DEWS2007 E9-2

# 時系列上の階層関係に注目した特徴抽出手法の検討

福田　遼平†　　大野　博之††　　稲積　宏誠††

† 青山学院大学大学院 理工学研究科 理工学専攻　〒 229–8558 神奈川県相模原市淵野辺 5–10–1
†† 青山学院大学 理工学部 情報テクノロジー学科　〒 229–8558 神奈川県相模原市淵野辺 5–10–1
E-mail: †{r-fukuda,oono,hiro}@ina-lab.it.aoyama.ac.jp

**あらまし**　時間情報を含むデータから期間ごとに情報を集計し，それらの関係に注目することにより，有用な知識を発見することができる．本研究ではカテゴリカル属性をもつ複数の時系列データを以上の観点から順序木で表現し，共通部分木を探すことで特徴を抽出する．そして通常非常に多く得られる特徴を，元のデータの部分木保有状況からグループ化し，その説明能力とクラスごとの特徴を把握する上での有効性を検討する．
**キーワード**　データマイニング，時系列データ，グラフマイニング，半構造データ

# An Examination of Feature Extraction Technique Focused on Hierarchical Relation from Time Series.

Ryohei FUKUDA†, Hiroyuki OONO††, and Hiroshige INAZUMI††

† Graduate school of Science and Engineering, Aoyama Gakuin University　Fuchinobe 5–10–1, Sagamihara-shi, Kanagawa, 229–8558 Japan
†† College of Science and Engineering, Aoyama Gakuin University　Fuchinobe 5–10–1, Sagamihara-shi, Kanagawa, 229–8558 Japan
E-mail: †{r-fukuda,oono,hiro}@ina-lab.it.aoyama.ac.jp

**Abstract**　Finding hierarchical relations from categorical time series data can be an effective way for feature extraction. To do this, we convert the data into the ordered tree structures, and get sub-patterns with wild cards using a tree mining algorithm. As a result, we usually get many sub-patterns. So, we will use clusters of sub-patterns made by these supports of the time series data, and find characteristics of the data. We will use this method for classification problem.
**Key words**　Data Mining,　Time Series Data,　Graph Mining,　Semi Structured Data

## 1. はじめに

従来から，時系列データからの知識発見については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や傾向を探す場合，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．この場合，必ずしも連続した期間ではなく，ある特定の条件を満たす断続的な期間についての特徴が注目されることもある．また，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ことが予想される．このような視点から，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中原，森田によって提案されている [1].

中原らの提案する分析手法は以下の通りである．

（1） 対象とするクレジットカード利用履歴データから，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．

（2） 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．

（3） 顧客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．そして，一度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を顧客属性データとともに属性項目とすることによって，決定木分析を行う．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．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．

そこで本稿では，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，あるいはなんらかの処理によってカテゴリカル属性に変換できるような時系列データに注目する．このような，各時系列データを木構造表現し，それらに共通して包含される部分木を抽出し，それを直接活用する方法を検討する．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．ただし，一般にこのような部分木は大量に抽出されるため，それらを用いた意味づけの方法が重要となる．

本稿では，抽出された部分木を部分パターンとし，それを分類問題に適用するための方法を提案する．そもそも，大量に得られる部分パターンは，すべて説明属性の候補となり得るが，同様の説明機能を有するものが多く存在することが予想される．すなわち，単独のパターンを直接分類のための説明属性に用いた場合，ほぼ同等の説明機能をもつ他のパターンを見逃してしまうだけでなく，選択された部分パターンから，クラスの特徴を知るうえでの核心的な部分を知ることも難しくなる．そこで，類似した説明機能をもつパターン集合をクラスタリングにより求め，求められたクラスタを決定木の属性として用いることで，これらの問題点を解決することを考える．その結果，分類能力だけでなく，説明能力の向上が期待される．本手法を，本学における IT 講習会合格履歴データに適用し，その有効性を検証する．

