

## 動的時間伸縮法に基づく時系列データからの決定木学習

山田 悠<sup>†</sup> 鈴木英之進<sup>†</sup> 横井 英人<sup>††</sup> 高林克日己<sup>††</sup>

<sup>†</sup> 横浜国立大学 大学院工学研究院 〒240-8501 横浜市保土ヶ谷区常盤台 79-5

<sup>††</sup> 千葉大学 医学部附属病院医療情報部 〒260-8677 千葉市中央区亥鼻 1-8-1

E-mail: †yuu@slab.dnj.ynu.ac.jp,suzuki@ynu.ac.jp,

††yokoih@telemed.ho.chiba-u.ac.jp,takaba@ho.chiba-u.ac.jp

**あらまし** 本論文では、時系列属性を含むデータから決定木を学習する新しい方式を提案する。時系列属性は、値と時刻のペアについてのシーケンスとして表される時系列データを値にとる属性であり、種々の実応用問題に頻出するために重要であると考えられる。われわれが提案する時系列決定木は、内部ノードに時系列データを持ち、時系列データに関する距離に基づいて例集合を分割する。最初に動的時間伸縮法に基づく基準例分割テストを定義し、次にこれを用いた決定木学習法を示す。実験の結果、提案手法は、他の手法に比較して理解しやすく正確な決定木を学習できることが分かった。さらに医療問題への適用の結果、時系列決定木は知識発見に有望であることが分かった。

**キーワード** 時系列決定木, 時系列データ, 動的時間伸縮法, 知識発見, 医療検査データ

## Decision-tree Induction from Time-series Data Based on Dynamic Time Warping

Yuu YAMADA<sup>†</sup>, Einoshin SUZUKI<sup>†</sup>, Hideto YOKOI<sup>††</sup>, and Katsuhiko TAKABAYASHI<sup>††</sup>

<sup>†</sup> Faculty of Engineering, Yokohama National University 79-5 Tokiwadai, Hodogaya, Yokohama,  
240-8501 Japan

<sup>††</sup> Division of Medical Informatics, Chiba University Hospital 1-8-1 Inohana, Chuo-ku, Chiba,  
260-8677 Japan

E-mail: †yuu@slab.dnj.ynu.ac.jp,suzuki@ynu.ac.jp,

††yokoih@telemed.ho.chiba-u.ac.jp,takaba@ho.chiba-u.ac.jp

**Abstract** This paper proposes a novel approach for learning a decision tree from a data set with time-series attributes. A time-series attribute takes, as its value, a sequence of values each of which is associated with a time stamp, and can be considered as important since it frequently appears in real-world applications. Our time-series tree has a time sequence in its internal node, and splits examples based on similarities between a pair of time sequences. We first define our standard example split test based on dynamic time warping, then propose a decision tree induction procedure for the split test. Experimental results confirm that our induction method, unlike other methods, constructs comprehensive and accurate decision trees. Moreover, a medical application shows that our time-series tree is promising in knowledge discovery.

**Key words** Time-series Decision Tree, Time-series Data, Dynamic Time Warping, Knowledge Discovery, Medical Test Data

### 1. はじめに

決定木 [4], [11], [12] は内部ノードで分割テストを行い葉でクラスを予測する木構造の分類子であり、帰納学習において頻繁に用いられている。決定木を学習する種々のアルゴリズムは欠落値やノイズに対する頑健性に優れ、数多くの応用領域で成

功を納めてきた。分割テストは、例えば名目属性に関しては通常、属性値に基づき例を該当する子ノードに割り振る。決定木学習法を種々の問題に適用するために、数値属性、木構造属性 [1], [2], および集合属性 [15] などに関する分割テストが提案されてきた。

