On Complementarity of Cluster and Outlier Detection Schemes

Zhixiang Chen[1], Ada Wai-Chee Fu[2], and Jian Tang[2]

[1] Department of Computer Science,
University of Texas-Pan American, Edinburg TX 78539 USA.
chen@cs.panam.edu
[2] Department of Computer Science and Engineering,
Chinese University of Hong Kong, Shatin, N.T., Hong Kong.
adafu@cse.cuhk.edu.hk,jtang@cse.cuhk.edu.hk

1 Introduction

We are interested in the problem of outlier detection, which is the discovery of data that deviate a lot from other data patterns. Hawkins [7] characterizes an outlier in a quite intuitive way as follows: *An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.*

Most methods in the early work that detects outliers independently have been developed in the field of statistics [3]. These methods normally assume that the distribution of a data set is known in advance. A large amount of the work was done under the general topic of clustering [6, 12, 15, 8, 17]. These algorithms can also generate outliers as by-products. Recently, researchers have proposed distance-based and density-based as well as connectivity-based outlier detection schemes [10, 11, 13, 4, 16], which distinguish objects that are likely to be outliers from those that are not based on the number of objects in the neighborhood of an object These schemes do not make assumptions about the data distribution.

In this paper, we want to find out if the indirect clustering approach such as [6] and the direct approach such as [10, 4, 16] are similar in the effects of outlier detection and also the cases where they may differ. When a direct approach and an indirect one have similar effects, we say that they are complementary. We consider the comparison of DBSCAN clustering method and the DB-Outlier, LOF and COF definitions of outliers. These methods are chosen based on their more superior powers on handling clusters or patterns of different shapes with no apriori distribution assumption. We believe that these methods are better equipped to handle the varieties of outlier natures. Some interesting discoveries are made. First, we show that DBSCAN and DB-Outlier approaches are almost complementary, and we also show an extension of the DB-outlier scheme so that it is complementary with DBSCAN. Second, we show that DBSCAN approach is complementary with density-based and connectivity-based outlier schemes within a density cluster or far away from some clusters. Finally, we show that there are cases where DBSCAN approach is not complementary with density-based and connectivity-based outlier schemes.

2 Density Based Clustering and Outlier Detection

Let $\mathcal{D}$ be a data set of points in a data space. For any $p \in \mathcal{D}$ and any positive value v, the v-neighborhood of p is $N_v(p) = \{o : dist(p, o) \leq v \ \& \ o \in \mathcal{D}\}$. For a given threshold $n > 0$, p is called a *core point* with respect to v and n (or *core point* if no confusion arises) if $|N_v(p)| \geq n$.

Given $v > 0$ and $n > 0$, a point p is *directly density-reachable* from a point q with respect to v and n, if $p \in N_v(q)$ and q is a core point. A point p is *density-reachable* from a point q with respect to v and n, if there is a chain of points $p_1, p_2, \ldots, p_s$, $p_1 = q$, $p_s = p$ such that p_{i+1} is directly density-reachable from p_i. A point p is *density-connected* to a point q with respect to v and n, if there is a point o such that both p and q are density-reachable from o with respect to v and n.

Definition 1 *(***DBSCAN Clusters and Outliers** *[6]). Let $\mathcal{D}$ be a set of objects. A cluster C with respect to v and n in $\mathcal{D}$ is a non-empty subset of $\mathcal{D}$ satisfying the following conditions: (1) Maximality: $\forall p, q \in D$, if $p \in C$ and q is density-reachable from p with respect to v and n, the also $q \in C$. (2) Connectivity: $\forall p, q \in C$, q is density-connected to p with respect to v and n. Finally, every object not contained in any cluster is an* outlier *(or a noise).*

3 Outlier Detection Schemes

Distance Based Outliers: Knorr and Ng [10] proposed a distance-based scheme, called $DB(n, v)$-outliers. Let $\mathcal{D}$ be a data set. For any $p \in \mathcal{D}$ and any positive values v and n, p is a $DB(n, v)$-outlier if $|N_v(p)| \leq n$, otherwise it is not. The weakness is that it is not powerful enough to cope with certain scenarios with different densities in data clusters [4]. Similar weakness is found in the scheme proposed by Ramaswamy, et al. [13], which is actually a special case of $DB(n, v)$-outlier.