## 2. 対象データの木構造表現と部分木抽出

### 2.1 対象とする時系列データ

本稿で対象とする時系列データは不定期な事象の生起にもとづく系列データである．

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．

### 2.2 時系列データの木構造化

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） 情報の集計

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） データの階層化

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．

このように木構造化した結果，元の時系列データは同一階層の兄弟関係ごとに時間関係が保持された順序木で表現されることになる．

### 2.3 部分木抽出

木構造で表したデータから共通部分木を抽出する．抽出手法には TreeMiner [2] を用いる．TreeMiner は複数の順序木に含まれる，頻出部分木をマイニングするアルゴリズムである．ただし，ここで木 s が木 t に含まれるとは，

i) s の全てのノードが t に存在する

ii) s においてノード $n_x$ がノード $n_y$ の親であれば，s の全ての枝 $b(n_x, n_y)$ に対して，t においては $n_x$ が $n_y$ の祖先となると定義し，このとき木 s は木 t の部分木としている．

このように定義することで共通ノードを親子関係からだけでなく，直接繋がっていない先祖関係のみの一致も許した柔軟な部分木を探し出すことができる．TreeMiner はこの定義を満たし，最小サポート値以上の木を全て取り出す．本稿では以降，木の包含についてはこの定義を用い，得られる頻出木を単に部分木と呼ぶこととする．

時系列データからの特徴と対応させた場合，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．例えば図 1 のように表現される事例データからは，図 2 に示すような部分木が抽出される．その結果，期間 B という大枠と，その後のいずれかの時期に C が起こるという特徴が発見されることとなる．

## 3. 部分木集合の要約

抽出された部分木から有用な特徴を探すために，本稿ではクラス分類に有用な属性を選択するという観点で取り組むこととし，抽出された部分木を決定木に用いる属性とする．しかし，抽出される部分木は通常多数存在し，ほぼ同等の分類能力を持つ属性が複数存在することになる．その結果，決定木生成においては，選択される属性は限られてしまい，有効な特徴を意味

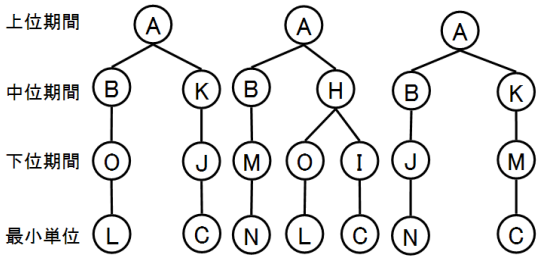


図 1　対象とする木構造

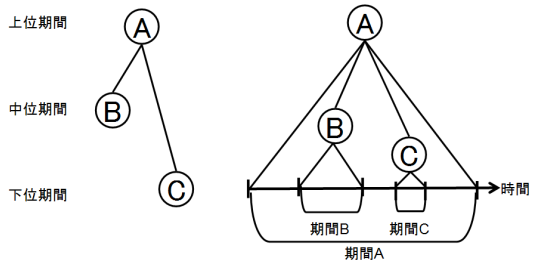


図 2　部分木例とその意味

する多くの部分木が活用されないままになってしまう．そこで，抽出された部分木について，まずクラス分類という観点から，類似性のある部分木同士でクラスタリングを行い，各クラスタを説明属性とする．クラスタの特性を見ることにより，クラス分類の要因となる時系列データの特徴を表現することが可能となる．ここで，部分木間の類似度は，各部分木が，元データから変換して生成される木構造全体の含有状況から算出する．

### 3.1　部分木のクラスタリング

M 個の木構造で表現された時系列データの事例から N 個の部分木が抽出されたとする．部分木 $S\{s_1, \cdots, s_N\}$ が元のデータを変換した木構造 $T\{t_1, \cdots, t_M\}$ のそれぞれに対して，保有される状況を表す行列 X を各クラスごとに作成する．$x_{ij}$ を $t_j$ が $s_i$ を含んでいるかを表す行列の成分とする．クラス a となる事例数を $M_a$ とすると，行列 $X_a$ は以下のようになる．

$$X_a = \begin{pmatrix} x_{1,1} & \cdots & x_{1,M_a} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,M_a} \end{pmatrix}, x_{ij} = \begin{cases} 0 & \text{if} \quad s_i \notin t_j \\ 1 & \text{if} \quad s_i \in t_j \end{cases}$$

