Segmental Gaussian Models による時系列パターン発見

今原 修一郎 † 佐藤 誠 ‡ 仲瀬 明彦 §
（株）東芝 研究開発センター † （株）東芝 研究開発センター ‡ （株）東芝 研究開発センター §

1. はじめに

時系列データ中から特徴的な波形（例えば高頻度で発生する波形）を時系列パターンとして自動的に発見する時系列パターン発見 [3] [4] は得られた時系列パターンないしはその特徴を利用することでパターン識別やデータ分析に使用できる。時系列データに全く同一の波形が含まれていることはまれであるため、時系列パターン発見では類似している波形の検出に時間方向や高さ方向の伸縮を考慮した波形のマッチングを行わなければならないという問題がある。この問題を解決するために提案されたマッチング方法として時系列パターンに 1 対 1 対応した確率モデル Deformable Markov Models [1] を使用する方法があり、柔軟なマッチングが可能になる。しかしこの手法は関数近似を行っているために元の波形を損なってしまうという問題点や、時系列クラスタリングに使用するには多くの計算量が必要という問題点がある。

本論文では、上記の問題点を解決するために Segmental Gaussian Models を提案し、この確率モデルを時系列パターンの保持に使用した時系列パターン発見方法を提案する。提案手法を株価データや生体センサデータに適用してその有効性の評価を行う。

2. Deformable Markov Models と Segmental Gaussian Models

Deformable Markov Models は与えられた時系列パターンを隠れマルコフモデル (HMM) のパラメータとして保持し、その尤度関数で波形との類似度を計算する。HMM の各状態を回帰関数 $f_i(\theta_i, t)$ が通用する区間であるとして持続時間 d_i と回帰変数 θ_i を推定している。この区間内では、時間方向や高さ方向の伸縮は考慮される。時系列データの区間分け（セグメンテーション）を明に含んでいるため、同一パターンだが異なるセグメンテーションであるような場合にも対応できる。しかし、回帰関数で近似を行っているために元の波形が複雑である場合にはその形状を損なってしまう。また、各状態の持続時間 d_i を推定しているために時系列クラスタリングとして使用するには多くの計算量が必要である。

この点を解決したものが Segmental Gaussian Models である。Deformable Markov Models と比較し、代表形状をモデルパラメータに含めたために複雑な形状を直接的に表現可能であること、持続時間 d_i の推定をせずに [2] 等のセグメンテーション手法で得た区間長を採用することで高速化したこと、モデル構造として単純な Gaussian Models を採用したために同一パターンだが異なるセグメンテーションという場合に対応できなくなったこと、の 3 点を特徴とする。次節ではこの Segmental Gaussian Models を時系列パターンの保持方法として用いた時系列パターン発見方法を構成する。

Time Series Pattern Discovery by Segmental Gaussian Models

†TOSHIBA Corp., IMAHARA Shuichiro
‡TOSHIBA Corp., SATO Makoto
§TOSHIBA Corp., NAKASE Akihiko

3. Segmental Gaussian Models による時系列パターン発見方法

提案する時系列パターン発見方法の概略は次の通りである。時系列データを [2] 等の手法でセグメンテーションし、分割した区間が L 個連続したもの全てを Segmental Gaussian Models で保持する。この確率モデルをクラスタとみなしてボトムアップ型のクラスタリングを行い、最終的に得られた確率モデルの代表波形を融合回数順に時系列パターンとして出力する。（図 1）

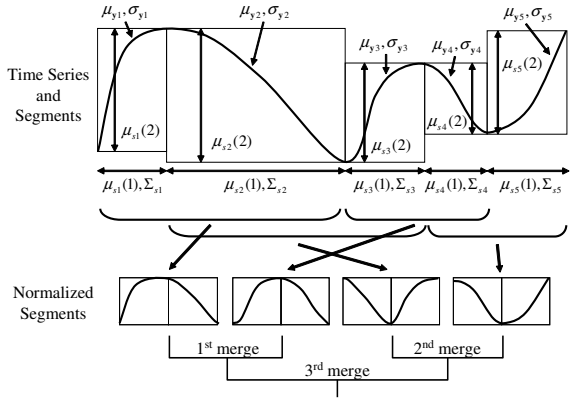

図 1: 時系列パターン発見アルゴリズム概要 (全体の区間数:5、時系列パターンの区間数:2)

