1F5-06

時系列決定木による分類学習

Classification by Time-series Decision Tree

山田 悠[*1] 鈴木 英之進[*2] 横井 英人[*3] 高林 克日己[*3]
Yuu YAMADA Einoshin SUZUKI Hideto YOKOI Katsuhiko TAKABAYASHI

[*1]横浜国立大学大学院工学府
Faculty of Engineering, Graduate School, Yokohama National University

[*2]横浜国立大学大学院工学研究院
Faculty of Engineering, Graduate School, Yokohama National University

[*3]千葉大学医学部附属病院医療情報部
Division of Medical Infomatics, Chiba University Hospital

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.

1. はじめに

決定木 [Quinlan93] 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，および集合値属性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，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 [Keogh01]．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，有効でないと危惧される．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，相違度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す．5. 節は結論である．

連絡先: 山田悠，横浜国立大学大学院工学府物理情報工学専攻，〒 240-8501 横浜市保土ヶ谷区常盤台 79-5，Tel: 045-339-4135，Fax: 045-339-4148，E-mail: yuu@slab.dnj.ynu.ac.jp

2. 時系列データの分類学習

2.1 問題定義

時系列データは，時間の経過とともに記録された計測値を時間順に並べた系列データである．本論文では，計測値を等間隔に補間したデータを扱う．データ集合は，n 個の例から構成され，各例は m 個の属性とクラス属性によって構成される．各属性は l 点から構成される 1 個の時系列データをとる時系列属性である．クラス属性は名目属性であり，属性値をクラスと呼ぶ．

時系列データからの分類学習では，入力データ集合から，例のクラスを予測する分類モデルの導出を目的とする．本論文では，分類モデルとして可読性に優れる決定木を考える．

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[*1]．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 [Keogh99, Keogh00]．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) [Sakoe78] 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 1 に，ユークリッド距離と DTW に基づく相違度を算出する際の，時系列データペアの対応例を示す．

時系列データ $\boldsymbol{A} = a_1, a_2, \ldots, a_I$ と $\boldsymbol{B} = b_1, b_2, \ldots, b_J$ 間の DTW に基づく距離 $G(\boldsymbol{A}, \boldsymbol{B})$ を求めることを考える．$\boldsymbol{A}, \boldsymbol{B}$

[*1] 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．

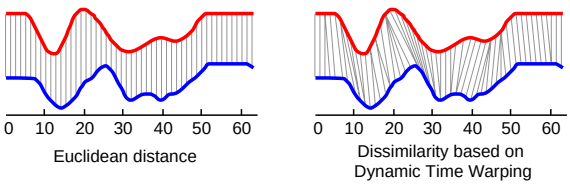

図 1: ユークリッド距離と DTW に基づく相違度を算出する際の，時系列データペアの対応例

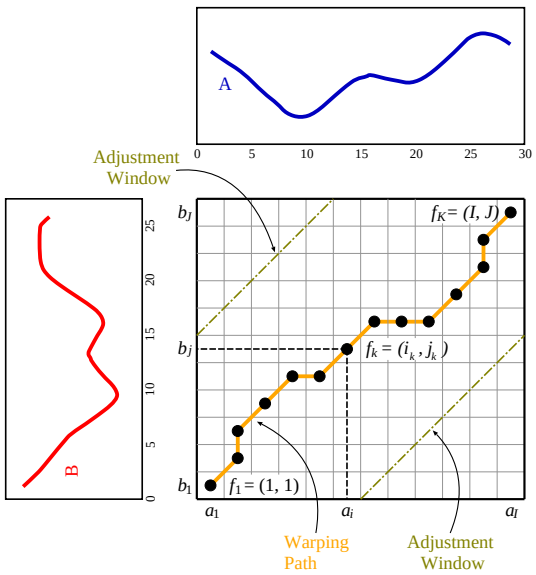

図 2: ワーピングパスの例

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 $f_k = (i_k, j_k)$ の系列で表す．

$$\boldsymbol{F} = f_1, f_2, \ldots, f_K \tag{1}$$

図 2 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．$\boldsymbol{F}$ の評価関数 $\Delta(\boldsymbol{F})$ は式 (2) で表される．

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

$\Delta(\boldsymbol{F})$ の値が小さいほど，$\boldsymbol{A}, \boldsymbol{B}$ 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，$\Delta(\boldsymbol{F})$ を $\boldsymbol{F}$ 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 $\Delta(\boldsymbol{F})$ が加法性を満たすため，$\Delta(\boldsymbol{F})$ を最小化する問題は可能な $\boldsymbol{F}$ 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

