ABHIYANTRIKI
An International Journal of Engineering & Technology
(AIJET)
Vol. 2, No. 1 (January, 2015) http://www.aijet.in/ eISSN: 2394-627X

Co-Location and Segregation Pattern Mining by Means of Statistical Importance

Amrita A. Kulkarni*
Department of CSE, GHRAET, Nagpur
RTM Nagpur University, Nagpur, India

Deepak Kapgate
Department of CSE, GHRAET, Nagpur
RTM Nagpur University, Nagpur, India

Abstract

Spatial data mining is the extraction of spatial patterns and knowledge which is extracted from large amount of data. In case of interaction pattern relations between objects gives two patterns which are co-location and segregation. Subsets of features whose instances are co-located together are co-location pattern. Subsets of features whose instances are not commonly co-located together are segregation pattern. The paper focuses on finding on these types of patterns with help of statistical model. Spatial auto-correlation i.e. the correlation among values of a single variable strictly referable to their relatively close location is considered. Under null hypothesis a threshold is considered if value is greater than assumed threshold then null hypothesis is rejected. Different prevalence measures are compared on real dataset i.e. minPI, maxPI and avgPI.

Keywords*: Spatial Neighbourhood, Spatial Data Mining, Co-Location Segregation, Statistical Importance.*

****Author for correspondence*** amritakulkarni04@gmail.com

1. Introduction

Large amounts of data has been collected and stored in large data bases by database technologies and data collection techniques. For some applications only a small amount of the data in the databases is needed. This data is called knowledge or information. Data mining is the process of extracting knowledge from these large databases. Data mining is also called knowledge discovery in databases or KDD process. Data mining is concepts and techniques for revealing interesting data patterns hidden in huge data sets. Data mining is developed in a many field of study, including database technology, machine learning, artificial intelligence, neural network, information retrieval etc. Data mining should be applicable to the various kinds of data and databases used in many applications which include relational databases, transactional databases, data warehouses, object- oriented databases, and special application oriented databases such as spatial databases, temporal databases, multimedia databases, and time-series databases.

Spatial data, which means data related to space. As the ability to capture and store information expands, spatial context has emerged as an increasingly important part of discovering knowledge in large amounts of data. The motivation for knowledge discovery in spatial domain is driven by

the fact that global statistics seldom provide useful insight and that most relationships in spatial datasets are geographically regional. Spatial data mining, can be also said as, spatial mining, is data mining as applied to the spatial databases. Spatial data are the data that have spatial or location component, and they show the information, which is more complex than classical data. Spatial data types, spatial relationships are stored into spatial database. Spatial data mining includes various tasks. These include spatial classification, spatial association rule mining, spatial clustering, characteristic rules and trend detection etc.

Interaction pattern mining can lead to important domain related insights in areas such as ecology, biology, epidemiology, earth science, and transportation. Pattern mining is data mining method that find existing pattern in data. In spatial domains, interaction between features generates two types of interaction patterns: co-location and segregation patterns. Co-location pattern can be defined as subset of objects which are frequently located together in closed neighbourhood. Mining of spatial co-location patterns problem can be related to various application domains. E.g. in location based services, various services are requested by service subscribers from their mobile PDA' with location devices such as GPS. Some type of services may be requested in proximate geographic area such as finding bank ATM's near college campus. Segregation patterns, representing negative interactions, can be defined as subsets of Boolean spatial features whose instances are infrequently seen to be located at close spatial proximity.

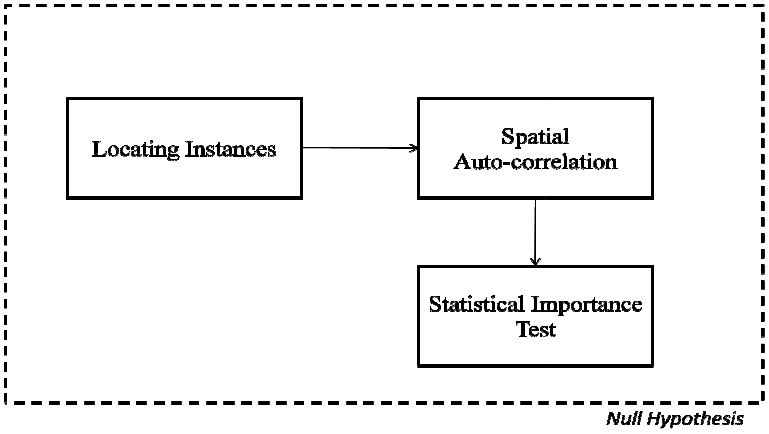

Fig. 1: Block diagram

2. Related Work

Techniques for finding co-location pattern are mostly based on association analysis and spatial statistics. Different approaches in association analysis are classified on its transaction datasets i.e. transaction free and transaction based approaches. In statistical method spatial autocorrelation is taken into account.

