Privacy preserving data mining in spatiotemporal databases based on mining negative association rules

K S Ranjith[1], Dr. Geetha Mary A[2].

School of Computing Science and Engineering, VIT University, Vellore 632 014, Tamil Nadu, India
E-mails: ksranjith2000@gmail.com , geethamary@vit.ac.in

Abstract. In the real world, most of the entities are involved with space and time, from any starting point to the end point of the space. The conventional data mining process is extended to the mining knowledge of the spatiotemporal databases. The major knowledge is to mine the association rules in the spatiotemporal databases, the traditional approaches are not sufficient to do mining in the spatiotemporal databases. While mining the association rules the privacy is the main concern. This paper proposed privacy preserved data mining technique for spatiotemporal databases based on the mining negative association rules and cryptography with low storage and communication cost. In the proposed approach first, the partial support for all the distributed sites is calculated and then finally the actual support was calculated to achieve privacy preserve data mining. The mathematical calculation was done and proved that this approach is best for mining association rules for spatiotemporal databases.

Keywords: Data Mining, Association Rules, Privacy-Preserving, Spatiotemporal Databases, Distributed Databases.

1 Introduction

Data Mining is a vast research area where it has applications on almost in every field like irrigation, finance, industry, etc., Data Mining is to identify rules, patterns or functions which occur from a large amount of databases.[1] It has applications like mining association rules, classification, production, clustering. Among them, mining association rules is a trendy research area, specifically; users can extract the frequent patterns based on that can take effective decisions. Nowadays, due to the IoT nodes, the data is distributed.[2] Distributed databases may be horizontal, vertically, hybrid partitioned databases.[3] distributed data mining of large databases may lead to problems like local mining may not be global[5]. Due to the connection of each and every object that surrounds us create a massive amount of data. Mining that kind of distributed data is a major challenge. The data is related to spatiotemporal data. Spatiotemporal data is a data which is related to both space and time. Finding association rules which are interesting may disclose some patterns for selective marketing, forecasting the weather or financial or medical diagnosis, decision supports [4].

Spatiotemporal data has a large space of data that are stored on many computers. Some examples of that kind of data stocks exchanging, environmental data, medical data. Searching is similar patterns in a spatiotemporal database is a very essential in a data mining [6]. The existing distance data mining algorithms aren't fully designed and not implemented for the recent trend of data generating systems like

IoT nodes [7]. Many databases are distributed in real time and many of the algorithms will not suits due to processing power memory sizes [9].

While sharing the data the privacy of two parties is very essential. So, privacy-preserving mining for distributed databases is very essential [10]. The privacy-preserving techniques may classify on the dimensions like data/rule hiding, data mining algorithm, data distribution, privacy presentation [11].Spatial Databases consist of shape, size, distance, position, etc [13]. In Association Rules, Mining Association Rules was founded by satisfying the predefined minimum concept and support given by the database. Temporal data was obtained by monitoring processes and workflows and registering events the spatial data is obtained by Robotics, CAD, GIS, Mobile computing, computer vision etc [16]. The metrics distance and nonmetric shape, direction etc for spatial and the metrics before, after for temporal need to be taken for spatiotemporal databases [17].

The collection of facts, observations, raw data, etc., which is stored in an organized and systemized manner that facilitates to extract the required data as per user requirement. An example of a database is an insurance database. Client details, policy details, etc. Distributed database is a collection of databases which has its own processing unit and can be managed by the distributed database manager. In this environment, the database can be partitioned vertically, horizontally and hybrid mode. In the Spatiotemporal database, the data is related to the attributes space and time. Both the parameters used to describe the state of the object, change of the pattern, related patterns for any object. Extracting the useful information, knowledge, and database is called a data mining. The knowledge can be association rules, clusters, etc. the major technique in data mining are classification and prediction.

1.1 Mining Association Rules:

Let a, b be an item sets which is subset to the item set i_1, i_2, i_3 from distributed $ddb_1, ddb_2 \ldots\ldots ddb_n$. Association rules consist of 2 steps. In the first step, frequent patterns will be generated. In the second step, by using the minimum support and minimum confidence may mine the association rules.

