不確実データ集合に対する距離に基づく外れ値検出

郭　楽† 北川　博之†,‡ 天笠　俊之†,‡
†筑波大学大学院システム情報工学研究科　〒305-8573 茨城県つくば市天王台 1-1-1
‡筑波大学大学院計算科学研究センター　〒305-8573 茨城県つくば市天王台 1-1-1
E-mail: †kakuraku1986@kde.cs.tsukuba.ac.jp, ‡{kitagawa,amagasa}@cs.tsukuba.ac.jp

あらまし　外れ値検出は，データ集合中で他の大多数のデータから大きく離れた値を抽出する技術で，クレジットカードの不正利用やセンサの異常値の検出など，様々な分野で応用されている．一方で，実世界において，センサの測定の誤差により，測ったデータが曖昧性を含む「不確実データ」になる状況がしばしば生じる．その「不確実データ」集合に対して外れ値検出を行う必要がある．そこで本研究では不確実データ集合に対する，距離に基づく手法を拡張した手法を提案し，実験により有効性を評価する．

キーワード　外れ値検出，不確実データ，データマイニング，DB 外れ値

Distance-Based Outlier Detection over Uncertain Data

Le　GUO†　Hiroyuki KITAGAWA†,‡ and Toshiyuki AMAGASA†,‡
†Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennoudai, Tsukuba-shi, 305-8573, Japan
‡Center for Computational Sciences, University of Tsukuba
1-1-1 Tennoudai, Tsukuba-shi, 305-8573, Japan
E-mail: †kakuraku1986@kde.cs.tsukuba.ac.jp, ‡{kitagawa,amagasa}@cs.tsukuba.ac.jp

Abstract　Outlier detection is an important data mining technique, and it discovers outliers which have features that differ profoundly from other objects or values. On the other hand, the new ways of collecting data such as the development of sensor devices have resulted in enormous amounts of uncertain data. The uncertainty added to the data points may make true outliers masked. This paper introduces a distance-based approach to outlier detection for uncertain datasets. We present experimental results suggesting the effectiveness of the method.

Key words　outlier detection, uncertainty data, data mining, DB-Outlier

1. はじめに

近年，巨大なデータから知識や有用な情報を発見するデータマイニングの技術が注目されている．中でも，オブジェクト集合中において，他のオブジェクトから大きく異なる特徴や値をもつオブジェクトを検出する「外れ値検出」の技術は，クレジットカードの不正利用の検出，ネットワークの不正行為の検出，医療や保険業界における不正請求検出など様々な分野での応用が期待されている．これまでにも、統計に基づく外れ値検出[1],[2],距離に基づく外れ値検出[5]、クラスタリングによる外れ値の検出[6],[7]等、様々な手法が確定データ集合に対する外れ値検出手法として、提案されている。

一方で，近年、センシングデバイスの普及や GPS 応用の進展により、曖昧性を含む「不確実なデータ」に対する処理の要求が高まっている．実世界において不確実データを扱う状況としておおまかに次の二つが考えられる．一つはそもそも正確な値を測ることが不可能な場合である．例えば，モバイルデータ

ベースでは，移動位置を知るために GPS が一般に利用されるが，必ずしもすべての状況において GPS が正しい現在位置を測定できるとは限らない．そのような状況で，GPS で測定された情報は曖昧なものとなる．二つめは正確なデータを公開したくない場合である．例えば，インターネットで自分のプライバシーを保護するために，一部の情報を隠すなど，意図的に不確実性を含ませた上で，情報提供を行うことがある．

不確実データが大量に存在しているため，このようなデータ集合に対する処理の要求が高まっている．そこで本研究は不確実性を有するデータ集合に対して外れ値検出を行う手法を提案する．具体的には基本的な外れ値検出手法である，距離に基づく外れ値検出（Distance-based outlier，DB 外れ値）[3]～[5]を基に不確実性を扱うための手法を提案する．

本稿の構成は次の通りである。2.で関連研究について述べる。3.で準備として DB 外れ値の定義と DB 外れ値検出アルゴリズムについて述べる。4．で提案手法を示し、5．で提案手法の有用性を示した実験とその結果を示す。最後に 6．で本稿をまとめる。

2.　関連研究

本節では既存の外れ値検出の研究について述べる。

確定データ集合に対する外れ値検出手法として様々な手法がこれまでに提案されている．代表的な手法としては，本研究で用いる DB 外れ値検出手法の他[6]に，統計に基づく手法[1]，クラスタリングによる手法[2]，密度を基にした手法[3]等がある．本研究でベースとする DB 外れ値検出は，他の手法に比べて，より単純で一般性のある手法である．

不確実データに対する外れ値検出手法としては，次のような手法が提案されている．Aggarwal ら[4]は，各オブジェクト値の出現確率が確率密度関数（PDF）で与えられる場合の密度に基づく外れ値検出手法を提案した．Wang らは，各オブジェクトの存在が確率的に決まる場合の距離に基づく外れ値検出手法を提案している[5].

