Differentially Private Analysis of Outliers

Rina Okada(✉), Kazuto Fukuchi, and Jun Sakuma

University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan
{rina,kazuto}@mdl.cs.tsukuba.ac.jp, jun@cs.tsukuba.ac.jp

Abstract. This paper presents an investigation of differentially private analysis of distance-based outliers. Outlier detection aims to identify instances that are apparently distant from other instances. Meanwhile, the objective of differential privacy is to conceal the presence (or absence) of any particular instance. Outlier detection and privacy protection are therefore intrinsically conflicting tasks. In this paper, we present differentially private queries for counting outliers that appear in a given subspace, instead of reporting the outliers detected. Our analysis of the global sensitivity of outlier counts reveals that regular global sensitivity-based methods can make the outputs too noisy, particularly when the dimensionality of the given subspace is high. Noting that the counts of outliers are typically expected to be small compared to the number of data, we introduce a mechanism based on the smooth upper bound of the local sensitivity. This study is the first trial to ensure differential privacy for distance-based outlier analysis. The experimentally obtained results show that our method achieves better utility than global sensitivity-based methods do.

Keywords: Differential privacy · Outlier detection · Smooth sensitivity

1 Introduction

Data mining technologies are now becoming increasingly influential in our daily life. When data mining is processed over personal data collected from individuals, the acquired knowledge might be used to infer private information. In this paper, we investigate differentially private outlier analysis.

Outlier detection is a task to identify instances that are apparently distant from the remaining instances. The objective of differential privacy [3] is to prevent adversaries from learning of the presence (or absence) of any particular instance from released information. Outlier detection and privacy protection are therefore intrinsically conflicting tasks. It presents a challenging difficulty. To overcome this difficulty, instead of identifying outliers, we consider reporting information which helps to recognize the occurrence of anomalous situations. More specifically, we examine the problem of counting outliers that appear in a given subspace with a guarantee of differential privacy.

A. Appice et al. (Eds.): ECML PKDD 2015, Part II, LNAI 9285, pp. 458–473, 2015.
DOI: 10.1007/978-3-319-23525-7_28

Related Works. We introduce existing studies of privacy aspects of outlier analysis. Secure multiparty computation (SMC) is a cryptographic tool that facilitates the evaluation of a specified function over their private inputs jointly, while maintaining these inputs as private. Vaidya et al. [20] introduced a SMC for distance-based outlier detection from horizontally and vertically partitioned private databases using random shares. Xue et al. [21] investigated a SMC for spatial outlier detection. Dung et al. [1] presented a SMC for distance-based outlier detection with the Mahalanobis distance. Li et al. [12] presented a SMC for density-based outlier detection. The objective of these works is to detect outliers securely without mutually sharing privately distributed data; privacy invasion caused by observing detected outliers is not considered.

Studies of differential privacy for outlier analysis are few, presumably because of its intrinsic difficulty, as described. Only one report in the literature [5] describes a study that considers the differential privacy of outlier analysis. This study was conducted to detect anomalous changes from a time series under a guarantee of differential privacy. The objective of this study is closely related to ours, whereas this method releases a one-dimensional time series with differential privacy; outlier detection is applied to the released data as a post process. Consequently, the approach differs from ours.

Lui et al. [14] introduced a novel privacy notion, outlier privacy, as a generalization of differential privacy. Outlier privacy measures an individual's privacy parameter by how much of an "outlier" the individual is. The objective of this study is to define privacy using the notion of outliers, but not for differentially private outlier analysis.

Our Contribution. We examine the problem of counting outliers that appear in a given subspace with a guarantee of differential privacy (Section 2). Randomization of query responses based on the global sensitivity analysis is the most straightforward approach for realization of differential privacy [4]. We derive the lower and upper bound of the global sensitivity of outlier counts (Section 4.1). From the derived bounds, we reveal that the global sensitivity-based randomization can make the outputs too noisy, particularly when the dimensionality of the given subspace is high. We specifically examine the observation that the counts of outliers are expected to be small compared to the number of data in typical datasets. Taking advantage of this, we develop a randomization mechanism for the counts of outliers based on the smooth upper bound of local sensitivity [18] (Section 4.2). A randomization mechanism based on the smooth upper bound typically has better utility because of its data-dependency. However, its evaluation is often costly. To alleviate this, we provide an efficient algorithm for evaluation of the smooth upper bound for counting outliers (Section 4.2). We demonstrated our methods with synthesized datasets and real datasets (Section 5). The experimentally obtained results demonstrate that our methods achieve better utility than that achieved using global sensitivity-based methods.

2 Differential Privacy

Let $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \in \mathbb{R}^{d \times N}$ be a database. An *analyst* issues a query $q : \mathbb{R}^{d \times N} \to \mathcal{T}$; then the database returns an output, where $\mathcal{T}$ denotes the range of the outputs. *Differential privacy* measures the privacy breach of database X caused by releasing output $T \in \mathcal{T}$ with no assumptions of the background knowledge of adversaries. The outputs are typically modified using a randomization *mechanism* $\mathcal{A} : \mathbb{R}^{d \times N} \to \mathcal{T}$ before release to preserve differential privacy.

Let $H(X, X') = |\{i : \boldsymbol{x}_i \neq \boldsymbol{x}'_i\}|$ denote the Hamming distance, the number of different records in X and X'. If $H(X, X') = 1$, then it can be said that X and X' are neighbor databases. In the following, we presume $|X| = |X'| = N$. Then, mechanism $\mathcal{A}$ guarantees (ϵ, δ)-differential privacy if, $\forall X' : H(X, X') = 1$ and $\forall T \subseteq \mathcal{T}$,

$$Pr[\mathcal{A}(X) \in T] \leq e^{\epsilon} Pr[\mathcal{A}(X') \in T] + \delta.$$

The parameter ϵ and δ are designated as privacy parameters. Randomization based on the global sensitivity is the most straightforward realization of differential privacy for continuous outputs [3].