Density Based Outliers: Breuning, et al. [4] proposed a density-based outlier detection scheme as follows. Let $p, o \in \mathcal{D}$. Let k be a positive integer, and *k-distance*(o) be the distance from o to its k-th nearest neighbor. The reachability distance of p with respect to o for k is $\textit{reach-disk}_k(p, o) = max\{\textit{k-distance}(o), dist(p, o)\}$. The reachability distance smoothes the fluctuation of the distances between p and its *"close"* neighbors. The local reachability density of p for k is:

$$lrd_k(p) = \left(\frac{\sum_{o \in N_{\textit{k-distance}(p)}(p)} \textit{reach-dist}_k(p, o)}{|N_{\textit{k-distance}(p)}(p)|} \right)^{-1}.$$

That is, $lrd_k(p)$ is the inverse of the average reachability distance from p to the objects in its k-distance neighborhood. The local outlier factor of p is

$$LOF_k(p) = \frac{\sum_{o \in N_{\textit{k-distance}(p)}(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_{\textit{k-distance}(p)}(p)|}.$$

The LOF value measures how strong an object can be an outlier. A threshold on LOF value can be set to define outliers.

Connectivity Based Outliers: The connectivity based outlier detection scheme was proposed in Tang et al. [16]. This scheme is based on the idea of differentiating *"low density"* from *"isolativity"*. While low density normally refers to the fact that the number of points in the *"close"* neighborhood of an object is (relatively) normalsize, isolativity refers to the degree that an object is *"connected"* to other objects. As a result, isolation can imply low density, but the other direction is not always true.

Definition 2 *Let $P, Q \subseteq \mathcal{D}$, $P \cap Q = \emptyset$ and $P, Q \neq \emptyset$. We define $dist(P, Q) = min\{dist(x, y) : x \in P \ \& \ y \in Q\}$, and call $dist(P, Q)$ the distance between P and Q. For any given $q \in Q$, we say that q is the nearest neighbor of P in Q if there is a $p \in P$ such that $dist(p, q) = dist(P, Q)$.*

Definition 3 *Let $G = \{p_1, p_2, \dots, p_r\}$ be a subset of $\mathcal{D}$. A set-based nearest path, or SBN-path, from p_1 on G is a sequence $\langle p_1, p_2, \dots, p_r \rangle$ such that for all $1 \leq i \leq r-1, p_{i+1}$ is a nearest neighbor of set $\{p_1, \dots, p_i\}$ in $\{p_{i+1}, \dots, p_r\}$.*

In the above, if the nearest neighbor is not unique, we can impose a pre-defined order among the neighbors to break the tie. Thus an *SBN*-path is uniquely determined.

Definition 4 *Let $s = \langle p_1, p_2, \dots, p_r \rangle$ be an SBN-path. A set-based nearest trail, or SBN-trail, with respect to s is a sequence $\langle e_1, \dots, e_{r-1} \rangle$ such that for all $1 \leq i \leq r-1$, $e_i = (o_i, p_{i+1})$ where $o_i \in \{p_1, \dots, p_i\}$, and $dist(e_i) = dist(o_i, p_{i+1}) = dist(\{p_1, \dots, p_i\}, \{p_{i+1}, \dots, p_r\})$.*

Again, if o_i is not uniquely determined, we should break the tie by a pre-defined order. Thus the *SBN*-trail is unique for any *SBN*-path.

Definition 5 *Let $G = \{p_1, p_2, \dots, p_r\}$ be a subset of $\mathcal{D}$. Let $s = \langle p_1, p_2, \dots, p_r \rangle$ be an SBN-path from p_1 and $e = \langle e_1, \dots, e_{r-1} \rangle$ be the SBN-trail with respect to s. The average chaining distance from p_1 to $G - \{p_1\}$, denoted by $ac\text{-}dist_G(p_1)$, is defined as*

$$ac\text{-}dist_G(p_1) = \sum_{i=1}^{r-1} \frac{2(r-i)}{r(r-1)} \cdot dist(e_i).$$

Definition 6 *Let $p \in \mathcal{D}$ and k be a positive integer. The connectivity-based outlier factor (COF) at p with respect to its k-neighborhood is defined as*

$$COF_k(p) = \frac{|N_k(p)| \cdot ac\text{-}dist_{N_k(p)}(p)}{\sum_{o \in N_k(p)} ac\text{-}dist_{N_k(o)}(o)}.$$

A threshold on COF can be set to define outliers.