時系列決定木は，内部ノードに基準となる属性時系列を持つ決定木であり，基準例分割テストによって例集合を分割していく．基準例分割テスト $\sigma(e, a, \theta)$ は，基準例 e，属性 a，類似閾値 θ から構成され，基準となる時系列データとの相違度に基づいて例集合を分割する．e の a に関する値を $e(a)$ で表すと基準例分割は，例集合 $e_1, e_2, \cdots, e_n$ 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，θ 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を選択した．ただし，θ 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る．$\sigma(e, a, \theta)$ 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 σ.gr と σ.gap は，分割テスト σ のそれぞれ利得比と gap を表す．

Procedure: *standardExSplit*
Input: 例集合 $e_1, e_2, \cdots, e_n$
Return value: 最良の分割テスト σ

```
1   σ.gr = 0
2   Foreach(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)) をキーとして整列し，ε_1, ε_2, ···, ε_n とする
5       Foreach(ε_1, ε_2, ···, ε_n の θ によるギロチンカット分割 σ')
6         If σ'.gr > σ.gr
7           σ = σ'
8         Else If σ'.gr == σ.gr And σ'.gap > σ.gap
9           σ = σ'
10  Return σ
```

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 [Berka02]．手話データ[*2] 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 [Hettich99]．

*2 元版は http://www.cse.unsw.edu.au/~ waleed/tml/data/

慢性肝炎データにおいては，肝生検[*3] 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，検査値の増減が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 箇所の電極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，提案した基準例分割テストを用いる時系列決定木の DTW の窓幅を $r = 0.1$ とし，枝刈り方式を悲観的枝刈り [Mingers89] 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 に基づく相違度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相違度基準 $H(e_i, e_j)$ 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 に基づく相違度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 [*4]．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，ぞれぞれ 14 属性，28 属性のデータ集合である．

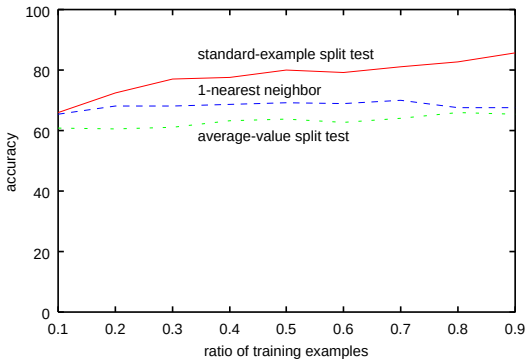

図 3: EEG データにおける各手法の学習曲線

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．

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相違度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 に基づく相違度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．

決定木のサイズに関しては，基準例分割テストよりも平均値決定木の方が大きな木を作る．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，テスト例数にほぼ比例する．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

*3 肝臓の線維化程度を調べるために，直接肝臓に器具を挿入し肝組織を採取する検査．

*4 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

	正答率 (%)				時間 (s)				サイズ (個)			
手法	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
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

	正答率 (%)				時間 (s)				サイズ (個)			
手法	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
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

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 3 に示す．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．

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 [Suzuki02]．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 [Kashima02]) 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．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，時系列データを含むデータ集合において従来の学習法と比べ正確性と可読性の優位性が示された．時系列決定木は，系列データの形を考慮した分類モデルを作成することが可能であり，新しい知識の獲得に大きく貢献できると期待される．

参考文献

[Berka02] P. Berka: ECML/PKDD 2002 Discovery Challenge, Download Data about Hepatitis, *http://lisp.vse.cz/challenge/ ecmlpkdd2002/*, 2002 (current September 28th, 2002).

[Hettich99] 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.

[Kashima02] H. Kashima and T. Koyanagi: Kernels for Semi-structured Data, *Proc. Nineteenth International Conference on Machine Learning*, pp. 291–298, 2002.

[Keogh99] 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.

[Keogh00] 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.

[Keogh01] 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.

[Mingers89] J. Mingers: "An Empirical Comparison of Pruning Methods for Decision Tree Induction", *Machine Learning*, Vol. 4, pp. 227–243, 1989.

[Quinlan93] J. R. Quinlan: *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, Calif., 1993.

[Sakoe78] 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.

[Suzuki02] 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.