following 4 steps:

(1) First, we materialize the neighborhood relationship between instances, and then we can get all row-instances of size-2 co-location patterns.

(2) For a candidate fuzzy pattern, we find the co-location pattern consisting of all the non-recurring features in the fuzzy candidate pattern. Then, we find the row-instances that have all the fuzzy features of the fuzzy candidate pattern from the table-instance of the co-location pattern, these row-instances are the row-instances of the candidate fuzzy pattern. For example, for the fuzzy pattern {C.Cu(L), C.Zn(L), D.Cu(M)}, the row-instances whose attributes satisfy C.Cu(L), C.Zn(L) and D.Cu(M) are found in all row-instances of pattern {C, D} as row-instances of candidate fuzzy pattern.

(3) If there is only one feature in the size-2 candidate fuzzy pattern, in this feature, the instances that satisfy all the fuzzy features in the candidate fuzzy pattern are found as the row-instances of the candidate fuzzy pattern. For example, for the fuzzy pattern {B.Cu(M), B.Zn(H)}, in the instances of feature B, find the instances that satisfy both Cu(L) and Zn(M) simultaneously as the row-instances of the fuzzy pattern {B.Cu(M), B.Zn(H)}.

4.3. Time performance analysis

The time cost of our proposed method is mainly divided into three parts: generating candidate fuzzy co-location patterns, generating all row-instances through connection and finding table-instances of the candidates. The computational complexity of generating candidate patterns is O(*C*), where *C* is the number of candidates generated by our method; The computational complexity of generating table-instances of candidates is O(*R*), where *R* is the number of row-instances generated by connection operations; The computational complexity of finding the table-instance of the candidates is O(*T*), *T* is the number of all row-instances found for all candidates.

5. Experimental Evaluation

In this section, in order to verify the effectiveness and efficiency of our proposed algorithm, a wide range of experiments have been done. We used a method similar to [6] to generate synthetic data sets. We also used the real data set of heavy metal content of topsoil samples in Quanzhou to evaluate our method.

All our experiments are carried out on C#, The Intel PC we use has windows 10 operating system and Intel Core i5-4258@ 2.40GHz and 8GB memory.

5.1. Results on synthetic data sets

First of all, we set the size of the instance distribution area to *D*D*, We divide the entire region into grids of size *min_dist*min_dist*, where, *min_dist* is the distance threshold of instance proximity. After determining the number of features and instances, we determine the number of instances of each feature according to Poisson distribution.

(1) The effect of distance threshold on results

We change the distance threshold to find the effect of different distance threshold on the execution time and the pruning rate. As shown in Figure 7, we find that with the increase of distance thresholds, the execution time of our method will continue to increase, because with the increase of distance thresholds, the number of neighbor instances will increase, so the time consumption will also increase.

(2) The effect of the prevalence threshold on the results

We change the prevalence threshold to find the effect of different prevalence threshold on the execution time and the pruning rate. As shown in Figure 8, we find that the execution time of our methods decreases with the increase of the prevalence threshold. In the case of higher prevalence thresholds, our method can prune a lot of candidate fuzzy co-location patterns, so when the prevalence threshold increases, the time consumption of our method will reduce rapidly.

(3) The effect of the number of fuzzy features on the results

We change the number of fuzzy features to find the effect of different number of fuzzy features on execution time and the pruning rate. As shown in Figure 9, we find that the execution time of our method first increases and then decreases. This is because when the number of fuzzy features increases from 18 to 36, the number of candidate fuzzy co-location patterns increases rapidly, and the generation of row-instances of candidate patterns also increases. When the number of fuzzy features continues to grow (when the number of fuzzy features increases from 36 to 90), the average number of row-instances of candidate patterns will be greatly reduced. Our proposed method has downward closure, so it can quickly prune the absolute non-prevalent candidate patterns. The number of fuzzy features is equal to the number of features multiply the number of fuzzy attributes multiply the number of fuzzy sets. In this experiment, the number of fuzzy attributes is 2, the number of fuzzy sets is 3, and the number of features selected are 3,6,9,12,15 respectively.

(4) The effect of the number of instances on the results

We change the number of instances to find the effect of different instance number on execution time and the pruning rate. As shown in Figure 10, we find that with the increase of the number of instances, the execution time of our method continues to increase. This is because in a region, with the increase of the number of instances, the number of neighbor instances will also increase, that is, the number of row-instances of candidate patterns will increase, and the time consumption will also increase greatly.

