頻出部分文字列に基づく階層型隠れマルコフモデルの構造推定

若林　啓†　三浦　孝夫†

† 法政大学 工学研究科　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †kei.wakabayashi@gs-eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

あらまし　統計的なデータ解析において, 対象の情報源を適切に表現するような確率モデルの構造を, サンプルデータから自動的に推定することは重要な課題である. 本研究では, 階層構造を持つ確率過程モデルである階層型隠れマルコフモデル（Hierarchical Hidden Markov Model, HHMM）の構造を推定する手法を示す. HHMM では, 上位階層の状態が再帰的に HHMM のサブモデルを持つことができる. ここでは, サンプルデータ中で頻出の部分系列に対して順次サブモデルを与えることで, HHMM 全体の構造の推定を行う手法を提案する. また, 実験により, 本手法を用いることで系列データの情報源の分布を適切にモデル化できることを示す.

キーワード　モデル選択, 階層型隠れマルコフモデル, 編集距離, ドットプロット

Topology Estimation for Hierarchical Hidden Markov Models based on Frequent Subsequences

Kei WAKABAYASHI† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †kei.wakabayashi@gs-eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

Abstract　Estimation of topology of probabilistic models provides us with an important technique for many statistical language processing tasks. In this investigation, we propose a new topology estimation method for *Hierarchical Hidden Markov Model* (HHMM) that generalizes Hidden Markov Model (HMM) in a hierarchical manner. HHMM is a stochastic model which has powerful description capability compared to HMM, but it is hard to estimate HHMM topology because we have to give an initial hierarchy structure in advance on which HHMM depends. In this paper we propose a recursive estimation method of HHMM submodels by using frequent similar subsequence sets. We show some experimental results to see the effectiveness of our method.

Key words　Model Selection, Hierarchical Hidden Markov Model, Edit Distance, DotPlot

1. 前　書　き

近年, 計算機資源の充実化に伴い, 大量のサンプルデータの解析による知識抽出技術が様々な分野で注目されている. このような統計的な知識処理技術では，対象の情報源の確率分布を学習するために確率モデルを用いることが多い [6]. 確率モデルの学習は，モデルの自由パラメタをサンプルデータの尤度を最大にするように決定するが，モデルの構造は事前に与えなければならない. 対象の情報源に対して適切な確率モデル構造を与える問題を，モデル選択という [10].

モデル選択では，一般に尤度をモデルの良さの尺度として用いるのは適切ではない. 尤度は学習データについての当てはまりの良さを表すため，パラメタが多く表現能力が高いモデルは，学習データをより精密にモデル化し尤度を大きくすることができる. しかし，有限なサンプルデータである学習データを過度に学習したモデルは，異なるサンプルについての当てはまりを悪化させる. 尤度に基づくモデル構造の評価では，この過学習が本質的な問題となる. 赤池情報量基準（AIC）は，対象の情報源から得られるであろう未知のデータの分布を近似的に推定することで，学習サンプルへの過度な適応を避けることができるモデル評価尺度である [1].

しかし，モデルが非観測の変数を含む場合，AIC を最大化することで適切なモデルが得られない場合があることが知られている [4], [11]. このようなモデルでは，AIC で仮定する，パラメタ空間上での尤度に対する漸近正規性を満たさない特異点が存在するため，AIC を適用できない. このため，AIC 最大化に代わるモデルの構造推定手法が必要とされている.

非観測の変数を含む確率モデルは，多くの人工知能分野で応用されている. 隠れマルコフモデル (Hidden Markov Model; HMM) は，非観測の状態がマルコフ遷移しながら観測値を確率

