MINING CO-LOCATION PATTERNS FROM SPATIAL DATA USING RULEBASED APPROACH

G.Priya*, N.Jaisankar and M.Venkatesan

School of Computing Sciences and Engineering, VIT University, Vellore-14, Tamilnadu, India
gpriya@vit.ac.in
njaisankar@vit.ac.in
mvenkatesan@vit.ac.in

Abstract- Co-location pattern is a group of spatial features/events that are frequently co-located in the same region. The co-location pattern discovery process finds the subsets of features frequently located together. Co-location rules are identified by spatial statistics or data mining techniques. A co-location algorithm has been used to discover the co-location patterns which possess an ant monotone property. This algorithm includes a pruning technique to make the item set to get only the most interesting patterns.

Keywords: Co-location patterns, participation index, pruning.

INTRODUCTION

Spatial data mining refers to the extraction of spatial relationships and other interesting patterns not explicitly stored in spatial data sets. It has a wide application in various areas such as geographic information systems, geo-marketing, database exploration, medical imaging, image processing, traffic control, environmental studies. A challenge to spatial data mining is the exploration of efficient spatial data mining techniques due to the large amount of spatial data, and spatial access methods.

Spatial data, like geographic (map) data, very large-scale integration (VLSI) or computer aided design data, and medical or satellite image data contain spatial-related information. Spatial data may be represented in raster format or vector format in which the raster format is consisting of n-dimensional bit maps or pixel maps and the roads, bridges, buildings, and lakes are represented as or overlays of basic geometric constructs, such as points, lines, polygons, and networks formed by these components.

There are two main parts are involved in spatial data mining process they are co-location pattern and the rule. Spatial co-location patterns represent the subsets of Boolean spatial features whose instances are located in close geographic proximity.This process finds the subsets of features frequently located together. For example, the analysis of an ecology data set may reveal symbiotic species. The differences between spatial data mining and classical data mining are mainly related to data input, statistical foundation, output patterns, and computational process.

Generally spatial association rules can be used to mine spatial data for patterns. Spatial association rules require a special reference spatial feature. This is transaction based and transactions are defined around instance of special spatial feature. It uses spatial predicates as item types. But decomposing spatial data into transactions may alter

patterns. So usage of co-location mining increases the efficiency of finding the interesting patterns. In this spatial data is continuous. Rules are generated for point data in space. This uses neighborhood definition and spatial joins.

LITERATURE REVIEW

Co-location patterns represent subsets of Boolean spatial features whose instances are often located in close geographic proximity. The association rules are derived using the Apriori algorithm.

A distance-based approach was proposed called k-neighbouring class sets. In this the number of instances for each pattern is used as the prevalence measure, which does not possess an anti-monotone property by nature. However a non-overlapping instance constraint can be used to get the anti-monotone property for this measure. In contrast an event centric model was developed which does away with the non-overlapping instance constraint. It also defined a prevalence measure called the participation index. This measure possesses the desirable anti-monotone property [1].

Spatial Knowledge Discovery in Databases (SKDD)is the process of identifying valid, novel, useful, and understandable patterns from large spatial datasets. Spatial Data Mining (SDM) is the core of the SKDD process, involving the inferring of algorithms that explore the geo-data develop models and discover significant patterns.[2][3] The 2 –item set 3-item set analysis in which co-location algorithm is more efficient than Apriori algorithm. Association rule-based approaches can be classified into transaction-based approaches and distance-based approaches. Transaction-based approaches focus on defining transactions over space. Apriori-like algorithm can be used for that type.It can be defined by a reference-feature centric model. In this model, transactions are created around instances of one user specified spatial feature. [4]

The strong interactions involving rare spatial features are often marked off in previous methods, since they require frequent co-occurrences of all features in the co-location patterns. Many measures are based on the measures of frequency or minimum participation ratio where rare events are unfavourable. Even though we have a good measure for co-location patterns in the presence of rare spatial features, it is still challenging to find all the patterns efficiently. [5].

A spatial association rule is a rule which describes the implication of one or a set of features by another set of features in spatial databases. For example a rule likes "most big hotels in chennai are close to the marina beach "is a spatial association rule.[6]The association rules are considered interesting if they satisfied both a minimum support threshold and confidence threshold.[7]. Association rule mining finds interesting association or correlation relationships among a large set of data items. [8]

spatial data mining can be categorized based on the kinds of rules to be discovered in spatial databases. A spatial characteristic rule is a general description of a set of spatial related data. Spatial discriminate rule is the general description of the contrasting or discriminating features of a class of spatial related data from other classes [9][10].Existing algorithms used for mining of spatial data are all association rule based. All spatial association rule based algorithms need a special reference spatial feature. This is transaction based where transactions are defined around instance of the special spatial feature. In such algorithms spatial predicates are used as item types. But decomposing the spatial data into transactions has a great probability of altering the patterns.