Support(S): (a=>b) = $\sum(aUb)/n$

Confidence (C) :(a=>b) = $(\sum(aUb)/\sum a$

1.2 Mining Association rules in spatiotemporal database:

Spatiotemporal database introduces the concepts task-relevant objects, reference objects, and concept hierarchy. The concept hierarchy is a taxonomy of different concepts from high-level concepts to generalizations on low-level concepts based on the attributes that oriented. The example is shown in the figure. Thus, the spatiotemporal object which is relevant to the task is called the task-relevant object. The main subject of the description is called reference object [13]. The spatiotemporal association rules may be in the form of a→b (t, s%, c %). Here 'a' and 'b' is the set of predicates. S% and c% are support and the confidence and t is the time step. The negative support is denoted as 1-P(A intersection B) in a database, $ddb_1, ddb_2, \ldots\ldots, ddb_n$ the negative confidence can be denoted as 1-P(a/b) where 'a' is satisfied by the member of the ddb_1, when 'b' is satisfied by the same member of ddb_1,. The conjunction of 'm' single spatiotemporal predicates is called m-predicate, m>=1.

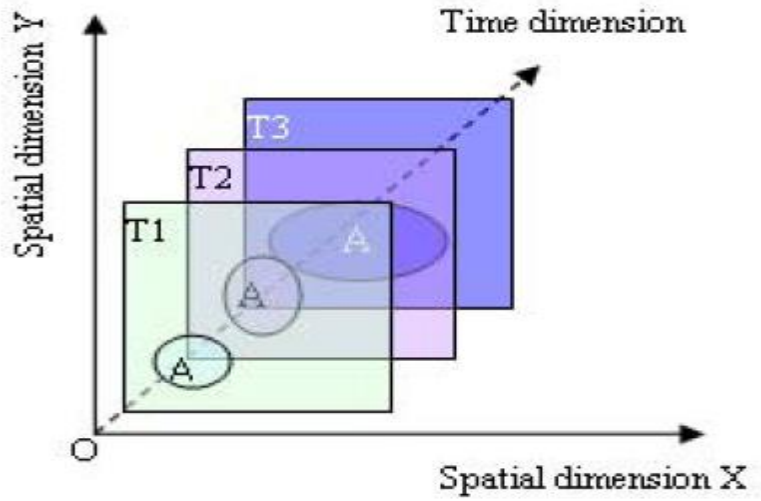

Fig. 1 Spatial-Temporal states and Process [3]

Each element of a spatiotemporal association rule is called a predicate. It is a conjunction of the spatial relations which includes the topological, direction, distance relations and the temporal relations like after, before etc., spatiotemporal association rule may have a different start time but should have the same timestamp. The examples of the spatiotemporal association rule may be 'western specific version warm pool' and rainfall of the south-east of China. The town references object and water resources, tour a resource, traffic facilitates as task-relevant objects are shown below Fig. 2. Privacy-Preserving is an issue that occurs both in centralized and distributed databases. In the centralized database, it will be in single place and multiple users are able to access the same database. Here the privacy-preserving is done to mining the sensitive information from different users whereas in the distributed database the databases will be in different places, the privacy-preserving mining will be global mining where each and every individual site information data need to hide so that every site can be able to access the global results that are useful for the analysis of data based on the user requirement.

