SPATIAL OUTLIER DETECTION TECHNIQES

Nalin Chaudhary[1], Dr. Kalpana Sharma[2], Manoj Kumar [3], Arjit Tomar[4]

[1] *Assistant Prof., CSE, Bhagwant University, Rajasthan, India*
[2] *Assistant Prof., CSE, Bhagwant University, Rajasthan, India*
[3] *Assistant Prof., CSE, Bhagwant University, Rajasthan, India*
[4] *Assistant Prof., CSE, Bhagwant University, Rajasthan, India*

ABSTRACT

Outliers can be defined as observations which appear to be inconsistent with the remainder of the dataset. They deviate too much from other observations. Outlier detection is a data mining technique like classification, clustering, and association rules. Recently, a few studies have been conducted on spatial outlier detection for large datasets.

This paper focuses on the question how Outlier can be detected. There are many known algorithms for detecting outliers, but most of them are not fast enough when the underlying probability distribution is unknown, the size of the data set is large, and the number of dimensions in the space is high. There are, however, applications that need tools for fast detection of outliers in exactly such situations. Planners are concerned about development; it provides a decision support for development process.

Keyword: - *Outlier*, *development*, *detection*, *GIS etc.*

1. INTRODUCTION

Spatial Data Mining Techniques has been used to reveal valuable information from large spatial data sets in many real applications. Spatial objects cannot be simply abstracted as isolated points. Such techniques have been used in many Real life applications like Geographical information system (GIS), Climate prediction, fire detection and etc. These techniques may also be used in real life applications such as to detect less developed region based upon parameters like size, population density, sex ratio, literacy rate and etc.

An important problem that appears often when analyzing data involves identifying irregular or abnormal data points called outliers. This problem broadly arises under two scenarios: when outliers are to be removed from the data before analysis, and when useful information or knowledge can be extracted by the outliers themselves. Outlier Detection in the context of the second scenario is a research that has attracted significant attention in a broad range of useful applications. For example, in credit card transaction data, outliers might indicate potential fraud; in network.

2. OUTLIER DETECTION APPROACHES

The existing approaches to outlier detection can be classified into five categories:

a. Distribution based
b. Depth-based
c. Clustering-based
d. Distance based
e. Density-based

Clustering-based approaches detect outliers as by-products [5]. Some clustering algorithms such as CLARANS, DBSCAN [2] [3], CURE [4] have the capability of handling exceptions. However, since the main objective of the clustering algorithms is to discover clusters, they are not developed to optimize outlier detection.

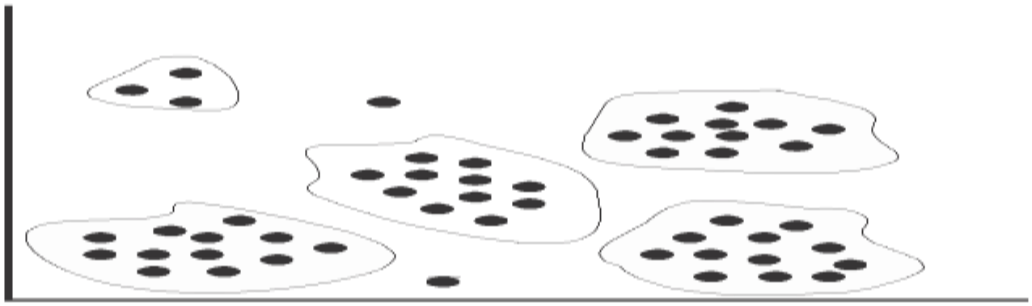

Fig-1: An example of clusters of points

3. PROBLEM FORMULATION

Taking a real data set, as there are several disticts in Haryana, where variation in growth and development is noticed. Taking the inspiration from that, we have collected a dataset of non spatial and spatial attributes for each district. Further SOutlier detection algorithm is applied on the dataset to detect the region which requires more development to be made.

Suppose we have a dataset of n districts
D= (d1, d2,.....dn)
d_i= districts in the state of Haryana or sites which are spatially distributed.
In our problem n=21,
Where d is district with r spatial attributes and m non spatial attributes as given below.

$d_i = (S_1, S_2....S_r, A_1, A_2 .... A_m)$

Where S_1,... S_r is spatial attributes and A_1,.....Am is non spatial attributes.
We want to find out a set of j sites called as spatial outlier sites say SO_j such that $SO_j \in D$ and j<n.
c data, outliersmight represent potential intrusion attempts.