的に出力する，系列データの確率モデルである [12]．HMM は音声認識や文字認識，DNA 解析，文章解析，系列分類などに広く用いられる [8]．HMM は，並列 HMM や差分学習型 HMM [9] など様々な拡張がされているが，マルコフ性を仮定するため，過去のデータがもつ文脈情報を明示的に与えることができない点が指摘されてきた．Fine らは，HMM で扱うことのできない時刻の離れた状態同士の階層的な構造を扱う，階層型隠れマルコフモデル (Hierarchical Hidden Markov Model; HHMM) を提案している [3]．HHMM は，各状態がそれぞれ HHMM のサブモデルを状態出力として持つことができる．この特性から，HHMM ではひとつの状態が複数の観測値に対応することが可能であり，過去のデータの大域的な特徴を現在の観測値に反映することができる．特に時系列データにおいては，直前の状態しか未来に影響を与えない HMM と比較して大きな表現能力の違いといえる．HHMM は，階層の深さや状態数といった構造をあらかじめ与えることで，サンプルデータの尤度を極大化するパラメタ学習ができる．各サブモデルは親をひとつしか持たないため，サブモデルのトポロジーは木構造で表わされる．Fine らは，文書モデルや手書き文字認識モデルとして，HMM よりも優れた表現能力があることを実験により示している．HHMM は系列データのモデル化としてより優れたモデルとして期待されているが，HMM よりも構造が複雑でパラメタの数も多く，対象の情報源に適切な HHMM 構造を推定するのが困難という問題がある．

本研究では，与えられたサンプルデータを用いて，HHMM の構造を推定する手法を提案する．従来のモデル選択手法の多くは，モデル構造の候補を列挙し，その中から何らかの情報量基準を最大にするモデルを選択する．HHMM の構造推定では，このようなモデル選択手法を適用することは２つの理由から困難である．一つは，HHMM が非観測の変数を含むモデルのため，AIC などの従来の情報量基準を適用することができない．もう一つは，HHMM では取り得る構造の数が膨大であり，全ての可能な構造について情報量基準値を求めることが困難である．このことから，候補を列挙することなくモデル構造を推定する新たなアプローチが必要である．

本稿では，サンプルデータに出現する部分系列の頻度を用いてサブモデルを逐次的に決定することで，モデル構造を列挙することなく直接 HHMM の構造を推定する．ここでの基本的なアイデアは，真の HHMM 構造にあるサブモデルが存在するならば，そのサブモデルに特徴的な出力が高い頻度で出現しているという推論である．しかし，特徴的な出力は確率的なゆれを伴う．このため，ここでは系列の編集距離に基づいて系列の類似度を定義し，この類似度に基づいて部分系列集合の頻度を定義する．本稿ではこの定義に基づいて，HHMM の構造推定を，頻出な部分系列集合を発見する問題に帰着させて解決する手法を示す．

2 章では階層型隠れマルコフモデルについて述べる．3 章では本研究で提案する類似部分系列集合発見による HHMM の構造推定手法を説明し，そのアルゴリズムについて述べる．4 章で実験結果を示し,5 章で結びとする．

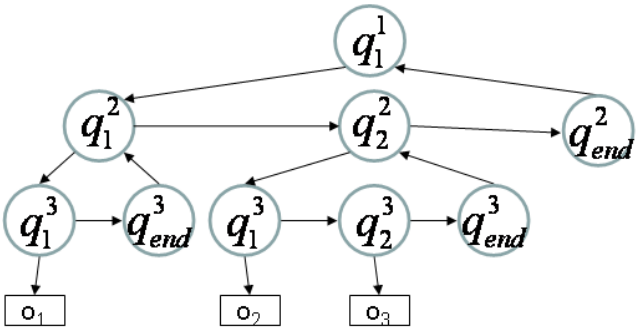

図 1　階層型隠れマルコフモデル

2. 階層型隠れマルコフモデル

階層型隠れマルコフモデル (Hierarchical Hidden Markov Model; HHMM) は，系列の確率分布を与える確率過程モデルである．図 1 に，HHMM の構造を示す．HHMM は状態が階層構造を持っている．ここでは階層 d の状態を q_i^d とする．状態は，観測値を出力する**生成状態**と，サブモデルを出力する**内部状態**に分類される．図 1 では，q_1^1，q_2^1，q_2^2 が内部状態，q_1^3，q_2^3 は生成状態である．階層 d の内部状態 q^d は，出力するサブモデル内の初期状態確率分布 $\pi^{q^d}(q_i^{d+1})$ と，サブモデル内の状態 q_i^{d+1} から状態 q_j^{d+1} の遷移確率分布 $a_{ij}^{q^d}$ をパラメタとして持つ．階層 d の生成状態 q^d は，出力する観測値 k の確率分布 $b^{q^d}(k)$ をパラメタとして持つ．HHMM は，構造とこれらのパラメタによって特徴づけられる．

