3B8-2

高次部分状態連鎖モデルサイズ決定手法の検討

Methods to Determine Appropriate Size of HISC Model

大西 智之[*1] フォンベト グエン[*1] 鷲尾 隆[*1]
Noriyuki Ohnishi Nguyen Phuong Viet Takashi Washio

[*1]大阪大学産業科学研究所
The Institute of Scientific and Industrial Research Osaka University

On the trend of massive sensor installation to large scale scientific, engineering and social systems for their fine resolution and sensitive analyses, massive dimensional time series observations consisting of more than thousands of variables have become widely available. Therefore, a HIgh order Substate Chain (HISC) model was proposed to capture the total dynamics of a complex system underlying the data in our laboratory. However, a method to select its appropriate model size has not been developed yet. The proposed approach indicates an appropriate model size based on some infomation criteria named AIC and MDL, and enabled to compose Model for a HIgh-order Substate Chain (HISC) which have high ability to accurately predict next substates. The performance of the proposed approach is demonstrated through the modeling of synthetic and real world data.

1. はじめに

近年，ユビキタスセンシング技術の高度化に伴い，社会インフラ [1] や工学的システム [2] などの様々なシステムに組み込まれているセンサーの数は，急速に増加しつつある．また，科学的計測技術の進歩によって，膨大な変数の大量の時系列データが集められつつある．そのような大規模センシングシステムにおいては，通信コストを節約するため，多数のセンサーが事象駆動的に信号を発信し観測データを出力する．従って，システムが出力するデータは，非常に高次元な変数を含むトランザクション時系列であるものが多くなってきている．しかも，トランザクション時系列データは，多数のセンサーから出力されるノイズを含めた値の組み合わせから成る．このため，高次元かつノイズを含む変数の値の組み合わせ爆発により，観測データ中のトランザクションは，膨大な種類数となる．通常のマルコフ過程モデリングのように，膨大な種類数のトランザクションをそのまま対象状態としてモデル化しようとすると，同じく膨大な状態組み合わせに関する状態遷移規則を同定する必要があり，モデリングの計算量が膨大になると共に，モデルのオーバーフィッティングやノイズによる誤差の増大を排除できない．この状況において，時系列データから統計的モデルを構築し，有用な情報や知識を抽出するために，対象の状態を適切に反映したシステムの動的な状態連鎖モデルを同定することが重要となる．この問題を克服するために，当研究室では，HIgh-order Substate Chain (HISC) モデル [3] を提案した．しかし，HISC モデルの予測精度は，モデルの詳細度によって変化する．モデルが大き過ぎると遷移規則がノイズを含み，データにオーバーフィッティングし，モデルの予測制度が低下する．逆に，小さすぎると将来を妥当に予測するモデルの能力が低くなる．従って，課題として，効率的に適切な HISC モデルの大きさを決定することが必要である．

本研究の目的は上記の課題を克服し，高次元時系列データに関して高次部分状態連鎖モデルを同定する際，その適切なモデルサイズの決定を可能とする手法を検討することである．また，検討した手法についていくつかのデータを用いて有効性検証を行う．

連絡先: 連絡先:大阪大学産業科学研究所
〒 567-0047 大阪不茨木市美穂ヶ丘 8-1
Email:ohnishi@ar.sanken.osaka-u.ac.jp

2. HISC モデル

大規模なシステムにおいて，以下のような，各センサーの名称と出力値の組み合わせをアイテムとする集合のトランザクション X_t で構成されるトランザクション時系列データ D を対象とする．

$$D = \{X_t | t = 1, \cdots, n\}$$
$$\text{ただし, } X_t = \{item_1^t, \cdots, item_{m_t}^t\}$$

はじめに，D からそれを表す従来の高次マルコフ連鎖モデル [4] を同定することを考える．高次マルコフ連鎖モデルは以下の状態遷移確率によって与えられる．

$$P_{\{s_1 \cdots s_\ell\}, s_0} = P(X_t = s_0 | X_{t-1} = s_1, \cdots, X_{t-l} = s_\ell)$$

