



ISSN 2278 – 0211 (Online)

Spatial Outlier Detection Approaches and Methods: A Survey

T. Surya

Research Scholar, NGM College, Pollaci, India

B. Azhagusundari

Assistant Professor, NGM College, Pollaci, India

Abstract:

An outlier is any object which is inconsistent with the remaining objects in a database in data mining the outlier detection plays an interesting and important role because the removal of false outliers may affect the mined results to a greater extent if it is important information needed for analysis. The Spatial outliers are locations which are significantly different from their neighborhoods even though they are not much deviated from the entire population. It helps in finding out the local instabilities of objects when compared with other objects in spatial data. The spatial exploration becomes important because it is applied in many applications like weather prediction, clinical traits, geospaced information processing etc. The detection of spatial outliers is necessary for analysis in this area. This paper presents a survey and study of spatial outliers, its approaches, detection methods and algorithms with their complexity along with their pros and cons.

Key words: Outliers, approaches, methods, algorithms

1. Introduction

One of the steps in Knowledge Discovery in Databases (KDD) is Datamining. The Datamining is defined as the process of searching, analyzing and sifting through large amounts of data to find out the relationships, patterns and significant statistical correlation [3]. It is the process of finding hidden patterns in databases by applying data analysis and discovery algorithms with acceptable computational efficiency and limitations. In short it is the process of finding the relevant and useful information from Databases [12]. The SDBS is a Spatial Database Management System for managing huge spatial data which may be point objects or spatially extended in 2D or 3D space or in some high dimensional vector space [7]. Since a large amount of data obtained from Satellite images, X-ray Crystallography and other automatic equipment are kept in this SDBS, the knowledge discovery becomes very important.

The relationships and characteristics existing implicitly in spatial databases and the discovery of such are known as spatial datamining [8]. Examining the spatial data is difficult because it is unrealistic and costly. The aim of spatial datamining is to automate the discovery process for finding the necessary spatial patterns, identifying the relationship between the spatial and non-spatial data and to reorganize. The outlier detection problem arises as an interesting one. An outlier is an observation in dataset which appears to be inconsistent with the remainder of that dataset. Even though the outliers are considered as an error (or) noise they also may carry any important information [12]. Hence it is necessary to identify outliers before modeling and analysis or else it may lead to misspecification models, biased parameter estimation and incorrect results. The outlier detection methods is applied in applications like credit card fraud detection, severe weather prediction, clinical trials, data cleansing, voting irregularity analysis, geographic information system, athlete performance analysis and various other tasks. In spatial mining the outlier detection is the process of identification of items, events or observations which does not fall under expected stability.

2. Outlier Detection Approaches

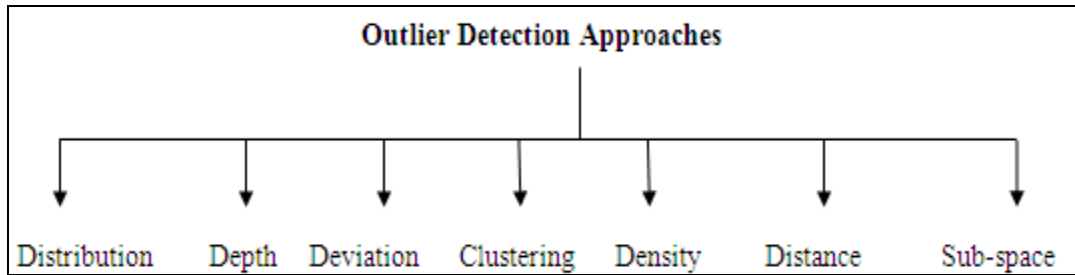


Figure 1

The outlier detection methods [15] are broadly categorized under the outlier detection approaches which includes the following:

- *Distribution-based approach* using standard statistical distributions.
- *Depth-based approach* which binds data object into an multidimensional information space.
- *Distance based approach* which calculates the proportion of database object which is at specified distance from the target object
- *Clustering-based approach* in which outliers are obtained as by-products in the end of clustering. The Distribution based and Depth based comes under the statistical approach. The Distribution method must assume the dataset to own some probability distribution and the Depth based method is not efficient for handling the high dimensional spatial data. Thus the Distance based approach came into existence solving the above problems
- *.Deviation-based approach* in which the characteristics of object are identified and the object which deviates from this is considered as outliers.
- *Density based* which depends on local outlier factor of each point which further depends on the local density of its neighborhood.
- *Sub-space based approach* in which the outliers are detected by observing the density distribution of projections from data. Patterns are used in different sub spaces to define outliers in high dimensional space. The outliers can be efficiently computed if some multi-dimensional index structures are used.

3. Outlier Detection Taxonomy

Broadly the outlier detection methods are classified as follows

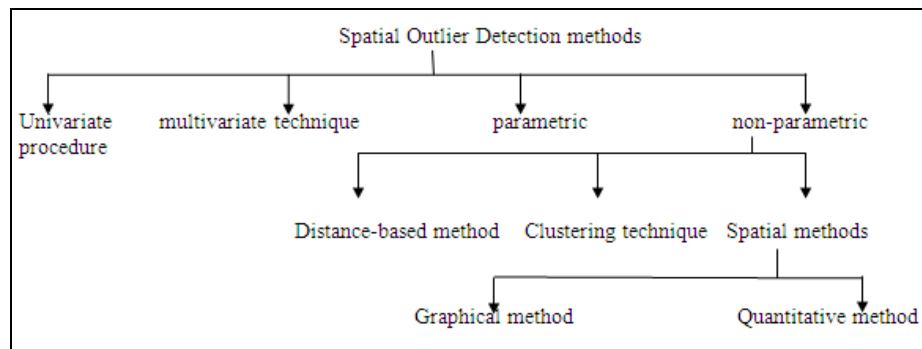


Figure 2

3.1. Univariate Technique

It relies on the assumption of underlying known distribution of data which is assumed to be identically and independently distributed.

3.2. Multivariate Technique

Here, observation alone cannot be detected as outliers and the detection is possible only when analysis is performed and iterations among different variables are compared within the class of data. The dataset with many outliers or outlier clusters undergoes the Masking and Swamping effect: Masking effect-Here an outlier masks the second outlier. In the presence of first outlier the second outlier is not considered as an outlier [15].It is considered as an outlier only after the deletion of first outlier. The masking occurs only when a cluster of outlying observations are skewed for mean and covariance and if the resulting distance of outlying point is small from the mean. Swamping effect: Here, one outlier swamps the second only if second is considered as an outlier in the presence of first. Swamping occurs when outlying observations skew mean and covariance towards it and away from outlying instance and if the distance to the mean is large making them look like outliers.

3.3. Parametric or Statistical procedure

It either assumes a known underlying distribution of observations or based on statistical estimate of unknown distribution parameters. But this is unsuitable for high dimensional datasets and arbitrary datasets without knowledge of underlying

3.4. Non-Parametric Methods

It does not assume any underlying model for data. There are three related classes under this as follows

- *Distance Based Method*: It is based on local distance measures and are capable of handling large databases
- *Clustering Technique Methods*: Here cluster of small size can be considered as clustered outliers. To identify both high and low density patterns a method is proposed where the clustering technique class is further divided into hard classifiers and soft classifiers. The hard classifiers partition the data into 2 non-overlapping sets: outliers and non-outliers. The soft classifiers offers a ranking by assigning each data on outlier classification factor too find the degree of outlyingness.
- *Spatial Methods*: It searches for extreme instabilities with respect to neighbouring values although it may not be significantly different from the entire population. It is closely related to clustering methods. It comes under the bi-partite multidimensional tests. It separates the spatial attributes from non-spatial attributes. The spatial attributes helps to characterize the location, neighborhood and distance. The non-spatial attribute dimensions are used for the comparison of spatially referenced objects with its neighbors. Under this category there are 2 types of tests: 1. *Graphical test*: It is based on the visualization of spatial data which indicates the spatial outliers and 2. *Quantitative test*: It provides a precise test to distinguish the outliers from the remainder of data.