HHMM では，ひとつの観測値に対して各階層の状態がそれぞれひとつ対応する．例えば，図 1 の観測値 o_2 に対応する状態は，最上位の第１階層が状態 q_1^1，第２階層が状態 q_2^2，第３階層が状態 q_1^3 である．時刻が経過すると，最下位の階層の状態が現在の状態にのみ依存して遷移し，最下位の階層の状態にのみ依存して観測値を出力する．時刻 3 では，第３階層の状態が q_1^3 から q_2^3 に遷移し，観測値 o_3 が状態 q_2^3 にのみ依存した確率分布に従って出力される．HHMM では，第１階層を除き，必ずひとつの**最終状態** q_{end}^d を持つ．最終状態に遷移すると，そのサブモデルは終了し，サブモデルを出力した状態が別の状態に遷移する．ここでは，観測値 o_3 を出力した後，第３階層の状態が q_2^3 から最終状態 q_{end}^3 に遷移し，q_2^2 のサブモデルは終了する．q_2^2 は遷移確率分布に従って別の状態に遷移する．第２階層の最終状態 q_{end}^2 に遷移し，最上位階層のサブモデルが終了した時，その HHMM 全体の出力は終了とする．

HHMM では，以下の３つの問題を扱う.（１）モデル λ と観測値の系列 O が与えられたとき，λ が O を出力する確率を求める問題.（２）モデル λ と観測値の系列 O が与えられたとき，λ が O を出力する際に最も尤度の大きい状態の系列を求める問題.（３）HHMM の構造と観測値の系列 O が与えられたとき，最も O の尤度を大きくするモデル λ のパラメタを求める問題. Fine らは，これらの問題の解となるアルゴリズムを提案している．問題（１）については，動的計画法を用いた**一般化前向き・後向きアルゴリズム**が提案されており，状態の総数を N，与えられた系列の長さを T として，$O(NT^3)$ で確率を求められる

ことを示している．問題（2）については，同様に動的計画法を用いて $O(NT^3)$ で状態系列を求める**一般化** *Viterbi* **アルゴリズム**を提案している．問題（3）については，状態系列の推定とパラメタの推定を交互に繰り返す EM アルゴリズムを用いた解として，**一般化** *Baum-Welch* **アルゴリズム**を提案している．これは HMM の Baum-Welch アルゴリズムと同様，パラメタの局所最適解を与えるアルゴリズムであり，解となるパラメタは与える初期値に依存する．

HHMM は系列の階層構造を扱えるため，HMM と比較して精密なモデル化が可能である．しかし，HHMM はサブモデルの関係を木構造として事前に与える必要があるため，構造の決定が難しいという問題がある．これまでの研究では，人手により構造を決定し，サンプルデータからパラメタの学習を行っていた．人手による構造の決定は，対象の情報源の特性についてよく知っている必要があり，手法の汎用化を困難にしている．本稿では，情報源からのサンプルデータに出現する頻出部分系列に着目することで HHMM の構造を推定する手法について論じる．

3. 頻出部分系列に基づく HHMM 構造推定

本章では，与えられたサンプルデータから HHMM の構造を自動的に推定する手法について述べる．ここでは，サブモデルを再帰的に推定することで HHMM 全体の構造を推定する．本研究の基本的なアイデアとして，サブモデルが存在するならば，サンプルデータ中にはそのサブモデルの出力である部分系列が頻出であると仮定する．しかし，HHMM は確率過程モデルであり，モデルの出力が常に特定の系列であることは通常ない．このため，部分系列の抽出における系列同士の比較では，完全一致に限らず，ある程度のミスマッチを許容する．本研究では，HHMM の構造推定を**頻出部分系列集合**に基づいて行う．HHMM は，再帰的に HHMM のサブモデルをもつため，頻出の部分系列を抽出することはサブモデル構造を推定することに対応する．本節では，まず部分系列集合の頻度を定義し，頻出部分系列集合の発見が容易な問題でないことを論じる．その後，頻出部分系列集合の抽出をドットプロット手法に基づいて効率的に行う手法を示す．