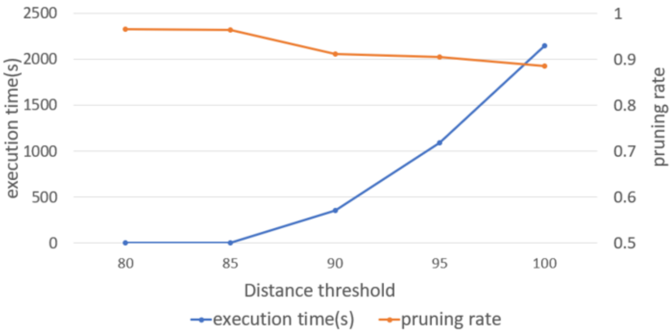


Figure 7. Execution time and the pruning rate varies with the distance threshold

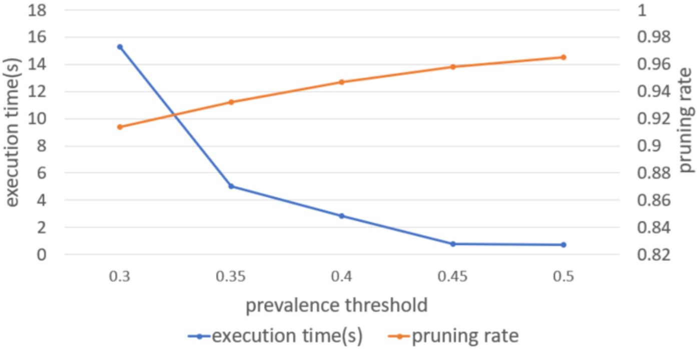


Figure 8. Execution time and the pruning rate varies with the prevalence threshold

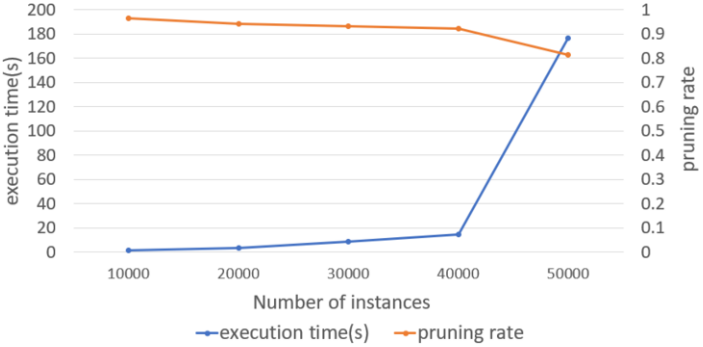


Figure 9. Execution time and the pruning rate varies with the number of instances

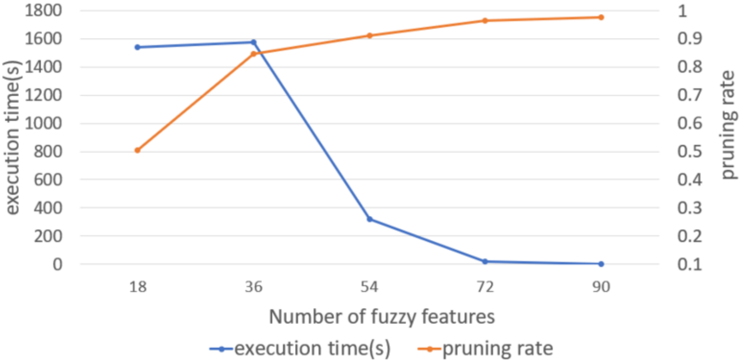


Figure 10. Execution time and the pruning rate varies with the number of fuzzy features

5.2. Results on real data sets

In this section, we use the spatial distribution data set of heavy metals in the topsoil of Jiayuguan City as the real data set to test the effectiveness of our method. As shown in Figure 11, it is divided into four functional areas: industrial area, agricultural area, living area and Gobi area. According to the proportion of the length of the area, we project these data on the space of 850 * 670.