時系列データは、時間順に計測した値を並べたデータであり、

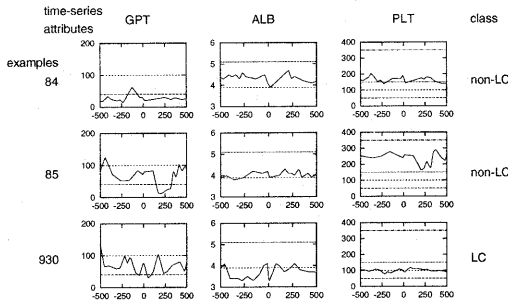


図1 時系列属性から構成されるデータ集合

Fig. 1 Data set which consists of time-series attributes

経済, 商業, 科学, 工業, および医学など, 種々の分野で頻出し重要視されている [9]. 時系列データを平均値などの統計量で置き換える手法は, 時系列データの構造を無視しており, 有効でないといわれる。一方, 時系列データの構造を扱う手法は, 時系列データを他の形式に変換するアプローチと, 時系列データの形を陽に扱うアプローチに分類できる。前者には, 時系列データをフーリエ変換などを用いて周波数領域に写像する手法や, カーネル関数などを用いて高次元空間に写像する手法が含まれる。後者は, 時系列データ同士のマッチングや距離を用いる手法が多い。なお時系列データの構造を扱う手法は, 時系列データ全体を対象とするアプローチと, 一部だけを対象とするアプローチに分類することも出来る。

本論文では分類モデルの分かりやすさを重視し, 時系列データの形を陽に扱い時系列データ全体を対象とする決定木の分割テストを提案する。時系列決定木は, この分割テストを用いる新しい決定木であり, 距離基準として動的時間伸縮法 (DTW) を用いる。以下, 2. 節において時系列データの分類学習問題を定義し, 3. 節において時系列決定木を提案する。4. 節では実データを用いて提案手法の有用性を示し, 特に慢性肝炎データへの応用を用いて本手法が医師の知識に近い分類モデルをほとんど領域知識を用いないで発見できたことを紹介する。5. 節は結論である。

## 2. 時系列データの分類学習

### 2.1 問題定義

時系列データ  $A$  は, 時間の経過とともに記録された計測値  $a_1, a_2, \dots, a_I$  を, 時間順に時刻と共に並べた系列データである。本論文では計測値が等間隔時間 (=1) で得られる場合を扱う。以下においては見やすさを重視し, 時系列データを曲線で表示する場合がある。

データ集合  $D$  は  $n$  個の例  $e_1, e_2, \dots, e_n$  から構成され, 各例  $e_i$  は  $m$  個の属性  $a_1, a_2, \dots, a_m$  とクラス属性  $c$  によって記述されるとする。属性  $a_j$  は値に 1 個の時系列データをとる時系列属性である<sup>(注1)</sup>。クラス属性  $c$  は名目属性であり, 属性

(注1): 名目属性, 連続値属性, 木構造属性 [1], [2] および集合属性 [15] など, 他の種類の属性を加えることも問題ないが, 紙面の都合上考えない

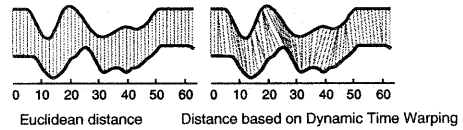


図2 ユークリッド距離と DTW に基づく距離を算出する際の, 時系列データペアの対応例

Fig. 2 Correspondence of a pair of time sequences in a Euclidean distance and a DTW-based distance

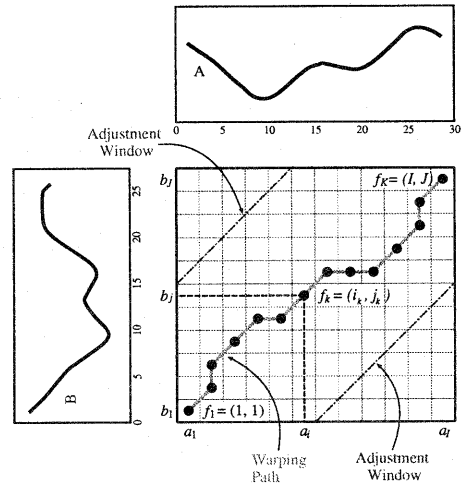


図3 ワーピングパスの例

Fig. 3 Example of a warping path

値をクラスと呼ぶ。図 1 に, 時系列データから構成されるデータ集合の例を示す。

時系列データからの分類学習では, 入力データ集合  $D$  から, 例  $e$  のクラス  $c(e)$  を予測する分類モデルの導出を目的とする。本論文では, 分類モデルとして決定木 [4], [11], [12] を考える。

### 2.2 動的時間伸縮法に基づく距離

本論文では時系列データのペアについて, 同時刻における計測値の差についての絶対和と定義した距離を, ユークリッド距離と呼ぶ<sup>(注2)</sup>。ただしユークリッド距離は, 計測値数が異なる時系列データのペアに適用できない上に, 人間の直観に反する結果を生じてしまう場合がある [7], [8]。これは, 人間は時系列データの形を柔軟に認識できるのに対し, この方法では時間方向の対応が固定化されるためである。