3.1 頻出部分系列集合の定義

本稿では，部分系列の類似度を**編集距離**を用いて定義し，類似度が大きいほど同じ HHMM のサブモデルから出力されやすいと考える．系列 s_1 と s_2 の編集距離は，s_1 に対して，シンボルの挿入，削除，不一致置換，一致の 4 種類の編集操作を行って s_2 に一致させるような編集操作系列に基づいて定義する．例えば，$s_1 = \{ABDAEF\}$，$s_2 = \{BCDBE\}$ のとき，s_1 を s_2 に一致させるためには次の編集操作を適用すればよい．

$$op = \{A \text{ の削除}, B \text{ の一致}, C \text{ の挿入}, D \text{ の一致}, \\ A \text{ を } B \text{ に置換}, E \text{ の一致}, F \text{ の削除 }\}$$

操作系列に基づいて，編集距離は次のように定義する．

$$d(s_1, s_2) = \frac{\sum_i Cost(op_i)}{|op|}$$

ここでは，$Cost(x)$ は編集操作 x が一致操作であればコスト 0，それ以外の操作であればコスト 1 を返す関数とする．$|op|$ は編集操作の数である．最適な編集操作系列は，動的計画法を用いて $O(mn)$(m,n はそれぞれ比較する系列の長さ) で求められることが知られている．ここでは系列の類似度を，編集距離を用いて次のように定義する．

$$sim(s_1, s_2) = 1 - d(s_1, s_2)$$

本稿では，集合内の要素が互いに高い類似度を持つような部分系列の集合を HHMM のサブモデルに対応させる．有用な部分系列の集合は頻出であると考える．ここでは系列集合 S の頻度 $f(S)$ を次のように定義する．

$$f(S) = \sum_{i=1}^{|S|} \prod_{j=1}^{|S|} sim(i, j)$$

$f(S)$ では，各要素は集合内のそれぞれの要素との類似度の積で重み付けされた頻度を持っており，集合全体の頻度を重み付けされた頻度の和によって定義する．集合内の要素が全て同じとき，$f(S)$ は集合の要素数 $|S|$ と一致する．ある閾値 σ について，$f(S) \geqq \sigma$ のとき，部分系列集合 S は頻出であるという．

頻出な部分系列集合を発見することは，計算量の観点から容易な問題ではない．特に，全ての部分系列から，集合として頻出となるような部分系列の組み合わせを求めるのは膨大な計算量となる．ここで，閾値 σ が与えられ，部分系列集合 S が頻出であるとき，$M = max_{i,j} sim(s_i, s_j)$ が満たす条件を考える．このような M を求めれば，全ての要素の組み合わせで類似度が M より大きくなるように集合を与えることで，要素数に対して頻度が単調増加する保証が得られる．S 内の全ての類似度が M だったとき，ちょうど $f(S) = \sigma$ となるような M が，$max_{i,j} sim(s_i, s_j)$ の下限である．このとき，

$$\begin{aligned} f(S) &= \sum_{i=1}^{|S|} \prod_{j=1}^{|S|} M \\ &= \sum_{i=1}^{|S|} M^{|S|-1} \\ &= |S| M^{|S|-1} = \sigma \end{aligned}$$

両辺対数を取って

$$log|S| + (|S| - 1) logM = log\sigma$$

ここで，$|S|$ について微分して 0 とおくと，

$$\begin{aligned} logM + \frac{1}{|S|} &= 0 \\ |S| &= \frac{1}{-logM} \end{aligned}$$

