Unsupervised Outlier Detection Using Reverse Neighbors Counts

Mr.K.SIVA KUMAR
Asst.Prof,Department of CSE
RVR&JC College of Engineering
Chowdavaram,Guntur,AP.
sivakumarkommerla@gmail.com

Mr.P.RAMAKRISHNA
Asst.Prof,Department of CSE
RVR&JC College of Engineering
Chowdavaram,Guntur,AP.
mails4prk@gmail.com

Mr.G.MANTHRU.NAIK
Asst.Prof,Department of CSE
SRI MITTAPALLI Institute Of Technology for Women
Tummalapalem,Guntur,AP.
manthru1979@gmail.com

***Abstract*—**Due to curse of dimensionality, there are various challenges to detect outliers in high-dimensional data. The distance concentration in high-dimensional data hinders the detection of outliers by making distance-based methods label all points as almost equally good outliers. In this paper, we prove that such a view is quite simple; by showing that distance-based methods produce more conflicting outlier scores in high-dimensions. Moreover, we show that high dimensionality can have a different impact, by reexamining the reverse neighbors in the context of unsupervised outlier-detection. We provide awareness of how some points known as anti-hubs appear occasionally in k-NN lists of other points. We then describe the relation between anti-hubs and outliers. By figuring out the classic k-NN method, the angle-based methods for high-dimensional data and the density-based local outlier factor on numerous synthetic and real-world data sets, we present novel insight into the efficiency of reverse neighbor counts in unsupervised outlier detection.

Keywords— outlier, reverse neighbor counts, anti-hubs, curse of dimensionality.

I. INTRODUCTION

In many data analysis tasks a large number of variables are being recorded or sampled. One of the first steps towards obtaining a coherent analysis is the detection of outlaying observations. Although outliers are often considered as an error or noise, they may carry important information.

Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis.

The definition of an outlier often depends on hidden assumptions regarding the data structure and the applied detection method.

Similarly an outlier as “an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism”.

Outlier detection is a process of identifying these patterns which do not have regular behavior. Detection of outliers is widely used these days. The interest in outliers is strong since they may constitute critical and actionable information in various domains, such as intrusion and fraud detection, and medical diagnosis.

The task of detecting outliers can be classified as supervised, semi-supervised, and unsupervised, depending on the existence of labels for outliers and regular instances. Unsupervised methods are more widely applied as other

categories require accurate and representative labels that are often expensive to obtain. Unsupervised methods include distance-based methods that mainly rely on a distance measure or similarity in order to detect outliers.

Our motivation is based on the following factors:

1) It is essential to understand how the increase of dimensionality impacts outlier detection. The actual challenges posed by the "curse of dimensionality" differ from the commonly accepted view that every point becomes an almost equally good outlier in high-dimensional space. We will present further evidence which challenges this view, motivating the (re)examination of methods.

2) Reverse nearest-neighbor counts have been proposed in the past as a method for expressing outlierness of data points but no insight apart from basic intuition was offered as to why these counts should represent meaningful outlier scores. Recent observations that reverse-neighbor counts are affected by increased dimensionality of data warrant their reexamination for the outlier-detection task. In this light, we will revisit the ODIN method.

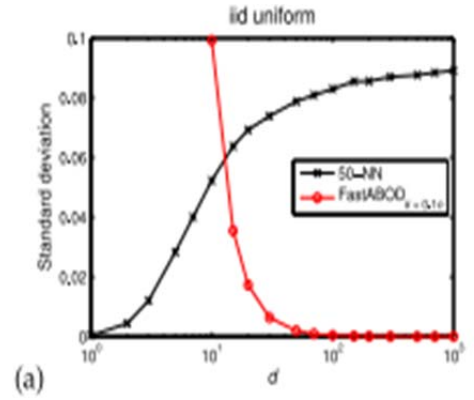
Our contributions can be summarized as follows:

1) We discuss the challenges faced by unsupervised outlier detection in high-dimensional space, in Section 3. Despite the general impression that all points in a high-dimensional data set seem to become outliers [9], we show that unsupervised methods can detect outliers which are more pronounced in high dimensions, under the assumption that all data attributes are meaningful, i.e. not noisy.