ここで，ℓ はモデルの次数を表し，M をアイテム集合のラベルを表す要素の有限集合 $s_0, \cdots, s_\ell \in M$ とする．D の高次マルコフ連鎖モデルを同定することは，それぞれのトランザクションを s_i としてラベル付けすることである．しかし，膨大な高次元データの場合，センサーの組み合わせおよび出力値とノイズの多様性により M は爆発し，モデル化することが困難である．更に，各々のトランザクション全体を状態とするため、トランザクションの部分集合がシステムの状態を反映する場合には，その部分状態間の遷移モデルを同定することが難しい．

この問題点に対し，トランザクション全体を状態とせずに，トランザクション時系列データ D から部分状態間の主要遷移規則を抽出し，確率的なモデルを構築する高次部分状態連鎖 (HISC) モデリング手法が当研究室で開発された．以下にその定式化を示す．

$S = S_1 \cdots S_l$ をアイテム集合 $\{S_1, \cdots, S_l\}$ からなるトランザクション時系列とする．そして，S のカウンター χ_s^t を $S_1 \subseteq X_t, \cdots, S_l \subseteq X_{t-l+1}$ の場合は $\chi_s^t = 1$，その他の場合は $\chi_s^t = 0$ とする．ただし，$S_i = \phi$ のときは，$S_i \subseteq X_{t-i+1}$ とする．D に S が表れる頻度 $\chi_s = \sum_t \chi_s^t$ と D に長さ ℓ の S が表れる可能性がある最大頻度:$L_\ell = n - (\ell - 1)$ から，D に表れる S の支持度 (sup) と S の直後に表れるアイテム集合 S_0 の確信度 (conf) を

$$\sup(S) = \frac{\chi_s}{L_\ell},\ conf(S_o | S) = \frac{\chi_{S_0 S}}{\chi_S}$$

と定義する．

最小支持度 (minsup) を設定すると，HISC モデルは，確率は時間不変という仮定のもとで $\sup(S_0 S) \geq \min\sup$ を満たす次の遷移確率で与えられる．

$conf(S_0|S) \approx P(S_0 \subseteq X_t|S_1 \subseteq X_{t-1}, ..., S_\ell \subseteq X_{t-\ell})$
ここで ℓ は現在部分状態への遷移に影響を与える過去状態の時間ステップ数の上限であり，モデル次数と呼ばれる．トランザクション時系列 $S_0S_1\cdots S_\ell$ は頻出なので，その部分系列 $S'_0\cdots S'_{\ell'}$ $(\subseteq S_0\cdots S_\ell)$ も頻出である．ここで，$S'_0\cdots S'_{\ell'} \subseteq S_0\cdots S_\ell$ は，$S'_0 \subseteq S_i \wedge \cdots \wedge S'_{\ell'} \subseteq S_{i+\ell'} (\ell' = 0, \cdots, \ell, i = 0, \cdots \ell - \ell')$ の関係を表している．従って，HISC モデルに前述の遷移確率に対応する状態遷移規則が含まれていれば，遷移確率 $P(S'_0 \subseteq X_t|S'_1 \subseteq X_{t-1}, \cdots, S'_{\ell'} \subseteq X_{t-\ell'})$ に対応する状態遷移規則も含まれる．従って，HISC モデルは，マルコフモデルではなく，従属関係にある部分状態の遷移規則を多く含んでいる．HISC モデルでは，図 1 のような順序木の一種であるトライデータ構造が用いられている．トライデータ構造は、各枝において，それぞれノードが左側から右側に時系列に並んでいる．

