時系列データベースにおける特徴パタンの抽出

Feature Pattern Extraction from Time-series Database

杉村博 Hiroshi SUGIMURA 松本一教

Kazunori MATSUMOTO

神奈川工科大学 工学研究科 情報工学専攻

Course of Information and Computer Sciences, Graduate School of Engineering, Kanagawa Institute of Technology

This paper proposes a classification method based on feature patterns of time series data. Typically, the most common form of this type of data visualization is the line chart, which uses position encodings for both time and value. Many similar data are stored into database by many observers or sensors. The discovery of knowledge for classification on such data is important. We describe two methods; extraction method for feature patterns from each time series data and classification method based on extracted feature patterns. In order to measure importance of extracted patterns, we develop a method that applies TF*IDF weight that is often used in text mining to time series data. Extracted feature patterns are used as attributes for machine learning. We measure classification accuracy by using real medical data in experiment.

1. はじめに

時系列データは時間にしたがって記録された数値のシーケンスであり、このようなデータは様々な分野で頻出する. 代表的には株価、医療のセンサデータ、視聴率などがあり、本研究はこのようなデータに焦点をあわせ、クラス分類のための知識を抽出することを目的としている.

時系列データをクラス分類する技術としてサポートベクタマシン (SVM) を用いた手法 [Kampouraki 09] があげられるが、SVM によって作成された分類器は人間にとっては理解が困難で、このようなモデルは、たとえば医療データをもとにして患者に説明する場合に扱いにくい情報となる.

論文 [杉山 08] では株価データに存在する複数のテクニカル指標を用いて、各時系列データをクラスタリングすることでそれらの代表的なパタンを見つけ出し、代表パタンを用いた決定木学習方法によって未来予測する手法を提案している。この方法は1つの株価銘柄に対して未来状態を予測するための知識を抽出できるが、複数の時系列データの違いや特徴を示すパタンを抽出できない。

そこで、本論文では複数の時系列データが存在するデータベースを用いて、これらの時系列データの特徴を表すパタンを抽出する方法について述べる。また、この方法によって獲得した特徴パタンを利用する1つの方法についても述べる。

2. 特徴パタン抽出

1つの時系列データに着目すれば、単純に頻出する部分時系 列データの形は重要であるが、データベース内の多くのデータ に頻出する場合に、その形はある1つの時系列データの特徴 と呼ぶにはふさわしくない。そこで本論文では、このような複 数のデータが存在するデータベースから、ある1つのデータ の特徴といえる時系列データの形を抽出する。本論文ではこの ような形を特徴パタンと呼ぶ。

図1は特徴パタン抽出の概要を示す. データベースに格納された時系列データからスライドウィンドウによって部分時系

連絡先: 杉村博, 神奈川工科大学 工学研究科 情報工学専攻, 神奈川県 厚木市 下荻野 1030, 046-291-3199, 046-291-3199, hiroshi.sugimura@gmail.com

列データを切り出す. 切り出された全ての部分時系列データを クラスタリングする. このようにして各時系列データはクラス タの系列となる. 各時系列データからその時系列データで頻出 し,他の時系列データでは非頻出であるクラスタを特徴パタン として抽出する.

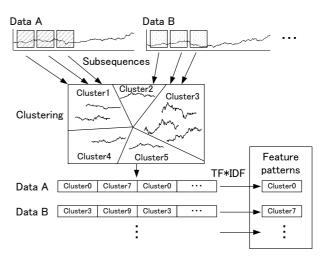


図 1: 特徴パタンの抽出

2.1 クラスタリング

切り出された部分時系列データから、クラスタリング用いて代表的な部分時系列データを求める。クラスタリングのために、本研究では k-means を用いる。k-means は次の目的関数を最小化する分割最適化クラスタリングの代表的な手法である。

$$\operatorname{Err}(\{X_i\}) = \sum_{i}^{k} \sum_{\mathbf{x} \in X_i} D(\mathbf{x}, \bar{\mathbf{x}}_i)$$

ただし、データ集合 X はベクトルで表現されたデータ \mathbf{x} の集合、クラスタ X_i はデータ集合の網羅的で互いに素な部分集

合, $\bar{\mathbf{x}}_i$ は X_i 中のセントロイド, $D(\mathbf{x}, \bar{\mathbf{x}}_i)$ はベクトル \mathbf{x} と $\bar{\mathbf{x}}_i$ の距離を計算する関数である.

一般に k-means では、ベクトル \mathbf{x} と \mathbf{x}_i の距離をユークリッド距離によって計算する。しかし、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、人間の直感に反する結果を生じてしまう場合がある [山田 03]. 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。そこで、システムは時系列データの距離計算のために Dynamic Time Warping (DTW) を使う.

