Journal of King Saud University – Computer and Information Sciences (2018) 30, 510–523

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Incremental topological spatial association rule mining and clustering from geographical datasets using probabilistic approach

Y. Jayababu [a,*], **G.P.S. Varma** [b], **A. Govardhan** [c]

[a] *Department of CSE, Pragati Engineering College, Surampalem, Andhra Pradesh 533437, India*
[b] *S R K R Engineering College, Andhra Pradesh, India*
[c] *University College of Engineering, Jawaharlal Nehru Technological University Hyderabad, India*

Received 26 March 2016; revised 20 December 2016; accepted 21 December 2016
Available online 24 December 2016

KEYWORDS

Spatial database;
Association rule mining;
Topological support;
Probabilistic approach;
Dynamic database

Abstract Due to the dynamic updating of real time spatial databases, the preservation of spatial association rules for dynamic database is a vital issue because the updates may not only invalidate some existing rules but also make other rules relevant. Consequently, the dynamic updating of spatial rules was handled by many researchers through the incremental association rule mining algorithm. Accordingly, in this paper we have developed an incremental topological association rule mining of geographical datasets using probabilistic approach. Initially, the spatial database is read out and it is passed through probability-based incremental association rule discovery algorithm to mine the topological spatial association rules. Once the rules are mined from the spatial database, the assumption here is that the database is dynamically updating for every time interval. In order to handle this dynamic nature, the proposed incremental topological association rule mining process is used in this paper. Here, the candidate topological rule generation is done from the spatial association rules using the topological relations such as, nearby, disjoint, intersects and inside/outside and the topological support is calculated using the proposed probabilistic topological support model. Finally, the spatial clustering is performed based on the mined spatial rules. From the experimentation, we proved that the maximum accuracy reached by the proposed method is 83.14% which is higher than the existing methods, which is defined as the ratio of the occurred rules and total number of topological data objects.

* Corresponding author.
E-mail addresses: jayababuy2015@gmail.com (Y. Jayababu), gpsvarma@yahoo.com (G.P.S. Varma), govardhan_cse@yahoo.co.in (A. Govardhan).
Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

1. Introduction

The collection of a large amount of spatial data is gathered by various developing fields such as remote sensing, e-commerce, and other data collection tools. Due to the huge amount of spatial data, the extraction of information from the spatial database (Ding et al., 2006) is one the most

http://dx.doi.org/10.1016/j.jksuci.2016.12.006
1319-1578

challenging task. In order to overcome the above challenge, the automated information is discovered from the spatial data, which leads to favorable fields of data mining or knowledge discovery in databases (KDD) Mukhopadhyay et al., 2014; Wu et al., 2014; Jiang et al., 2015. With the help of KDD database useful information can be retrieved from the data such as previously unknown values, hidden information and uncertain values (Clementini et al., 2000; Frawley et al., 1991). The classification of spatial data mining (Shyu et al., 2006; Laube et al., 2008; Guo et al., 2015; Ding et al., December 2008; Dao and Thill, 2012) is performed based on the types of rules, which have been located in the spatial database. Basically, the spatial association rule can be represented as $X \rightarrow Y$, in which the representation of X and Y shows the predicate set. A few of the aforementioned predicate sets contain spatial data. Basically, the large database contains various association relationships but some association relationships are not occurring regularly based on the concepts of minimum support and minimum confidence. In a set of spatial objects S, the support pattern of A is the probability that a member of spatial object S satisfies the pattern A.

Then, the confidence measure of pattern A and B is the probability that the pattern B occurs when the pattern A occurs. The threshold value is given by the user to determine the strong spatial association rules (Koperski and Han, 1995; Han and Fu, 1995; Dong et al., 2012). Basically, the spatial database is used to accumulate and control the spatial objects. The spatial object consists of two components such as descriptive component and spatial component. The components based on the spatial data mainly include their geometry, which is based on the type of point, line, surface, etc. Based on the topological relationships, the spatial objects are related to each other. The topological relationship (Pascucci et al., 2011; Fang et al., 2010; Doraiswamy et al., 2014) is based on the representation of a spatial extent by a set of points and composition of three subsets such as boundary, interior and exterior. In this paper we have taken the topological relations, which have been used as spatial predicates for complex objects. The spatial predicates are named as adjacent, within, close and overlap. A set of topological relations for geographical datasets using probabilistic approach is used in this paper. In order to reduce the computational overhead the topological relations are used in the complex objects. The generated spatial database is passed through the topological relations to mine the topological spatial association rules. Basically, the spatial database is dynamically updated for every time interval due to nature. Here, a probability-based dynamic discovery of rules is performed for the newly added database and the preservation of the important spatial rules are computed based on the probability (Mohamed and Refaat, 2011) of occurrence in the existing and new database.

The organization of the paper is as follows: Literature review is presented in Section 2. The problem definition and contributions of the paper are presented in Section 3. Proposed methodology: Incremental topological association rule mining of geographical datasets using probabilistic approach is presented in Section 4. In Section 5, the experimental results and the Performance analysis of topological spatial rule mining are presented. Finally, we conclude this paper in Section 6.

2. Literature review

Literature presents various techniques for spatial association rules' mining and clustering. Here, we present the review of different works. Koperski and Han (1995) 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 relationship among a set of spatial and possibly some nonspatial 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 faces issues when we incorporate the multiple concepts into the mining algorithm without much computational overhead. Clementini et al. (2000) have used objects with broad boundaries, the concept that absorbs all the uncertainty by which spatial data were commonly affected and allows 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 are completely missed due to the random initialization.

Shyu et al. (2006) 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. (2008) 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 to make 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.

Qin et al. (2008) have proposed an efficient approach to derive association rules from spatial data using Peano count tree (P-tree) structure. 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 FP-growth and Apriori algorithms. Even though the tree-based mining algorithm are effective than Apriori, the memory requirement to store the tree structure is high as compared with the candidate-based methods. Dao and Thill (2012) 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 does not consider the diversity constraints and automatic thresholding to group the data objects. But, the requirements of significant rules and distance function are missing in this paper.

Krista and Borut (2009) have presented an agglomerative hierarchical clustering algorithm for spatial data. It discovered clusters of arbitrary shapes which 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 post processing phase. The properties of the algorithm were demonstrated by examples. The algorithm was also adapted to the streaming algorithm for clustering large spatial datasets. It proves poor in offering high quality of the clustering solutions, capability of discovering concave/deeper and convex/higher regions, their robustness to outlier and noise. Pilevar and Sukumar (2005) have proposed a clustering method, GCHL (a Grid-Clustering algorithm for High-dimensional very Large spatial databases) by combining a density-grid based clustering with 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 requires at most a single scan through the data. But, this algorithm does not consider the kernel space for clustering the data objects and also, it utilized the traditional distance measurement for clustering.

3. Problem definition and contributions of the paper

- Let a spatial database, SDB, $S_i; 0 \leqslant i \leqslant n$ include 'n' spatial objects with different attribute values. The first challenge of finding useful information is formulated as searching problem that two series of thresholds should filter out the most useful information which can be discovered from the spatial database.
- Then, the second challenge is based on the topological rule mining algorithm. Due to dynamic nature, the topological spatial rules should be updated based on the added database. Basically, the spatiotemporal datasets are used to carry distance and other topological information based on the geometric and temporal computation. The topological relationships such as disjoints, meets, overlaps, contains, covers, intersects and equals between two spatial objects can change over time.
- Another important challenge is based on the dynamically updated spatial database for every time interval. Due to the dynamic updating of real time spatial databases, the preservation of spatial association rules for dynamic database is a vital issue because the updates may not only invalidate some existing rules but also make other rules relevant.

