差分プライバシを保証した外れ値検出

岡田　莉奈† 　福地　一斗† 　佐久間　淳††,†††

† 筑波大学システム情報工学研究科　〒 305-8573 つくば市天王台 1-1-1
†† 筑波大学システム情報系　〒 305-8573 つくば市天王台 1-1-1
††† 科学技術振興機構 CREST　〒 102-0076 東京都千代田区五番町 7K's 五番町
E-mail: †{rina,kazuto}@mdl.cs.tsukuba.ac.jp, ††jun@cs.tsukuba.ac.jp

あらまし　外れ値検出はデータ解析における重要な問題である．しかし，特異な値を外れ値として検出することそのものがプライバシ侵害を招きかねず，外れ値検出におけるプライバシ保護の実現は困難な問題である．本研究では，外れ値の個数を検出するクエリに注目する．個数のような統計値からプライバシを保護する指標として差分プライバシ [1] がある．我々は外れ値の個数を検出するクエリに対して差分プライバシを保証する方法を検討する．差分プライバシを保証するために大域敏感度によるメカニズムが一般的に用いられる．しかしながら，外れ値検出にこの一般的なメカニズムを用いると出力結果の有用性を下げてしまう．この問題を克服するために，大域敏感度の代わりに平滑敏感度 [2] を用いる手法を理論的に示し，差分プライバシを満たしつつ高い有用性を持つ出力結果を得るようにする．実験では大域敏感度を用いたときと平滑敏感度を用いたときの出力結果の有用性を比較する．

キーワード　外れ値検出，差分プライバシ，平滑敏感度

1. はじめに

データベース内の異常検知は重要な問題であり，異常検知は外れ値を検出することで実現可能である．しかし，しばしばデータベースには個人の機密情報が含まれており，データの取り扱いには注意が必要となる．外れ値の目的は，外れ値となるインスタンスの特定であるが，外れ値が明らかになることによって個人情報が漏れる可能性がある．そこで，データベースはプライバシを保護した外れ値の開示をしなければならない．

データベースのプライバシを保護しながら分析を行なう場合，プライバシ保護方法として匿名化と摂動法の 2 種類がある．匿名化とは，対象としているデータベース内に同じインスタンスが複数存在するようにすることによって，あるインスタンスが特定される確率を 1 以下に低める方法である．この確率を $\frac{1}{k}$ に低める場合，k 匿名化 [3] という．同じインスタンスが複数存在するようにするためにはインスタンスの一般化を行なうことが多い．一方，摂動法は，データベースとの統計的クエリ応答が対話的である際に用いられるプライバシ保護の枠組みである．その中でも出力摂動法では，データベース内の情報には何も変更を加えずに出力結果に対してノイズを加える．外れ値は他のインスタンスとは性質が異なるものなので，同じインスタンスを複数存在させるような匿名化は適用することができない．そこで，我々は出力摂動法の一つである差分プライバシ [1] を用いて外れ値検出を行なう．

外れ値の個数検出の例として，外れ値検出をセンサデータを通じたシステム監視における特異事象の発生検出に利用する場合を考える．このとき，外れ値検出の目的をその特異さを把握することであると解釈すれば，外れ値となるインスタンスを一意に特定する必要は必ずしもない．このように外れ値の発生個数や外れ値を引き起こしたデータ属性の把握などができれば十分な場合がある．また，プライバシ保証下でのデータ解析が必要な例として，外れ値検出を個人の安全保護を目的とした行動監視に使うような場合を考える．緊急時にはプライバシ保護よりも安全保護の方が重要視される．ただし，だからといって通常時における個人の詳細な行動情報の継続的監視にプライバシ上の懸念がある．緊急事態発生の検出はプライバシ保護を考慮しつつ行い，一度緊急事態発生が検出された後にはプライバシ保護を考慮せずに個人を特定し，必要な措置を測ることが望ましい．本研究ではこのような応用を踏まえ，プライバシ保護をしつつ，より正確な外れ値検出の両立を目指す初段階として，外れ値個数の検出クエリから差分プライバシを保証する．

1.1 関連研究

本節では，本研究に関連する外れ値検出に関する研究，差分プライバシに関する研究，差分プライバシを保証した異常検出に関する研究の順に紹介する．

外れ値検出方法は大きく分けて三種類ある．一種類目は，統計に基づく外れ値検出である．統計に基づく外れ値検出には様々あるが，その中でもグラブス検定 [4] がよく使われる．グラブス検定では正規分布を仮定する．1 検定で外れ値を 1 つ検出し，検出した外れ値をデータセットから取り除き，外れ値がなくなるまで繰り返す．このとき対立仮説を “そのデータセット内に外れ値が存在しない”，帰無仮説を “そのデータセット内に少なくともひとつは外れ値が存在する” として棄却検定を行う．しかし，グラブス検定では多次元ベクトルを扱うことが困難である．二種類目は，教師付き学習よる外れ値検出である．教師付き学習による外れ値検出も様々あるが，その中でも SVM を用いた外れ値検出 [5] が有名である．これは，過去の学習データセットから外れ値と正常値の境界線を学習し，今後の予測を行うというものであるり，多次元ベクトルにも対応できる．しかし，教師付き学習では大量のラベル付き学習データ

セットが取得することが必要であり，外れ値は正常値に比べて入手が困難である．三種類目は，距離に基づく外れ値検出である．距離に基づく外れ値検出も様々あるが，これまで多くの研究 [6,7] が Edwin らの論文 [8,9] 内の外れ値の定義を使用している．この外れ値の定義は一般的であり，任意の距離空間で使用できる．このような調査から我々は本研究で Edwin らの距離の定義を用いることにした．

差分プライバシ [1] は，Dwork らによって提案されたプライバシ保護指標の一つである．差分プライバシを保証していれば，統計値から個人のプライバシが保護されている．これまでに様々な統計値から個人のプライバシを保護する研究が行われている [1,2,10]．[1] では，平均・総和・分散・ヒストグラムクエリに対する差分プライバシの適用方法が示されており，[2] では，中央値クエリに対する差分プライバシの適用方法が示されており，[10] では，最尤推定量クエリに対する差分プライバシの適用方法が示されている．このように差分プライバシは統計値の公開において理論的に健全なプライバシ定義として広く認知されつつあるので，本研究でもプライバシ保護指標として差分プライバシを用いる．

