知 能 と 複 雑 系 130－1
（２００２．１１．１２）

多重サイコロモデルを用いたEM法による慢性肝炎データ医療検査結果の類型化

渡辺　健志[1]　　鈴木　英之進[1]
横井　英人[2]　　高林　克日己[2]

概要

本論文ではアクティブマイニング研究の一例として，EM法に基づいた慢性肝炎データにおける医療検査結果の類型化を示す．医療の専門家を交えた試行錯誤の末，類型を表現する確率モデルとして多重サイコロモデルを提案するに至った．実験によって得られた類型の大部分は，医学的背景から見て説明が容易かつ明確であり，本手法は患者の予測や理解に関して有望であることが分かった．

Prototyping Medical Test Results in Chronic Hepatitis Data with the EM Algorithm on Multi-Dice Models

Takeshi Watanabe[3]　**Einoshin Suzuki**[3]
Hideto Yokoi[4]　**Katsuhiko Takabayashi**[4]

Abstract

This paper presents an example of active mining endeavor in which we estimate prototypes of medical test results in chronic hepatitis data with the EM algorithm. As a result of trials and errors with medical domain experts, we have come to invent a multi-dice model, which represents a probabilistic model of the prototype. Experiments show that most of the obtained prototypes can be interpreted easily and clearly in the medical context, and our proposed method is promising in both recognition and understanding of a patient.

[1] 横浜国立大学大学院工学府物理情報工学専攻電気電子ネットワークコース
[2] 千葉大学医学部附属病院医療情報部
[3] Department of Electrical and Computer Engineering; Division of Advanced Physics, Electrical and Computer Engineering, Graduate School of Yokohama National University
[4] Division for Medical Informatics, Chiba-University Hospital

1 導入

典型的な医療検査結果は，日付，患者 ID，検査値から構成され，1 日に 1 人の患者が受ける検査数は全検査項目数に比較してきわめて小さいという特徴がある．そのため一見，医療検査結果集合は相関ルール発見 [1] に頻繁に用いられるトランザクションデータ集合として扱えるように考えられる．しかし，われわれは未検査の扱いと正常値の区別に特別な配慮が必要という結論に至った．本稿でアクティブマイニングの対象となる慢性肝炎データ [2] は多くの医療検査結果を含み，既存の学習・発見手法では効果的な解析ができない．

この問題を解決するため，われわれ機械学習研究者と医療専門家は，EM 法に基づく混合確率分布推定 [4] を用いて検査結果を類型化することを試みた．医療専門家の評価により，類型は患者レベルでのデータの理解，明確化に有用であることが示された．本稿では試行錯誤の過程と今回の結果の医療分野への貢献を示す．

2 混合多項分布推定

2.1 問題定義

患者 1 人が 1 日の特定回[5] に受けた検査結果集合を 1 事例とする．A は検査項目集合であり，$A = \{a_1, a_2, \cdots, a_{n(A)}\}$ と表される．ただし $a_l, n(A)$ はそれぞれ検査項目，検査項目数である．この問題の入力はデータ集合 X であり，$n(X)$ 個の事例 $X = \{\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_{n(X)}\}$ で構成される．ここで x_i は属性値ベクトルを表し，$n(A)$ 個の値 $\mathrm{x}_i = \{v_i(a_1), v_i(a_2), \cdots, v_i(a_{n(A)})\}$ で構成される．ただし各 $v_i(a_l)$ は事例 x_i における検査 a_l の値を示す．出力は集合 K であり，$n(K)$ 個の類型 $K = \{k_1, k_2, \cdots, k_{n(K)}\}$ で構成される．事例 x_i の生起確率は次で与えられる．

$$p(\mathrm{x}_i) = \sum_{j=1}^{n(K)} p(\mathrm{x}_i|k_j)p(k_j) \tag{1}$$

ただし $p(k_j)$ は類型 k_j の生起確率を表し，$p(\mathrm{x}_i|k_j)$ は k_j のときに x_i の起こる条件付き確率を表す．

[5] データ集合 [2] において，同じ患者 ID，日付，回数をもつトランザクション集合

2.2 EM 法

本稿では混合確率分布推定に，EM 法 [4] を用いる．この手法は山登り法 [6] によって $p(k_j)$ と $p(\mathrm{x}_i|k_j)$ の最尤値を推定する．これらの最尤値は負の対数尤度 ε を最小化する値として定義できる．

$$\varepsilon = -\sum_{i=1}^{n(X)} \ln \left(\sum_{j=1}^{n(K)} p(\mathrm{x}_i|k_j)p(k_j) \right) \tag{2}$$

EM 法の手順は，次の通りである．

1. $p(k_j)$ と $p(\mathrm{x}_i|k_j)$ の初期値を決定する．
2. ベイズ則から $p(k_j|\mathrm{x}_i)$ を計算する．

$$p(k_j|\mathrm{x}_i) = \frac{p(\mathrm{x}_i|k_j)p(k_j)}{p(\mathrm{x}_i)} \tag{3}$$

$$\text{where } p(\mathrm{x}_i) = \sum_{j=1}^{n(K)} p(\mathrm{x}_i|k_j)p(k_j) \tag{4}$$

