Y. Jayababu*, G.P.S. Varma and A. Govardhan

Mining Spatial Association Rules to Automatic Grouping of Spatial Data Objects Using Multiple Kernel-Based Probabilistic Clustering

DOI 10.1515/jisys-2016-0044
Received April 16, 2016; previously published online August 6, 2016.

Abstract: With the extensive application of spatial databases to various fields ranging from remote sensing to geographical information systems, computer cartography, environmental assessment, and planning, discovery of interesting and hidden knowledge in the spatial databases is a considerable chore for classifying and using the spatial data and knowledge bases. The literature presents different spatial data mining methods to mine knowledge from spatial databases. In this paper, spatial association rules are mined to automatic grouping of spatial data objects using a candidate generation process with three constraint measures, such as support, confidence, and lift. Then, the proposed multiple kernel-based probabilistic clustering is applied to the associate vector to further group the spatial data objects. Here, membership probability based on probabilistic distance is used with multiple kernels, where exponential and tangential kernel functions are utilized. The performance of the proposed method is analyzed with three data sets of three different geometry types using the number of rules and clustering accuracy. From the experimentation, the results proved that the proposed multi-kernel probabilistic clustering algorithm achieved better accuracy as compared with the existing probabilistic clustering.

Keywords: Spatial database, association rule mining, clustering, probabilistic clustering, kernel.

1 Introduction

It is critical to develop tools for the discovery of interesting knowledge from large spatial databases because huge amounts of spatial data are gathered by remote sensing, e-commerce, and other data collection tools. So, it is necessary to develop an automated knowledge discovery scheme, leading to a hopeful field called data mining or knowledge discovery in databases (KDD) [10, 15, 20]. KDD is defined as the significant retrieval of hidden, uncertain, and possibly useful information from data [3, 7]. Spatial data mining [4, 5, 8, 12, 18] is classified depending on the types of rules to be discovered in spatial databases. A spatial association rule will be of the form $X \rightarrow Y$, where X and Y indicate the predicate sets and a few of them are spatial. Several association relationships may present in a large database, but some will occur infrequently. The idea of minimum support and minimum confidence are presented to concentrate on the frequently occurring patterns. The support of pattern A in a set of spatial objects S is the probability that a member of S satisfies pattern A and the confidence of A → B is the probability that pattern B occurs if pattern A occurs. The user can denote the thresholds to enclose the strong rules to be discovered [11].

***Corresponding author: Y. Jayababu,** Department of CSE, Pragati Engineering College, Surampalem, Andhra Pradesh 533437, India, e-mail: jayababuy2015@gmail.com
G.P.S. Varma: SRKR Engineering College, Andhra Pradesh, India
A. Govardhan: University College of Engineering, Jawaharlal Nehru Technological University, Hyderabad, India

Rules mined from the spatial database can then be used for a large number of applications such as clustering, classification, feature weighting, and indexing for retrieval improvement. Spatial clustering [9, 17] is a subset of clustering [21]. Tobler [19] stated the first law of geography: “Everything is related to everything else, but near things are more related than distant things.” Spatial clustering is a key technique for spatial data mining and spatial data analysis. The goal of spatial clustering is to partition spatial data into a series of meaningful subclasses, called spatial clusters. The spatial objects in the same cluster are related to each other, and in different clusters, the spatial objects are dissimilar. Existing spatial clustering techniques can be classified into two categories. One category is obtained from the spatial point pattern statistical analysis field, such as spatial autocorrelation analysis and spatial scan techniques. The second category is developed in spatial data mining field using partitioning algorithms and hierarchical algorithms [13].

When the algorithms are applied in geospatial data, two aspects are usually considered. One is the efficiency because the geospatial data are usually large. The other one is the adjustment on the algorithm to adapt in real applications [6, 16, 22], such as geographical information systems (GISs) and so on. GISs support a broad range of spatial queries that are used to support location-based studies. Such systems are intended to store, retrieve, control, analyze, and map geographical data [6]. In this paper, spatial association rules are mined to automatic grouping of spatial data objects using multiple kernel-based probabilistic clustering (PC). The input of the proposed system is a spatial database, which is in the vector format (shape file), that contains the geometric location with related attribute information. The spatial database is read out and passed through a spatial association rule mining algorithm, which can mine the important spatial association rules using support and confidence measures. The mined spatial association rules are then processed to form an associate vector to further group the spatial data objects. Grouping of spatial data objects is done using the proposed multiple kernel-based PC algorithm (MKPCA), which is the hybridization of multiple kernels with PC.

The organization of the paper is as follows. The problem definition and contributions of the paper is presented in Section 2. The proposed method of mining of spatial association rules to automatic grouping of spatial data objects using multiple kernel-based PC is presented in Section 3. In Section 4, the experimental results and the comparative analysis with the existing methods are presented. Finally, the conclusion is presented in Section 5.