2) In Section 4 we consider anti-hubs and how these anti-hubs relate to outlierness of points. We consider low dimensional settings and extend our view to the full range of neighborhood sizes.

3) Basing on the relation between outliers and anti-hubs in both high-dimensional and low-dimensional settings in Section 5, by using k-occurrence we analyze two ways for expressing the outlierness of points, starting with the ODIN method.

4) Lastly, in Section 6 we illustrate our experiments with synthetic and real data sets, demonstrating the benefits of the methods, and the conditions in which the benefits are expected.



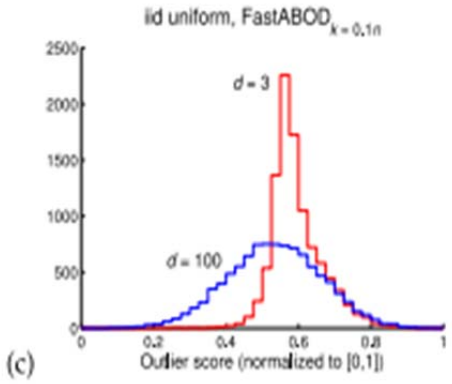


Fig 1: Outlier scores versus dimensionality d for uniformly distributed data in $[0,1]^d$: (a) Standard deviation; (b) Histogram of 50-NN scores; (c) Histogram of normalized ABOD scores.

II. RELATED WORK

The scope of our investigation is to examine:
(1) Unsupervised methods.
(2) Point anomalies, i.e., particular points which can be outliers without considering contextual or collective information
(3) Methods that produce outlier score for each point and detect the outliers basing on these scores.
The most popular and frequently used methods are based mainly on nearest neighbors that assume the outliers are those points which appear far from rest of the points. These methods depend on a distance or similarity measure to find the neighbors. Euclidean distance is the most widely used distance.

III. OUTLIER DETECTION IN HIGH DIMENSIONS

In this section we revisit the commonly accepted view that unsupervised methods detect every point as an almost equally good outlier in high-dimensional space [9]. In [10] this view was challenged by showing that the exact opposite may take place: as dimensionality increases, outliers generated by a different mechanism from the data tend to be detected as more prominent by unsupervised methods, assuming all dimensions carry useful information.

We present an example showing that this can happen even when no true outliers exist, in the sense of originating from a different distribution than other points.

Example 3.1: Let us observe n =10,000 d-dimensional points, whose components are independently drawn from the uniform distribution in range [0,1]. We employ the classic k-NN method [3] (k = 50; similar results are obtained with other values of k). We also examine ABOD [19] (for efficiency reasons we use the Fast ABOD variant with k = 0:1n), and use standard deviation to express the variability in the assigned outlier scores.

Fig. 1a illustrates the standard deviations of outlier scores against dimensionality d. Let us observe the k-NN method first. For small values of d, deviation of scores is close to 0, which means that all points tend to have almost identical outlier scores. This is expected, because for low d values, points that are uniformly distributed in $[0,1]^d$ contain no prominent outliers. This assumption also holds as d increases, i.e., still there should be no prominent outliers in the data. Nevertheless, with increasing dimensionality, for k-NN there is a clear increase of the standard deviation. This increase indicates that some points tend to have significantly smaller or larger outlier scores than others. This can be observed in the histogram of the outlier scores in Fig. 1b, for d=3 and d=100. In the former case, the vast majority of points have very similar scores. The latter case, however, clearly shows the existence of points in the right tails of the distributions which are prominent outliers, as well as points on the opposite end with much smaller scores.
The ABOD method, on the other hand, exhibits a completely different trend in Fig. 1a, with the deviation of its scores quickly diminishing as dimensionality increases, which makes it appear that the method is severely "cursed" by the dimensionality. However, ABOD was specifically designed to take advantage of high dimensionality and shown to be very effective in such settings (cf. Section 6), meaning that the shrinking variability observed in Fig. 1a says little about the expected performance of ABOD, which ultimately depends on the quality of the produced outlier rankings [10]. However, when scores are regularized by logarithmic inversion and linearly normalized to the [0,1] range [25], a trend similar to k-NN can be observed, shown in Fig. 1c.