Global Sensitivity. Presuming that the output domain of query q is in $\mathbb{R}^p$, then randomization based on the global sensitivity [3] provides a mechanism that guarantees differential privacy for queries of any type, as long as its global sensitivity is evaluable. The ℓ_2 global sensitivity of query $q : \mathbb{R}^{d \times N} \to \mathbb{R}^p$ is defined by $GS_q = \max_{X, X' : H(X, X') = 1} \|q(X) - q(X')\|_2$ where $\| \cdot \|$ denotes ℓ_2 norm of vectors. Given the global sensitivity GS_q for query q, the following mechanism $\mathcal{A}$ that randomizes the output of the query by eq. (1) provides (ϵ, δ)-differential privacy [2]:

$$\mathcal{A}_q(X) = q(X) + Y, \tag{1}$$

where Y is an sample drawn from the Gaussian distribution with mean 0 and variance $\frac{GS_q^2 \cdot 2 \log(2/\delta)}{\epsilon^2}$.

Smooth Sensitivity. For some functions, the global sensitivity can be impractically large even when the sensitivities are small with almost all neighboring pairs. This large sensitivity occurs because it is evaluated as the greatest difference of outputs among possible neighboring pair of databases. For example, the global sensitivity of median is N, the whole sample size, but this arises only in a pathological situation. Randomization based on the smooth sensitivity [18] enables the use of moderate sensitivity for such overly sensitive queries. For a given database X, the ℓ_2 *local sensitivity* for query q is defined as the greatest difference of outputs for $\forall X'$ s.t. $X' : H(X, X') = 1$:

$$LS_q(X) = \max_{X' : H(X, X') = 1} ||q(X) - q(X')||_2.$$

It is noteworthy that $GS_q = \max_{X \in \mathbb{R}^{d \times N}} LS_q(X)$ holds.

Nissim et al. presented the *smooth sensitivity* [18], which is a class of smooth upper bounds to the local sensitivity. Given $\beta > 0$, the smooth sensitivity of query $q : \mathbb{R}^{d\times N} \to \mathbb{R}^p$ is defined by

$$S^*_{q,\beta}(X) = \max_{X' \in \mathbb{R}^{d\times N}} (LS_q(X') \cdot e^{-\beta H(X,X')}).$$

[18] also showed that adding noise proportional to the smooth sensitivity yields a differentially private mechanism if the noise distribution satisfies some properties. Let Y be a noise generated from the Gaussian distribution with mean 0 and variance 1. Let $S_{q,\beta}$ be a β-smooth upper bound of query q. Then, if $\alpha = \frac{\epsilon}{5\sqrt{2\ln 2/\delta}}$ and $\beta = \frac{\epsilon}{4(p+\ln 2/\delta)}$, mechanism $\mathcal{A}_q$ guarantees (ϵ, δ)-differential privacy [18]:

$$\mathcal{A}_q(X) = q(X) + \frac{S_{q,\beta}(X)}{\alpha} \cdot Y.$$

3 Problem Statement

Our objective is to analyze outliers that are included in a private database in a differentially private manner. Outlier detection is a problem to identify an instance that is significantly distant from other instances. Therefore, the result of outlier detection is fundamentally privacy-invasive in terms of differential privacy. In order to understand the behavior of the outliers in the target dataset without identifying outliers, we investigate counting outliers in a given subspace under the constraint of differential privacy.

3.1 Counting Outliers

In this study, we use distance-based outliers [9]. Presuming that records are real-valued vectors, $\boldsymbol{x}_i \in \mathbb{R}^d$, and letting $X = \{\boldsymbol{x}_i\}_{i=1}^N$ denote the database, we let $S \in \{1, 2, \ldots, d\}$ denote a subspace. The Euclidean distance between $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ in subspace S is denoted by $dist_S(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{\sum_{i\in S}(x_i - y_i)^2}{|S|}}$ [7]. Let $r > 0$ and $k \in \{1, \ldots, N\}$. Then, the set of neighborhood vectors of $\boldsymbol{x}$ in subspace S is defined by

$$N_S(X, r, \boldsymbol{x}) = \{\boldsymbol{y} \in X : dist_S(\boldsymbol{x}, \boldsymbol{y}) \le r, \boldsymbol{x} \neq \boldsymbol{y}\}.$$

With this definition of the neighboring vectors, the outliers in subspace S are defined by

$$O_S(X, k, r) = \{\boldsymbol{x} \in X : |N_S(X, r, \boldsymbol{x})| < k\}.$$

Then, the task of the outlier count is to find the number of outliers in S:

$$q_{count}(X, k, r, S) = |O_S(X, k, r)|.$$

If the subspace is not specified, then $O(X, k, r)$ denotes the set of outliers in the full space. Distance-based outliers are definable with any type of object and distance defined for the corresponding objects, but we presume that the objects are represented as real vectors and that the Euclidean distance is used as the distance definition.

3.2 Differential Privacy of Outlier Analysis

We introduce several typical scenarios of differentially private outlier analysis using query q_{count}.