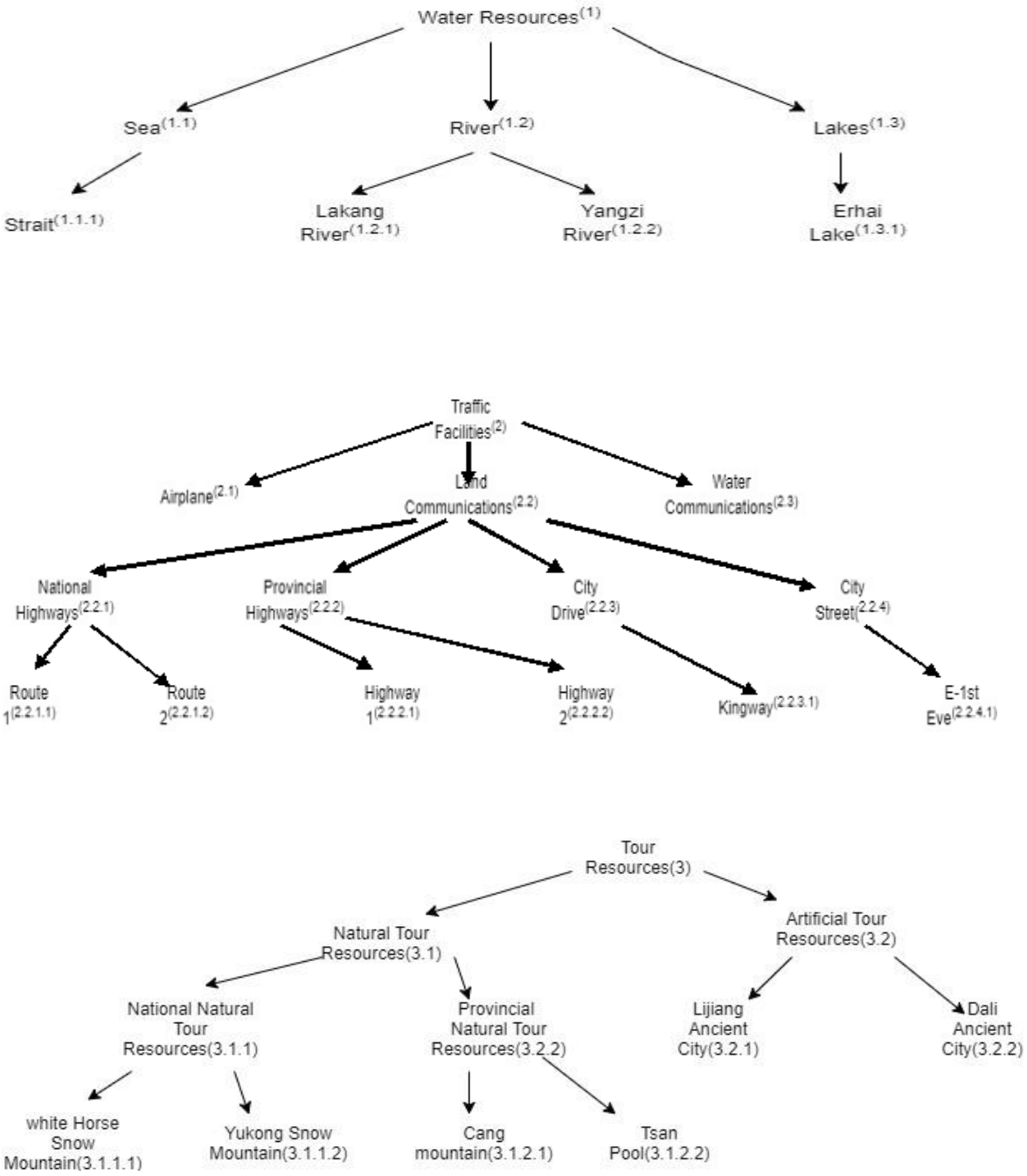

Fig. 2 The town references object and water resources, tour a resource, traffic facilitates as task-relevant objects.

2 Related Work

In [1], presented algorithms that use to build polygons and polylines from a range of ultrasonic data i.e. spatial databases. In [8], presents how to control the data from large databases of spatiotemporal by using the cultural algorithms with evolution programs. The paper [10] presents preserving mining of association rules on a horizontally partitioned data. In [18] Survey of privacy-preserving, data mining was done. In Kantarcioglu, et al, [14] privacy preserving association rules mining problem was investigated on horizontally distributed partitioned data. In [15], Kantarcioglu, et al, proposed the enhanced Kantarcioglu and Clifton's scheme for privacy preserving. Rakesh agarwal,et-al [1], proposed a procedure for privacy-preserving, data mining.The objective of the procedure is to develop a model for aggregated data in data mining without accessing the private information of individual data in the records.by using reconstruction decision tree procedure authors developed the model for aggregated data. In [2],yehuda lindell,preserves a protocol to secure multiparty databases by using the id 3 algorithm for providing privacy preserving data mining In [3], j.mennis demonstrates that how to mine the association rules from spatiotemporal database by considering the case study of U.S.A region on urban growth and the results have been produced that which shows the land cover and socioeconomic changes in the places Denver,U.S.A (1970.1990).