4. Spatial Outlier Detection Algorithms

4.1. CLARANS (Clustering Large Application Based on RANdomised Search)

It is a Clustering based approach. It falls under the partitioning method of clustering algorithms. It helps in identification of spatial structures. CLARANS is found to be very effective and efficient method in spatial datamining. CLARANS can handle both point and polygon objects efficiently It uses mostly the k-medoid partitioning algorithm. It is a main memory clustering technique. The runtime of CLARANS on objects will be less than its runtime on whole database [2]. CLARANS is a bounded and randomized search strategy for improving initial pattern.

Algorithm: CLARANS takes some neighbours dynamically. The clustering process is carried out by searching a graph where every node is a potential solution which is a set of k-medoids. After obtaining a local optimum CLARANS starts with a new node randomly selected for finding next optimum[3]. It has no explicit notion of noise instead it splits clusters if they are relatively large or if closer to some other cluster. CLARANS can help outlier detecting algorithms efficiently by splitting the clusters in large databases.

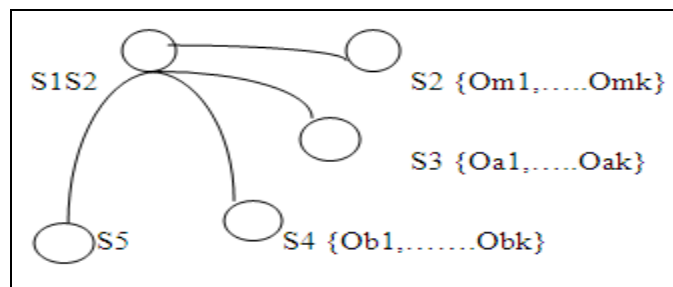


Figure 3: Random Search in CLARANS

4.2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

It falls under the Density based approach. It is a density-based clustering algorithm. DBSCAN identifies clusters of arbitrary shape. It is used for identifying clusters in k-dimensional pointsets. It can be applied for both 2D and 3D Euclidean space as to some high dimensional feature space. The key idea here is at each point of cluster the neighbourhood of given radius has to contain at least a minimum number of points that is the density must exceed some threshold.

Algorithm: The algorithm is based on the notion of *density reachability*. Any point q is *directly density-reachable* from a point p if it is not far away than a given distance ϵ . q is called *density-reachable* from p if there is a sequence p_1, \dots, p_n of points with $p_1 = p$ and $p_n = q$ where each p_{i+1} is directly density-reachable from p_i . q might lie on the edge of a cluster with less neighbors to count as dense itself. This would stop the process of finding a path that stops with the first non-dense point. p and q are density-connected if there is a point o such that both p and q are density-reachable from o . DBSCAN requires two parameters, the threshold value ϵ (eps) and the minimum number of points required to form a cluster (minPts). It starts with an arbitrary starting point. From this point the ϵ -neighborhood is retrieved. It is then analysed and if it contains sufficiently many points, a cluster is started. Otherwise, the point is indicated as noise. This process continues until the density-connected cluster is completely found. Then, again a new unvisited point is retrieved and processed which may lead to new cluster or noise.

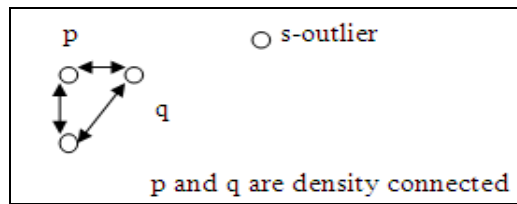


Figure 4

Runtime complexity: $O(n \log n)$ without index structure it is $O(n^2)$. memory: $O(n^2)$

4.3. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

It is also a clustering based approach and falls under the hierarchical clustering with agglomerative method. Birch has the ability to incrementally and dynamically cluster multi-dimensional data in an attempt to produce the best clustering for a given set of resources. Birch requires only a single scan of the database. Birch is the first clustering algorithm proposed in the database area to handle noise.