Scenario 1. Given threshold k and radius r, presume that the objective is to inspect that the outliers exists in the given dataset. The analyst issues a query $z = q_{count}(X, k, r)$; then checking $z > \theta$ yields the final result where θ denotes a prescribed threshold parameter for outlier counts. Let $z' = q_{count}(X', k, r)$. For guarantee of (ϵ, δ)-differential privacy, we require, for $\forall X' : H(X, X') = 1$ and $\forall T \in \mathcal{T}$,

$$Pr[T = \mathcal{A}(z)] \leq e^{\epsilon} Pr[T = \mathcal{A}(z')] + \delta.$$

Scenario 2. Let the data dimension be $d = 3$. Given threshold k and radius r, presume that the objective is to identify the subspaces that cause the largest numbers of outliers. Then, the target subspace set is $\mathcal{S} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$. The analyst issues query $q_{count}(X, k, r, S_i)$ for each $S_i \in \mathcal{S}$. Let $z_i = q_{count}(X, k, r, S_i)$. For the guarantee of (ϵ, δ)-differential privacy, we require, $\forall X' : H(X, X') = 1$ and $\forall T \in \mathcal{S}$,

$$Pr[T = \mathcal{A}(z_1, \ldots, , z_7)] \leq e^{\epsilon} Pr[T = \mathcal{A}(z_1, \ldots, z_7)] + \delta.$$

4 Differentially Private Count of Outliers

As explained in this section, we investigate the problem of differentially private count of outliers in a given subspace. The discussion herein holds for any subspace including the full space. Therefore, for this discussion, we presume that the outlier is counted in the full space.

4.1 Difficulties in Global Sensitivity Method

Analytical evaluation of the global sensitivity of determination of q_{count} is not trivial, partly because it needs the kissing number. The kissing number K_d is the largest number of hyperspheres with same radius in $\mathbb{R}^d$ that can touch equivalent hyperspheres with no intersections [15–17]. The kissing numbers in $d = 1$ and $d = 2$ are readily derived respectively as $K_1 = 2$ and $K_2 = 6$ (see Fig. 1 for $K_2 = 6$). However, finding the kissing number in $d \geq 3$ is not trivial. In addition, the kissing number in general dimensions remains as an open problem [15–17]. We derive the upper and lower bound of the global sensitivity of q_{count} presuming that the kissing number in general dimensions is given.

Theorem 1 (Upper and lower bound on the global sensitivity of q_{count}). *Let K_d be the kissing number in $\mathbb{R}^d$. Then, the upper and lower bound on the global sensitivity of q_{count} is*

$$\min(N, 2dk + 1) \leq GS_{q_{count},d}(k) \leq \min(N, kK_d + 1). \tag{2}$$

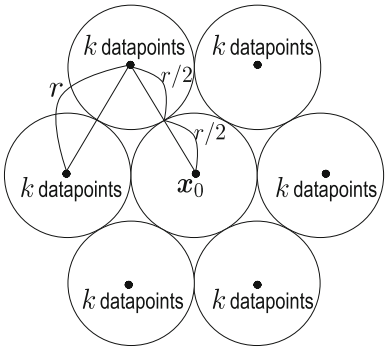


Fig. 1. This figure shows an example of the upper bound of the global sensitivity in two dimension. Six surrounding hyperspheres can be packed around the center hypersphere because the kissing number is $K_2 = 6$. We here suppose k datapoints exist at the center of each surrounding hypersphere and no datapoint exists at $\boldsymbol{x}_0$, the center of the center hypersphere. Then, kK_2 outliers become inliers by adding a point to $\boldsymbol{x}_0$. Suppose the added point is an outlier, Then, the added point can be changed from an outlier to an inlier, too. The upper bound of the global sensitivity for two dimension is thus $kK_2 + 1 = 6k + 1$.

Sketch of Proof. The lower bound is trivial so we omit the proof. We show the sketch of the proof for the upper bound. Suppose the radius of the center hypersphere and the hyperspheres touching the center hyperspheres (referred to as the surrounding hyperspheres) are $r/2$. Let $\boldsymbol{x}_0$ be the center of the center hypersphre. Note that intersection between the surrounding hyperspheres does not exist. We further suppose k datapoints exist at the center of each surrounding hypersphere. These datapoints are outliers by definition, and become inliers by adding a point to the center $\boldsymbol{x}_0$ of the center hypersphere. By definition of the kissing number, the number of the surrounding hyperspheres that do not touch or intersect mutually is at most K_d. No more surrounding hyperspheres can be packed around x_0, so $kK_d + 1$ is the upper bound of the outlier count. Since the global sensitivity is at most N, we can conclude that $GS_{q_{count},d}(k) \leq \min(N, kK_d + 1)$.

We empirically investigate the tightness of the bound in low dimensions. In $d = 1$ and $d = 2$, the global sensitivity is given respectively as $GS_{q_{count},1}(k) = 2k+1$ and $GS_{q_{count},2}(k) = 5k+1$. Noting that $K_1 = 2$ and $K_2 = 6$, the bound is tight in $d = 1$ but not in $d = 2$. Fig. 2 shows the upper and lower bounds of the global sensitivity of q_{count} evaluated using known upper bounds on the kissing number [15–17]. As the figure shows, the upper bound of the global sensitivity grows exponentially with respect to the dimensionality, which indicates that the guarantee of differential privacy by perturbation based on the global sensitivity can be impractical, especially when the dimensionality of the target subspace is large.

The global sensitivity can be prohibitively large simply because the global sensitivity is evaluated considering the worst case. However, one can typically

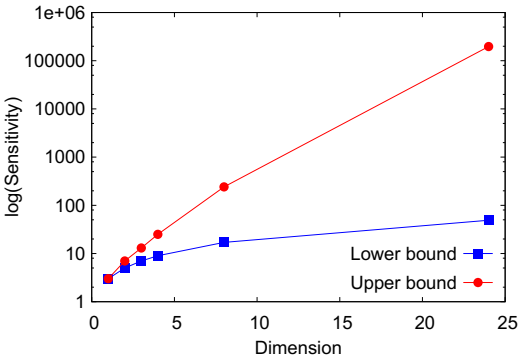


Fig. 2. The bounds of the global sensitivity for counting outliers

expect that the number of outliers in the database is much smaller than the number of instances. To improve the utility of the count query, we introduce the smooth sensitivity, which is a sensitivity definition depending on the database.