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

- We have developed spatial association rule mining method which is suitable for three different types of geometries such as, point, line and polygon through candidate generation process with three different constraint measures.
- We have proposed probabilistic topological support (PTS) method (Miller, 2007) to perform the topological process. Here, the candidate topological rule generation is done from the spatial association rules using the topological relations such as, nearby, disjoint, intersects and inside/outside and the topological support is calculated using the proposed PTS model, which is defined as the ratio of the occurred rules and total number of topological data.

4. Proposed methodology: Incremental topological association rule mining of geographical datasets using probabilistic approach

The block diagram representation of proposed methodology is shown in Fig. 1, in which the dynamic preservation of important topological spatial rules is presented. The steps involved in the proposed system are described as follows: (i) Initially, the spatial database is considered as an input of the proposed system. Here, the spatial database is represented in the form of vector format. (ii) Then, the spatial database is read and passed through spatial association rule mining algorithm (Zaki, 2000) to mine the important spatial association rules using probability threshold and relationship threshold measurement. (iii) Once the rules are mined from the spatial database, the assumption here is that the database is dynamically updating for every time interval. So, in order to handle dynamic nature, the topological spatial rules should be updated but the full scan with the whole database is not allowed. (iv) Here, the generated spatial rule is based on the following predicates such as nearby, disjoint, intersects and inside/outside. (v) Then, the topological spatial association rules of previous and current database are combined. (vi) Finally, the spatial rules are updated based on the probabilistic approach.

4.1. Data pre-processing

At first, the input of the system is considered as spatial data, which is stored in the shape file format. The shape file format defined as the digital vector format which is used to store the geometric location using the information based on the attributes. The spatial database can be represented as follows:

$$SDB = S_i \quad 0 \leqslant i \leqslant n \tag{1}$$

The geometric locations can be represented using different spatial representations such as point, line or polygon etc. Basically, the spatial objects are represented in the form of geometry and its related coordinate information is in the form of X and Y coordinates. The geometrical representation of spatial data objects can be expressed as follows:

$$S_i = \{L_x, L_y, A_1, A_2, A_3 \ldots A_m\} \tag{2}$$

The attribute information of the spatial data object can be generated based on the data object characteristics. Once the shape file is read, the attribute information is converted to a transaction data format. Using the transaction form of database, the mining process (Agrawal and Srikant, 1994) is carried out easily through the rule mining algorithm. Then, the set of features are generated using the unique attribute value from every attribute and stored in the transaction database. The column value of the transaction database is selected from the

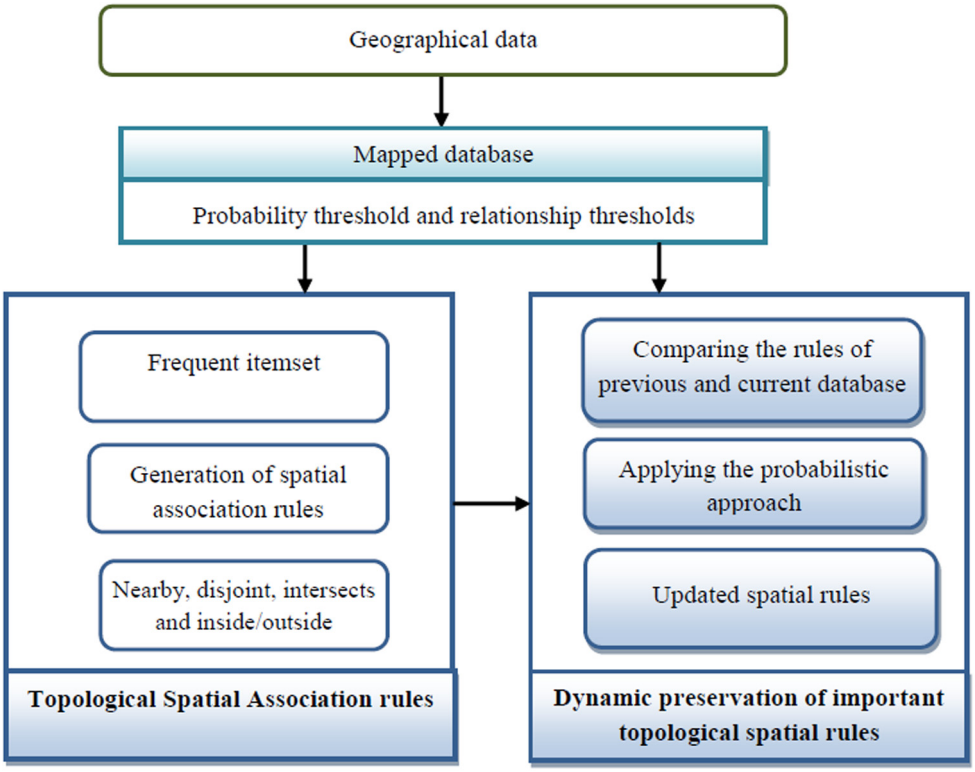


Figure 1 Block diagram representation of proposed methodology.

feature value of every unique attribute value. Then, the row of transaction database is considered as spatial data objects, in which the representation values are based on the presence of feature values in the corresponding spatial data objects.

4.2. Mining of spatial association rules

After generating the transaction database, it is given as input to the apriori algorithm. The apriori algorithm (Agrawal et al., 1993; Fang et al., 2010) is defined as association rule mining algorithm, which is used for frequent mining process. The apriori algorithm can be generated in two phases such as frequent itemset generation and association rule generation. Initially, the frequent itemsets are generated based on the feature value which has more frequency among spatial data objects. Here, the frequent combination of feature values can be filtered using the constraint measure, called support. Secondly, the confidence measure is calculated based on the association rule that is generated from the frequent itemset. The representation of standard spatial association rule can be represented as follows:

$$SAR = A_1 \cap A_2 \cap A_m \Rightarrow B_1 \cap B_2 \cap B_m \tag{3}$$

where, the subset of antecedent feature values in the spatial database is defined as A_1, A_2, A_m and the subset of consequent feature value can be denoted as B_1, B_2, B_m.. From the above spatial association rule (SAR), we can say, if the existing value of antecedent feature values A_1, A_2, A_m same as consequent feature value of B_1, B_2, B_m. In addition, the antecedent feature values are co-exists in most of the spatial data objects. The Rule SAR maintains the spatial data base with spatial association rule measures such as support S and confidence C.

$$Supp\,(A_1 \cap A_2 \cap A_m) = \frac{Number\,of\,spatial\,objects\,satisfy\,A_1 \cap A_2 \cap A_m}{Total\,number\,of\,spatial\,objects\,in\,SDB}$$

$$Conf\,(A_1 \cap A_2 \cap A_m \Rightarrow B_1 \cap B_2 \cap B_m) = \frac{\sup(A_1 \cap A_2 \cap A_m \cap B_1 \cap B_2 \cap B_m)}{\sup(A_1 \cap A_2 \cap A_m)}$$

Then, the other spatial data objects measurement like lift is also used to calculate the importance level of association rules. This lift measure is given as follows:

$$lift(A_1 \cap A_2 \cap A_m \Rightarrow B_1 \cap B_2 \cap B_m) = \frac{\sup(A_1 \cap A_2 \cap A_m \cap B_1 \cap B_2 \cap B_m)}{\sup(A_1 \cap A_2 \cap A_m) \times \sup(B_1 \cap B_2 \cap B_m)}$$