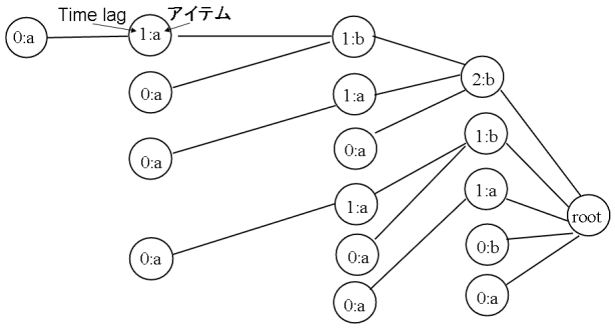

図 1: トライデータ構造

HISC モデルの詳細度は，どれだけ過去の状態の影響を状態遷移に含めるかを表すモデル次数と共に，どれだけの頻度で生起する部分状態やそれらの間の状態遷移規則までをモデルに含めるかを決める最小支持度で決定される．このうち，通常はモデル次数は対象を支配するダイナミクスに関する知見から決定できる場合が多いが，最小支持度はモデルが良好な予測精度をもたらすように決める必要がある．最小支持度を低くすると，ノイズアイテムや主要ではないアイテムを遷移規則に含み，データにオーバーフィッティングするので，モデルの精度が低下する．逆に，最小支持度を高くすると，遷移規則の数が少なくなるので，将来を妥当に予測するモデルの能力が低くなる．現状では，データに関して効率的に適切な HISC モデルの詳細度を決定する手法は未確立である．我々は，遷移規則の数，つまりモデルサイズと将来を予測する能力の両方を評価し，適切なモデルサイズを決定する手法を検討する．

3. モデル評価基準

自己回帰モデルの適切な大きさを評価する指標として，AIC [5] が代表的である．一方，ベイジアンネットワークなど，人工知能の予測モデルのモデル選択規準としては MDL 原理 [6, 7, 8] が用いられる．

3.1 AIC

AIC(Akaike Information Criterion) は，確率モデルにより，どれほど将来を妥当に近似予測できるかを評価する．確率モデルのパラメータ値を最尤推定によって決めた確率分布は予測分布と呼ばれるが，この予測分布が真の分布の近似としてどれほど妥当であるかを測ることにより，真の分布と予測分布の誤差を定義し AIC を導出する．

以下のような，入力変数ベクトル $X_t = (x_{t_1}, \cdots, x_{t_M})^T$ と，出力変数 Y_t で構築される時系列データ D を対象とする．

$$D = \left\{ X_t = (x_{t_1}, \cdots, x_{t_M})^T, Y_t \,|t = 1, \cdots, n \right\}$$

D を支配する真のモデル形式は一般には不明なことが多いが，ここではモデル形式が f であると仮定する．データから推定されるモデルを $\hat{f}$ としたときの，このモデルによる Y_t の予測状態 $\hat{Y}_t$ は，$\hat{Y}_t = \hat{f}(X_t)$ によって与えられる．従って，観測状態 Y_t と予測状態 $\hat{Y}_t$ の予測誤差は，$e_t = Y_t - \hat{Y}_t$ となる．ただし，$e_t; t = 1, 2, \cdots, n$ は平均 0，分散一定とする．予測誤差 e_t の分散の推定値を $\hat{\sigma}^2$ とし，誤差分布をガウス近似すれば，

$$g(Y_t|\hat{f}, \hat{\sigma}^2) = (2\pi)^{-\frac{1}{2}} (\hat{\sigma}^2)^{-\frac{1}{2}} \exp\{-\frac{\left(Y_t - \hat{f}(X_t)\right)^2}{2\hat{\sigma}^2}\}$$

が Y_t の予測分布となる．

複数の予測分布 $g(Y_t\,|f_{\,1}, \sigma_1^2), g(Y_t\,|f_{\,2}, \sigma_2^2), \cdots$ が与えられたとき，最も良いモデルを選択するために，カルバック・ライブラーの情報量を用いると，計算は省略するが，分布 $g\left(Y_t|\hat{f}, \hat{\sigma}^2\right)$ の平均対数尤度

$$E_X[E_y\{\log g(Y_t|\hat{f}, \hat{\sigma}^2)\}] = \int g(Y_t|f, \sigma^2) \log g(Y_t|\hat{f}, \hat{\sigma}^2) dy$$