a) Data Mining

First type in this approach is transaction free approach [5] which does not produce its transaction type data. Mohammad Akbari et al. [3] transactions free approach that considers fuzzy definition of neighborhood. Regional co-location Pattern Mining algorithm which generates co-location candidates, calculates neighborhood value for each co-location instances based Fuzzy neighborhood. New participation ratio is also developed to find involvement of features. Jin Soung Yoo et al. [8], joinless co-location mining algorithm based on star neighborhood partition model, which separates neighboring objects using star relationship, is used. A partial join co-location mining algorithm modified as coarse filtering scheme introduced to decrease multiple join operation. Clique partition model which separates objects that have clique relationship is used in partial join co-location. Yan Huang et al. [9], focused on finding co-location pattern with

rare event which is missed due to equal participation of features. A maximal participation ratio (maxPR) as measure which shows that if maxPR is high co-location pattern contains high rare events. Yan Huang et al. [11] proposed concept of proximity neighborhood to mine co-location patterns which is transaction free approach. Prevalence-Based Pruning, Multi-resolution pruning techniques is also specified. Hui Xiong et al. [10] transaction free approach to mine extended spatial objects. A buffer based model for co-location pattern discovery used. EXCOM algorithm, will give set of co-location patterns, and apply pruning. Shashi Shekhar et al., [12] proposes concept of user specified neighborhood instead of transactions of specified group of items. Co-location miner algorithm to find co-location patterns is introduced. B. Arunasalam et al. [13] have mined positive and negative patterns in complex spatial relationship, later use statistical technique to determine if these relationships are substantial. To mine complex relationship NP_MaxPI algorithm is introduced.

Second approach is transaction based mining [5]; G. Kiran Kumar et al. [4], classification of instances of features according to its neighborhood has three types event-centric, Reference Feature Centric & window centric model. This paper proposed technique as Hierarchical Window Centric Model in which set of spatial features separated into four windows and window centric model applied on each window independently. Seung Kwan Kim et al. [5] introduces framework which is transaction based approach i.e. generates its own transaction data type, this paper need two algorithms MAximal Clique Enumerator (MACE) algorithm used to find maximal cliques and Apriori algorithm to generate frequent patterns also Generate_Neighboring_Graph algorithm used to find the neighborhood of every spatial object. Venkatesan, M. et al. [6] processes data in form of co-ordinates which generates instances of features. Later by calculating participation index and pruning technique co-location pattern is found. Event centric model focuses on subsets of spatial features probably to occur in a locality around instance. Xiangye Xiao et al. [7] recognised instances of a candidate to obtain its prevalence in the test steps. In order to reduce the computational cost of recognizing these instances, a density based approach is presented. The objects are divided into partitions and identifying instances. A dynamic upper limit of the prevalence for a candidate is maintained. If the current upper limit becomes less than a threshold, stop recognizing its instances in the remaining partitions.

b) Statistical Approach

Sajib Barua et al. [1] proposed method which takes spatial autocorrelation into consideration and extract co-location and segregation pattern. To find these patterns statistical significance test is used. The prevalence measure, participation index (PI), is calculated using null hypothesis and general observed data. Randomization test performed to generate positive and negative values. They also introduced a neighborhood sampling approach using a grid based space partitioning. Jundong Li et al. [2] proposed co-location mining algorithm which indicate statistical significant co-locations in datasets. At first buffer is used to demonstrate affected area near an object, second transaction dataset is generated, and at last statistical significant co-location is identified.