4.3. Mining of topological spatial association rules using input data and spatial association rules

Generalization-based spatial data mining methods are used to discover the relationship between the spatial and non-spatial objects. However, the reflecting structure of the spatial objects cannot be discovered by spatial data mining methods. Basically, the spatial association rules contain different kinds of spatial predicates, which represent the topological relationship between the spatial objects such as adjacent-to, close-to, inside, intersection, near-by, etc. In addition, the spatial predicates can be represented in the form of spatial orientation such as left, right, south, east, north, etc. More over the spatial

predicates contain some information that has been based on the distance such as close-to, far-away, etc. The preliminary concepts based on the spatial association rule are discussed as follows:

Definition 1. In definition 1, two kinds of thresholds can be introduced such as minimum support and minimum confidence. Here, the minimum support is defined as the proportion of transactions in the database based on the number of itemsets. In a dataset S, the support of conjunction of predicates $A_1 \wedge A_2 \wedge \cdots \wedge A_m$ can be denoted as $\sigma(A/S)$. The confidence rule of $A \rightarrow B$ can be represented in a set S. The representation of confidence rule can be denoted as $\varphi(A \rightarrow B/S)$ that is defined as the ratio of $\sigma(A \wedge B/S)$ Vs $\sigma(A/S)$. From definition 2, we can understand that B is satisfied by a set of object member S when A is satisfied by the same member of object set S. Here, 1-predicate is defined as single predicate and the conjunction of k single predicates is defined as the k-predicates.

Definition 2. Basically, the large spatial database is used to generate the large number of spatial association rules. However, people are interested in the spatial patterns based on the two concepts like large support and high confidence. Consequently, the spatial association rule is generated with large supports and high confidence, which is named as strong rule. Here, the spatial association rule can be represented as follows:

$$A_1 \wedge A_2 \wedge \cdots \wedge A_m \rightarrow B_1 \wedge B_2 \wedge \cdots \wedge B_n \ (\%c)$$

where, $A_1 \wedge A_2 \wedge \cdots \wedge A_m$ is defined as the least predicates, $B_1 \wedge B_2 \wedge \cdots \wedge B_n$ is defined as the spatial predicates and the confidence of the spatial association rule can be represented as c%.

Definition 3. If the support of A is greater than its minimum support threshold, a set of predicate A is considered as large in set S at level k. Accordingly, the confidence rule $A \rightarrow B/S$ is high when its confidence is greater than its corresponding confident threshold.

Definition 4. If the predicate $A \wedge B$ is large in set S, the following confidence rule $A \rightarrow B/S$ is considered as strong and high in set S.

4.3.1. Topological relations considered

Basically, the spatial relation is used to specify the location of objects in space corresponding to the reference object. The commonly used spatial relations are named as topological relations, directional relations and distance relations. The spatial relations based on the topological rules are discussed in this section. The topological model is based on the Dimensionally Extended nine-Intersection Model (DE-9IM) standard, in which the spatial relationship between the two regions is described. Based on the DE-9IM model, the spatial relations are not affected by rotation, translation and scaling process. Some of the topological relations can be expressed as follows:

(i) Equals: equality is one of the important relations in the topological concept. It is used to analyze whether the particular condition already exists in a set or not. The equal's relation is shown below:

$$S_1 \equiv is\_a(p, l\arg e\_town) \wedge equals(p, q) \wedge is\_a(q, water)$$

where, the two spatial objects such as p and q are topologically equal.

(ii) Disjoint: disjoint is defined as no common points, which is used to form the set of disconnected geometrics.

$$S_2 \equiv is\_a(p, school) \wedge disjoint(p, q) \wedge is\_a(q, beach)$$

where, the spatial points of p and q are placed in disjoint manner, in which p denotes the school and q denotes the beach area.

(iii) Intersects: The two spatial data objects p and q are placed in the overlapped manner. Here, the spatial data objects p and q contain the common area that can be shown below:

$$S_3 \equiv is\_a(p, l\arg e\_town) \wedge inter\sec ts(p, q) \wedge is\_a(q, small\_city)$$

The above equation shows the intersection of both spatial objects p and q does not contain the null area.

(iv) Touches: Here, the first spatial object p touches the next spatial object q, in which at least one boundary point is considered in common. However, the spatial object does not contain any interior points.

$$S_4 \equiv is\_a(p, l\arg e\_town) \wedge touches(p, e) \wedge is\_a(e, highway)$$

(v) Covers: one of the spatial objects q lies inside the spatial object p, in which no point of q is placed in the exterior of p. Here, every point of q covers the interior points of p that can be represented as follows:

$$S_5 \equiv is\_a(p, world) \wedge covers(p, r) \wedge is\_a(r, country)$$

4.3.2. Spatial rule mining algorithm

In order to analyze the relational and transactional databases, the data mining algorithms are used for recent researches. Most of the data mining methods are used in the spatial objects with exactly familiar location and finite accuracy. Due to the different sources, the spatial information is affected by various uncertainties such as incompleteness, instability, indistinctness, imprecision, and error (Worboys, 1998). Here, the technique is used to mine the spatial association rules when the spatial information is inaccurate. In order to calculate the strong spatial association rules, the topological relations are proposed in this paper. The topological relations are used to reduce the computational complexity. However, the topological relations generate inefficiency due to intersection models. The spatial data mining process can be classified based on the types of rules that can be discovering in the spatial databases. The mining algorithm contains three processing steps such as candidate topological rule generation from spatial association rules, finding of topological support using spatial database and filtering of rules.

Step 1: Candidate topological rule generation from spatial association rules

Initially, the task relevant spatial objects are extracted from the spatial database based on the spatial association rules. In this paper, the spatial data such as tsunami and concord data

are considered as a geographical dataset. All these types of geographical datasets are stored in the relevant spatial database, which have been further used in the mining process. The storage of relevant database can be accomplished by the spatial query implementation. The search space is partitioned based on the topological intersection values, in which the space is searched at a time for level one. The searching process is starting from the iteration between the candidate generation and evaluation phase. Candidate generation phase is based on the refinement process, in which one or more allowed spatial objects to the pattern to be refined. Here, the spatial association rules are used to generate the support, confident and lift values of spatial objects. The candidate evaluation phase is used to compare the candidate generated spatial pattern with minimum support threshold. From the comparison results we can say, the minimum support patterns are rejected in the candidate evaluation phase based on the support threshold. For example, consider 50 geographical areas with tsunami based on the spatial association rules. All these objects collected from the tsunami areas are stored into the relevant spatial database for further mining process. The example of spatial association rule is represented as follows:

$$R_1 \Rightarrow if\, p\, is\, geographical\_area \wedge p\, is\, heavy\, rain \wedge p\, contains\, sea\ (55\%, 70\%)$$

$$R_2 \Rightarrow if\, p\, is\, geographical\_area \wedge p\, is\, heavy\, flood \wedge p\, contains\, river\ (15\%, 60\%)$$

where, the measures of support and confidence can be represented as $(S\%, C\%)$. From the generated spatial association rule, the minimum support rules only considered for further processing. While considering the minimum support as 25%, the association rules are considered above this minimum value. Accordingly, the spatial association rule R_2 can be neglected for further process. Because, the support value is calculated from the second rule is obtained as below 25%.

