D-031

階層表現可能な時系列データからの有用な特徴抽出の試み

A Step towards Useful Feature Extraction from Time Series Data by Hierarchical Structure

福田　遼平 †　　大野　博之 ‡　　稲積　宏誠 ‡
Ryohei Fukuda　　Hiroyuki Oono　　Hiroshige Inazumi

1. はじめに

現在，時系列データからデータマイニング手法を用いて知識を発見するための多くの研究が行われている．複数の時系列データから共通する特徴を探す場合，細部は完全に一致していなくても，ある期間に注目して情報をまとめた場合には共通する傾向が存在する場合がある．特に時系列上の事象は階層的に表現できる場合が多く，その活用も重要なテーマといえる．

そこで，注目すべき期間単位にデータを整理し，再配置することで同一の時系列データに対してさまざまな分析が可能となる．例えば，クレジットカード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 [1]．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し，その結果を用いた効率的な決定木生成方法を提案する．

2. 時系列データからの特徴抽出

提案手法では，まず時系列データを階層化して木構造で表現し，木構造で表したデータから共通部分木を抽出する．その後，抽出された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．なお，本稿で対象とする時系列データは不等間隔に記録された，１属性のカテゴリカルデータである．

2.1 階層表現可能な時系列データ

時系列の情報を扱う多くの問題には，例えば年，季節，月などのようにまとまった期間が背景として存在する．さらにある月の情報と別の月の情報などのように同じ観測期間での情報や，ある年の情報とその年のある季節の情報などのように異なる観測期間での情報に何らかの関係が存在することがある．

本稿で用いる時系列データはこのような期間ごとの情報から関係を探すことができる，あるいは探すことに意味があるデータである．その方法として期間ごとの情報を階層的に表現して分析するため，背景として存在する期間のうち，長期の期間は，短期の期間を包含する必要がある．その結果，同一階層の兄弟関係ごとに時間関係が保持されることになる．

2.2 時系列データの木構造化と部分木抽出

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・後半などが問題背景として存在する場合，あるいは説明属性として注目できると仮定される場合，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，期間名とそれらの期間で集約されたデータによりノードラベルを決定する．
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い，それにラベルを付すこと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ことになる．

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．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からである．これを実現する方法として，TreeMiner[2] を用いる．

3. クラス分類に用いる属性の生成

ここでは，抽出された部分木の示す傾向を属性として，時系列データのクラス分類を行う方法を考える．しかしながら対象とする時系列データに一定以上含まれている部分木は，その含有率を低くすると膨大なものとなるであろうし，含有率を高くすると極端に減少することになる．クラス分類に有用な部分木を直接的に発見することは非常に困難であるので，本稿では抽出された多くの部分木を用いたクラス分類を考える必要がある．ところが，このような部分木は，ほとんど同じようなクラス分類能力をもつような冗長なものや，属性間に相関が存在するものが多数含まれることが予想される．その結果，正確なクラス分類や高い分類属性の説明機能が期待できない．

そこで本稿では抽出された部分木について，まず類似性の高い部分木同士でクラスタリングを行い，クラスタに含まれる部分木集合を改めて属性として考える．その結果，クラス分類において同様に機能する部分木の組み合わせのいずれが重要であるかに注目することになる．クラスタリングで使用する部分木間の類似度は，各部分木を含む時系列データ分布の類似性に基づいて算出する．またクラスタリング手法に k-means 法を用いることとする．

4. 分析

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 IT 講習会合格履歴データを用いる．これは，５種類のテスト項目からなり，それぞれ基本操作 (B)，文書作成 (W)，表計算 (E)，プレゼン

