On The Relationships Between Clustering and Spatial Co-location Pattern Mining

Yan Huang
University of North Texas
[huangyan]@cs.unt.edu

Pusheng Zhang *
Microsoft Corporation
[pzhang]@microsoft.com

Abstract

The goal of spatial co-location pattern mining is to find subsets of spatial features frequently located together in spatial proximity. Example co-location patterns include services requested frequently and located together from mobile devices (e.g., PDAs and cellular phones) and symbiotic species in ecology (e.g., Nile crocodile and Egyptian plover). Spatial clustering groups similar spatial objects together. Reusing research results in clustering, e.g. algorithms and visualization techniques, by mapping co-location mining problem into a clustering problem would be very useful. However, directly clustering spatial objects from various spatial features may not yield well-defined co-location patterns. Clustering spatial objects in each layer followed by overlaying the layers of clusters may not applicable to many application domains where the spatial objects in some layers are not clustered.

In this paper, we propose a new approach to the problem of mining co-location patterns using clustering techniques. First, we propose a novel framework for co-location mining using clustering techniques. We show that the proximity of two spatial features can be captured by summarizing their spatial objects embedded in a continuous space via various techniques. We define the desired properties of proximity functions compared to similarity functions in clustering. Furthermore, we summarize the properties of a list of popular spatial statistical measures as the proximity functions. Finally, we show that clustering techniques can be applied to reveal the rich structure formed by co-located spatial features. A case study on real datasets shows that our method is effective for mining co-locations from large spatial datasets.

[1]This work was done when the second author was with the University of Minnesota.

1 Introduction

Advanced data collecting tools such as global positioning systems and earth observing systems are accumulating increasingly large spatial datasets. For example, NASA's Earth Observing System (EOS) has been producing more than a terabyte of data each day since 1999. Now more powerful, reliable, and inexpensive location-enabled mobile devices are generating large geo-referenced datasets. These spatial data are considered to contain nuggets of valuable information in the form of interesting and potentially useful patterns that were previously unknown. The automatic discovery of such patterns is being widely investigated with the use of spatial data mining [14, 15] techniques.

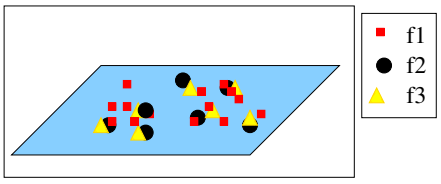


Figure 1. Illustration of Spatial Co-location Patterns. Symbols represent different spatial features. Spatial features • and △ tend to be located together.

Mining *spatial co-location patterns* [16, 13, 3] is an important spatial data mining task with broad applications. Spatial co-location patterns represent subsets of the spatial features whose objects are often located in close geographic proximity. Examples include symbiotic species, e.g., the Nile crocodile and Egyptian plover in ecology, frontage roads and highways in metropolitan road maps, and co-located services frequently requested together from mobile devices (e.g., PDAs and cellular phones) in location-based services.

In Figure 1, there are three types of spatial objects denoted by different symbols. As can be seen, objects of • and △ tend to be located together in space. A spatial feature,

Proceedings of the 18th IEEE International
Conference on Tools with Artificial Intelligence (ICTAI'06)
0-7695-2728-0/06 $20.00 © 2006 IEEE

e.g. West Nile disease, drought, and car accident, refers to the conceptual abstraction of a set of similar spatial objects. Each spatial feature has a set of objects in a spatial framework, e.g. there are many objects of West Nile disease across the world. A co-location is a set of spatial features whose objects tend to locate together in spatial proximity. A spatial framework is a mapped set of geographic regions that supports agency programs or studies [12]. Formally, the spatial co-location mining problem is defined as follows: We are given a database $\mathcal{D}$ of spatial objects in a spatial framework S. Let $\mathcal{F} = \{f_1, f_2, \ldots, f_K\}$ be a set of spatial features. Each spatial object has a spatial feature and consists of the following fields: *object-id, location,* and *spatial-feature*, where *spatial-feature* $\in \mathcal{F}$. We denote the set of objects O that are associated with an spatial feature $f \in \mathcal{F}$ as $f.O$. The problem of finding spatial co-location patterns is to find subset of spatial features whose objects tend to locate together in spatial proximity.

Clustering [10] divides a set of objects into several groups where objects in the same group are most similar to each other while objects from different groups are most dissimilar from each other. The first attempt of the spatial co-location mining efforts may be applying clustering techniques to help us discover co-locations, e.g., we may find mini-clusters and then analyze the resulting mini-clusters to find co-located spatial features. However, interpreting mini-cluster of spatial objects to derive patterns among spatial features is challenging.