After generating the relevant spatial data base, the topological relation such as within, overlap, touches, intersects and disjoint are applied to the spatial dataset of all the geographical area based on the filtered minimum bounding rectangle (MBR) predicates such as heavy rain, sea, river, cyclone, etc. Then, the candidates generate the topological relations based on the generated relevant spatial database. Here, the candidate generates all possible topological relations based on the generated spatial association rules. Some of the candidate rules are generated from the above mentioned spatial rule that is represented as follows:

$$CR_{11} \Rightarrow is\_a(p, geographical\_area) \wedge with\_in(p, heavy\, rain) \wedge close\_to(p, sea)$$

$$CR_{12} \Rightarrow is\_a(p, geographical\_area) \wedge close\_to(p, heavy\, rain) \wedge with\_in(p, sea)$$

$$CR_{13} \Rightarrow is\_a(p, geographical\_area) \wedge adjacent\_to(p, heavy\, rain) \wedge with\_in(p, sea)$$

$$CR_{14} \Rightarrow is\_a(p, geographical\_area) \wedge covers(p, heavy\, rain) \wedge inside(p, sea)$$

where, CR is defined as the candidate generated topological rule and the number of candidate rules generated based on the first rule (R_1) can be represented as $CR_{11}, CR_{12}, CR_{13}, CR_{14} \cdots etc.$

Step 2: Finding of topological support using spatial database

Once the candidate topological rules are generated, the topological support is calculated for every generated candidate rules such as $CR_{11}, CR_{12}, CR_{13} \cdots etc.$

$$PTS(R) = \frac{O(R)}{T(D)} \Rightarrow P(R) \tag{4}$$

where, the probabilistic topological support can be denoted as $PTS(R)$, which is defined as the ratio of occurred rules $O(R)$ to the total number of topological data $T(D)$ and $P(R)$ is defined as the probability of rule R within the database. Then, the topological support is calculated from the generated candidate rules that are represented as follows:

$$TCR_{11} \Rightarrow is\_a(p, geographical\_area) \wedge with\_in(p, heavy\, rain) \wedge close\_to(p, sea)(34\%)$$

$$TCR_{12} \Rightarrow is\_a(p, geographical\_area) \wedge close\_to(p, heavy\, rain) \wedge with\_in(p, sea)(72\%)$$

$$TCR_{13} \Rightarrow is\_a(p, geographical\_area) \wedge adjacent\_to(p, heavy\, rain) \wedge disjoint(p, sea)(49\%)$$

$$TCR_{14} \Rightarrow is\_a(p, geographical\_area) \wedge covers(p, heavy\, rain) \wedge inside(p, sea)(88\%)$$

where, TCR is defined as the topological support of particular candidate rule. The PTS value of TCR_{11}, TCR_{12}, TCR_{13} and TCR_{14} is measured as 34%,72%,49% and 88% respectively.

Step 3: Filtering of rules

Basically, the filtering process is used to reduce the unwanted computations and increase the speed of the computation. Here, the filtering process can be done based on the threshold process such as probability threshold and relationship threshold which is given by the user. After generating the topological support, the filtering rules are applied to select the strong spatial association rules generated by the candidates. When the threshold value of topological support is fixed as 50%, the candidate rules are selected based on the fixed threshold. If the value of topological support is above 50%, the particular association rule is stored in the spatial database. Accordingly, from the above calculated topological support, the first candidate rule and third candidate rule are eliminated and the second and fourth candidate association rules are stored in the spatial database. Then, the filtered candidate association rules are represented as follows: (see Fig. 2)

$$TCR_{12} \Rightarrow is\_a(p, geographical\_area) \wedge close\_to(p, heavy\, rain) \wedge with\_in(p, sea)(72\%)$$

$$TCR_{14} \Rightarrow is\_a(p, geographical\_area) \wedge covers(p, heavy\, rain) \wedge inside(p, sea)(88\%)$$

4.4. Incremental topological association rule mining

After generating the topological support using spatial dataset, the assumption here is that the spatial database is updated dynamically for every time interval due to environmental changes. Due to the changes of spatial data, the topological support also gets changed. This is one of the major challenges faced by researchers. Basically, the whole update of topological support makes the computational complexity. In order to reduce this computational complexity, the incremental topological association rule mining is proposed in this paper. Here,

```
Input : SDB
Output : mined value
Parameters: SDB→spatial database
            P_DB→ predicate database
            MBR→ minimum boundary rectangles
Begin
        Relevant _SDB = extract _task _relevant _objects
        MBR_P database = calculate MBR_P based on the relevant _SDB
        MBR_P ===MBR predicates
        P-DB = filter with min _supp(min _supp _threshold, MBR_P database)
for (L=1, L!=max& P_DB!=null, L++)
        {
        P_DB=calculate _P(L, P_DB, relevant _SDB)
        P_DB=filter with min _supp(min _sup(L), P_DB)
        mine _association rules(P_DB)
}
End
```

Figure 2 Algorithmic description of mining spatial association rules.

the newly updated topologic support is added with existing generated topological support. By doing this process, the computation time is reduced and reduces the complexity of computation also.

Step 1: Mining topological spatial association rules from static database

This section shows the mining of topological spatial association rules from the static database. Here, the spatial database values are not changed dynamically. From the static database, the topological spatial association rules are mined incessantly. Then, the mined static spatial association rules are stored in the rule set 1. At first, 100 spatial association rules are stored in the static database. For example, the topological spatial association rules from static database are represented as follows:

$$TCR_{12} \Rightarrow is\_a(p, geographical\_area) \wedge close\_to(p, heavy\,rain) \wedge with\_in(p, sea)(72\%)$$
$$TCR_{14} \Rightarrow is\_a(p, geographical\_area) \wedge covers(p, heavy\,rain) \wedge inside(p, sea)(88\%)$$

Then, the spatial association rules are updated due to environmental changes. So, the number of spatial association rules is changed as 110, here 10 rules are newly updated due to dynamic nature. Due to this updating process, the topological relations are affected easily. In order to handle this situation, the updated rules are taken out and the spatial association rules and topological support are applied. The spatial association rules of updated data can be expressed as follows:

$$UR_1 \Rightarrow if\,p\,is\,geographical\_area \wedge p\,is\,heavy\,flood \wedge p\,contains\,sea\ (51\%, 79\%)$$
$$UR_2 \Rightarrow if\,p\,is\,geographical\_area \wedge p\,is\,cloudy \wedge p\,contains\,mountains\ (19\%, 46\%)$$

