平成 15 年度　卒業論文概要

入学年度	学生番号	氏名
平成 12 年度	12117676	柴田 智幸

卒業研究題目　時空間データを対象とする基準例分割による分類学習

和田・犬塚　研究室

1. はじめに

クラス分類問題とは, クラスが未知の事例に対しクラスを予測して分類する問題である．分類するための分類子を生成するには, クラスが既知の訓練事例から学習をおこなう．この分類子を生成するための学習方法として決定木学習などがある．

決定木学習は, 内部ノードで分割テストをおこない葉でクラスを予測する木構造の分類子である．出力される決定木という分類子は人間にも理解しやすく, 大規模データにも適用できるアルゴリズムが多数あることから, データマイニングの手法としても利用されている．

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る．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．時系列データを扱う既存の手法として, 時系列決定木 [1] がある．これは, 木のノードに一つの事例への近さによる分割テストを配置するものである．

本論文では, 前述した時系列決定木の扱えなかった時空間データすなわち, 多次元データの時系列データや時系列データと非時系列データを複合したデータ集合を扱うことを可能にした分類モデルの導出を提案する．また, そのデータ集合を扱うことによる可読性の向上についても考える．

2. データマイニング

データマイニングは, 大規模なデータの中に潜んでいる価値ある情報を掘り出すことを目的としたデータ処理技術である．

また, データは多くの場合, 大規模であり, 質が悪く, 複雑な形をしているため, 前処理の段階で解析できるようにデータクレンジングをすることは重要である．

3. 時系列決定木 [1]

動的時間伸縮法 (DTW)

時系列のデータのペアに関する相違度計算法である．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．このため動的時間伸縮法は, 計測値数が異なる時系列データのペアにも適用できる上に, 結果がより人間の直観に合致することがわかる．

基準例分割テスト

基準例分割テストを $\sigma(ex, a, \theta)$ と表現する．ここで,ex は一つの事例であり基準例と呼ぶ,a は属性,θ は閾値である．事例 ex の属性 a に関する時系列データを $ex(a)$ で表すとすると基準例分割は, 事例集合 $\{ex_1, ex_2, \ldots, ex_n\}$ を, $S(ex(a), ex_i(a)) < \theta$ を満たす事例 e_i から構成される事例集合 $E_1(ex, a, \theta)$ とそれ以外の事例集合 $E_2(ex, a, \theta)$ に分ける．ここで S は DTW による距離である．

[1] の手法は基準例分割テストを木のノードに配置した決定木を与えている．ここで, 分割の選択基準は利得比基準を選択している．

4. 提案手法

本論文では次の三つのことを提案する．

(1). 時空間データにおける基準例分割テスト

時空間データとは, 各次元ごとになっている時系列データを一つの多次元空間を表す時系列データとして扱うデータである．この時空間データを一つの時空間属性として, 基準例分割テストで用いる．この属性の相違度の算出は, 各時間毎に示す各次元の値を用い, 各次元毎でユークリッド距離を求め, その距離の合計を相違度とする．

(2). 非時系列データを複合するデータにおける基準例分割テスト

非時系列データとは, 一つの実数値をもつ数値属性のデータである．この非時系列データに対しても基準例分割テストを用いた．非時系列データの相違度の算出は, 非時系列属性における互いの値の差の絶対値によって算出される．また最良の基準例分割テストを求めるとき非時系列属性は, 次のように扱うことにした．時系列属性の中で最大の利得比を持つ属性を求める．その後, 非時系列属性の中で最大の利得比の属性を求める．求めた両属性の利得比を比較しより大きい方の属性を選択する．非時系列データに対しても基準例分割テストを用いた．

(3). データクレンジング

データ集合を分類する前に, 通常データクレンジングをおこなう．このデータクレンジングの際に, 提案手法では線形補間をおこなうことを提案する．線形補間とは, 異なる 2 点の間の値を求める方法である．

5. 実装と実験結果

提案手法を実装し, 実験をおこなった．用いたデータ集合は, データマイニングのベンチマークデータ集合の手話データである．手話データは, 非時系列データを含んでいなかったので, 非時系列データを扱えることを示すため各属性の時系列データの平均値を求め, 属性とした．

評価方法として,*leave-one-out* 法と 20 回の 5 交差検定を用いた．提案手法の結果に時系列決定木 [1] の結果を追加したものを表 1 に示す．また, サイズは決定木の内部ノード数の平均である．

表 1:実験結果

手法（入力データ）	*leave-one-out* 法		20 回の 5 交差検定法	
	正答率 (%)	サイズ (個)	正答率 (%)	サイズ (個)
既存の手法 (時系列)[1]	86.3	38.7	85.9	28.3
提案手法 (時空間)	84.0	24.0	83.5	20.8
提案手法 (+非時系列)	82.3	24.1	61.7	22.0

既存の手法と比較して, 決定木のサイズが小さくなり知識として分かりやすい表現になっている．5 交差検定より,*leave-one-out* 法の結果の方が正答率が高いのは,*leave-one-out* 法の方が訓練事例が多いため基準例に選ばれる事例が訓練集合にあるためだと考えられる．

6. まとめ

本論文では, 時系列決定木では扱えなかったデータを扱えるアルゴリズムを提案し, 実際に実装をおこない扱えることを実験で示した．提案手法を用いることで, 従来法より決定木のサイズを小さくし可読性を向上させることができることも示した．また時空間データを扱えることで, より系列データの形を考慮した分類モデルの生成をおこなうことが可能であり, 新しい知識の発見に貢献できると期待される．

今後の課題として, 導出した決定木の内部ノードにある基準例の属性が, 時空間データで三次元以上の空間の場合の表現の方法を考える必要がある．また, 実験を手話データのみでしかおこなっていないため, 様々な時系列データ集合に適用し検証する必要がある．

参考文献

[1] 山田 悠, 鈴木 英之進, 横井 英人, 高林 克日己,: 時系列決定木による分類学習. The 17th Annual Conference of the Japanese Society for Artificial Intellegence,2003