Volume 3, Issue 8, August 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Spatial Data Mining

Dr. A. Padmapriya,
M.C.A., M.Phil., Ph.D.
Department of Computer Science and Engineering, Alagappa University, Karaikudi, INDIA.

N.Subitha,
(M.Phil. Research Scholar)
Department of Computer Science and Engineering, Alagappa University, Karaikudi, INDIA.

Abstract: The research of spatial data is in its infancy stage and there is a need for an accurate method for rule mining. Association rule mining searches for interesting relationships among items in a given data set. This paper enables us to extract pattern from spatial database using k-means algorithm which refers to patterns not explicitly stored in spatial databases. Since spatial Association mining needs to evaluate multiple spatial relationships among a large number of spatial objects, the process could be quite costly. An interesting mining optimization method called progressive refinement can be adopted in spatial association analysis. The method first mines large data sets roughly using a fast algorithm and then improves the quality of mining in a pruned data set. The aim of this research is to analyze the spatial pattern and to check the diffusion of dengue fever cases using k-mean algorithm.

Keywords: spatial data mining, k-mean, spatial relationship, dengue fever.

I. Introduction

The process of KDD is interactive and iterative, involving several steps such as data selection, data reduction, data mining, and the evaluation of the data mining results. The heart of the process, however, is the data mining step which consists of the application of data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data.

While a lot of research has been conducted on knowledge discovery and data mining in relational databases, only a few works deal with knowledge discovery in *spatial* databases for an introduction to spatial databases,Finding implicit regularities, rules or patterns hidden in spatial databases is an important task, e.g. for geo-marketing, traffic control or environmental studies. A spatial database contains objects which are characterized by a spatial location and/or extension as well as by several non-spatial attributes. Dengue fever, and especially the life-threatening form - DHF is an infectious mosquito-borne disease that places a heavy burden on public health systems in Malaysia as well as on most of the tropical countries around the world.

Various environmental factors such as rainfall, temperature, living conditions, demography structure

Domestic waste management and population distribution are important in determining the mosquito survival and reproduction Dengue fever (DF) is a mosquito-borne acute febrile viral disease characterized by sudden onset, fever, intense headache, myalgia, loss of appetite, rash and some non-specific signs and syndromes[14].

The female mosquito "Aedes aegypti"was found to be the most efficient vector. The Aedes female becomes infected when she takes the blood meal from an infected person within the viraemic phase of illness. The extrinsic incubation time of the Aedes female is about 8-12 days. After the extrinsic incubation period, the Aedes female is able to transmit the dengue virus to a human through her bite. The incubation within a human takes about three to 14 days (average 4-7 days). The diffusion of dengue depends on the interaction between vector, parasite and human in the natural environment. The dengue virus cannot be transmitted directly from human to human. Effective vector control is the only solution for dengue control and prevention in situations where vaccines are unavailable. With its powerful analysis, modelling and mapping capabilities, GIS systems may serve as a decision-support tool for epidemic investigation, monitoring, simulation, prediction, prevention and resource allocation[2].

II. Motivation of The Research Work

Application of Spatial Data Mining for Agriculture

The concept is applied in the area of agriculture where giving the temperature and the rainfall as the initial spatial data and then by analyzing the agricultural meteorology for the enhancement of crop yields and also reduce the crop losses based on the k-means algorithm [2]. The data objects can be considered of any dimensions, for the simplicity here we have considered two dimensions. The data objects are clustered or grouped based on the principle of maximizing intra class similarity and minimizing interclass similarity. Each cluster can be viewed as class of objects from which rules can be derived.

International Journal of Computer Applications (0975 – 8887) Volume 15– No.2, February 2011.

Spatial clustering in geographical information systems

Clustering is the task of grouping the objects of a database into meaningful subclasses (that is, clusters) so that the members of a cluster are as similar as possible whereas the members of differ-end clusters differ as much as possible

from each other. Applications of clustering in spatial data-bases are, e.g., the detection of seismic faults by grouping the entries of an earthquake catalog or the creation of thematic maps in geographic information systems by clustering feature vectors. We can support clustering algorithms by our database primitives if the clustering algorithm is based on a "local" cluster condition, i.e. if it constructs clusters by analyzing a restricted neighborhood of the objects. Examples are the density-based clustering algorithm DBSCAN (Ester et al., 1996) as well as its generalized version GDBSCAN (Sander et al., 1998).