動的時間伸縮法 (DTW) [13] は, 時系列データのペアに関する相違度計算法であり, 時系列データにおける 1 点のデータをもう片方の時系列データにおける複数点のデータに対応づけられるため, 時間方向の非線形な伸縮を許容できる。図 2 に, ユークリッド距離と DTW に基づく距離を算出する際の, 時系列データペアの対応例を示す。

時系列データ  $A = a_1, a_2, \dots, a_I$  と  $B = b_1, b_2, \dots, b_J$  間の DTW に基づく距離  $G(A, B)$  を求めることを考える。 $A, B$  の対応づけをワーピングパスと呼び,  $I \times J$  平面上の格子点

(注2): 本稿では異常値に関する耐性を考え, 2 乗和ではなく絶対和を用いる。

$f_k = (i_k, j_k)$  の系列で表す.

$$\mathbf{F} = f_1, f_2, \dots, f_K \quad (1)$$

図 3 にワーピングパスの例を示す.

$a_{i_k}$  と  $b_{j_k}$  との距離を  $d(f_k) = |a_{i_k} - b_{j_k}|$  で表す.  $\mathbf{F}$  の評価関数  $\Delta(\mathbf{F})$  は式 (2) で表される.

$$\Delta(\mathbf{F}) = \frac{1}{I+J} \sum_{k=1}^K d(f_k) w_k \quad (2)$$

$\Delta(\mathbf{F})$  の値が小さいほど,  $\mathbf{A}, \mathbf{B}$  が類似していることになる. 極端な伸縮を防ぐために整合窓 ( $|i_k - j_k| \leq r$ ) の条件下で,  $\Delta(\mathbf{F})$  を  $\mathbf{F}$  に関して最小化する. ただし,  $w_k$  は  $f_k$  に関する正の重みで  $w_k = (i_k - i_{k-1}) + (j_k - j_{k-1})$  であり,  $i_0 = j_0 = 0$ .

この場合  $\Delta(\mathbf{F})$  が加法性を満たすため,  $\Delta(\mathbf{F})$  を最小化することは可能な  $\mathbf{F}$  を全て調べなくても解くことができる. 通常, 時間計算量が  $O(IJ)$  である動的計画法が用いられる.

### 2.3 従来の学習法の問題点

従来の決定木学習手法は時系列属性を想定していないため, 時系列データを含むデータ集合に適用する場合, データの前処理が必要となる. 最も単純な方法の一つとして, 時系列データを計測値の平均値で置き換える方法が考えられる. ただしこの方式は時系列データの構造を無視しており, 例えば形が大きく異なる時系列データが類似すると見なされてしまう欠点がある.

時系列データのペアに関して距離が定義されていることから, 決定木学習手法ではなく最小近傍法を用いて分類学習を行うことも考えられる. ただし, 最小近傍法は怠惰な学習 (lazy learning) であるために分類モデルが存在せず, 学習結果が分かりにくいという欠点がある.

## 3. 時系列決定木

### 3.1 基準例分割テスト

上記の問題を考慮し, われわれは時系列データからの分類学習の分類モデルとして, 時系列決定木を提案する. 時系列決定木は, 内部ノードに時系列データを持ち, 時系列データに関する距離に基づいて例集合を分割する. 内部ノードに時系列データを持つため, 時系列データの形を陽に扱うことができ, 可読性にも優れると考えられる. 距離基準は, 2.2 節で述べた理由により, DTW に基づく.

この分割テストは, 基準例分割テストと呼ばれ, 基準となる時系列データに基づいて例集合を分割する. 基準例分割テスト  $\sigma(e, a, \theta)$  は, 基準例  $e$ , 属性  $a$ , 類似閾値  $\theta$  から構成される.  $e$  の  $a$  に関する値を  $e(a)$  で表すと基準例分割は, 例集合  $e_1, e_2, \dots, e_n$  を,  $G(e(a), e_i(a)) < \theta$  を満たす例  $e_i$  から構成される例集合  $S_1(e, a, \theta)$  とそれ以外  $S_2(e, a, \theta)$  に分ける. この分割を,  $\theta$  によるギロチンカットとも呼ぶことにする.