外れ値検出における差分プライバシの保証に関する研究はその実現の困難さから多くない．差分プライバシを保証した異常検出に関する研究 [11] では，時系列データの一次元インスタンスに対して差分プライバシを保証した異常検出が行えるフレームワークを提案している．このフレームワークでは，フィルタリングという操作によってクエリ結果に一度ノイズを加えたものに対して出力結果の有用性が向上させる．一方，我々の手法は定常データの多次元インスタンスに対して差分プライバシを保証した外れ値検出を行うため，[11] とは取り組む状況が異なる．

1.2 貢　　献

定常データの多次元インスタンス (以降，ベクトルと呼ぶ) に対して外れ値検出を行う．我々の知る限り，定常データの多次元ベクトルの差分プライバシを保証しつつ外れ値検出を行う研究は本研究が初めてである．差分プライバシを達成するために外れ値の個数検出クエリ (以降，外れ値個数クエリと呼ぶ) 結果に加えるノイズを抑制するために平滑敏感度 [2] を用いることに注目した．本研究の貢献は次の三つである．一つ目は，差分プライバシを達成するための外れ値個数クエリに対する大域的敏感度の下限を示し，一般的に解析が困難な局所的敏感度の上限，平滑敏感度の上限を理論的に示したことである．二つ目は，計算量を緩和したことである．外れ値個数クエリに対する局所敏感度と平滑敏感度の計算量は指数計算量となる．そこで，指数計算量を緩和するアルゴリズムを提案し，次元数を d とすると $\mathcal{O}(d^2)$ 程度の計算量に抑えられるようにしたことである．三つ目は，平滑敏感度の上限を用いることは高い有用性を持つ出力結果を得られることを実験的に示したことである．

本稿の構成は，次のようになっている．まず，2. 章と 3. 章で本研究で用いる外れ値検出と差分プライバシの紹介をする．次に，4. 章で外れ値個数検出クエリに対して差分プライバシ適用するために用いる敏感度について理論評価し，5. 章で敏感度を用いた際の有用性評価をする．最後に 6. 章でまとめと今後の課題について述べる．

2. 外れ値検出

外れ値検出として距離に基づいた方法 [8] がよく使われており，我々も文献 [8] 内に記載されている外れ値の定義を用いる．データベース $X = \{\boldsymbol{x}_i\}$，ベクトル $\boldsymbol{x}_i \in \mathbb{R}^d (d$ は次元数), データベースサイズ $|X| = N$ とする．このとき，外れ値の定義を定義 1 に示す．以降，$d(\cdot,\cdot)$ はユークリッド距離関数とする．

［定義 1］（外れ値） 式 (1) に示すベクトル集合 $O(X,k,r)$ 内に含まれるベクトルは外れ値である．

$$O(X,k,r) = \{\boldsymbol{x}_i \in X || \{\boldsymbol{x}_j \in X \backslash \boldsymbol{x_i} | d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leqq r\}| < k\} \tag{1}$$

ただし，k と r は，ユーザ指定のパラメータだが，$k \geqq 1$，$r > 0$ とする．

本稿では，データベース内の外れ値でないベクトルを正常値と呼ぶ．また，今回，我々が注目している外れ値の個数を検出するクエリ $q_{out} : X \to \mathbb{Z}$ を式 (2) に示す．

$$q_{out}(X,k,r) = |O(X,k,r)| \tag{2}$$

次章以降では，この外れ値個数クエリからプライバシ保護する方法について議論する．

3. 差分プライバシ

差分プライバシはデータベースとの統計的クエリ応答を対話的に行うことを想定した場合に用いられるプライバシ保護指標である．情報取得者が統計的クエリ $q : X \to \mathcal{T}$ をデータベースに問い合わせ，データベースが統計値を返答をする際にプライバシを保護するノイズを加えるメカニズム $\mathcal{A} : \mathcal{T} \to \mathcal{T}$ を用いる．ただし，$\mathcal{T}$ は出力結果の範囲である．本研究で外れ値の個数を検出する際に用いる (ϵ,δ)-差分プライバシ [12] について紹介する．そのために，ハミング距離を導入する．ハミング距離を定義 2 に示す．

［定義 2］（ハミング距離 [1]） データベース X と X' のハミング距離はベクトル $\boldsymbol{x}_i$ と $\boldsymbol{x}'_i$ が異なる数，つまり，

$$H(X,X') = |\{i : \boldsymbol{x}_i \neq \boldsymbol{x}'_i\}| \tag{3}$$

である．X と X' は $H(X,X') = 1$ のとき近隣データベースである．ここで，$|X| = |X'| = N$ とする．

(ϵ,δ)-差分プライバシでは，近隣データベースを用いる．(ϵ,δ)-差分プライバシの定義を定義 3 に示す．

［定義 3］（(ϵ,δ)-差分プライバシ） $H(X,X') = 1$ なる $\forall(X,X')$，$\forall o \in \mathcal{T}$ について式 (4) を満たせば，メカニズム $\mathcal{A}$ は (ϵ,δ)-差分プライバシを保証する．

$$Pr[\mathcal{A}(X) \in o] \leqq e^{\epsilon} Pr[\mathcal{A}(X') \in o] + \delta \tag{4}$$

ただし，X と X' は任意の近隣データベースである．また，ϵ と δ はプライバシパラメータであり，$\delta = 0$ のとき ϵ-差分プライバシ [1] となる．

本研究では，プライバシ指標として定義 3 に示した (ϵ, δ)-差分プライバシを用いる．3.1 節，3.2 節では，クエリ結果がプライバシに及ぼす影響力である敏感度について紹介する．

3.1 大域敏感度

敏感度の一つとして，一般的に大域敏感度 [1] が用いられる．大域敏感度を定義 4 に示す．以降では，出力範囲を $\mathbb{R}^d$ とする．

［定義 4］（大域敏感度） データベースのドメインを $\mathcal{D}$ とすると，クエリ $q : \mathcal{D}^N \to \mathbb{R}^d$ に対する大域敏感度は，式 (5) である．

$$GS_q = \max_{X, X' : H(X, X') = 1} ||q(X) - q(X')|| \tag{5}$$

大域敏感度を用いて式 (4) を満たすガウシアンメカニズムがある．このガウシアンメカニズムが (ϵ, δ)-差分プライバシを満たすことを定理 1 に示す．

［定理 1］（大域敏感度を用いたガウシアンメカニズム [12]） クエリ $q : \mathcal{D}^N \to \mathbb{R}^d$ とすると，式 (6) に示すメカニズムは (ϵ, δ)-差分プライバシを満たす．

$$\mathcal{A}_q(X) = q(X) + Y \tag{6}$$

ただし，Y は平均 0，分散 $\frac{GS_q^2 \cdot 2 \log(2/\delta)}{\epsilon^2}$ の正規分布から $i.i.d.$ にサンプリングしたノイズである．