Table - 1: Typical non-spatial parameters

HARYANA	**25353081**	**13505130**	**11847951**	**19.9**	**877**
Ambala	1136784	604044	532740	12.1	882

STATE	**TOTAL POPULATION**				
DISTRICT	**MALE**	**FEMALE**	**0<6**	**GROWTH RATE**	**SEX RATIO**
Bhiwani	1629109	864616	764493	14.3	884
Faridabad	1798954	961532	837422	31.7	871
Fatehabad	941522	494834	446688	16.8	903
Gurgaon	1514085	817274	696811	73.9	853
Hisar	1742815	931535	811280	13.4	871
Jhajjar	956907	514303	442604	8.7	861
Jind	1332042	712254	619788	12.0	870
Kaithal	1072861	570595	502266	13.4	880
Karnal	1506323	798840	707483	18.2	886
Kurukshetra	964231	510370	453861	16.8	889

Mahendragarh	921680	486553	435127	13.4	894
Mewat	1089406	571480	517926	37.9	906
Palwal	1040493	553704	486789	25.5	879
Panchkula	558890	298919	259971	19.3	870
Panipat	1202811	646324	556487	24.3	861
Rewari	896129	472254	423875	17.1	898
Rohtak	1058683	566708	491975	12.6	868
Sirsa	1295114	683242	611872	16.0	896
Sonipat	1480080	798948	681132	15.7	853
Yamunanagar	1214162	646801	567361	16.6	877

Table - 2: Typical non-spatial parameters (cont.)

State/distict	**%age0-6**	**sex ratio**	**LITERACY RATE**		
			total	**male**	**female**
HARYANA	**13.0**	**830**	**76.6**	**85.4**	**66.8**
Ambala	10.9	807	82.9	88.5	76.6
Bhiwani	12.6	831	76.7	87.4	64.8
Faridabad	13.2	842	83.0	89.9	75.2
Fatehabad	12.6	845	69.1	78.1	59.3
Gurgaon	13.1	826	84.4	90.3	77.6
Hisar	12.1	849	73.2	82.8	62.3
Jhajjar	12.1	774	80.8	89.4	71.0
Jind	12.4	835	72.7	82.5	61.6
Kaithal	12.6	821	70.6	79.3	60.7
Karnal	12.9	820	76.4	83.7	68.3
Kurukshetra	12.0	817	76.7	83.5	69.2
Mahendragarh	11.9	778	78.9	91.3	65.3
Mewat	22.3	903	56.1	73.0	37.6
Palwal	16.5	862	70.3	82.6	56.4
Panchkula	11.7	850	83.4	88.6	77.5
Panipat	13.7	833	77.5	85.4	68.2
Rewari	12.5	784	82.2	92.9	70.5
Rohtak	11.9	807	80.4	88.4	71.2
Sirsa	11.9	852	70.4	78.6	61.2
Sonipat	12.7	790	80.8	89.4	70.9
Yamunanagar	11.8	825	78.9	85.1	72.0

Table - 3: Typical Spatial Parameters

District name	Total area	longitude	latitude

4. PROPOSED SYSTEM

The proposed methodology is discussed step by step in figure3 given below.

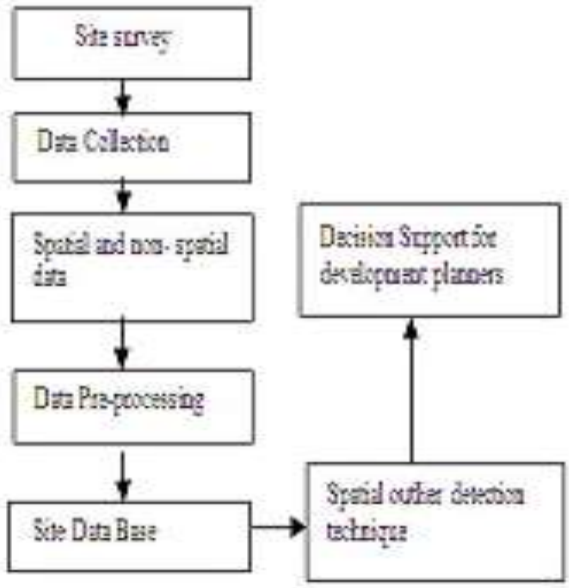

Figure 3. Methodology for spatial outlier detection

Now we are in position to apply a suitable spatial outlier detection algorithm. Many outlier detection algorithms are available. We discuss following two such important algorithms.

4.1 Mean Algorithm

1. Given the spatial data set $X = \{x_1, x_2, x_n\}$, predefined threshold θ, attribute function f, and the number k of nearest neighbours

2. for each fixed j ($1 \leq j \leq q$); standardize the attribute function fj, i.e., $fj(xi) \leftarrow fj(x_i) - \mu fj$

Σfj for i = 1, 2, n.