CLARANS: A Method for Clustering Objects for Spatial Data Mining

Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. To this end, this paper has three main contributions. First, we propose a new clustering method called CLARANS, whose aim is to identify spatial structures that may be present in the data. Experimental results indicate that, when compared with existing clustering methods, CLARANS is very efficient and effective. Second, we investigate how CLARANS can handle not only points objects, but also polygon objects efficiently. One of the methods considered, called the IR approximation, is very efficient in clustering convex and non-convex polygon objects. Third, building on top of CLARANS, we develop two spatial data mining algorithms that aim to discover relationships between spatial and nonspatial attributes. Both algorithms can discover knowledge that is difficult to find with existing spatial data mining algorithms.

IEEE Transactions On Knowledge And Data Engineering, Vol. 14, No. 5, September/October 2002

Spatial Data Mining using Cluster Analysis

The main objective of the spatial data mining is to discover hidden complex knowledge from spatial and not spatial data despite of their huge amount and the complexity of spatial relationships computing. However, the spatial data mining methods are still an extension of those used in conventional data mining. Spatial data is a highly demanding field because huge amounts of spatial data have been collected in various applications, ranging from remote sensing, to geographical information systems (GIS), computer cartography, environmental assessment and planning, etc.[6]. Spatial data mining tasks include: spatial classification, spatial association rule mining, spatial clustering, characteristic rules, discriminant rules, trend detection. Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. All the members of the cluster has similar features. Members belong to different clusters has dissimilar features. Several clustering methods for spatial data mining include; PAM, CLARA, CLARANS, SD(CLARANS), NSD(CLARANS).

International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 4, August 2012.

III. PROPOSED SYSTEM

The finding of frequent item sets is done by Cluster Analysis. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on the attribute values describing the objects. One such clustering method is partitioning method, in which it creates an initial set of k partitions, where parameter k is the number of partitions to construct; then it uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. It is the most well-known commonly used centroid based technique that takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting intracluster similarity is but intercluster similarity is low. Clustering similarity is measured in regard to the mean value of the objects in cluster which can be viewed cluster's centroid or center of gravity.

IV. ALGORITHM

K-Means Clustering

K-Means Algorithm: The algorithm for partitioning, where each cluster's center is represented by mean value of objects in the cluster.

Input:k: the number of clusters,
D: a data set containing n objects.

Output: A set of k clusters.

Method:

1. arbitrarily choose k objects from D as the initial cluster centers.
2. **repeat**
3. (re)assign each object to the cluster to which the object is most similar, based on the mean value of the objects in the cluster;.
4. update the cluster means,i.e. calculate the mean value of the objects for each cluster;
5. **until** no change;

Algorithm explanation

The k-means algorithm which is used in this paper, randomly selects *k* number of objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and the cluster mean. Then it computes new mean for each cluster. This process iterates until the criterion function converges. The algorithm attempts to determine k partitions that minimize the squared error functions. The method is relatively scalable and efficient in processing large data sets because the computational complexity of the algorithm is O (nkt), where n is the total number of objects, k is the number of clusters, and t is the number of iterations [11]. The method often terminates at the local optimum. K-means is the most

well-known commonly used centroid based technique that takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting intra cluster similarity is but inters cluster similarity is low. Clustering similarity is measured in regard to the mean value of the objects in cluster which can be viewed cluster's centroid or center of gravity.

K-Means Clustering – Example

For example, if our class (decision) attribute is Fever *Type* and its values are: Dengue, bacteria, etc. - these will be the classes. They will be represented by cluster1, cluster2, etc. However, the class information is never provided to the algorithm. The class information can be used later on, to evaluate how accurately the algorithm classified the objects.
Example 1.