が大きいモデルを選択すべきことが分かる．この式は n をパラメータ推定に使うトランザクション個数，M をモデルのパラメータ数をすると，

$$E_X[E_y\{\log g(Y_t|\hat{f}, \hat{\sigma}^2)\}] = E_X[\frac{1}{n}\sum_{t=1}^{n} \log g(Y_t|\hat{f}, \hat{\sigma}^2) - \frac{M+1}{n}$$

で表され，これを $-2n$ 倍した量

$$AIC = -2\sum_{t=1}^{n} \log g(Y_t|\hat{f}, \hat{\sigma}^2) + 2(M+1)$$
$$= n \log\left(2\pi\hat{\sigma}^2\right) + n + 2\,(M+1)$$

が AIC である．AIC が最小であるモデルが望ましい．

3.2 MDL 原理

MDL(Minimum Description Length) 原理は，データを最も効率良く圧縮できるかを評価する．記述長は，モデルの記述長とそのモデルとデータの誤差の記述長の和で表される．データを最も短く記述するために，どのようなモデルを選ぶのが良いかをモデル選択の基準とする．単純なモデルを選択すると，モデルの記述長は短くなるが，データとモデルの分布との誤差が大きくなる．逆に，データに合わせたモデルを選択すると，モデルの記述長は大きくなるが，データとモデルの分布の誤差が小さくなる．従って，合計の記述長が最小になるモデルを選択することが重要である．以下に，記述長の式を導出する．

圧縮しようとする時系列データを

$$D = \left\{ X_t = (x_{t_1}, \cdots, x_{t_M})^T \,|t = 1, \cdots, n \right\}$$

とする．あるモデルパラメータ $\hat{\theta}$ を与えたときの，D を表す確率密度関数を $f\left(D; \hat{\theta}\right)$ とする．あるモデルクラス M_k の確率密度関数は，

$$f(D; M_k) = \frac{f\left(D; \hat{\theta}\right)}{C_k}$$

によって与えられる．ただし，

$$C_k = \int_{\hat{\theta} \in \Omega} f\left(D; \hat{\theta}\right) dD$$

とする．Ω は，$\hat{\theta}$ の値の定義域を表す集合とする．両辺の対数を取ることにより，

$$-\log \hat{f}(D; M_k) = -\log f\left(D; \hat{\theta}\right) + \log C_k$$

となる．右辺第一項は，あるパラメータ $\hat{\theta}$ が与えられたときの状態 D の確率密度関数 $f\left(D; \hat{\theta}\right)$ を誤差と考えることができ，誤差の記述長を表す．右辺第二項は，全ての状態の確率密度関数を積分しており，モデルを表すと考えることができ，モデルの記述長を表す．従って，左辺は，モデルの記述長とそのモデルとデータの誤差の記述長の和である．

4. 評価方法

人工データと実データについて説明し，さらに 3 章の評価基準を人工データと実データに適用するための方法を述べる．

4.1 人工データ

I を人工データに含まれる全てのアイテムの集合とする．I から確率的に $|S|$ 個のアイテムを選択し，それらのアイテムを集合とするトランザクションを S とする．ただし，$|S|$ は，平均 $\overline{|S|}$，分散 1 のガウス確率に従う．S の集合を SS，モデルの次数を ℓ，状態遷移規則を R，R の集合を RS とする．R は，以下の規則により，確率的に作成される．

$$R = \left\{S_0^1, \cdots, S_0^c\right\} S_1 ... S_l$$

ただし，$S_i (1 = 1, \cdots, \ell)$ を SS の要素とし，c を次状態の候補数とする．データの最初は，トランザクション時系列 $X_1 \cdots X_\ell$ から開始する．次に，条件部が $S_1 \subseteq X_1 \wedge \cdots \wedge S_\ell \subseteq X_\ell$ を満たすときに，R の確率 $P_R(S_0^i)$ で次状態 S_0 を導く．ただし，$\sum_{i=1}^{c} P_R(S_0^i) = 1$ とする．最後に，I からノイズとなるアイテム集合 N を確率的に選択し，$X_{\ell+1} = S_0 \cup N$ とする．ただし，$|N|$ は，平均 $\overline{|N|}$，分散 1 のガウス確率に従う．この過程を，トランザクション時系列データ $D = \{X_t | t = 1, ..., n\}$ が生成されるまで繰り返す．デフォルト設定として，$|I| = 5000$，$|SS| = 1000$，$|RS| = 10$，$c = 3$，訓練データは n =1000 ステップ時間，テストデータも n =1000 ステップ時間とする．