各クラスの部分木保有状況を表した行列ごとに部分木を対象としてクラスタリングを行う．部分木間の類似度基準として，部分木が各事例に同時に含まれる場合を最も類似性が高いと考える．すなわち，各クラスの部分木保有行列の成分 $i, j$ に対して，その距離 $d_a(i, j)$ を次のように定義する．

$$d_a(i, j) = \sum_{l=1}^{M_a} \delta_a(x_{il}, x_{jl})$$

ただし

$$\delta_a(x_{il}, x_{jl}) = \begin{cases} 0 & \text{if} \quad x_{il} = x_{jl} = 1 \\ 1 & \text{if} \quad x_{il} = x_{jl} = 0 \\ 2 & \text{if} \quad x_{il} \neq x_{jl} \end{cases}$$

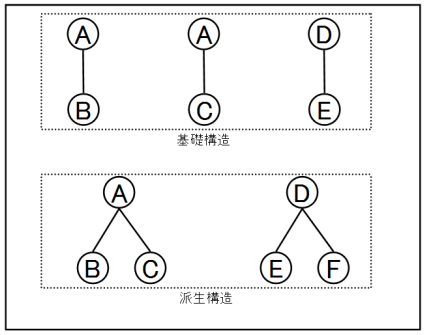


図 3　クラスタ内の基礎構造と派生構造

とする．この結果，各クラスについて，保有されている事例数のレベルごとに部分木集合が分類されることになる．

この類似度を用いて各クラスごとに部分木集合のクラスタリングを行う．今回は群平均法で階層的クラスタリングを行い，クラスタ数は任意に決定するものとする．

### 3.2　決定木分析

クラスごとに生成されたクラスタ $C_a\{c_{a1}, c_{a2}, \cdots, c_{ak}\}$ は，クラス a の事例のある部分集合に対して共通に含まれる部分木集合を意味している．

これら全てのクラスごとに求められたクラスタを分類属性として決定木分析を行う．その際，属性値の決定方法が問題となる．ただし，決定木分析を行う目的は，クラス分類に寄与する部分木あるいは部分木集合の特定であり，さらにそれが定性的な説明機能をもつことが望まれる．そこで，各クラスタの特徴を基礎構造と派生構造に分けて考えることとする．

基礎構造とは，クラスタ内のいかなる部分木も自分自身の部分木とはならないものである．派生構造とは，必ず基礎構造となる部分木を包含する部分木である．例えば図 3 に示す 5 つの部分木が同一クラスタに含まれているとする．この場合には，3 つの基礎構造と 2 つの派生構造から構成されることになる．そこで，説明属性である各クラスタの属性値を基礎構造を全て含むか否かの 2 値とする．その結果，決定木分析によって，まず基礎構造により特徴を説明し，その具体的な実現例を，派生構造を用いて補足説明することが可能となる．

## 4.　実　　験

本稿で提案する手法を，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に適用し，その有効性を検討する．

### 4.1　分析データ

IT 講習会合格履歴データは，5 種類のテスト項目からなり，それぞれ 4 種類の基本操作 (B)，3 種類の文書作成 (W)，3 種類の表計算 (E)，2 種類のプレゼンテーション (P)，総合 (M) 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り，全項目を合格することで単位を取得する仕組みとなっている．その結果，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（単位取得者）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．なお，このデータは合格履歴のみからなり，不合格の履歴は記載されていない．

表 1 に示すように，合格履歴は合格日が記載されたカテゴリカル属性のデータとなっている．また，日常的な学生の行動が関係していることから，受講生は学期・週・曜日などの期間において受験姿勢が違うことが推測される．

しかし，網羅的に示されたデータから年間を通した受験姿勢の変化を見ることは難しいため，各期間を関連性にもとづいて木構造化した後に分析することが，この問題に対する一つの解決策ではないかと考えている．そこで，各階層ごとのノードラベルは以下のように定義して木構造化する．

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（序盤: 4 週，中盤: 4 週，終盤: 4 週）
曜日レベル:　曜日 ID（月火，水，木金）
受験項目:　B，W，E，P，M

図 4 に木構造化した学生の合格状況の例を示す．このような木構造化を行った後，図 5 に示す部分木が抽出されたとする．これは，曜日に関係なく後期の序盤か中盤に P を合格し，終盤に M を合格することを意味している．図中右の表は部分木のルートから葉ノードまでのパス上を通るノードのラベルを 1 行で表し，時間順に並べたものとなる．