本研究ではオブジェクトの各属性値の取り得る値の不確実性を考慮した場合を対象に，最も基本的な外れ値検出手法である距離に基づく外れ値検出手法の拡張した手法を提案する．

3.　距離に基づく外れ値検出手法

本節では、本研究で用いる外れ値の定義と、既存の確定データに対する DB 外れ値検出アルゴリズムを説明する。

3.1.　DB 外れ値

本研究では、Edwin M. Knorr らにより提案された DB 外れ値を外れ値の定義として用いる[6].

［定　義　1］

N 個の k 次元オブジェクト $O_1 \ldots O_N$ からなるオブジェクト集合 S において，オブジェクト O_i が DB(p, D)外れ値であるとは，O_i からの距離が D より大きい範囲に，S 中の $\lceil pN \rceil$ 個以上のオブジェクトが存在するということである．

以下では，オブジェクト O_i から距離が D 以下の範囲を O_i の D 近傍と呼ぶ．$M = N(1-p)$ を用いると，O_i が DB(p, D)外れ値であるとは，オブジェクト O_i の D 近傍内のオブジェクト数（O_i を含む）が M 個以下ということと等価である．

k = 2 の場合の DB 外れ値の例を図 1 に示す．各点はオブジェクトを表す．図 1 の左上の円，右下の円はそれぞれオブジェクト O_1，O_2 の D 近傍を表す．p = 0.9，N = 30 とする．従ってもしあるオブジェクトの D 近傍内のオブジェクト数が $M = 3$（$M = 30 * (1 - 0.9)$）以下であれば，そのオブジェクトは DB 外れ値である．図 1 では，O_1 の D 近傍内のオブジェクト数は 3（ ≤M ）であるので，O_1 は DB 外れ値である．一方，O_2 の D 近傍内のオブジェクト数は 9（ ＞M ）であるので O_2 は DB 外れ値ではない．

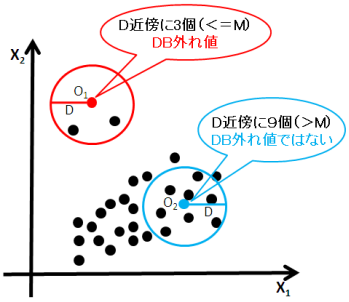

図 1　DB 外れ値の例：N=30, p=0.9

3.2.　セルに基づく外れ値検出アルゴリズム

N 個の k 次元オブジェクト $O_1 \ldots O_N$ からなるオブジェクト集合 S 中の DB 外れ値を検出する最も単純な方法は，各オブジェクト $O_i (1 <= i <= N)$ に対して O_i と $O_p (1 <= p <= N, i \neq p)$ の距離 $d(O_i, O_p)$ を順次計算し，O_i の D 近傍のオブジェクト数を数えて判断する方法である．しかし，このような単純な方法で外れ値検出を行うと，$O(N^2)$ 回の距離計算が必要となり，全体の計算量は膨大になる．

距離計算の回数を減らし，全体の計算量を削減するために，セルを用いたアルゴリズム(Cell-Based Algorithm) が Knorr らにより提案された[5]．セルを用いたアルゴリズムでは，距離計算の回数を減らすために，オブジェクトが存在する空間をセルで分割して計算を行う．

3.2.1 セル構造

N 個の k 次元オブジェクト $O_1 \ldots O_N$ は, $X_1,\ldots,X_k$ 軸からなる k 次元空間中の点として表現できる. この k 次元空間中を一辺の長さが $l = \frac{D}{2\sqrt{k}}$ (対角線の長さが $\frac{D}{2}$) のセルに分割する. X_1 軸について x_1 番目, ..., X_k 軸について x_k 番目のセルを $C_{x_1,\cdots,x_k}$ と表す. なお以降は, 特にセルの各軸に対する添字を明記する必要がない場合は, セルを単に C_x と略記する. なお, 隣接するセルの境界線にあるオブジェクトは, $x_1, \ldots, x_k$ の値の小さい方のセルに含まれると見なす.

この時, 次の性質 1 が成り立つ.

[性　質　1]

同一セル中のすべての 2 オブジェクト間の距離は $\frac{D}{2}$ 未満である.

セル C_x 内に存在する 2 つオブジェクト O_i, O_p 間の距離が最も長くなるのは, O_i, O_p 間の距離 $d(O_i, O_p)$ がセルの対角線の長さ $\frac{D}{2}$ と等しくなる場合である. 従って性質 1 が成り立つ.

次に, $C_{x_1 \ldots x_k}$ の L_1 近傍範囲を定める.

[定　義　2]

$C_{x_1 \ldots x_k}$ の L_1 近傍 $L_1(C_{x_1 \ldots x_k})$ を, 以下のように定義する.

$$L_1\left(C_{x_1,\cdots,x_k}\right) = \left\{ C_{u_1,\ldots u_k} \mid |u_i - x_i| \le 1 (1 \le i \le k), C_{u_1,\ldots u_k} \ne C_{x_1,\ldots x_k} \right\}$$

定義 2 から, 次の性質 2 が成り立つ.

[性　質　2]

$C_{u_1 \ldots u_k} \in L_1(C_{x_1 \ldots x_k})$, $O_i \in C_{x_1 \ldots x_k}$, $O_p \in C_{u_1 \ldots u_k}$ の時, $d(O_i, O_p) < D$.