Here, UR_1 and UR_2 denote the first and second updated rules respectively. After generating the spatial association rule, all possible topological relations are calculated using candidate generation phase that can be shown below:

$$CUR_{11} \Rightarrow is\_a(p, geographical\_area) \wedge adjacent\_to(p, heavy\,flood) \wedge intersets(p, sea)$$
$$CUR_{12} \Rightarrow is\_a(p, geographical\_area) \wedge close\_to(p, heavy\,flood) \wedge with\_in(p, sea)$$
$$CUR_{13} \Rightarrow is\_a(p, geographical\_area) \wedge in(p, heavy\,flood) \wedge touch(p, sea)$$
$$CUR_{14} \Rightarrow is\_a(p, geographical\_area) \wedge covers(p, heavy\,flood) \wedge inside(p, sea)$$

where, CUR is defined as the candidate topological rule based on the updated dataset. After generating the candidate topological rule, the topological support is generated for the updated spatial database. Some of the topological support representations of the updated spatial database can be expressed as follows:

$$TCUR_{11} \Rightarrow is\_a(p, geographical\_area) \wedge adjacent\_to(p, heavy\,flood) \wedge intersets(p, sea)(25\%)$$
$$TCUR_{12} \Rightarrow is\_a(p, geographical\_area) \wedge close\_to(p, heavy\,flood) \wedge with\_in(p, sea)(47\%)$$
$$TCUR_{13} \Rightarrow is\_a(p, geographical\_area) \wedge in(p, heavy\,flood) \wedge touch(p, sea)(63\%)$$
$$TCUR_{14} \Rightarrow is\_a(p, geographical\_area) \wedge covers(p, heavy\,flood) \wedge inside(p, sea)(77\%)$$

where, the topological support of candidate generated rule for updated spatial dataset is defined as ($TCUR$. After generating the topological support for updating datasets, the filtering process is done based on the probability threshold and relationship threshold. Here, the threshold value is fixed as 50% to filter the topological support relations. Based on the fixed threshold value, the filtered topological support of updated database can be represented as follows:

$$TCUR_{13} \Rightarrow is\_a(p, geographical\_area) \wedge in(p, heavy\,flood) \wedge touch(p, sea)(63\%)$$
$$TCUR_{14} \Rightarrow is\_a(p, geographical\_area) \wedge covers(p, heavy\,flood) \wedge inside(p, sea)(77\%)$$

Step 2: Merging of topological spatial association rules of static database and updated database

After generating the spatial association rule and topological support of updated spatial database $TCUR$, which is merged with already existing static database TCR. Before the merging process, the available static and updated databases are shown below:

$$TCR_{12}^{S} \Rightarrow is\_a(p, geographical\_area) \wedge close\_to(p, heavy\,rain) \wedge with\_in(p, sea)(72\%)$$

$$TCR_{14}^{S} \Rightarrow is\_a(p, geographical\_area) \wedge covers(p, heavy\,rain) \wedge inside(p, sea)(88\%)$$

$$TCUR_{13}^{D} \Rightarrow is\_a(p, geographical\_area) \wedge in(p, heavy\,flood) \wedge touch(p, sea)(63\%)$$

$$TCUR_{14}^{D} \Rightarrow is\_a(p, geographical\_area) \wedge covers(p, heavy\,flood) \wedge inside(p, sea)(77\%)$$

where, TCR^S is defined as the static database and $TCUR^D$ is defined as the dynamic database. Then, the merging process is done between both static and updated databases that can be represented as follows:

$$PTS(TCR^S_{12}, TCR^S_{14}) \iff PTS(TCUR^D_{13}, TCUR^D_{14})$$

While performing the merging process, the updated database is searched in the static database whether it is available or not. Initially, the updated topological association rule is merged with the static topological rule. Here, the merging process can be calculated as follows:

$$PTS(R_1) = \left(\frac{UPTS_S(R_{12}) + PTS_S(R_{12})}{2}\right) = \left(\frac{0+72}{2}\right) = 36\%$$

$$PTS(R_2) = \left(\frac{UPTS_D(R_{14}) + PTS_S(_{14})}{2}\right) = \left(\frac{77+88}{2}\right) = 82.5\%$$

$$PTS(R_3) = \left(\frac{UPTS_S(R_{13}) + PTS_S(R_{13})}{2}\right) = \left(\frac{0+63}{2}\right) = 31.5\%$$

where, $PTS_D(R_1)$ is defined as the dynamic probabilistic topological rule based on rule one (R1) which is calculated to be 36%. Here, the first static topological rule TCR^S_{12} is matched with the dynamic database. Then find the average value of TCR^S_{12} and matched value. If no value is matched with the TCR^S_{12} means consider the matched value as zero. Then, the average between the probabilistic topological support of TCR^S_{12} value and its matched value is calculated.

Step 3: Filtering

After performing the merging process, the filtering process is applied to the dynamic database based on the probability threshold and user threshold. When the threshold value is fixed as 75%, the spatial association rules above the threshold value only stored in the spatial database. After filtering, the final association rules stored in the spatial database is represented as follows:

$$TCR_{14} \Rightarrow is\_a(p, geographical\_area) \wedge covers(p, heavy\,rain) \wedge inside(p, sea)\,(88\%)$$

4.5. Apply the spatial rules of the static database to spatial data clustering

These steps read out spatial feature matrix as input and the clustering is carried out using the three dimensional variable which contains x, y and associate vector value. The clustering is done using the MKPCA method (Gladis et al., 2015). The most conventional method for clustering is k-means clustering 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 probabilistic clustering algorithm. The probabilistic clustering 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. The major advantages of kernel space are that the data can be easily separable if the range of data values are exposed in different interval using kernel function. This separation makes the algorithm more easy to find the similar data points. The main characteristic of Kernel Methods, however, is their distinct approach to this problem. Kernel methods map the data into higher dimensional spaces in the hope that in this higher-dimensional space the data could become more easily separated or better structured. Here, we select exponential and tangential kernel for converting data into kernel space. The exponential kernel behaves almost non-linearly and the higher-dimensional projection is possible with the non-linear power. The tangential function maintains good regularization and the decision boundary will be highly non-sensitive to noise in training data.

Let assume that SF be 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, membership probability $P_j(y)$ is computed for the every spatial data object with the centroids using the following mathematical formulae.

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

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

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

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

5. Results and discussion

The experimental results of the proposed method are discussed in this section and the proposed incremental spatial association rules are discussed in detail with three different geographical dataset such as tsunami and concord dataset.

5.1. Experimental set up

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 OS. *Datasets utilized:* The datasets are taken from MATLAB tool and the

description of those datasets are given in Table 1. *Tsunamis*: These data are from Global Tsunami Database which is collected by U.S. National Geospatial Data Center (NGDC) with National Oceanic and Atmospheric Administration (NOAA). These data contain the following attributes such as, Year, Month, Day, Hour, Minute, Second, Val_Code, Validity, Cause_Code, Cause, Eq_Mag, Country, Location, Max_Height, Iida_Mag, Intensity, Num_Deaths, Desc_Deaths with the spatial attributes like, geometry, X, Y. *Concord:* These data are distributed by MassGIS (Geographic and Environmental Information (MassGIS) to the NAD83 datum. This data set was constructed by concatenating Massachusetts Highway Department road shapefiles for the Maynard and Concord USGS Quadrangles. The following attributes like, streentname, RT_number, Class, Admin_type, length are retained with spatial attributes like, geometry, bounding box, X, Y. *Landareas:* These data are collected from worldmap region. This worldmap(region) sets up an empty map axes with projection and limits suitable to the part of the world specified in region. Region can be names of continents, countries, and islands as well as 'World', 'North Pole', 'South Pole', and 'Pacific'. The spatial attributes are geometry, bounding box, X, Y, country name.

Evaluation metrics: To analyze the performance of the spatial rule mining algorithm, a number of rules mined are utilized.

5.2. Experimental results

This section presents the experimental results of the proposed spatial data mining method. Table 2 shows the sample data objects of tsunamis data.

Table 3 shows the sample spatial rules mined from the tsunamis data. From the table, rules state that most of the spatial data objects are affected with tsunami due to the cause of earth quake. Also, if the spatial data objects are closely affected with earthquake and tsunami, then cause code and validity code are one and four. If spatial data objects are nearer to cause code one, then most of the spatial data object face earthquake and tsunami. Table 4 shows the sample topological rules mined from the tsunamis data. Table 5 shows the sample rule mined from the concord data. Table 6 shows the sample rule mined from the landareas data.