3 Proposed Framework

Due to rapid growth of data generated notes the demand for extracting knowledge in every fields like industries, education, financial sectors, etc is increasing. Its necessary to collect data for all generating nodes then need to store and provide the required patterns depending upon the user perspective. Storing data can be done by different an organization that maintains to store the data. Many data mining techniques are exists to extract the knowledge from different databases. The fundamental goals of data mining techniques are prediction or description. [12] [13] [14]. It has some techniques like clustering, classification, association rules, etc. among them association rules have many applications which is used to generate the relationship between attributes from databases. Association rules can be mined by user specified measures, minimum and minimum support.

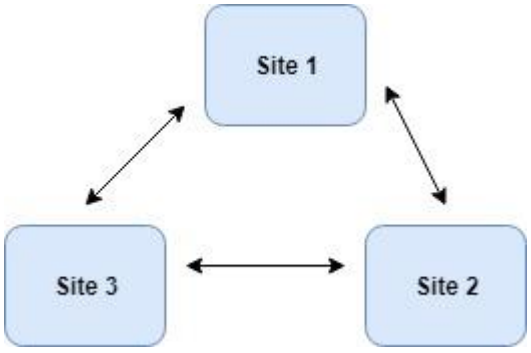

Fig. 3 Communication among Sites and Distributed Mining [2].

To find the association rules, firstly, we find the frequent patterns from various transactions in a database; secondly, we find the association rules by using user-specified support and the confidence. Here the authors considered the different distributed databases like $ddb_{1,}$ $ddb_{2...}$ $ddb_{n.}$

4. Proposed Algorithm:

The process of the proposed algorithm is the item sets i_1, i_2,… i_n.for each sites $ddb_{1,}$ $ddb_{2,...}ddb_n$, the transactions $t_{1,}t_{2,....}$ $T_{n.}$ for site $ddb_{1,}$ the transaction $t_{1,}$ $t_{2,}$t . for site $ddb_{2.....}$ is considered. From the transactions of each and every site, by using the distributed FP growth algorithm (DFPGA) Calculate the frequent patterns. Due to DFPG algorithm, no candidate set was generated and it reduces consumption of memory, number of iterations, duplicates of data, etc when compared to DMA (distributed mining of association rules). After generating the frequent patterns by using the negative support and confident may mine the negative association rules support at every site that monitors the remaining sites which is in the distributed environment assume site is coordinates, the negative association rules support partial support is sent to the coordinator site finally the coordinator site will find actual support and mine the global association rules for the distributed environment.

Algorithm:

Step 1: START
Step 2: Consider the distributed databases $ddb_{1,}$ $ddb_{2, ...}$ $ddb_{n.}$
Step 3: Select the sites from $ddb_{1,}$ $ddb_{2 ...}$ $ddb_{n.}$
Step 4: give unique number for every site and apply cryptography.
Step 5: generate the frequent patterns by DFPGA.
Step 6: Calculate the negative support and confidence by 1-P(A intersection B), 1-P(a/b) .
Step 7: Arrange each site in ring architecture; each site has its own longitude and latitude with respect to the time (given a identification number for every site, that indicates the space and time).
Step 8: calculate the negative partial support.
Step 9: Send negative partial support to all other sites by Ps=1-[x.support-(ddb_1*minimum support) +encrypted identification number].
Step 10: now all the sites send the value to the coordinator (s1)
Step 11: now coordinator site 1 subtract the cryptographic value and calculates the actual support by As=1-($\sum_{i=1}^{n} \neg p$)
Step 12: now site 1 broadcasts the actual support to every site that presents in the distributed environment.
Step 13: END

