# Outlier Detection Algorithms in Data Mining

Jingke Xi
*School of Computer Science and Technology,*
*China University of Mining and Technology, XuZhou, JiangSu, 221008, China*
*xjk@cumt.edu.cn*

## Abstract

*Outlier is defined as an observation that deviates too much from other observations. The identification of outliers can lead to the discovery of useful and meaningful knowledge. Outlier detection has been extensively studied in the past decades. However, most existing research focuses on the algorithm based on special background, compared with outlier detection approach is still rare. This paper mainly discusses and compares approach of different outlier detection from data mining perspective, which can be categorized into two categories: classic outlier approach and spatial outlier approach. The classic outlier approach analyzes outlier based on transaction dataset, which can be grouped into statistical-based approach, distance-based approach, deviation-based approach, density-based approach. The spatial outlier approach analyzes outlier based on spatial dataset that non-spatial and spatial data are significantly different from transaction data, which can be grouped into space-based approach and graph-based approach. Finally, the paper concludes some advances in outlier detection recently.*

## 1. Introduction

Data mining is a process of extracting valid, previously unknown, and ultimately comprehensible information from large datasets and using it for organizational decision making [1]. However, there a lot of problems exist in mining data in large datasets such as data redundancy, the value of attributes is not specific, data is not complete and outlier [2].

Outlier is defined as an observation that deviates too much from other observations that it arouses suspicions that it was generated by a different mechanism from other observations [3]. The identification of outliers can lead to the discovery of useful and meaningful knowledge and has a number of practical applications in areas such as transportation, ecology, public safety, public health, climatology, and location based services. Recently, a few studies have been conducted on outlier detection for large dataset [4]. However, most existing research focuses on the algorithm based on special background, compared with outlier detection approach is still rare. This paper mainly discusses about outlier detection approaches from data mining perspective. The inherent idea is to research and compare achieving mechanism of those approaches to determine which approach is better based on special dataset and different background.

The rest of this paper is organized as follows. Section 2 reviews related work in outlier detection. Section 3 discusses and compares approach of outlier detection which can be categorized into two approaches: classic outlier approach based on transaction dataset and spatial outlier approach based on spatial dataset. The classic outlier approach can be grouped into statistical-based approach, distance-based approach, deviation-based approach, density-based approach. The spatial outlier approach can be grouped into space-based approach and graph-based approach. Recent advances in outlier detection are provided in Section 4. Finally, Section 5 concludes with a summary of those outlier detection algorithms.

## 2. Previous Work

The classic definition of an outlier is due to Hawkins [3] who defines "an outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism".

Most approaches on outlier mining in the early work are based on statistics which use a standard distribution to fit the dataset. Outliers are defined based on the probability distribution. For example, Yamanishi et a1. used a Gaussian mixture model to describe the normal behaviors and each object is given a score on the basis of changes in the model [5].

Knorr et al. proposed a new definition based on the concept of distance, which regard a point p in data set as an outlier with respect to the parameters K and λ, if no more than k points in the data set are at a distance λ or less than p [6].

Arning et a1. proposed a deviation-based method, which identify outliers by inspecting the main characteristics of objects in a dataset and objects that "deviate" from these features arc considered outliers [7].

978-0-7695-3497-8/08 $25.00 © 2008 IEEE
DOI 10.1109/IITA.2008.26

Breunig et al. introduced the concept of local outlier, a kind of density-based outlier, which assigns each data a local outlier factor LOF of being an outlier depending on their neighborhood [2]. The outlier factors can be computed very efficiently only if some multi-dimensional index structures such as R-tree and X-tree [8] are employed. A top-n based local outlier mining algorithm which uses distance bound micro-cluster to estimate the density was presented in [9].

Lazarevic and Kumar proposed a local outlier detection algorithm with a technique called "feature bagging" [10].

Shekhar et al. [11] proposed the definition of spatial outlier: "A spatial outlier is a spatially referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood".

Kou et al. developed spatial weighted outlier detection algorithms which use properties such as center distance and common border length as weight when comparing non-spatial attributes [12]. Adamet al. proposed an algorithm which considers both spatial relationship and semantic relationship among neighbors [13]. Liu and Jezek proposed a method for detecting outliers in an irregularly-distributed spatial data set [14].

## 3. Outlier Detection Approach

Outlier detection has been extensively studied in the past decades and numerous approaches have been developed. These approaches can be mainly classified into two categories: classic outlier approach and spatial outlier approach. The classic outlier approach analyzes outlier based on transaction dataset, which can be grouped into statistical-based approach, distance-based approach, deviation-based approach, density-based approach. The spatial outlier approaches analyze outlier based on spatial dataset, which can be grouped into space-based approach and graph-based approach, as illustrated in Figure 1.