### 4.2　クラス間の特徴抽出

A 学科と B 学科の合格者（それぞれ 80 名）の合格状況を木構造表現した後に部分木を抽出し，これを用いて受験傾向の違いがどこにあるかについて，学科を目的属性とした決定木分析を行った．

#### 4.2.1　基礎構造からの分析

部分木抽出の過程では，両学科の学生の木構造を同時に抽出対象として，サポート 30%で部分木を抽出した．次に，得られた部分木に対して，クラスごとにクラスタリングを行った．クラスタは抽出された 2818 個の部分木からそれぞれ 30 ずつ生成し，合計 60 属性を説明属性とした．これを用いて生成された決定木（決定木 1）を図 6 に示す．ただし，葉ノードはクラスと（正解クラスの事例数）/（ルールをみたす全事例数）を表す．決定木分析には C4.5 をベースにして作られた，weka の J4.8 を用いた．

次に，両学科間での受験姿勢について大まかな分析を行うために，図 7 で示すような，葉ノードの最小事例数を 10 にして枝狩りを行った決定木を用いて検討することとする．表 8 は，4 つのクラスタ内の，基礎構造が示す受験傾向である．両学科間の特徴は基礎構造の部分木から次の 4 つの受験傾向で表すことができる．

受験傾向 1:　前期の比較的早い週に基本操作 2 項目か文書作成項目を合格．中盤にも何かを受験．週の前半に受験．遅くとも中盤までに文書作成を合格．
受験傾向 2:　表計算項目を学期の終盤に合格．前期中盤の週に受験し，その後文書作成項目を合格．
受験傾向 3:　学期中盤の週に受験．週の後半に受験．この時期に文書作成，表計算，プレゼンテーション項目を合格．
受験傾向 4:　後期の序盤，中盤に受験．後期の中盤以前の週後

表 1　学生の合格履歴例

| 認定日 | 合格項目 |
|---|---|
| 5/10(火) | B |
| 5/17(火) | B |
| 5/18(水) | B |
| 5/23(月) | B |
| 5/31(火) | W |
| 5/31(火) | W |
| 6/28(火) | W |
| 6/28(火) | E |
| 6/28(火) | E |
| 6/28(火) | P |
| 7/4(月) | P |
| 7/5(火) | E |
| 11/22(火) | M |