OVERVIEW OF THE PROPOSED WORK

Mining spatial co-location patterns is an important spatial data mining task. A spatial co-location pattern is a set of spatial features that are frequently located together in spatial proximity. The previous studies on co-location pattern mining emphasize frequent co-occurrences of all the features involved. This marks off some valuable patterns involving rare spatial features. One dominant obstacle is that the maximal participation ratio is not monotonic with respect to co-location pattern containment relation. Thus, the conventional Apriori-like pruning technique cannot be applied. Without proper pruning, there could be many possible combinations.

One method of preventing the loss of rare patterns is to use distance based algorithms for mining the spatial data. These methods increase the efficiency of finding the interesting patterns. In these methods the spatial data is continuous. The rules are generated for point data. Co-location mining is a distance based mining algorithm. Co-location patterns represent subsets of Boolean spatial features whose instances are often located in close geographic proximity.

There is one important observation about co-location patterns with rare spatial features, "even though the participation index of the whole pattern could be low there must be some spatial feature with high participation ratio."So these distances based algorithms have high efficiency of finding the rare and interesting patterns. The

transaction based algorithms use support and confidence for pruning the interesting patterns. The proposed system uses a measure called the participation index. This measure possesses the desirable anti-monotone property. In this paper distance based approach is used to find the co-location patterns from the spatial data. The participation index is used to prune the data to accept only an interesting pattern.
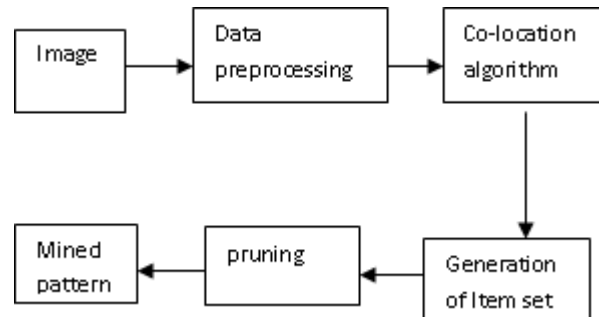
Overview of the Proposed System

Figure1 General Architecture of proposed system

The proposed system consists of a satellite image. The image is processed in MATLAB where the instance is identified by colour identification, and also the coordinates for the instances are retrieved which are stored in a text file. The co-location algorithm is used to generate item sets from those co-ordinates. When an algorithm is applied the co-ordinates are mapped in a grid map. The distance between the instances is calculated. The 2-item sets are calculated by comparing the neighbouring grid spaces and they are pruned if patterns don't have minimum participation index. 3-item sets are calculated by using non –pruned items. Depending upon the participation index an interesting patterns are identified after pruning. The architecture proposed will be feasible for the co-location spatial data mining.

Data Pre-Processing: The data is collected form an image using MATLAB. This raw data is processed using image processing where each item/ objects is differentiated by either colour or shape of it each item instances are given unique instance numbers..Data got from image processing is in rough format. So it should be parsed using either a text parser and convert the raw data into data in specified format which is given below in the tabular column.

Table I Sample Data

X coordinates	Y coordinates	Type
123	443	1
222	657	2
897	892	3
645	432	2
307	692	3
112	334	1

Grid Conversion: If the coordinates which read from data pre-processing is converted into grid coordinates then it will be easy to find the collocation pattern. A specific grid value is taken according to size of image and the coordinate values stored in vectors are converted into grid coordinates and stored in another vector which stores grid value of both X an Y coordinate together.

Applying Co-location Algorithm: After converting the data in to grid value, whole data is stored in double array. The collocation pattern is identified using participation index and pruning index value. The steps for co-location algorithm are given below   Step 1: For each element in first row of array.

Step 2: Compare each and every element in next row.

Step3: Find the difference in grid coordinates.

Step 4: Check if the difference is greater than neighbor.

Step 5: Mark true if they are collocated.

Step 6: End the loop.

Step 7: Calculate participation index by diving Number of true s by total number of instances and initialize pruning index.

Step 8: Compare participation index value and Pruning index and consider only the item sets that are above the pruning index.