IV. ANTI-HUBS AND OUTLIERS

In this section, we consider anti-hubs as a special category of points in high-dimensional spaces. We also explore the relation between anti-hubs and outliers detected by unsupervised methods in the context of varying neighborhood size k.

A. Anti-hubs: *Definition and Causes*

The existence of anti-hubs is a direct consequence of high dimensionality when neighborhood size k is small compared to the size of the data.

Distance concentration refers to the tendency of distances in high-dimensional data to become almost indiscernible as dimensionality increases, and is usually expressed through a ratio of a notion of spread and magnitude of the distribution of distances of all points in a data set to some reference point. If this ratio tends to 0 as dimensionality goes to infinity, it is said that distances concentrate.

Let us define the notions of k-occurrences, hubs and anti-hubs.

Definition1 (k-occurrences): Let D be a finite set of n points. For point $x \in D$ and a given distance or similarity measure, the number of k-occurrences, denoted $N_k(x)$, is the number of times x occurs among the k nearest neighbors of all other points in D. Equivalently, $N_k(x)$ is the reverse k-nearest neighbor count of x within D.

Definition 2 (hubs and anti-hub): For $q \in (0,1)$, hubs are the points $x \in D$ with the highest values of $N_k(x)$. For $p \in (0,1)$, $p < 1 - q$, anti-hubs are the points $x \in D$ with the lowest values of $N_k(x)$.

Under widely applicable assumptions, for $k \ll n$, as dimensionality increases the distribution of N_k becomes skewed to the right, with variance increasing, resulting in the emergence of hubs that appear in many more k-NN lists than other points, and conversely anti-hubs that appear in a much lower number of k-NN lists (possibly 0). These extreme cases of hubs and anti-hubs are the main points of our interest.

To get a clear idea through which hubs and anti-hubs emerge, let us consider representation of distances in the data space via the hyper-sphere centered at some reference point. The data mean is selected as the reference point. Some data points will lie very close to the hyper-sphere surface, some will lie considerably below the surface and some will be situated considerably above.

Since the data mean is taken to be the center of the hyper-sphere, as dimensionality increases the points considerably below the surface tend to become closer, in relative terms, to other points in the data set, becoming hubs.

Conversely, points considerably above the surface of the hyper-sphere become anti-hubs, while points near to the surface, i.e., the "regular" points, tend to have a close to expected value of N_k (which is k).

It follows that the principal mechanism which generates hubs and anti-hubs is spatial centrality: when a point is closer to the center, the distances to its neighbors become smaller. Conversely, when a point is farther away from the center, the distances to its nearest neighbors become larger. Spatial centrality becomes amplified as dimensionality increases, producing more pronounced hubs and anti-hubs.

Example 3.1: To illustrate the effect of centrality on formation of hubs and anti-hubs, let us consider Spearman's rho and Kendall's tau-b correlations between Euclidean distance to the data center and N5 scores, for the same iid uniform data used in the previous example. As dimensionality increases, stronger correlation emerges, increasing from 0:02 Spearman/0:014 Kendall in the three-dimensional case to 0:8 Spearman/0:63 Kendall for 20 dimensions and 0:867 Spearman/0:715 Kendall for the 100-dimensional setting (weaker Kendall than Spearman correlations are predominantly due to ties in the integer N5 scores). This implies that points closer to the center tend to become hubs. We made analogous observations with other values of k n, and combinations of data distributions and distance measures for which hubness occurs.