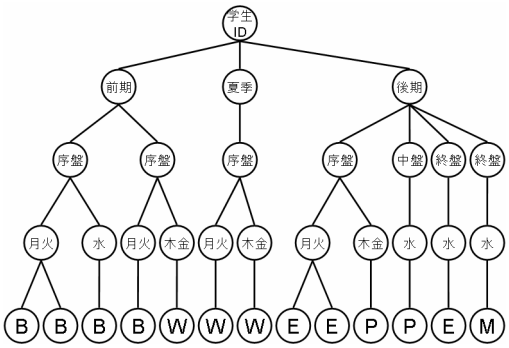


図 4　受講生 1 人の合格状況を表す木構造

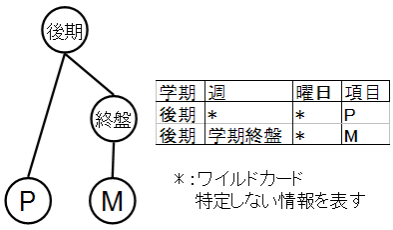


図 5　抽出される部分木例

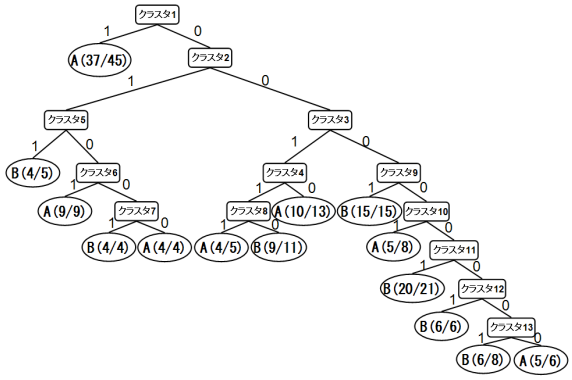


図 6　決定木 1: クラスタ数 60

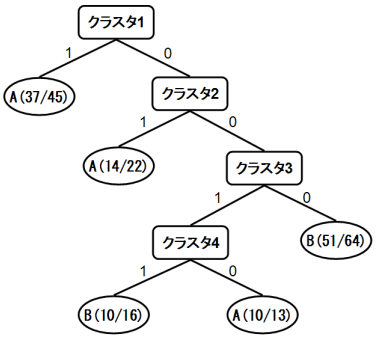


図 7　決定木 1: クラスタ数 60，葉ノードの最小事例数 10

表 2　クラスタ内の基礎構造が示す受験傾向（決定木 1）

| クラスタ ID | 部分木 ID | 学期 | 週 | 曜日 | 項目 |
|---|---|---|---|---|---|
| 1 | 1-1 | 前期 | 序盤 | * | B |
| 部分木数 4 | | 前期 | 序盤 | * | B |
| | | 前期 | * | 月火 | ＊ |
| | | 前期 | * | * | W |
| | 1-2 | 前期 | * | * | B |
| | | 前期 | * | * | B |
| | | 前期 | * | 月火 | * |
| | | 前期 | * | * | W |
| | | 前期 | 中盤 | * | * |
| 2 | 2-1 | * | 終盤 | 月火 | E |
| 部分木数 20 | 2-2 | 前期 | 中盤 | * | * |
| | | 前期 | * | * | W |
| 3 | 3-1 | * | 中盤 | 木金 | * |
| 部分木数 8 | 3-2 | * | 中盤 | * | E |
| | 3-3 | * | * | 木金 | P |
| | 3-4 | * | * | 木金 | W |
| | 3-5 | * | * | 木金 | E |
| 4 | 4-1 | 後期 | 序盤 | * | * |
| 部分木数 34 | 4-2 | 後期 | * | * | E |
| | | 後期 | 中盤 | * | * |
| | 4-3 | 後期 | * | 木金 | * |
| | | 後期 | 中盤 | * | * |
| | 4-4 | 後期 | 中盤 | * | * |
| | | 後期 | 中盤 | * | * |
| | 4-5 | * | 中盤 | * | P |