Algorithm: In a set of N d -dimensional data points, the clustering feature CF of the set is defined as the triple, where LS is the linear sum and SS is the square sum of data points. Clustering features are organized in a CF tree with two parameters: branching factor B and threshold T . Each and every non-leaf node contains at most B entries of the form $[CF_i, child_i]$, where $child_i$ is a pointer to its i th child node and CF_i the clustering feature representing the associated subcluster. Any leaf node contains at most L entries each of the form $[CF_i]$. It also has two pointers prev and next used to chain all leaf nodes together. The tree size depends on the threshold parameter T . A node is required to fit in the size P . P determines B and L and P can be varied. Each entry in a leaf node is not a single data point but a subcluster.

At first it scans all data and builds an initial memory CF tree using the given amount of memory. Next it scans all the leaf entries in the initial CF tree to rebuild a smaller CF tree, while removing outliers and grouping subclusters into larger ones. Any existing clustering algorithm is used to cluster all leaf entries and here the agglomerative hierarchical algorithm is applied to the subclusters. By this a set of clusters is obtained that captures major distribution pattern in the data. That is a point which is too far from its closest seed can be treated as an outlier.

4.4. Index Based Algorithms (for finding all DB (p,D) outliers)

According to this algorithm, let N be the number of objects in dataset T , and let F be the underlying distance function that gives the distance between any pair of objects. For an object 0 , the D -neighbourhood of 0 contains the set of objects $Q \in T$ that are within distance D of 0 . The fraction p is the minimum fraction in T that must be outside the D -neighbourhood of an outlier. let M be the maximum number of objects within the D -neighbourhood of an outlier, With values given for p and D , the problem of finding all DB (p, D)-outliers can be solved by answering a nearest neighbour or range query centred at each object 0 . The range search with fradius D for each object 0 . Once the $(M + 1)$ neighbours are found in the D -neighbourhood, the search stops, and 0 is declared a non-outlier; otherwise, 0 is an outlier. The procedure for finding all DB (p, D)-outliers has a worst case complexity of $O(k N^2)$.

4.5. Moran Scatterplot Method

It is a graphical method which is a plot of normalized attribute value given by $(Z(f(i)) = [(f(i) - \mu_f) / \sigma_f])$ against the neighborhood average of normalized attribute values $(w - z)$ where w is row normalized neighborhood matrix. The graph indicates a spatial association of dissimilar values in the upper left and lower right quadrants that is low value surrounded by high value neighbours (p and q) and high value surrounded by low value(s). Some points surrounded by unusually high or low value neighbor is identified and are treated as spatial outliers.

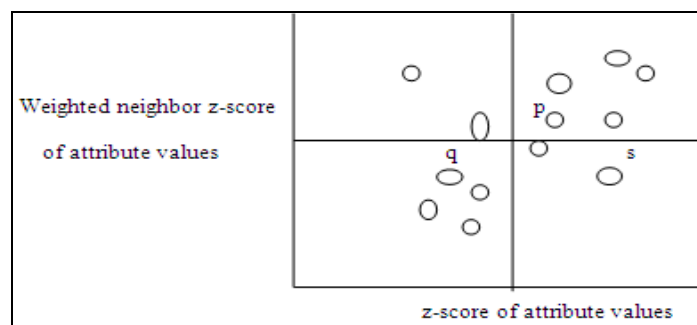


Figure 5

ALGORITHM	OUTLIER DETECTION APPROACH	COMPLEXITY
CLARANS	Clustering -based approach	Linearly proportional to no of object ³
DBSCAN	Density- based approach	$O(n \log n), O(n^2)$
BIRCH	Clustering- based approach	$O(n^2 \log n)$
INDEX BASED METHOD	Distance- based approach	$O(kN^2)$
MORAN SCATTERPLOT	Deviation- based approach	Optimal memory and time complexity

