時系列データに対する効果的な外れ値検出

石田 梢† 北川 博之‡

†筑波大学第三学群情報学類 ‡筑波大学大学院システム情報工学研究科
‡筑波大学大学院計算科学研究センター

1 はじめに

データ集合中の他から逸脱した特徴や値をもつオブジェクトを検出する外れ値検出はデータマイニングの重要な手法の一つであり，これまでに統計に基づく手法 [1] や距離に基づく手法 [2] など，様々な手法が静的なデータ集合に対して提案されてきた．一方で，センサデバイスやユビキタスコンピューティングの進展により時系列に変化するストリームデータは増大し，これらにに対するデータマイニングは重要となっている．

そこで我々は時系列で変化するデータに対して連続的な外れ値検出を効率的に行う手法を，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 (Distance-based outlier)[2] を基に提案する．

本稿では，時刻 t_j における各オブジェクトの属性値の集合を S^j とする．データストリームとして (S^1,S^2, $\cdots$,S^m)(ただし現在時刻を t_m とする) があるときに，連続的に S^m 中の外れ値を検出する．

上記を行うための最も単純な方法は，新しい属性値集合 S^m が到着するたびに S^m に対して上記のような静的な手法を用いて外れ値を検出することである．しかし，外れ値検出は計算コストが大きいため，毎時間単位ごとにこれを繰り返すのは効率が悪い．

そこで，一般に属性値集合 S^m は直前の S^{m-1} が変化したものでありその変化があまり大きくない場合も多いことを利用し，S^{m-1} における外れ値検出の結果を利用した差分計算を用いて，効率的に S^m における外れ値検出を行う手法を提案する．

2 距離に基づく外れ値検出手法

2.1 外れ値の定義

データ集合 S 中のオブジェクト O が $DB(p,D)$ 外れ値であるとは，O からの距離が D より大きい範囲に S のうち p 以上の割合のオブジェクトが存在するということである．ただし $M(=N(1-p))$ とし，O からの距離が D 以下の範囲を D 近傍と呼ぶ．

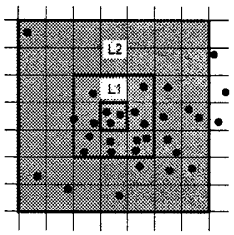

図 1: 2 次元のセル構造の例

2.2 外れ値検出アルゴリズム (Cell-Based Algorithm)

上記の定義にそのまま従い，データ集合 S 中の全オブジェクトに対して外れ値検出を行うと，大量の距離計算を行うため計算量は膨大になる．そこで提案されたのがデータ集合にセル構造を取り入れてオブジェクト毎の計算の数を減らす Cell-Based アルゴリズム [2] である．

Cell-Based アルゴリズムでは，オブジェクトを一辺の長さが $l = \frac{D}{2\sqrt{k}}$(k は次元数) のセルに量子化をし，セル毎に外れ値の判定を行う．そのために各セル C の L_1 近傍，L_2 近傍を定義する．

[L_1 近傍] C に隣接したセルを C の L_1 近傍とする．

[L_2 近傍] C にの全オブジェクトに対し必ず D 近傍外であるセル以外のセルで，C と L_1 近傍でないセルを L_2 近傍とする．

従って，C 内の全オブジェクトに対して，L_1 近傍の範囲は必ず D 近傍内であり，L_2 近傍よりも外の範囲は D 近傍外である．2 次元の例を図 1 に示す．

そこで以上の定義に基づき，次に示すラベル付けを行い，セル毎の外れ値判定を行う．

C 内に M 以上のオブジェクトが存在するセル (*red*) と $C \cup L_1$ 内に M 以上のオブジェクトが存在するセル (*pink*) 内の全オブジェクトは外れ値ではない．一方，$C \cup L_1 \cup L_2$ に M 以上のオブジェクトが存在しないセル (*yellow*) 内の全オブジェクトは外れ値である．従っ

Effective Outlier Detection over Time-Series Data
Kozue ISHIDA†, and Hiroyuki KITAGAWA‡
†College of Information Sciences
‡Graduate School of Systems and Information Engineering
‡Center for Conputational Sciences
University of Tsukuba
†‡Tennodai1-1-1, Tsukuba-shi, Ibaraki, 305-8573 Japan