B. The Relation between Anti hubs and Outliers

We can generally categorize outlier detection methods into global and local approaches. We can decide whether a data object is outlier or not based on the complete database or only on a selection of data objects i.e., global or local scope.

For example, by raising the value of k when using the classic k-NN outlier detection method, one increases the set of data points used to determine the outlier score of the point of interest, moving from a local to a global notion of outlierness, and ending in the extreme case when k = n-1. Likewise, raising k when determining reverse nearest neighbors, i.e., anti-hubs, raises the expected size of reverse-neighbor sets (while their size can still vary amongst points).

Since anti-hubs have been defined as points with the lowest Nk values, we can explore the relation between Nk scores and outlierness by measuring the correlation between Nk values and outlier scores produced by unsupervised methods.

For the data in Example 3.1, we measured the Kendall tau-b correlation coefficient between inverse Nk values and the outlier scores computed by the k-NN and ABOD methods (for efficiency reasons we use Fast ABOD [19] with k=0.1 n=1,000). The measured correlations are plotted in Figs. 2a and 2b, together with the correlation between inverse Nk values and the distance to the data set mean (Fig. 2c) for two values of dimensionality: low (d = 2) and high (d = 100). Furthermore, we consider two portions of points for computing correlations: all points (p = 100%) and p = 5% of points with the highest distance from the data set mean as the strongest outliers. It can be seen that for the high dimensional case correlations for p = 100% are very strong for a wide range of k values, with the exceptions being very low (close to 1) and very high values (close to n = 10,000). For p = 5% agreement between Nk and k-NN/ABOD still exists, but is notably weaker. This means that Nk scores can be considered a feasible alternative to established k-NN and ABOD outlier scoring methods, since on one hand they produce very similar rankings overall, but on the other hand the rankings of the strongest outliers produced by Nk values do not completely agree with the established methods, suggesting that Nk is not redundant compared to them. The suitability of Nk for expressing outlierness is supported by the strong correlations with the "ground truth" shown in Fig. 2c.

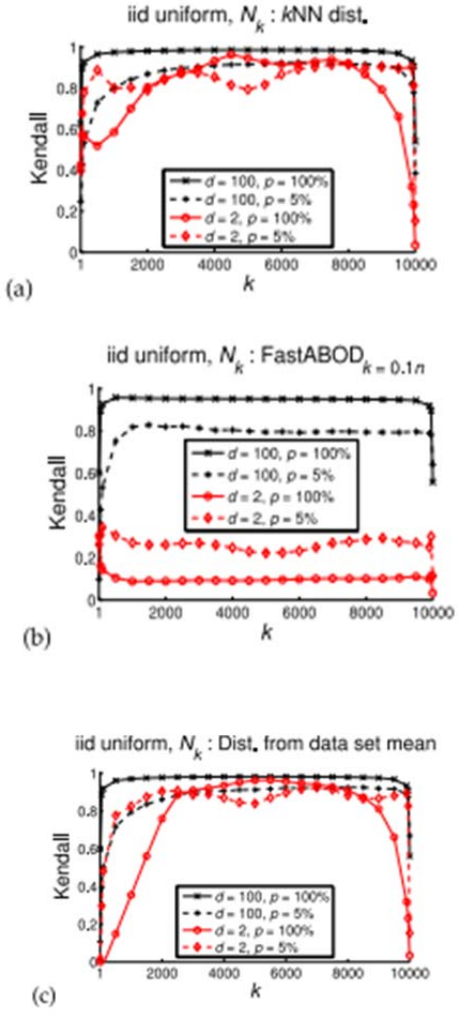


Fig 2: Correlation between Nk values and outlier scores (a, b), and the distance from the data set mean as the "ground truth" (c), for iid uniform random data (n = 10,000 points).

V. OUTLIER DETECTION BASED ON ANTI-HUBS

In this section we consider methods for outlier detection based on anti-hubs described in previous sections. Natural outlier scoring based on Nk counts was used in the ODIN method [11]. Since we do not consider

threshold parameters, and apply normalization to the scores, we will reformulate this method as AntiHub, which defines the outlier score of point x from data set D as a function of Nk(x), and is given in Algorithm 1.