これより，$|S| = \frac{1}{-logM}$ のとき，$f(S) - \sigma$ が最小になる．元の式に代入すると，

$$\frac{1}{-logM} M^{\frac{1}{-logM} - 1} = \sigma$$

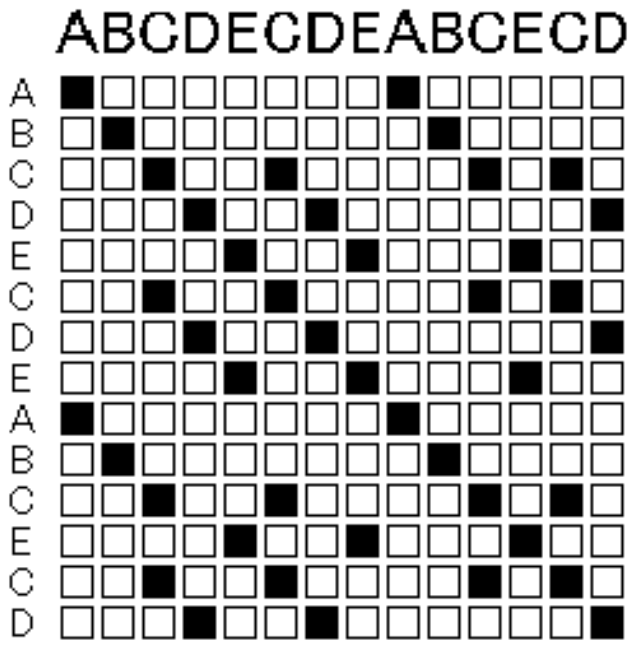

図 2　系列 $\{ABCDECDEABCECD\}$ のドットプロット

を得る．これは解析的に解くのは難しいが，M について 0.0 から 1.0 の間で単調増加であるため，近似解を求めることができる．例えば，$\sigma = 4$ のときは $M = 0.9032$，$\sigma = 3$ のときは $M = 0.8683$ である．

3.2　ドットプロット

本稿では，与えられた系列から，$f(S)$ がある閾値 σ を超えるような部分系列集合を抽出する問題を考える．系列の部分系列を全て列挙し，それぞれの類似度を計算するのは，系列長 n に対して少なくとも $O(n^4)$ の計算量である．この計算量は現実的ではない．本研究では，**ドットプロット**に基づいて頻出部分系列集合を $O(n^2)$ の計算量で近似的に抽出する手法を提案する．

ドットプロットは，系列の類似性を 2 次元平面上のプロットにより視覚的に表したものである．図 2 に，系列 $s = \{ABCDECDEABCECD\}$ のドットプロットを示す．縦軸と横軸には s のシンボルを並べる．図 2 で黒い四角で表した平面上のプロットは，縦軸と横軸のシンボルが一致していることを意味する．縦軸と横軸には同じシンボル列があるため，左上から右下への対角線上には必ずプロットが存在する．また，一致する部分系列の位置では，プロットが左上から右下への線となる．例えば，部分系列 $\{CDE\}$ は，s 中で先頭から 3 番目と 6 番目に出現するが，これは平面上で (6,3)，(3,6) の位置から始まる斜めのラインに対応する．

部分系列 $\{ABC\}$ に対応する (1,9) から始まる斜めのラインと，$\{ECD\}$ に対応する (9,12) から始まるラインは，1 座標のギャップがあるひとつのラインとみなすことができる．これは，s 中で 9 番目から出現する部分系列 $s_{9,14} = \{ABCECD\}$ が，1 番目から出現する部分系列 $s_{1,7} = \{ABCDECD\}$ から，削除の編集操作を一回行った系列であるといえる．本稿では，このとき縦軸に対応する系列 $s_{1,7}$ をパターンと呼び，横軸に対応する系列 $s_{9,14}$ を共起系列と呼ぶ．ドットプロットの平面上でこのラインを検出すれば，1 番目から出現する部分系列は，ラインの開始点と終点の縦軸座標から分かり，9 番目から出現する部分系列は，ラインの開始点と終点の横軸座標から分かる．また，このときの系列類似度 $sim(s_{1,7}, s_{9,14})$ は，

$$\frac{\text{ライン中のプロットの数}}{max(|s_{1,7}|, |s_{9,14}|)}$$

と一致する．この方法で求める類似度は，編集距離において最適な編集操作系列中に挿入操作と削除操作が両方含まれているときに定義と一致しない．しかし，本稿では高速化のため，この方法で得た値を類似度の近似値として用いる．