て，*red*，*pink*，*yellow* 以外のオブジェクトを持つすべてのセル (*white*) にのみオブジェクト毎の距離計算による外れ値判定を行えばよい．これにより距離計算が減り，大幅な計算時間の短縮となる．

3 連続的DB外れ値検出アルゴリズム

3.1 準備

Cell-Based アルゴリズムを基に連続的な外れ値検出の手法を提案する．提案手法では，初期集合 S^1 を除く2回目以降の属性値集合を対象とする．ただし，S^1 に対する外れ値検出は Cell-based アルゴリズムを使用し，各セルは L_1 近傍のオブジェクト数 $CountL_1$，L_2 近傍のオブジェクト数 $CountL_2$，ラベル情報を持つこととする．

また，t_m における属性値集合 S^m は，S^{m-1} と，t_{m-1} から t_m の間に値に変化したオブジェクト (移動オブジェクト) の集合 $\Delta = \{O_r^m, O_s^m, ...\}$ から作成される．

3.2 外れ値検出対象セル

Δ から S^m の外れ値を検出する際の処理が必要とされる最小限のセルを示す．図2〜5はそれぞれの例を表している．ただし対象は中央のセルであり，赤い点は移動オブジェクトを表す．

A. セルを移動した移動オブジェクトを持つセル

B. L_1 または L_2 近傍にセルを移動した移動オブジェクトを持つセル

C. セル内に移動オブジェクトを持ち，L_2 近傍に *white* セルが存在するセル

D. セル内に移動オブジェクトを持つる *white* セル

A，B はセル内のオブジェクト数，$CountL_1$ または $CountL_2$ が変わり，ラベルが変わる可能性がある．C は L_2 近傍の *white* セル内のオブジェクト，D はセル内の移動オブジェクトのオブジェクトごとの外れ値判定がそれぞれ変更される可能性がある．

以上の最小限のセルのみに対して外れ値判定を行う．このように外れ値判定を行うセルを限定することで冗長な計算を省き，効率化を図る．

4 実験

提案手法による処理速度の向上を，移動体のシミュレーションデータを用いて検証した．比較対象は単位時間毎に Cell-Based アルゴリズムを行う手法とした．

図6は全オブジェクト数を10000，移動オブジェクトの平均速度を 1.5[m/s]，$D = 15$[m]，$p = 0.9995$ とし，単位時間 (1sec) あたりに移動する移動オブジェクトの割合を変化させて，処理時間を検証した．移動オブジェクトの割合が100% に達しても提案手法 は比較手法の 66% 程度の処理時間に抑られることがわかる．

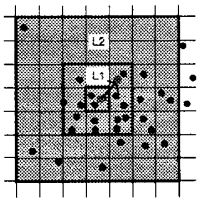

図 2: A

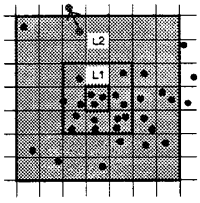

図 3: B

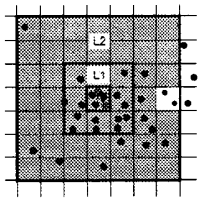

図 4: C

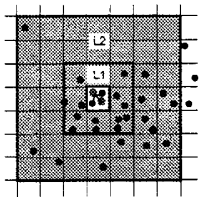

図 5: D

5 おわりに

本論文では時系列データに対する連続的な外れ値検出手法を提案し，実験により提案手法の妥当性を評価した．

謝辞

本研究の一部は科学研究費補助金特定領域研究（# 19024006）による．

参考文献

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

[2] Edwin M. Knorr, Raymond T.Ng and Vladimir Tucakov, "Distance-based outliers: algorithms and applications", The VLDB Journal, 2000.

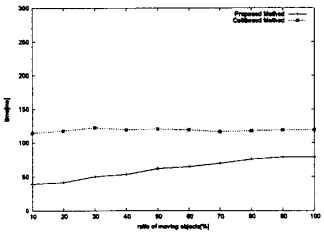

図 6: 実験結果