セル C_x に含まれるオブジェクト O_i、C_x の L_1 近傍のセル C_u に含まれるオブジェクト O_p 間の距離が最大になるのは、O_i, O_p 間の距離 $d(O_i, O_p)$ がセルの対角線長の 2 倍と等しくなる場合である。従って性質 2 は成り立つ.

[定　義　3]

$C_{x_1 \ldots x_k}$ の L_2 近傍 $L_2(C_{x_1 \ldots x_k})$ を, 以下のように定義する.

$$L_2\left(C_{x_1,\cdots,x_k}\right) = \left\{ C_{u_1,\ldots,u_k} \mid |u_i - x_i| < \left\lceil 2\sqrt{k} \right\rceil (1 \le i \le k), \; C_{u_1,\cdots u_k} \notin L_1\left(C_{x_1,\cdots,x_k}\right), C_{u_1,\cdots u_k} \ne C_{x_1,\cdots,x_k} \right\}$$

定義 3 から, 次の性質 3 が成り立つ.

[性　質　3]

$C_{u_1 \ldots u_k}$ は $C_{x_1 \ldots x_k}$ の L_1、L_2 近傍以外のセルとする. このとき $C_{u_1 \ldots u_k} \ne C_{x_1 \ldots x_k}$ であり, それぞれのセルに含まれるオブジェクトを $O_p \in C_{u_1 \ldots u_k}$, $O_i \in C_{x_1 \ldots x_k}$ の時, $d(O_i, O_p) > D$.

C_x 内オブジェクト O_i と C_x, C_x の L_1 近傍のセル, L_2 近傍のセルに含まれないオブジェクト O_p の距離が最小になるのは, 定義 3 から必ず $d(O_i, O_p) > \left| 2\sqrt{k} \right| l$ が成り立つ. 従って $\left\lceil 2\sqrt{k} \right\rceil l \ge 2\sqrt{k} l = 2\sqrt{k} \frac{D}{2\sqrt{k}} = D$ であることにより, 性質 3 は成り立つ.

$C_{x_1 \ldots x_k}$ 内のオブジェクト数を n_0, L_1 ($C_{x_1 \ldots x_k}$) 内のオブジェクト数を n_1, L_2 ($C_{x_1 \ldots x_k}$) 内のオブジェクト数を n_2 とする. 性質1～3 から次の性質4 が成り立つ.

[性　質　4]

(1) $n_0 > M$ ならば, $C_{x_1 \ldots x_k}$ 内の全オブジェクトはDB外れ値ではない.

(2) $n_0 + n_1 > M$ ならば, $C_{x_1 \ldots x_k}$ 内の全オブジェクトはDB外れ値ではない.

(3) $n_0 + n_1 + n_2 \le M$ ならば, $C_{x_1 \ldots x_k}$ 内の全オブジェクトはDB外れ値である.

ここで, (1) が成り立つ場合はred, (1) が成り立たず(2) が成り立つ場合は pink, (1), (2)が成り立たず (3) が成り立つ場合は yellow とセル $C_{x_1 \ldots x_k}$ を色付けする. 各セルに対してこのような方法で色付けする処理を行った後に, 未だ色付けされていないオブジェクトが存在するセルがある場合, そのセルを white に色付けする.

3.2.2 アルゴリズム

アルゴリズム 1 はセルを用いたアルゴリズムである. セル C_x の n_0, n_1, n_2 の値をそれぞれ $n_0(x)$, $n_1(x)$, $n_2(x)$ と表す. 1-2 行目で各セル C_x の $n_0(x)$の値を初期化する. 3-4 行目でオブジェクト O_i が含まれるセルを特定する. 5-7 行目では $n_0 > M$ であるセルを red とする(性質 4(1)). 8-11 行目では L_1 近傍に red セルが存在するセルを pink とする. これは, 任意のセル C_x の L_1 近傍に red セルが存在する場合, C_x は必ず性質 4(2) を満たすからである. 14-15 行目で red もしくは pink と判定されなかったセルで,かつ, 性質 4(2)を満たすセルを pink とする. 18-19 行目で性質 4(3)を満たすセルを yellow とし, yellow とされたセル内の全オブジェクトを外れ値とする (性質 4(3)). 最後に 20-27 行目で, 19 行目までの処理で red, pink, もしくは yellow と判定されなかったセル C_w を white とし, white と判定されたセルに対して 22-27 行目にてセルに含まれる各オブジェクト O_i に DB 外れ値判定を行う.

アルゴリズム 1 : セルを用いたアルゴリズム

1. **for each** セル C_x do
2. 　$n_0(x) \leftarrow 0$
3. **for each** オブジェクト O_i **do**
4. 　O_i を含むセル C_x を特定し, $n_0(x)$ を 1 増やす.