In this paper, we investigate the relationships between clustering and spatial co-location mining. We propose a framework for mining co-locations based on spatial feature clustering. Rather than clustering at the spatial object level, the spatial feature clustering approach clusters spatial features directly. In order to use clustering techniques, we define proximity of two spatial features by summarizing their spatial objects embedded in a continuous spatial framework via various techniques. We investigate the properties of the proximity measures in light of desired properties of a similarity measure in clustering. Furthermore, we summarize the properties of a list of popular spatial statistical measures as proximity functions. Finally, we show that clustering techniques can be applied to reveal the rich structure formed by co-located spatial features. An experimental study on real datasets shows the effectiveness of the proposed framework on the discovery of spatial co-location patterns from spatial databases.

The goal of our investigation is to link co-location mining with clustering. Because clustering is a area with relatively fertile results, we are hoping that the work on clustering, e.g. visualization, fast algorithms, schemes that find different shaped clusters, can be reused by spatial co-location mining to reveal the rich structure formed by co-located spatial features.

The remainder of the paper is organized as follows. In Section 2, we review related work. Section 3 presents our proposed spatial feature clustering based co-location mining framework. An experiment study on real dataset is reported in Section 4. We summarize our work and discuss future directions in Section 5.

2 Related Work

Due to the close relationships between clustering and co-location mining, finding co-location patterns via clustering, the focus of this paper, has attracted much attention [7, 6]. Clustering have been mainly applied in two ways: layer based clustering and mixed clustering as illustrated in Figure 2 and Figure 3 respectively.

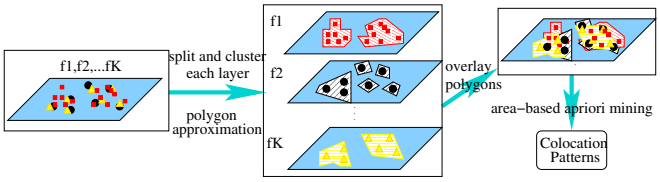


Figure 2. Layer Based Clustering Approach

A layer based clustering approach [7, 6] clusters each layer/spatial feature of spatial objects first, approximates each cluster by a polygon and then overlays the polygons from all layers together. The overlapping area of the polygons from all spatial features in a co-location pattern represents how co-located the corresponding spatial features are. Mining co-location patterns is equivalent to finding the subset of spatial features whose object clusters, approximated by polygons, overlap for large areas. Given X and Y as sets of layers, a clustered spatial co-location rule is defined as $X \Rightarrow Y(CS, CC\%)$, for $X \bigcap Y = \emptyset$, where CS is the clustered support, defined as the ratio of the area of the cluster (region) that satisfies both X and Y to the total area of the study region S, and $CC\%$ is the clustered confidence, which may be interpreted as $CC\%$ of areas of clusters (regions) of X intersect with areas of clusters(regions) of Y. This technique may not be applied to applications where one or more spatial feature layers are not clustered. Also note that there is a distinction between spatial co-location patterns and spatial co-location rules. Spatial co-location pattern refers to subsets of spatial features frequently located in spatial proximity while a spatial co-location rule is a conditional probability of observing one set of spatial features given that the other set of spatial features are already observed. We focus on spatial co-location patterns in this paper.

Mixed clustering approach clusters all objects regardless of their spatial features under a unified spatial framework. Since each cluster contains objects of different spa-

Proceedings of the 18th IEEE International
Conference on Tools with Artificial Intelligence (ICTAI'06)
0-7695-2728-0/06 $20.00 © 2006 IEEE

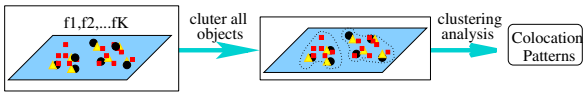


Figure 3. Mixed Clustering Approach

tial features, resulting clusters do not represent spatial co-locations. Association pattern mining [2] algorithms may be applied to derive co-location patterns among spatial features. Cluster sizes, overlapping property, and shapes are likely impact the quality of the mined co-location patterns.

Various attempts have also been made to model spatial co-location patterns without a clustering perspective. In spatial statistics, functions such as, *G nearest neighbor function, F nearest neighbor function, J function, Ripley's K function and its variations* exist to measure self or pairwise distance based point patterns [3]. Measures of correlation may not be naturally extended to measure co-locations among more than two spatial features. However, we may be able to use these measures to define similarity between spatial features and then apply clustering schemes. This paper is the first step in extending co-locations with more than two features using spatial statistical measures and investigating the related issues.