Step 9: The items that are pruned out in n-item set calculation are ruled out in n+1- item set calculation.

Step 10: n+1-item set calculation is carried out n same way as n-item set calculation with one more loop than n-item set.

RESULTS AND DISCUSSION

Figure 2 shows the result showing different images showing detected different objects after giving the path of actual image to be processed in MATLAB.
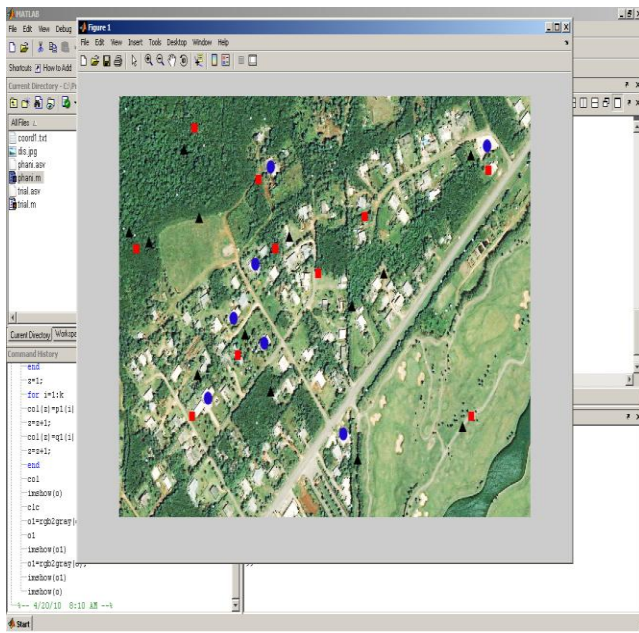

Figure 2 Spatial Image

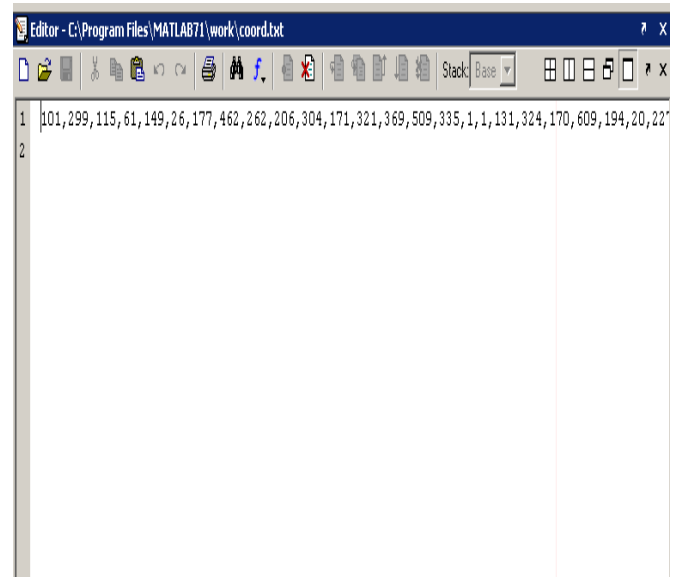Figure 2 shows the x and y coordinates of objects which are used in an image.

Figure 3 Coordinate of Spatial Image

Graphical User Interface (Heading 2)

Figure 5 shows that the output of the MATLAB is browsed and the coordinates are processed in the following screen.
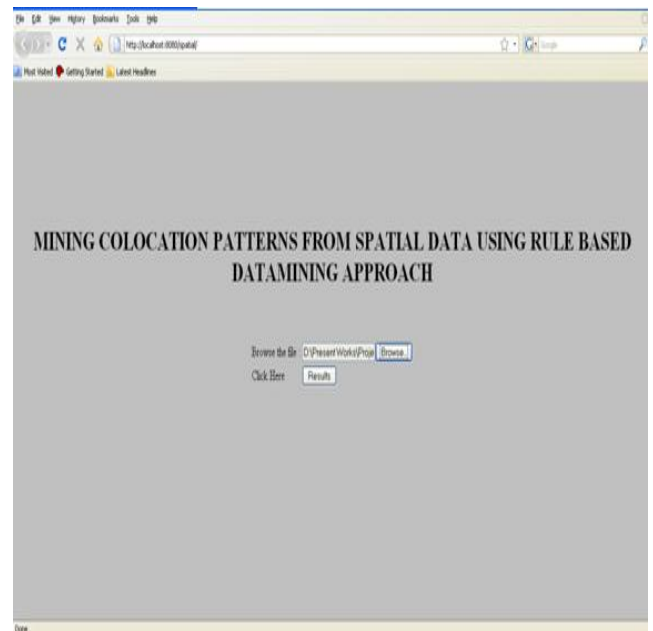