3.3　頻出部分系列集合の抽出

本手法では，与えられた系列 D について，ドットプロットの平面上を一度走査することで頻出部分系列集合を近似的に求める．

まず，パターン集合 P を空集合とし，行を表す変数を $l = 1$ とする．ドットプロット平面上の l 行目を読み，P にパターン $D_{l,l}$ を加える．全ての l 行目のプロットを，$D_{l,l}$ の共起系列とする．$l \geqq 2$ のときは，全ての i についてパターン $D_{i,l-1}$ をコピーしてパターン $D_{i,l}$ とし，P に加える．全ての l 行目のプロット (x, l) について，パターン $D_{i,l}$ の全ての共起系列と比較する．共起系列の末尾の座標を (a, b) とすると，プロットの座標が $x > a \wedge x \leqq a + (l - b)$ を満たす最小の a をもつプロットを共起系列の末尾に加える．これは，共起系列に対して削除または置換の編集操作のみを行うことを意味する．挿入操作が必要な系列の場合はここでは類似性を検出できないが，ドットプロット平面の対称性により，x 行目の走査で検出できる．この操作を全ての l について実行することで，全ての部分系列パターンを列挙し，その共起系列を抽出することができる．

本手法では，頻出部分系列集合内全ての系列の類似度が，類似度の最大値の下限 M 以上であると仮定する．この仮定により，類似度が M を下回るような共起系列を除去できるため，空間効率を改善することができる．各 l 行目の操作の最後に，全ての i について，パターン $D_{i,l-1}$ の全ての共起系列の類似度を計算し，類似度が M よりも小さい共起系列を除去する．

得られたパターン集合 P から頻出部分系列集合を抽出する．P の要素を系列長でソートし，系列長が長い順に，共起系列集合にパターン系列を加えた部分系列集合 S の頻度 $f(S)$ を求める．この時点では共起系列同士の類似度は計算できないが，ドットプロット上で共起系列がパターンとなっている行が存在するため，それぞれの共起系列を P から探し，共起系列同士の類似度を求める．もし P 内に見つからない組み合わせが存在した場合，実際に当該系列の編集距離を動的計画法により求めて類似度を求める．全ての S の要素同士の類似度が求まったら，$f(S)$ を求め，$f(S) \geqq \sigma$ であれば，S を頻出部分系列集合とする．P から S の要素の系列と重なった系列を含むパターンを除去する．この操作を P の全ての要素で行い，頻出部分系列集合の集合を得る．

抽出した全ての頻出部分系列集合 S について，S の要素の系列を単純に結合して，以上の手法を適用する．このとき，σ は

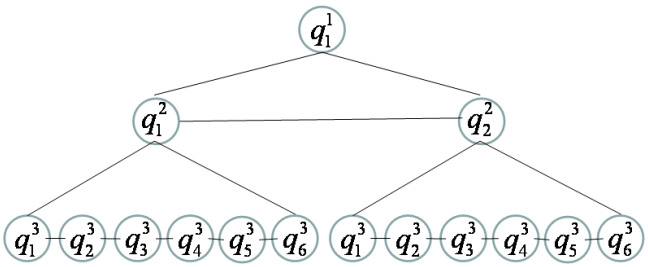

図 3　実験データを生成する HHMM

S の頻度 $f(S)$ とする．ただし，パターン系列および共起系列は結合した境界を越えないものとする．この操作を再帰的に適用することで，階層構造を持つ頻出部分系列集合を抽出する．

最後に，抽出された階層構造を持つ頻出部分系列集合を用いて，HHMM の構造を推定する．ここでは，それぞれの頻出部分系列集合に HHMM のサブモデルを対応させる．このときのサブモデルの状態数は，

$$\frac{\sum_{i=1}^{|S|} |S_i|}{|S|}$$

とする．

4. 実　　験

4.1 実験方法

本稿では，人工データを用いて提案手法の評価を行う．HHMM 構造の抽出に用いる系列データとして，図 3 に示す HHMM を用いて長さ 100 までの系列データを 3 つ乱数により生成する．この HHMM は最大の深さが 3 で，12 個の生成状態と 3 個の内部状態を含んでおり，6 種類のシンボルを出力する．生成状態は 2 つのサブモデルに含まれており，それぞれのサブモデルは $\{0,1,2,3,4,5\}$，$\{3,1,5,0,2,4\}$ という系列を出力する確率が高い．