4.2 Local Sensitivity and Smooth Sensitivity

For convenience of discussion later, several notations are introduced here. Given radius r, $deg(\boldsymbol{x})$ denotes the size of neighborhoods of $\boldsymbol{x}$:

$$deg(X, r, \boldsymbol{x}) = |N(X, r, \boldsymbol{x})|.$$

We say that the degree of $\boldsymbol{x}$ is k if $deg(X, r, \boldsymbol{x}) = k$. A set of vectors in X whose degree is exactly k is denoted as

$$V(X, k, r) = \{\boldsymbol{x} \in X : deg(\boldsymbol{x}) = k\}.$$

Unless specifically stated otherwise, the radius r and target database X is fixed. Therefore, they are omitted as $deg(\boldsymbol{x})$ and $V(k)$. Finally, a set of degree-k neighborhoods of $\boldsymbol{x}$ in X is denoted as

$$CV(X, \boldsymbol{x}, k, r) = B(\boldsymbol{x}, r) \cap V(k),$$

where $B(\boldsymbol{x}, r)$ denotes the sphere with radius r and centered at $\boldsymbol{x}$.

Local Sensitivity. Given database X, let X_1 be a database s.t. $H(X, X_1) = 1$. Then, following the definition of the local sensitivity in Section 2, the local sensitivity of q_{count} is defined as

$$LS^{(0)}_{q_{count}}(X, k, r) = \max_{X_1 : H(X, X_1) = 1} \|q_{count}(X_0, k, r) - q_{count}(X_1, k, r)\|.$$

Exact evaluation of the exact local sensitivity is intractable. Instead, the following theorem gives the upper bound of the local sensitivity.

Theorem 2. *Given X, the local sensitivity of q_{count} for X is bounded above as*

$$LS^{(0)}_{q_{count}}(X,k,r) \leq \max\left\{\max_{\boldsymbol{x}\in X}\{|CV(X,\boldsymbol{x},k,r)|\}, \max_{\boldsymbol{x}\in\mathbb{R}^d}\{|CV(X,\boldsymbol{x},k-1,r)|\}\right\}+1.$$

Proof. $CV(X,\boldsymbol{x},k,r)$ is the set of non-outliers that become outliers if $\boldsymbol{x}$ is removed; $CV(X,\boldsymbol{x},k-1,r)$ is the set of outliers that become inliers if a vector is placed at $\boldsymbol{x}$. Thus, if vector $\boldsymbol{x}_0 \in X$ is moved to $\boldsymbol{x}'_0$, the number of outliers increases by $|CV(X,\boldsymbol{x}_0,k,r)|$ by removing $\boldsymbol{x}_0$ and the number of inliers decreases by $|CV(X,\boldsymbol{x}'_0,k-1,r)|$ by adding $\boldsymbol{x}'_0$. With this understanding, the local sensitivity is given as:

$$\begin{aligned}
&LS^{(0)}_{q_{count}}(X,k,r)\\
&= \max_{X_1:H(X,X_1)=1} \|q_{count}(X,k,r) - q_{count}(X_1,k,r)\|\\
&\leq \max_{\boldsymbol{x}_0\in X,\boldsymbol{x}'_0\in\mathbb{R}^d} |CV(X,\boldsymbol{x}_0,k,r) \setminus CV(X,\boldsymbol{x}'_0,k-1,r)| + 1\\
&\leq \max_{\boldsymbol{x}_0\in X,\boldsymbol{x}'_0\in\mathbb{R}^d} \max\{|CV(X,\boldsymbol{x}_0,k,r)|, |CV(X,\boldsymbol{x}'_0,k-1,r)|\} + 1\\
&= \max\left\{\max_{\boldsymbol{x}\in X}\{CV(X,\boldsymbol{x},k,r)\}, \max_{\boldsymbol{x}'_0\in\mathbb{R}^d}\{CV(X,\boldsymbol{x},k-1,r)\}\right\}+1.
\end{aligned}$$

Naive evaluation of the local sensitivity is intractable. An algorithm to evaluate this upper bound is presented in Section 4.3.

Smooth Sensitivity. Given database X, let X_t be a database s.t. $H(X,X_t) = t$. By definition, the smooth sensitivity of q_{count} is given as

$$S^*_{q_{count}}(X) = \max_{t=0,1,\ldots,N} e^{-t\beta} LS^{(t)}_{q_{count}}(X),$$

where

$$LS^{(t)}_{q_{count}}(X) = \max_{X_t:H(X,X_t)=t} LS^{(0)}_{q_{count}}(X_t).$$

The function $LS^{(t)}_q(X)$ returns the largest local sensitivity among the datasets of which t records differ from X. Similarly to $LS^{(0)}_{q_{count}}(X)$, exact evaluation of $LS^{(t)}_{q_{count}}(X)$ is intractable because the variation of X_t can increase exponentially with respect to t. Instead, we derive the upper bound on $LS^{(t)}_{q_{count}}(X)$ using $CV(X,\boldsymbol{x},k,r)$.

Theorem 3. *Given X, for $t \geq 0$, $LS^{(t)}_{q_{count}}(X)$ is bounded above as*

$$LS^{(t)}_{q_{count}}(X) \leq \max_{\boldsymbol{x}\in\mathbb{R}^d}\left\{\max\{C^{(t)}(X,\boldsymbol{x},k,r), C^{(t)}(X,\boldsymbol{x},k-1,r)\}+t+1\right\}, \quad (3)$$

where

$$C^{(t)}(X, \boldsymbol{x}, k, r) = \left| \bigcup_{i=-t}^{t} CV(X, \boldsymbol{x}, k+i, r) \right|.$$

For the proof of this theorem, we use the following helper lemma.