半に受験し，表計算項目合格．学期中盤の週にプレゼンテーション項目を合格．

以上の受験傾向と決定木のルールから，両学科の特徴として以下の受験傾向が挙げられる．

A 学科

- 受験傾向 1 をもつ
- 受験傾向 1 をもたず，受験傾向 2 をもつ
- 受験傾向 1，2，4 をもたず，受験傾向 3 をもつ

B 学科

- 受験傾向 1，2 をもたず，受験傾向 3，4 をもつ
- 受験傾向 1，2，3 をもたない

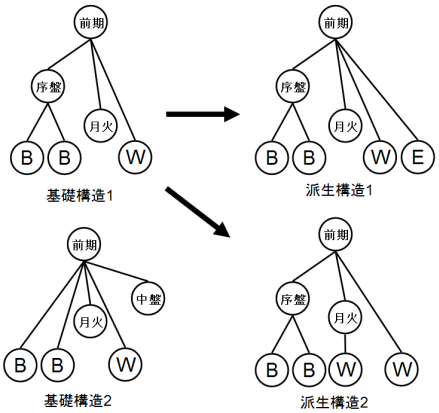


図 8　クラスタ 1 の部分木

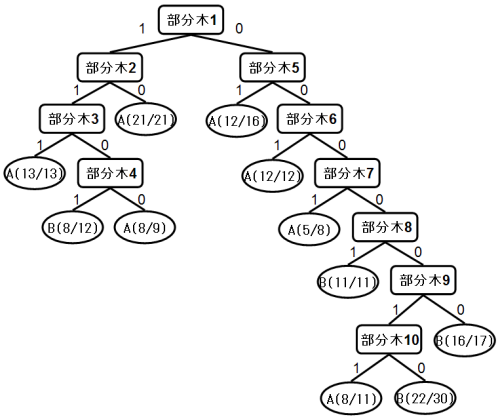


図 9　決定木 0: 部分木を直接属性に使用

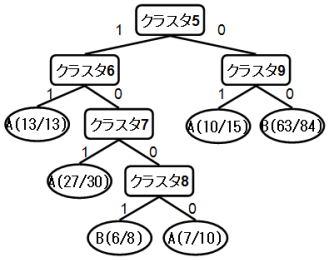


図 10　決定木 2: クラスタ数 100，葉ノードの最小事例数 8

#### 4.2.2　派生構造からの分析

クラスタ 1 には全部で 4 つの部分木が存在する．その基礎構造と派生構造を図 8 で示す．例えば，派生構造 1 は，基礎構造 1，2 を含む事例中，90%に含まれている．このような場合にはこの性質である中盤以降に表計算に合格するという傾向も利用することができる．

このように，必要に応じて基礎構造と派生構造を組み合わせて利用することで，説明能力の向上が期待できる．部分木の組み合わせによる特徴表現を実現していることが本提案の利点といえる．

### 4.3　部分木を直接決定木属性に用いた場合との比較

部分木そのものを決定木属性とした場合と本手法についての比較を行う．図 9，10 は部分木そのものを属性とした場合の決定木（決定木 0）と，クラスタ数 100，サポート 30%としたと

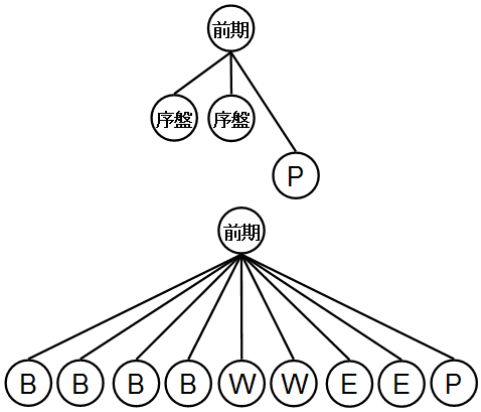


図 11　クラスタ 5 の派生構造例

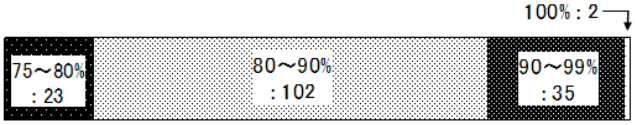


図 12　クラスタ 5 の基礎構造をもつ学生内での派生構造のサポート

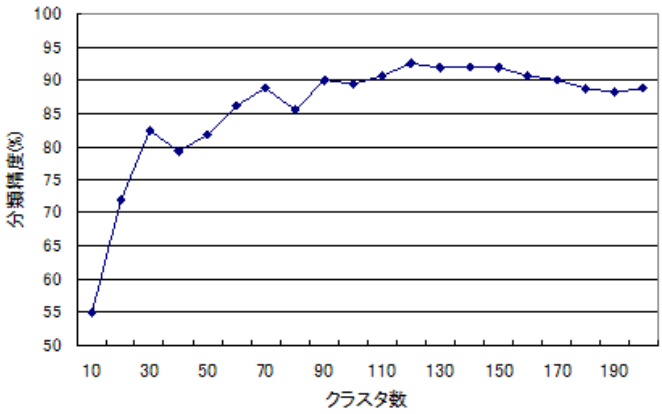


図 13　クラスタ数と分類精度

きの決定木（決定木 2）である．

ここで，決定木 0 において分類属性となった部分木に注目する．表 3 に示すとおり，決定木 0 の分類属性となった部分木は，ルートノードで選択された部分木 1 を除いて，決定木 1，2 では出現していないことがわかる．決定木 0 ではこれら部分木の組み合わせによるプロダクションルールにより受験傾向を示すことになる．

一方，決定木 2 の基礎構造による受験傾向は表 4 に示すとおりである．この 2 つの決定木では，ルートノードで選択された属性として前期のプレゼンテーション項目の合格が選択されている．ただし，クラスタ 5 にはその派生構造となる部分木が 162 個存在し，部分木 5-1 に類似する部分木が数多く存在することがわかった．これは，前期プレゼンテーション項目を 1 回受験したことに加えて，非常に多くの付加情報を含んでいることが示されており，部分木 1 は，そのほんの一例であることを意味する．派生構造の例を図 11 に示す．