5. **for each** セルC_x **do**
6. **if** $n_0(x) > M$ **then**
7. C_x をred とする．
8. **for each red** セルC_r **do**
9. **for each** C_r のL_1 近傍のセルC_u
10. **if** C_uがred でない場合**then**
11. C_u をpink とする．
12. **for each** セル内にオブジェクトが存在し，色がついていないセルC_w **do**
13. $n_1(w) \leftarrow \sum_{C_i \in L_1(C_w)} n_O(i)$
14. **if** $n_0(w) + n_1(w) > M$ **then**
15. C_w をpink にする．
16. **else**
17. $n_2(w) \leftarrow \sum_{C_i \in L_2(C_w)} n_O(i)$
18. **if** $n_0(w) + n_1(w) + n_2(w) \leq M$ **then**
19. C_wをyellowとし,C_w内の全オブジェクトを外れ値とする．
20. **else**
21. C_w をwhite とする．
22. **for each** オブジェクトO_i $\in C_w$ **do**
23. Count $\leftarrow n_0(w) + n_1(w)$
24. **for each** オブジェクトO_p $\in$ C_u,C_u $\in L_2(C_w)$ **do**
25. **if** dist$(O_i,O_p) \leq D$ **then**
26. Count を1増やし，Count > MになったらO_iは外れ値ではないと判断し，次のオブジェクトの処理(22 行目）を開始する．
27. O_iを外れ値とする．

4. 提案手法

本節では，我々が提案する不確実データ集合に対する外れ値検出の手法を示す．まず，提案手法で使われる定義について述べ，次にアルゴリズムについて説明する．

4.1 不確実領域

本研究で扱うオブジェクトは不確実性を含むため，測定された位置に実際のオブジェクトが存在するとは限らない．ここではオブジェクトが存在しうる範囲を不確実領域と定義する．

[定　義　4]

オブジェクトが存在する可能性のある領域を不確実領域と呼ぶ．逆に，オブジェクトはこの領域外に存在する確率は0 である．

本研究では，不確実領域を測定されたオブジェクトが中心点としてからの距離がεの範囲とする．さらに，本研究においては，簡単化のため，オブジェクトはその不確実領域内に最も基本的な一様に分布するものとする．2 次元の時，不確実領域は半径εの円形になる．

2 次元の不確実データの例を図 2 に示す．点 O にあるオブジェクトに対して，半径εの円内の領域が O の不確実領域である．他のオブジェクトも同様に円形の不確実領域を持つ．オブジェクト O の不確実領域の中心点を，オブジェクト O の観測値と以下では呼ぶ．

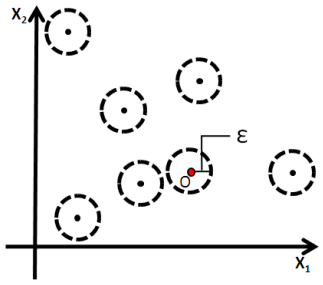

図 2　2 次元の不確実データ集合の例

4.2 PDB (Probabilistic Distance-Based) 外れ値

不確実領域を持つオブジェクト集合に対して，外れ値を以下のように定義する．

[定　義　5]

オブジェクト A がオブジェクト O のD近傍内に存在する確率を Pr(D: O, A)で表す．このとき，

$$\sum_A \Pr(D:O,A) \leq N(1-p)$$

を満たすときオブジェクト O が PDB 外れ値である．

ここで,$\sum_A \Pr(D:O,A)$オブジェクト O の D 近傍内に存在するオブジェクト数の期待値を計算している．

PDB 外れ値の例を図 3 に示す．それぞれの小さい円は各オブジェクトの半径εの不確実領域を表す．オブジェクトの観測値が中心にあるが，その値はεの範囲の不確実性を有する．オブジェクトO_1とO_2の真の値はそれぞれ O_1'と O_2'にあるとする．大きい円は中心点O_1'とO_2'のD近傍を表す．ただし，ε=0.3, D=2.0，p=0.714, N=7 とする．もし D 近傍内のオブジェクト数の期待値が M = 2（M = 7×(1 - 0.714)）以下であれば，オブジェクトは PDB 外れ値である．いま，左上のオブジェクトO_1の D 近傍のオブジェクト数の期待値は 1.075（ < M ）であるので，O_1は PDB 外れ値である．一方，下のオブジェクトO_2の D 近傍のオブジェクト数の期待値は 5.535（ >= M ）であるので PDB 外れ値ではない．

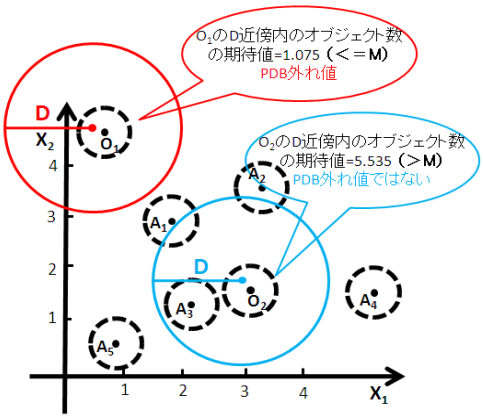

図 3　2 次元 PDB 外れ値の例

4.3　セルベースの PDB 外れ値検出手法

不確実データに対する外れ値検出では，期待値を導出するため確率計算が必要であり，単なる距離計算であった通常の外れ値検出処理よりも高コストである．このため各オブジェクトに対してナイーブな処理を行うのは極めて効率が悪い．ここでは Cell-based アルゴリズムを不確実データに拡張し，検出処理の効率化を図る．

4.3.1　　セル構造

各次元を $X_1,\ldots,X_k$ 軸とする k 次元空間中に存在する半径 ε の不確実領域を持つオブジェクトを考える．このとき k 次元空間を一辺の長さが $l=\frac{D-2\varepsilon}{2\sqrt{k}}$ のセルに分割する．各セルを c 各次元 X_i の原点からの位置 x_i を用いて $C_{x_1,\cdots x_k}$ と表す．このとき各セルは以下の性質を満たす．