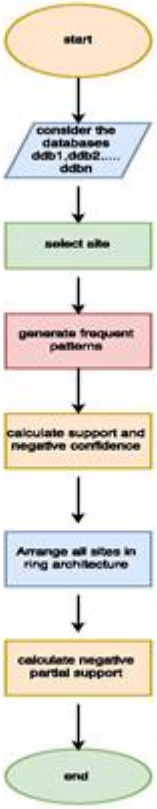

Fig. 4 Flow Chart for Proposed Alogorithm

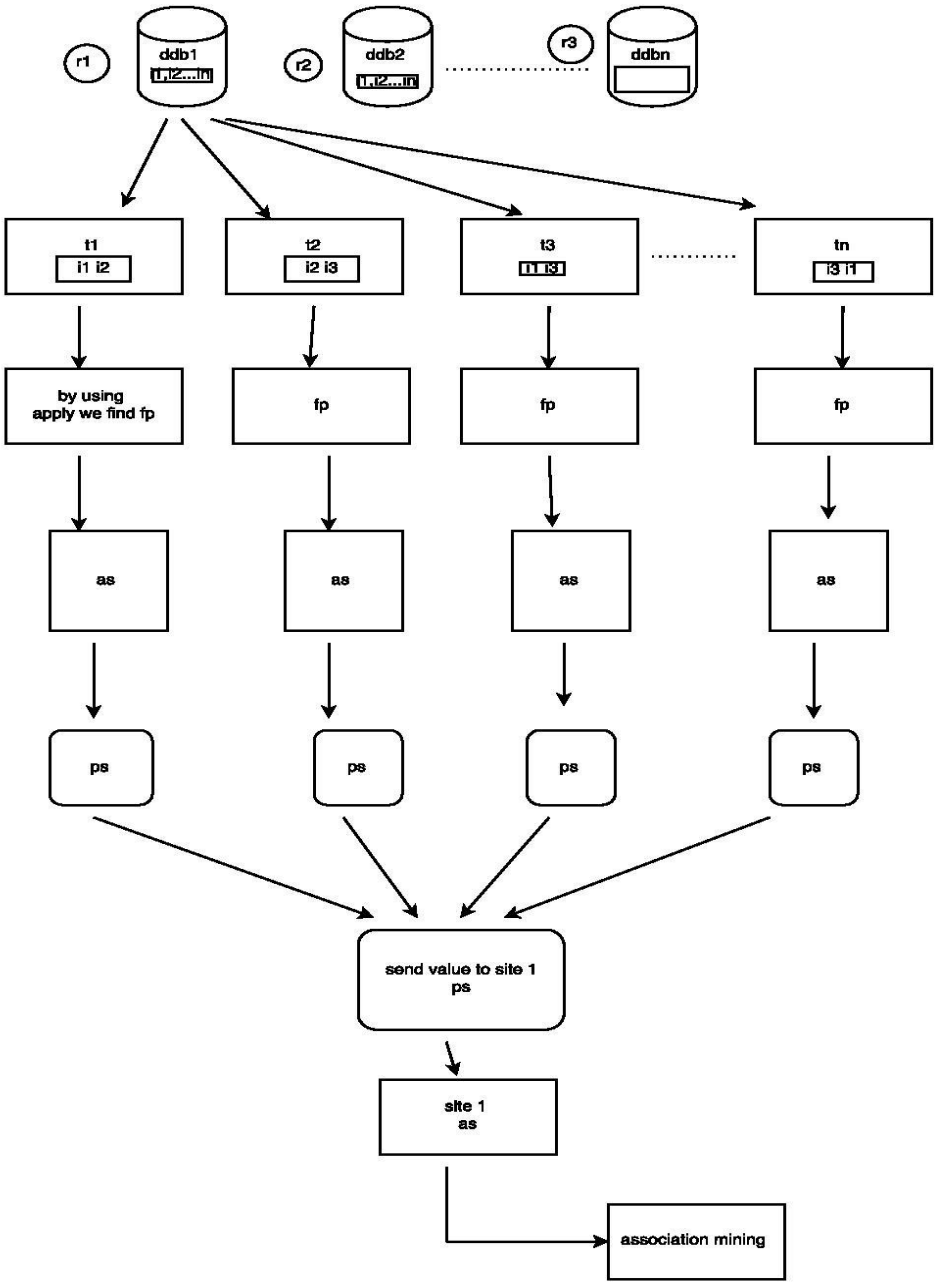

Fig. 7 The Process of Association Mining in Spatiotemporal Databases.

5 Anylitical results

To perform the Analysis Nepal schools dataset was considered, it contains the spatiotemporal data with the differet attributes but with the serial number and the spatiotemporal id (cental-1, East-2, etc….). in this the different spatiotemporal id is taken as the separate site. So for the dataset Nepal has total 6 sites, i.e. six different space and the time ids. Threshold support 40% is considered to analyze the spatio temporal database for all the sites that where each site has a unique number. First each sites calculate the support count using the DFPGA, the negative partial support by using the defined formula, the cryptographic id wad added and then sent to all the other sites with same process in a ring manner to the coordinator site, finally the actual support was calculated and extracted the association rules globally with privacy preserving.

Site1: Negative (-ve) Support

DB1=5.3, Encrypted value=1, Support Count=4/75= 0.05
Support = 1- Support Count= 0.95
Partial Support (Ps) =1-[x.support-(ddb_1*minimum support) +encrypted identification number].
Partial Support (Ps) = 1-[0.95-(5.3*0.4) +1]
Partial Support (Ps) =1-(-0.17)
Partial Support (Ps) =1.17

Site2: Negative (-ve) Support

DB1=25.3, Encrypted value= 2, Support Count=19/75= 0.25
Support = 1- Support Count= 0.75
Partial Support (Ps) =1-[x.support-(ddb_1*minimum support) +encrypted identification number].
Partial Support (Ps) = 1-[0.75-(25.3*0.4) +2]
Partial Support (Ps) =1-(-7.37)
Partial Support (Ps) =8.37

Site3: Negative (-ve) Support

DB1=21.3 , Encrypted value=3 , Support Count= 16/75=0.21
Support = 1- Support Count= 0.79