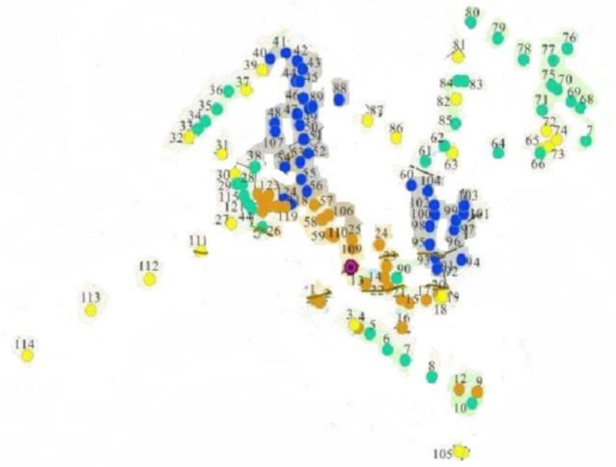


Figure 11. data set of heavy metals in the topsoil of Jiayuguan City

Table 3. Mining results of the real data set

Prevalent fuzzy co-location pattern	**Prevalence tendency degree**	**Lower participation index and Upper participation index**	**Participation index in traditional method**
{A.Cu(L), C.Cu(L)}	0.6125	0.0709,0.6621	0.2312
{A.Cu(L), D.Cu(L)}	0.2328	0.036,0.3801	0.1169
{A.Cu(M), C.Cu(M)}	0.2636	0.1567,0.3513	0.2384
{D.Cu(M), D.Zn(H)}	1	0.4781,0.9135	0.6599
{D.Cu(L), D.Zn(M)}	0.9096	0.2414,0.8896	0.6681
{A.Zn(L), A.Cu(M), C.Zn(L)}	0.2636	0.1567,0.3513	0.2384
{A.Zn(L), C.Zn(L), D.Cu(M)}	0.2167	0.1749,0.3346	0.2492
{B.Cu(L), B.Zn(L), C.Zn(L)}	0.8586	0.2597,0.5448	0.4342
{B.Cu(M),B.Zn(L),C.Cu(M) ,C.Zn(M)}	0.7039	0.1804,0.5844	0.3358
{B.Cu(M),B.Zn(L),C.Cu(M) ,C.Zn(L)}	1	0.3933,0.738	0.5726

In the real data set, we set the distance threshold to 50 and the prevalence threshold to 0.3, 112 prevalent fuzzy co-location patterns (26 absolutely prevalent fuzzy co-location patterns and 86 fuzzy co-location patterns with prevalence tendency degree) were mined according to our measurement method, some of which are shown in Table 3, and 63 prevalent fuzzy co-location patterns were mined according to the traditional measurement method. The sampling points of functional area A are represented by blue dots, the sampling points of functional area B are represented by green dots, the sampling points of functional area C are represented by brown dots, the sampling points of functional area D are represented by yellow dots. A, B, C and D represent industrial area, agricultural area, residential area and Gobi area respectively in this data set. For example, the fuzzy pattern {A.Cu(M), C.Cu(M)} indicates that the industrial area with middle copper content and the residential area with middle copper content are prevalently located together, and the prevalence tendency degree of this pattern is 0.2636. {B.Cu(M), B.Zn(L), C.Cu(M), C.Zn(M)} indicate agricultural areas with low zinc content and middle copper content, residential areas with middle zinc content and middle copper content tend to be often located together, and the degree of this tendency is 0.7039.

When the fuzzy membership function is uncertain, our method can find the fuzzy co-location pattern that is always prevalent, which is helpful for us to find the stable prevalent fuzzy co-location pattern. Our method can also find the fuzzy co-location pattern that has a certain tendency to be prevalent, which is helpful for us to find the potential fuzzy co-location pattern and predict the occurrence of some prevalent fuzzy co-location patterns. However, the method of mining fuzzy co-location pattern based on type-1 fuzzy membership function can only divide the fuzzy patterns into prevalent fuzzy patterns and non-prevalent fuzzy patterns according to the participation index and prevalence threshold. When there is uncertainty in the fuzzy membership function, this method cannot judge which fuzzy patterns are stable prevalent, which fuzzy patterns have certain tendency are prevalent.

6. Conclusion

Traditional spatial co-location pattern mining methods can not directly mine fuzzy co-location patterns. Therefore, in this paper, we propose a fuzzy co-location pattern mining algorithm based on type-2 fuzzy sets and join-based algorithm to mine prevalent fuzzy co-location patterns from spatial instances with attribute information. Since the fuzzy attributes of spatial data are uncertain, we propose the concepts of upper and lower bound participation ratios of fuzzy features, upper and lower bound participation indexes of fuzzy co-location patterns, to measure the prevalence degree of fuzzy patterns. We also propose a pruning strategy, which effectively prunes the absolute non-prevalent fuzzy co-location patterns in the process of mining prevalent fuzzy co-location patterns.