(ϵ, δ)-差分プライバシは定義 3 より，プライバシパラメータ ϵ や δ を小さくするほどプライバシをより厳しく保証することができる．しかし，定理 1 より，ϵ を小さくすると分散が大きくなり，クエリ結果に加えるノイズが大きくなるので，出力結果の有用性が低くなる．つまり，ϵ と出力結果の有用性にはトレードオフの関係がある．また，一般的にはデータベース中の外れ値の個数はデータ数に比べてごくわずかであるため，ノイズの分散が大きくなることを避けなければならない．しかし，大域敏感度はクエリ q にのみ依存して表されるために，中央値のように大域敏感度が非常に大きくなるクエリもある．そこで，敏感度をデータベースに依存させることによってクエリ結果に加えるノイズを小さく抑える方法として平滑敏感度 [2] を用いることがある．平滑敏感度を次節で紹介する．

3.2 平滑敏感度

平滑敏感度を紹介するために局所敏感度を紹介する．局所敏感度は，あるデータベース X が与えられたときに X の近隣データベース X' に対する敏感度のことである．局所敏感度を定義 5 に示す．

［定義 5］（局所敏感度） クエリ $q : \mathcal{D}^N \to \mathbb{R}^d$ に対する局所敏感度は，式 (7) である．ただし，データベースのドメインを $\mathcal{D}$ とする．

$$LS_q(X) = \max_{X' : H(X, X') = 1} ||q(X) - q(X')|| \tag{7}$$

しかし，文献 [2] にもあるように局所敏感度は (ϵ, δ)-差分プライバシを満たすことができない．よって，局所敏感度を用いて平滑敏感度を考える必要がある．平滑敏感度を定義 6 に示す．

［定義 6］（平滑敏感度） $\beta > 0$ のとき，与えられたデータベース X に対するハミング距離が t であるデータベースを X' とすると，クエリ $q : \mathcal{D}^N \to \mathbb{R}^d$ に対する平滑敏感度は，式 (8) である．ただし，データベースのドメインを $\mathcal{D}$ とする．

$$S^*_{q,\beta}(X) = \max_{X' \in \mathcal{D}^N} (LS_q(X') \cdot e^{-\beta H(X, X')}) \tag{8}$$

平滑敏感度を用いて式 (4) を満たすガウシアンメカニズムがある．このメカニズムが (ϵ, δ)-差分プライバシを満たすことを定理 2 に示す．

［定理 2］（平滑敏感度を用いたガウシアンメカニズム [2]） クエリ $q : \mathcal{D}^N \to \mathbb{R}^d$ とすると，$\alpha = \frac{\epsilon}{5\sqrt{2 \ln 2/\delta}}, \beta = \frac{\epsilon}{4(d + \ln 2/\delta)}$ のとき，式 (9) に示すメカニズムは (ϵ, δ)-差分プライバシを満たす．

$$\mathcal{A}_q(X) = q(X) + \frac{S^*_{q,\beta}(X)}{\alpha} \cdot Y \tag{9}$$

ただし，Y は平均 0，分散 1 の正規分布から $i.i.d.$ にサンプリングしたノイズである．

4. 章ではこれらの大域敏感度，局所敏感度，平滑敏感度の定義を用いる．実験では式 (6)，式 (9) に示したメカニズムを用いる．

4. 外れ値個数の敏感度解析

本節では，外れ値個数クエリ q_{out} の敏感度評価を行うために大域敏感度，局所敏感度，平滑敏感度を導く．

4.1 外れ値個数の大域敏感度

以降の議論で用いるので，次数の概念を定義 7 に示す．

［定義 7］（次数） ベクトル $\boldsymbol{x}_i \in X$ の次数 $deg(\boldsymbol{x}_i)$ を式 (10) とする．

$$deg(\boldsymbol{x}_i) = |\{\boldsymbol{x}_j \in X | d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leqq r, i \neq j\}| \tag{10}$$

任意次元の外れ値クエリの大域敏感度の導出は未解決問題である接吻数 [13, 14] と関連しており，導出が困難である．そこで，大域敏感度の下限を導出する．定義 4 に従って外れ値個数クエリの大域敏感度の下限を定理 3 に示す．

［定理 3］（外れ値個数クエリの大域敏感度の下限） d 次元のユークリッド空間の点を考えると，外れ値個数クエリの大域敏感度 $GS_{q_{out}}$ の下限は式 (11) のようになる．

$$GS_{q_{out}}(k) \geqq \min(N, 2d(k-1) + 1) \tag{11}$$

証明:まず，外れ値の個数の変化は最大で N 個なので，$2d(k-1)+1$ が N を超える場合は N であることは自明である．次に，$2d(k-1)+1$ が N を超えない場合について証明する．任意のデータベース X から X の任意の近隣データベース X' に変化する際の外れ値の個数の最大変化量を考える．X から X' への変更は，あるベクトル $\boldsymbol{x}_0$ が元の位置 $\boldsymbol{x}_0$ から別の位置 $\boldsymbol{x}'_0$ への移動と解釈できる．d 次元のユークリッド空間において，第 i 次元の値のみが r(または $-r$) であり，それ以外の次元の値が 0 である点を p_i^+(または p_i^-) とする．任意の異なる 2 点 $p_i^{\pm}, p_j^{\pm}$ 間の距離は $\sqrt{2}r$ 以上である．ここで，原点にベクトル $\boldsymbol{x}_0$ が 1 つ，各点 $p_i^{\pm}$ に $k-1$ 個のベクトルが重なって存在するデータベース X を考える．ただし，$|X| = 2d(k-1)+1$ とする．このとき，$\boldsymbol{x}_0$ を X 内の他のどのベクトルとも r 以上離れている点に移動させると，各点 $p_i^{\pm}$ に重なって存在する