†青山学院大学大学院　理工学研究科　理工学専攻
‡青山学院大学　理工学部　情報テクノロジー学科

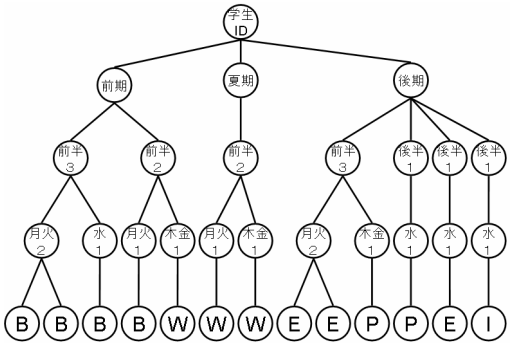

図 1: 受講生 1 人の合格状況を表す木構造

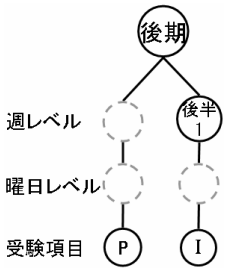

図 2: 共通部分木例 (A 学科サポート 34%)

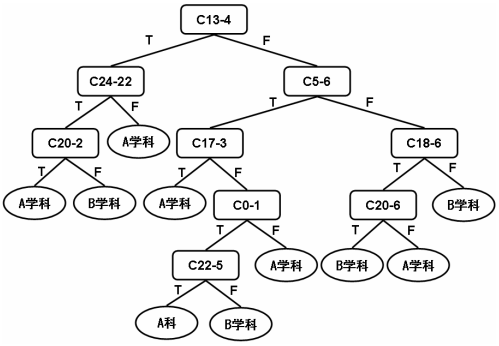

図 3: 全ての部分木を属性とした決定木

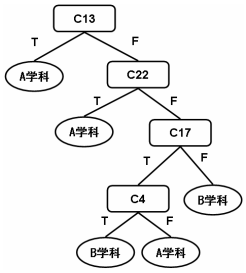

図 4: 部分木集合を属性とした決定木

テーション (P)，総合 (I) に分けられる．また，受験順序は指定されおらず，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の違いなどの発見が期待される．さらに，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．

受講生一人のデータに対して一つの木を作成する．各階層ごとのノードラベルは以下のように定義したうえで，対象とするデータを木構造化した例を図 1 に示す．なお，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．

ルート: 学生 ID

学期レベル: 期 ID（前期，夏期，後期）

週レベル: 週 ID（学期前半，中盤，後半）+ 合格数

曜日レベル: 曜日 ID（月火，水，木金）+ 合格数

受験項目: B，W，E，P，I

次に各学生の合格状況を表す木構造から共通する部分木を抽出した．図 2 は A 学科 81 名中，サポート 34%で抽出された部分木の例である．この部分木は，後期の期間内にプレゼンテーション項目を 1 項目合格し，かつ後期後半に総合項目 1 項目のみを合格という学生の傾向を示している．ただし，プレゼンテーション項目を合格した週と曜日，あるいは総合項目を合格した曜日については特に共通性はないことを示している．

次に A 学科，B 学科の合格者からサポートがそれぞれ 30%の部分木を抽出し，これらを用いて各学生の所属学科を決定クラスとした決定木分析を行った．決定木アルゴリズムには C5.0 を用いた．図 3 は抽出された部分木を全ての決定木属性に用いた場合であり，図 4 は両学科から抽出された計 515 個の部分木から k-means 法を用いて生成したクラスタを決定木属性に用いた場合の結果である．クラスタ数は 30 とし，属性値は各クラスタ内の部分木の半数以上が受験生の木構造に含まれている場合を T とし，そうでない場合を F とした．また，分類精度は全ての部分木を属性に用いた場合が 65.3%，部分木集合を属性とした場合が 74.9%となった．

図 3 における決定木のルートとなった部分木は，クラスタ 13 内の部分木の一つである．このクラスタ 13 には 16 の部分木が含まれており，またクラスタ 13 の部分木を 1 つでも含む学生は，この部分木集合の全てを含むことが分かった．そして，それらを全て含むことは前期期間内に全ての項目に合格していることを意味していた．

このように，部分木を直接決定木属性に用いた場合，クラスタ内のいずれかの部分木を選択することになり，上記の受験傾向を見逃す可能性がある．生成した新規属性は，クラス分類に有用な特徴を選択したものであることが分かる．

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．また，抽出された部分木からクラスタリングを用いて生成した部分木集合を属性として定義した．さらに実験を通してその有用性を示した．

参考文献

[1] 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).

[2] 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).