[性　質　5]

観測値を同じセル中に持つ任意の 2 オブジェクト間の距離は D 未満である．

2 次元の例を図 4 に示す．中央のセルに着目すると，観測値を同一セルに持つ時オブジェクト間の距離が最大になるのは対角の位置にある時であり，その時の観測値同士のセルの対角線の長さ $\frac{D-2\varepsilon}{2}$ となる．従って 2 オブジェクト間の距離は D 未満となるため性質 1 は明らかである．

次に $C_{x_1,\cdots x_k}$ の L_1，L_2 近傍について説明する．

[定　義　6]

L_1 近傍は 3.2.1 節における定義と同じである．

$$L_1\left(C_{x_1,\cdots x_k}\right)=\left\{C_{u_1,\cdots u_k}\,|\,|u_i-x_i|\le 1(1\le i\le k),C_{u_1,\cdots u_k}\ne C_{x_1,\cdots x_k}\right\}$$

[性　質　6]

$C_{u_1\ldots u_k}$ を $C_{x_1\ldots x_k}$ L_1 近傍のセルとし，オブジェクト $Q\in C_{u_1\ldots u_k}$，$O\in C_{x_1\ldots x_k}$ を考える．このとき任意の OQ 間の距離は D 未満である．

2 次元の例を図 5 に示す．中央のセルについて 2 つのオブジェクト間の観測値が L_1 近傍内で最大になるのはオブジェクトの観測値が O と Q に場合である．これはセルの対角線の 2 倍と等しい．この時，O と Q の真値の間の最大距離は D 未満である．

[定　義　7]

L_2 近傍を定義する式は以下のように変わる．

$$L_2\left(C_{x_1,\cdots x_k}\right)=\left\{C_{u_1,\ldots,u_k}\,\middle|\,|u_i-x_i|<\left\lceil\frac{(D+2\varepsilon)\times 2\sqrt{k}}{D-2\varepsilon}\right\rceil(1\le i\le k),\ C_{u_1,\cdots u_k}\notin L_1\left(C_{x_1,\cdots x_k}\right),C_{u_1,\cdots u_k}\ne C_{x_1,\cdots x_k}\right\}$$

[性　質　7]

$C_{u_1\ldots u_k}$ は $C_{x_1\ldots x_k}$ の L_1，L_2 近傍以外のセルとする．このとき $C_{u_1\ldots u_k}\ne C_{x_1\ldots x_k}$ であり，それぞれのセルに含まれるオブジェクトを $R\in C_{u_1\ldots u_k}$，$O\in C_{x_1\ldots x_k}$ を考える．このとき任意の O R 間の距離は D より大きい．

2 次元の例を図 6 に示す．O と R の観測値間の距離が最小になるのが図の状態であり，L_1, L_2 層の厚さを n とすると，

$$n\ge\frac{(D+2\varepsilon)*2\sqrt{k}}{D-2\varepsilon}\,l=\frac{(D+2\varepsilon)*2\sqrt{k}}{D-2\varepsilon}\,\frac{D-2\varepsilon}{2\sqrt{k}}=D+2\varepsilon$$

である．この時，O と R の真値の最小距離は D より大きい．

[性　質　8]

（1）　C 内に観測値を持つオブジェクト数が M より多ければ，観測値を C 内に持つ全てのオブジェクトは PDB 外れ値ではない．

（2）　$C\in L_1(C)$ に観測値を持つオブジェクト数が M より多ければ，観測値を C 内に持つ全てのオブジェクトは PDB 外れ値ではない．

（3）　$C\in L_1(C)\ \in L_2(C)$ に観測値を持つオブジェクト数が M 以下であれば,C 内に観測値を持つ全てのオブジェクトは PDB 外れ値である．

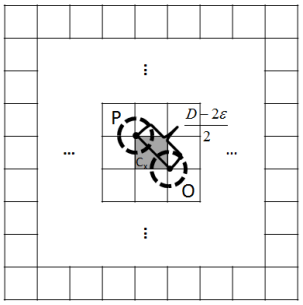

図 4　性質 5

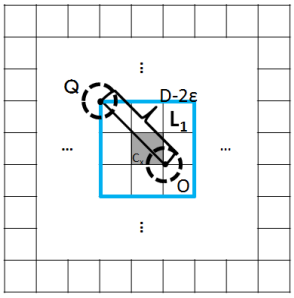

図 5　性質 6

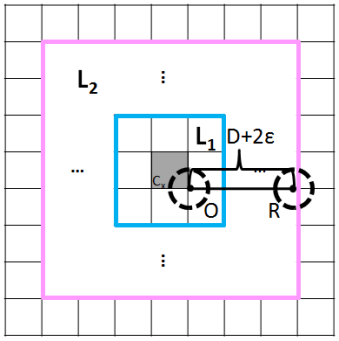

図 6　性質 7

4.3.2　　セルと用いたアルゴリズム

アルゴリズム2はセルを用いたアルゴリズムである．セル C_x の n_0，n_1，n_2 の値をそれぞれ $n_0(x)$，$n_1(x)$，$n_2(x)$ と表す．1-2行目で各セル C_x の $n_0(x)$ の値を初期化する．3-4行目でオブジェク