2 Literature Review and Problem Statement

The literature presents various techniques for spatial association rules mining and clustering. Here, we present a review of different works. Koperski and Han [11] have proposed an efficient method for mining strong spatial association rules in geographic information databases. A spatial association rule is a rule indicating certain association relationships among a set of spatial and possibly some non-spatial predicates. A strong rule indicates that the patterns in the rule have relatively frequent occurrences in the database and strong implication relationships. Several optimization techniques were explored, including a two-step spatial computation technique (approximate computation on large sets and refined computations on small promising patterns), shared processing in the derivation of large predicates at multiple concept levels, etc. This work suffered with incorporating multiple concepts into the mining algorithm without much computational overhead. Clementini et al. [3] have used objects with broad boundaries, the concept that absorbs all the uncertainty by which spatial data were commonly affected and allow computations in the presence of uncertainty without rough simplifications of the reality. The topological relations between objects with a broad boundary can be organized into a three-level concept hierarchy. The progressive refinement approach was used for the optimization of the mining process. Even though the rule mining process utilizes the optimization algorithm, the mining for accurate spatial rules is completely missed due to the random initialization.

Shyu et al. [18] have customized the data mining algorithms using visual content and potential objects extracted from geospatial image databases with other relevant information, such as text-based annotations. Queries utilizing the mining results were also discussed in this paper. These mining and query processing algorithms play an important role in GeoIRIS (Geospatial Information Retrieval and Indexing System). The query processing for the multiple concept and topological relations pose manual preparation for the rule mining processes. Laube et al. [12] have investigated the support and confidence measures for spatial and spatio-temporal data mining. Using fixed thresholds to determine how many times a rule that uses proximity is satisfied seems too limited. It allowed the traditional definitions of support and confidence but does not allow making the support stronger if the situation is "really close," as compared to "fairly close." The traditional measure of support and confidence are not suitable to mine the spatial rules if they considered the topological relations.

Ding et al. [5] have proposed an efficient approach to derive association rules from spatial data using a Peano count tree (P-tree) structure. The P-tree structure provided a lossless and compressed representation of spatial data. Based on P-trees, an efficient association rule mining algorithm PARM with fast support calculation and significant pruning techniques was introduced to improve the efficiency of the rule mining process. The P-tree-based association rule mining (PARM) algorithm was implemented and compared with the FP-growth and Apriori algorithms. Even though the tree-based mining algorithm is more effective than Apriori, the memory requirement to store the tree structure is high as compared with the candidate-based methods. Dao and Thill [4] have proposed a comprehensive framework and library of algorithms of spatial analysis and visual analytics to resolve this fundamental challenge. The framework was the first attempt in delivering a complete geospatial knowledge discovery framework using spatial association rule mining. The spatial analytics do not consider the diversity constraints and automatic thresholding to group the data objects. However, the requirements of significant rules and distance function are missing in this paper.

Zalik and Zalik [21] have presented an agglomerative hierarchical clustering algorithm for spatial data. It discovered clusters of arbitrary shapes that may be nested. The algorithm uses a sweeping approach consisting of three phases: sorting is done during the preprocessing phase, determination of clusters is performed during the sweeping phase, and clusters are adjusted during the postprocessing phase. The properties of the algorithm were demonstrated by examples. The algorithm was also adapted to the streaming algorithm for clustering large spatial data sets. It proved poor in offering high-quality clustering solutions, in capability of discovering concave/deeper and convex/higher regions, and in robustness to outliers and noise. Pilevar and Sukumar [17] have proposed a clustering method, GCHL (a Grid-Clustering algorithm for High-dimensional very Large spatial databases), by combining density-grid-based clustering with an axis-parallel partitioning strategy to identify areas of high density in the input data space. The algorithm worked well in the feature space of any data set. The method operated on a limited memory buffer and required at most a single scan through the data. However, this algorithm did not consider the kernel space for clustering the data objects, and it also utilized the traditional distance measurement for clustering.

Problem Definition:

- Let a spatial database, SDB, S_i; $0 \leq i \leq$ n, including "n" spatial objects with different attribute values. The first challenge of finding useful information is formulated as a searching problem that two series of thresholds should filter out the most useful information that can be discovered from the spatial database.
- The second challenge is to apply the association rules to perform spatial clustering in SDB, consisting of n spatial data points located in p-dimensional real space of $S_i \in \mathbf{R}^p$. The challenge here is to find the k cluster centers from n spatial data points to divide those points into k clusters.
- The third challenge taken here is to find the distance values among data points and cluster centers. In PC, Euclidean distance is used to find the distance matrix to compute the probability of data assignment. However, further improvement is possible when the clustering algorithm is considered kernel space to find the probability assignment. So, redefining of distance measure is indispensable to perform the improved spatial clustering.

Contributions of the paper: The above challenges are dealt with the following contributions made in the research:

- We have developed a spatial association rule mining method that is suitable for three different types of geometries, such as point, line, and polygon, through a candidate generation process with three different constraint measures.

- Spatial association rules are combined with spatial attributes to perform the clustering process where the multiple kernels are utilized to find the distance measurement among the cluster centers and data space. Here, multiple kernel space is computed for finding the distance matrix to obtain the probability of belonging to spatial objects through cluster centers. Through this, we developed a new clustering algorithm called MKPCA.

3 Proposed Method of Mining of Spatial Association Rules to Automatic Grouping of Spatial Data Objects Using Multiple Kernel-Based PC

This paper presents a methodology for spatial association rule mining and the subsequent process of spatial data clustering. The input of the proposed system is the spatial database that is in the vector format. The spatial database is read out and passed through the spatial association rule mining algorithm, which can mine the important spatial association rules using support and confidence measures. The mined spatial association rules are then processed to form an associate vector to further group the spatial data objects. Grouping of spatial data objects is done using the proposed MKPCA, which is the hybridization of multiple kernels with PC. The block diagram of the proposed methodology is shown in Figure 1.