Table 1: Algorithms with Its Approaches and Complexity

ALGORITHM	PROS	CONS	EFFECT ON SPATIAL OUTLIERS
CLARANS	Handles polygonal objects, based on randomized se ach, not affected by increasing dimentionality, no need of distance function.	Min-memory clustering technique, not much efficient due to more i/o operations	Has notion to outliers to some extent but clustering is main work.SDCLARANS-efficient in SDM
DBSCAN	Handles arbitrarily shaped clusters, single-link effect reduced, can be used in databases involving region queries	Distance metric used is not suited for high dimensional data, inefficient to varied density clusters	Robust to outliers in arbitrary shaped clusters
BIRCH	Make full use of available memory to derive sub clusters thereby reducing I/O costs	Does not consider every data point is important and hence does not scan all points currently	Has notion to spatial outliers
INDEX BASED METHOD	Feasible for datasets with many attributes	Needs more memory space for indexing	Scales outliers efficiently than depth-based methods
MORAN SCATTERPLOT	Feasible and efficient finding outliers by plotting graphs	More efficient and has less negative effect	Strongly robust to spatial outliers

Table 2: Algorithms –Pros & Cons and Their Effect on Outliers

5. Conclusion

The Spatial outlier detection is very much essential to find out the noise in dataset. These outliers must be checked whether they are true or false outliers since the elimination of false outliers may affect the final analysis results and the presence of true outliers also makes confusion. In any datamining process the elimination of inconsistent values or outliers itself makes the process easier for further analysis and in spatial datamining it is very much essential since abundant data is involved in the processing. This paper discusses the approaches, methods and some algorithms with their comparisons in spatial outlier detection.

6. References

1. Raymond T.Ng,Jaiwei Han-canada,"CLARANS-A method for clustering objects for spatial data mining".
2. Martin Ester,Hans-Peter Kriegel,Jorg Sander,Xiaowei Xu,"A Density based algorithm for Discovering Clusters in Large Spatial Databases with noise",CiteSeer -1996.
3. Xu, Xiaowei, et al. "Clustering and knowledge discovery in spatial databases."Vistas in Astronomy 41.3 (1997): 397-403.
4. Ester, Martin, et al. "Clustering for mining in large spatial databases." KI 12.1 (1998): 18-24.
5. Parimala, M., Daphne Lopez, and N. C. Senthilkumar. "A survey on density based clustering algorithms for mining large spatial databases.
6. Shekhar, Shashi, and Sanjay Chawla. Spatial databases: a tour. Vol. 2003. Upper Saddle River, NJ: prentice hall, 2003.
7. Koperski, Krzysztof. Spatial data mining. Diss. Simon Fraser University, 1999..
8. Ng, Raymond T., and Jiawei Han. "Efficient and Effective Clustering Methods for Spatial Data Mining." Proc. of. 1994.
9. Geographic data mining and knowledge discovery. Chapman & Hall/CRC, 2009.
10. He, Zengyou, et al. "A fast greedy algorithm for outlier mining." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2006. 567-576.
11. Ester, Martin, Hans-Peter Kriegel, and Jörg Sander. "Spatial data mining: A database approach." Advances in spatial databases. Springer Berlin Heidelberg, 1997
12. Knox, Edwin M., and Raymond T. Ng. "Algorithms for mining distancebased outliers in large datasets." Proceedings of the International Conference on Very Large Data Bases. 1998.

13. A Survey of datamining approaches:Algorithm and Architecture-Arvind Sharma,HS Dat,RK Gupta
14. Ben-Gal, Irad. "Outlier detection." Data Mining and Knowledge Discovery Handbook. Springer US, 2010. 117-130.
15. Shekhar, Shashi, Chang-Tien Lu, and Pusheng Zhang. "A unified approach to detecting spatial outliers." *GeoInformatica* 7.2 (2003): 139-166.
16. Schneider, Christian S. Jensen Markus, and Bernhard Seeger Vassilis J. Tsotras. "Advances in Spatial and Temporal Databases."