2F1-04

マルコフモデルに基づく時系列データからの知識発見

Knowledge Discovery from Sequential Data based on Markov Model

湯上 伸弘　吉田 由起子　小林 健一　太田 唯子
Nobuhiro Yugami　Yukiko Yoshida　Kenichi Kobayashi　Yuiko Ohta

(株)富士通研究所
FUJITSU Laboratories Ltd.

Mining frequent patterns is an effective approach to discover useful knowledge from various kinds of databases such as POS data and web access log. However, it usually leads huge number of frequent patterns and it is difficult to understand the meanings of discovered patterns and to select useful ones. This paper proposes a new knowledge discovery tool, Action Browser, to resolve this difficulty. It summarizes and visualizes characteristics of the database focusing on orderings of events and helps a user to restrict the range of frequent patterns mining to filter out useless patterns.

1. はじめに

本稿では, 顧客毎の販売履歴や WEB のアクセスログ等の時系列データから有用なパターンを発見するための分析支援ツール Action Browser を提案する. 時系列データ中に一定の頻度以上で出現する頻出パターンの発見では, 既に様々な手法が提案されている[Mannila 97]. これらにおける最大の問題点のひとつは, 通常膨大な数のパターンが発見されてしまうことである. しかもそれらのほとんどは, パターン内のイベントの相関が低く統計的な意味がないものや既によく知られているパターン, 未知ではあっても分析の目的に合わないパターン等有用性が低く, いわば不要なパターンである. それらのなかから, 少数の実際に役に立つパターンを見つけ出すことは容易ではない. パターンの興味度等の評価基準[Brin 97]を工夫することである程度までは不要なパターンの数を減らすことができるが, それだけでは十分でない場合が多い. そのため, できるだけ不要なパターンを避け, 有用なパターンの発見を容易にするためには, 分析者が自分の持つ知識や分析の目的に合わせて注意深く分析の範囲や条件を設定する必要がある. Action Browser は, 分析対象の時系列データが持つ特徴を, 分析者からの要求に応じてインタラクティブに示すことで, 分析者がデータ中の様々な性質を理解するのを助け, 適切な分析範囲や条件を見つけるのを支援する.

2. 時系列データ

本稿で扱う時系列データは, 各種のイベントとその発生時刻の組の集合である. 同じイベントが複数回数発生したり, 同時刻に複数のイベントが発生したりしてもよい. また, 複数の時系列データが与えられてもよい. イベントのカテゴリが定義されていてもよい. 図 1に販売履歴の分析における例を示す. この例では, おにぎりやコロッケ等の商品の販売がイベント, 販売日がイベントの発生時刻であり, 顧客毎に複数の系列が存在する. おにぎりやコロッケは同じカテゴリ食料品に属し, ジュースやビールは飲料というカテゴリに属している.

顧客 ID	販売日	商品（イベント）
030011	2003/5/18	食料品／おにぎり
	2003/5/19	食料品／コロッケ
	2003/5/23	飲料／ジュース
030012	2003/5/30	食料品／幕の内弁当
	2003/5/30	飲料／ビール
	2003/6/13	飲料／ビール
	2003/6/14	雑貨／週刊誌
……	……	……

図 1 時系列データの例（販売履歴）

3. Action Browser

Action Browser の目的は, 分析者からの要求に応じて分析対象のデータの特徴を様々な粒度で様々な側面から可視化することで, 分析者がデータのもつ性質を短時間で把握し, 分析の目的にあったパターンを求めるための適切な分析条件や分析範囲の設定を容易にすることである. そのため, マルコフモデルに基づいてイベント間の遷移を可視化する機能や, 対象範囲に応じてイベントのクラスタリングを行うことで可視化する粒度を自動的に調整する機能, 可視化する範囲を詳細化していく際のナビゲーション機能, 他の分析ツールとの連携機能等を提供している. 以下でこれらの機能について説明する.

3.1 イベント間の遷移関係の可視化

Action Browser の基本的な機能は, 2節で説明した時系列データから, 各イベントの出現確率とイベント間の遷移確率からなるマルコフモデルを構築し, イベントをノード, 確率の大きい遷移をアークとする有向グラフとして可視化することである[小林 03]. 有向グラフの可視化はマグネティック・スプリング・モデル[三末 94]を用いて行う. 図 2は Action Browser による表示例である. 2つのイベント間の遷移確率は, それらが連続しておきる確率ではなく, 一方のイベントの発生したときに, 一定時間以内にもう一方も発生する割合として計算する. これは, ある程度時間を置いて発生するイベント間の関係も分析したいからである.

マルコフモデルの学習と表示を, 対象とするイベントの種類や発生時刻の範囲等を変えて繰り返すことにより, 分析者は,

連絡先：湯上 伸弘, (株)富士通研究所, 〒211-8588 川崎市中原区上小田中 4-1-1, yugami@flab.fujitsu.co.jp

分析の目的に関連する規則性を見つけ出すためにはどのようなイベントやカテゴリを考慮する必要があるかを短時間で判断できる.