Figure 4 Browse file for user

The figure 5 shows the grid values of different instances of different objects.
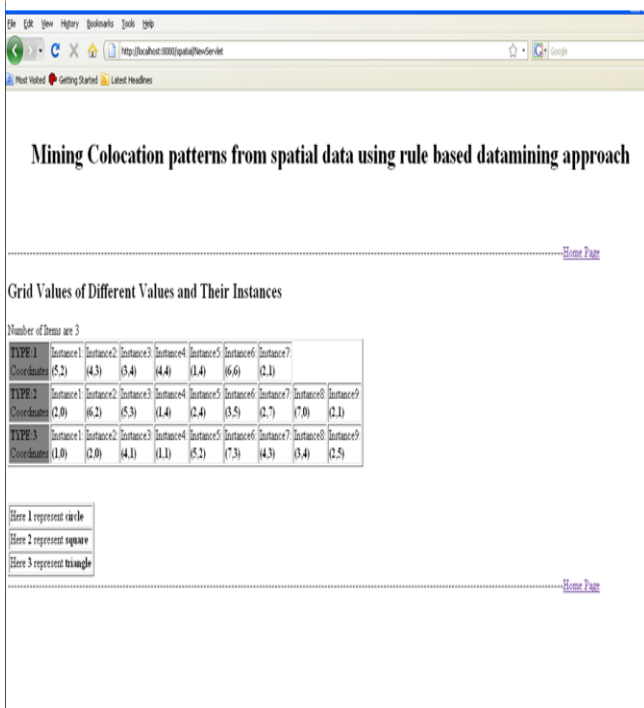
Figure 5 Grid Coordinates

Figure 6 shows the final predictions of the algorithm which represents the various entities which are collocated together. In this paper the coordinates for various objects present in the spatial image is read. Then the coordinate values are converted into readable format. Once it is transformed into grid values our proposed system will find the collocation patterns using collocation algorithm.

Figure 6 Co-location pattern detection

CONCLUSIONS

In this paper, the proposed system formalized the co-location problem and showed the similarities and differences between the collocation rules problem and the classic association rules. The proposed system found that which are the item sets are collocated. Our proposed system plan to examine statistical methods, spatial data types, such as line segments and polygons, extend the co-location mining framework to handle continuous features. If locations of features change over time, it is possible for us to identify some spatiotemporal association patterns.

REFERENCES

[1] Huang, Y., Shekhar, S. and Xiong, H, "Discovering Colocation Patterns from Spatial Data Sets: A General Approach" IEEE Transactions on Knowledge and Data Engineering, 16, (12) pp. 1472-1485, 2004.

[2] He YueShun; Li Xiang;" A Study of Spatial Data Mining Technique Based on Web Management and Service Science", . MASS '09. International Conference , 1 – 4,2009

[3] Tie-li Yang; Ping-Bai; Yu-Sheng Gong;" Spatial Data Mining Features between General Data Mining". Education Technology and Training, 2008. and 2008 International Workshop on Geosciences and Remote Sensing. ETT and GRS541-544,2008:

[4] M.H.Margahny and A.A.Shakour"Scalable Algorithm for Mining Association Rules Mining algorithms "AIML Journal, Volume (6), Issue (3),September, 2006

[5] Huang, Y., Pei, J. and Xiong, H. "Mining Co-location Patterns with Rare Events from Spatial Data Sets", Geoinformatica 10:239-260.2006

[6] K.Koperski and J.Han, "Discovery of Spatial Association Rules in Geographic Information Databases," Proc.FourthInt'l Symp. Spatial Databases, 1995.

[7] C. Gyorodi, R. Gyorodi. "Mining Association Rules in Large Databases". Proc.of Oradea EMES'02: 45-50, Oradea, Romania, 2002.

[8] J. Han, M. Kamber, "Data Mining Concepts and Techniques", MorganKaufmann Publishers, San Francisco, USA, 2001, ISBN 1558604898.

[9] W. Lu.J. Han and B. C. Ooi." Discovery of General Knowledge in Large Spatial Databases". In Proc Far East Workshop on Geographic Information Systems pp.275-279 Singapore. June 1993

[10] R. Ng and J. Han "Efficient and effective clustering method for spatial data mining." In Proc Int. Conf. VLDB. Pp-144-155, Santiago. Chile, Sept 1994.