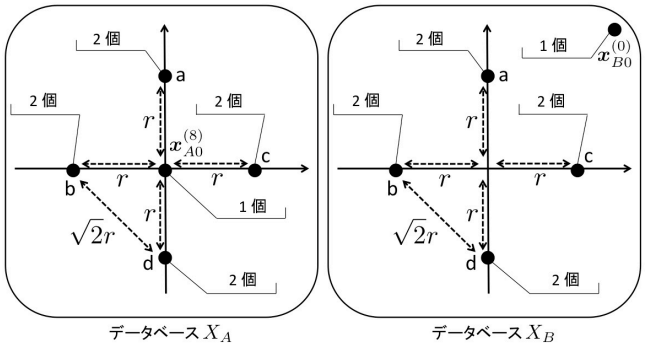

図 1　$GS_{q_{out}}$ の下限の例 (2 次元)

$2d(k-1)$ 個のベクトルは全て外れ値となり，これが外れ値個数の変化量となる．さらに，$\boldsymbol{x}_0$ 自身が正常値から外れ値 (あるいはその逆) に変化しうることを考慮し，1 を加えると，式 (11) となる．□

定理 3 を用いて，$GS_{q_{out}}$ の下限を求める例を例 1 に示す．

［例 1］　$d=2$，$k=3$ のとき，図 1 のような場合が考えられる．図 1 のデータベース X_A には●印で示したベクトルのみ存在するとする．X_A は，原点にベクトル $\boldsymbol{x}_{A0}^{(8)}$(括弧内は次数) が 1 個，原点から r 離れた点 a，b，c，d にベクトルが 2 個ずつ存在する．このとき，点 a，b，c，d から r より遠い点に $\boldsymbol{x}_{A0}$ が移動した X_A の近隣データベース X'_A を考えると，X_A では全てのベクトルの次数が k 以上だが，X'_A では全てのベクトルの次数が k 未満であるため，外れ値個数クエリの大域敏感度の下限は $2\times 2(3-1)+1=9$ となる．また，図 1 のデータベース X_B も●印で示したベクトルのみ存在するとする．X_B は，点 a，b，c，d から r より遠い点にベクトル $\boldsymbol{x}_{B0}^{(0)}$ が 1 個，点 a，b，c，d にベクトルが 2 個ずつ存在する．このとき，原点に $\boldsymbol{x}_{B0}$ が移動した X_B の近隣データベース X'_B を考えると，X_B では全てのベクトルの次数が k 未満だが，X'_B では全てのベクトルの次数が k 以上であるため，外れ値個数クエリの大域敏感度の下限は $2\times 2(3-1)+1=9$ となる．♢

外れ値個数クエリの大域敏感度の下限を示すことはできたが，$GS_{q_{out}}$ の下限を用いたとしても次数に対して線形的に増加する．大域敏感度はデータに依存せずに導くため，証明に用いたような病的な例も含むことから，必要以上に大きな敏感度になってしまう．そこで，敏感度として特に病的な例を考えることのない与えられたデータベースに依存させることによって導出できる平滑敏感度を用いることを考える．平滑敏感度を用いることは，病的な例を含めないで考えるわけではないが，一般的にデータベース内の外れ値はごくわずかであるため，平滑敏感度を用いることが適切であると考えた．

4.2　外れ値個数の局所敏感度

外れ値個数クエリの平滑敏感度を導く前に，定義 5 に従って外れ値個数クエリの局所敏感度 $LS_{q_{out}}$ を導く．局所敏感度や平滑敏感度は，与えられたデータベースに依存して敏感度が決まる．まず，$LS_{q_{out}}$ で用いる 2 つの外れ値候補数関数 OC と正常値候補数関数 IC を定義 8 に示す．

［定義 8］(外れ値候補数関数 OC と正常値候補数関数 IC)

次数が k である X_0 の元の集合を $V(k)\subseteqq X_0$ とする．また，データベース X_0 のドメインを $\mathcal{D}\subseteqq\mathbb{R}^d$ とする．点 $\boldsymbol{x}$ を中心とした半径 r の超球を $B(\boldsymbol{x},r)$ とする．$B(\boldsymbol{x},r)$ 内に含まれる次数 k のベクトル集合を $B(\boldsymbol{x},r)\cap V(k)$ とする．このとき，関数 OC と関数 IC を式 (12) と式 (13) とする．

$$OC(X_0,k,r)=\max_{\boldsymbol{x}_i\in X_0}|B(\boldsymbol{x}_i,r)\cap V(k)| \tag{12}$$

$$IC(X_0,k,r)=\max_{\boldsymbol{x}\in\mathcal{D}}|B(\boldsymbol{x},r)\cap V(k-1)| \tag{13}$$

$$=\max_{b(\boldsymbol{x},r)\in\mathcal{B}}|b(\boldsymbol{x},r)\cap V(k-1)| \tag{14}$$

ただし，$\mathcal{B}$ は $V(k)$ または $V(k-1)$ の部分集合を含む半径 r 以下の超球集合である．

関数 OC はデータベース中のベクトル $\boldsymbol{x}_i$ が移動すると外れ値になるベクトルの最大の個数を返し，式 (12) はデータベースのドメイン $\mathcal{D}$ 中の点 $\boldsymbol{x}$ へベクトルが移動すると正常値となる最大の個数を返す．このように関数 OC は離散集合 X_0 で定義される．一方，式 (13) は連続集合 $\mathcal{D}$ 上で定義される．よって，その評価は自明ではないため，式 (14) のようにして求める．超球集合 $\mathcal{B}$ は，$V(k)$ または $V(k-1)$ の部分集合を含む半径 r 以下の超球集合なので，式 (14) は式 (13) で数え上げるべきベクトルを数えることができる．本節では，式 (14) 内の超球集合 $\mathcal{B}$ を列挙するアルゴリズムを示す．このアルゴリズムによって出力された超球集合を式 (14) に用いることによって，式 (14) は算出可能になる．