Partial Support (Ps) =1-[x.support-(ddb_1*minimum support) +encrypted identification number].
Partial Support (Ps) = 1-[0.79-(21.3*0.4)+3]
Partial Support (Ps) =1-(-4.73)
Partial Support (Ps) =5.73

Site4: Negative (-ve) Support

DB1= 12, Encrypted value=4, Support Count= 9/75=0.12
Support = 1- Support Count= 0.88
Partial Support (Ps) =1-[x.support-(ddb_1*minimum support) +encrypted identification number].
Partial Support (Ps) = 1-[0.88-(12*0.4)+4]
Partial Support (Ps) =1-0.08
Partial Support (Ps) =0.92

Site5: Negative (-ve) Support

DB1=16, Encrypted value=5 , Support Count= 12/75=0.16
Support = 1- Support Count= 0.84
Partial Support (Ps) =1-[x.support-(ddb_1*minimum support) +encrypted identification number].
Partial Support (Ps) = 1-[0.84-(16*0.4)+5]
Partial Support (Ps) =1-(-0.56)

Partial Support (Ps) =1.56

Site6: Negative (-ve) Support

DB1= 20, Encrypted value= 6, Support Count= 15/75=0.2
Support = 1- Support Count=0.8
Partial Support (Ps) =1-[x.support-(ddb_1*minimum support) +encrypted identification number].
Partial Support (Ps) = 1-[0.8-(20*0.4)+6]
Partial Support (Ps) =1-(-1.2)
Partial Support (Ps) =2.2

Actual Support (As) =1-($\sum_{i=1}^{n} \neg p$)
As=1-(19.95)
Actual Support (As) = -18.95

6 Conclusion

This paper presents privacy preserving association rule mining in the distributed environment. The cryptographic technique was used for encrypting the identification number of every distributed site for privacy purpose. Locally, all the sites calculate the negative support and confidence, partial support then partial support was sent to the coordinator site. So that coordinator site calculates the actual support and broadcasts to all sites and finds the association rules. The example for distributed environment is spatiotemporal databases. The proposed algorithm will best fit for the spatiotemporal databases to mine association rules.