3.2 イベントのクラスタリング

イベントの種類が数百あるいはそれ以上になると, 個々のイベントをノードとしたグラフを表示することは現実的ではないし, たとえそれを表示しても, そこから分析者がイベント間の関係を把握することは困難である. そこで Action Browser は, イベント間の遷移確率分布が持つ情報をできるだけ損なわないようにイベントのクラスタリングを行い, 表示もクラスタ単位で行うことで分析の粒度を自動的に調整する. 遷移確率分布の持つ情報量は, 実際の遷移確率分布と各イベントが互いに独立に発生すると仮定したモデルにおける遷移確率分布との差を Kullback-Leibler 距離を用いて評価する. すなわち, イベント間の相関をできるだけ残すようにクラスタリングを行う. クラスタの発生確率を $P(A)$, データ中のクラスタ A から B への遷移確率 $P(A\rightarrow B)$, 各イベントが独立に発生すると仮定した場合の遷移確率を $Q(B)$ (独立性が成立する場合遷移確率は A に依存しない)とすると, 遷移確率分布の持つ情報量は

$$\sum_A P(A)\sum_B \{P(A\rightarrow B)\log P(A\rightarrow B)/Q(B) + (1-P(A\rightarrow B))\log(1-P(A\rightarrow B))/(1-Q(B))\}$$

となる. この評価関数を最大化するようにクラスタリングを行うことにより, 互いに強い相関を持つイベントの集合や同一のイベントへ遷移する可能性の高いイベントの集合が同一のクラスタにまとめられる. その結果, 多数のイベントを対象とした場合には, 多くのイベント間に共通して現れる特徴を, 逆に対象イベントを絞り込んで分析した場合には, 個々のイベント単位での詳細な特徴を表示することができる.

3.3 分析のナビゲーション

3.2 節で述べたように, 多数のイベントを対象としている場合にはイベントのクラスタリングが行われ, クラスタ単位で可視化が行われる. そのため, 2つのクラスタ内の一部のイベント間について高い相関があっても, それらのクラスタ間にはアークが表示されない可能性がある. Action Browser は, このような相関を見逃すのを防ぐために, アークとして, クラスタに含まれるイベント／カテゴリ間の遷移の最大値を表示することで分析者の注意を促し, 分析範囲を詳細化していく場合のナビゲーションを行う.

3.4 異なるデータ間の比較

実際の多くの分析では, あるデータの持つ特徴を調べるだけでは不十分で, それを他のデータと比較することが要求される. 例えば商品の販売履歴の分析で特定の商品を購入した人々の特徴を調べる場合, その商品を買った人で頻出する特徴であっても, それが商品を購入していない人にも多く観察される場合には意味がない. Action Browser は, 分析対象のデータと比較対象のデータとの遷移確率分布の差を Kullback-Leibler 距離で評価し, 差への寄与の大きい遷移を表示するアークとして選択したり, クラスタリングを行う際の評価関数として用いたりする. これにより, 特定のデータでのみ観察される特徴を抽出・表示することができる.

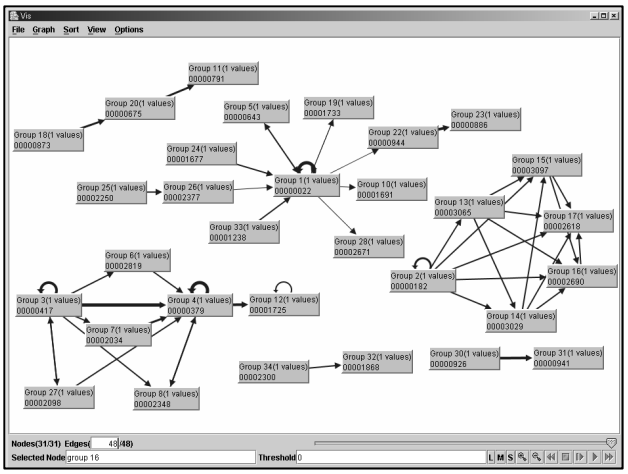

図 2 Action Browser の実行例

3.5 データ検索／頻出パターン分析との連携

Action Browser の画面上から直接データの検索や頻出パターン発見等の他の分析ツールを呼び出すことができる. これにより, Action Browser で観察される特徴を実際のデータや頻出パターンで確認して分析を進めることができる.

4. まとめ

本稿では, 時系列データを対象とした分析支援ツール Action Browser を紹介した. Action Browser を使うことにより, 短時間で分析対象のデータの特徴を把握し, 効率よく分析を進めることができると期待される. ただし現時点では, 実際にどの程度分析にかかる時間を短縮できるか, また得られる知識の質をどの程度向上させることができるかについての評価は不十分であり, 今後実際のデータへの適用を通じて有効性の検証を行う必要がある.

参考文献

[Brin 97] Brin, S., Matwani, R. and Silverstein, C: Beyond market baskets: generalizing association rules to correlations, Proceedings of SIGMOD, pp265-276(1997).

[Mannila 97] Mannila, H., Toivonen, H. and Verkamo, A. I.: Discovery of Frequent Episodes in Event Sequences, Data Mining and Knowledge Discovery, Vol.1 No.3, pp259-289(1997).

[小林 03] 小林健一他: マルコフモデルに基づく時系列データからの知識発見ツール Action Browser, 人工知能学会基礎論研究会資料集 SIG-FAI51, 2003.

[三末 94] 三末和男, 杉山公造: マグネティック・スプリング・モデルによるグラフ描画法について, 情報処理学会研究報告「グループウェア」No.007-003, 1994.