次に，定義 5 に従って外れ値個数クエリの局所敏感度 $LS_{q_{out}}$ を定理 4 に示す．

［定理 4］(外れ値個数クエリの局所敏感度の上限)　定義 8 に示した関数 OC，関数 IC を用いると，外れ値個数クエリの局所敏感度 $LS_{q_{out}}$ の上限は式 (15) となる．

$$LS_{q_{out}}(X_0,k,r)\leqq\max\{OC(X_0,k,r),IC(X_0,k,r)\}+1 \tag{15}$$

証明:局所敏感度は，与えられたデータベース X_0 から X_0 に対する任意の近隣データベース X'_0 に変化する際の外れ値の個数の最大変化量である．この変化量は，X_0 内のベクトル $\boldsymbol{x}_0$ が $\boldsymbol{x}'_0$ に変化するときの “外れ値の増加量” と “正常値の増加量” の絶対値の差となる．よって，

$$\begin{aligned}&LS_{q_{out}}\\&=|OC(X_0,k,r)-IC(X_0,k,r)|+1\\&\leqq\max\{OC(X_0,k,r),IC(X_0,k,r)\}+1\end{aligned} \tag{16}$$

□

定理 4 を用いて，$LS_{q_{out}}$ の上限を求める例を例 2 に示す．

［例 2］　$d=2$，$k=1$ のとき，図 2 のような場合が考えられる．図 2 のデータベース X_A には●印で示したベクトルのみ存在するとする．さらに $\boldsymbol{x}_{A0}$ 以外のベクトルは全て互いに r より遠いとする．このとき，X_A において $\boldsymbol{x}_{A0}$ を他の全てのベクトルと r より遠い点に移動した X_A の近隣データベース X'_A を考えるこ

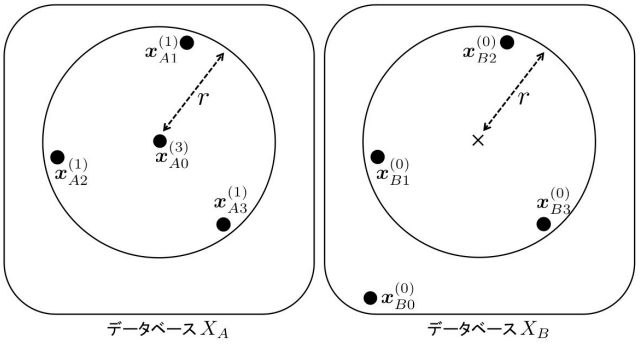

図 2　$LS_{q_{out}}$ の上限の例 (2 次元)

とは関数 OC の値を最大にする場合に相当する．よって，X_A では $\boldsymbol{x}_{A0}$ を中心とした半径 r の円内に含まれる次数が k であるベクトルが X'_A では全て次数が $k-1$ になるため，外れ値個数クエリの局所敏感度の上限は $OC(X_A,k,r)+1=3+1=4$ となる．また，データベース X_B にも●印で示したベクトルのみ存在するとする．さらに X_B も $\boldsymbol{x}_{B0}$ 以外のベクトルは全て互いに r より遠いとする．X_B において $\boldsymbol{x}_{B0}$ を他の全てのベクトルと遠くても r 離れた点に移動することは関数 IC の値を最大にする場合に相当する．よって，X_B では印を中心とした半径 r の円内に含まれる次数が $k-1$ であるベクトルが X'_B では全て次数が k になるため，外れ値個数クエリの局所敏感度の上限は $IC(X_B,k,r)+1=3+1=4$ となる．♢

式 (12) と式 (14) の計算量評価を行う．式 (12) に関しては，ベクトル $\boldsymbol{x}_i \in X_0$ を中心とした半径 r の超球内に存在する $V(k)$ に含まれるベクトルの個数を数えあげるので，計算量は $\mathcal{O}(dN^2)$ となる．式 (14) を解くためには $B(\boldsymbol{x},r)$ を列挙しなければならないく，式 (14) の計算量は $B(\boldsymbol{x},r)$ の列挙アルゴリズム (Algorithm1) から導く．Algorithm1 は再帰関数になっており，入力がベクトル集合，出力が超球のリストである．内部ではさらに，Smallest-Enclosing-Ball(seb) アルゴリズム [15] を使う．seb アルゴリズムの計算量は，$\mathcal{O}(d^2)$ である．Algorithm1 は，外れ値検出パラメータ r とベクトル集合 $V(k-1)$ を入力にとり，超球リストを出力する．$V(k-1)$ 内のベクトルをベクトル配列 P(インデックスは 1 から始まるものとする) に格納し，一つずつ増減させながら超球候補点集合 C を作成する．そして，C を seb アルゴリズムへ渡すことによって超球 b を生成し，その b が半径 r 以下の超球になっていれば超球リスト B に加える．よって，Algorithm1 の計算量は再帰分の $\mathcal{O}(|2^{V(k-1)}|)$ と Smallest-Enclosing-Ball アルゴリズムの $\mathcal{O}(d^2)$ から $\mathcal{O}(d^2|2^{V(k-1)}|)$ となる．このアルゴリズムは列挙に基づくため，指数計算量になる．しかし，一般的に外れ値の対象となる典型的なデータセットではベクトル集合 $V(k-1)$ のサイズは数個程度であることが想定され，現実的な時間で計算可能であることが期待できる．評価実験の際には，seb アルゴリズムの Java コード [16] を用いた．

4.3　外れ値個数の平滑敏感度

与えられたデータベース X_0 から t 個ベクトルを移動させたデータベース X_t への変換の際に外れ値の変化量に影響するベクトル数を数え上げるための関数 out，関数 in を式 (17)，式 (18) のように用意する．

Algorithm 1：$b(\boldsymbol{x},r)\in\mathcal{B}$ 列挙アルゴリズム $E(r,P,C,i)$

Input

外れ値検出パラメータ r, ベクトル配列 P,
超球候補点集合 C, インデックス i

Output

超球リスト B

```
E(r, V(k−1), ∅, 1)
subroutine:E
if i ≧ P.size then
    return ∅
else
    超球 b = seb(C)
    B1 ← E(r, P, C ∪ {P[i]}, i + 1)
    B2 ← E(r, P, B, C, i + 1)
    if B1.size > 0 then
        return B1 ∪ B2
    else if b.radius ≦ r then
        return B2 ∪ {b}
    else
        return B2
    end if
end if
end subroutine
```

$$out(X_0,\boldsymbol{x},k,r,t)=\sum_{i=0}^{t}|B(\boldsymbol{x},r)\cap V(k+i)| \tag{17}$$

$$in(X_0,\boldsymbol{x},k,r,t)=\sum_{i=1}^{t}|B(\boldsymbol{x},r)\cap V(k-i)| \tag{18}$$