AIC を適用する上で，先ず，モデルサイズについて述べる．AIC の評価式の右辺第三項 $(M+1)$ は，条件部 M 次と結論部 1 次の合計であり，モデルサイズを表している．各状態遷移規則に含まれる現状態 S_0 のノードが状態遷移規則の結論部に使われる頻度は，

$$\frac{\text{現状態 } S_0\text{のノードの支持度}}{\text{全データ数}}$$

に比例しているので，各現状態 S_0 のノードを 1 個と数えるのではなく (現状態 S_0 のノードの支持度)/(全データ数) 個と数えることにより，モデルサイズを考えることができる．従って，$M+1$ に相当するモデルサイズは，

$$\sum_{\text{現状態 } S_0\text{のノード } \in T} \frac{\text{現状態 } S_0\text{のノードの支持度}}{\text{全データ数}}$$

で与えられる．ただし，T はトライデータ構造に含まれる条件部と結論部のノードの集合とする．次に，予測誤差について述べる．観測状態を S_r，予測状態を S_e とする．時刻 t の S_r と S_e の予測誤差 $E_t(S_r, S_e)$ は，コサイン類似性尺度 $CM_t(S_r, S_e)$ [9] を用いて，

$$E_t(S_r, S_e) = 1 - CM_t(S_r, S_e)$$
$$(0 \leq E_t(S_r, S_e) \leq 1)$$

と表される．各アイテムのデータ出現確率を $p(item)$ としたとき，その情報量 $Info(item) = -\log(p(item))$ から，

$$E_t(S_r, S_e) = 1 - \frac{\sum_{item \in S_r \cap S_e} Info^2(item)}{\sqrt{\sum_{item \in S_r} Info^2(item)} \sqrt{\sum_{item \in S_e} Info^2(item)}}$$

となる．従って，誤差分散は，

$$\sigma^2 = \frac{1}{n} \sum_{t=1}^{n} \{E_t(S_r, S_e)\}^2$$

と求めることができる．

次に，MDL の適用方法について説明する．モデルの記述長は，AIC の場合と同様に考えることにより，

$$\log\left(\sum_{\text{現状態 } S_0\text{のノード } \in T} \frac{\text{現状態 } S_0\text{のノードの支持度}}{\text{全データ数}}\right)$$

と表される．誤差の記述長は，各時刻ステップ t の予測状態確率密度関数を前述の $CM_t(S_r, S_e)$ とし，時間ステップでその平均を考え，更に時系列データ $D\,(|D| = n)$ 全体での確率密度関数を考えることにより，

$$-n \log\left(\frac{1}{n} \sum_{t=1}^{n} CM(S_r, S_e)\right)$$

と表される．

4.2 実データ

実データは，カリフォルニア大学アーバイン校の知識発見データベース [10] に保管されている Pioneer-1 Mobile Robot の直進データと回転データを用いる．直進データは，ロボットが前進や後進するときの速度や壁との距離など測定している．回転データは，曲率半径を一定に固定して動くときの右タイヤと左タイヤの速度などを測定している．Pioneer-1 Mobile Robot の動作は 36 個のセンサーによってモニタされ，時系列の数値を出力する．従って，そのセンシングシステムから出力される直進データと回転データは，トランザクション時系列データとなる．現状の高次部分状態連鎖モデルは，各センサー出力が記号の場合しか扱えないので，データ前処理として，各センサー s の出力分布 $[s_{\min}, s_{\max}]$ を $disc$ 等分に離散化する．離散化された範囲を $[l_i^s, u_i^s](i = 1, \cdots, disc, l_1^s = s_{\min}, u_{disc}^s = s_{\max})$ とする．そして，各離散化区間出力値 x_t のアイテムを $< s : x_t >$ とする．