Association rule-based approaches to mining spatial co-location patterns may be divided into transaction-based approaches and distance-based approaches. Transaction based approaches focus on defining transactions over space so that an Apriori-like algorithm [2] can be used. Transactions over space may be defined by a reference-feature centric model [11]. Under this model, transactions are created around objects of one user-specified spatial feature. The association rules are derived using the Apriori [2] algorithm. The rules found are all related to the reference feature. A distance-based approach was proposed concurrently by Morimoto [13] and us [16]. Morimoto defined distance-based patterns called k-neighboring class sets. In his work, the number of occurrence for each pattern is used as the prevalence measure. Our event centric model [16] provides a transaction-free approach by using the concept of proximity neighborhood. We proposed to use a participation index and conditional probability to measure the interestingness of the co-location patterns.

3 Proposed Approach

In this section, we propose our clustering based framework for spatial co-location mining. The basic idea is that instead of clustering spatial objects we cluster spatial features. The proposed framework is as shown in Figure 4. First, each spatial feature is represented by its spatial layer. Then the similarity/proximity between pairwise spatial features are derived from their spatial layers and are represented by a proximity matrix. Finally, spatial features are clustered based on the proximity matrix. The major challenge here is to define a meaningful proximity function with desired properties.

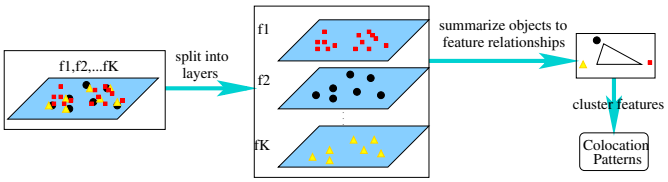


Figure 4. Feature Clustering Framework

3.1 Choosing Proximity Function For Pairwise Spatial Features

In clustering, a dissimilarity value based on a distance function defined in a set O is assigned to each pair of points in O to measure the distance between them. Example distance functions include Manhattan distance and Euclidean distance, which can be generalized to the Minkowski distance metric. The similarity between two objects is a numerical measure to quantify the degree to which the two objects are alike. The similarity measure must be higher when pairs of objects are more alike. Example similarity functions include cosine measure, Jaccard coefficient, and correlation.

However, for spatial co-location using clustering, the similarity function assumes the role of measuring how "co-located" the two spatial features are to each other. In other words, the similarity function becomes a *proximity function*. How to define such a proximity matrix? What are the desired properties of a proximity matrix?

In this section we propose a density ratio as the proximity function and investigate its properties. Proximity between a pair of spatial features can also be defined by various spatially meaningful measures based on their objects including *participation index [9], join selectivity, G nearest neighbor function, F nearest neighbor function, J function, Ripley's K function and its variations* [3]. We summarize their properties as a proximity function for co-location mining as well.

3.1.1 Density Ratio

We define *a density ratio* as the co-location function. Density is introduced first to facilitate the definition of density ratio.

Definition 1 (Density). *For a given spatial framework S, the density of an object set O in S is the average number of objects in O in each unit of S, i.e.* $density(O, S) = \frac{|\{o|o \in O \wedge o \in S\}|}{area(S)}$.

Proceedings of the 18th IEEE International
Conference on Tools with Artificial Intelligence (ICTAI'06)
0-7695-2728-0/06 $20.00 © 2006 IEEE

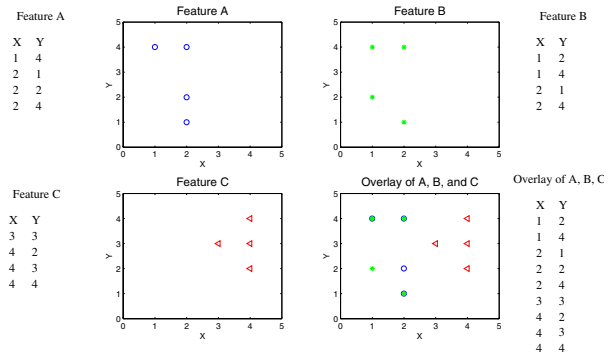


Figure 5. An Example Spatial Database

In Figure 5, there are four spatial features $\{A, B, C, D\}$ whose objects are represented by four spatial layers. The density of A in S is $density(A.O, S) = \frac{4}{25} = 0.16$, meaning there are about 0.16 objects of spatial feature A in each unit of the spatial framework S. And the density of B in S is $density(B.O, S) = \frac{4}{25} = 0.16$ as well.

Given two object sets O and O', if the average density of O' in the neighborhoods of objects in O is higher than the density of O' in the overall embedding space, then it is likely that objects in O' tend to be co-located with objects in O.

