3A3-3

# 階層化が可能な時系列データからの特徴抽出

Feature Extraction from Categorical Time Series Data by Hierarchical Structure

福田　遼平[*1]　大野　博之[*2]　稲積　宏誠[*2]
Ryohei Fukuda　Hiroyuki Oono　Hiroshige Inazumi

[*1]青山学院大学大学院　理工学研究科
Graduate school of Science and Engineering, Aoyama Gakuin University

[*2]青山学院大学　理工学部　情報テクノロジー学科
College of Science and Engineering, Aoyama Gakuin University

Tree mining algorithm is very useful and effective to analyze Web information as graph-based semi-structure pattern mining. Considering categorical time series data, each value can be often evaluated with the measure of time periods with hierarchical representation. Therefore, such kinds of data are represented as semi-structure patters. We propose new method to extract effective patterns including wild cards from categorical time series data using tree mining algorithms.

## 1. はじめに

われわれを取り巻くさまざまな分野では，大量の蓄積データから有用な知識を獲得する試みがなされており，それらを実現するためのデータマイニング技術の研究が盛んに行われている．特に，属性・属性値ペアとして構造化されたデータに対しては，命題論理のクラスに属する決定木分析など，非常に多くの取り組みがなされている．一方，それら属性値間の関係構造を厳密に評価するためには，述語論理のクラスに属する手法が求められる．しかしながら，現実的な取り組みとしては，述語論理のクラスに準じた能力を有するグラフ構造表現されたデータに対するグラフマイニング手法が特に有用とされている．

グラフマイニング手法は，近年盛んに検討されており，AGM や gSpan および GBI などが提案されている．また，木構造データに限定したものとしては，FREQT や TreeMiner[Zaki 02] および TreeFinder[Termier 02] などが提案されている．ただし，これらの手法を用いるにあたっては，グラフ探索アルゴリズムそのものの計算量の問題から，大量データへの適用を阻んできたが，多くの改良が施され，徐々に実用性が増してきている．また一方，その有用性が強調されているにもかかわらず，その応用領域としては，化学構造式であったり，XML などの Web ページを対象としたものなど，明示的にグラフ構造や木構造で表現された対象に限定されているのが現状である．

そのような中で，クレジット利用履歴データを用いて，期間を考慮しながら購買金額を木構造でコード化し，あるアクションを起こす顧客を識別するための部分パターン抽出方法が提案されている [中原 05]．また，そこで抽出されたパターンを説明変数として用いることで，汎用的な手法である決定木モデルの精度向上に寄与するであろうとの見通しが示されている．ただし，ここでの抽出は遺伝的アルゴリズムを利用したものであり，木構造データを直接マイニングしたものではない．

本稿では，問題に応じて期間ごとに階層的に特徴づけが行えるようなカテゴリカル属性をもつ時系列データを対象とし，各時系列データを木構造表現し，それらに共通に包含される部分木を直接的に抽出することによって，有用な部分パターンの抽出が可能であることを示す．さらに，その適用領域の拡大について検討する．

連絡先: 福田遼平，青山学院大学大学院理工学研究科
〒 229-8558 神奈川県相模原市淵野辺 5-10-1
Email:r-fukuda@ina-lab.it.aoyama.ac.jp

## 2. 購買履歴データからの部分パターン抽出

クレジットカード利用履歴データの分析は以下のように示されている [中原 05].

1. 提供されたデータは年度ごとに分割し，初年度に一括払いだけを利用し，次年度にリボ払い・分割払いをした顧客（リボ併用顧客）と，初年度及び次年度にいずれも一括払いだけを利用した顧客（一括選好顧客）に分類する．
2. 観測期間を「日（平日と休日)」「週」「月」「季節」「年」を考慮して購買金額を集計する．この「日（平日と休日)」「週」「月」「季節」「年」により階層構造が構成される．
3. 顧客ごとのデータ表現として，各観測期間ごとに当該顧客の利用総額を基準として平均値からのずれにより利用金額区分（3 値) によるコード化を行う．
4. 各顧客の全体の各観測期間ごとの利用総額と比較して，平均値からのずれによる利用金額区分（3 値）によりコード化を行う．

これにより，各顧客情報は，各ノードが観測期間と 2 種類の利用金額区分によりコード化され，特徴付けられた深さ 5 の木構造で表現されることになる．いったん木構造表現した顧客情報の各ノードを遺伝子列に変換する．これを用いて，一括選好顧客のサポートを最小化（最大化）し，リボ併用顧客のサポートを最大化（最小化）する 2 目的最適化問題での解を有効な部分パターンとして GA を用いた解の探索を行っている．この部分パターンは，完全に連続していないものも対象とする．すなわち，部分的にワイルドカードコードを持つことを許している．このようにして得られた部分パターンは，データからすでに判明している顧客属性データとともに属性項目とされるため，決定木分析が可能となる．その結果，部分パターンを説明変数として加えることによってモデル精度の改善が実現できたとの報告がなされている．

このように，木構造表現された対象に対して，ワイルドカードコードを持たせながら共通部分パターンを見つけることは，親子関係か先祖関係を満足する 2 項関係を共有する共通部分