実データによる HISC モデリングへの AIC の適用方法について説明する．モデルサイズについては，4.1 節で述べた方法と同様であるが，予測誤差については，時間ステップ t における $Precision$ と $Recall$ を下式で与える．

$$Precision(X_t, S_t) = 1 - \sqrt{\frac{\sum_{<s:x_t> \in X_t, <s:[l_i^s, u_i^s]> \in S_t} e(<s:x_t>, <s:[l_i^s, u_i^s]>)}{|S_t|}}$$

$$Recall(X_t, S_t) = 1 - \sqrt{\frac{\sum_{<s:x_t> \in X_t, <s:[l_i^s, u_i^s]> \in \quad S_t} e(<s:x_t>, <s:[l_i^s, u_i^s]>)}{|X_t|}}$$

ただし，

$$e(< s : x_t >, < s : [l_i^s, u_i^s] >) = \left(\frac{x_t - (u_i^s + l_i^s)/2}{s_{max} - s_{min}}\right)^2$$

とする．これら 2 式より，時刻 t の予測誤差を

$$E_t = \sqrt{(1 - Recall)^2 + (1 - Precision)^2}$$

と定義する．従って，誤差分散は，

$$\sigma^2 = \frac{1}{n} \sum_{t=1}^{n} \{E_t\}^2$$

と求めることができる．

次に，実データによる HISC モデリングへの MDL の適用方法について説明する．モデルの記述長は，4.1 節と同様である．誤差の記述長については，確率密度関数を $\sqrt{Recall^2 + Precision^2}$ とし，時間ステップでその平均を考え，更に時系列データ $D\,(|D| = n)$ 全体での確率密度関数を考えることにより，

$$-n \log\left(\frac{1}{n} \sum_{t=1}^{n} \sqrt{Recall^2 + Precision^2}\right)$$

と表される．

5. 検証実験

3 章の手法を 4 章の人工データと実データに適用し，AIC と MDL による高次部分状態連鎖モデルの決定特性の分析を行う．

5.1 人工データによる検証

デフォルト設定のパラメータを用い，他のパラメータ，モデル次数 ℓ，主要なアイテムの個数 $|\overline{S}|$，ノイズなアイテムの個数 $|\overline{N}|$ を変化させて検証を行った．各設定を# ℓ-$|\overline{S}|$-$|\overline{N}|$ とラベル付ける．CrossValidation で評価した誤差分散，AIC，MDL が各々最小となる最小支持度を表に示す．CrossValidation で評価した誤差分散は，訓練データから得られる結果の正しさを検証するために，モデリングに用いたデータとは別にテストデータを用い，４章と同様の方法で求めた．この CrossValidation で誤差分散を最小にする最小支持度が，モデルの最適な詳細度に対応すると考えられる．ただし，CrossValidation 法は多くの計算時間や余分なテストデータを要するため，実際のモデル適用現場では実用性に問題があると考えられる．

表 1: 人工データの結果

# ℓ-$\overline{\lvert S\rvert}$-$\overline{\lvert N\rvert}$	CrossValidation	AIC	MDL
#1-3-7	0.02	0.02	0.03
#3-3-7	0.01	0.15	0.01
#4-3-7	0.02	0.15	0.03
#3-2-7	0.01	0.13	0.02
#3-3-7	0.01	0.15	0.01
#3-4-7	0.01	0.13	0.06
#3-3-2	0.01	0.14	0.02
#3-3-7	0.01	0.15	0.01
#3-3-14	0.03	0.16	0.02

AIC はモデルサイズ項の影響が非常に大きいので，CrossValidation と AIC により得られる各最小支持度の差は大きい．しかし，#1-3-7 は HISC モデルのサイズが小さいので，AIC はモデルサイズ項の影響が小さい．従って，この場合は，CrossValidation と AIC により得られる各最小支持度は一致する．また，MDL は，誤差項の影響が非常に大きいため，CrossValidation と MDL により得られる各最小支持度の差は小さく，より適切なモデル詳細度に対応する最小支持度を推定できることが判る．よって，AIC より MDL の方が適している．