4 Complementarity of DB-outlier and DBSCAN

When the clustering approach and outlier detection approach both give the same result about a data point (as outlier or non-outlier) we say that they are *complementary*. Since both techniques typically require some parameter settings, it is of interest to see if there exist some parameter settings for each approach so that the methods are complementary. In this section, we first show that the DBSCAN clustering scheme and the DB-outlier detection scheme are *almost* complementary. We then propose an extended DB-outlier detection scheme and show that it is complementary with the DBSCAN clustering scheme.

Theorem 1 *If there is a parameter setting for DBSCAN clustering scheme to detect clusters and outliers, then there is a parameter setting for the DB-outlier detection scheme such that the following is true: For any object $p \in \mathcal{D}$, if DBSCAN identifies p as an outlier then DB-outlier detection scheme also identifies it as an outlier. (Note that this implies that if DB-outlier scheme detection scheme identifies p not to be an outlier, then DBSCAN identifies it not to be an outlier (i.e., inside some cluster).)*

Proof. Let a parameter setting for DBSCAN be v and n. For any object $p \in \mathcal{D}$, if DBSCAN identifies p as an outlier, then we have by definition that $|N_v(p)| < n$. When we choose the same parameter setting v and n for the DB-outlier detection scheme, it identifies p as an outlier too, because $|N_v(p)| < n$. □

It is easy to see that objects identified by DBSCAN to be inside clusters can be identified as outliers by the DB-outlier detection scheme with the same parameter setting. In order to avoid such in-complementarity on border objects, we propose the following extension of the DB-outlier detection scheme:

Definition 7 (EDB-Outliers). *Given any object p in a data set $\mathcal{D}$, p is an* extended distance-based outlier*, denoted as EDB-outlier, with respect to v and n, if $|N_v(p)| < n$ and $\forall q \in N_v(p)$, $|N_v(q)| < n$.*

Following the work in [10], one can easily design an EDB-outlier detection scheme to detect EDB-outliers with respect to the parameter setting of v and n. The following result shows that EDB-outliers and DBSCAN-outliers are complementary.

Theorem 2 *If there is a parameter setting for DBSCAN to detect clusters and outliers, then there is a parameter setting for EDB-outlier detection scheme such that the following is true: For any object $p \in \mathcal{D}$, DBSCAN identifies p as an outlier if and only if EDB-outlier detection scheme also identifies it as an outlier.*

Proof. Let a parameter setting for DBSCAN be v and n. For any object $p \in \mathcal{C}$, if DBSCAN identifies p as an outlier, then we have by definition $|N_v(p)| < n$

and p is not density reachable from any core object in $\mathcal{D}$ with respect to v and n. The latter property means that $\forall q \in N_v(p)$, q is not a core object, i.e., $|N_v(q)| < n$. Hence, when we choose the same parameter setting of v and n for the EDB-outlier detection scheme, it identifies p as an outlier too.

On the other hand, if the EDB-outlier detection scheme identifies p as an outlier with respect to v and n, then we have by definition $|N_v(p)| < n$ and $\forall q \in N_v(p)$, $|N_v(q)| < n$. Now consider that the same parameter setting of v and n is chosen for DBSCAN. Suppose that p is inside a cluster C, then by definition there is a core object $q \in C$ such that p is density-reachable from q with respect to v and n. I.e., there is sequence of objects $q_1 = q, q_2, \ldots, q_s, q_{s+1} = p$, $q_i \in C$, such that q_{i+1} is directly density-reachable from q_i. In particular, p is directly density-reachable from q_s. By definition, $|N_v(q_s)| \geq n$ and $p \in N_v(q_s)$. Thus, we have $|N_v(q_s)| \geq n$ and $q_s \in N_v(p)$, a contradiction to the given fact that p is an EDB-outlier with respect to v and n. The above argument implies that p must not be in any cluster with respect to v and n. Therefore, DBSCAN will identify p as an outlier with respect to v and n. □

5 Complementarity of COF and LOF

Let $\mathcal{D}$ be a data set. For any integer $k > 0$ and any object $p \in \mathcal{D}$, we use $N_k(p)$ to denote $N_{\textit{reach-dist}_k(p)}(p)$ for convenience. In the following lemma, C can be viewed as a cluster. We first consider complementarity inside a density cluster.

