APTIKOM Journal on Computer Science and Information Technologies
Vol. 2, No. 1, 2017, pp. 1~7
ISSN: 2528-2417
DOI: 10.11591/APTIKOM.J.CSIT.41

An Improved Approximation Algorithm for Co-location Mining in Uncertain Data Sets using Probabilistic Approach

M. Sheshikala*[1], D. Rajeswara Rao[2], Md. Ali Kadampur[3]
[1,3]Department of Computer Science & Engineering, SR Engineering College, India
[2]Department of Computer Science & Engineering, KL University, India
*Corresponding author, e-mail: marthakala08@gmail.com

Abstract

In this paper we investigate colocation mining problem in the context of uncertain data. Uncertain data is a partially complete data. Many of the real world data is Uncertain, for example, Demographic data, Sensor networks data, GIS data etc.,. Handling such data is a challenge for knowledge discovery particularly in colocation mining. One straightforward method is to find the Probabilistic Prevalent colocations (PPCs). This method tries to find all colocations that are to be generated from a random world. For this we first apply an approximation error to find all the PPCs which reduce the computations. Next find all the possible worlds and split them into two different worlds and compute the prevalence probability. These worlds are used to compare with a minimum probability threshold to decide whether it is Probabilistic Prevalent colocation (PPCs) or not. The experimental results on the selected data set show the significant improvement in computational time in comparison to some of the existing methods used in colocation mining.

Keywords*: probabilistic approach, colocation mining, spatial data set, prevalent colocation, PPC*

1. Introduction

Basically colocation mining is the sub-domain of data mining. The research in colocation mining has advanced in the recent past addressing the issues with applications, utility and methodsof knowledge discovery. Many techniques inspired by data base methods (Join based, Join-less, Space Partitioning, etc.,) have been attempted to find the prevalent colocation patterns in spatial data. Fusion and fuzzy based methods have been in use. However due to growing size of the data and computational time requirements highly scalable and computationally time efficient framework for colocation mining is still desired. This paper presents a computational time efficient algorithm based on Probabilistic approach in the uncertain data.

Consider a spatial data set collected from a geographic space which consists of features like birds (of different types), rocks, different kinds of trees, houses, which is shown in Figure 4. From this the frequent patterns on a spatial dimension can be identified, for example, <*bird, house*> and <*tree, rocks*>, the patterns are said to be colocated and they help infer a specific eco-system. This paper presents a computationally efficient method to identify such prevalent patterns from spatial data sets. Since the object data is scattered in space (spatial coordinates) extractinginformation from it is quite difficult due to complexity of spatial features, spatial data types, and spatial relationships.

For example, a cable service provider may be interested in services frequently requested by geographical neighbours, and thus gain sales promotion data. The subscriber of the channel is located on wide geographical positions and has wide ranging interest/preferences. Further in the process of collecting data there may be some missing links giving rise to uncertainty in the data. From the data mining point of view all this adds to complexity of analysis and needs to be handled properly. The paper addresses the uncertainty and data complexity issues in finding prevalent colocations.

The paper includes 1. The methods for finding the exact Probabilistic Prevalent colocations (PPCs). 2. Developing a dynamic programming algorithm to find Probabilistic Prevalent colocations (PPCs) which dramatically reduces the computation time. 3. Results of application of the proposed method on different data sets. The remaining paper is organized as follows: In Section-1, we discuss the introduction, and related work is discussed in Section-2. In section-3 we discuss the definitions, and a block diagram to show the complete flow to find PPCs are discussed in section-4,

Received November 30, 2016; Revised January 8, 2017; Accepted February 16, 2017

In section-5 we discuss dynamic- programming algorithm for finding all Probabilistic Prevalent Colocations. We show the experiment results in Section-6.Finally, in section-7 we suggest future work.