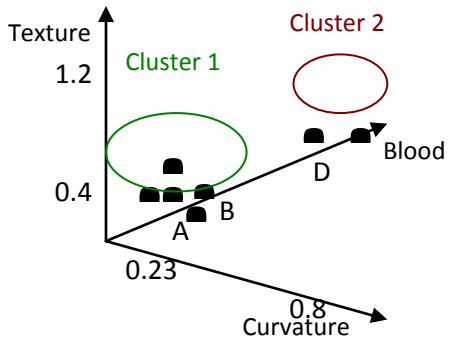
The way we do that, is by plotting the objects from the database into space. Each attribute is one dimension:
After all the objects are plotted, we will calculate the distance between them, and the ones that are close to each other – we will group them together, i.e. place them in the same cluster.

	Curvature	**Texture**	**Blood Consumption**	**Fever Type**
x1	0.8	1.2	A	Dengue
x2	0.75	1.4	B	Dengue
x3	0.23	0.4	D	Bacteria
x4 . .	0.23	0.5	D	Bacteria

Problem: Cluster the following eight points (with (x, y) representing locations) into three clusters A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9). Initial cluster centers are: A1 (2, 10), A4 (5, 8) and A7 (1, 2). The distance function between two points a=*(x1, y1)* and b=*(x2, y2)* is defined as: $\rho(a, b) = |x2 - x1| + |y2 - y1|$. Use k-means algorithm to find the three cluster centers after the second iteration.

Solution:

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2,10)				
A2	(2, 5)				
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				



First we list all points in the first column of the table above. The initial cluster centers – means, are (2, 10), (5, 8) and (1, 2) - chosen randomly. Next, we will calculate the distance from the first point (2, 10) to each of the three means, by using the distance function:

point mean1
x1, y1 x2, y2
(2, 10) (2, 10)
ρ(a, b) = |x2 – x1| + |y2 – y1|
ρ(point, mean1) = |x2 – x1| + |y2 – y1|
= |2 – 2| + |10 – 10|
= 0 + 0
= 0

point mean2
x1, y1 x2, y2
(2, 10) (5, 8)
ρ(a, b) = |x2 – x1| + |y2 – y1|
ρ(point, mean2) = |*x2 – x1*| + |*y2 – y1*|
= |5 – 2| + |8 – 10|
= 3 + 2
= 5

point mean3
x1, y1 *x2, y2*
(2, 10) (1, 2)
ρ(a, b) = |*x2 – x1*| + |*y2 – y1*|

ρ(point, mean2) = |*x2 – x1*| + |*y2 – y1*|
= |1 – 2| + |2 – 10|
= 1 + 8
= 9

Initial Iteration Table(1).

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2,10)	0	5	9	1
A2	(2, 5)				
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				

So, which cluster should the point (2, 10) be placed in? The one, where the point has the shortest distance to the mean – that is mean 1 (cluster 1), since the distance is 0.
Cluster 1 Cluster 2 Cluster 3
(2, 10)
So, we go to the second point (2, 5) and we will calculate the distance to each of the three means, by using the distance function:
point mean1
x1, y1 *x2, y2*
(2, 5) (2, 10)
ρ(a, b) = |*x2 – x1*| + |*y2 – y1*|
ρ(point, mean1) = |*x2 – x1*| + |*y2 – y1*|
= |2 – 2| + |10 – 5|
= 0 + 5
= 5
point mean2
x1, y1 *x2, y2*
(2, 5) (5, 8)
ρ(a, b) = |*x2 – x1*| + |*y2 – y1*|
ρ(point, mean2) = |*x2 – x1*| + |*y2 – y1*|
= |5 – 2| + |8 – 5|

= 3 + 3
= 6

point mean3
x1, y1 *x2, y2*
(2, 5) (1, 2)

$\rho(a, b) = |x2 - x1| + |y2 - y1|$
$\rho(point, mean2) = |x2 - x1| + |y2 - y1|$
$= |1 - 2| + |2 - 5|$
$= 1 + 3$
$= 4$

So, we fill in these values in the table:
Iteration 1 Table(2).

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2,10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				

So, which cluster should the point (2, 5) be placed in? The one, where the point has the shortest distance to the mean – that is mean 3 (cluster 3), since the distance is 0.