Definition 2 (Density Ratio). *For two object sets O and O', and a given neighborhood function $N(o)$ over an object o in the object set O, the density ratio of O' around O is defined as: $densityRatio(O, O') = \frac{average_{o \in O}(density(O', N(o)))}{density(O', S)}$ where S is the spatial embedding space.*

The numerator represents the average density of objects in set O' that are co-locate with objects in set O. The denominator of the ratio is the base density of O' objects in the embedding space. In other words, $densityRatio(O, O')$ is the average density of objects in O' in the neighborhood of an object in O over the base density of O' in the embedding space. Based on the definition of density ratio, the density ratio between the object sets of two spatial features f_1 and f_2 is $densityRatio(f, f') = densityRatio(f.O, f'.O) = \frac{average_{o \in f.O}(density(f'.O, N(o)))}{density(f'.O, S)}$. In Figure 5, the density ratio of spatial features A and B is $densityRatio(A, B) = \frac{average_{o \in A.O}(density(B.O, N(o)))}{density(B.O, S)} = \frac{\frac{2}{\pi\sqrt{2}^2} + \frac{2}{\pi\sqrt{2}^2} + \frac{2}{\pi\sqrt{2}^2} + \frac{2}{\pi\sqrt{2}^2}}{4} / 0.16$, assuming the neighborhood function $N(o)$ for an object o is an Euclidean distance no greater than $\sqrt{2}$.

If the density of objects of spatial feature f' in the neighborhood of an object of spatial feature f is higher than f''s overall density, the density ratio will be greater than 1 and spatial feature f' is said to be co-located with spatial feature f. On the contrary, if the density of objects of spatial feature f' in the neighborhood of an object of spatial feature f is lower than f''s overall density, then the density ratio will be less than 1 and spatial feature f is repelling spatial feature f'. Otherwise, density ratio is close to 1 and spatial features f and f' are independently distributed.

3.1.2 Properties of Density Ratio

The desired properties for a similarity function sim in clustering are: (i) non-negativity: $sim(x, y) \geq 0, \forall x, y \in O$, (ii)commutativity: $sim(x, y) = sim(y, x), \forall x, y \in O$, and (iii) reflexivity: $sim(x, y) = m \Leftrightarrow x = y$ where m is the largest possible value of the similarity function [17]. This means that x is most similar to itself and if two objects are most similar in the object set, they are the same object. Our goal is to use density ratio or some similar measures (will be discussed in section 3.2) to define the proximity of two spatial features and obtain a proximity matrix for the clustering. We will look at the properties of the density ratio in light of the designed properties of a similarity function in clustering.

Property 1 (Non-negative). *Density ratio is non-negative.*

By definition, the density ratio is non-negative. The value of a density ratio is a positive number with 1 indicating complete independent distribution of the two involved spatial features. The larger the number, the stronger the co-location is between the two spatial features. The smaller the number, the stronger the repulsion is.

Property 2 (Commutative). *For a symmetric neighborhood function, density ratio is commutative.*

Proof. We need to prove for a symmetric neighborhood function $N(o)$ and any two spatial features f and f', $densityRatio(f.O, f'.O) = densityRatio(f'.O, f.O)$. Let S be the spatial embedding space. Let $number(O, S')$ denote the number of objects from an object set O in a space S'. So, $density(O, S') = \frac{number(O, S')}{area(S')}$. Now,

$$
\begin{aligned}
&densityRatio(f.O, f'.O) \\
&= \frac{average_{o \in f.O}(density(f'.O, N(o)))}{density(f'.O, S)} \\
&= \frac{average_{o \in f.O}(\frac{number(f'.O, N(o))}{area(N(o))})}{\frac{number(f'.O, S)}{area(S)}} \\
&= \frac{\sum_{o \in f.O}(\frac{number(f'.O, N(o))}{area(N(o))}) / number(f.O, S)}{\frac{number(f'.O, S)}{area(S)}}
\end{aligned}
\tag{1}
$$

Because $\sum_{o \in f.O}(number(f'.O, N(o)))$ is number of object pairs $|pairs_{N(o)}(f.O, f'.O)|$, with one from $f'.O$ and one from $f.O$, that are within neighborhood defined by the symmetric neighborhood function $N(o)$, $\sum_{o \in f.O}(number(f'.O, N(o))) =$