3. Null Hypothesis

Hypothesis testing or significance testing is a method for testing a claim or hypothesis about a parameter in a population, using data measured in a sample. The method of hypothesis testing can be summarized in four steps. We will describe each of these four steps:

Step 1: State the hypotheses: here we assume the spatial autocorrelation among objects.
Step 2: Set the criteria for a decision: the probability (p-value) of obtaining a PI-value at least as extreme as the observed PI-value if the features were spatially independent of each other.
Step 3: Compute the test statistic.
Step 4: Make a decision.

a) Basic Definitions
Participation Ratio:
Consider different objects A, B and C is located in space such that each has no. of instances as A=2, B=3 and C=7

Participation ratio = no. of instances located together/ total no. of instances
here PR w.r.t. A = 2/2
PR w.r.t. B = 2/3
PR w.r.t. C = 3/7

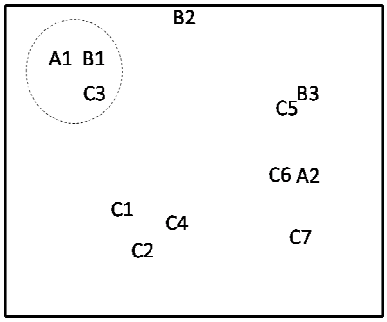

Fig. 2: Participation ratio

Participation Index:
$PI = \min\{PR_A, PR_B, PR_C\}$
However in this paper we have considered values of maxPI and avgPI

b) Spatial Auto-correlation
Spatial autocorrelation is the correlation among values of a single variable strictly attributable to their relatively close location positions on a two-dimensional surface, introducing a deviation from the independent observations assumption of classical statistics. Spatial autocorrelation exists because real-world phenomena are typified by orderliness, pattern, and systematic concentration, rather than randomness.

Nearest Neighbour Distance:
To generate spatially auto-correlated features nearest neighbour distance technique is used. In NND, data sets are divided into six different regions according to their occurrence in that area as north, east, west, south, central and south-central region. A random object is chosen from preferred region and all objects' latitude and longitude locations present in that region are subtracted. Hence, we get list of smallest distance from selected object. The objects which have closest distance from randomly selected object are clustered together. These values are similar and occur one another hence spatially auto-correlated.

c) Statistical Importance Test
Let PI1 denote the participation index of the inclusive data, and let PI2 denote the participation index of data set generated under nearest neighbour distance approach. Then we estimate, using the distribution of PI-values under the null model p-value.

To test importance of p-value obtained it is checked weather P<=0.05. 0.05 is level of significance Level of significance, refers to a criterion of judgment upon which a decision is made regarding the value stated in a null hypothesis. The criterion is based on the probability of

obtaining a statistic measured in a sample if the value stated in the null hypothesis were true. In behavioural science, the criterion or level of significance is typically set at 5%.

When the probability of obtaining a sample mean is less than 5% if the null hypothesis were true, then we reject the value stated in the null hypothesis. For example, If a typical p value = 0.05 is used, there is 5% chance that a spurious co-location or a segregation is reported.

4. Result

As mentioned the proposed work consists of comparison between three different prevalence measure i.e. minPI, maxPI, and avgPI. Different combinations of animal and tree species are detected as co-location and segregation pattern.