トOiが含まれるセルを特定する．5-7行目ではn0＞Mであるセルをredとする(性質8(1))．8-11行目ではL1近傍にredセルが存在するセルをpinkとする．これは，任意のセルCxのL1近傍にredセルが存在する場合，Cxは必ず性質8(2)を満たすからである．14-15行目でredもしくはpink 判定されなかったセルでかつ，性質8(2)を満たすセルを pinkとする．18-19行目で性質8(3)を満たすセル内の全オブジェクトを外れ値とする（性質8(3)）．最後に20-27行目で，19行目までの処理でred，pink，もしくは判定されなかったセルCw をwhiteとし，whiteと判定されたセルに対して22-27 行目にてセルに含まれる各オブジェクトOi にPDB外れ値判定を行う．まず22 行目にて，Cw中のOiのD近傍内のオブジェクト数Countにn0(w) + n1(w)を代入する．これは，CwのL1近傍のセルに含まれるオブジェクトは必ずOiのD近傍内に存在することが保障されているからである（性質5）．次に，Oi とL2近傍のセルに含まれるオブジェクトOpとの距離を計算し，OpがOiのD近傍内に存在する確率を計算し，OiのD近傍内にあるオブジェクト数の期待値（CountとOpがOiのD近傍内に存在する確率の和）がMを超えた時点でOiは外れ値ではないと判定する．L2近傍の全オブジェクトを走査しても，Oi のD近傍のオブジェクト数の期待値がM個以下であれば，Oiは外れ値であると判断する．

アルゴリズム2: セルを用いたアルゴリズム

1. **for each** セルCx do
2. n0(x) ← 0
3. **for each** オブジェクトOi **do**
4. Oi を含むセルCx を特定し，n0(x) を 1 増やす.
5. **for each** セルCx **do**
6. **if** n0(x) > M **then**
7. Cx をred とする.
8. **for each red** セルCr **do**
9. **for each** Cr のL1 近傍のセルCu
10. **if** Cuがred でない場合**then**
11. Cu をpink とする.
12. **for each** セル内にオブジェクトが存在し，色がついていないセルCw **do**
13. $n_1(w) \leftarrow \sum_{C_i \in L_1(C_w)} n_O(i)$
14. **if** n0(w) + n1(w) > M **then**
15. Cw をpink にする.
16. **else**
17. $n_2(w) \leftarrow \sum_{C_i \in L_2(C_w)} n_O(i)$
18. **if** n0(w) + n1(w) + n2(w) ≤ M **then**
19. Cw 内の全オブジェクトを外れ値とする.
20. **else**
21. Cw をwhite とする.
22. **for each** オブジェクトOi ∈ Cw **do**
23. Count ← n0(w) + n1(w)
24. **for each** オブジェクトOp ∈ Cu,Cu ∈L2(Cw) **do**
25. **if** dist(Oi,Op) ≤ D **then**
26. オブジェクトOp がオブジェクトOiのD近傍に存在する確率を計算し，計算した結果とCountの和がM以下であれば，Oi は外れ値ではないと判断し，次のオブジェクトの処理(22 行目）を開始する.
27. Oi を外れ値とする.

4.3.3 D 近傍内のオブジェクトの確率計算

まず，性質 6 により，観測値が L1 近傍内に存在するオブジェクトが D 近傍に含まれる確率は 1 である．さらに性質 7 により，観測値が L_2近傍外に存在するオブジェクトが D 近傍に含まれる確率は 0 である．このため，観測値が L_2近傍に存在するオブジェクトについてのみ D 近傍に存在する確率を計算する．

2 次元の例を図 7 に示す．それぞれの小さい円はオブジェクト O，A の不確実領域である．一番大きい円は O の不確実領域の中心に対する D 近傍を表す. 二つの円が重なっている黒い部分はAの不確実領域のうち，D 近傍に入る部分を表す．この際，A が O の D 近傍に存在する確率を，この重なっている部分の面積とオブジェクト A の不確実領域の面積の割合を用いて計算する．計算する一般式は次のようになる．

$$\Pr(D:O,A)=\int_{\varepsilon(O)} \frac{S(x,y)}{A\text{の不確実領域の面積}} P(x,y)dxdy$$

ただし，ε(O)はオブジェクト O の半径 ε の不確実領域，S(x,y)はオブジェクト A の不確実領域のうち，O の D 近傍に入る部分の面積、 P(x, y)はオブジェクト O が x, y に存在する確率を表す確率密度関数を表す.