Lemma 1. *Let $t \geq 0$ be an integer, and let X and X_t be databases such that $H(X, X_t) = t$. Then, for any $\boldsymbol{x} \in \mathbb{R}^d$, threshold k, and radius r,*

$$|CV(X_t, \boldsymbol{x}, k, r)| \leq \left| \bigcup_{i=-t}^{t} CV(X, \boldsymbol{x}, k+i, r) \right| + t.$$

Proof. We first consider the case $t = 1$. Suppose $\boldsymbol{x} \in X$ is moved from $\boldsymbol{x}$ to $\boldsymbol{x}_1$, and X_1 is given as $X_1 = X \setminus \{\boldsymbol{x}\} \cup \{\boldsymbol{x}_1\}$. The degree of records in $X \setminus \{\boldsymbol{x}\}$ around $\boldsymbol{x}$ decreases by one by removing $\boldsymbol{x}$, and the degree of records in $X \setminus \{\boldsymbol{x}\}$ around $\boldsymbol{x}_1$ increases by one by adding $\boldsymbol{x}_1$. Since the degree of the records in $V(X, k+1, r)$ and $V(X, k-1, r)$ may become k in X_1, $V(X_1, k, r)$ is thus a subset of $V(X, k+1, r) \cup V(X, k, r) \cup V(X, k-1, r) \cup \{x_1\}$. When $t > 1$, for the same reason, $V(X_t, k, r)$ is a subset of $\bigcup_{i=-t}^{t} V(X, k+i, r) \cup \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_t\}$ where $\boldsymbol{x}_1, ..., \boldsymbol{x}_t$ are the records moved from X to X_t. Thus, the size of $CV(X_t, \boldsymbol{x}, r, k)$ is bounded above as

$$\begin{aligned}
|CV(X_t, \boldsymbol{x}, r, k)| &\leq \left| B(\boldsymbol{x}, r) \cap \left\{ \bigcup_{i=-t}^{t} V(X, k+i, r) \cup \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_t\} \right\} \right| \\
&\leq \left| \bigcup_{i=-t}^{t} B(\boldsymbol{x}, r) \cap V(X, k+i, r) \right| + |\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_t\}| \\
&\leq \left| \bigcup_{i=-t}^{t} CV(X, \boldsymbol{x}, k+i, r) \right| + t.
\end{aligned}$$

Sketch of Proof (of Theorem 3). From Theorem 2 and exchangeability of max, letting

$$\begin{aligned}
C^{(t)}_{\text{out}}(X, k, r) &= \max_{X_t : H(X, X_t) = t} \max_{\boldsymbol{x} \in X_t} |CV(X_t, \boldsymbol{x}, r, k)| \text{ and} \\
C^{(t)}_{\text{in}}(X, k-1, r) &= \max_{X_t : H(X, X_t) = t} \max_{\boldsymbol{x} \in \mathbb{R}^d} |CV(X_t, \boldsymbol{x}, r, k-1)|
\end{aligned}$$

yields

$$LS^{(t)}_{q_{count}}(X) \leq \max\{C^{(t)}_{\text{out}}(X, k, r), C^{(t)}_{\text{in}}(X, k-1, r)\} + 1.$$

We derive the bound on $C^{(t)}_{\text{out}}(X, k, r)$ using $C^{(t)}_{\text{in}}(X, k-1, r)$, and the bound on $C^{(t)}_{\text{in}}(X, k-1, r)$ using Lemma 1.

4.3 Efficient Computation of Smooth Sensitivity Bound

For randomization by the mechanism of Theorem 3, it is necessary to evaluate the smooth upper bound. Naive evaluation of the smooth upper bound of eq. (3) is intractable because it requires an exhaustive search over continuous domain to evaluate $LS^{(t)}_{q_{count}}(X)$. To alleviate this, we first show an efficient algorithm that evaluates the upper bound of $LS^{(t)}_{q_{count}}(X)$ shown derived by Theorem 3. Then using the algorithm, we derive the algorithm that calculates the smooth sensitivity upper bound.

Algorithm for Local Sensitivity Bound. To evaluate the upper bound of $LS^{(t)}_{q_{count}}(X)$, we need to calculate

$$\max_{\boldsymbol{x}\in\mathbb{R}^d} C^{(t)}(X,\boldsymbol{x},k,r) = \max_{\boldsymbol{x}\in\mathbb{R}^d} \left| \bigcup_{i=-t}^{t} V(X,k+i,r) \cap B(\boldsymbol{x},r) \right|, \text{ and} \tag{4}$$

$$\max_{\boldsymbol{x}\in\mathbb{R}^d} C^{(t)}(X,\boldsymbol{x},k-1,r) = \max_{\boldsymbol{x}\in\mathbb{R}^d} \left| \bigcup_{i=-t}^{t} V(X,k+i-1,r) \cap B(\boldsymbol{x},r) \right|. \tag{5}$$

Letting $P = \bigcup_{i=-t}^{t} V(X,k+i,r)$ (resp. $P = \bigcup_{i=-t}^{t} V(X,k+i-1,r)$), we can obtain the value of eq. (4) (resp. eq. (5)) by finding the largest subset $C \subseteq P$ that is enclosed by a ball with radius r. To check whether or not a given subset $C \subseteq P$ is enclosed by the ball, we use the algorithm that solves the *smallest enclosing ball* (seb) problem [6]. The goal of the problem is to find the smallest ball that encloses the given points. The given subset $C \subseteq P$ is enclosed by a ball with radius r if $\mathrm{seb}(C) \leq r$ where $\mathrm{seb}(C)$ denotes the radius of the resultant ball of the smallest enclosing ball problem of C.