$|pairs_{N(o)}(f.O, f'.O)| = \sum_{o \in f'.O}(number(f.O, N(o)))$. Further more, for any two objects $o \in f.O$ and $o' \in f'.O$, $area(N(o)) = area(N(o'))$ for a symmetric neighborhood function. So, we have

$$
\begin{aligned}
&\frac{\sum_{o \in f.O}(\frac{number(f'.O,N(o))}{area(N(o))})/number(f.O,S)}{\frac{number(f'.O,S)}{area(S)}} \\
=&\frac{\frac{|pairs_{N(o)}(f.O,f'.O)|}{area(N(o))})/number(f.O,S)}{\frac{number(f'.O,S)}{area(S)}} \\
=&\frac{\sum_{o \in f'.O}(\frac{number(f.O,N(o))}{area(N(o))})/number(f.O,S)}{\frac{number(f'.O,S)}{area(S)}} \\
=&\frac{average_{o \in f'.O}(\frac{number(f'.O,N(o))}{area(N(o'))})}{\frac{number(f.O,S)}{area(S)}} \\
=&\frac{average_{o \in f'.O}(density(f.O,N(o)))}{density(f.O,S)} \\
=&densityRatio(f'.O, f.O)
\end{aligned}
\tag{2}
$$

□

The commutative property is important because most clustering algorithms only worked on symmetric similarity matrices. Next we look at the value range of a density ratio. The smallest value for a density ratio $densityRatio(f, f')$ is 0, which happens when no objects of spatial feature f' is in the neighborhood of an object of spatial feature f.

Property 3 (Maximum Value). *For two spatial features f and f', and a given neighborhood function $N(o)$ over an object o in $f.O$, the maximum value of the density ratio $densityRatio(f.O, f'.O)$ is $\frac{area(S)}{area(N(o))}$.*

Proof. From the proof of Property 2, we have $densityRatio(f.O, f'.O) = \frac{\frac{|pairs_{N(o)}(f.O,f'.O)|}{area(N(o))})/number(f.O,S)}{\frac{number(f'.O,S)}{area(S)}}$. The maximum value for $|pairs_{N(o)}(f.O, f'.O)|$ is $number(f.O, S) \times number(f'.O, S)$. This happens when every object o of spatial feature f is in the neighborhood $N(o)$ of every object of spatial feature f' (objects $f.O$ and objects in $f'.O$ form a complete bipartite graph under the neighborhood function $N(o)$). So, the maximum value of the density ratio $densityRatio(f.O, f'.O)$ is $\frac{area(S)}{area(N(o))}$. □

Property 4 (Non-reflexive). *Density ratio is not reflexive.*

Proof. We prove by a counter example that the statement: $f = f' \Rightarrow densityRatio(f.O, f'.O) = m$ is false, where m is the largest possible value of the density ratio using Figure 5. In Figure 5, assuming the neighborhood function $N(o)$ for an object o is Euclidean distance no greater than $\sqrt{2}$, the density ratio of spatial feature A and A itself is $densityRatio(A.O, A.O) = \frac{average_{o \in A.O}(density(A.O,N(o)))}{density(A.O,S)} = \frac{\frac{1}{\pi\sqrt{2}^2}+\frac{1}{\pi\sqrt{2}^2}+\frac{1}{\pi\sqrt{2}^2}+\frac{1}{\pi\sqrt{2}^2}}{\frac{4}{0.16}} = \frac{25}{4\pi\sqrt{2}^2}$. From Property 3, we know that the maximum possible value for density ratio in this example is $m = \frac{area(S)}{area(N(o))} = \frac{25}{\pi\sqrt{2}^2}$. So, $densityRatio(A.O, A.O) \neq m$.

We can also prove that if $densityRatio(f.O, f'.O) = m$, f may not equal to f'. Again, we prove by an counter example. In Figure 5, assuming the neighborhood function $N'(o)$ for an object o is Euclidean distance no greater than $\sqrt{10}$, the density ratio of spatial features A and B is $densityRatio(A.O, B.O) = \frac{average_{o \in A.O}(density(B.O,N'(o)))}{density(B.O,S)} = \frac{\frac{4}{\pi\sqrt{2}^2}+\frac{4}{\pi\sqrt{2}^2}+\frac{4}{\pi\sqrt{4}^2}+\frac{4}{\pi\sqrt{4}^2}}{\frac{4}{0.16}} = \frac{25}{\pi\sqrt{10}^2}$. From Property 3, we know that the maximum possible value for density ratio in this example is $m = \frac{area(S)}{area(N'(o))} = \frac{25}{\pi\sqrt{10}^2}$. So,$densityRatio(A.O, B.O) = m$ although $A \neq B$. □

Reflexivity means an object is more similar to itself than to any other object and if two objects are the same, then their similarity is the maximum. Reflexivity is not a desired property for a proximity function of two spatial features because the proximity of a feature f with itself measures how clustered the objects of the spatial feature is, i.e. self proximity but not self similarity.

3.2 Other Proximity Functions and Their Properties

We illustrated density ratio as a proximity function for clustering based co-location mining and proved its properties. Other measures in spatial statistics may also be used to define a proximity function. Table 1 summarizes the definitions of these measures and their properties (we omit the proves for the lack of space). These measures are for self or pairwise point patterns with a distance predicate.