2. Related Work

Many methods have been extensively explored in order to find the prevalent colocations in spatially precise data. Some of these methods are:

2.1. Space Partitioning Method

This approach finds the neigh-boring objects of a subset of features. It finds the partition center points with base objects and decomposes the space from partitioning points using a geometric approach and then finds a feature within a distance threshold from the partitioning point in each area. This approach may generate incorrect colocation patterns, because it may miss some of the colocation instances across partition areas.

2.2. Join-Based Approach

This approach finds the correct and complete colocation instances, first it finds all neighboring pair objects (of size 2) using a geometric method, the method finds the instance of size k(> 2) colocations by joining the instances of its size k-1 sub1``set colocation where the first k-2 objects are common. This approach is computationally expensive with the increase of colocation patterns and their instances.

2.3. Join-Less Approach

The join-less approach puts the spatial neighbor relationship between instances into a compressed star neighborhood. All the possible table instances for every colocation pattern were generated by scanning the star neighborhood, and by 3-time filtering operation. This join-less colocation mining algorithm is efficient since it uses an instance look-up schema instead of an expensive spatial or instance join operation for identifying colocation table instances, but the computation time of generating colocation table instances will increase with the growing length of colocation pattern.

2.4. CPI-tree Algorithm

This algorithm proposed by Wnag et al in [11] developed in new structure called CPI-tree (colocation pattern instance tree) which could materialize the neighbor relationships of spatial data sets, and find all the table instances recursively from it. This method gives up Apriori like model, (i.e.) to generate size-k prevalence colocations after size (k-1) prevalence colocations, but Apriori candidate generate-test method reduces the number of candidate sets significantly and leads to performance gain.

3. The Basic Definitions

3.1. Spatial Data

Spatial data also known as geo-spatial data is the information which identifies the geographic location of features and boundaries on Earth, such as Forests, Oceans etc., Usually Spatial data is stored in terms of numeric values.

3.2. Colocation Mining

It is the process of finding patterns that are colocated in nearby regions.Co-location rule process finds the subsets of features whose instances are frequently located together in geographic space. It is found that classical data mining techniques are often inadequate for spatial data mining and different techniques need to be developed. For this we discuss the co-location pattern mining over spatial data sets. Many important applications use colocation mining. For example:

1. NASA (studying the climatologically effects, land use classification),
2. National Institute of Health (predicting the spread of disease),
3. National Institute of Justice (finding crime hot spots),
4. Transportation agencies (detecting local instability in traffic).

3.3. Spatial Colocation Mining:

It is a group of spatial features whose instances are frequently located around the geographic space. Let F= $\{f_1, f_2, \ldots\ldots\ldots\ldots f_n\}$ be the set of features and Z= $\{P_1, P_2, \ldots\ldots\ldots\ldots\ldots, P_n\}$

where $\{P_1, P_2, \ldots\ldots\ldots\ldots, P_n\}$ are the subsets of features $\{f_1, f_2, \ldots\ldots\ldots\ldots f_n\}$ Let T be the threshold set {d, min_ *prev*, P_m} then C ∈ Z such that for C, T is valid. For example from the Figure 1 we can identify the features and instances related in a spatial data set.

