木構造データから有効なパターンを抽出するための グラフマイニングに関する研究

中原　孝信
（大阪府立大学大学院経済学研究科経済学専攻　現所属・同大学院経済学研究科経済学専攻）
指導教員　森田裕之 助教授

1. はじめに

本論文は，データマイニング研究の１つの方向であるグラフマイニングを応用し，ビジネスデータに対し適用可能な手法を提案している．グラフマイニングは，複雑な関係性を保存したままマイニングすることが可能であり，ビジネスデータに適用した場合，時間の捉え方によって異なる意味を持つ顧客の購買行動を，複数の観点から表現することが可能となる．

既存のグラフマイニング手法は，主に化合物を対象とした化学構造式への適用[2]や，WEBページなどのXML形式のデータに適用[1]されており，ビジネスデータに対する適用は，ほとんど行われていないのが現状である．グラフマイニングの計算は非常に複雑であることが知られており，このことが，大規模なデータを有するビジネス領域での適用を妨げる要因の１つになっている．

提案手法では，まずトランザクション形式であるID付きPOSデータを，複数の期間を同時に考慮して木構造で表現する．そして，そこからメタ戦略の１つである遺伝的アルゴリズムを応用した方法で，有効な部分パターンを抽出する．最終的に，抽出された部分パターンを決定木分析の説明変数として利用することで，判別モデルを生成する．

提案手法の適用事例として，某クレジットカード会社と，某スーパーマーケットのID付きPOSデータへ提案手法を適用し，有効性を検証している．その結果，２種類のデータともに，抽出したパターンを説明属性に利用することで，利用しない場合に比べてモデルの精度が向上しており，また，解釈可能なパターンを決定木モデルの上位に出現させることに成功した．このことから提案手法は，有効であると考えられる．

2. グラフマイニングを応用した手法の提案

提案する手法は，まず複数の時間を同時に考慮できる木構造によりID付きPOSデータを表現する（以下，ヒストリカル・ツリーと呼ぶ）．そこから，遺伝的アルゴリズム（GA）を応用したパターン抽出の方法について提案している．木構造への変換は，複数の期間（例えば季節，月，週，平日休日）で数値データを集計し，それらの集計値を少数のコードに写像する．そして期間の関係を壊さずに図１上部のような木構造で表現する．コードの写像は，集計値が個人にとって通常利用するような値かどうかという個人の感じる程度を表す個人コードと，他の顧客全体の集計値と比較して，どの程度かという全体コードの２種類のコードを１セットとし，複数のコードセットに対応可能な木構造による表現方法を提案している．

次に，GAにより部分パターンを抽出するために，遺伝子列へ変換する．図１下部は，木構造で表現されたコードを深さ優先探索順に遺伝子列に変換したものである．生成されるコードは，２種類のコードと，位置コードの３つからなる遺伝子列である．位置コードは，木の深さに合わせた番号をつけることで識別する．

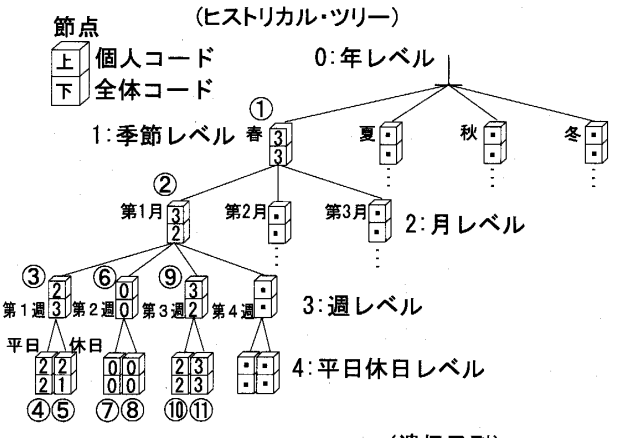

図１　ヒストリカル・ツリーと遺伝子列

そして，このようなコード化を各顧客全員に対して施し，2つの顧客集合の違いを識別できる有効な部分パターンをGAにより抽出する．このとき有効なパターンとは，一方の顧客集合のサポートを最大化（最小化）し，他方の顧客集合のサポートを最小化（最大化）するようなパターンであり，2つの2目的最適化問題での最適解が有効な部分パターンだと考え，GAにより解の探索を行う．