Fig. 3 represents the visualization of input data. Fig. 3.a and b shows the visualization of tsunamis data and concord data respectively.

Table 1 Description of datasets.

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

Table 3 Sample rule mined from the tsunamis 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}

Table 4 Sample topological rule mined from the tsunamis data.

Close_to{Validity|definite tsunami} → with_in{Cause|Earthquake}^adjacent_to {Val_Code|4
Covers{Cause|Earthquake} → close_to{Validity|definite tsunami}
Inside{Cause|Earthquake}^adjacent_to{Validity|definite tsunami} → disjoint{Cause_Code|1}^nearby{Val_Code|4}

Table 5 Sample rule mined from the concord data.

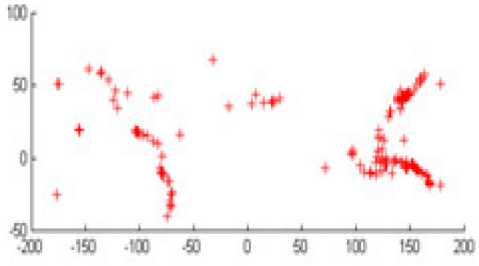
{{RT_NUMBER|0} = > {ADMIN_TYPE|0
{ADMIN_TYPE|0} = > {RT_NUMBER|0}
{RT_NUMBER|0} = > {CLASS|5}
{CLASS|5} = > {RT_NUMBER|0}

Table 6 Sample rule mined from the landareas data.

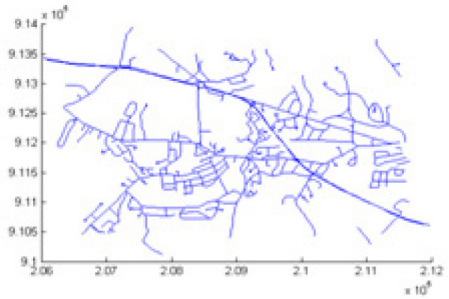
{X |-180} = > {CLASS |Antarctica}
{Y|-84} = > {CLASS | Africa and Eurasia}
{{X|-95}{Y|-100} → {CLASS | Great Britain}

Table 2 Sample dataset (tsunamis).

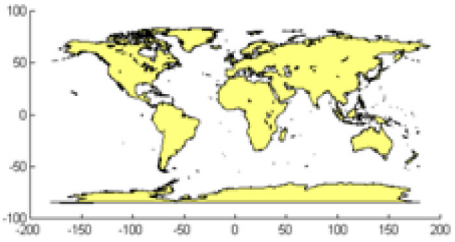
Geometry	X	Y	year	Month	Validity
'Point'	128.300000000000	−3.80000000000000	1950	10	'questionable tsunami'
'Point'	−156	19.5000000000000	1951	8	'definite tsunami'
'Point'	157.950000000000	−9.02000000000000	1951	12	'questionable tsunami'
'Point'	143.850000000000	42.1500000000000	1952	3	'definite tsunami'
'Point'	−155	19.1000000000000	1952	3	'definite tsunami'



(a) Visualization of tsunamis data



(b)Visualization of concord data



(c)Visualization of landareas data

Figure 3 Visualization of input data.

5.3. Performance analysis of topological spatial rule mining

This section shows the performance analysis of the proposed topological spatial rule mining using spatial measures. We have considered three input datasets such as tsunamis, concord and landareas datasets. Then, the results are calculated based on the various values of topological measures such as support, confidence and percentage.

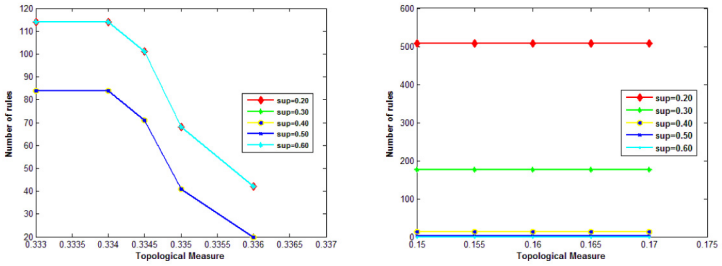
5.3.1. Spatial rules mined from input data using various support measures

This section shows the mining of spatial rules from the input data which has been selected from the tsunami and concord dataset. Fig. 4 shows the mining of spatial rules based on the various support measure using tsunami data. While increasing the topological measure, the number of rules used for mining process is reduced. Here, thresholds are varied from 0.2 to 0.6 and the results are analyzed with five different support thresholds. For the support of 0.2, the proposed method obtained 114 rules while the topological measure is maintained as 33% in tsunamis data. For the support value of 0.5, the number of rules used in the propose method varies from 84 to 20 based on the topological measures. From the Fig. 4(a), we can conclude that the number of rules are reduced by increasing the topological measures.

Fig. 4(b) shows the number of spatial rules mined for concord data using support measures. Here, the thresholds vary from 0.2 to 0.6 and the topological measures vary from 15% to 17%. When the support measure is considered as 0.4, the proposed method generated only two numbers of rules. When the support is taken as 0.2, the number of rules obtained in the proposed method is 508. Furthermore, increasing the support measures from 0.2 to 0.3, the number of rules used in the proposed method is getting reduced as 176 from 508. From the Fig. 4(b), we can say, the changing of topological measures does not affect the number of rules. Here, the numbers of rules are maintained constantly by increasing the topological measures. The reason behind this property is that the support has more influence on mining the rules than the topological measure. We can see that, for various support values, the number of rules is different. However, the numbers of rules used in the proposed method are reduced by increasing the support threshold. Finally, we can conclude that, while the increasing values of support threshold from 0.2 to 0.6, the number of rules used in the proposed method is getting reduced from 508 to zero value.

5.3.2. Spatial rules mined from input data using various confidence measures

In this section, the performance analysis of the proposed spatial rule mining method is discussed and the results are obtained for various values of confidence measures using tsunami dataset. While fixing the confidence value, the generated rules are calculated based on the topological measures. Here, the values of topological measures are changed from 33.3% to 33.6%. When the number of confidence value is fixed as 0.24, the number of rules obtained in the proposed method is 302. Then for the same confidence value, the generated rule is altered based on the topological measures. When the topological measure is changing from 33.4% to 33.5%, the number of rule generation is reducing from 157 to 84. For the confidence value 0.22, the proposed method does not generate any kind of rules. From the Fig. 5(a), we can conclude, the numbers of generated rules are increased based on the various topological measures.



(a) Rules mined for tsunami data (b) Rules mined for concord data

Figure 4 Spatial rules mined from input data using various support measure.

Fig. 5(b) shows the mining of spatial rules for concord dataset using various confidence measures. Here, the numbers of generated rules are maintained constantly for all of the topological measures. Here, thresholds are varied from 0.2 to 0.25 and the results are analyzed with five different confidence thresholds. Here, the more numbers of rules are generated for the confidence value 0.2. Then further increasing the confidence threshold as 0.21, the number of confidence rules is generated as 176. Again the confidence threshold is increased to 0.22, 13 numbers of rules are generated for the corresponding topological measures. Here, the numbers of rules is independent of the topological measures. The topological measurement values can be changed from 15% to 17%. The number of generated rule in the proposed method is reduced based on the increasing value of confidence measures. The less number of rules is generated based on the following confidence measure such as 0.22, 0.24 and 0.25. From the figure, we prove that the proposed method maintained constant rules based on the topological measures.