3.1 Data Preprocessing

The spatial data taken as input for the system, SDB, S_i; $0 \le i \le n$, is stored in shapefile format, which is the digital vector format to store the geometric location with related attribute information. The geometry can be a point, line, or polygon, usually. Every spatial data object, $S_i = \{L_x, L_y, A_1, A_2, A_3 \dots A_m\}$, is in the form of these geometries and its associated coordinate information, which is in the form of X and Y coordinates. Every spatial data object will have attribute information based on the characteristic of the data. After reading the shapefile, this information is directly converted to a transaction form that is easily minable through the rule

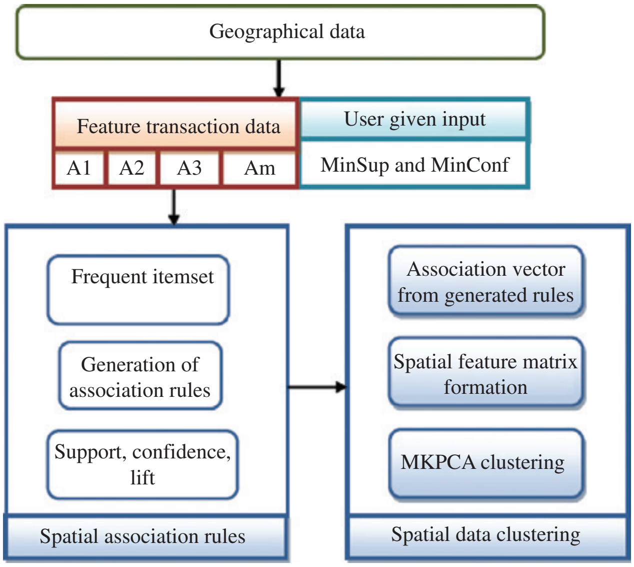

Geographical data
Feature transaction data
A1
A2
A3
Am
User given input
MinSup and MinConf
Frequent itemset
Generation of association rules
Support, confidence, lift
Spatial association rules
Association vector from generated rules
Spatial feature matrix formation
MKPCA clustering
Spatial data clustering

Figure 1: Block Diagram.

mining algorithm. The transaction database contains a set of features that are obtained through the unique attribute value from every attribute. Every unique attribute value is considered as feature value, which is the column value of the transaction database. Every row is a spatial data object and the values in the transaction data represent the presence of feature values in the corresponding spatial data objects.

3.2 Mining of Spatial Association Rules

The transaction form of database is given as input to the association rule mining algorithm, called the Apriori algorithm [1]. The two important phases of association rule mining are frequent itemset generation and association rule generation. At first, every feature value that has higher frequency among spatial data objects is considered a frequent itemset. Here, the constraint measure, called support, is utilized to filer out the frequent combination of feature values. In the second step, association rules are generated from the frequent itemset based on the constraint measure, called confidence.

The standard spatial association rule has the form $AR = a_1 \cap a_2 \cap a_m \Rightarrow b_1 \cap b_2 \cap b_m$, where a_1, a_2, a_m (antecedent) and b_1, b_2, b_m (consequent) are subsets of feature values in spatial database. The rule AR is given as "if a_1, a_2, a_m exist, then it is likely that b_1, b_2, b_m also co-exist in most of the spatial data objects." The rule AR maintains in the spatial database SDB with support S and confidence C, if S% of spatial data objects in SDB contains the attribute value a_1, a_2, a_m and b_1, b_2, b_m, and C% of the spatial data objects that contain a_1, a_2, a_m also contain b_1, b_2, b_m.

$$\mathrm{Supp}(a_1 \cap a_2 \cap a_m) = \frac{\text{Number spatial objects satisfy } a_1 \cap a_2 \cap a_m}{\text{Total number of spatial objects in SDB}}, \tag{1}$$

$$\mathrm{conf}(a_1 \cap a_2 \cap a_m \Rightarrow b_1 \cap b_2 \cap b_m) = \frac{\sup(a_1 \cap a_2 \cap a_m \cap b_1 \cap b_2 \cap b_m)}{\sup(a_1 \cap a_2 \cap a_m)}. \tag{2}$$

In addition to support and confidence, other measures like lift are also utilized to find the importance level of association rules. This measure is given as follows:

$$\mathrm{lift}(a_1 \cap a_2 \cap a_m \Rightarrow b_1 \cap b_2 \cap b_m) = \frac{\sup(a_1 \cap a_2 \cap a_m \cap b_1 \cap b_2 \cap b_m)}{\sup(a_1 \cap a_2 \cap a_m) \times \sup(b_1 \cap b_2 \cap b_m)}. \tag{3}$$

3.3 Construction of Spatial Feature Matrix

A spatial feature matrix, SF, is an intermediate data space generated for clustering algorithm with the location information and the associate vector generated from the association rules. The total attributes to be presented in the spatial feature matrix are location information (X and Y coordinates) and associate vector, AV. The associate vector is generated by finding the attribute value of spatial data objects presented in the association rules. A spatial data object, S_i, can obtain the value in the associate vector field only if mined association rules contain the corresponding attribute feature value. For example, the attribute value of a spatial data object is filled with association rules and the value in the associate vector is filled out. The final spatial feature matrix has $n+1$ attributes like SF = $\{A_1, A_2, A_3, \dots A_n, \mathrm{AV}\}$.