Lemma 3 *Given a subset C of a data set $\mathcal{D}$ and an integer $k > 0$, assume that $N_k(p) \subseteq C$ and there is a positive value d such that $\forall p \in C$, $d - \epsilon \leq \textit{reach-dist}_k(p) \leq d + \epsilon$ for a very normalsize fixed positive value $0 < \epsilon < d$. Assume further that there exists a positive value $f > 0$ such that for any $p, q \in C$, $f - \psi \leq \textit{dist}(p, q) \leq f + \psi$ for some positive normalsize value ψ with $0 < \psi < f$. Then, $\forall p \in C$, we have*

$$\frac{d-\epsilon}{d+\epsilon} \leq LOF_k(p) \leq \frac{d+\epsilon}{d-\epsilon}, \textit{ and } \frac{f-\psi}{f+\psi} \leq COF_k(p) \leq \frac{f+\psi}{f-\psi}. \tag{1}$$

Proof. $\forall p \in C$, since $N_p \subseteq C$, we have $\forall q \in N_k(p)$, $q \in C$, hence $d - \epsilon \leq \textit{reach-dist}_k(q) \leq d + \epsilon$. By definition, $\frac{1}{lrd_k(p)} = \frac{\sum_{q \in N_k(p)} \textit{reach-dist}_k(p,q)}{|N_k(p)|}$. Hence, $d - \epsilon \leq lrd_k(p) \leq d + \epsilon$. Since $LOF_k(p) = \frac{\sum_{q \in N_k(p)} \frac{lrd_k(q)}{lrd_k(p)}}{|N_k(p)|}$, we have $\frac{d-\epsilon}{d+\epsilon} \leq LOF_k(p) \leq \frac{d+\epsilon}{d-\epsilon}$, this implies the left part of (1).

Given any $p \in C$, let $s = \{e_1, \ldots, e_{r-1}\}$ be the SBN-trail with respect to the SBN-path from p on $N_k(p)$. It follows from the definition and the given conditions that $f - \psi \leq \textit{dist}(e_i) \leq f + \psi$ for $i = 1, 2, \ldots, r-1$. Hence, by definition we have

$$\textit{ac-dist}_k(p) = \sum_{i=1}^{r-1} \frac{2(r-i)}{r(r-1)} \cdot \textit{dist}(e_i) \leq \sum_{i=1}^{r-1} \frac{2(r-i)}{r(r-1)} \cdot (f + \psi) = f + \psi.$$

Similarly, we have

$$ac\text{-}dist_k(p) = \sum_{i=1}^{r-1} \frac{2(r-i)}{r(r-1)} \cdot dist(e_i) \geq \sum_{i=1}^{r-1} \frac{2(r-i)}{r(r-1)} \cdot (f-\psi) = f - \psi.$$

Thus,

$$\text{COF}_k(p) = \frac{|N_k(p)|_{ac\text{-}dist_k}(p)}{\sum_{o \in N_k(p)} ac\text{-}dist_k(o)} \leq \frac{f+\psi}{f-\psi}; \text{COF}_k(p) = \frac{|N_k(p)|_{ac\text{-}dist_k}(p)}{\sum_{o \in N_k(p)} ac\text{-}dist_k(o)} \geq \frac{f-\psi}{f+\psi}$$

Hence, these together implies the right part of (1). □

Theorem 4 *Let C be any cluster in a data set $\mathcal{D}$ satisfying the conditions in Lemma 1. We can choose a parameter setting $d+\epsilon$ and k for the DBSCAN clustering scheme so that all points in C will be identified as cluster points, i.e., non-outliers. We can choose a parameter setting of k and $\frac{d-\epsilon}{d+\epsilon}$ for the LOF-outlier detection scheme so that it identifies all points in C as cluster points. Finally, we can choose a parameter setting of k and $\frac{f-\psi}{f+\psi}$ for the COF-outlier detection scheme so that it identifies all points in C as cluster points as well.*