関数 out と関数 in は，超球 $B(\boldsymbol{x},r)$ との共通部分に含まれる次数 $k\pm i$ のベクトル集合 $V(k\pm i)$ の個数を返す．以降では，式 (17) と式 (18) に示した外れ値候補数関数 OC と正常値候補数関数 IC を式 (20) に示す外れ値上限候補数関数 $\overline{OC}$，式 (22) に示す正常値上限候補数関数 $\overline{IC}$ のように拡張したものを用いる．式 (19)，式 (21) は $t=0$ のとき，式 (12)，式 (14) と同義になる．外れ値候補数関数と正常値候補数関数は，データベース X_0 とハミング距離が t となるデータベース X_t を考えるために関数 in と関数 out を用い，最後に $\max_{\boldsymbol{x}\in X_0}(out(\cdot)+in(\cdot))$(あるいは $\max_{\boldsymbol{x}\in\mathcal{D}}(out(\cdot)+in(\cdot))$) を考えることでデータベース X_t の局所敏感度を表現した．このように外れ値候補数や正常値候補数は式 (19) や式 (21) を用いて求めるべきだが，式 (19) や式 (21) は厳密な評価が困難なため，我々はその上限，外れ値上限候補数関数 $\overline{OC}$，正常値上限候補数関数 $\overline{IC}$ を式 (20)，式 (22) として用いる．

$$\max_{\boldsymbol{x}_i\in X_0}\left(out(X_0,\boldsymbol{x},k,r,t)+in(X_0,\boldsymbol{x},k,r,t)\right) \tag{19}$$

$$\leqq \max_{\boldsymbol{x}_i\in X_0}\{out(X_0,\boldsymbol{x},k,r,t)\}+\max_{\boldsymbol{x}\in\mathcal{D}}\{in(X_0,\boldsymbol{x},k,r,t)\} \tag{20}$$

$$\triangleq \overline{OC}(X_0,k,r,t)$$

$$\max_{\boldsymbol{x}\in\mathcal{D}}(out(X_0,\boldsymbol{x},k-1,r,t)+in(X_0,\boldsymbol{x},k-1,r,t)) \quad (21)$$

$$\leqq \max_{\boldsymbol{x}_i\in X_0}\{out(X_0,\boldsymbol{x},k-1,r,t)\}+\max_{\boldsymbol{x}\in\mathcal{D}}\{in(X_0,\boldsymbol{x},k-1,r,t)\} \quad (22)$$

$$\triangleq \overline{IC}(X_0,k,r,t)$$

この関数 $\overline{OC}$ と関数 $\overline{IC}$ を用い，移動させた $t+1$ 個のベクトルが全て正常値から外れ値 (あるいはその逆) に変化しうることを考慮すると式 (8) の $LS_q(X')$ は，式 (23) となる．

$$LSq_{out}(X_t) \quad (23)$$
$$\leqq \max\{\overline{OC}(X_0,k,r,t),\overline{IC}(X_0,k,r,t)\}+t+1$$

外れ値個数クエリの平滑敏感度として式 (23) を用いると，平滑敏感度の上限は式 (24) のようになる．

$$S^*_{q_{out},\beta}(X_0)=\max_{X_t\in\mathcal{D}^N}\{LS(X_t)\cdot e^{-\beta t}\} \quad (24)$$
$$\leqq \max_{\substack{X_t\in\mathcal{D}^N\\ :0\leqq t\leqq N}}\{(\max(\overline{OC}(X_0,k,r,t),\overline{IC}(X_0,k,r,t))+t+1)\cdot e^{-\beta t}\}$$

式 (19) から式 (20) にすることや式 (21) から式 (22) にすることは，有用性を犠牲にしているが，プライバシは保証したままである．しかし，我々の実験では有用性が過度に悪くならなかったことを式.

確認した (24) を用いて，$S^*_{q_{out},\beta}$ の下限を求める例を例 3 に示す.

［例 3］ $d=2$, $k=1$ のとき，図 3 のような場合が考えられる．図 3 のデータベース X_A には●印で示したベクトルのみ存在するとする．X_A と X_B の $\boldsymbol{x}_{A0}$ と $\boldsymbol{x}_{B0}$ 以外のベクトルは全て互いに r より遠いとする. このとき，X_A において $\boldsymbol{x}_{A4}$ を他の全てのベクトルと r より遠い点に移動することは X_A のハミング距離 $t=1$ のデータベース X_{A1} を考えることに相当し，さらに X_{A1} の $\boldsymbol{x}_{A0}$ を他の全てのベクトルと r より遠い点に移動した X_{A1} の近隣データベース X'_{A1} を考えることは関数 $\overline{OC}$ の値を最大にする場合に相当する．よって，X_A では $\boldsymbol{x}_{A0}$ を中心とした半径 r の円内に含まれる次数が k と $k+1$ であるベクトルが X'_A では全て次数が $k-1$ になるため，$t=1$ のときの外れ値個数クエリの局所敏感度の上限は $\overline{OC}(X_A,k,r,1)+1=2+1=3$ となる．また，図 3 のデータベース X_B にも●印で示したベクトルのみ存在するとする．X_B において $\boldsymbol{x}_{B4}$ を▲印の点に移動させることは X_A のハミング距離 $t=1$ のデータベース X_{B1} を考えることに相当し，さらに X_{B1} の $\boldsymbol{x}_{B0}$ を × 印に移動した X_{B1} の近隣データベース X'_{B1} を考えることは関数 $\overline{IC}$ の値を最大にする場合に相当する．よって，X_B では $\boldsymbol{x}_{B0}$ を中心とした半径 r の円内に含まれる次数が $k-1$ と $k-2$ であるベクトルが X'_{B1} では全て次数が k になるため，$t=1$ のときの外れ値個数クエリの局所敏感度の上限は $\overline{IC}(X_B,k,r,1)+1=2+1=3$ となる．♢

式 (20) と式 (22) の計算量評価は，式 (12) と式 (14) と同様の考え方で，それぞれ $\mathcal{O}(dN^2)$，$\mathcal{O}(d^2|2^{\cup_{i=-t}^{t}V(k-1+i)}|)$ となる．しかし，関数 $\overline{IC}$ の評価は t に対して指数的に増加するこ