Algorithm 1 shows the recursive algorithm that calculates eq. (4) or eq. (5) for given $P = \bigcup_{i=-t}^{t} V(X,k+i,r)$ or $P = \bigcup_{i=-t}^{t} V(X,k+i-1,r)$. $P[i]$ denotes the i-th element of the set P. Algorithm 1 searches for the largest subsets $C \subseteq P$ that is enclosed by a ball with radius r with the breadth-first search. In the algorithm, the calls of seb can be skipped for efficiency by using the fact that the radius of the enclosing ball of C_2 is larger than one of C_1 if $C_1 \subseteq C_2 \subseteq P$. The computational cost of Algorithm 1 is $\mathcal{O}(2^{|P|})$ of the calls of seb.

Algorithm for Smooth Sensiticity Bound. Algorithm 1 costs exponential time with respect to $|P|$ and the size of P increases monotonically as t increases. However, because of exponential decrease of $e^{-t\beta}$, maximization of $e^{-t\beta} LS^{(t)}_{q_{count}}(X)$ is attained by small t in most cases. Taking account of this property, we provide Algorithm 2 that calculates the smooth sensitivity bound with avoiding evaluation of $LS^{(t)}_{q_{count}}(X)$ of large t.

Proposition 1. *For any t and $t' < t$, $LS^{(t)}_{q_{count}}$ is bounded above as*

$$LS^{(t)}_{q_{count}}(X) \leq \min\{N, \max\{U^{(t)}_{t'}(X,k,r), U^{(t)}_{t'}(X,k-1,r)\} + t + 1\},$$

Algorithm 1. Calculation of $\max_{\boldsymbol{x}\in\mathbb{R}^d} C^{(t)(X,\boldsymbol{x},k,r)}$(eq. (4) and eq. (5))

Input: Records P and radius r.
Output: The value of eq. (4) or eq. (5).
Initialization: $C = \emptyset$ and $i = 1$

```
Function E(r, P, C, i)
    br ← 0
    if C ≠ ∅ then
        br ← seb(C)
    end
    if br ≤ r then
        m ← |C|
        if i ≤ |P| then
            b1 ← E(r, P, C ∪ {P[i]}, i + 1)
            b2 ← E(r, P, C, i + 1)
            m ← max{m, b1, b2}
        end
        return m
    end
    else
        return 0
    end
end
```

where

$$U_{t'}^{(t)}(X,k,r) = \max_{\boldsymbol{x}\in\mathbb{R}^d} C^{(t')}(X,\boldsymbol{x},k,r) + \left| \bigcup_{i\in\{-t,\ldots,-t'-1\}\cup\{t'+1,\ldots,t\}} V(X,k+i,r) \right|.$$

Sketch of Proof. For any database X, because the number of outliers does not exceed the number of the records in X, the local sensitivity is less than N. In addition, using the fact that $CV(X,\boldsymbol{x},k,r) \subseteq V(X,k,r)$ for any $\boldsymbol{x}\in\mathbb{R}^d$, we can derive $\max_{\boldsymbol{x}\in\mathbb{R}^d} C^{(t)}(X,\boldsymbol{x},k,r) \le U_{t'}^{(t)}(X,k,r)$ for any t and $t' < t$.

Using the bound in Proposition 1, we have the upper bound of $e^{-t\beta} LS_{q_{count}}^{(t)}(X)$ as

$$\begin{aligned} e^{-t\beta} LS_{q_{count}}^{(t)}(X) \le & e^{-t\beta} \min\{N, \max\{U_{t'}^{(t)}(X,k,r), U_{t'}^{(t)}(X,k-1,r)\} + t + 1\} \\ =: & S_{\mathrm{UB}}^{t',t}(X). \end{aligned}$$

Letting $S_{\mathrm{UB}}^{t}(X) = \max_{i=1,\ldots,N-t} S_{\mathrm{UB}}^{t,t+i}(X)$, we can obtain the following proposition.

Proposition 2. *If there exists* U_T *such that* $\max_{t=0,\ldots,T} e^{-t\beta} LS_{q_{count}}^{(t)}(X) \le U_T$ *and* $S_{\mathrm{UB}}^{T}(X) \le U_T$*, then* $S_{q_{count}}^{*}(X) \le U_T$.

Algorithm 2. Calculation of the smooth sensitivity of q_{count}

Input: Database X, threshold k, radius r and smooth parameter ϵ.
Output: The smooth sensitivity upper bound of query q_{count} for database X.
Initialization: $S_{\max} = 0$ and
$\max_{\boldsymbol{x}\in\mathbb{R}^d} C^{(-1)}(X, \boldsymbol{x}, k, r) = \max_{\boldsymbol{x}\in\mathbb{R}^d} C^{(-1)}(X, \boldsymbol{x}, k-1, r) = 0.$

for $t = 0$ *to* N **do**
  Calculate S_{UB}^{t-1} by Proposition 2
  if $S_{\mathrm{UB}}^{t-1} \leq S_{\max}$ **then**
    return $S_{\max}$
  end
  $S_{\max} \leftarrow \max\{S_{\max}, e^{-t\beta} LS_{q_{count}}^{(t)}(X)\}$
  Store $\max_{\boldsymbol{x}\in\mathbb{R}^d} C^{(t)}(X, \boldsymbol{x}, k, r)$ and $\max_{\boldsymbol{x}\in\mathbb{R}^d} C^{(t)}(X, \boldsymbol{x}, k-1, r)$ for calculating S_{UB}^{t} in next loop
end
return $S_{\max}$

Proof. If $S_{\mathrm{UB}}^{T}(X) = \max_{i=1,\ldots,N-T} S_{\mathrm{UB}}^{T,T+i}(X) \leq U_T$, since $e^{-t\beta} LS_{q_{count}}^{(t)}(X) \leq S_{\mathrm{UB}}^{T,t}(X)$ for any $t > T$, we have $e^{-t\beta} LS_{q_{count}}^{(t)}(X) \leq U_T$, $\forall t > T$. Thus, we have $\max_{t=0,\ldots,T} e^{-t\beta} LS_{q_{count}}^{(t)}(X) \leq U_T$ and $\max_{t>T} e^{-t\beta} LS_{q_{count}}^{(t)}(X) \leq U_T$.