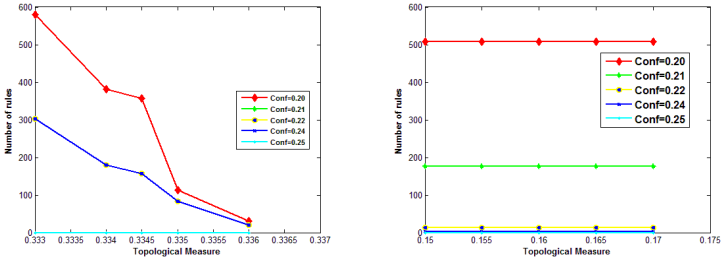
5.3.3. Spatial rules mined from input data using various incremental percentages

This section shows the performance analysis of spatial rule mining based on the various incremental percentages. Fig. 6 (a) shows the mining of spatial data which are collected from the tsunami dataset, in which the mining process is carried out based on the various incremental percentages and the changes of topological measures. Here, some of the incremental percentages do not generate the number of rules. From the figure, we can say, while changing the percentage value as 70, there are no rules generated. Further increasing the percentage as 80, the number rules are generated as 114. While fixing the percentage value as 50, the proposed method generated the 114 for 33.4% of topological measure. Further increasing the topological measure as 33.6%, the generated rules are reduced as 84%. Again increasing the topological measure from 33.65 to 33.8%, the generated rules are maintained constantly. Finally, the rule generation becomes zero for the topological measurement value of 35%.

Fig. 6(b) shows the performance analysis of the proposed spatial rule mining method for concord data using various incremental percentages. Here, various number of percentage values generate the various number of spatial rules. Here, five percentage thresholds are considered for mining spatial rules for concord dataset. While increasing the topological measures, the numbers of generated rules are getting reduced linearly. For the percentage value of 50, the proposed method obtained 508 spatial rules, which is analyzed based on the topological measures. For the same percentage value 50, the generated spatial rules are changed like 508 to 216 based on the topological measures. Again the numbers of generated rules are reduced from 216 to 28, based on the corresponding topological measure of 22%. When the number of percentage measure is fixed as 90%, the proposed method obtained 461 rules. Then, the generated rules are changed based on the topological measures in the range of 17% to 23%. From the Fig. 6 (b) we can prove that, the number of rules generated in the proposed method reduced based on the topological measures.

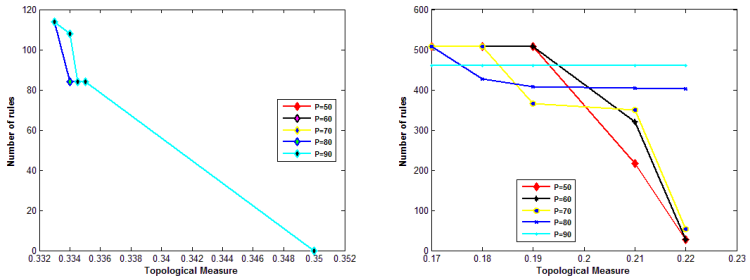
5.4. Performance analysis of spatial clustering

The analysis of the proposed MKPC spatial data clustering method performed on the static database is discussed here. The performance of proposed MKPC is compared with Probabilistic clustering (PC) (Ben-Israel and Iyigun, 2008), K-medoid clustering (Pilevar and Sukumar, 2005) and grid



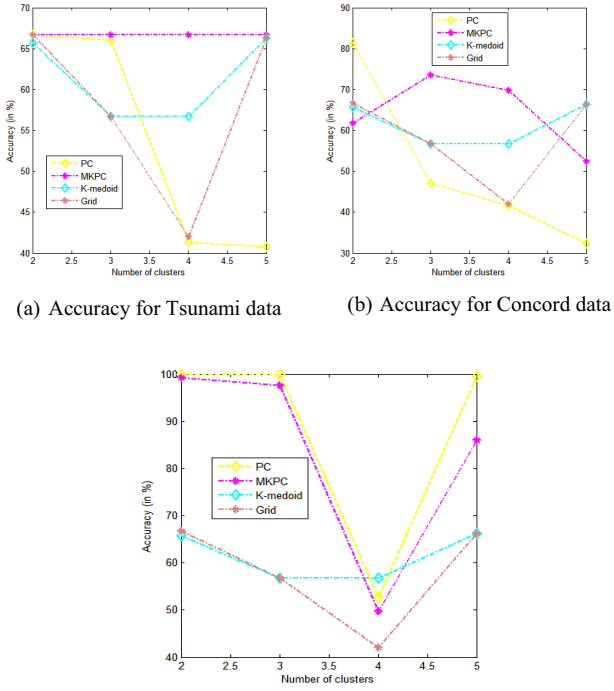
(a) Rules mined for tsunami data (b) Rules mined for concord data

Figure 5 Spatial rules mined from input data using various confidence measures.



(a) Rules mined for tsunami data (b) Rules mined for concord data

Figure 6 Spatial rules mined from input data using various incremental percentages.



(a) Accuracy for Tsunami data (b) Accuracy for Concord data

(c) Accuracy for landareas data

Figure 7 Clustering accuracy of three input datasets.

clustering (Zhang and Couloigner, 2005). Here, three input data such as, tsunamis, landareas and concord are given as input to the proposed method and the results obtained are plotted in Fig. 7a, 7.b and 7.c respectively. The experimentation is conducted 10 times for the random centroids and the average performance is taken for comparison. From Fig. 7a, the proposed MKPC method obtained 66.6% while the 65.1% are obtained by the PC method when the number of cluster is three. For the same input number of cluster, K-medoid and grid clustering methods obtained the accuracy of 56.7%. When we increase the number of clusters, the proposed method shows the same accuracy value. From Fig. 7b, for concord data, we obtain the accuracy of 72.1% and 46.2% for the proposed and existing methods when we fix the number of cluster as three. Here, the existing K-medoid and grid clustering obtained the accuracy of 56.7%. From Fig. 7c, for landareas data, we obtain the accuracy of 100% and 100% for the proposed and existing methods when we fix the number of cluster as two. For the same number of clusters, the K-medoid and grid clustering methods obtained the accuracy of 68% and 67% respectively. From Fig. 7, we prove that the proposed MKPC algorithm provided better accuracy as compared with Probabilistic clustering (PC).

Table 7 shows the statistical analysis of clustering accuracy on three datasets. In Table 7 and 8, bold indicates the better performance. From the table, we understand that the proposed MKPC clustering outperformed the existing algorithms in Tsunami and Concord datasets in terms of mean and variance. The average performance of the proposed MKPC in terms of clustering accuracy is 66.66% which is high compared with the existing PC, K-medoid and grid clustering. Also, the variance of the proposed MKPC is less as compared with other clustering algorithm. This ensured that the proposed algorithm does not have much change on the performance even though the centroids are initialized randomly. The performance comparison of the concord dataset shows that the proposed MKPC clustering obtained the 64.4% which is higher as compared with other clustering algorithms. While analyzing the

Table 7 Statistical analysis of clustering accuracy on three datasets.

		Tsunami	Concord	Landareas
PC	Mean	53.70	50.65	**88.08**
	Variance	213.63	458.55	556.47
MKPC	Mean	**66.66**	**64.40**	83.14
	Variance	**0**	**86.86**	531.13
K-medoid	Mean	61.35	61.35	61.35
	Variance	28.89	28.89	**28.89**
Grid	Mean	57.92	57.92	57.92
	Variance	134.08	134.08	134.08

Table 8 Pair-wise statistical test of algorithms on clustering accuracy.