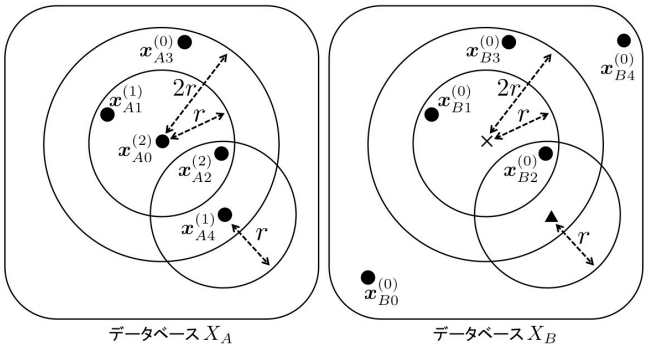

図 3 $S^1_{q_{out},\beta}(X)$ の上限の例 (2 次元)

Algorithm 2：$S^*_{q_{out},\beta}(X)$ を効率よく求める方法

Input

データベース X, 外れ値検出パラメータ k,r

Output

クエリ q_{out} の平滑敏感度 $S^*_{q_{out},\beta}(X_0)$

Initialization

リスト $S_{\text{past}}=\emptyset$

1: **for** $t=0$ to N **do**
2: S^t_{UB} を求める
3: **if** $S_{\text{past}}\neq\emptyset$ and $\max_{0\leqq i<t}S_{\text{past}}(i)>S^t_{\text{UB}}(X_0)$ **then**
4: **return** $\max_{0\leqq i<t}S_{\text{past}}(i)$
5: **end if**
6: $S_{\text{past}}(t)=S^t_{q_{out},\beta}(X_0)$
7: **end for**
8: **return** $\max_{0\leqq i\leqq N}S_{\text{past}}(i)$

とから t の値が大きいときは工夫する必要がある．そこで，平滑敏感度を求める際の効率化を行うことでこの問題を解消する．平滑敏感度の上限は指数的に単調減少するため，平滑敏感度は小さい t で与えられることが期待できる．さらに，t の値が小さければ列挙関数 E の再帰回数や $|\cup_{i=0}^{t}V(k+i)|$ が小さいことから現実的な時間で計算を行うことが可能である．そこで，t が小さい値から順に平滑敏感度を計算する．ステップ t での平滑敏感度を式 (25) に示す.

$$S^t_{q_{out},\beta}(X_0) \quad (25)$$
$$=(\max(\overline{OC}(X_0,k,r,t),\overline{IC}(X_0,k,r,t))+t+1)\cdot e^{-\beta t}$$

データベース X_0 とハミング距離 t であるデータベース X_t の近隣データベース $\forall X'_t$ に対して $LS_{q_{out}}(X'_t)\leqq N$ なので，$S^t_{q_{out},\beta}(X_0)\leqq e^{-t\beta}N\triangleq S^t_{\text{UB}}(X_0)$ である．このステップ t での上界 $S^t_{\text{UB}}(X_0)$ を用い，計算時間量が大きい関数 in をなるべく計算せずに $S^*_{q_{out},\beta}(X_0)$ を求める方法を提案する．この方法を Algorithm2 に示す．指数的に単調減少する S^t_{UB} は $S^t_{q_{out},\beta}(X_0)$ の上界となるので，$i<t$ である $S^i_{q_{out},\beta}(X_0)$ よりも $S^t_{\text{UB}}(X_0)$ が小さければ，$S^i_{q_{out},\beta}(X_0)(i<t)$ の最大値が $S^*_{q_{out},\beta}(X_0)$ となる．

5. 実　　験

本章では，(ϵ,δ)-差分プライバシのノイズ規模を決定するク

表 1　評価対応表

	事実が外れ値	事実が正常値
予測が外れ値	TP	FP
予測が正常値	FN	TN

エリ q_{out} に対する大域敏感度と平滑敏感度が与えるクエリ結果に加えるノイズを比較をする．5.1 節では実験設定を述べ，5.2 節ではその実験設定下での結果を示し，その結果に対して考察する．

5.1　実験設定

本節では，実験概要，データセット，各種パラメータの設定について述べる．実験概要は，外れ値個数クエリの大域敏感度の下限を用いる際と平滑敏感度の上限を用いる際のクエリ結果に加えるノイズの大きさを比較することである．データセットは，UCI Machine Learning Repository から 2 つのデータセットを選んだ．一つ目は，ionosphere データセットである．これは，電離層中から跳ね返りがあるかどうかのレーダデータベースであり，次元数 34，サンプル数 351 である．ラベルは “good”(225 個) と “bad”(126 個) の二種類である．“bad” を外れ値とし，“good” のベクトル全てと “bad” の 10 ベクトルを合わせて外れ値検出用データセットとした．二つ目は，wdbc データセットである．これは，乳がんのデータベースであり，次元数 $d = 32$，サンプル数 $N = 569$ である．ラベルは，“無害”(357 個) と “有害”(212 個) の二種類である．“有害” を外れ値とし，“無害” のベクトル全てと “有害” の 10 ベクトルを合わせて外れ値検出用データセットとした．両データセット共に各属性が平均 0，分散 1 になるようにスケーリングした．このような外れ値検出用データセットの構築の仕方は，文献 [17,18] と同様にした．外れ値検出パラメータ k は，文献 [19] と同様に $k = 5$ とした．パラメータ r は，式 (26) に示す正確度 A が最も高くなるときの r を用いる．式 (26) 内の TP, FP, FN, TN は表 1 の各セルに対応する．

$$A = \frac{TP + TN}{TP + FP + FN + TN} \tag{26}$$

5.2　結果と考察

ionosphere データセットでは正確度 A が最良値を得た $r = 0.3$ を，wdbc データセットで正確度 A が最良値を得た $r = 1.3$ を用いた．$k = 5$，$\epsilon = 0.5$ における外れ値個数クエリの平滑敏感度の上限を計算し，ノイズを生成した結果を図 5.2，表 5.2 に示す．図 5.2，表 5.2 内の横軸はプライバシパラメータ ϵ であり，縦軸である出力結果は真のクエリ結果 (今回はどちらのデータセットにおいても真の外れ値は 10 個) に外れ値個数クエリの大域敏感度の下限を用いたときのノイズの標準偏差 σ_{GS} または平滑敏感度の上限を用いたときのノイズの標準偏差 σ_S を加えたものを表示している． 図 5.2，図 5.2 より，プライバシパラメータ $\epsilon = 0.2, 0.3$ と非常に厳しいプライバシ保証環境においても外れ値個数クエリの平滑敏感度の上限を用いた際のノイズの標準偏差 σ_S とクエリ結果 $q_{out}(\cdot)$ の比 $\sigma_S / q_{out}(\cdot)$ が小さいことが分かる．このことから，平滑敏感度の上限を用いることはプライバシを保証しつつ得られる結果の有用性も高いと言える．