われわれは分割の選択基準として, 最も頻繁に用いられている方法の一つである利得比基準 [12] を選択した. ただし,  $\theta$  によるギロチンカットでは属性毎に分割可能な点が  $n-1$  個と多いため, 利得比基準が最大となる分割点が複数個存在する場合が多い. この場合, 左右の子ノードに分割される例集合

の時系列データが, 似ていないほど良い分割であると考え.  $\sigma(e, a, \theta)$  の  $gap(e, a, \theta)$  は,  $S_1(e, a, \theta)$  内で  $G(e(a), e_i(a))$  が最大となる例  $e'$  と  $S_2(e, a, \theta)$  内で  $G(e(a), e_j(a))$  が最小となる例  $e''$  に関して,  $G(e'(a), e''(a))$  と定義される. 利得比基準が最大となる分割点が複数個存在する場合,  $gap(e, a, \theta)$  が最大となる分割を選択する.

以下に, 最良の基準例分割を求める手続き *standardExSplit* を示す. ただし  $\sigma.gr$  と  $\sigma.gap$  は, 分割テスト  $\sigma$  のそれぞれ利得比と  $gap$  を表す.

**Procedure:** *standardExSplit*

**Input:** 例集合  $e_1, e_2, \dots, e_n$

**Return value:** 最良の分割テスト  $\sigma$

```

1   $\sigma.gr = 0$ 
2  Foreacn(例  $e$ )
3    Foreach(時系列属性  $a$ )
4      現ノードの全ての例  $e_i$  を,  $G(e(a), e_i(a))$  をキーとして整理し,  $e_1, e_2, \dots, e_n$  とする
5      Foreach( $e_1, e_2, \dots, e_n$  の  $\theta$  によるギロチンカット分割  $\sigma'$ )
6        If  $\sigma'.gr > \sigma.gr$ 
7           $\sigma = \sigma'$ 
8        Else If  $\sigma'.gr == \sigma.gr$  And  $\sigma'.gap > \sigma.gap$ 
9           $\sigma = \sigma'$ 
10 Return  $\sigma$ 

```

### 3.2 クラスタ例分割テスト

上記の基準例分割から, 2 本の基準時系列データに基づく分割が自然に思い付く. 次の分割テストは, クラスタ例分割テストと呼ばれ, 2 個の中心例  $e', e''$  と選択属性  $a$  から構成される. クラスタ分割テストは, 例集合  $e_1, e_2, \dots, e_n$  を,  $d(e'(a), e_i(a)) < d(e''(a), e_i(a))$  を満たす例  $e_i$  から構成される例集合  $S_1(e', e'', a)$  とそれ以外  $S_2(e', e'', a)$  に分ける. なお分割の選択基準は, 基準例分割テストと同じである.

以下に, 最良のクラスタ分割テストを求める手続き *clusterExSplit* を示す.

**Procedure:** *clusterExSplit*

**Input:** 例集合  $e_1, e_2, \dots, e_n$

**Return value:** 最良の分割テスト  $\sigma$

```

1   $\sigma.gr = 0$ 
2  Foreach(例のペア  $e', e''$ )
3    Foreach(時系列属性  $a$ )
4      各例  $e_i$  を,  $e', e''$  の内近い方に割り当て, この分割を  $\sigma'$  とおく
5      If  $\sigma'.gr > \sigma.gr$ 
6         $\sigma = \sigma'$ 
7      Else If  $\sigma'.gr == \sigma.gr$  And  $\sigma'.gap > \sigma.gap$ 
8         $\sigma = \sigma'$ 
9    Return  $\sigma$ 

```

## 4. 実験

### 4.1 実験条件

提案手法の有効性を調べるために、実データを用いた実験を行う。慢性肝炎データは、千葉大学医学部附属病院から提供していただいた医療データであり、ECML/PKDD 2002 Discovery Challenge 国際ワークショップでのデータマイニングコンテストでも用いられた [3]。手話データ<sup>(注3)</sup>と EEG データは、データマイニングのベンチマークデータ集合である [5]。