$$\mathrm{AV}(S_i) = \begin{cases} 1 \mathit{if}\ AR \in S_i \\ 0 \mathit{if}\ AR \notin S_i \end{cases}. \tag{4}$$

3.4 Proposed MKPCA for Spatial Data Clustering

This step reads out the spatial feature matrix as input, and the clustering is carried out using the three-dimensional variable that contains x, y and the associate vector value. The clustering is done using the proposed MKPCA method that is newly devised in this work. The most conventional method for clustering is k-means clustering [14], which partitions data space into k partitions based on the iterative procedure. The problem of finding the distance and specifying the group information by k-means algorithm provides a chance of developing a PC algorithm. PC [2] is about finding the membership probability based on the probabilistic distance. Here, the probabilistic distance is further modified with multiple kernels, where exponential and tangential kernel functions are utilized.

Assume that SF is the spatial feature matrix to be given as input for the proposed MKPCA. At first, initial centroids, $C=\{c_1, c_2, c_3 \ldots. c_k\}$, are randomly generated based on the given user input about the number of cluster required (k). Once the centroids are randomly generated, the membership probability, $P_j(y)$, is computed for every spatial data object with the centroids using the following mathematical formulae:

$$P_j(y)=\frac{D(y)}{d_j(y)},\ j=1,2,\ \ldots\ n, \tag{5}$$

$$D(y)=\frac{\overset{n}{\underset{j=1}{\pi}} d_j(y,\ c_j)}{\overset{k}{\underset{i=1}{\pi}}\underset{j\neq i}{\pi} d_j(y,\ c_j)}, \tag{6}$$

$$d_j(y,\ c_j)=\exp\left(\frac{-ED(y,\ c_j)^{\wedge 2}}{\sigma^2}\right)+\tanh(ED(y,\ c_j)), \tag{7}$$

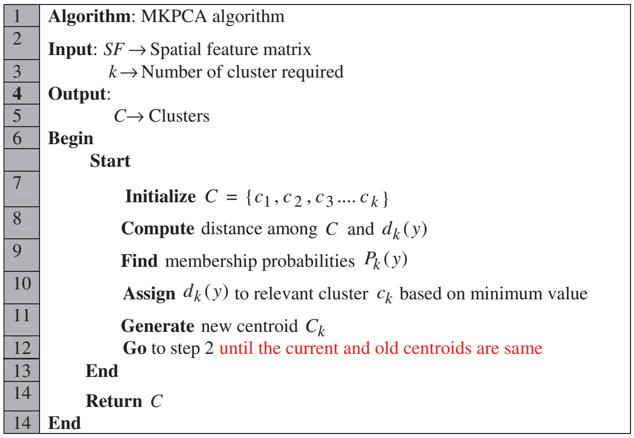
where y indicates the spatial data objects, $D(y)$ indicates the distance matrix of the input spatial data, and d_j is distance between spatial data objects with cluster centroids. In the above formulae, exponential and tangential functions are utilized to convert data space into kernel space after finding the Euclidean distance between spatial data objects and centroids.

Then, new centroids, C_k, are again generated based on the following formulae:

$$C_k=\sum_{i=1,2,n}\left(\frac{m_k(y_i)}{\sum_{j=1,2,n} m_k(y_j)}\right)y_i, \tag{8}$$

$$m_k(y_i)=\frac{p_k(y_i)^2}{d_k(y_i,\ c_k)}, \tag{9}$$

where $m_k(y_i)$ is the membership degree of the data point y_i and $P_k(y_i)$ is the membership probability. Based on the new centroids, the membership probability is computed and this procedure is repeated until the current and old centroids are the same. Figure 2 shows the pseudo-code of MKPCA. From the figure, we understand that the cluster centroids are initialized in the first iteration, and then membership probabilities are determined using the formulae given in Eqs. (5), (6), and (7), where Eq. (7) is newly developed by combining the exponential and tangential functions. Then, based on the minimum distance, the spatial data points are grouped. This process is repeated again after computing new centroids based on the equation given in Eq. (8), until there is no change in the cluster centroids.



```
1   Algorithm: MKPCA algorithm
2   Input: SF → Spatial feature matrix
3          k → Number of cluster required
4   Output:
5          C → Clusters
6   Begin
        Start
7           Initialize $C = \{c_1, c_2, c_3 .... c_k\}$
8           Compute distance among $C$ and $d_k(y)$
9           Find membership probabilities $P_k(y)$
10          Assign $d_k(y)$ to relevant cluster $c_k$ based on minimum value
11          Generate new centroid $C_k$
12          Go to step 2 until the current and old centroids are same
13      End
14      Return C
14  End
```

Figure 2: Pseudo Code of MKPCA.

4 Results and Discussion

The experimental results of the proposed method are discussed in this section, and the performance of the method in rule mining and clustering is also discussed in detail with two different metrics.