Proposition 2 shows that if the largest upper bound in Theorem 3 for $t = 0, ..., T$ can be bounded above by $S_{\mathrm{UB}}^{T}(X)$, then the calculation of the upper bound in Theorem 3 for $t > T$ can be skipped. Algorithm 2 shows the calculation of the smooth sensitivity of q_{count} with this skip by following Proposition 2.

5 Experiments

In this section, we show the empirical evaluation of the utility of the mechanism for counting outliers query.

5.1 Settings

We used a synthetic dataset and a real dataset (adult). The synthetic dataset consists with 50 samples of 2 dimensional real vectors. The dataset contains 45 inliers which are sampled from $\mathcal{N}(\mathbf{0}, \mathbb{I})$ where $\mathbb{I}$ represents an identity matrix. The 5 outliers are sampled from $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\mu_1 = \mu_2 = 20$ and Σ is a diagonal matrix such that $\Sigma_{11} = \Sigma_{22} = 100$.

A real dataset (adult) was chosen from UCI Machine Learning Repository [13]. We removed two categorical attributes, "category" and "fnlwgt". The dataset was scaled so that the average and variance of each attribute is 0 and 1, respectively. The dataset is originally prepared for classification tasks. For our outlier analysis, following [19,22], 45 samples with the positive label are treated as inliers and 5 samples with negative labels were treated as outliers (See Table 1 for the detail). We changed the privacy parameter from $\epsilon = 0.1$ to 0.9; δ was

Table 1. Sumarry of datasets

	synthetic	adult
The number of outliers	5	5
The number of inliers	45	45
The number of samples N	50	50
Dimension d	2	7
Treshold k	3	3
Radious r	1.1	0.35

fixed as $\delta = 0.01$. See Table 1 for the parameters of the outliers. We partitioned the instances into two classes: one is "true", indicating the instance detected as an outlier; the other is "false". For each dataset, we tuned the radius r so that the *Accuracy* given by eq. (6) is maximized:

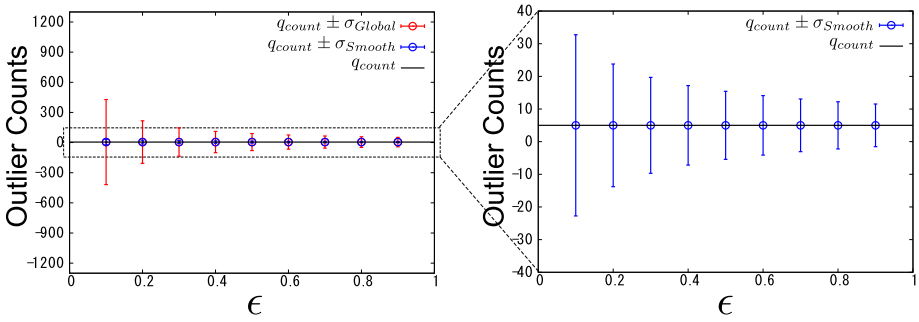
$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}, \tag{6}$$

where TP, TN, FP and FN respectively denote true positive, true negative, false positive, and false negative. For implementation, we used [11] to solve the smallest enclosing ball problem.

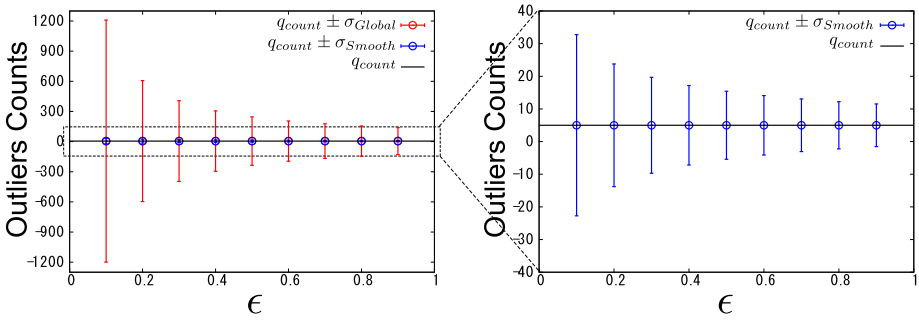
5.2 Count Outliers

Following the Scenario 1 described in Section 3.2, we evaluated the utility of the mechanisms of q_{count} on the synthetic dataset. As the criterion of the utility of the mechanisms, we show the standard deviation of the noise added to the query. We compared the standard deviation of the noise of the mechanism based on the smooth sensitivity upper bound in eq. (3) with the mechanism based on the global sensitivity lower bound in eq. (2). Fig. 3 shows the output values and the standard deviations for each mechanism in various ϵ. In Fig. 3, "Global" and "Smooth" respectively present the global sensitivity-based mechanism and the smooth sensitivity-based mechanism.

It is apparent that the standard deviation of the noise of the smooth sensitivity-based mechanism is significantly lower than that of the global sensitivity-based mechanism. Indeed, the standard deviation of the noise of global sensitivity-based mechanism is approximately 10-30 times larger than that of the smooth sensitivity-based mechanism even though the global sensitivity-based mechanism uses the lower bound. In addition, the smooth sensitivity-based mechanism achieves the noise of which standard deviation is lower than 7 for $\epsilon \geq 0.7$ for each datasets. The reason why we got these results is our approach depends only on the number of outliers, not on the number of dimensions. From these results, we can conclude that our framework is sufficiently practical in this setting.