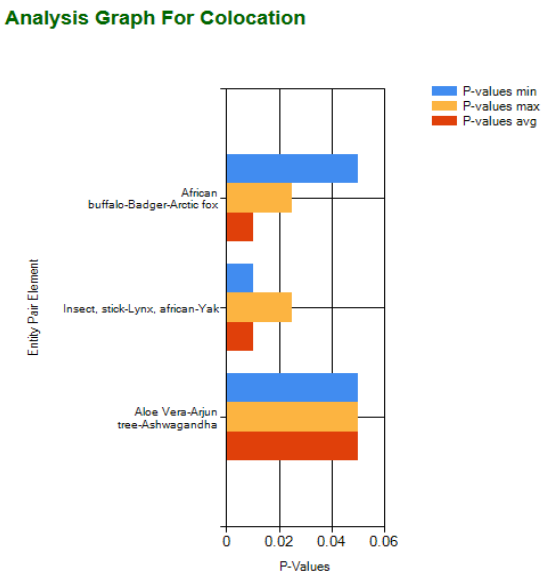

Fig. 3: Analysis Graph for Co-Location

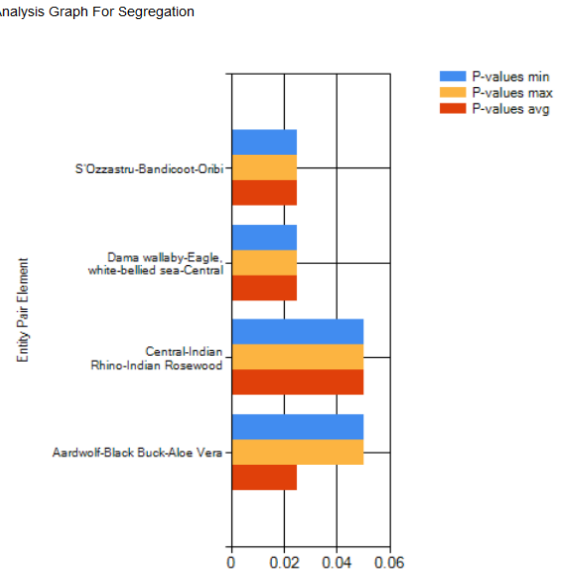

Fig. 4: Analysis Graph for Segregation

5. Conclusion

The paper covers designing an appropriate null model for a statistical test, implementing nearest neighbouring distance for the spatial dependencies measured from the data finding statistically important patterns. The future work may include, using different datasets such as urban features e.g. residences, hospitals, schools, restaurants, crime events e.g. assaults, drunk driving, robberies. Studying and comparing pruning techniques to reduce unwanted data.

References

[1] Sajib barua et al. Mining Statistically Significant Co-Location and Segregation Patterns. IEEE transactions on knowledge and data engineering. 26 (5). May 2014.

[2] Jundong Li et al. Discovering Statistically Significant Co-Location Rules in Datasets with Extended Spatial Objects. DaWaK 2014: 124-135.

[3] Mohammad Akbari et al. New Regional Co-Location Pattern Mining Method Using Fuzzy Definition of Neighborhood. ACSIJ Advances in Computer Science: an International Journal. Vol. 3, Issue 3, No.9 , May 2014.

[4] G. Kiran Kumar et al. Hierarchical Window Centric Method of Modeling Spatial Co-Location Patterns on Spatial Database. IJARCSSE. Volume 3, Issue 11, November 2013.
[5] Seung Kwan Kim et al. A Framework of Spatial Co-Location Pattern Mining for Ubiquitous GIS. Multimedia Tools Application Springer Science + Business Media, LLC 2012.
[6] Venkatesan, M. et al. Event Centric Modeling Approach in Co-location Pattern Analysis From Spatial Data. International Journal of Database Management Systems. Vol.3, No.3, 125-133, August 2011.
[7] X. Xiao et al. Density based co-location pattern discovery. In: Proceedings 16th ACM Int. Adv. GIS, Irvine, CA, USA, 2008, pp. 250–259.
[8] J. S. Yoo et al. A Joinless Approach For Mining Spatial Co-Location Patterns. IEEE transactions on Knowledge & Data Engineering. Vol. 18, No. 10, pp. 1323–1337, Oct. 2006.
[9] Y. Huang, et al. Mining Co-Location Patterns With Rare Events from Spatial Data Sets. Geo-Informatica, Vol. 10, No. 3, pp. 239–260, 2006.
[10] Hui Xiong et al. A Framework For Discovering Co-location Patterns In Data Sets with Extended Spatial Objects. Geoinformatica, 2006.
[11] Y. Huang et al. Discovering Co-location Patterns from Spatial Data Sets: A General Approach. IEEE transactions on Knowledge and Data Engineering. Vol. 16, No. 12, pp. 1472–1485, Dec. 2004.
[12] Shashi Shekhar et al. Discovering Spatial Co-location Patterns: A Summary of Results. Symposium on Large Spatial Databases - SSD, pp. 236-256, 2001.
[13] B. Arunasalam, et al. Striking two birds with one stone: Simultaneous mining of positive and negative spatial patterns. In: Proceeding of 5th SIAM ICDM, 2005, pp. 173–182.
[14] K. Koperski et al. Discovery of spatial association rules in geographic information databases. In: Proceedings 4th International SSD, Portland, ME, USA, 1995, pp.47-66.
[15] Ch. N. Santhosh Kumar et al. Spatial Data Mining using Cluster Analysis. International Journal of Computer Science & Information Technology. Vol. 4, Issue. 4, pp.71-77, 2012.
[16] Sruthi K Surendran et al. Multi-Resolution Pruning Based Co-Location Identification In Spatial Data. IOSR Journal of Computer Engineering. Volume 16, Issue 2, pp 1-5, 2014.