Proof. For any $p \in C$, it follows from the conditions of Lemma 3 that $reach\text{-}dist_k(p) \leq d+\epsilon$. By the definition of reachability distance, $N_k(p)$ will have at least k objects. Since $N_k(p) \subset N_{d+\epsilon}(p)$, we have $|N_{d+\epsilon}(p)| \geq k$. Hence, p is a core object with respect to $d+\epsilon$ and k. It follows from the definition of DBSCAN clusters that p will be identified as an object inside a cluster, hence a non-outlier. From (1) of Lemma 1 we have $\frac{d-\epsilon}{d+\epsilon} \leq LOF_k(p)and\frac{f-\psi}{f+\psi} \leq COF_k(p)$. This means that when the parameter setting of k and $\frac{d-\epsilon}{d+\epsilon}$ is chosen for the LOF-outlier detection scheme, (Precisely, k is used to define the reachability distance, and $\frac{d-\epsilon}{d+\epsilon}$ is a threshold for the LOF values to select outliers.) p will be identified as a non-outlier. Similarly, when the parameter setting of k and $\frac{f-\psi}{f+\psi}$ is chosen for the COF-outlier detection scheme, p will be identified a non-outlier as well. □

Next we show some cases where LOF and COF are both complementary with DBSCAN in detecting points outside some clusters as outliers. For any two sets of objects A and B, let $dist(A,B) = min\{dist(x,y) : x \in A \ \&\ y \in B\}$. Again, for any object p, we let $N_m(p)$ denote the m-reachability neighborhood $N_{reach\text{-}dist_m(p)}$ of p.

Theorem 5 *Given two subsets O and C of a data set $\mathcal{D}$, let $d = \min\{dist(x,y) : x \in O \& y \in C\}$, and $m = |O|$. Assume that $\forall o \in O$, $N_m(o) = O$, and $N_{2m}(o) - O \subset C$. Moreover, $\forall p \in C$, $lrd_{2m}(p) \geq \frac{4}{d}$ and $ac\text{-}dist_{2m}(p) < 1$; and for $\forall p \in O$, $\frac{2d}{3} \leq ac\text{-}dist_{2m}(p) \leq d$. Then, there exist parameter settings for DBSCAN, LOF and COF respectively such that each of the three methods will identify O as outliers.*

Proof. First, let r be the diameter of O, i.e., $r = max\{dist(x,y) : x \in O \& y \in O\}$. Since for any $o \in O$, $N_m(o) = O$, we have $reach\text{-}dist_m(o) \leq r$ and $N_m(o) = N_r(o) = O$. When the parameter setting of r and $m+1$ is chosen for the DBSCAN clustering scheme, then o is not a core object. Since $N_r(o) = O$, o is not directly reachable from any object outside O with the distance r. Hence, DBSCAN will identifies O as outliers.

For any $o \in O$, given conditions $N_m(o) = O$ and $N_{2m}(o) - O \subset C$ it follows that, for any $p \in N_{2m}(o)$, $reach\text{-}dist_{2m}(o,p) = max\{2m\text{-}distance(p), dist(o,p)\} \geq d$. Thus, $lrd_{2m}(o) = \frac{|N_{2m}(o)|}{\sum_{p \in N_{2m}(o)} reach\text{-}dist_{2m}(p,o)} \geq \frac{1}{d}$. Hence,

$$LOF_{2m}(o) = \frac{\sum_{p \in N_{2m}(o)} \frac{lrd_{2m}(p)}{lrd_{2m}(o)}}{|N_{2m}(o)|} = \frac{d(\sum_{p \in O-\{o\}} lrd_{2m}(p) + \sum_{p \in C \cap N_{2m}(o)} lrd_{2m}(p))}{|N_{2m}(o)|}$$
$$\geq \frac{d(\sum_{p \in O-\{o\}} \frac{1}{d} + \sum_{p \in C \cap N_{2m}(o)} \frac{4}{d})}{|N_{2m}(o)|} = \frac{|O - \{o\}| + 4|C \cap N_{2m}(o)|}{|N_{2m}(o)|} > 1.$$

Therefore, when the parameter setting of $2m$ and 1 is chosen for the LOF-outlier detection scheme, each object in O will be identified as an outlier.