Algorithm 1: AntiHub1

Input:

Distance measure "dist"

Ordered data set D = (x1; x2; ... ; xn) where xi ϵ D, for i ϵ {1; 2; ... ; n}

No. of neighbors k ϵ {1; 2; ...}

Output:

Vector s = (s1; s2; ... ; sn) , where si is the outlier score of xi, for i ϵ {1; 2; ... ; n}

Steps:

1) Let x be our point of interest. We find out the nearest neighbors of all other data objects except x.

2) Then we calculate the number of times x has occurred among the nearest neighbors of all other points in D with respective to dist.

3) This value is the Nk occurrence of point x i.e. Nk(xi).

4) The outlier score of point x from data set D is a function of Nk(x).

5) We repeat the above process for each and every data object.

Discrimination of scores represents a notable weakness of the AntiHub method.

In order to add more discrimination, one approach could be to raise k, possibly to some value comparable with n.

But the approach raises two concerns:

(1) With increasing k the notion of outlierness moves from local to global, thus if local outliers are of interest they can be missed;

(2) k values comparable with n raise issues with computational complexity.

We propose a simple heuristic method AntiHub2, which refines outlier scores produced by the AntiHub method by also considering the Nk scores of the neighbors of x, in addition to Nk(x) itself.For each point x, AntiHub2 proportionally adds (1-α).Nk(x) to α times the sum of Nk scores of the k nearest neighbors of x, where α ϵ [0,1]. To aggregate the neighbors' scores, we select summation as a simple way, while other heuristics such as averaging are also possible. The proportion α is automatically determined by maximizing discrimination between outlier scores of the strongest outliers, and controlled by two user-provided parameters: the ratio of strongest outliers for which to observe discrimination and step size when searching for the best a value. Algorithm 2 describes the method in more detail.

Algorithm 2: AntiHub2

Input:

- Distance measure dist
- Ordered data set D = (x1, x2, ... , xn), where xi ϵ D, for i ϵ {1, 2, ... , n}
- No. of neighbors k = {1, 2, ...}

Output:

- Vector s = (s1,s2, ... , sn) , where si is the outlier score of xi, for i ϵ (1,2, ... , n)

Temporary variables:

- AntiHub scores a
- Sums of nearest neighbors' AntiHub scores "ann"
- Proportion α ϵ [0; 1]
- (Current) raw outlier scores ct,t

Steps:

1: Assign a := AntiHub1

2: For each i ϵ (1, 2, ... , n)

Assign anni :=Σ j ϵ NNdist(k;i) aj, where NNdist(k; i) is the set of indices of k nearest neighbors of xi

3: Now add (1-α).Nk(x) to α times the sum of Nk scores of the k nearest neighbors of x. i.e, cti =(1-α) ai +α.anni

4: The outlier score of point x from data set D is a function of cti

---

VI. EXPERIMENTAL EVALUATION

The goals of our experimental evaluation are

1) To examine the effectiveness of outlier detection methods proposed in the previous section.
2) To examine the behavior of the methods with respect to the k parameter.

The main aim of this section is to further support that reverse-neighbor relations can be effectively applied to detect outliers in both high- and low-dimensional settings.

A. Experimental Procedure

I) Methods:

We consider the two methods described in the previous section, denoted AntiHub1(k) and AntiHub2(k), where k is the used number of nearest neighbors. We will always assume Euclidean distance. For convenience, k may be referred to as a fraction of data set size n.

We examine the k-NN method as the main criteria for comparison [3]. For the second baseline we select ABOD [19] as it exploits the properties of high-dimensional data.
We will also include LOF [30] as a classic representative of density-based methods. Finally, we will include the influenced outlierness method (INFLO) [24], a density-based method which also makes use of reverse nearest neighbors (through a symmetric neighborhood relationship).