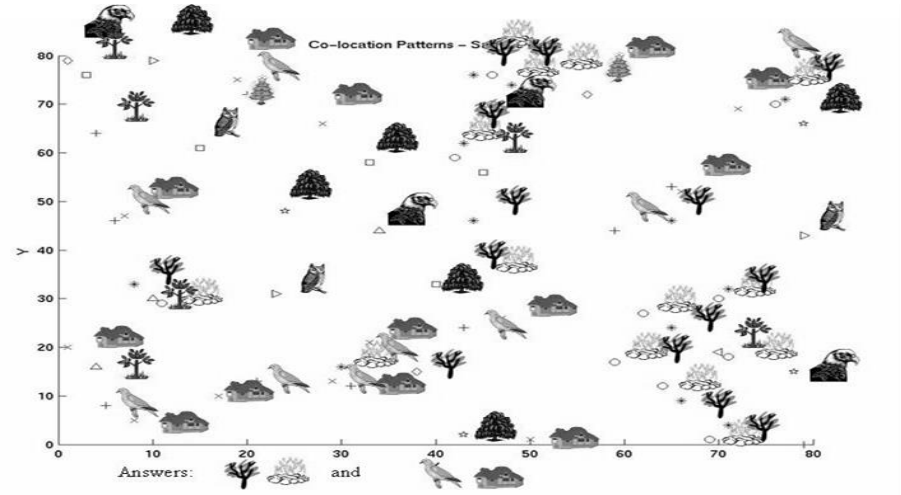

Figure 1. Example of Spatial Colocation data

From the Figure 1 we can identify that there are different types of features like tree, Bird, Rocks and House and we have instances for the features like trees which are of various types of trees, and Birds which are like Eagle, Sparrow, Owl, and the Features like rock and house are having only one kind of instance. From the Figure we can conclude that rocks and a type of tree is colocated, Sparrow and house are colocated. From the Figure 1 we can identify that there are different types of features like tree, Bird, Rocks and House and we have instances for the features like trees which are of various types of trees, and Birds which are like Eagle, Sparrow, Owl, and the Features like rock and house are having only one kind of instance. From the Figure 1 we can conclude that rocks and a type of tree is colocated, Sparrow and house are colocated.

3.4. Instance of a Feature:

The instances of a feature are the existential probability of the instance in the place location. If F is a feature then $F.i$ is an instance.

3.5. Spatially Uncertain Feature:

A spatial feature contains the spatial instances, and a data set Z containing spatially uncertain features is called spatially uncertain data set. If Z is a data set then set of features is A, B, C. Shown in Figure 2.

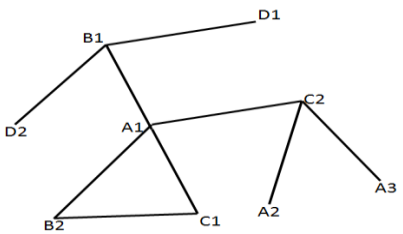

Figure 2. Distribution of example spatial Instance

3.6. Probability of Possible Worlds

For each colocation of k-size, c={$f_1, f_2, \ldots\ldots\ldots f_n$of each instance $F.i$ there are two different possible worlds (i) one among them is that the instance is present (ii) and the other is absent. Take the set of features F={$f_1, f_2, \ldots\ldots\ldots f_n$} and the set of instances S={$S_{f_1}, S_{f_2,} \ldots\ldots\ldots\ldots, S_{f_n}$}, where $S_{f_i}(1 <= i <= k)$ is the set of instances in S and there are $2^{|S|}= 2^{\left|s_{f_1}, s_{f_2,\ldots\ldots\ldots\ldots\ldots, s_{fn}}\right|}$possible worlds at most. Each Possible world w is associated with a probability P (w) that is the true world, where P (w) > 0.

3.7. Neib_tree

The Neib_tree is constructed for the Figure 2 which indicates the existence of the path from one feature to the other. If there is a path it indicates that a table instance is existing. This Neighboring tree eliminates the duplicates can be seen in Figure 3.

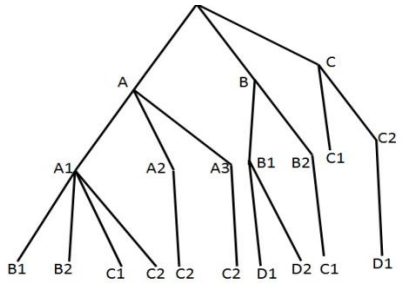

Figure 3. Neib_tree for Figure 2.

4. Block Diagram

Basic flow of co-location pattern mining: In this section, we present a flow diagram which describes the flow of identifying the Probabilistic Prevalent colocations. Given a Spatial data set, a neighbour relationship, and interest measure thresholds the basic colocation pattern mining involves 4 steps as in Figure 3.

