



合,  $\bar{x}_i$  は  $X_i$  中のセントロイド,  $D(\mathbf{x}, \bar{x}_i)$  はベクトル  $\mathbf{x}$  と  $\bar{x}_i$  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  $\bar{x}_i$  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 を用いて特徴パターンを抽出する. 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.

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

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ーンを用いて, 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ー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ーンとの相違度は 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	$C_1$
$S_1$	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附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 型肝炎の特徴パターンを抽出し, その特徴パター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 クラスを用い, 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ーンを抽出する. 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ーンを抽出し, その特徴パターンを用いて分類する手法を提案した. 抽出する特徴パターンは 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ー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)