木を発見することに他ならない．そこで，2 目的最適化という観点からの候補パターンの絞込みの問題を課題としつつ，まずワイルドカードを含む部分パターン発見について，グラフマイニングによる取り組みについて検討することとする．

## 3. 時系列データからの特徴抽出

ここでは複数存在するカテゴリカル属性あるいは適当な処理によって離散化された数値属性をもつ時系列データを仮定し，共通する特徴パターンを抽出することを目的とする．また，時系列データを木構造で表現し，木構造で表したデータから共通部分木を抽出する．そして抽出された共通部分木から時系列上での共通部分を解釈するという手順をとる．

時系列データの木構造化は以下の手順で行う．

1. 注目する期間の決定
   注目する期間は問題背景と仮説に基づき，分析者が決定する．例えば平日・休日，月の前半・後半などによる違いがデータに表れていると考えられる場合，これらの期間に注目する．

2. ノードの作成
   時系列データを注目した期間ごとに分割し，期間とそれらの期間で集約された情報をノードラベルとしてコード化する．

3. データの木構造化
   注目した期間について階層表現を行う．これにもとづいて期間ごとのノードにより木構造表現する．ただし，同一階層で複数の表現が存在する場合（例えば平日・休日と，週の前半・後半など）には，同一の問題に対して異なる木構造表現を作成する．

次に，木構造化されたデータから共通部分木を抽出する．ただし，ここでの部分木とは，その親子関係あるいは先祖関係のいずれかが対象とする木に共通に含まれているものと定義する．したがって，本稿における共通部分木抽出アルゴリズムは，TreeMiner や TreeFinder を用いるものとする．

先祖関係を考慮した部分木を探すことで，その部分木を，その最上位の期間内にワイルドカードを含む共通パターンとみなすことができる．

## 4. 実験

本稿で用いる時系列データとしては，本学において 13 項目のテストに合格することが義務づけられている情報スキル合格履歴データを用いる．これは，5 種類のテスト項目（基本操作，文書作成，表計算，プレゼンテーション，総合）からなり，受験順序は指定されていない．また，1 日で何科目も受験可能である．そこで，学生個々の合格履歴データからは，その取り組みのペースや受験順序の違い，また最終合格者と不合格者の特徴の違い，所属学部や学科による取り組み方の違いなどの発見が期待される．従って，これを用いて指導方針や運営の改善に役立てていくことなどが考えられる．

受講生一人分のデータに対して一つの木構造を作成する．対象とするデータを木構造化した例を図 1 に示す．

各ノードのラベルは以下の通りである．

**ルート:** 学生 ID

**学期レベル:** 期 ID（前期，夏期，後期）

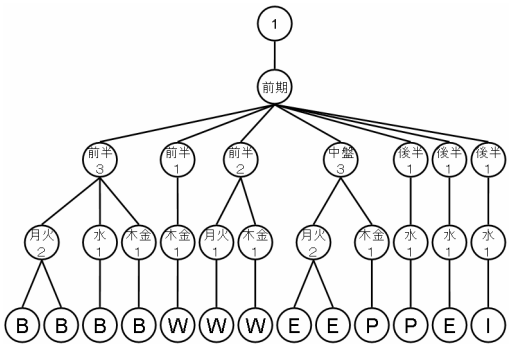


図 1: 受講生 1 人の合格状況を表す木構造

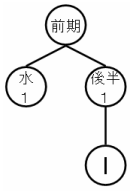


図 2: 共通部分木

**週レベル:** 週 ID（学期前半，中盤，後半）+ 合格数

**日レベル:** 曜日 ID（月火，水，木金） + 合格数

**葉ノード:** 項目 ID

合格履歴データの木構造化は必ずしも一通りだけではなく，同じデータに異なる木構造化手法を用いることもできる．

共通部分木の抽出には半構造データマイニング手法である TreeFinder を用いた．

抽出される部分木の例を図 2 に示す．これは前期に全項目を合格した学生から抽出された共通部分木の例である．この部分木から前期の水曜日に 1 項目合格し，さらに前期後半に総合 1 項目のみを合格した週が存在することを示している．また，総合項目を何曜日に合格したかは特定していない．

## 5. まとめ

本稿では時系列データを階層表現し，共通する特徴を抽出する方法を示した．これは，部分的にワイルドカードを含む時系列上の特徴パターンを自動的に抽出するものである．これを分類問題に活用していくためには，特徴抽出のための戦略を部分木抽出アルゴリズムに組み込ませる必要がある．抽出精度の向上と抽出した情報の活用が今後の課題である．

## 参考文献

[Zaki 02] M.J.Zaki: Efficiently mining frequent trees in a forest, *Proc. of the 8th International Conference on knowledge Discovery and Data Mining,* pp.71-80 (2002).

[Termier 02] A. Termier, M-C Rousset, MSebag.: TreeFinder: a First Step towards XMLDataMining, *IEEE ICDM'02*, pp.450-457 (2002).

[中原 05] 中原孝信, 森田裕之：GA を用いた購買履歴データからの有効部分パターンの抽出, 日本オペレーションズ・リサーチ学会　2005 年秋季研究発表会, pp.214-215(2005).