4.1 Experimental Setup

- *Platform:* The proposed spatial data mining method is implemented using MATLAB 8.2.0.701 (R2013b) with a system configuration of 2 GB RAM Intel processor and 32-bit operating system.
- *Data sets utilized:* The data sets are taken from the MATLAB tool, and the descriptions of those data sets are given in Table 1.
- *Tsunami*: These data are from the Global Tsunami Database, which is collected by the U.S. National Geospatial Data Center with the National Oceanic and Atmospheric Administration. These data contain the following attributes: Year, Month, Day, Hour, Minute, Second, Val_Code, Validity, Cause_Code, Cause, Eq_Mag, Country, Location, Max_Height, Iida_Mag, Intensity, Num_Deaths, and Desc_Deaths, with the spatial attributes like geometry, *X*, *Y*.
- *Concord:* These data are distributed by the Massachusetts Office of Geographic and Environmental Information to the NAD83 data. This data set was constructed by concatenating Massachusetts Highway Department road shapefiles for the Maynard and Concord US Geological Survey quadrangles. Attributes such as Streetname, RT_number, Class, Admin_type, and length are retained with spatial attributes like geometry, bounding box, *X*, *Y*.
- *Landareas:* These data are collected from the world map region. This world map (region) sets up empty map axes with projection and limits suitable to the part of the world specified in region. The region can be a name of continents, countries, and islands, as well as “World,” “North Pole,” “South Pole,” and “Pacific.” The spatial attributes are geometry, bounding box, *X*, *Y*, and country name.

Table 1: Description of Data Sets.

	Tsunami	Concord	Landareas
Number of spatial data objects	162	609	537
Geometry type	Point	Line	Polygon
Number of attributes	21	9	5

– *Evaluation metrics:* To analyze the performance of the spatial rule mining algorithm, the number of rules mined is utilized. To analyze the performance of spatial data clustering, the clustering accuracy, which refers to the degree of closeness of measurement of a quantity to its actual value, is utilized. Clustering accuracy is the measure of closeness of the cluster formed by the proposed algorithm to the required value, which means how accurate the members of a cluster are. The clustering accuracy is computed by using the formula below:

$$CA=\frac{1}{n}\sum_{i=1}^{g} S_i, \tag{10}$$

where n is the number of spatial data points, g is the number of classes (ground truths), and S_i is number of spatial data points occurring in both cluster i and its corresponding class.

4.2 Experimental Results

This section presents the experimental results of the proposed spatial data mining method. Table 2 shows the sample data objects of the Tsunami data. Here, spatial objects are stored as point data. Every point is represented as x and y coordinates, which are the location of the data object. Then, the attributes belonging to the Tsunami data of the corresponding object are explained with different attributes like year, month, and validity. Validity is an attribute that gives information about whether the tsunami is questionable or definite. The first object indicates that the tsunami occurred in the location of (128.3, −3.8) and the year of occurrence, month of occurrence, validity are 1950, October, and "questionable tsunami," respectively.

Table 3 shows the sample spatial rules mined from the Tsunami data. From the table, the rules state that most of the spatial data objects are affected by the tsunami due to an earthquake. Also, if the spatial data objects are closely affected by the earthquake and tsunami, then the cause code and the validity code are 1 and 4, respectively. If spatial data objects are nearer to cause code 1, then most of the spatial data objects face an earthquake and tsunami.

Figure 3 represents the visualization of input data and the clustering results. Figure 3A shows the visualization of Tsunami data, and the clustered results are shown in Figure 3B. Here, two clusters are indicated with diamond and circle symbols. Figure 3C gives the visualization of the Concord data and in Figure 3D, the clustered results of Concord data are shown. Here, two cluster outputs are shown using turquoise and red colors.

Table 2: Sample Data Set (Tsunami).

Geometry	*X*	*Y*	Year	Month	Validity
Point	128.300000000000	−3.80000000000000	1950	10	"Questionable tsunami"
Point	−156	19.5000000000000	1951	8	"Definite tsunami"
Point	157.950000000000	−9.02000000000000	1951	12	"Questionable tsunami"
Point	143.850000000000	42.1500000000000	1952	3	"Definite tsunami"
Point	−155	19.1000000000000	1952	3	"Definite tsunami"

Table 3: Sample Rule Mined from the Tsunami Data.

{Validity|definite tsunami} = > {Cause|Earthquake}{Val_Code|4}
{Cause|Earthquake} → {Validity|definite tsunami}
{Cause|Earthquake}{Validity|definite tsunami} → {Cause_Code|1}{Val_Code|4}
{Val_Code|4}{Validity|definite tsunami} → {Cause_Code|1}{Num_Deaths|0}
{Cause_Code|1}{Val_Code|4} →{Desc_Deaths|0}{Validity|definite tsunami}
{Val_Code|4}{Validity|definite tsunami} → {Cause_Code|1}{Cause|Earthquake}
{Cause_Code|1}{Val_Code|4} → {Cause|Earthquake}{Validity|definite tsunami}