慢性肝炎データにおいては、肝生検<sup>(注4)</sup>を受けて線維化程度が分かっている患者について、肝硬変（線維化程度が F4 あるいは肝生検結果が LC）であるかそれ以外であるかを予測する問題を設定した。患者によって検査回数が大きく異なるため、肝生検の前後各 500 日間において検査回数が 10 以上の時系列データだけを用いた。この結果、肝硬変は 30 例、それ以外のクラスは 34 例となった。検査間隔も大きく異なるため、隣接する 2 点間の線形補間を用いて 10 日おき 101 点の時系列データに変換した。医師のアドバイスにしたがい、重要であることが分かっている 14 属性 (GOT, GPT, ZTT, TTT, T-BIL, I-BIL, D-BIL, T-CHO, TP, ALB, CHE, WBC, PLT, HGB) を用いた。なお、検査値の増減も重要となる場合も考えられるため、各属性を次の手順で変換した属性を追加した、28 属性のデータ集合も用意した。属性  $a$  の時系列データ  $e(a)$  において、最大値と最小値をそれぞれ  $l(e, a)$ ,  $s(e, a)$  とする。変換前後の値をそれぞれ  $e(a, t)$ ,  $e'(a, t)$  とすると、

$$e'(a, t) = e(a, t) - \frac{l(e, a) + s(e, a)}{2}. \quad (3)$$

手話データは、95 種類の単語を 5 人が複数回にわたり発話し、その際の手の位置と動きを記した時系列的データである。本実験では 95 種類の単語の内、Norway, spend, lose, forget, boy の 5 種類をクラスとし、各クラスにつき 70 個の例を用いた。属性は提供されている 15 個中の 9 個 (x, y, z, roll, thumb, fore, index, ring, little) を用い、各時系列データの長さは 50 である。EEG データは、77 人のアルコール中毒患者とそうでない 43 人の患者に関する脳波のデータである。3 パターンの状況に関して、頭皮の 64 箇所から得られた時系列データが記されている。属性数は 192、時系列長は 255 である。

実験では、3. 節で提案した基準例分割テストを用いる時系列決定木とクラスタ例分割テストを用いる時系列決定木を、DTW の窓幅を  $r = 0.1$  とし、枝刈り方式を悲観的枝刈り [10] にして用いた。比較対象として、時系列データの構造を無視する決定木と、時系列データの形を陽に扱う怠惰な学習法を選択した。前者は、前処理で時系列データを平均値で置き換え、その結果出来る数値属性から構成されるデータ集合から決定木を学習する。後者は、DTW に基づく距離基準を用いる最小近傍法である。公平に比較するために、前者では悲観的枝刈りを用い、後者では DTW の窓幅を  $r = 0.1$  とした。最小近傍法の性能は、

用いる距離基準に大きく依存することが知られている。試行錯誤の結果、各時系列属性  $a_i$  における DTW の結果値の最大値  $q(a_i)$  を用いて正規化する距離基準  $H(e_i, e_j)$  が、高正答率を示すので採用された。ただし  $q(a_i) \geq \forall j \forall k G(e(a_j), e(a_k))$  である。

$$H(e_i, e_j) = \sum_{k=1}^m \frac{G(e_i(a_k), e_j(a_k))}{q(a_i)} \quad (4)$$

なお、提案手法においても、属性間の DTW 値を比較するための正規化を用いた手法を用意して実験した。ただし、この場合の正規化は正答率を下げる場合が多かったため、次節の結果からは省いている。

### 4.2 実験結果

評価方法として leave-one-out と、20 回の 5 交差検定を用いた実験結果を、それぞれ表 1, 2 に示す。ただし、サイズは学習された決定木の平均ノード数を表し、時間は熱心な学習 (eager learning) と怠惰な学習を比較する都合上、訓練時間とテスト時間の合計の平均値である<sup>(注5)</sup>。計算機の CPU は Pentium IV 3 GHz、メモリは 1.5G bytes である。肝炎 1, 2 は 4.1 節で述べた、それぞれ 14 属性、28 属性のデータ集合である。

正答率に関しては、基準例分割テストを用いる時系列決定木と最小近傍法が比較的良い結果を示している。これらは、時系列データの形を陽に扱う手法が、実験で用いた学習手法に適しているためだと考えられる。逆に平均値決定木は、時系列データの形を無視するために正答率が高い場合が無く、手話と EEG ではきわめて正答率が低い。最小近傍法は手話データでの正答率が高く、EEG データでの正答率が低い。これは、手話データでは全ての属性がクラス分類に有用であるのに対し、EEG データではクラス分類に関して有用でない属性が多く含まれているためだと考えられる。最小近傍法は、前処理に属性選択が必要であることが確認された。さらに最小近傍法は、2.3 節で述べたように怠惰な学習であるために分類子がなく、時系列決定木に比較して学習結果の可読性が悪い。この欠点は、医療など学習結果の解釈が重要である分野においては致命的であると考えられる。なおクラスタ例分割テストも時系列データの形を陽に扱うが、ほぼ全ての場合において基準例分割テストよりも正答率が低い。これは、クラスタ例分割テストが主に時系列データの形に基づいて例集合を分割するために、異なるクラスに属する例が混在する子ノードを作りやすいためだと考えられる。