(a) The result of the synthetic dataset



(b) The result of the adult dataset

Fig. 3. Experimental results for the global sensitivity-based mechanism and the smooth sensitivity-based mechanism on each dataset. The right panel is obtained by scaling the left panel so that the error bars of the smooth sensitivity-based mechanism are visible. The horizontal axis denotes the privacy parameter ϵ. The vertical axis denotes the output value of the query without randomization. The error bars denote the standard deviation of the noise added by the mechanisms.

6 Conclusion and Future Works

We present the differentially private distance-based outlier analysis for the query that counts outliers in a given subspace. Taking advantage of the smooth sensitivity [18], the resulting output of the mechanism can be less noisy than that of the global sensitivity-based mechanism. Although the evaluation of the smooth upper bound is often costly, we provide an efficient algorithm for the evaluation of the smooth upper bound for the problem for outlier counting. This paper describes an initial step towards differentially private outlier analysis, and the experimental evaluation is performed with relatively small-size datasets. In our algorithm, we invoke the smallest enclosing ball algorithm that takes as input the power set of instances. Because of this construction, we need a more efficient algorithm for application to larger size datasets.

Subspace discovery for outlier analysis has been investigated as a major topic of outlier detection [7,8,10]. Differentially private subspace discovery can be

achieved by issuing count queries sequentially to each subspace; however, the number of subspaces increases exponentially with respect to the dimensionality, which costs a large amount of privacy budget. An efficient mechanism for subspace discovery is left as an area of the future work.

Acknowledgments. This research was supported by KAKENHI 24680015, JST CREST *Advanced Core Technologies for Big Data Integration*, and the program *Research and Development on Real World Big Data Integration and Analysis* of the MEXT, Japan.

References

1. Bao, H.T., et al.: A distributed solution for privacy preserving outlier detection. In: Proceedings of the 2011 Third International Conference on Knowledge and Systems Engineering, pp. 26–31. IEEE Computer Society (2011)
2. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: privacy via distributed noise generation. In: Vaudenay, S. (ed.) EUROCRYPT 2006. LNCS, vol. 4004, pp. 486–503. Springer, Heidelberg (2006)
3. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006)
4. Dwork, C., Smith, A.: Differential privacy for statistics: What we know and what we want to learn. Journal of Privacy and Confidentiality **1**(2), 2 (2010)
5. Fan, L., Xiong, L.: Differentially private anomaly detection with a case study on epidemic outbreak detection. In: Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops, pp. 833–840. IEEE Computer Society (2013)
6. Fischer, K., Gärtner, B., Kutz, M.: Fast smallest-enclosing-ball computation in high dimensions. In: Di Battista, G., Zwick, U. (eds.) ESA 2003. LNCS, vol. 2832, pp. 630–641. Springer, Heidelberg (2003)
7. Keller, F., Müller, E., Böhm, K.: Hics: high contrast subspaces for density-based outlier ranking. In: IEEE 28th International Conference on Data Engineering (ICDE 2012), Washington, DC, USA (Arlington, Virginia), 1–5 April, 2012, pp. 1037–1048. IEEE Computer Society (2012)
8. Keller, F., Müller, E., Wixler, A., Böhm, K.: Flexible and adaptive subspace search for outlier analysis. In: 22nd ACM International Conference on Information and Knowledge Management, CIKM 2013, San Francisco, CA, USA, October 27 - November 1, 2013, pp. 1381–1390. ACM (2013)
9. Knorr, E.M., Ng, R.T.: Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the 24rd International Conference on Very Large Data Bases. pp. 392–403. VLDB 1998, Morgan Kaufmann Publishers Inc., San Francisco, CA (1998)
10. Knorr, E.M., Ng, R.T.: Finding intensional knowledge of distance-based outliers. In: Proceedings of the 25th International Conference on Very Large Data Bases, pp. 211–222. VLDB 1999, Morgan Kaufmann Publishers Inc., San Francisco, CA (1999)
11. Kutz, M., Kaspar, F., Bernd, G.: A java library to compute the miniball of a point set. https://github.com/hbf/miniball, last Accessed Time: February 2, 2015

12. Li, L., Huang, L., Yang, W., Yao, X., Liu, A.: Privacy-preserving lof outlier detection. Knowledge and Information Systems **42**(3), 579–597 (2015)
13. Lichman, M.: UCI machine learning repository (2013). http://archive.ics.uci.edu/ml
14. Lui, E., Pass, R.: Outlier privacy. In: Dodis, Y., Nielsen, J.B. (eds.) TCC 2015, Part II. LNCS, vol. 9015, pp. 277–305. Springer, Heidelberg (2015)
15. Mittelmann, H.D., Vallentin, F.: High-accuracy semidefinite programming bounds for kissing numbers. Experimental Mathematics **19**(2), 175–179 (2010)
16. Musin, O.R.: The kissing problem in three dimensions. Discrete & Computational Geometry **35**(3), 375–384 (2006)
17. Musin, O.R.: The kissing number in four dimensions. Annals of Mathematics **168**(1), 1–32 (2008)
18. Nissim, K., Raskhodnikova, S., Smith, A.: Smooth sensitivity and sampling in private data analysis. In: Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing, pp. 75–84. STOC 2007. ACM, New York (2007)
19. Pham, N., Pagh, R.: A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 877–885. KDD 2012. ACM, New York (2012)
20. Vaidya, J., Clifton, C.: Privacy-preserving outlier detection. In: The Fourth IEEE International Conference on Data Mining, pp. 233–240. IEEE Computer Society, Brighton (2004)
21. Xue, A., Duan, X., Ma, H., Chen, W., Ju, S.: Privacy preserving spatial outlier detection. In: Proceedings of the 9th International Conference for Young Computer Scientists, pp. 714–719. IEEE Computer Society (2008)
22. Zhang, K., Hutter, M., Jin, H.: A new local distance-based outlier detection approach for scattered real-world data. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 813–822. Springer, Heidelberg (2009)