The method of mining fuzzy co-location pattern based on type-2 fuzzy membership function has great practical significance. For example, it can help us find out whether the industrial areas will cause heavy metal pollution to the surrounding residential areas, and then find out whether the pollutants in the industrial area are related to the epidemic of the surrounding residents; In the distribution of urban facilities, we can see whether the per capita consumption level of the commercial areas is related to the housing prices of the surrounding residential areas, and so on. Our method also has some limitations: generating table-instances of candidate fuzzy patterns require a lot of connection operations, so the efficiency is not high; It does not consider how to mine fuzzy co-location patterns when different features have different attributes.

In the future work, we consider that there are several directions for further researches:

(1) Propose new fuzzy co-location pattern mining algorithms to improve the efficiency of the fuzzy co-location pattern mining method.

(2) Use more real data sets to verify the effectiveness of our proposed methods, for example, a data set of tumor diseases in a certain area and pollutant emissions from

surrounding factories.

References

[1] J.S. Yoo, S. Shekhar, M. Celik. A join-Less approach for co-location pattern mining: a summary of results, in: Proc. IEEE ICDM, 2005, pp. 813-816.

[2] Y. Huang, S. Shekhar, H. Xiong. Discovering colocation patterns from spatial datasets: a general approach, IEEE Trans. Knowl. Data Eng. 2004, 16 (12): 1472-1485.

[3] X. Bao and L. Wang. A clique-based approach for co-location pattern mining. 2019, 490 (2019): 244-264.

[4] A. Berry, R. Pogorelcnik. A simple algorithm to generate the minimal separators and the maximal cliques of a chordal graph, Inf. Process. Lett. 2011, 111 (11): 508-511.

[5] Cenk Ulu, Müjde Güzelkaya, Ibrahim Eksin. Granular type-2 membership functions: A new approach to formation of footprint of uncertainty in type-2 fuzzy sets. 2013, 13(8): 3713-3728.

[6] F. Wang, H. Mo. Some basic questions about type-2 fuzzy sets. Acta Automatica Sinica,2017, 43(07): 1114-1141.

[7] Z. Wang, Kwong-Sak Leung, George J. Klir. Applying fuzzy measures and nonlinear integrals in data mining. Fuzzy Sets and Systems, 2005, 156(3): 371-380.

[8] F. E. Petry, L. Zhao. Data mining by attribute generalization with fuzzy hierarchies in fuzzy databases. Fuzzy Sets and Systems, 2009, 160(15): 2206-2223.

[9] C. Lee, H. Pan. Performance enhancement for neural fuzzy systems using asymmetric membership functions. Fuzzy Sets and Systems 2009, 160 (7): 949-971.

[10] J. M. Mendel and H. Wu. "Type-2 Fuzzistics for Symmetric Interval Type-2 Fuzzy Sets: Part 1, Forward Problems," IEEE Transactions on Fuzzy Systems, 2006, 14(6): 781-792.

[11] T.W. Liao. A Procedure for the Generation of Interval Type-2 Membership Functions from Data, Applied Soft Computing Journal http://dx.doi.org/10.1016/j.asoc.2016.09.034.

[12] E. Kayacan, A. Sarabakha, S. Coupland, et al. Type-2 fuzzy elliptic membership functions for modeling uncertainty[J]. Engineering Applications of Artificial Intelligence, 2018, 70:170-183.

[13] P. Wu, L. Wang, Y. Zhou. Research on mining spatial co-location patterns with fuzzy attributes. Computer Science and Exploration, 2013, 7(04): 348-358.

[14] L. Wang, Y. Bao, J. Lu. et al. A new join-less approach for co-location pattern mining. In Proc. of the IEEE 8th International Conference on Computer and Information Technology (CIT2008), Piscat-away, NJ: the IEEE Computer Society Press, 2008: 197-202.

[15] L. Wang, L. Zhou, J. Lu. et al. An Order-clique-based Approach for Mining Maximal Co-locations Information Sciences, 2009, 179 (19): 3370-3382