Cluster 1 Cluster 2 Cluster 3
(2, 10) (2, 5)

Analogically, we fill in the rest of the table, and place each point in one of the clusters:

Iteration 1(Final Table)

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)	12	7	9	2
A4	(5, 8)	5	0	10	2
A5	(7, 5)	10	5	9	2
A6	(6, 4)	10	5	7	2
A7	(1, 2)	9	10	0	3
A8	(4, 9)	3	2	10	2

Cluster 1 Cluster 2 Cluster 3
(2, 10) (8, 4) (2, 5)
(5, 8) (1, 2)
(7, 5)

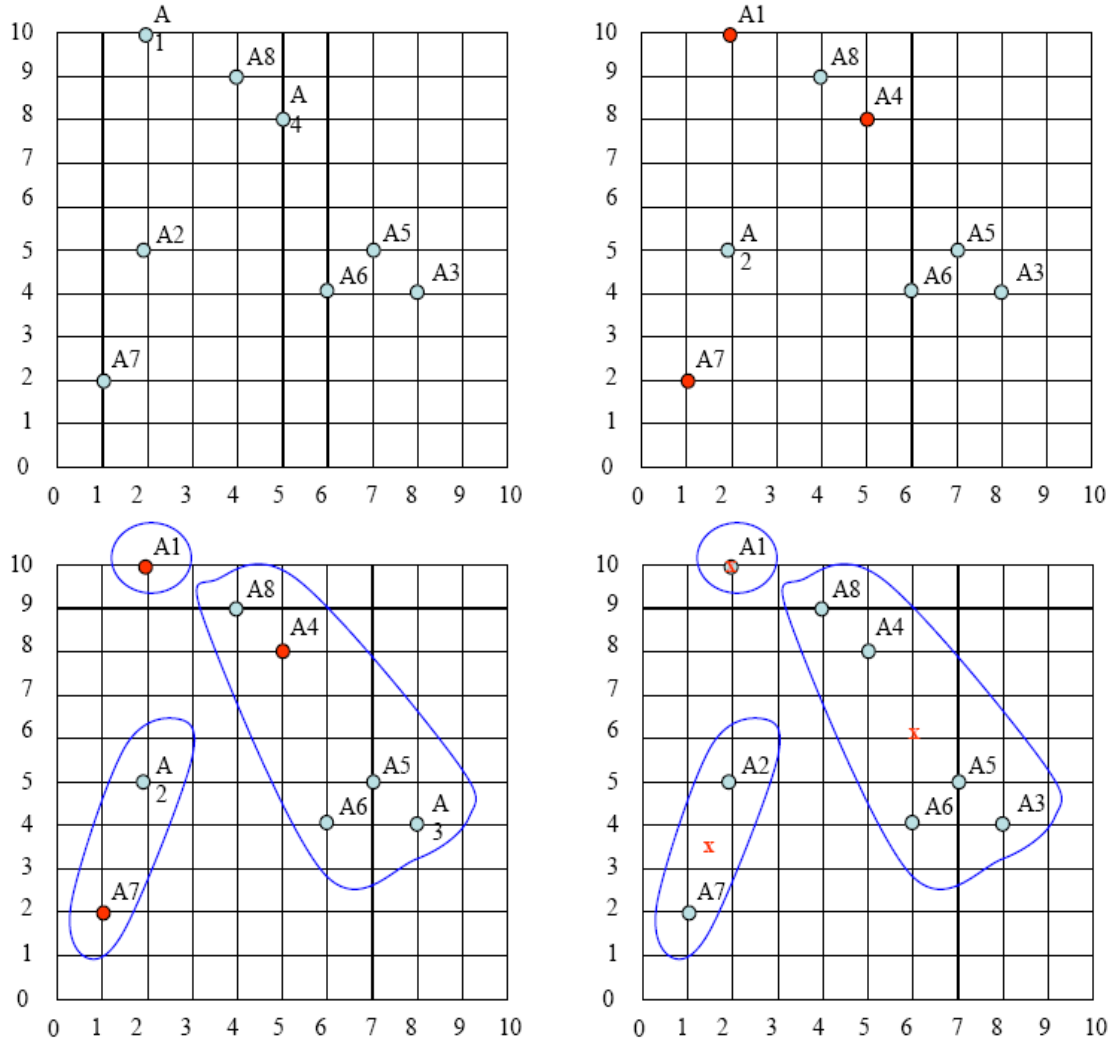
(6, 4)
(4, 9)

Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.
For Cluster 1, we only have one point A1(2, 10), which was the old mean, so the cluster center remains the same.
For Cluster 2, we have ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)
For Cluster 3, we have ((2+1)/2, (5+2)/2) = (1.5, 3.5)

new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

b) centers of the new clusters:
C1= (2, 10), C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6), C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

c)



The initial cluster centers are shown in red dot. The new cluster centers are shown in red x. That was Iteration1 (epoch1). Next, we go to Iteration2 (epoch2), Iteration3, and so on until the means do not change anymore.

V. EXPERIMENTAL STUDY

In this paper we employ data collected from two different columns. 1. Peoples checked by dengue fever. 2. Peoples ected by dengue fever. During Jan 2011 - Mar 2012 201 members were checked out of 70 members were affected.
Then during Jan 2012 – Apr 2013 422 members were checked out of 519 members were affected.

	Jan2011–Mar2012	Apr2012-Mar2013
Checked peoples	201	70
Affected peoples	622	519

VI. CONCLUSION

Clustering is an efficient way of reaching information from raw data and k-means are the basic methods for it. The current method to find the distance between data points and the clusters is Euclidean distance. This method gives circular cluster to identify the dengue patients. In this way we are able to save the computational cost significantly by reducing the number of comparisions with means and also by the least use to Euclidean formula. The results showed that our method can perform clustering operation much faster than the classical ones.

REFERENCES

[1] Chi-Farn Chen; Ching-Yueh Chang; Jiun-Bin Chen "Spatiial knowledge discovery using spatial datamining method",Geoscience and Remote Sensing Symposium, ieee International Volume 8, Issue 25, Page(s): 5602 - 5605 July 2005

[2] Davenhall, B. (2002) Esri ArcUser Online. ArcUser July-September2002: Health geography.http://www.esri.com/news/arcuser/0702/overview.html[accessed 20 November 2003]

[3] Ester M., Frommelt A., Kriegel H.-P., and Sander J. 1998 "Algorithms for Characterization and Trend Detection in Spatial Databases", Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, New York City, NY, pp. 44-50.

[4] [FPM 91] Frawley W.J., Piatetsky-Shapiro G., Matheus J.: "Knowledge Discoveryin Databases: An Overview", in: Knowledge Discovery in Databases,AAAI Press, Menlo Park, 1991, pp. 1-27.Reduction of distance computations

[5] J. Han, Y. Cai, and N. Cercone, "Knowledge Discovery inDatabases: an Attribute-Oriented Approach,"Proc. 18th Conf.Very Large Databases,pp. 547–559, 1992

[6] W. Lu, J. Han, and B. C. Obi. Discovery of General Knowledge in Large Spatial Databases. In Proc. Far East Workshop on Geographic Information Systems pp. 275-289, Singapore, June 1993

[7]. Jiawei Han, Member and Yongjian Fu, Member, "Mining Multiple-Level Association Rules in Large Databases" , ieee transactions on knowledge and data engineering, vol 11, no.5, September/October, 2000.

[8] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis. JohnWiley & Sons, 1990

[9] Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R. and Mostashari, F. (2005) A space-time permutation scan statistics for disease outbreak detection. PLOS Medicine, 2(3): e59.

[10] Lawson, A.B (2001) Statistical methods in spatial epidemiology. An introductory guide to disease mapping. England: John Wiley & Son, Ltd.

[11] J.B Mac Queen in 1967 "Clustering Algorithm for spatial data mining: An Overview"2012.

[12] Sander J., Ester M., Kriegel H.-P., and Xu X. 1998 "Density-Based Clustering in Spatial Databases: A New Algorithm and its Applications", Data Mining and Knowledge Discovery, an International Journal, Kluwer Academic Publishers, Vol.2, No. 2.