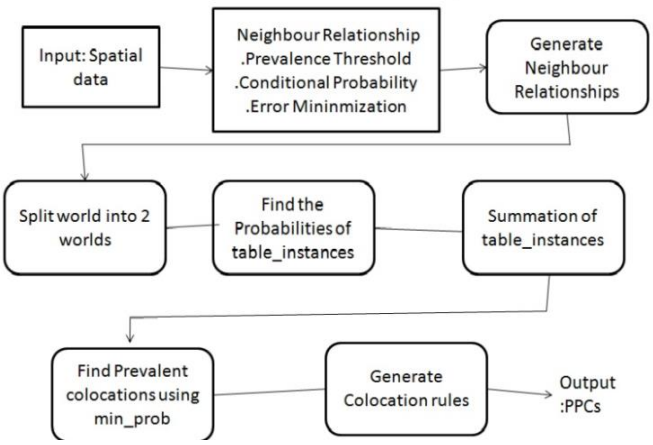

Figure 4. Block diagram to find the PPCs

First candidate colocation patterns are generated and the colocation instances and spitted into two worlds from the spatial data set. Next, find the probabilities using minimum prevalence and compute summation of table instances of each colocation, Next find prevalent colocation using minimum probability.

5. The Basic Algorithm

The algorithm (Algorithm-1) is designed to find allPPCs with (min_prev, min_prob) pairing. The algorithm uses dynamic approach where in it prunes out the candidates which are not prevalent and works on the reduced search space to find the PPCs. It uses an approximation approach by accepting an initial error that would be tolerated in finding the PPCs and thereby speeds up the process of finding the PPCs. The algorithm is presented below

Algorithm-1

Input:

$F = (f_1, f_2, \ldots\ldots\ldots, f_n)$ a set of Spatial Features;
S: A spatially uncertain data set;
min_prev: A minimum prevalence threshold;
min_prob: A minimum Probability Threshold;
e: An Approximation error;
Probability of table instances:

$P_r^{(c,f_1)}[0,0]_w = 1$;
$P_r^{(c,\bar{f}_1)}[0,0]_w = 1$;
$P_r^{(c,f_1)}[0,0]_w = 0\ (1 < i < 1)$;
$P_r^{(c,\bar{f}_1)}[0,0]_w = 0\ (1 < i < 1)$;

Output:-

(min_prev, min_prob) PPCs.

Begin

1) Read approximation error **e**.
2) if e=1 STOP
3) else
4) Call Neib_tree_gen (F,S,NHR); // to identify table instances.
5) Assign $P1 = F, k = 2$;
6) While (not empty P_{k-1}
7) and $k \leq n$) do
a) for each colocation $"W"$ of size $'k'$ compute Probabilities of worlds from equation-3:
b) Split W intoW_1 and W_2 where $W_1 = f_1.j > (f_2.j, f_3.j \ldots\ldots, f_n.j) W_2 = f_2, f_3 \ldots\ldots\ldots\ldots, f_n,$ and $W_2 \subseteq$ w;
c) for each set w=($f_1.l, \ldots\ldots\ldots, f_n.l$) compute Probability of table _instances as equation-4:.
d) for each w compute Prevalence Probability $P(PR^R(c) \geq min\ prev)_{w_{1+w}}$ as equation-5:
e) Compute the summation of all Prevalence Probabilities
$PPs = PPs + (P1 + P2 + \cdots\ldots\ldots\ldots + Pn)$
f) if ($PPs \leq min\_prob$) then c=c-C_k;
g) P_k=sel_prev_colocation(C_k,min_prev, min_prob);
h) $k = k + 1$;
i) $end\ while$;
8) STOP;
9) Return ($P2 \cup P3 \ldots\ldots\ldots\ldots \cup Pn$)

End.