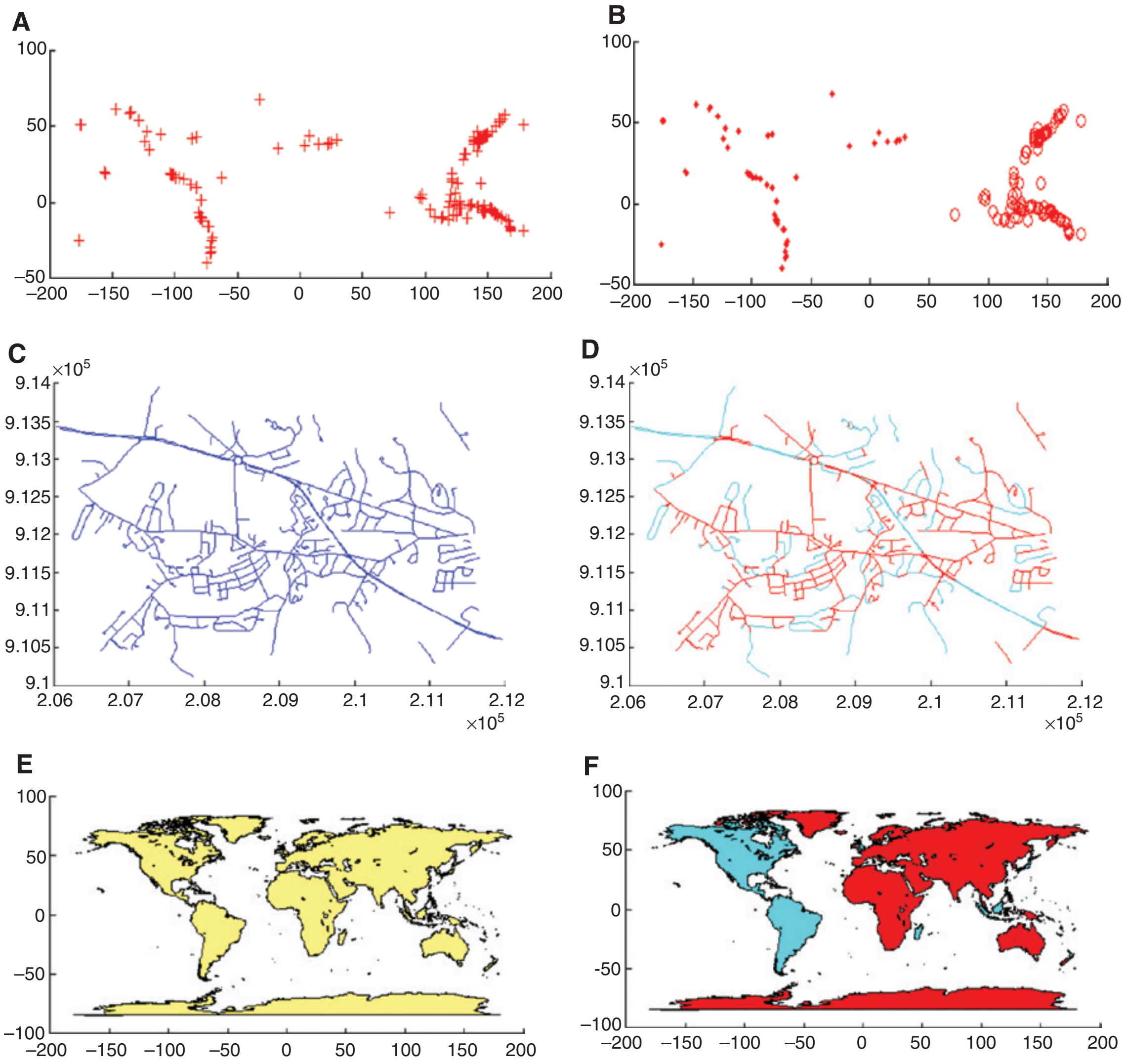


Figure 3: Visualization of Input Data and the Clustering Results.
(A) Visualization of Tsunami data. (B) Clustering results of Tsunami data. (C) Visualization of Concord data. (D) Clustering results of Concord data. (E) Visualization of Landareas data. (F) Clustering results of Landareas data.

Figure 3E shows the visualization of the Landareas data, and in Figure 3F the clustered results of Landareas data are shown. Here, two cluster outputs are shown using turquoise and red colors.

4.3 Performance Analysis of Spatial Rule Mining

The performance analysis of the proposed spatial rule mining method is discussed in this section. Here, two input data, such as Tsunami and Concord, are given as input to the proposed method. The results are obtained for various values of the support threshold. Figure 4A shows the rules mined for Tsunami data. Here, thresholds are varied from 0.2 to 0.5, and the results are analyzed with two different confidence thresholds. For the confidence of 0.2, the proposed method obtained 656 rules while the 580 rules are obtained for the confidence of 0.4 in Tsunami data. When we increase the support threshold, more important rules are obtained by the proposed method. Figure 4B shows the rules mined for Concord data. Similarly, for Concord data, we obtained only four concise rules for all the thresholds, which we took for performance comparison. This shows that the

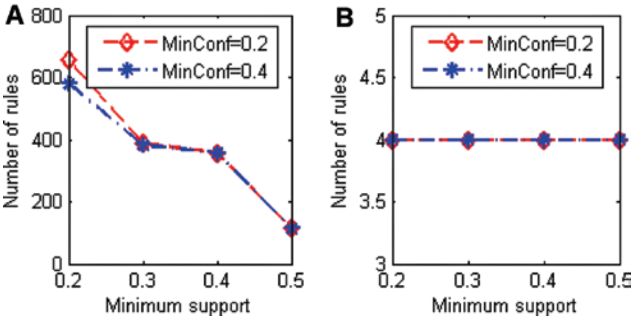


Figure 4: Spatial Rules Mined from Input Data.
(A) Rules mined for Tsunami data. (B) Rules mined for Concord data.

frequent occurrence of spatial data objects for different features level is very low. Also, the co-occurrence of patterns for Landareas is very small, so the rules are not obtainable even for the smaller thresholds.