この HHMM を用いて乱数により生成されたデータを図 5 の上部に示す．実験データの一部のドットプロットを図 4 に示す．

この実験データから，頻出部分系列集合を抽出する．頻度の閾値は $\sigma = 3.0$ とした．

4.2 実験結果

図 5 に，実験データから抽出された頻出部分系列集合を示す．図 5 は，最上部の四角内に実験データの系列を示し，そこから抽出された頻出部分集合を線で結んだ四角内に示している．系列の下線は，抽出された箇所を示す．この結果から，本手法により完全一致ではない系列の集合を頻出部分系列集合として抽出できていることが分かる．データを生成した HHMM の構造と比較すると推定した階層が多いが，最下位の階層ではデータを生成した HHMM と類似した結果が得られている．

図 6 に，抽出した頻出部分系列集合に基づいて推定した HHMM の構造を示す．ここでは，大きく分けて 2 つのサブモデルが抽出されている．これらのサブモデルは真の HHMM と比較して深い階層を持つが，それぞれ真の HHMM の 2 つのサブモデルに対応しているといえる．

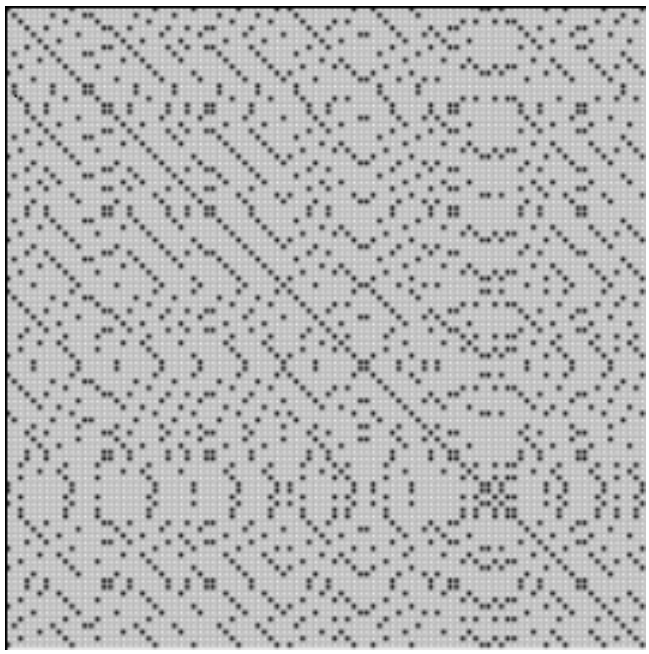

図 4　実験データのドットプロット

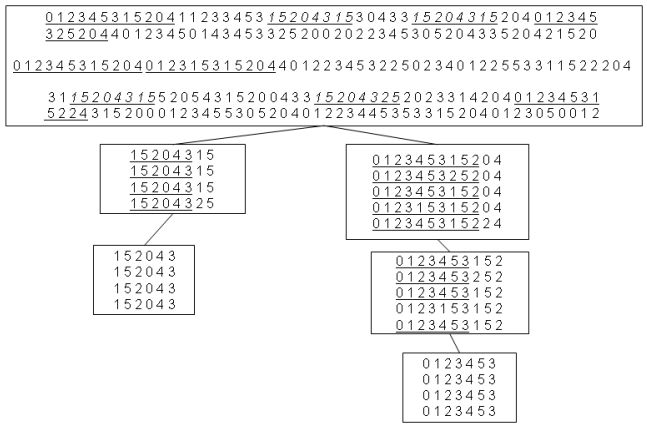

図 5　抽出した頻出部分系列集合

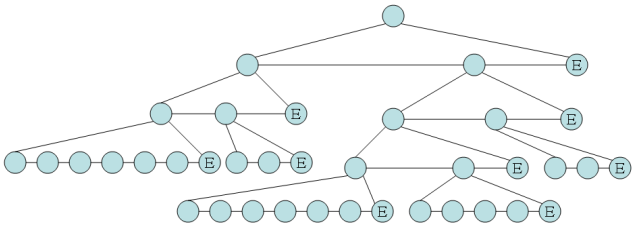

図 6　推定した HHMM 構造

4.3 考　　察

実験結果から，提案手法について考察を行う．まず，抽出された部分系列集合について考察する．集合内の系列を見ると，どの要素も十分に類似した部分系列集合であることが分かる．一方，実験データから抽出されなかったシンボルが多い．これは，頻度を類似度の積に基づいて定義したため，要素数を増やすよりも集合内の類似性を重視したことによる．これにより，十分に類似していない部分系列は抽出されず，構造の推定には用いられない．しかし，本研究では HHMM の構造を抽出する