3.1 Segmental Gaussian Model

セグメンテーションによって時系列を区間に分け、連続した L 区間を取り出して尤度 $\prod_{i=1}^{L} Q_i(\boldsymbol{s}_i, \boldsymbol{y}_i)$ を計算する。この尤度はスケールに関する確率 $p_{\boldsymbol{s}_i}(\boldsymbol{s}_i)$ と形状に関する確率 $p_{\boldsymbol{y}_i}(\boldsymbol{y}_i|\boldsymbol{s}_i)$ から構成される。

$$\prod_{i=1}^{L} Q_i(\boldsymbol{s}_i, \boldsymbol{y}_i) = \prod_{i=1}^{L} p_{\boldsymbol{s}_i}(\boldsymbol{s}_i) p_{\boldsymbol{y}_i}(\boldsymbol{y}_i|\boldsymbol{s}_i)$$

$$p_{\boldsymbol{s}_i}(\boldsymbol{s}_i) = (2\pi)^{-\frac{d_s}{2}} |\Sigma_{\boldsymbol{s}_i}|^{-\frac{1}{2}} \exp(-\frac{1}{2} (\boldsymbol{s}_i - \boldsymbol{\mu}_{\boldsymbol{s}_i})^T \Sigma_{\boldsymbol{s}_i}^{-1} (\boldsymbol{s}_i - \boldsymbol{\mu}_{\boldsymbol{s}_i}))$$

$$p_{\boldsymbol{y}_i}(\boldsymbol{y}_i|\boldsymbol{s}_i) = (2\pi\sigma_{\tilde{\boldsymbol{y}}_i}^2)^{-\frac{d_y}{2}} \exp(-\frac{1}{2\sigma_{\tilde{\boldsymbol{y}}_i}^2} (\boldsymbol{f}(\boldsymbol{y}_i, \boldsymbol{s}_i) - \boldsymbol{\mu}_{\boldsymbol{y}_i})^T (\boldsymbol{f}(\boldsymbol{y}_i, \boldsymbol{s}_i) - \boldsymbol{\mu}_{\boldsymbol{y}_i}))$$

$\boldsymbol{s}_i$ の次元数は d_s、$\tilde{\boldsymbol{y}}_i$ の次元数は d_y である。d_s は長さ、高さのみを考えるので通常は 2 となる。スケールに関するパラメータは長さと高さ (の差分) を組にした

平均ベクトル $\boldsymbol{\mu}_{\boldsymbol{s}_i}$ と共分散行列 $\Sigma_{\boldsymbol{s}_i}$ である。形状に関するパラメータはスケール $\boldsymbol{s}_i$ と元の波形 $\boldsymbol{y}_i$ を正規化関数 $\boldsymbol{f}(\boldsymbol{y}_i, \boldsymbol{s}_i)$ を使用して正規化した波形 $\tilde{\boldsymbol{y}}_i$ の平均ベクトル $\boldsymbol{\mu}_{\boldsymbol{y}_i}$ と分散 $\sigma^2_{\boldsymbol{y}_i}$ である。パラメータにバイアスは含まれていないため高さの違いはこのモデルでは区別されない。正規化関数で正規化して形状からスケールに関する影響を取り除き、$\boldsymbol{y}_i$ と $\mu_{\boldsymbol{y}_i}$ の次元数を d_y に正規化する。$\sigma^2_{\boldsymbol{y}_i}$ は対角行列である共分散行列 $\Sigma_{\boldsymbol{y}_i}$ の対角成分である。各区間ごとに独立していると仮定しているので、区間ごとの尤度 $Q_i(\boldsymbol{s}_i, \boldsymbol{y}_i)$ の積を取ると、連続した L 個の区間に対する Segmental Gaussian Model の尤度 $\prod_{i=1}^{L} Q_i(\boldsymbol{s}_i, \boldsymbol{y}_i)$ となる。

3.2 確率モデルの融合

ボトムアップクラスタリングは最も類似するクラスタを融合していく。確率モデル i を構成する元データを確率モデル j に入れた尤度の平均と i と j を入れかえた場合の尤度の平均のうち小さい方を代表値とすると、最も類似する確率モデル i, j はその値が最も大きい組とする。

確率モデル i, j を融合するために構成している元データから再度パラメータを推定する。パラメータ推定式は尤度関数 $\prod_{j=1}^{n_1+n_2} \prod_{i=1}^{L} Q_i(\boldsymbol{s}_i, \boldsymbol{y}_i)$ を最尤推定すれば良いが、得られた式では融合するたびに構成している元データを全て使用するため効率が著しく良くない。そこで融合前のパラメータのみを使用して融合後の確率モデルのパラメータを再推定する次の式を使用する。