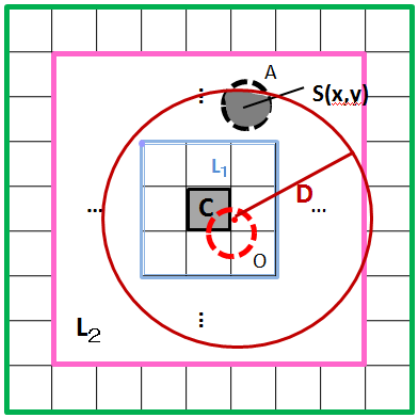

図 7　D近傍に存在する確率

実際には，O の位置は不確実領域内で不定である．一般の k 次元の場合について，オブジェクト A がオブジェクト O の D 近傍内に存在する確率を計算する一般式は次のようになる。

$$\Pr(D:O,A)=$$
$$\int\cdots\int \delta[d(O,A)\le D]P(x_{1O},\cdots x_{kO})P(x_{1A},\cdots x_{kA})$$
$$dx_{1O}\ldots x_{kO}\,dx_{1A}\ldots x_{kA}$$

ただし，D は D 近傍を与える距離，$d(O,A)$ は O の真値と A の真値の距離，δ( )はその距離が D 以下の時 1，それ以外の

時 0 を返す関数, $P(\chi_{1O,}\cdots\chi_{kO})$, $P(\chi_{1A,}\cdots\chi_{kA})$ はそれぞれオブジェクト O と A が, $(\chi_{1O,}\cdots\chi_{kO}),(\chi_{1A,}\cdots\chi_{kA})$ に存在する確率を表す密度関数である

5. 実験

本章は提案手法の評価実験について述べる。まずは、提案手法により検出されたデータを評価する実験について述べる．次は、不確実性を考慮した時の処理時間への影響の実験について述べる。

実験に使用したのは，AMD Athlon™64 × 2 Dual Cure Processor 3800+2GHz の CPU と 1982MB のメインメモリを持つ Microsoft Windows XP マシンであり，すべてのアルゴリズムの実装は Java1.6.0 02 で行った．

5.1 実験データ

使用した人工データ 1 は, 各オブジェクトの現在位置情報（x 座標と y 座標の値）をオブジェクトの状態として持つ不確実データとした．オブジェクトの分布は，オブジェクトの x 座標，y 座標がそれぞれ一様分布に従うように制定し，幅 50 の枠の外側が密に，幅 150 の枠の内側が疎になるように設定し，基本的に内側に存在するオブジェクトが外れ値オブジェクトとして検出しやすいように設定した．なお，全オブジェクト数は 1300 であり, 幅 50 の枠の外側に 1200 個オブジェクトと幅 150 の枠の内側に 100 個オブジェクトである．範囲は 1000×1000 である．

使用した人工データ 2 は, 各オブジェクトの現在位置情報（x 座標と y 座標の値）をオブジェクトの状態として持つ。オブジェクトの分布は，オブジェクトの x 座標，y 座標がそれぞれ正規分布に従うように制定し，中央分布が密に，周辺部分が疎になるように設定した．

5.2 提案手法により検出されるデータ

5.2.1 実験概要

提案手法（データの不確実性を考慮した手法）を用いた場合と従来手法（データの不確実性を考慮しない手法）を用いた場合について外れ値検出されたオブジェクトの比較を行った．

具体的には以下の通りである．

まず，元データに ε に基づく 0 から δ の範囲でゆらぎを与えて生成した 10 組のランダムデータセットに対して，従来手法を用いて，外れ値を検出する．次に，元データに対して，提案手法を用いて，外れ値を検出する．

5.2.2 実験結果

表 1､表 2 は人工データ 1 と人工データ 2 それぞれに対して、従来手法と提案手法により検出された外れ値の比較を示している．外れ値 ID は検出されたオブジェクトの ID である．検出率は, その ID を持っているオブジェクトが 10 組のデータ集合において外れ値として検出された割合を表している．

表 1 は D=120, M=5, δ=20 をパラメータとした．従来手法では 10 個オブジェクトを外れ値として検出した．提案手法では 7 個オブジェクトが検出された. また, 提案手法は検出率 0.9 以上のオブジェクトのみ検出した．表 2 は D=16, M=20, δ=6 をパラメータとした．従来手法では 6 個オブジェクトを外れ値として検出した. 提案手法では 5 個オブジェクトが検出された. また, 提案手法は検出率 0.7 以上のオブジェクトのみ検出した．

結果をみると，提案手法により，検出回数が多いオブジェクトがより多く検出された．逆に，従来手法では検出回数が多いオブジェクトだけではなく，検出回数が低いオブジェクトも検出された．すなわち，データの不確実性を考慮しない従来手法よりデータの不確実性を考慮した本研究の提案手法の方が，より安定な外れ値の検出が実現できる．

表 1　人工データ 1 の実験結果

外れ値ID	検出率
1211	1.0
1217	1.0
1228	1.0
1247	1.0
1267	1.0
1295	1.0
1293	0.9
1269	0.6
1254	0.5
1207	0.3