実行時間に関しては、時系列データを平均値に変換する決定木学習法が速い。これは、平均値決定木は DTW に基づく時系列間の距離を用いないためである。逆に言えば、他の手法においては DTW に基づく距離を用いることが計算時間が長い場合があることの原因となっている。DTW に基づく距離を用いる手法の中では、クラスタ例分割テストを用いる時系列決定木が、他の手法に比べ極端に計算時間が長い場合がある。これは、前記の理由によりこの分割テストが異なるクラスに属する例が

(注3) : 元版は <http://www.cse.unsw.edu.au/~waleed/tml/data/>

(注4) : 4.3 節で説明する

(注5) : 時間は正確には、全ての例ペア間の DTW を求める時間と leave-one-out の時間あるいは 20 回の 5 交差検定の時間を足し、学習回数で割った値である

表 1 leave-one-out を用いた場合の実験結果  
Table 1 Results of experiments with leave-one-out

手法	正答率 (%)				時間 (s)				サイズ (個)			
	肝炎 1	肝炎 2	手話	EEG	肝炎 1	肝炎 2	手話	EEG	肝炎 1	肝炎 2	手話	EEG
基準例分割	79.7	85.9	86.3	70.0	0.8	1.4	63.3	96.8	9.0	7.1	38.7	16.6
クラスタ例分割	65.6	73.4	85.4	63.3	1.3	1.3	1876.4	1300.5	9.4	7.2	42.8	23.4
平均値決定木	73.4	70.3	35.7	52.5	0.0	0.1	0.2	2.9	10.9	11.4	47.4	61.9
最小近傍法	82.8	84.4	97.8	60.8	0.2	0.4	0.1	47.5	N/A	N/A	N/A	N/A

表 2 20 × 5 交差検定を用いた場合の実験結果  
Table 2 Results of experiments with 20 × 5-fold cross validation

手法	正答率 (%)				時間 (s)				サイズ (個)			
	肝炎 1	肝炎 2	手話	EEG	肝炎 1	肝炎 2	手話	EEG	肝炎 1	肝炎 2	手話	EEG
基準例分割	71.1	75.6	85.9	63.8	0.5	0.8	28.3	51.1	8.3	7.5	28.3	13.4
クラスタ例分割	69.1	69.0	86.2	58.5	0.6	1.5	966.9	530.0	7.9	7.4	36.0	19.9
平均値決定木	73.0	70.7	34.9	51.3	0.0	0.0	0.2	2.5	10.1	10.0	41.1	40.1
最小近傍法	80.9	81.5	96.6	61.8	2.2	4.4	9.3	1021.9	N/A	N/A	N/A	N/A

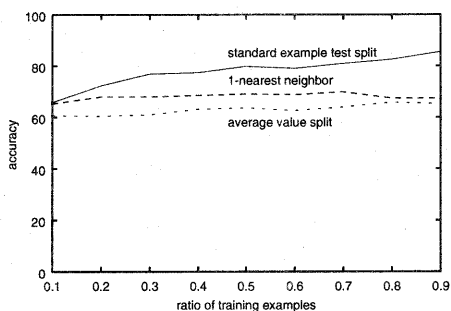


図 4 EEG データにおける各手法の学習曲線

Fig. 4 Learning curve of each method for the EEG data

混在する子ノードを作りやすく、その結果枝刈りを始める際の決定木がきわめて大きくなる場合があるためだと考えられる。

決定木のサイズに関しては、基準例分割テストとクラスタ例分割テストの傾向は、後者がやや大きいものの、似ている。一方平均値決定木は、さらに大きな木を作る。これは、平均値決定木が異なるクラスに属する例が混在する子ノードを作りやすいためだと考えられる。

表 1 と表 2 における計算時間の違いは、主に訓練例数とテスト例数が原因である。最小近傍法は、分類子を構築しない怠惰な学習であるため、実行時間はテスト時間に等しく、テスト例数に比例する。その結果、leave-one-out を用いる方が 20 × 5 交差検定を用いる場合に比較して計算時間が短く、決定木学習法とは逆の傾向を示す。