References

1. Getta, J.R., McKerrow, L. and McKerrow, P.J.: The application of database mining techniques to data fusion in spatial databases. In Proceeding of 1st Australian Data Fusion Symposium, pp. 135-140. IEEE(1996).
2. Sahu, A.K., Kumar, R. and Rahim, N.: Mining Negative Association Rules in Distributed Environment. In *proceedings International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 934-937,. IEEE(2015).
3. Zhang, X., Su, F., Du, Y. and Shi, Y.,: Association rule mining on spatio-temporal processes. In *proceedings 4th International Conference on Wireless Communications, Networking and Mobile Computing*, pp. 1-4. IEEE(2008)..
4. Neerugatti, V. and Reddy, R.M., 2017. A Survey on Secure Connectivity Techniques for Internet of Things Environment. *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)*, *4*(3), (2017).
5. Cheung, D.W., Ng, V.T., Fu, A.W. and Fu, Y.: Efficient mining of association rules in distributed databases. *IEEE Transactions on Knowledge & Data Engineering*, 1(6), 911-922 (1996).
6. Cheung, D.W., Han, J., Ng, V.T., Fu, A.W. and Fu, Y.: A fast distributed algorithm for mining association rules. In *Fourth International Conference on Parallel and Distributed Information Systems*, pp. 31-42. IEEE (1996).
7. Chen, M.S., Han, J. and Yu, P.S.: Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, *8*(6), 866-883 (1996).

8. Neerugatti, V. and Reddy, R.M.: An introduction, reference models, applications, open challenges in Internet of Things. *International Journal of Modern Sciences and Engineering Technology (IJMSET) (2017)*.
9. Abraham, T. and Roddick, J.F.: Survey of spatio-temporal databases. *GeoInformatica*, *3*(1), 61-99 (1999).
10. Wang, C., Huang, H. and Li, H.: A fast distributed mining algorithm for association rules with item constraints. In *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions*, Vol. 3(1), pp. 1900-1905. IEEE (2000).
11. Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y. and Theodoridis, Y.: State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, *33*(1),50-57 (2004).
12. Bertino, E., Fovino, I.N. and Provenza, L.P.: A framework for evaluating privacy preserving data mining algorithms. *Data Mining and Knowledge Discovery*, *11*(2), 121-154 (2005).
13. Bertino, E., Fovino, I.N. and Provenza, L.P.: A framework for evaluating privacy preserving data mining algorithms. *Data Mining and Knowledge Discovery*, *11*(2), pp.121-154 (2005).
14. Wang, L., Xie, K., Chen, T. and Ma, X.: Efficient discovery of multilevel spatial association rules using partitions. *Information and Software Technology*, *47*(13), 829-840 (2005).
15. Gurevich, A. and Gudes, E.: Privacy preserving data mining algorithms without the use of secure computation or perturbation. In *proceeding 10th International Database Engineering and Applications Symposium (IDEAS'06)* , 121-128, IEEE (2006).
16. Chang, C.C., Yeh, J.S. and Li, Y.C.: Privacy-Preserving Mining of Association Rules on Distributed Databases (2006)
17. Kotsiantis, S. and Kanellopoulos, D.: Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 71-82 (2006).
18. Andrienko, G., Malerba, D., May, M. and Teisseire, M.: Mining spatio-temporal data. *Journal of Intelligent Information Systems*, *27*(3),187-190 (2006).
19. Wang, J., Luo, Y., Zhao, Y. and Le, J.: A survey on privacy preserving data mining. In *proceeding First International Workshop on Database Technology and Applications*, 111-114, IEEE (2009).