3.3 Co-location and Clustering

Once we have the proximity matrix based on the proximity function defined, we can apply clustering algorithm on spatial features. Clustering techniques may be broadly classified into partition based, hierarchical, density-based, grid-based and model-based schemes [8]. Resulting clusters may be globular, such as those found by partition-based K-means or complete link hierarchical clustering algorithms or arbitrary shaped, such as those found by

Measure	Definition	Non-negativity	Commutativity	Reflexivity	Range
$G_{ij}(d)$	the proportion of objects of a spatial feature i for which the distance to the nearest object of the spatial feature j is less than or equal to d.	Yes	No	No	$[0, 1]$
$F_i(d)$	the proportion of all the objects of spatial feature i for which the distance to the nearest object of itself is less than or equal to d.	N/A	N/A	N/A	$[0, 1]$
$J_{ij}(d)$	$J_{ij}(d) = (1 - G_{ij}(d))/(1 - F_j(d))$	Yes	No	No	$[0, \infty)$
$K_{ij}(d)$	average number of objects of spatial feature i within distance d of a randomly chosen object of spatial feature j divided by the number of objects of spatial feature j in unit area	Yes	Yes	No	$[0, area(S)]$
$densityRatio_{ij}(d)$	density of objects of spatial feature i within distance d of an object of spatial feature j divided by the density of objects of spatial feature i in S	Yes	Yes	No	$[0, \frac{area(S)}{\pi d^2}]$

Table 1. Proximity Measures for Spatial Features i and j in a Spatial Framework S

density-based DBScan [5] or single link hierarchical clustering algorithms. Also clusters may be overlapping, such as those found by the model-based EM algorithm or non-overlapping, such as those found by most partition based and hierarchical clustering algorithms.

We have showed that the major difference between a proximity function in co-location mining and a similarity function in clustering is that the reflexivity is not desired in a proximity function. Thus, the clusters found by different clustering schemes may be interpreted by combining the meaning of the clusters with the meaning of the proximity function used. For example, a globular cluster found using the K-means algorithm may be interpreted as a set of spatial features with strong tendency to occur as a group in proximity, and an arbitrary shaped cluster found using DBScan may be interpreted as a set of spatial features which forms a chain of interactions represented by the shape of the cluster.

4 Experiment and Results of a Case Study

We use a case study to investigate the co-locations found by the proposed clustering-based framework to show its usefulness. In the experiment, we explore co-locations of spatial features using the proposed clustering-based framework on a dataset available at the Digital Chart of the World Data Server[18]. The experiment design includes three components as shown in Figure 6: the generation of the map layers for spatial features, the construction of a proximity matrix for the spatial features, and clustering of the spatial features based on the proximity matrix. Each cluster represent a set of co-located features.

The dataset includes 12 spatial features in Texas, which emphasizes landmarks important from flying altitudes as shown in Table 2. The spatial coordinates, such as longitude and latitude, were extracted from the raw data for each spatial feature in Texas using the ArcGIS Toolkit [4]. The spatial distribution of objects for each spatial feature corresponds to a map layer. For example, the map layers for dense population and all airports in Texas are shown in Figure 7. The investigation of the co-locations of these spatial features are of critical importance in this application to identify subsets of spatial features with significant spatial interactions.

SF ID	Spatial Features	Description
1	Dense Population	Densely populated places
2	High Drainage	Perennial inland water
3	Drainage Supplement	Nonperennial inland water
4	Hypsography	Locations with precise spot elevations
5	Low Altitude	Low altitude points
6	Mines	Locations of mines
7	Aeronautical Civil	Civil airport locations
8	Cultural Landmark	Cultural landmark places
9	High Altitude	High altitude locations
10	All Airports	All airport locations
11	Water Rights	Surface water allowed to be used by the public
12	Outfalls	Outlet of a body of water

Table 2. Spatial Features in the Experimental Data.

The pairwise density ratios values were calculated and

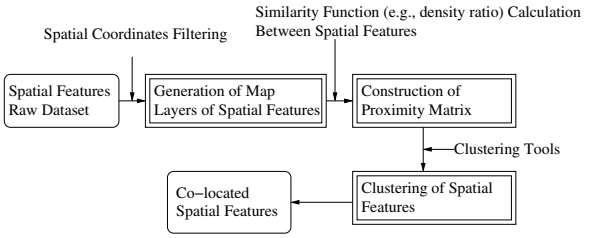


Figure 6. Feature Clustering Framework

Proceedings of the 18th IEEE International
Conference on Tools with Artificial Intelligence (ICTAI'06)
0-7695-2728-0/06 $20.00 © 2006 IEEE

visualized in Figure 8 (a). The proximity matrix was constructed for spatial features based on the pairwise density ratio values. In the proximity matrix, each row/column of the matrix represents a spatial feature and the value at the (*i*,*j*) entry of the matrix indicates the similarity between the *i*th and *j*th features with respect to the degree of their proximity. Off-the-shelf clustering tools can be applied to group co-located spatial features based on the proximity matrix.