$$P(W) = \prod_{i=1}^{n} \left(\prod_{(e \in s_{f_i}) \in w} P(e) * \prod_{(e \in s_{f_i}) \ni w} (1 - P(e)) \right) \quad (3)$$

$$P^{(c,f_1)}[i,j]_w = \begin{cases} P^{(c,f_1)}[i,j]_w if, f_1.j \in table_{instance_{w \cup f_1^j}}(c) \\ P^{(c,f_1)}[i,j-1]_w.(1-p_j) + P^{(c,f_1)}[i-1,j-1]_w.p_j \\ \quad if, f_1.j \in table_{instance_{w \cup f_1^j}}(c) \\ \quad and \geq j.min\_previstrue \\ \quad 0\ otherwise, \end{cases}$$

$$P^{(c,\bar{f}_1)}[i,j]_w = \begin{cases} P^{(c,f_1)}[i,j-1]_w \ \ if, f_1.j \notin table_{instance_{w \cup f_1^j}}(c) \\ P^{(c,\bar{f}_1)}[i,j-1]_w . (1-p_j) \ + \ P^{(c,\bar{f}_1)}[i-1,j-1]_w . p_j \\ \qquad\qquad otherwise, \end{cases} \tag{4}$$

$$\left(\sum_{i=1}^{l_1} P^{(c,f_1)}[i,j]_w \left(\sum_{j=0}^{\frac{|1-\min\_prev|}{\min\_prev} * i} P^{(c,\overline{f_1})}[j,l_1]_w \right) \right) \tag{5}$$

6. Results

The results are compared against a data set given in the Table 1 which consists of 7 features with an average of 2 instances. From Table 1 we get 2 PPCs when min_prev = 0.4 and min_prob = 0.4 and d=150, and = 0.001 and those PPCs are {1, 3} and {4, 5}, the result can be seen in the following Figure 5.

Table 1. A Synthetic Sample Data Set

Features	X-Coordinates	Y- Coordinates	Probability
0	328	1362	0.5
0	190	1140	0.4
0	392	1220	0.9
1	290	1264	0.1
1	330	1480	1
2	260	1278	0.1
3	185	1440	0.1
3	320	1500	0.4
3	330	1500	0.7
4	150	1580	0.1
4	150	1300	1
5	225	1300	1
5	260	1530	0.1
6	220	1650	0.4
6	60	1590	1

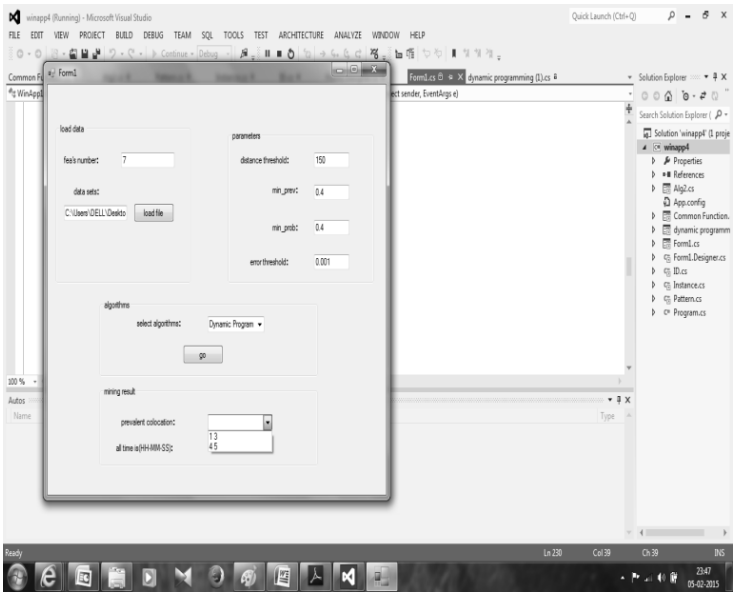

Figure 5. PPCs for Table 1 with $min\_prev = 0.4$ and $min\_prob = 0.4$, d=150,and $\varepsilon = 0.001$

From Figure 6, it is proved that the computation time for the improved approximation algorithm works well when compared to dynamic algorithm.