図 12 ではクラスタ 5 の派生構造が，基礎構造を示す学生内でどれぐらいサポートされているかを示してたものである．図

表 3　部分木を属性とした決定木が示す受験傾向

| 部分木 ID | 決定木 1,2 との関係 | 学期 | 週 | 曜日 | 項目 |
|---|---|---|---|---|---|
| 1 | クラスタ 5 | 前期 | * | * | P |
| | の派生構造 | 前期 | * | * | P |
| 2 | 含まれない | 前期 | 序盤 | * | * |
| | | 前期 | * | * | B |
| | | 前期 | 中盤 | * | * |
| | | 前期 | 中盤 | 月火 | * |
| 3 | 含まれない | * | * | 水 | B |
| 4 | 含まれない | 前期 | 序盤 | * | * |
| | | 前期 | * | * | W |
| | | 前期 | * | * | W |
| | | 前期 | * | * | W |
| | | 前期 | * | 木金 | * |
| 5 | 含まれない | 前期 | * | 月火 | * |
| | | 前期 | * | * | W |
| | | 前期 | 中盤 | * | * |
| | | 前期 | 終盤 | * | * |
| 6 | 含まれない | 前期 | 序盤 | * | * |
| | | 前期 | 中盤 | * | * |
| | | 前期 | * | * | E |
| 7 | 含まれない | 前期 | 序盤 | 月火 | * |
| | | 前期 | * | * | W |
| | | 前期 | * | * | W |
| | | 前期 | * | * | W |
| 8 | 含まれない | 前期 | * | * | B |
| | | 前期 | * | * | B |
| | | 前期 | * | * | B |
| | | 前期 | * | * | B |
| | | 前期 | * | * | W |
| | | 前期 | * | * | W |
| 9 | 含まれない | 後期 | * | * | E |
| | | 後期 | * | * | E |
| | | 後期 | * | * | P |
| 10 | 含まれない | 前期 | * | * | B |
| | | 前期 | * | * | B |
| | | 前期 | * | * | B |
| | | 前期 | * | * | B |

に示すように大半の派生構造は基礎構造をもつ学生内で 80%以上サポートされていることがわかる．決定木属性として単独の部分木を使用した場合，もし同一クラスタ中の部分木の一つが選択された場合には，残りの部分木はその後決定木で選択されない可能性が高く，これらの特徴を見逃してしまうことになる．

さらに，図 13 にサポート値一定（30%）にしたままで，クラスタ数を変化させた場合の決定木の分類精度を示す．一定以上のクラスタ数に対しては，ほぼその分類精度が安定していることがわかる．

### 4.4　受験傾向の分析

最終的な分析は，以上行った決定木分析の結果を踏まえて行われることが望まれる．例えば，決定木 1 で求められた受験傾向 1 について，前期の平均合格項目数は A 学科が 9.76，B 学科が 6.35 であり，A 学科の学生は前期に多くの項目を合格し，B 学科の学生は多くの項目を後期に合格していることからも裏づ

表 4　クラスタ内の基礎構造が示す受験傾向（決定木 2）

| クラスタ ID | 部分木 ID | 学期 | 週 | 曜日 | 項目 |
|---|---|---|---|---|---|
| 5 | 5-1 | 前期 | * | * | P |
| 部分木数 163 | | | | | |
| 6 | 6-1 | * | 終盤 | 水 | * |
| 部分木数 1 | | | | | |
| 7 | 7-1 | 前期 | 序盤 | * | B |
| 部分木数 3 | | 前期 | * | * | B |
| | | 前期 | * | 月火 | * |
| | | 前期 | * | * | W |
| 8 | 8-1 | 前期 | * | * | B |
| 部分木数 2 | | 前期 | * | 月火 | * |
| | | 前期 | * | * | B |
| 9 | 9-1 | 前期 | 序盤 | 月火 | * |
| 部分木数 61 | | 前期 | 終盤 | * | * |
| | 9-2 | 前期 | * | 月火 | * |
| | | 前期 | * | * | B |
| | | 前期 | 終盤 | * | * |
| | 9-3 | 前期 | * | 月火 | * |
| | | 前期 | 中盤 | * | * |
| | | 前期 | 終盤 | * | * |
| | 9-4 | 前期 | * | 月火 | * |
| | | 前期 | 中盤 | * | W |
| | | 前期 | * | 月火 | * |
| | 9-5 | 前期 | * | 月火 | * |
| | | 前期 | * | 月火 | * |
| | | 前期 | 終盤 | * | * |
| | 9-6 | 前期 | * | 月火 | * |
| | | 前期 | * | * | W |
| | | 前期 | 終盤 | * | * |