基準例分割テストは、訓練例が少ない 20 × 5 交差検定の場合に正答率が大きく下がる場合が多いが、これは基準例に選ばれる例が訓練例集合に現れないためであると考えられる。この仮説を検証するために、例数が最も多い EEG データにおける学習曲線を図 4 に示す。ただしクラスタ例分割テストは表 1, 2 の結果が悪いために省略した。図より、基準例分割テストは例

数が小さい場合に正答率が低下する度合いが比較的大きいことが分かる。この実験より、提案手法は極端に例数が少ないデータには用いない方が良いことが分かる。

#### 4.3 慢性肝炎データからの知識発見

慢性肝炎は、ウイルス感染により肝細胞が炎症を起こし傷害される疾患である。炎症が長期に持続した場合、肝硬変と呼ばれる終末像を呈する。肝硬変に至る道程においては線維化と呼ばれる指標が進行度を表し、線維化の全くない F0 から肝硬変を表す F4 の 5 段階がある。線維化は、直接肝臓に器具を挿入し肝組織を採取する肝生検によって診断するが、肝生検は出血などの危険性を伴い短期入院を必要とするため、頻回に検査できない。したがって、この検査と同等以上の線維化指標が、血液検査のような簡便な検査で代替できれば大きな意義がある。

本節の実験で得られる分類子は、この実現につながる可能性がある。われわれは 1 回目の肝生検だけを対象とし、肝硬変 (LC) かそうでない (NLC) をクラスとする、51 例 (内 LC は 21 例、非 LC は 30 例) 14 属性から構成されるデータ集合を得た。このデータ集合を、慢性肝炎データ 0 と呼ぶ。慢性肝炎データ 0 から学習された基準例分割テストを用いる時系列決定木を図 5 に示す。ただし、属性の次の括弧内に記述されているのは患者番号であり、葉は多数クラスを予測するものとする。グラフ中の横点線は、検査値の分類 (normal や high など) が変わる境界値を表す。

学習された時系列決定木とその分割テストで用いられている時系列データは医師らの関心と呼び、重要な発見であると認められた。この時系列決定木は、まず CHE (コリンエステラーゼ) で肝臓の予備能を調べ、予備能が劣っている患者を肝硬変と予測する。それ以外の患者については、GPT (ALT) で炎症の度合を調べ、度合が激しい者を肝硬変と予測する。それ以外の者については、ALB (アルブミン) で肝臓の別の予備能を調べ、優劣で肝硬変でないかそうであることを予測する。この判断基準は医師が日常行っている血液検査の解釈に非常に合致して

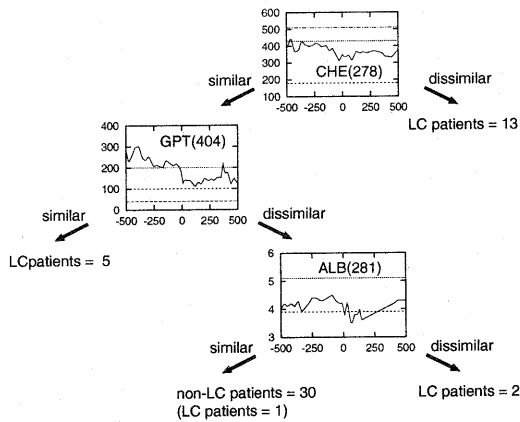


図 5 最初の肝生検を対象とする慢性肝炎データから学習された時系列決定木

Fig. 5 Time-series decision tree learned from chronic hepatitis data of the first biopsies

表 3 慢性肝炎データで leave-one-out を用いた実験結果

Table 3 Results of experiments from the chronic hepatitis data set using leave-one-out

手法	正答率 (%)	時間 (s)	サイズ (個)
基準例分割	88.2	0.5	6.9
クラス例分割	62.7	0.5	8.5
平均値決定木	82.4	0.0	10.1
最小近傍法	74.5	0.1	N/A

いた。肝炎に関与するという属性を設定したこと以外にはほとんど医学知識を使用しないアルゴリズムであるにも関わらず、妥当性の高い結果を出したとの評価を受けた。

慢性肝炎データ 0 に、各種手法を適用した実験結果を表 3 に示す。われわれの決定木は可読性に優れる上に、予測正答率が 88% と最も正確であり、この結果も医師にとって評価された。さらに、leave-one-out で誤予測された患者を検討した所、それらのほとんどが例外的な患者であることが判明した。われわれが提案する基準例分割テストを用いる時系列決定木は、例外的な患者の特定にも有用であることが分かった。