We employ the two-way search method [24], always using the default threshold value of M = 1. For LOF and INFLO we use the implementations provided by the environment for developing KDD-applications supported by index-structures (ELKI) [31], while for other methods we use our own implementations.
In all the experiments performed, we used areas under curve (AUC), which is a standard way to measure the effectiveness of outlier-detection methods [4], [18], [32].

II)Data Sets

To perform experiments we used synthetic as well as real data sets. Synthetic data sets are employed because we can modify crucial parameters, such as dimensionality and distribution of data. Comparison among different outlier detection methods is also performed with real data sets summarized in Table 1, which shows the number of points (n), dimensionality (d), skewness of the distribution of N10 (SN10), the percentage of points labeled as outliers, and data set source.

The associated class labels are used only for evaluation purposes. Unless otherwise stated in Table 1, we designate the minority class as outliers. All real data sets are z-score standardized. The SN10 values are included

following the approach from [14], in order to give an indication of data set intrinsic dimensionality which affects the degree distribution of the k-NN graph (for k n).
Despite of the majority of data sets having high (embedding) dimensionality d, only us-crime appears to be intrinsically high-dimensional (with SN10 > 1), with churn, ctg3, ctg10 and nba-allstar-1973-2009 being of moderate intrinsic dimensionality, and other data sets having low intrinsic dimensionality.

B.*Experimental Results*

I) Experiment with Synthetic Data

We randomly generate n = 10,000 points divided into two clusters of equal size, by drawing 5,000 points from the multivariate normal distribution with mean -1 and independent components with standard deviation 0.1, and the other 5,000 points from the normal distribution with mean 1 and component standard deviation 1. We consider dimensionalities 2 and 100. For each cluster, we take 5% of points with the largest distance from their cluster center, move them even farther from the center by 20% of the distance, and designate them as outliers.

The results of applying outlier detection methods are shown in Fig. 3. The chart in Fig. 3a shows the 2-dimensional setting, while Fig. 3b depicts 100-dimensional data. Both charts show the AUC of outlier detection methods as a function of parameter k, with the exception of ABOD which is used in its Fast ABOD variant with k = 0.1n. To facilitate a better view of both local and global case, a logarithmic horizontal axis is used.

In all cases k-NN and ABOD do not perform particularly well. AntiHub is able to achieve very good performance, with the only difference between the low- and high-dimensional settings being in the value of k where good performance is achieved: in low dimensions values between 100 and 1,500 are all sufficient, while in high dimensions it performs best for k ranging from 500 to just below 5,000. Density-based methods LOF and INFLO expectedly perform very well, but it is important to note that AntiHub exhibits robustness to different densities in a very simple and natural fashion, without explicitly modeling density.