$$
\begin{aligned}
\alpha &= n_1/(n_1+n_2), \quad \beta = n_2/(n_1+n_2) \\
\boldsymbol{\mu}_{\boldsymbol{s}} &= \alpha\mu_{\boldsymbol{s}_1} + \beta\mu_{\boldsymbol{s}_2}, \quad \boldsymbol{\mu}_{\boldsymbol{y}} = \alpha\boldsymbol{\mu}_{\boldsymbol{y}_1} + \beta\boldsymbol{\mu}_{\boldsymbol{y}_2} \\
\Sigma_{\boldsymbol{s}} &= \alpha(\Sigma_{\boldsymbol{s}_1} + (\mu_{\boldsymbol{s}} - \mu_{\boldsymbol{s}_1})(\mu_{\boldsymbol{s}} - \mu_{\boldsymbol{s}_1})^T) \\
&+ \beta(\Sigma_{\boldsymbol{s}_2} + (\mu_{\boldsymbol{s}} - \mu_{\boldsymbol{s}_2})(\mu_{\boldsymbol{s}} - \mu_{\boldsymbol{s}_2})^T) \\
\sigma^2_{\boldsymbol{y}} &= \alpha(\sigma^2_{\boldsymbol{y}_1} + (\boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{\mu}_{\boldsymbol{y}_1})^T(\boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{\mu}_{\boldsymbol{y}_1})/d_{\boldsymbol{y}}) \\
&+ \beta(\sigma^2_{\boldsymbol{y}_2} + (\boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{\mu}_{\boldsymbol{y}_2})^T(\boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{\mu}_{\boldsymbol{y}_2})/d_{\boldsymbol{y}})
\end{aligned}
$$

添え字に $1, 2$ があれば融合前、添え字に $1, 2$ がなければ融合後のパラメータを表し、n は元データ数を表す。

4. 実験

1983/1/4〜2003/8/25 の日足株価データの 10 点移動平均、及び、腕時計型生体センサの加速度データを用いて本手法の有効性を検証した。

本実験では、融合回数が 3 以上の確率モデルの数（小さすぎると性能が極端に悪化）、尤度関数に自分自身を構成している元データを入れた時の尤度の平均と標準偏差（前者が大、後者が小で良好）、確率モデルの代表パターンとその元データのグラフ（ばらつき小で良好）の 3 種類の評価方法を使用して手法の性能を測定する。

4.1 形状の有無の評価

形状の有効性を確認するため $\boldsymbol{\mu}_{\boldsymbol{y}}$ に元の形状を入れた場合と直線を入れた場合について株価データで実験した。

表 1 に結果を示す。形状無しの場合は形状有りの場合と比較して標準偏差が大きくなっている。微妙な形状を区別できないため、集まる波形のばらつきが大きいからだと思われる。これは図 2 からも理解できる。

形状	パターン数	平均	標準偏差	最大	最小
有	30	31.4	4.96	37.5	22.2
無	21	34.6	13.7	48.2	5.53

表 1: 形状の効果 (株価, 区間数 1, 融合 200 回)

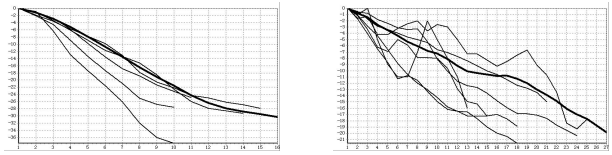

図 2: 代表パターン (太線) と元の波形 (株価, 区間数 1, 融合 200 回)、(左) 形状有り、(右) 形状無し

4.2 他種データによる有効性の評価

本手法が株価データに特化したアルゴリズムでない事を示すため、他の種類のデータで実験を行った。図 3 より、類似波形が集まっていることが分かる。

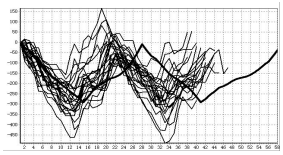

図 3: 代表パターン (太線) と元の波形 (生体センサ, 区間数 4, 融合 500 回)

5. まとめ

時系列パターンの保持方法として、Segmental Gaussian Models という確率モデルを提案した。この確率モデルの特徴は、元の波形をモデルパラメータにしたことで波形が複雑な場合でもマッチングできること、Deformable Markov Model ベースのために時間方向と高さ方向の伸縮を考慮した波形のマッチングが可能であること、同一パターンが異なるセグメンテーションを行った場合の処理を省略することで計算量を低くしていること、省略した状況に非対応になったこと、の 4 点である。

また、この Segmental Gaussian Models を時系列パターンの保持方法として使用し、時系列パターン発見手法を構成した。この手法は、最終的に得られた確率モデルの代表波形を出現頻度順に時系列パターンとして出力する。これにより、大量の時系列データから高頻度で発生する波形を時系列パターンとして発見し、データ分析やデータ識別を行うことができる。

参考文献

[1] X. Ge and P. Smyth: Deformable Markov model templates for time-series pattern matching. *Proc. KDD-2000*, pp.81-90, 2000.

[2] E. Keogh and P. Smyth: A probabilistic approach to fast pattern matching in time series databases. *Proc. KDD'97*, pp.24-30, 1997.

[3] G. Das, K. Lin, H. Mannila, G. Renganathan, and P. Smyth: Rule discovery from time series. *Proc. KDD'98*, pp.16-22, 1998.

[4] E. Keogh, S. Lonardi, and W. Chiu: Finding surprising patterns in a time series database in linear time and space. *Proc. KDD-2002*, pp 550-556, 2002.