ことが目的であるため，ここではシンボルの抽出率を上げる必要はなく，正確な頻出部分系列集合を抽出することのほうが重要であると考える．このことから，本実験で抽出されなかったシンボルが多いことは問題ではないと考える．

抽出された部分系列集合の要素は，系列長が全て同じである．これは，長さが異なる系列が集合内にあると，集合全体の頻度を低下させることによる．この特性は HHMM の構造推定としては望ましくない．なぜなら，サブモデルの状態の遷移は一般に一定の回数で終わるわけではなく，出力の系列長は異なることが多いためである．実際に，実験データで抽出されなかった系列の多くは，同じシンボルが複数続くなど長さが異なるものが多くみられる．このことは，編集距離に基づく類似度と，HHMM のサブモデルに所属する確率が，異なる長さの系列に対して大きく異なることに起因する．

推定された構造としては大きく 2 つのサブモデルに分かれている．これは実験データを生成した HHMM と同様の特性であることから，頻出部分系列集合の抽出が隠れた構造を推定する手法として有用であることを示しているといえる．

5. 結　　論

本研究では，頻出部分系列集合の抽出に基づいた系列データの階層型の構造推定手法を提案した．本手法により，階層型隠れマルコフモデルから出力された系列データの構造が推定できることを実験により示した．

本稿では抽出した部分系列集合を直接 HHMM のサブモデルとして決定する手法を示したが，モデル構造の候補を絞ることを目的とすれば，情報量基準によるモデル選択アプローチを HHMM 構造に有効に適用するための手法としての利用も考えられる．

文　　献

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans*, 19(6).

[2] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 994, Washington, DC, USA, 1997. IEEE Computer Society.

[3] S. Fine and Y. Singer. The hierarchical hidden markov model: Analysis and applications. In *MACHINE LEARNING*, pages 41–62, 1998.

[4] H. Katsuyuki. On the problem in model selection of neural network regression in overrealizable scenario. *Neural Computation*, 14(8):1979–2002, 2002.

[5] S. Luhr, H. H. Bui, S. Venkatesh, and G. A. West. Recognition of human activity through hierarchical stochastic learning. *Pervasive Computing and Communications, IEEE International Conference on*, 0:416, 2003.

[6] C. Manning, , C. D. Manning, H. S. Utze, M. The, M. Press, and L. Lee. Foundations of statistical natural language processing, 1999.

[7] C. Vogler and D. Metaxas. Parallel hidden markov models for american sign language recognition. *Computer Vision, IEEE International Conference on*, 1:116, 1999.

[8] K. Wakabayashi and T. Miura. Topics identification based on event sequence using co-occurrence words. In *NLDB '08: Proceedings of the 13th international conference on Natural Language and Information Systems*, pages 219–225, Berlin, Heidelberg, 2008. Springer-Verlag.

[9] K. Wakabayashi and T. Miura. Data stream prediction using incremental hidden markov models. In *DaWaK '09: Proceedings of the 11th International Conference on Data Warehousing and Knowledge Discovery*, pages 63–74, Berlin, Heidelberg, 2009. Springer-Verlag.

[10] 下平英寿, 久保川 達也, 竹内 啓, 伊藤 秀一：統計科学のフロンティア 3 モデル選択，岩波書店，2004.

[11] 甘利俊一, 尾関智子, 朴慧暎, 階層モデルにおける学習と推論-特異構造を持つ統計モデル, 電子情報通信学会論文誌,Vol. J85-DV,No. 5,pp.701-708,2002.

[12] 北 研二: “確率的言語モデル”, 東京大学出版会, 1999.