従来手法

外れ値ID	検出率
1211	1.0
1217	1.0
1228	1.0
1247	1.0
1267	1.0
1295	1.0
1293	0.9

提案手法

表 1　人工データ 2 の実験結果

外れ値ID	検出率
836	1.0
1378	1.0
1383	1.0
321	0.8
82	0.7
1238	0.3

従来手法

外れ値ID	検出率
836	1.0
1378	1.0
1383	1.0
321	0.8
82	0.7

提案手法

5.3 不確実性を考慮した時の処理時間への影響

5.3.1 実験概要

D, M の値を変化させて, 不確実性を考慮した場合と不確実性を考慮しない場合に対して処理時間を計測した．

5.3.2 実験結果と考察

図 8、図 9 は，人工データ 1 に対して、D の値と M の値と処

理時間の関係を表している．図 10、図 11 は、人工データ 2 に対して、D の値と M の値と処理時間の関係を表している．

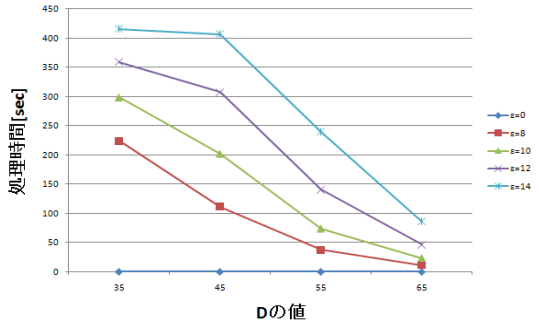

図 8　人工データ 1: Ｄの値と処理時間の関係

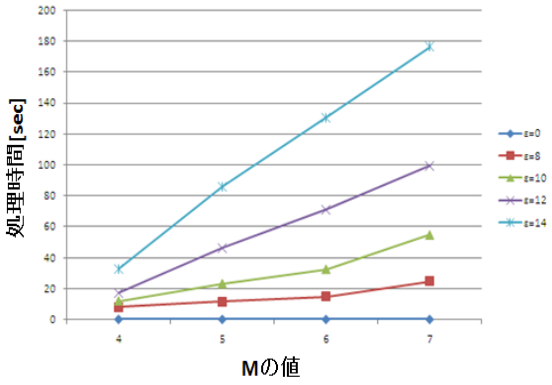

図 9　人工データ 1: Ｍの値と処理時間の関係

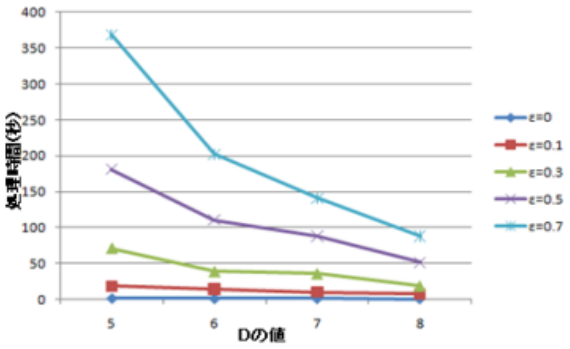

図 10　人工データ 2: Ｄの値と処理時間の関係

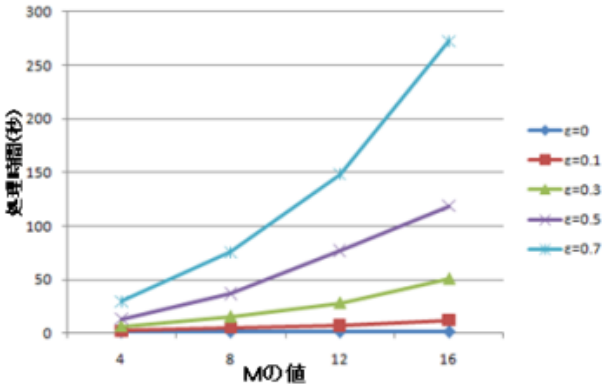

図 11　人工データ 2:Mの値と処理時間の関係

図 8～11 の結果から、近傍の大きさ D が小さい程，セルの数が増えるため，処理時間が増加することが分かる．　また，不確実領域の半径 ε が大きい程，その影響は大きい．

近傍にはオブジェクト数 M が増加に伴い，確率計算が必要なオブジェクトペアが増加するため，処理時間が増加する．

また，不確実領域の半径 ε が大きい程，その影響は大きい．

6.　まとめと今後の課題

本研究は不確実データ集合に対する距離に基づいた外れ値検出手法を提案した．さらに人工データを用いた実験を行い，提案手法の有効性を評価した．また不確実性が大きい程，処理時間を増加することが示された．

今後の課題として，計算処理の効率化について，検討して予定である．

謝辞

本研究の一部は科学研究費補助金特定領域研究(# 21013004)による．

参考文献

[1] V. Barret and T. Lewis, Outliers in statistical data, Wiley, Chichester, 2001.

[2] R. T. Ng and J. Han, “Efficient and effective clustering methods for spatial data mining, ” VLDB, 1994.

[3] M. M. Breuning, H. P. Kriegel, R. T. Ng and J. Sander, ” Lof: identifying density-based local outliers, ” SIGMOD, 1998.

[4] Charu C. Aggarwal and Philip S. Yu, ”Outlier detection with uncertain data, ” SIAM, 2007.

[5] Bin Wang, Gang Xiao, Hao Yu and Xiaochun Yang, ”Distance-based outlier detection on uncertain data, ”CIT, 2009.

[6] E. M. Knorr, R. T. Ng and V. Tucakov, ” Distance-based outliers：algorithms and applications, ” VLDB, 2000.

[7] Kozue Ishida and Hiroyuki Kitagawa, ” Detecting current outliers: continuous outlier detection over time-series data streams, ”DEXA, 2008.

[8] 郭　楽, 天笠　俊之,　北川　博之, ”不確定性を有するデータ集合に対する外れ値検出, ”情報処理学会第 72 回全国大会.