4.4 Performance Analysis of Spatial Clustering

The performance analysis of the proposed MKPC spatial data clustering method is discussed here. The performance of the proposed MKPC is compared with that of PC [2], sweep-line clustering [21], and grid clustering [17]. Here, three input data, such as Tsunami, Landareas, and Concord, are given as inputs to the proposed method, and the results obtained are plotted in Figure 5A, B, and C, respectively. From Figure 5A, the proposed MKPC method obtained 66.6%, while the 41.9% is obtained by the PC method when the number of clusters is three. For the same input number of clusters, the sweep-line and grid clustering methods obtained an accuracy of 56.7%. When we increased the number of clusters, the proposed method showed the same accuracy value. From Figure 5B, for Concord data, we obtained accuracies of 63.7% and 47.2% for the proposed and existing methods when we fixed the number of clusters to three. Here, the existing sweep-line and grid clustering obtained an accuracy of 52.7% and 31.5%, respectively. From Figure 5C, for Landareas data, we obtained an accuracy of 100% for both the proposed and existing methods when we fixed the number of clusters to two. For the same number of clusters, the sweep-line and grid clustering methods obtained an accuracy of 84.9% and 82.9%, respectively. From Figure 5, we prove that the proposed MKPC algorithm provided better accuracy as compared with PC.

5 Conclusion

This paper presented a spatial data mining method for spatial association rule mining and further carried out spatial clustering. Initially, a spatial database that is in the format of a shape file is taken for spatial association rule mining algorithm, where support, confidence, and lift measures are applied to constrain the rules generation process. Once the spatial rules are generated, an associate vector is formed to further group the spatial data objects. Here, the MKPCA algorithm is newly developed by extending the PC with the help of a multiple kernel-based distance measure. The finding of the membership probability was further modified with multiple kernels, where exponential and tangential kernel functions are utilized. For the experimentation, three data sets from three geometry types, such as point, line, and polygon, are taken from the MATLAB tool, and the performance of the proposed method is analyzed according to the number of rules and the clustering accuracy. The experimental outcome proved that the proposed MKPCA achieved better accuracy as compared with the existing method. In the future, an optimization algorithm can be integrated with the proposed algorithm for rapid estimation of cluster centroids for improving the performance of the clustering process.