### 3.1. Classic Outlier

Classic outlier approach analyzes outlier based on transaction dataset, which consists of collections of items. A typical example is market basket data, where each transaction is the collection of items purchased by a customer in a single transaction. Such data can also be augmented by additional "items" describing the customer or the context of the transaction. Commonly, transaction data is relative to other data to be simple for the outlier detection. Thus, most outlier approaches are researched on transaction data.

(1) Statistical Approach

Statistical approaches were the earliest algorithms used for outlier detection, which assumes a distribution or probability model for the given data set and then identifies outliers with respect to the model using a discordancy test. In fact, many of the techniques described in both Barnett and Lewis [15] and Rousseeuw and Leroy [16] are single dimensional. However, with the dimensions increasing, it becomes more difficult and inaccurate to make a model for dataset.

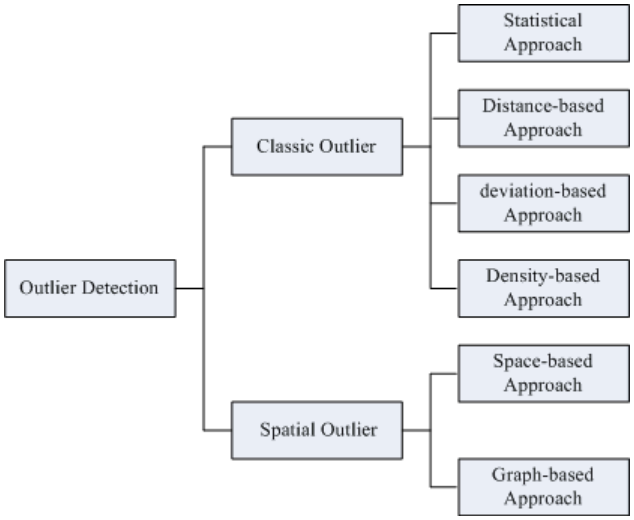


**Figure 1 Outlier Detection Approach**

(2) Distance-based Approach

The concept of distance-based outlier relies on the notion of the neighborhood of a point, typically, the k-nearest neighbors, and has been first introduced by Knorr and Ng [6, 17]. Distance-based outliers are those points for which there are less than k points within the distance in the input data set. This definition does not provide a ranking of outliers and needs to determine an appropriate value of the parameter.

Ramaswamy et al. [18] modified the definition of outlier introduced by Knorr and Ng and consider as outliers the top n point's p whose distance to their k-th nearest neighbor is greatest. To detect outliers, a partition-based algorithm is presented that, first, partitions the input points using a clustering algorithm and, then, prunes those partitions that cannot contain outliers.

The distanced-based approach is effective in rather low dimensions, because of the sparsity of high dimensional points, the approach is sensitive to the parameter λ and it is hard to figure out a-priori. As the dimensions increase, the method's effect and accuracy quickly decline.

(3) Deviation-based Approach

Arning et a1. proposed a deviation-based method, which identify outliers by inspecting the main characteristics of objects in a dataset and objects that "deviate" from these features arc considered outliers [19].

(4) Density-based Approach

The density-based approach estimates the density distribution of the data and identifies outliers as those lying in low-density regions. Breunig et al. [2] assign a

local outlier factor (*LOF*) to each point based on the local density of its neighborhood, which is determined by a user-given minimum number of points (*MinPts*). Papadimitriou et al. [20] present *LOCI* (Local Correlation Integral) which uses statistical values based on the data itself to tackle the issue of choosing values for *MinPts*.

Density-based techniques have the advantage that they can detect outliers that would be missed by techniques with a single, global criterion. However, data is usually sparse in high-dimensional spaces rendering density-based methods problematic.

## 3.2. Spatial Outlier

For spatial data, classic approaches have to be modified because of the qualitative difference between spatial and non-spatial attributes. Spatial dataset could be defined as a collection of spatially referenced objects, such as roads, buildings and cities. Attributes of spatial objects fall into two categories: spatial attributes and non-spatial attributes. The spatial attributes include location, shape and other geometric or topological properties. Non-spatial attributes include length, height, owner, building age and name. A spatial neighborhood of a spatially referenced object is a subset of the spatial data based on the spatial dimension using spatial relationships, e.g., distance and adjacency. Comparisons between spatially referenced objects are based on non-spatial attributes [21].