## 5. おわりに

機械学習アルゴリズムを現実のデータに適用する場合、量質形の問題が立ちはだかることが多い [14]。データは大量にあり、質が悪く、複雑な形をしていることが多い。近年、構造データからの分類学習 (例えば [6]) が重要性を増して来ているが、当然の流れであると考えられる。

尤度やベイズ決定などの確率論を用いた自動診断の歴史は古く、ROC 曲線なども古くから用いられている [16]。一方、1990 年代において、医療の情報化はそれまでの文字型データと数値型データから時系列波形や各種画像などのマルチメディアデータへと広がり、種々のメディアを駆使する診断全般の支援を可能としつつある [16]。本研究で提案した時系列データの形を陽

に扱う時系列決定木は、慢性肝炎データにおいて医師が有望と認める結果を出すことに成功した。マルチメディアデータにおける種々の構造を有効に扱う分類学習手法への道のりはまだ遠いと思われるが、本手法はこの目的に向けての重要な一歩であると思われられる。

## 謝 辞

本研究の一部は、文部科学省科学研究費特定領域研究「アクティブマイニング」の援助を受けている。

## 文 献

- [1] H. Almuallim, Y. Akiba, and S. Kaneda: "On Handling Tree-structured Attributes in Decision Tree Learning", *Proc. Twelfth International Conference on Machine Learning*, pp. 12-20, 1995.
- [2] H. Almuallim, Y. Akiba, and S. Kaneda: "An Efficient Algorithm for Finding Optimal Gain-ratio Multiple-split Tests on Hierarchical Attributes in Decision Tree Learning", *Proc. Thirteenth National Conference on Artificial Intelligence*, pp. 703-708, 1996.
- [3] P. Berka: ECML/PKDD 2002 Discovery Challenge, Download Data about Hepatitis, <http://lisp.vse.cz/challenge/ecmlpkdd2002/>, 2002 (current September 28th, 2002).
- [4] L. Breiman, J. Friedman, R. Olshen, and C. Stone: *Classification and Regression Trees*, Chapman and Hall, Belmont, California, 1984.
- [5] S. Hettich and S. D. Bay: The UCI KDD Archive <http://kdd.ics.uci.edu>, Irvine, CA: University of California, Department of Information and Computer Science, 1999.
- [6] H. Kashima and T. Koyanagi: Kernels for Semi-structured Data, *Proc. Nineteenth International Conference on Machine Learning*, pp. 291-298, 2002.
- [7] E. J. Keogh and M. J. Pazzani: "Scaling up Dynamic Time Warping to Massive Dataset", *Principles of Data Mining and Knowledge Discovery, LNAI 1704*, pp. 1-11, 1999.
- [8] E. J. Keogh and M. J. Pazzani: "Scaling up Dynamic Time Warping for Datamining Application", *Proc. Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 285-289, 2000.
- [9] E. J. Keogh: "Mining and Indexing Time Series Data", *Tutorial at The 2001 IEEE International Conference on Data Mining (ICDM)*, <http://www.cs.ucr.edu/~%7Eeamonn/tutorial.on.time.series.ppt>, 2001.
- [10] J. Mingers: "An Empirical Comparison of Pruning Methods for Decision Tree Induction", *Machine Learning*, Vol. 4, pp. 227-243, 1989.
- [11] S. K. Murthy: "Automatic Construction of Decision Trees from Data: A Multi-disciplinary Survey", *Data Mining and Knowledge Discovery*, Vol. 2, No. 4, pp. 345-389, 1998.
- [12] J. R. Quinlan: *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, Calif., 1993.
- [13] H. Sakoe and S. Chiba: "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", *IEEE Transaction on Acoustics, Speech, and Signal Processing*, Vol. ASSP-26, No. 1, pp.43-49, 1978.
- [14] 鈴木英之進:「データマイニング」, エンサイクロペディア情報処理 改訂 4 版, pp. 246-247, 情報処理学会編, オーム社, 2002.
- [15] F. Takechi and E. Suzuki: "Finding an Optimal Gain-ratio Subset-split Test for a Set-valued Attribute in Decision Tree Induction", *Proc. Nineteenth International Conference on Machine Learning*, pp. 618-625, 2002.
- [16] 田中博: 電子カルテと IT 医療, エム・イー振興協会, 2001.