In hierarchical clustering, it is not required to specify the number of clusters; therefore, it would be ideal for exploratory data analysis on large spatial data, e.g., the discovery of co-located spatial features. In the experiment, three hierarchical clustering algorithms [17], namely *single link*, *complete link*, and *group average*, were used as the clustering algorithms for spatial co-location mining[3]. The clustering results are visualized using dendrograms. A dendrogram consists of many U-shaped lines connecting spatial features in a hierarchical tree structure. The use of dendrograms to illustrate clustering results preserves the sequence of constructions of clusters with regard to the proximity.

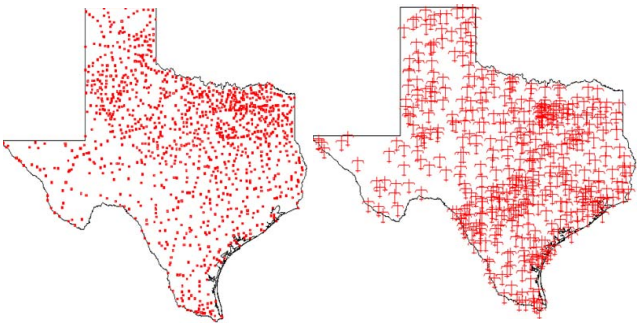


Figure 7. Examples of Map Layers for Spatial Features in Texas.(a)Dense Population.(b)All Airports.

The three clustering algorithms use different ways to define distances between clusters. Thus the co-location patterns using these algorithms may find different types of co-locations. The single link clustering algorithm takes the shortest distance from any member of one cluster to any member of the other cluster as the distance between one cluster and another cluster. Therefore, it finds the similarity between one cluster and another cluster to be equal to the greatest proximity of any feature of one cluster to any feature of the other cluster. The clustering results using the single link algorithm are illustrated in Figure 8 (a). As shown in Figure 8 (b), there are three significant co-location patterns: c_1 = {all airports (feature 10), civil airports (feature 7)}, c_2 = {all airports (feature 10), civil airports (feature 7), mines (feature 6)}, and c_3 = {low altitude points (feature 5), outfalls (feature 12)}. The first one seems obvious due to high overlays in the map layers. The second pattern means that there are airports near the mines. The third one signifies that most outfalls, e.g., the mouth of a river, have low altitudes, and this co-location pattern confirms the domain knowledge in hydrology.

The complete link clustering algorithm considers the distance between one cluster and another cluster to be equal to the longest distance from any member of one cluster to any member of the other cluster. The clustering results using the complete link algorithm are illustrated in Figure 8 (c). The co-location patterns c_1 and c_3 are also detected using the complete link algorithm. The co-location pattern c_2 found by the single link is not significant. As can be seen in Figure 8 (c), there are two additional co-location patterns, c_4 = {high drainage (feature 2), water rights (feature 11)} and c_5 = { mines (feature 6), high altitude (feature 9) }. The co-location pattern c_4 signifies that inland water mostly belongs to or is close to the surface water which is allowed to be used by the public. The interpretation of the co-location pattern c_5 needs to be further investigated by local geographers.

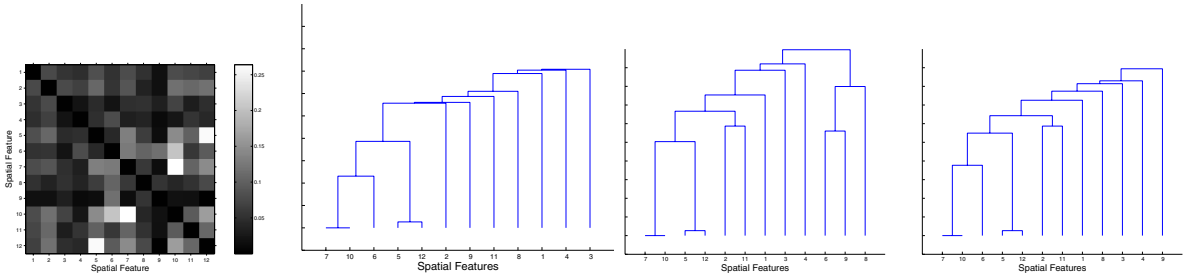
The group average clustering algorithm considers the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster. The clustering results using the group average algorithm are illustrated in Figure 8 (d). As can be seen, the co-location patterns c_1, c_2, c_3, and c_4 are found using the group average clustering algorithm.