Spatial outliers are spatially referenced objects whose non-spatial attribute values are significantly different from those of other spatially referenced objects in their spatial neighborhoods. Informally, a spatial outlier is a local instability, or an extreme observation with respect to its neighboring values, even though it may not be significantly different from the entire population. Detecting spatial outliers is useful in many applications of geographic information systems and spatial dataset [11, 21, 22].

The identification of spatial outliers can reveal hidden but valuable information in many applications, For example, it can help locate severe meteorological events, discover highway congestion segments, pinpoint military targets in satellite images, determine potential locations of oil reservoirs, and detect water pollution incidents.

(1) Space-based Approach

Space-based outliers use Euclidean distances to define spatial neighborhoods. Kou et al. developed spatial weighted outlier detection algorithms which use properties such as center distance and common border length as weight when comparing non-spatial attributes [12]. Adamet al. proposed an algorithm which considers both spatial relationship and semantic relationship among neighbors [13]. Liu et al. proposed a method for detecting outliers in an irregularly-distributed spatial data set [14].

(2) Graph-based Approach

Graph-based Approach uses graph connectivity to define spatial neighborhoods. Yufeng Kou et al. proposed a set of graph-based algorithms to identify spatial outliers, which first constructs a graph based on k-nearest neighbor relationship in spatial domain, assigns the non-spatial attribute differences as edge weights, and continuously cuts high- weight edges to identify isolated points or regions that are much dissimilar to their neighboring objects. The algorithms have two major advantages compared with the existing spatial outlier detection methods: accurate in detecting point outliers and capable of identifying region outliers [24].

## 4. Recent Advances in Outlier Detection

Along with the fast development of data mining technique, identification of outliers in large dataset has received more and more attention. Traditional outlier detection methods may not be efficiently applicable to large dataset. So some new methods are specially designed for special background.

(1) High Dimension-based Approach

High dimension space is a difficult problem for outlier detection. According to the criterion of the technique designed for high dimension proposed in the literature [22], a new method ODHDP based on the concept of projection is proposed in this paper, it can well deal with the sparsity of high dimensional points. The basic idea of the approach is to find the outliers by clustering the projections of data set. So, firstly, clustering the projections of data set in each dimension, and putting different weight to each dimension; secondly, selecting the dimension which has the maximum weight in the rest of dimensions for Descartes combination clustering in turn, then pruning the candidate clusters in which the number of the points is less than threshold, until all dimensions are scanned; thirdly, computing the similarity of the points in the remains based on their relationship with the clusters in full dimension, by which the outliers is distinguished from the remains [23].

(2) SVM-based Approach

A SVM-based outlier detection approach was proposed [25]. The method uses several models of varying complexity to detect outliers based on the characteristics of the support vectors obtained from SVM-models. This has the advantage that the decision does not depend on the quality of a single model, which adds to the robustness of the approach. Furthermore, since it is an iterative approach, the most severe outliers are removed first. This allows the models in the next iteration to learn from "cleaner" data and thus reveal outliers that were "masked" in the initial model.

Other outlier detection efforts include Support Vector approach [26], using Replicator Neural Networks (RNNs) [27], and using a relative degree of density with respect only to a few fixed reference points [28].

## 5. Conclusions

This paper mainly discusses about outlier detection approaches from data mining perspective. Firstly, we reviews related work in outlier detection. Next, we discuss and compare algorithms of outlier detection which can be categorized into two categories: classic outlier approach and spatial outlier approach. The classic outlier approach analyzes outlier based on transaction dataset, which can be grouped into statistical-based approach, distance-based approach, deviation-based approach, density-based approach. The spatial outlier approach analyzes outlier based on spatial dataset, which can be grouped into space-based approach, graph-based approach. Thirdly, we conclude some advances in outlier detection recently.

## References


[1] Yu, D., Sheikholeslami, G. and Zang, "A find out: finding outliers in very large datasets". *In Knowledge and Information Systems*, 2002, pp.387-412.

[2] Breunig, M.M., Kriegel, H.P., and Ng, R.T., "LOF: Identifying density-based local outliers." *ACM Conference Proceedings*, 2000, pp. 93-104.

[3] D. M. Hawkins, "Identification of Outliers". *Chapman and Hall*, London, 1980.

[4] Aggarwal, C. C., Yu, S. P., "An effective and efficient algorithm for high-dimensional outlier detection, *The VLDB Journal*, 2005, vol. 14, pp. 211-221.

[5] Yamanishi. K, Takeuchi. J ,and Williams. G On-line, "unsupervised outlier detection using finite mixtures with discounting learning algorithms". *In Proceedings of the Sixth ACM SIGKDDOO*, Boston, MA, USA, pp.320-324.

