社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－145 (21)
2008－ＦＩ－ 91 (21)
2008／6／20

トランザクションデータに対するより汎用的な外れ値検出の実現に向けて

成田 和世† 北川 博之†‡
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センター‡
narita@kde.cs.tsukuba.ac.jp, kitagawa@cs.tsukuba.ac.jp

Toward A More General Outlier Detection for Transaction Databases

Kazuyo Narita†, Hiroyuki Kitagawa†‡
Graduate School of Systems and Information Engineering, University of Tsukuba†
Center for Computational Sciences, University of Tsukuba‡

1. はじめに

外れ値検出は，データベース中の規則性から著しく逸脱しているデータを発見する技術である．多くの既存外れ値検出手法は連続値を属性値に持つ数値型レコードデータを対象としている．しかし一方で，世の中には連続値以外の情報を持つデータも多数存在している．例えば，POS データに代表されるトランザクションデータである．このようなデータから外れ値を検出する技術の開発はまだそれほど進展しておらず，より汎用的な手法の提案は重要であると考えられる．

我々はトランザクションデータに着目し，外れ値となるトランザクションを検出するフレームワークを提案する．本稿ではスペースの都合から，提案手法の概要と，提案手法が現在抱える問題点や今後の課題について簡単に説明する．

本稿の構成は次のようである．2.で説明に必要な準備知識を述べる．3.で提案手法を概説する．4.で提案手法の問題点や今後の課題について述べ，まとめる．

2. 準備

トランザクションデータ D はトランザクションの集合である．トランザクション $t \in D$ はアイテム集合である．アイテム集合 X の D におけるサポートを $\sup(X)$ とすると，与えられた最小サポート $msup$ に対して $\sup(X) \geq msup$ である X を頻出アイテム集合と呼ぶ．互いに疎なアイテム集合 X, Y に対して，記述 $X \rightarrow Y$ を相関ルールと呼ぶ．本稿では特に，非常に大きな最小確信度 $mconf$ が与えられた場合を考える．このときに，頻出アイテム集合 X, Y ($X \cap Y = \phi$) で作られる相関ルール $X \rightarrow Y$ の確信度 $\mathrm{conf}(X \rightarrow Y)$ が mconf 以上であるとき，$X \rightarrow Y$ を高確信度ルールと呼ぶことにする．

3. 提案手法

統計的に見て，高確信度ルール $X \rightarrow Y$ は X に対する Y の相関が強い規則性であると考えられる．このとき，$X \subseteq t$ であるトランザクション $t \in D$ が，$Y \not\subset t$ であれば t は X と Y 間の強い相関に反しており，外れ値でありそうだと考えられる．$|Y - t|$ が大きいほど相関に反している度合いは強い．この考えに基づきトランザクション t の外れ値度 $\mathrm{OD}(t)$ を導入するために，我々は相関性閉包という新しい概念を導入する．R を高確信度ルールの集合としたとき，トランザクション $t \in D$ の相関性閉包 t^+ は次のように求める．

$$t^0 = t$$
$$t^{i+1} = t^i \cup_{X \rightarrow Y \in R} \{e \mid e \in Y \wedge X \subseteq t^i \wedge e \notin t^i\}$$
$$t^+ = t^\infty$$

相関性閉包を用いると，$\mathrm{OD}(t)$ は次式となる．

$$OD(t) = |t^+ - t| / |t^+|$$

我々は与えられた閾値 mod に対して，$\mathrm{OD}(t) \geq mod$ である t を外れ値として全て検出する．このような外れ値を全て検出するためのアルゴリズムは，大きく二つの段階に分けられる．最初の段階で D から高確信度ルールの集合 R を生成し，次の段階で各 $t \in D$ の外れ値度 $\mathrm{OD}(t)$ を R を使って計算する．mod 以上の t を外れ値として出力する．アルゴリズムの詳細はスペースの都合で省略する．

4. おわりに

本稿ではトランザクションデータから外れ値となるトランザクションを検出するためのフレームワークについて簡単に概説した．スペースの都合から本稿では省略したが，実験の結果から，提案手法には以下の二つの問題点が挙げられる．一つは三つのパラメタを用意したことで，パラメタセンシティビティが発生している点である．提案手法に与えるパラメタセットによって，検出される結果が大きく変化する場合がある．二つ目の問題点は，計算処理に時間が掛かる点である．これについてはナイーブアルゴリズムを改良したアルゴリズムで，速度の向上が得られた．しかし，我々は今後の課題の一つに，オンラインで外れ値の検出を可能とすることを考えており，そのためにはやはりパラメタセンシティビティと処理速度の向上の問題は解決すべき事項と思われる．

発表では提案手法の詳細な説明をした後，問題点や課題に対する現時点での我々の対応案を述べ，それに関する活発な意見交換を行いたいと考えている．

参考文献

1) Kazuyo, N. and Hiroyuki, K.: Detecting Outliers in Categorical Record Databases Based on Attribute Associations, Proc. of APWeb 2008, pp. 111-123.
2) Kazuyo, N. and Hiroyuki, K.:Outlier Detection for Transaction Databases using Association Rules, Proc. of WAIM 2008. (to appear)