In summary, the experiment shows that the proposed framework using clustering techniques is useful for mining spatial co-location patterns. The significant spatial interactions between spatial features can be identified for further investigation by application domain experts. Different clustering algorithms may generate different types of spatial co-location patterns, and dendrograms can be used in exploratory data analysis to support interactive explorations of co-located spatial features.

5 Conclusion and Future Work

In this paper, we proposed a novel framework for co-location mining using clustering techniques. We showed that the proximity of two spatial features can be captured by summarizing their spatial objects embedded in a continuous space via various techniques. We also discussed the desired properties of a proximity function. Clustering tech-

[3]Hierarchical clustering tools in Matlab 6.5 [1], including single link, complete link, and group average, were used to cluster spatial features in this experiment. Due to the input format for hierarchical clustering tools in Matlab, the proximity matrix was normalized between 0 and 1 and transformed into a distance matrix (I - normalized proximity matrix). The distance matrix was used as the input into the Matlab hierarchical clustering tools.



(a) Proximity Matrix (b) Single Link (c) Complete Link (d) Group Average

Figure 8. Proximity Matrix and Dendrograms Using Different Clustering Algorithms.

Co-location ID	Co-location Pattern
c_1	{all airports (feature 10), civil airports (feature 7)}
c_2	{all airports (feature 10), civil airports (feature 7), mines (feature 6)}
c_3	{low altitude points (feature 5), outfalls (feature 12)}
c_4	{high drainage (feature 2), water rights (feature 11)}
c_5	{ mines (feature 6), high altitude (feature 9) }

Table 3. Co-location Patterns in Experimental Data.

niques can be applied to reveal the rich structure formed by co-located spatial features in spatial data. An experimental study on a real dataset shows that our proposed framework is effective for mining co-location patterns to identify the subsets of spatial features with significant spatial interactions.

In future work, we would like to investigate further how different clustering algorithms affect the results. We will also explore effective visualization techniques to assist the interactive discovery of spatial co-location patterns from large datasets. Furthermore, we plan to collaborate with domain experts to further investigate the spatial co-location patterns found in our experiment.

References

[1] The Mathworks Inc. Hierachical Clustering Tools in Matlab 6.5. http://www.mathworks.com/.

[2] R. Agarwal and R. Srikant. Fast Algorithms for Mining Association Rules. In *Proc. of the 20th Int'l Conference on Very Large Data Bases*, 1994.

[3] N. Cressie. *Statistics for Spatial Data*. Wiley and Sons, ISBN:0471843369, 1991.

[4] Environmental Systems Research Institute, Inc. ArcGIS Family. http://www.esri.com.

[5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *the Proc. of Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[6] V. Estivill-Castro and I. Lee. Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-referenced Crime Data. In *Proc. of the 6th International Conference on Geocomputation*, 2001.

[7] V. Estivill-Castro and A. Murray. Discovering Associations in Spatial Data - An Efficient Medoid Based Approach. In *Proc. of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1998.

[8] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

[9] Y. Huang, S. Shekhar, and H. Xiong. Discovering co-location patterns from spatial datasets: A general approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(12), to appear.

[10] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[11] K. Koperski and J. Han. Discovery of Spatial Association Rules in Geographic Information Databases. In *Proc. of the 4th International Symposium on Spatial Databases*, 1995.

[12] McMahon. Common Ecoregions Map . *Environmental Management*, 28, 2001.

[13] Y. Morimoto. Mining Frequent Neighboring Class Sets in Spatial Databases. In *Proc. ACM SIGKDD*, 2001.

[14] J. Roddick and M. Spiliopoulou. A Bibliography of Temporal, Spatial and Spatio-temporal Data Mining Research. *ACM Special nterest Group on Knowledge Discovery in Data Mining Explorations*, 1999.

[15] S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, ISBN: 0130174807, 2003.

[16] S. Shekhar and Y. Huang. Co-location Rules Mining: A Summary of Results. In *Proc. 7th Intl. Symposium on Spatio-temporal Databases*, 2001.

[17] P. Tan, M. Steinbach, and V. Kumar. Introduction to data mining. In preparation for publication.

[18] The Pennsylvania State University Libraries. Digital Chart of the World Data Server. http://www.maproom.psu.edu/dcw/.

Proceedings of the 18th IEEE International
Conference on Tools with Artificial Intelligence (ICTAI'06)
0-7695-2728-0/06 $20.00 © 2006 IEEE