[6] Knorr, E.M., Ng, R.T., "Finding Intentional Knowledge of Distance-Based Outliers", *Proceedings of the 25th International Conference on Very Large Data Bases*, Edinburgh, Scotland, pp.211-222, September 1999.

[7] Agarwal, D., Phillips, J.M., Venkatasubramanian, "The hunting of the bump: on maximizing statistical discrepancy". *In: Proc. 17th Ann. ACM-SIAM Symp*. on Disc. Alg. pp. 1137–1146 (2006).

[8] Berchtold, S., Keim, D., Kriegel, H.-P, "The X-tree: An efficient and robust access method for points and rectangles". *In: VLDB (1996)*.

[9] Jin, W., Tung, A.K.H., Han, J.W. "Mining Top-n Local Outliers in Large Databases". *In: KDD (2001)*.

[10] Lazarevic, A., Kumar" Feature Bagging for Outlier Detection". *In: KDD (2005)*.

[11] S. C. Shashi Shekhar, "Spatial Databases: A Tour. Prentice Hall", 2003.

[12] Y. Kou, C.-T. Lu, and D. Chen. "Spatial weighted outlier detection". *In Proceedings of the Sixth SIAM International Conference on Data Mining*,pp. 614–618, Bethesda, Maryland, USA, 2006.

[13] N. R. Adam, V. P. Janeja, and V. Atluri., "Neighborhood-based detection of anomalies in high-dimensional spatio-temporal sensor datasets". *In Proceedings of the 2004 ACM symposium on Applied computing*, Nicosia, Cyprus, 2004. pp. 576–583

[14] H. Liu, K. C. Jezek, and M. E. O'Kelly, "Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and gis". *International Journal of Geographical Information Science*, 15(8), 2001. pp.721–741.

[15] Barnett, V. & Lewis, T. (1994).,"Outliers in Statistical Data", *3rd edn*. John Wiley & Sons.

[16] Rousseeuw, P. & Leroy, A. (1996).,"Robust Regression and Outlier Detection", *3rd edn*. John Wiley & Sons.

[17] E. Knorr, R. Ng, and V. Tucakov, "Distance-Based Outlier: Algorithms and Applications," *VLDB* J., vol. 8, nos. 3-4 2000, pp. 237-253.

[18] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets," *Proc. Int'l Conf. Management of Data (SIGMOD '00)*, 2000, pp. 427-438.

[19] A. Arning, R. Agrawal, and P. Raghavan, "A Linear Method for Deviation Detection in Large Databases," Proc. Int'l Conf. Knowledge Discovery and Data Mining, 1996, pp. 164-169.

[20] Papadimitriou, S., Kitawaga, H., Gibbons, P., Faloutsos, C., "LOCI: Fast outlier detection using the local correlation integral", *Proc. of the Int'l Conf. on Data Engineering*, 2003.

[21] Chang-Tien Lu, Dechang Chen,Yufeng Kou, "Detecting spatial outliers with multiple attributes", *Tools with Artificial Intelligence, 2003.* Proceedings. 2003, pp.122–128.

[22] Aggarwal, C.C, Yu, P. "Outlier detection for high dimensional data", *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Santa Barbara, CA, 2001, pp. 37-47.

[23] Ping Guo, Ji-Yong Dai, Yan-Xia Wang, "Outlier Detection in High Dimension Based on Projection", Machine Learning and Cybernetics, 2006 International Conference, 2006,pp.1165 – 1169.

[24] Yufeng Kou, Chang-Tien Lu, Dos Santos, R.F." Spatial Outlier Detection: A Graph-Based Approach", ICTAI 2007, Volume 1, 2007,pp.281 – 288.

[25] E.M. Jordaan. Deployment of Robust Inferential Sensors, "Irtdusrriol application of Supper Vector Machines for Regression", Ph. D. thesis. Eindhaven University of Technology, 2002.

[26] Jordaan, E.M.;,Smits, G.F., " Robust outlier detection using SVM regression", Neural Networks, 2004. Proceedings. 2004,pp.2017 – 2022.

[27] Harkins, S., He, H., Williams, G., Baster, R., "Outlier Detection Using Replicator Neural Networks", *DaWaK'02*, 2002, pp. 170-180.

[28] Pei, Y., Zaiane, O., Gao, Y., "An Efficient Reference based Approach to Outlier Detection in Large Dataset", *IEEE Int'l Conference on Data Mining*, 2006.