5.2 実データによる検証

直進データは 6330 ステップ時間のトランザクション時系列データで，最初の 4000 ステップ時間を訓練データ，残りをテストデータとした．回転データは 2325 ステップ時間のトランザクション時系列データであり，最初の 1000 ステップ時間を訓練データ，残りをテストデータとした．離散化領域数 *disc* は先行研究に基づき 30,50,100.300 と設定する．人工データと同様に結果を示す．

表 2: 直進データと回転データの結果

	直進データ			回転データ		
disc	CrossValidation	AIC	MDL	CrossValidation	AIC	MDL
30	0.02	0.05	0.02	0.04	0.15	0.05
50	0.01	0.04	0.02	0.05	0.11	0.11
100	0.04	0.04	0.04	0.05	0.12	0.11
300	0.05	0.05	0.06	0.09	0.13	0.11

直進データに対する HISC モデルサイズは比較的小さく，モデルサイズ項と誤差分散項の両方が AIC に影響し，CrossValidation と AIC により得られる各最小支持度の差は小さい．さらに，disc を大きくすると，HISC モデルのサイズは小さくなるので，人工データ同様に最小支持度の差は小さくなる．また，MDL は誤差項の影響が非常に強いので，CrossValidation と MDL により得られる各最小支持度の差は小さい．

回転データの HISC モデルサイズは比較的小さく，モデルサイズ項の影響が大きいので，CrossValidation と AIC により得られる各最小支持度の差は大きい．ただし，disc を大きくすると，HISC モデルのサイズは小さくなるので，最小支持度の差は小さくなる．また，MDL は誤差項の影響が非常に強いので，CrossValidation と MDL により得られる各最小支持度の差は小さい．よって，AIC より MDL の方が適している．

6. おわりに

本稿では，AIC と MDL を人工データと実データの両方に適応し，性能評価実験を行った．その結果，MDL を用いることで，高次部分状態連鎖モデルの適切なモデルサイズを効率良く決定できることを確認した．今後の課題として，AIC や MDL を超えて，より適切な HISC モデルの詳細度を検討することが挙げられる．

参考文献

[1] [Beaudin 04]J. Beaudin, S. Intille, and E. M. Tapia, “ Lessons learned using ubiquitous sensors for data collection in real homesn, ” Extended Abstracts of the 6th Conf. on Human Factors in Computing Systems. (2004)

[2] [Sun 02]Z. Sun, R. Miller, G. Bebis, and D. DiMeo, “ A real-time precrash vehicle detection system, ”Proc. of the IEEE Workshop on Applications of Computer Vision. (2002)

[3] [Nguyen 06]Viet Phuong Nguyen and Takashi Washio, “ Modeling Dynamics of Massive Dimensional and Complex Systems ”, Proc. of the Int. Workshop on Data-Mining and Statistical Science (2006), pp.125-132,(2006).

[4] [Berchtold 02]A. Berchtold and A. E. Raftery, “The mixture transition distribution model for high-order markov chains and non-gaussian time series,” *Statistical Science, 17(3)*. (2002)

[5] 尾崎統, 北川源四郎 (2003), 時系列解析の方法, 朝倉書店.

[6] [Barron 98]A. Barron,J. Rissanen and Bin Yu, “ The minimum description length principle in coding and modeling ”(1998)

[7] 情報理論とその応用学会 (2000), 情報源符号-歪みのあるデータ圧縮, 培風館.

[8] 下平 英寿, 久保川 達也, 伊藤 秀一, 竹内 啓 (2004), モデル選択-予測・検定・推定の交差点, 共立出版.

[9] J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” *Morgan Kaufmann, San Francisco, CA, USA*. (2006)

[10] UCI-KDD-Data-Repository, “ http://kdd.ics.uci.edu/databases/pioneer/pioneer.data.html. ”