2.2 DTW

DTW 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。時間軸のずれのコストを q、値の不一致のコストを s としたとき、距離 g(i,j) の計算式は次のようになる。

$$g(i,j) = \min\{g(i,j-1) + q, g(i-1,j) + q, g(i-1,j-1) + s\}$$

2.3 TF*IDF

クラスタリングによって求められた代表的な部分時系列データから、TF*IDF を用いて特徴パタンを抽出する. TF*IDF は、テキストマイニングの分野でよく文書中の特徴的な単語を抽出するために用いられるアルゴリズムである. TF*IDF は次の式によって計算できる.

$$TF * IDF(w_i, t_j) = tf(w_i, t_k) \times \log \frac{N}{n}$$

このとき、単語頻度 $TF(w_i,t_k)$ は文書 t_k に頻出する単語 w_i の総数を計算する関数、N は文書の総数、そして n は w_i が出現する文書の総数である。本論文では、1 つの時系列データを文書とみなし、部分時系列データの形を単語としてみなしてこの手法を適用する。

3. クラス分類

本論文では抽出したパタンを用いて、時系列データを分類する。トレーニングデータのインスタンスは1つの時系列データに対する各特徴パタンとの相違度とクラスの組とする。時系列データと特徴パタンとの相違度はDTWによって計算する。このようにして作成したトレーニングデータの例を図2に示す。

Data	P_0	P ₁	P_2	Class
S ₀	50.2	102.4	49.0	C_1
S ₁ //www.	53.8	123.3	352.7	C_2
S ₂ ~~	221.2	88.6	133.7	C_2
S ₃ ~~~~~	300.5	72.8	104.0	C_3

図 2: トレーニングデータ

このトレーニングデータから決定木学習を用いて分類器を作成する.決定木学習は決定木という木構造のクラス分類のためのモデルを作成するアルゴリズムで、トレーニングデータを分割するための枝と分割後のクラスである葉をもつモデルを生成する.一定の基準を満たすまで再帰的に行い、クラスを予測する知識を木構造によって表現する.

4. 実験

使用したデータは PKDD-2005 Workshop on Discovery Challenge で使われたものと同じ千葉大学付属病院から提供 された慢性肝炎の検査データである. 本研究では血小板 (PLT) の時系列データのみを用いて, B型肝炎と C型肝炎の特徴パ タンを抽出し、その特徴パタンを用いて分類を行う. C型肝 炎については治療のためにインターフェロン (IFN) を投与し た患者がいるので、C型肝炎 with IFN と C型肝炎 without IFN の 2 クラスを用意し、合計 3 クラスに分類した. 部分時 系列データは幅20のスライドウィンドウによって抽出し、そ れに満たないデータは削除した. この部分時系列データをク ラスタ数を変えてクラスタリングして特徴パタンを抽出する. 決定木アルゴリズムには C4.5 を使い, 決定木の精度は 10 分 割交差検定によって求めた. 予測精度の比較のために、スライ ドウィンドウで獲得した時系列データを基にして他の機械学 習を使った場合についても実験した. 比較にはニューラルネッ トワーク (NN), サポートベクタマシン (SVM) を用いている. 表1に結果を示す.

表 1: 実験結果

	特徴抽出数	抽出時間 (hms)	精度 (%)
cluster 10	6	4h1m	57.59
cluster 20	11	7h58m	64.04
cluster 30	19	12h18m	77.83
NN	_	35m53s	54.48
SVM	_	3 s	54.37

5. おわりに

本論文では時系列データの特徴パタンを抽出し、その特徴パタンを用いて分類する手法を提案した。抽出する特徴パタンは1つの時系列データで頻出し、データベース全体では非頻出である部分時系列データの形のことである。本手法は従来のNNやSVMとくらべて特徴抽出に時間がかかるが、高い予測精度を手に入れることができる手法として活躍できると考えられる。また、この実験によって獲得した特徴パタンや決定木を専門家によって評価する予定である。

参考文献

[Kampouraki 09] Kampouraki, A., Manis, G., and Nikou, C.: Heartbeat time series classification with support vector machines, *IEEE transactions on information technology in biomedicine*, Vol. 13, No. 4, pp. 512–518 (2009)

[杉山 08] 杉山 喜昭, 平林 悟, 阿部 秀尚, 山口 高平: 時系列パターン抽出に基づく個人投資家意思決定支援システムの実現, 第22回 人工知能学会 全国大会論文集 (2008)

[山田 03] 山田 悠, 鈴木 英之進, 横井 英人, 高林 克日己: 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, *IPSJ SIG Notes. ICS*, Vol. 2003, No. 30, pp. 141-146 (2003)