Similarly, For any $o \in O$, let $y = |N_{2m}(o)|$, then $y \geq 2m$. It follows from the given conditions that $\mathrm{COF}_{2m}(o)$

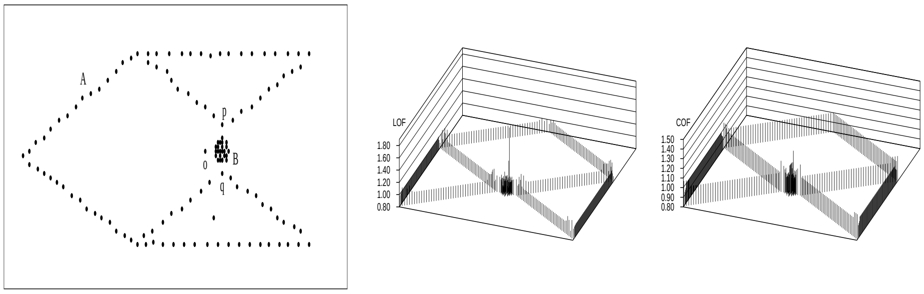
$$= \frac{|N_{2m}(o)| \cdot ac\text{-}dist_{2m}(o)}{\sum_{p \in N_{2m}(o)} ac\text{-}dist_{2m}(p)} \geq \frac{y \cdot \frac{2d}{3}}{\sum_{p \in O-\{o\}} ac\text{-}dist_{2m}(p) + \sum_{p \in N_{2m}(o) \cap C} ac\text{-}dist_{2m}(p)}$$
$$\geq \frac{y\frac{2d}{3}}{(m-1)d + (y-m+1)} \geq 1, \quad when\ d \geq 6.$$

Hence, when the parameter setting of $2m$ and 1 is chosen for the COF-outlier detection scheme, every object in O will be identified as outliers. □

6 Non-Complementarity

In this section, we shall show two non-complementarity results of DBSCAN cluster and outlier schemes LOF and COF, which reveal that in general DBSCAN scheme is not complementary with the LOF-outlier detection scheme, nor with the COF-outlier detection scheme. We use two approaches to obtain our results, one is by actual computation and the other is detailed analysis. The computing environment of our computation is a Dell Precision 530 Workstation with dual Xeon 1.5 GHz processors. In order to have more precise results, all arithmetic operations were carried out with 10 decimal digit precision.

Example 1. Let us consider a data set $\mathcal{D}_1$ consisting of data objects as shown in Fig. 1. A is the set composed of objects on the line patterns, and B is the set composed of all objects in the disk pattern. o refers to the object outside B. A has 402 objects such that the 1-distance of any object in it is $\sqrt{2}$. B has 44 objects such that the 1-distance of any object in it is less than $\sqrt{2}$ and the distance between A and B (which are respectively, p and q) is $\sqrt{2}$. Finally, the



(a) Data set (b) LOF Values (c) COF Values

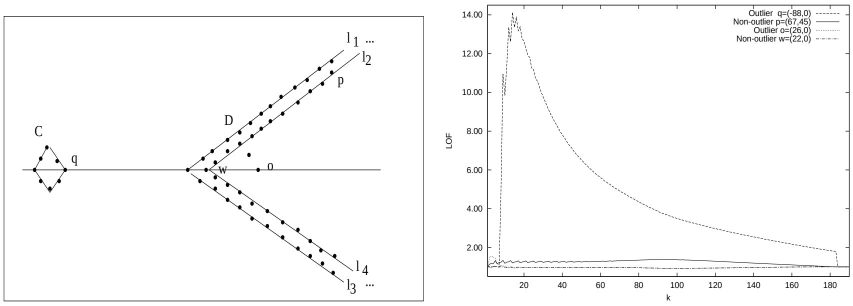
Fig. 1. LOF and COF Values for Example 1 ($k = 3$)

distance between o and B is $\sqrt{2}$, and the distance between o and A is greater than $\sqrt{2}$.

Non-Complementarity Result 1. *When the parameter setting of $k = 3$ and $threshold = 1.4$ is used, the LOF-outlier detection scheme identifies o as an outliers, and all objects in A or B as non-outliers. When the parameter setting of $k = 3$ and $threshold = 1.13$ is used, the COF-outlier detection scheme identifies o as an outliers, and all objects in A or B as non-outliers. Finally, for any parameter setting of v and n, the DBSCAN scheme will not be able to identify o as an outlier and objects in A or B as non-outliers.* □

The first two results are obtained through actual computation. We have implemented the LOF outlier detection algorithm in [4] and the COF outlier detection algorithm in [16]. We used k=3 to compute LOF values and the COF values. The results are shown in Figure 1. We found that the LOF value of o is 1.8146844215 and all other objects have LOF values the same or almost the same as 1, except several objects have LOF values larger than 1 and less than 1.4. Hence, the threshold 1.4 enables the LOF-outlier detection scheme to distinguish the outlier o from all the non-outliers in A or B. We also found that the COF value of o is 1.2057529817, and all the other objects have COF values the same or almost the same as 1, except several objects have COF values larger than 1 but less than 1.13. Hence, the threshold value 1.13 enables the COF-outlier detection scheme to distinguish the outlier o from all the other non-outliers.