3. For each spatial point xi, compute the k nearest neighbour set $NNk(x_i)$

4. For each spatial point x_i, compute the neighbourhood function g such that $gj(x_i)$ = average of the data set $\{fj(x): x \rightarrow NNk(xi)\}$, and the comparison function
$h(x_i) = f(x_i) - g(x_i)$.

5. Compute $d2(xi) = (h(xi) - \mu s)\, T\Sigma - 1s\, (h(x_i) - \mu s)$.

If d2 (x_i) > θ, x_i is a spatial outlier w.r.t. A.

4.2 Median Algorithm

1. Given the spatial data set X = {x_1, x_2, x_n}, predefined threshold θ, attribute function f, and the number of nearest neigh

2. for each fixed j ($1 \leq j \leq q$); standardize the attribute function fj, i.e., fj (x_i) ← fj (x_i) −μfj.
Σfj for i = 1, 2, . . . , n.

3. For each spatial point xi, compute the k nearest neighbor set NNk (xi) based on its spatial location.

4. For each spatial point xi, compute the neighborhood function g such that gj (xi) = median of the data set {fj(x): x → NNk (x_i)}, and the comparison function h (xi) = f (x_i) − g (x_i).

5. Compute d2 (x_i) = (h (x_i) − as) TΣ−1s (h (x_i) − μs).
If d2 (xi) > θ, xi is a spatial outlier w.r.t. A.

4. CONCLUSIONS

We have discussed SOutlier detection techniques to solve a real life problem. These SOutlier sites will be very useful for the strategic planners to make efficient decisions regarding development work in a region. The efficiency of the proposed system may also be improved using some other better outlier detection techniques. In future we will implement these techniques to detect outlier sites using real data set.

5. REFERENCES

[1]. M. M. Breunig, H-P. Kriegel, R. Ng, J. Sander, Lof: Identifying Density-Based Local Outliers, ACM SIGMOD Int. Conf. on Management of Data, Dallas, TX; 2000, pg. 93–104
[2]. M. Ester, H-P. Kriegel, J. Sander, X. Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, In: Proceedings of the 2nd Int. Conference on Knowledge Discovery and Data Mining, Portland, OR; 1996
[3]. M. Ester, H-P. Kriegel, J. Sander, X. Xu, Clustering for Mining in Large Spatial Databases, KI Journal (Artificial Intelligence), Special Issue on Data Mining 1998; 12 (1), pg. 18–24.
[4]. S. Guha, R. Rastogi, K. Shim, Cure: An Efficient Clustering Algorithm For Large Databases, In: Proc. Acmsigmod Int. Conf. on Management of Data, Seattle, WA; 1998. pp. 73–84.
[5]. A. Jain, M. Murty, P. Flynn, Data Clustering: A Review, ACM Computing Surveys, 1999, 31(3), pp. 264–323.
[6] T. Johnson, I. Kwok, R. Ng, Fast Computation of 2-Dimensional Depth Contours, In: Proc. 4th. Int.Conf. on KDD, New York, NY, 1998, pp. 224–228.
[7]. E. M. Knorr, R. T. Ng, Algorithms for Mining Distance-Based Outliers in Large Datasets, In: Proc. 24th Int. Conf. Very Large Data Bases, New York, NY; 1998, pp. 392–403.
[8]. E. M. KNORR, R. T. NG, V. TUCAKOV, Distance- Based Outliers: Algorithms and Applications, Journal: Very Large Data Bases, 2000, 8 (3-4), pp. 237–25
[9]. Arvind Sejwal, Application of Spatial Outlier Detection Technique to Detect Ambiguous Sites to Establish an Industry, Department Of Computer Science & Engineering, Ambala College of Engg & Applied Research, Devsthali, Ambala
[10]. www.censusindia.gov.in
[11]. Karmakers A, Syed M. Rahman,"Outlier Detection in Spatial Databases Using Clustering Data Mining", Sixth International Conference on Information Technology: New Generations-2009, DOI 10.1109/ITNG.2009.198, 2009 IEEE explore.pp 1657-1658.

[12]. Ma Yiming et al. "Toward Managing Uncertain Spatial Information for Situational Awareness Applications" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 10, OCTOBER 2008, pp 1408-1423.
[13]. S. Sotoodeh, "Hierarchical Clustered Outlier Detection in Laser Scanner Point Clouds", IAPRS Volume XXXVI, Part 3 / W52, 2007, pp 383-387.
[14]. Thomas Binu and G Raju , "A Novel Fuzzy Clustering Method for Outlier Detection in Data Mining" International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009, pp 161-165.