	PC	MKPC	K-medoid	Grid
PC	–	1	1	1
MKPC	1	–	**0.025**	**0.025**
K-medoid	1	**0.025**	–	0.1
Grid	1	**0.025**	0.1	–

performance of landareas data, the PC shows good performance in terms of mean value and K-medoid shows the better performance in terms of variance. Overall, the proposed MKPC outperformed in two datasets in terms of mean and variance.

Pair-wise statistical tests (p-test) is widely applied for statistical validation. The advantages of the p-test are that it tests all pair wise differences, is simple to compute, and reduces the probability of making a Type I error. It is also robust with respect to unequal group sample sizes. Pair-wise statistical tests are conducted here to evaluate the algorithmic performance by combining different algorithms. Here, statistical test is conducted on the clustering accuracy values by combining two different algorithms. Table 8 shows the p values of different combinations of algorithms. For the hypothesis testing, p value should be less than 0.1. From the table, we understand that the hybrid model rejects null hypothesis in most of the combinations by reaching the value of 0.025. The table again shows that the statistical test almost always returns lower p values for the proposed hybrid model than for other algorithms and more often rejects the null hypothesis. Overall, it is known that the proposed hybrid model is more likely to reject the null hypothesis.

6. Conclusion

In this paper, we proposed the incremental topological association rule mining of geographical datasets using probabilistic approach. Here, topological relations, such as nearby, disjoint, intersect, inside and outside were considered in the probability-based incremental association rule discovery algorithm, which utilized the probability threshold and relationship threshold to filter the spatial rules. In addition, the mining of the topological spatial association rules is dynamically processed using the probabilistic approach. Finally, the mined topological spatial rules were used for spatial data clustering. For the experimentation, three datasets, such as tsunami, landareas and concord data are utilized and the performance analysis of the proposed method is performed based on the various measures, such as support, confidence and incremental percentage. Also, the spatial clustering proved that the accuracy of the proposed method is 66.66%, 64.40% and 83.14% on tsunami, concord and landareas datasets respectively.

References

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules. In: Proceedings of International Conference of Very Large Datbase, pp. 487–499.

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings for ACM Sigmod International Conference on Management of Data, pp. 207–216.

Ben-Israel, Adi, Iyigun, Cem, 2008. Probabilistic D-clustering. J. Classif. 25 (1), 5–26.

Clementini, Eliseo, Di Felice, Paolino, Koperski, Krzysztof, 2000. Mining multiple-level spatial association rules for objects with a broad boundary. Data Knowl. Eng. 34, 251–270.

Dao, T.H.D., Thill, J-C., 2012. A comprehensive framework for spatial association rule mining. In: The 7th International Conference on Geographic Information Science, Columbus, OH, USA, pp. 18–21.

Ding, W., Eick, C., Wang, J., Yuan, X., 2006. A framework for regional association rule mining in spatial datasets. In: Proceedings of international conference on data mining, pp. 851–856.

Ding, Qin., Ding, Qiang., Perrizo, William., December 2008. PARM—An efficient algorithm to mine association rules from spatial data. IEEE Trans. Syst. Man Cybern. Part B: Cybern. 38 (6), 1513–1524.

Dong, Weishan, Li, Li, Zhou, Changjin, Wang, Yu, Li, Min, Tian, Chunhua, Sun, Wei, 2012. Discovery of Generalized Spatial Association Rules. In: proceedings of International Conference on Service Operations and Logistics, and Informatics, pp. 60–65.

Doraiswamy, Harish, Ferreira, Nivan, Damoulas, Theodoros, et al, 2014. Using topological analysis to support event-guided exploration in urban data. IEEE Trans. Visualization Comput. Graphics 20 (12).

Fang, Gang, Tu, Cheng-sheng, Xiong, Jiang, Wang, Zi-Quan, 2010. Spatial constraint topology association rules mining based on apriori. In: Proceedings of International Conference on Information Engineering and Computer Science, pp. 1–4.

Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J., 1991. Knowledge discovery in databases: an overview. Association for the Advancement of Artificial Intelligence, pp. 1–27.

Gladis, V.P., Rathi, Pushpa, Palani, S., 2015. Brain tumor detection and classification using deep learning classifier on MRI images. Res. J. Appl. Sci. Eng. Technol. 10 (2), 177–187.

Guo, Yi, Gao, Junbin, Feng, L., 2015. Random spatial subspace clustering. Knowl. Based Syst. 74, 106–118.

Han, J., Fu, Y., 1995. Discovery of multiple-level association rules from large databases. In: Proceedings of International Conference on Very Large Database, pp. 420–431.

Jiang, Zhe, Shekhar, S., Zhou, Xun, Knight, J., Corcoran, J., 2015. Focal-test-based spatial decision tree learning. IEEE Trans. Knowl. Data Eng. 27 (6), 1547–1559.

Koperski, Krzysztof, Han, Jiawei, 1995. Discovery of spatial association rules in geographic information databases. Adv. Spatial Databases Lect. Notes Comput. Sci. 951, 47–66.

Krista, R.Z., Borut, Z., 2009. A sweep-line algorithm for spatial clustering. Adv. Eng. Softw. 40, 445–451.

Laube, Patrick, de Berg, Mark, van Kreveld, Marc, 2008. Spatial support and spatial confidence for spatial association rules. Headway Spatial Data Handling Lect. Notes Geoinf. Cartography, 575–593.

Miller, H., 2007. Geographic data mining and knowledge discovery. handbook Geog. Inf. Sci., 352–366

Mohamed, M., Refaat, H., 2011. A fast parallel association rule mining algorithm based on the probability of frequent itemsets. Comput. Sci. Network Secur. 2 (5), 152.

Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., Coello Coello, C.A., 2014. A survey of multiobjective evolutionary algorithms for data mining: Part I. IEEE Trans. Evol. Comput. 18 (1), 4–19.

Pascucci, V., Weber, G., Tierny, J., Bremer, P.-T., Day, M., Bell, J., 2011. Interactive exploration and analysis of large-scale simulations using topology-based data segmentation. IEEE Trans. Visual Comput. Graphics 17 (9), 1307–1324.

Pilevar, A.H., Sukumar, M., 2005. GCHL: a grid-clustering algorithm for high-dimensional very large spatial data bases. Pattern Recogn. Lett. 26, 999–1010.

Qin, D., Qiang, D., William, P., 2008. PARM—An efficient algorithm to mine association rules from spatial data. IEEE Trans. Syst. Man. Cybern. Part B: Cybern. 38 (6), 1513–1524.

Shyu, Chi-Ren, Klaric, Matt, Scott, Grant, Mahamaneerat, Wannapa Kay, 2006. 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.

Worboys, M.F., 1998. Imprecision in finite resolution spatial data. GeoInformatica 2 (3), 257–279.

Wu, Xindong, Zhu, Xingquan, Wu, Gong-Qing, Ding, Wei, 2014. Data mining with big data. IEEE Trans. Knowl. Data Eng. 26 (1), 97–107.

Zaki, M., 2000. Scalable algorithms for association mining. IEEE Trans. Knowl. Data Eng. 12 (3), 372–390.

Zhang, Qiaoping, Couloigner, Isabelle, 2005. A new and efficient K-medoid algorithm for spatial clustering. Lect. Notes Comput. Sci. 3482, 181–189.