Recall that the distance between any two adjacent objects in A is $\sqrt{2}$ and $dist(A, B) = \sqrt{2}$. If the radius parameter v is less than $\sqrt{2}$, then the DBSCAN scheme will identify all objects in A as outliers. In order to identify objects in A as non-outliers, the radius parameter v must be greater than $\sqrt{2}$. Note that the distance between o and B is $\sqrt{2}$ and the 1-distance of any object in B is less than $\sqrt{2}$. Let $w \in B$ such that $dist(o, w) = \sqrt{3}$. For any radius parameter $v > \sqrt{2}$, o is directly reachable from w with respect to v. Hence, if DBSCAN scheme identifies w as an non-outliers then it also identifies o as an non-outlier. In summary, for any parameter setting of v and n, the DBSCAN scheme cannot distinguish o from objects in A if $v \leq \sqrt{2}$; it cannot distinguish o from objects in B if $v > \sqrt{2}$. Therefore, no parameter setting of v and n enables the DBSCAN scheme to identifies outliers and clusters for objects in $\mathcal{D}$.



(a) Data Set of Example 2 (b) LOF Values of Four Objects

Fig. 2. Illustrations of Example 2

EXAMPLE 2. Let us consider a data set $\mathcal{D}_2$ as shown in Fig. 2(a). $\mathcal{D}_2$ has 192 objects. Set C composed of eight objects on the border of a diamond pattern such that any two adjacent objects have a distance $\sqrt{2}$. Set D composed of 182 objects on four lines l_1, l_2, l_3 and l_3. l_1 and l_2 lie 45 degrees above the horizontal line, and l_3 and l_4 lie 45 degrees below the horizontal line. l_2. l_1 and l_3 meet at $(20, 0)$, and l_2 and l_4 meet at $w = (22, 0)$. Any two adjacent objects on any of the four lines have a distance of $\sqrt{2}$. Two additional objects lie below the l_2, the first is $(22, 4)$ and the second is $o = (26, 0)$. Let $q = (-88, 0)$ and $p = (67, 45)$ as shown in the figure. According to Hawkin's definition, is obvious that q and o are outliers, but p and w are not.

Non-Complementarity Result 2. *When a parameter setting of $v = \sqrt{2}$ and $n = 4$ is chosen, the DBSCAN scheme identifies objects in C (including q) and o as outliers and all other objects as non-outliers. However, for any parameter setting of k and threshold, the LOF-outlier detection scheme cannot identify outliers and non-outliers correctly.* □

Because objects in C lie at the border of a diamond shape and an equal distance of $\sqrt{2}$ separates any two adjacent objects in the diamond shape in C, the $\sqrt{2}$-neighborhood of any object in C has exactly 3 objects. Thus, every object in C is an outlier with respect to $v = \sqrt{2}$ and $n = 4$. It follows from the condition of $\mathcal{D}_2$, the $\sqrt{2}$-neighborhood of any object in D is exactly 4 except the six end objects $(20, 0), w, q, (66, 46), (66, -46)$ and $(67, -45)$. But, those six objects are reachable from some other points on the lines with respect to v, and so does the object $(22, 4)$. Hence, the DBSCAN identified all objects in D and $(22, 4)$ as non-outliers. Since $\sqrt{2}$-neighborhood of o has exactly 3 objects and o is not reachable from any core object with respect to v and pct (o is reachable from the non-core object $(22, 4)$), the DBSCAN identifies o as an outlier. In summary, with the parameter setting of $v = \sqrt{2}$ and $pct = 4$, the DBSCAN identifies outliers and clusters in $\mathcal{D}_2$ correctly.