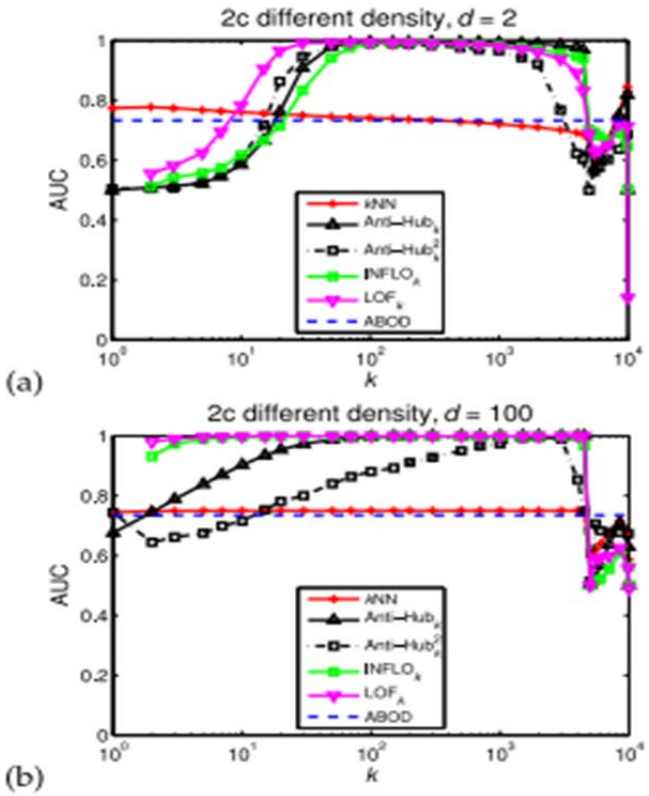


Fig. 3. Results for synthetic data with two cluster

VII. CONCLUSION

In this paper, we provided a unifying view of the role of reverse nearest neighbor counts in problems concerning unsupervised outlier detection, focusing on the effects of high dimensionality on unsupervised outlier-detection methods.

Based on the analysis, we formulated the Anti Hub method for detection of outliers, discussed its properties, and proposed a derived method which improves discrimination between scores.

The existence of hubs and anti-hubs in high-dimensional data is relevant to machine-learning techniques from various families: supervised, semi-supervised, as well as unsupervised.

In this paper we mainly focused on only unsupervised methods, but in future work it can be extended to supervised and semi-supervised methods as well. Another relevant topic is the development of approximate versions of Anti Hub methods that may sacrifice accuracy to improve execution speed.

Finally, secondary measures of distance/similarity, such as shared-neighbor distances warrant further exploration in the outlier-detection context.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Survey, vol. 41, no. 3, p. 15, 2009.

[2] P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection. Hoboken, NJ, USA: Wiley, 1987.

[3] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," SIGMOD Rec., vol. 29, no. 2, pp. 427–438, 2000.

[4] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data," in Proc. Conf. Appl. Data Mining Comput. Security, 2002, pp. 78–100.

[5] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," VLDB J., vol. 8, nos. 3–4, pp. 237–253, 2000.

[6] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in Proc. 7th Int. Conf. Database Theory, 1999, pp. 217–235.

[7] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in Proc. 8th Int. Conf. Database Theory, 2001, pp. 420–434.

[8] D. Franc¸ois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," IEEE Trans. Knowl. Data. Eng., vol. 19, no. 7, pp. 873–886, Jul. 2007.

[9] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in Proc. 27th ACM SIGMOD Int. Conf. Manage. Data, 2001, pp. 37–46.

[10] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," Statist. Anal. Data Mining, vol. 5, no. 5, pp. 363–387, 2012.

[11] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in Proc 17th Int. Conf. Pattern Recognit., vol. 3, 2004, pp. 430–433.

[12] J. Lin, D. Etter, and D. DeBarr, "Exact and approximate reverse nearest neighbor search for multimedia data," in Proc 8th SIAM Int. Conf. Data Mining, 2008, pp. 656–667.

[13] A. Nanopoulos, Y. Theodoridis, and Y. Manolopoulos, "C2P: Clustering based on closest pairs," in Proc 27th Int. Conf. Very Large Data Bases, 2001, pp. 331–340.

[14] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Hubs in space: Popular nearest neighbors in high-dimensional data," J. Mach. Learn. Res., vol. 11, pp. 2487–2531, 2010.

[15] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," SIGMOD Rec., vol. 29, no. 2, pp. 93–104, 2000.

[16] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," in Proc 19th IEEE Int. Conf. Data Eng., 2003, pp. 315–326.

[17] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in Proc 13th Pacific-Asia Conf. Knowl. Discovery Data Mining, 2009, pp. 813–822.

[18] H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek, "LoOP: Local € outlier probabilities," in Proc 18th ACM Conf. Inform. Knowl. Manage., 2009, pp. 1649–1652.

[19] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in Proc 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 444–452.

[20] M. E. Houle, H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek, € "Can shared-neighbor distances defeat the curse of dimensionality?" in Proc 22nd Int. Conf. Sci. Statist. Database Manage., 2010, pp. 482–500.