提案するGAの特徴は，エリート解（各世代の近似パレート解）集合をすべて保存して探索を行うこと，個人コードと全体コードを1セットとして用いる場合は，2つの親から1回の交叉で8つの子個体を作成していることである（Pセットのコードを用いる場合は，2^{2P+1}個である）．そして部分的なワイルドカードの利用を許容することで，完全に連続していない部分パターンを探索することも可能にしている点である．この方法により抽出された部分パターンを決定木分析の説明変数に利用することで，目的となる顧客集合を判別する．

3. 提案手法の適用事例

提案手法をクレジットカードとスーパーマーケットのID付きPOSデータにそれぞれ適用した．前者の分析目的は，一括払いに加えリボ払いを併用する顧客の購買特徴と，一括払いだけを利用する顧客の購買特徴の違いを識別し，将来リボ払いを併用するようになると期待される顧客の特徴を発見することである．

図2は，抽出したパターンと顧客属性などを説明変数に利用し，生成した決定木モデルである．判別精度は66%である．パターンを含めずに構築したモデルの精度は56%であり，パターンにより精度が向上した．また，決定木には抽出したパターンが上位の説明変数として出現している．パターン1や3は，月初の平日に利用するという購買行動を示しており，将来リボ払いを併用するであろう顧客の特徴として，これらのパターンが出現している．このようなパターンが出現する理由としては，当該クレジットカード会社の締め日が月末であり，月初に利用すると，返済までの期間が長期になることが考えられる．そして平日の利用であることを考えると，生活に必要なものを購買していると予想される．

スーパーマーケットの分析では，RFM分析により優良顧客を特定し，優良顧客の判別を目的に，あるカテゴリに限定した提案手法の適用例を示している．抽

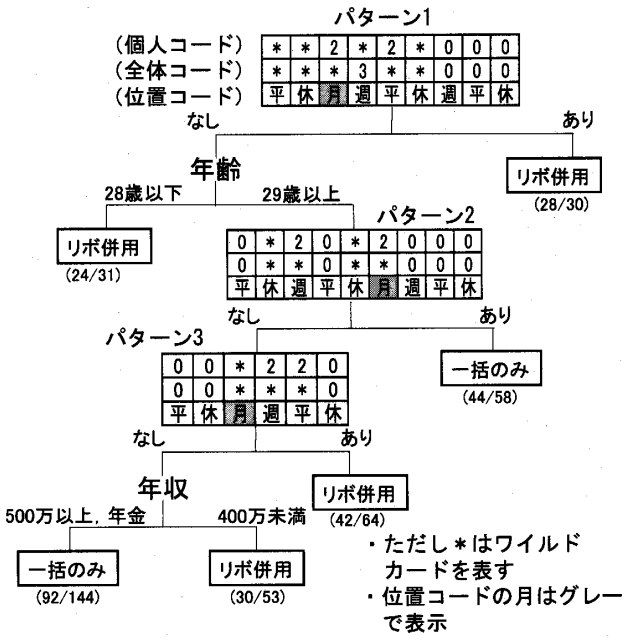

図2 決定木分析の結果

出された部分パターンは，212個であり，全て決定木分析の説明変数として利用し，将来優良顧客が一般顧客になるかを判別した．紙幅の都合上モデルの記述は省略させていただくが，モデルの精度は70%である．パターンを含めない場合の精度は64%であり，精度が6%向上している．また，この決定木モデルでも抽出したパターンは，モデルの上位に出現している．

計算実験の結果からいずれも判別精度が向上したこと，モデルの上位に抽出パターンが出現していることからも，提案手法は有効であると考えられる．

4. まとめ

グラフマイニングにおける1つの分析方法を提案し，2つの異なる種類のデータに対する計算実験から，その有効性を確認した．今後の課題は，一般化に向け適用事例を増やすことと，抽出された多くのパターンから，効率的により有効なパターンだけを選択する方法を考慮すべき点などが考えられる．

参考文献

[1] T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto and S. Arikawa. Efficient substructure discovery from large semi-structured data. In Proc. *SDM2002*, pp. 158-174, 2002.

[2] A. Inokuchi, T. Washio and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In *Proc. PKDD 2000*, pp. 13-23, 2000.