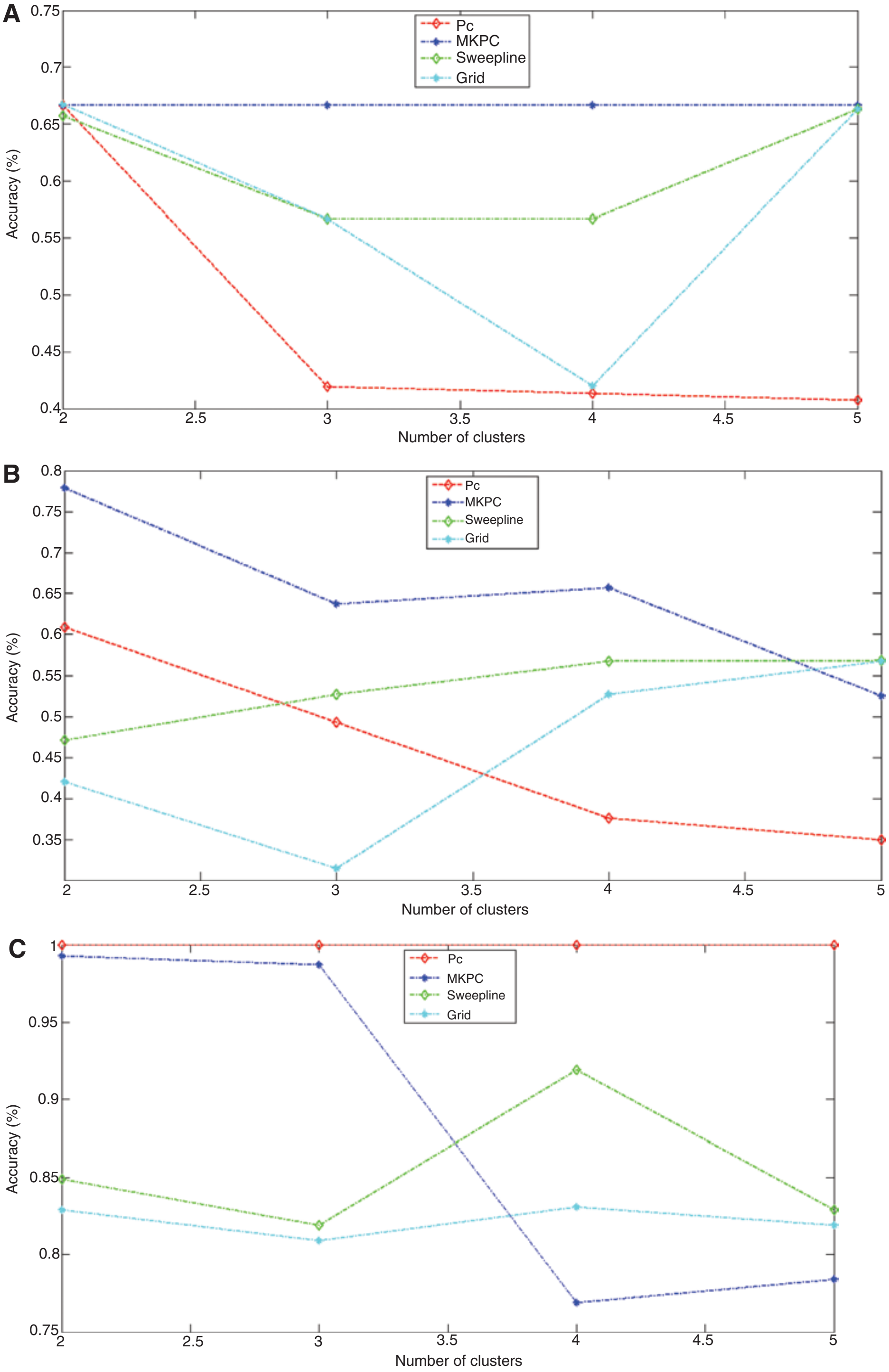


Figure 5: Clustering Accuracy of Three Input Data Sets.
(A) Accuracy for Tsunami data. (B) Accuracy for Concord data. (C) Accuracy for Landareas data.

Bibliography

[1] R. Agrawal, T. Imielinski and A. Swami, Mining association rules between sets of items in large databases, in: *Proceedings of the International Conference on Management of Data, ACM SIGMOD*, pp. 207–216, Washington, DC, May 1993.

[2] A. Ben-Israel and C. Iyigun, Probabilistic D-clustering, *J. Classif.* **25** (2008), 5–26.

[3] E. Clementini, P. Di Felice and K. Koperski, Mining multiple-level spatial association rules for objects with a broad boundary, *Data Knowl. Eng.* **34** (2000), 251–270.

[4] T. H. D. Dao and J. C. Thill, A comprehensive framework for spatial association rule mining, in: *The 7th International Conference on Geographic Information Science*, Columbus, OH, USA, September 18–21, 2012.

[5] Q. Ding, Q. Ding and W. Perrizo, PARM – an efficient algorithm to mine association rules from spatial data, *IEEE Trans. Sys. Man Cybern B Cybern.* **38** (2008), 1513–1524.

[6] B. Fan, A hybrid spatial data clustering method for site selection: the data driven approach of GIS mining, *Expert Syst. Appl.* **36** (2009), 3923–3936.

[7] W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus, Knowledge discovery in databases: an overview, in: *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. J. Frawley, eds., pp. 1–27, AAAI/MIT Press, Cambridge, MA, 1991.

[8] Y. Guo, J. Gao and L. Feng, Random spatial subspace clustering, *Knowl.-Based Syst.* **74** (2015), 106–118.

[9] U. Gupta and N. Ranganathan, A game theoretic approach for simultaneous compaction and equipartitioning of spatial data sets, *IEEE Trans. Knowl. Data Eng.* **22** (2010), 465–478.

[10] Z. Jiang, S. Shekhar, X. Zhou, J. Knight and J. Corcoran, Focal-test-based spatial decision tree learning, *IEEE Trans. Knowl. Data Eng.* **27** (2015), 1547–1559.

[11] K. Koperski and J. Han, Discovery of spatial association rules in geographic information databases, *Adv. Spat. Databases* **951** (1995), 47–66.

[12] P. Laube, M. Berg and M. van Kreveld, Spatial support and spatial confidence for spatial association rules, in: *Headway in Spatial Data Handling*, Lecture Notes in Geoinformation and Cartography, pp. 575–593, 2008.

[13] Q. Liu, M. Deng, Y. Shi and J. Wang, A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity, *Comput. Geosci.* **46** (2012), 296–309.

[14] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, in: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, University of California Press, Berkeley, CA, 1963.

[15] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay and C. A. Coello Coello, A survey of multiobjective evolutionary algorithms for data mining: part I, *IEEE Trans. Evolut. Comput.* **18** (2014), 4–19.

[16] E. Packer, P. Bak, M. Nikkila, V. Polishchuk and H. J. Ship, Visual analytics for spatial clustering: using a heuristic approach for guided exploration, *IEEE Trans. Vis. Comput. Graph.* **19** (2013), 2179–2188.

[17] A. H. Pilevar and M. Sukumar, GCHL: a grid-clustering algorithm for high-dimensional very large spatial data bases, *Pattern Recognit. Lett.* **26** (2005), 999–1010.

[18] C. R. Shyu, M. Klaric, G. Scott and W. K. Mahamaneerat, Knowledge discovery by mining association rules and temporal-spatial information from large-scale geospatial image databases, in: *Proceedings of IEEE International Conference on Geoscience and Remote Sensing Symposium*, pp. 17–20, 2006.

[19] W. R. Tobler, A computer model simulation of urban growth in the Detroit region, *Econ. Geogr.* **46** (1970), 234–240.

[20] X. Wu, X. Zhu, G. Q. Wu and W. Ding, Data mining with big data, *IEEE Trans. Knowl. Data Eng.* **26** (2014), 97–107.

[21] K. R. Zalik and B. Zalik, A sweep-line algorithm for spatial clustering, *Adv. Eng. Softw.* **40** (2009), 445–451.

[22] Q. Zhao, Y. Shi, Q. Liu and P. Franti, A grid-growing clustering algorithm for geo-spatial data, *Pattern Recognit. Lett.* **53** (2015), 77–84.