In order to obtain the result for LOF-outlier detection scheme, we ran the LOF outlier detection algorithm for $k = 1, 2, 3, \ldots, 191$ ($191 = |\mathcal{D}_2| - 1$) to compute LOF values for the four objects p, q, o and w. The LOF values are shown in Fig. 2(b). We found the following results: for $1 \leq k \leq 7$, $LOF_k(q) \leq\leq LOF_k(p)$; for $8 \leq k \leq 182$, $LOF_k(o) \leq\leq LOF_k(p)$; and for $183 \leq k \leq 191$, $LOF_k(o) \leq\leq LOF_k(w)$. This implies that any given parameter $threshold$ cannot separate both q and o from p and w. Hence, for any parameter setting of k and $threshold$, the LOF-outlier detection scheme cannot identify outliers and non-outliers in $\mathcal{D}_2$ correctly.

References

1. M. Ankerst, M. Breunig, H.P. Kriegel, and J. Sander: “OPTICS: Ordering points to identify the cluster structure”, Proc. of ACM-SIGMOD Conf., pp. 49-60, 1999.
2. A. Arning, R. Agrawal, P. Raghavan: ”A Linear Method for Deviation detection in Large Databases”, Proc. of 2nd Intl. Conf. On Knowledge Discovery and Data Mining, 1996, pp 164 - 169.
3. V. Barnett, T. Lewis: ”Outliers in Statistical Data”, John Wiley, 1994.
4. M. Breuning, Hans-Peter Kriegel, R. Ng, J. Sander: ”LOF: Identifying density-based Local Outliers”, Proc. of the ACM SIGMOD Conf., 2000.
5. W. DuMouchel, M. Schonlau: ”A Fast Computer Intrusion Detection Algorithm based on Hypothesis Testing of Command Transition Probabilities”, Proc.of 4th Intl. Conf. On Knowledge Discovery and Data Mining, 1998, pp. 189 - 193.
6. M. Ester, H. Kriegel, J. Sander, X. Xu: ”A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, Proc. of 2nd Intl. Conf. On Knowledge Discovery and Data Mining, 1996, pp 226 - 231.
7. T. Fawcett, F. Provost: ”Adaptive Fraud Detection”, Data Mining and Knowledge Discovery Journal, Kluwer Academic Publishers, Vol. 1, No. 3, 1997, pp 291 - 316.
8. S. Guha, R. Rastogi, K. Shim: ”Cure: An Efficient Clustering Algorithm for Large Databases”, In Proc. of the ACM SIGMOD Conf., 1998, pp 73 - 84.
9. D. Hawkins: ”Identification of Outliers”, Chapman and Hall, London, 1980.
10. E. Knorr, R. Ng: ”Algorithms for Mining Distance-based Outliers in Large Datasets”, Proc. of 24th Intl. Conf. On VLDB, 1998, pp 392 - 403.
11. E. Knorr, R. Ng: ”Finding Intensional Knowledge of Distance-based Outliers”, Proc. of 25th Intl. Conf. On VLDB, 1999, pp 211 - 222.
12. R. Ng, J. Han: ”Efficient and Effective Clustering Methods for Spatial Data Mining”, Proc. of 20th Intl. Conf. On Very Large Data Bases, 1994, pp 144 - 155.
13. S. Ramaswamy, R. Rastogi, S. Kyuseok: ”Efficient Algorithms for Mining Outliers from Large Data Sets”, Proc. of ACM SIGMOD Conf., 2000, pp 427 - 438.
14. N. Roussopoulos, S. Kelley, F. Vincent, ”Nearest Neighbor Queries”, Proc. of ACM SIGMOD Conf., 1995, pp 71 - 79.
15. G. Sheikholeslami, S. Chatterjee, A. Zhang: ”WaveCluster: A multi-Resolution Clustering Approach for Very Large Spatial Databases”, Proc. of 24th Intl. Conf. On Very Large Data Bases, 1998, pp 428 - 439.
16. J. Tang, Z. Chen, A. Fu, D. Cheung, “A Robust Outlier Detection Scheme in Large Data Sets”, PAKDD, 2002.
17. T. Zhang, R. Ramakrishnan, M. Linvy: ”BIRCH: An Efficient Data Clustering Method for Very Large Databases”, Proc. of ACM SIGMOD Intl. Conf., , 1996, pp 103 - 114.