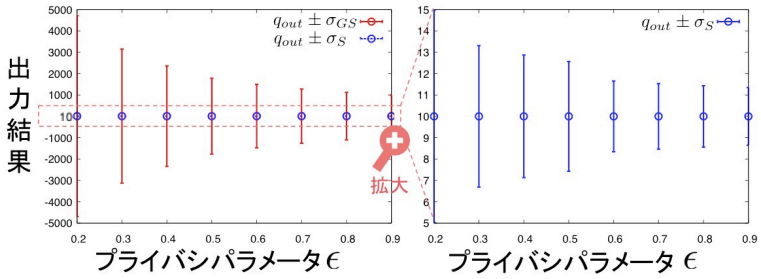

図 4　ionosphere データセットの実験結果 (クエリ結果 ± ノイズの標準偏差)

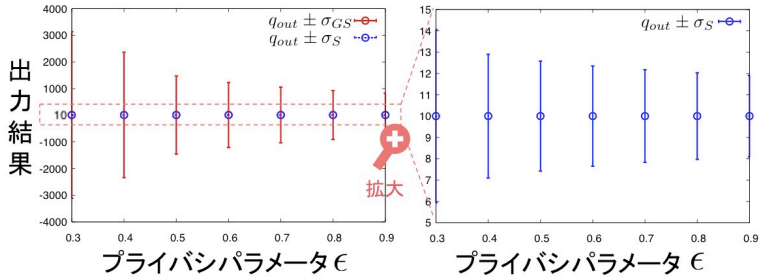

図 5　wdbc データセットの実験結果 (クエリ結果 ± ノイズの標準偏差)

6.　おわりに

本稿では，外れ値の個数検出に対する差分プライバシの適用を試みた．まず，大域敏感度を用いる場合と平滑敏感度を用いる場合を検討し，理論的成果として外れ値個数クエリの大域敏感度の下限，局所的敏感度の上限，平滑敏感度の上限を示した．これによって，大域敏感度を用いると下限でさえ，ノイズ規模が大きくなることが分かった．さらに，平滑敏感度を用いると差分プライバシを保証しつつ，クエリ結果に加えるノイズが抑制できることも分かった．次に，本稿では平滑敏感度の上限を求める際の計算量評価を行った．平滑敏感度の計算量は指数計算量であるが，平滑敏感度の上限を使用することや提案した効率化アルゴリズムによって現実的な時間計算時間で動作した．最後に，実験を通してプライバシを厳しく保証する状況において外れ値個数クエリの敏感度として大域敏感度の下限を用いるときよりも平滑敏感度の上限を用いるときのほうが出力結果が有用であることを示した．以上のことから，これまで対応することができなかった定常データの多次元ベクトルに対する外れ値の個数検出時のプライバシ保護の実現をした．

今後の課題は二つある．一つ目は，本稿で示した平滑敏感度の上限を用いた際のノイズをより抑制するためにより厳密な外れ値個数クエリの平滑敏感度解析評価を行ないたい．しかし，厳密な評価を行うほど計算量が爆発的に増加してしまうので，この問題を解決したい．二つ目は，データベースのドメイン内のうち，領域または次元に着目した外れ値の存在に関する解析に対する差分プライバシの保証を検討することである．これによって，対象とするデータベースの特異さを把握する．

謝　　辞

本研究は，JST CREST 「ビッグデータ統合利用のための次世代基盤技術の創出・体系化」領域および科学研究費 12913388 の助成を受けました．

文　　献

[1] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[2] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing*, STOC '07, pages 75–84, New York, NY, USA, 2007. ACM.

[3] Latanya Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, October 2002.

[4] Frank E. Grubbs. Sample criteria for testing outlying observations. *Ann. Math. Statist.*, 21(1):27–58, 03 1950.

[5] Bernhard Schlkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. New support vector algorithms, 2000.

[6] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data*, 6(1):3:1–3:39, mar 2012.

[7] Nabil M. Hewahi and Motaz K. Saad. Class outliers mining: Distance-based approach. *International Journal of Computer, Information, Systems and Control Engineering*, 1(9):2752 – 2765, 2007.

[8] Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24rd International Conference on Very Large Data Bases*, VLDB '98, pages 392–403, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[9] Edwin M. Knorr, Raymond T. Ng, and Vladimir Tucakov. Distance-based outliers: Algorithms and applications. *The VLDB Journal*, 8(3-4):237–253, feb 2000.

[10] Cynthia Dwork and Adam Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):2, 2010.

[11] Liyue Fan and Li Xiong. Differentially private anomaly detection with a case study on epidemic outbreak detection. In *ICDM Workshops*, pages 833–840, 2013.

[12] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay, editor, *EUROCRYPT*, volume 4004 of *Lecture Notes in Computer Science*, pages 486–503. Springer, 2006.

[13] J. H. Conway, N. J. A. Sloane, and E. Bannai. *Sphere-packings, Lattices, and Groups*. Springer-Verlag New York, Inc., New York, NY, USA, 1987.

[14] Peter Brass, William Moser, and Janos Pach. *Research Problems in Discrete Geometry*. Springer, 2005.

[15] Kaspar Fischer, Bernd Grtner, and Martin Kutz. Fast smallest-enclosing-ball computation in high dimensions. In *Proc. 11th European Symposium on Algorithms (ESA*, pages 630–641. SpringerVerlag, 2003.

[16] Kutz, Fischer Martin, Grtner Kaspar, and Bernd. A java library to compute the miniball of a point set.

[17] Ke Zhang, Marcus Hutter, and Huidong Jin. A new local distance-based outlier detection approach for scattered real-world data. *CoRR*, abs/0903.3257, 2009.

[18] Ninh Pham and Rasmus Pagh. 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*, KDD '12, pages 877–885, New York, NY, USA, 2012. ACM.

[19] M. Sugiyama and K. M. Borgwardt. Rapid distance-based outlier detection via sampling. In *Advances in Neural Information Processing Systems 26*, pages 467–475, Lake Tahoe, Nevada, USA, December 2013.