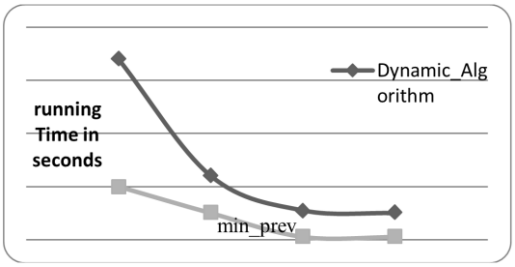

Figure 6. Varying min_prev and min_prob, d=150, and ε= 0.001

7. Conclusion

We have proposed a method for finding Probabilistic Prevalent Colocation in Spatially Uncertain data sets which are likely to be prevalent. We have given an approach in which the computation time is drastically reduced. Future Work can include the parallel computation for finding the Prevalent Colocation which are evaluated independently and this work can also be expanded to find the Probabilistic Prevalent colocations in other Spatially Uncertain data models, for example fuzzy data models and graphical spatial data. Further keeping in view the work can be extended to find the important sub functionalities in colocation mining to formulate colocation mining specific primitives for the next generation programmer which we can expect to evolve as a scripting language. In essence the scope of the work can cover data base technologies, parallel programming domain, graphical graph methods, programming language paradigms and software architectures.

References

[1] C.C. Aggarwal et al, *Frequent Pattern Mining with Uncertain Data*, Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 29-37, 2009.
[2] T. Bernecker, H-P Kriegel, M. Renz, F. Verhein, and A. Zuefle, *Proba-bilistic Frequent Itemset Mining in Uncertain Databases*, Proc. 15th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD '09), pp. 119-127, 2009. [05]
[3] C.-K. Chui, B. Kao, and E. Hung, *Mining Frequent Item sets from Uncertain Data*, Proc. 11th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), pp. 47-58, 2007.
[4] C.-K. Chui, B. Kao, *A Decremental Approach for Mining Frequent Item sets from Uncertain Data*, Proc. 12th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), pp. 64-75, 2008.
[5] Y. Huang, H. Xiong, and S. Shekar, *Mining Confident Co-Location Rules without a Support Threshold*, Proc. ACM Symp. Applied Com-puting, pp. 497-501, 2003.
[6] Y. Huang, S. Shekar, and H. Xiong, Discovering Co-Location Patterns from Spatial Data Sets: A General Approach, *IEEE Trans. knowledge and Data Eng*., vol. 16, no. 12, pp. 1472-1485, Dec. 2004.
[7] Y. Huang, J. Pei, and H. Xiong, "Mining Co-Location Patterns with Rare Events from Spatial Data Sets," *Geoinformatics*, vol. 10, no. 3, pp. 239-260, Dec. 2006.
[8] Y. Morimoto, *Mining Frequent Neighboring Class Sets in Spatial Databases*, Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 353-358, 2001.
[9] J.S. Yoo, S. Shekar,J. Smith, and J.P. Kumquat, *A Partial Join Approach for Mining Co-Location Patterns*, Proc. 12th Ann. ACM Int'l Workshop Geographic Information Systems (GIS), pp. 241-249, 2004.
[10] J.S. Yoo and S. Shekar, A Join less Approach for Mining Spatial Co-Location Patterns, *IEEE Trans. knowledge and Data Eng.(TKDE)*, vol. 18, no. 10, pp. 1323-1337, Dec. 2006.
[11] L. Wang, Y. Bao, J. Lu and J. Yip, *A New Join-less Approach for Co-Location Pattern Mining*, Proc. IEEE Eighth ACM Int'l Conf. Computer and Information Technology (CIT), pp. 197-202, 2008.
[12] L. Wang, H. Chen, L. Zhao and L. Zhou, *Efficiently Mining Co-Location Rules of Interval Data*, Proc. Sixth Int'l Conf. Advanced Data Mining and Applications, pp. 477-488, 2010.
[13] Q. Zhang, F. Li, and K. Yi, *Finding Frequent Items in Probabilistic Data*, Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 819-832, 2008.
[14] Wang, P. Wu, and H. Chen, Finding Probabilistic Prevalent Colocations in Spatially Uncertain Data Sets, *IEEE Trans. knowledge and Data Eng.(TKDE)*, vol. 25, no. 4, pp. 790-804, Apr. 2013.