けられている．決定木 1，2 では A 学科の学生に対して前期の早い時期での受験が顕著な特徴として示され，また，その一方で，B 学科を特徴付けるものとして，週の後半や，後期受験が示されており，両学科の受験姿勢の違いを，抽出されたパターンを用いてうまく表現されていることがわかる．

また決定木 2 のルートノードに選択されたクラスタ 5 では，前期中のプレゼンテーション項目の合格に関する受験傾向が示されており，これを満たす多くの学生が A 学科に分類されている．プレゼンテーション項目は，IT 講習会において総合項目の直前に学ぶ項目と位置づける学生が多く，この項目を合格する学生は，総合項目以外を全て合格している場合が多い．すなわち，これは早めに多くの項目を合格していることも示しているといえる．さらに，クラスタ 5 にはその派生構造として前期中に多数の項目を合格しているという部分木が存在していることからも裏づけされる．このように，基礎構造を中心として傾向を定めた後に派生構造により，さまざまな分析が可能となることから，実際の問題を反映した特徴を抽出することができたといえる．

## 5. ま と め

本稿では，時間情報を含むカテゴリカル属性のデータを，期間ごとの階層関係を用いて表現し，グラフマイニングアルゴリズムを用いて部分木で表現された共通する特徴の抽出を行った．さらに，抽出された特徴である部分木集合をクラスタリングし，生成されたクラスタ内の基礎構造を定義した後に，その有無による決定木学習を行った．抽出された特徴は，階層表現にもとづく，ある規則性を表現したものである．また，クラスタリングを行うことにより，類似した分類能力を有する部分木集合を分類属性として用いることが可能となり，単一の部分木のみの利用に比べて，説明能力が向上されるといえる．

実験からわかるとおり，クラスタリングのサイズを変更することによって，分類属性の性質も変化することがわかる．今後は，可変のクラスタリングサイズをうまく活用する方法，クラスタリングの基準の見直しによる精度向上，複数の階層表現を組み合わせることによる，多様な分析の実現などのユーザビリティの向上をめざしたい．

## 文　献


[1] 中原孝信，森田裕之 “ターゲット顧客を識別するためのクレジット購買履歴データを用いたパターン分析，” オペレーションズ・リサーチ，Vol51，No.2，pp.89-96，2006.

[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.

[3] A. Termier, M-C Rousset, MSebag, “TreeFinder: a First Step towards XMLDataMining,” *IEEE ICDM’02,* pp.450-457 ,2002.

[4] 浅井達哉，有村博紀 “半構造データマイニングにおけるパターン発見技法，” 電子情報通信学会，Vol.J87-D1，No.2，2004，pp.79-96.

[5] 神嶌 敏弘 “データマイニング分野のクラスタリング手法 (1)，” 人工知能学会誌，Vol.18，No.1，pp.59-65，2003.

[6] 金城敬太，尾崎知伸，澤井啓吾，古川康一 “相関ルールとネットワーク分析による時系列データからの知識獲得，” 第 19 回人工知能学会全国大会，2005.

[7] 福田遼平，大野博之，稲積宏誠 “階層化が可能な時系列データからの特徴抽出，” 第 20 回人工知能学会全国大会，2006.

[8] 福田遼平，大野博之，稲積宏誠 “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，”FIT2006(第 5 回情報科学技術フォーラム)，2006.

[9] 松田喬，吉田哲也，元田浩，鷲尾隆 “グラフ構造に着目した肝炎データからの知識発見，” 人工知能学会研究会資料，SIG-KBS-A201，pp.67-72，2002.

[10] 矢田勝俊，宮脇あすか，元田浩，鷲尾隆，羽室行信 “グラフマイニングのビジネス応用，” 人工知能学会研究会資料，SIG-KBS-A304，pp.81-86，2004.