   $p(\mathrm{x}_i|k_j)$ を求める手順は確率モデルによって異なり，詳細は 3 章の各モデルの説明で述べる．
3. $p(k_j)$ と $p(\mathrm{x}_i|k_j)$ を更新する．

$$p^{\mathrm{new}}(k_j) = \frac{1}{n(X)} \sum_{i=1}^{n(X)} p(k_j|\mathrm{x}_i) \tag{5}$$

   $p^{\mathrm{new}}(\mathrm{x}_i|k_j)$ を求める手順は確率モデルによって異なり，詳細は 3 章の各モデルの説明で述べる．
4. 収束するまで手順 2，3 を繰り返す．

3 提案手法

3.1 単一サイコロモデル

代表的な医療検査は正常値の範囲が指定されており，この範囲から外れる値を異常と判断する．ここで 1 と 0 をそれぞれ異常値とそれ以外の値と見なすと，医療検査結果は大部分の検査値が 0 である，疎な 2 値の表で表されるトランザクションデータに変換できる．

この節では，1 つの類型を 1 つの $n(A)$ 面サイコロで表し，これを単一サイコロモデルと呼ぶ．$n(A)$ は前述の通り検査項目数を示す．サイコロのある面が起こる確率は相当する検査が異常となる

確率を表す．多項分布 [5] は互いに排反で全て合わせると全事象となる事象を独立に複数回試行したときに各事象の起きる回数をモデル化する．例えばサイコロを投げたとき各面の出る確率をモデル化できる．多項分布は購買行動のプロファイリング [3] に用いられて成功したことから，どのような種類のトランザクションデータも効果的に解析できると思われた．単一サイコロモデルを用いて異常な医療検査値の類型を得るため，事例 x_i の検査 a_l の値 $v_i(a_l)$ を正常か異常かで分類する．

$$v_i(a_l) = \begin{cases} 0 & \text{(normal value)} \\ 1 & \text{(abnormal value)} \end{cases} \tag{6}$$

類型 k_j は各医療検査値が異常となる確率から構成される．

$$k_j = \{p_j(a_1), p_j(a_2), \cdots, p_j(a_{n(A)})\} \tag{7}$$

各 $p_j(a_l)$ は類型 k_j において医療検査 a_l が異常となる確率を表す．多項分布の定義 [5] から式 (3),(4) の条件付き確率 $p(\mathrm{x}_i|k_j)$ は以下[6] で与えられる．

$$p(\mathrm{x}_i|k_j) = \frac{n(A)!}{\prod_{l=1}^{n(A)} v_i(a_l)!} \prod_{l=1}^{n(A)} p_j(a_l)^{v_i(a_l)} \tag{8}$$

条件付き確率 $p(\mathrm{x}_i|k_j)$ は

$$p_j^{\mathrm{new}}(a_l|k_j) = \frac{\sum_{i=1}^{n(X)} p(k_j|\mathrm{x}_i) v_i(a_l)}{\sum_{i=1}^{n(X)} p(k_j|\mathrm{x}_i) \sum_{l=1}^{n(A)} v_i(a_l)} \tag{9}$$

を用いて更新される．

3.2 多重コインモデルと多重サイコロモデル

単一サイコロモデルは 2 つの欠点を持つ．1) 検査値の分類が正常もしくは異常という粗いものであることと，2) 正常値と未検査を同じと見なしてしまうことである．1 つ目の問題に対処するため，検査 a_l の値を次のように離散化する．$R(a_l) = \{r_1(a_l), r_2(a_l), \cdots, r_{n(a_l)}(a_l)\}$ ただし $n(a_l)$ と $r_m(a_l)$ は離散化したラベルの数と m 番目のラベルを表す．2 つ目の問題に対しては，未検査と他の値を分けて扱うことで対処する．事例 $\mathrm{x}_i = \{v_i(a_1), v_i(a_2), \cdots, v_i(a_{n(A)})\}$ において $v_i(a_l)$ は次のどれかの値をとる．

$$v_i(a_l) = \begin{cases} r_1(a_l) \\ r_2(a_l) \\ \vdots \\ r_{n(a_l)}(a_l) \\ - \quad \text{(untested)} \end{cases} \tag{10}$$

直観的に式 (10) は投げない事も許容したサイコロを投げる事象を表す．1 つのサイコロは 1 つの検査項目に当たり，$n(A)$ 個の検査モデルには $n(A)$ 個のサイコロを必要とする．

類型 k_j は

$$k_j = \{k_{j1}, k_{j2}, \cdots, k_{jn(A)}\} \tag{11}$$

$$\text{where } k_{jl} = \{p_j(r_1(a_l)),\ p_j(r_2(a_l)), \cdots, p_j(r_{n(a_l)}(a_l))\} \tag{12}$$

で表され，各 $p_j(r_m(a_l))$ は類型 k_j において検査 a_l の値が $r_m(a_l)$ となる確率を表す．

式 (3)，(4) における条件付き確率 $p(\mathrm{x}_i|k_j)$ は以下で与えられる．

$$p(\mathrm{x}_i|k_j) = \prod_{l=1}^{n(A)} p_j(v_i(a_l)) \tag{13}$$

ただし $v_i(a_l) = -$ の場合 $p_j(v_i(a_l)) = 1$ と見なす．

条件付き確率 $p(\mathrm{x}_i|k_j)$ は

$$p_j^{new}(r_m(a_l)|k_j) = \frac{\sum_{i=1}^{n(X)} p(k_j|\mathrm{x}_i) p(k_j) \gamma_1(\mathrm{x}_i, k_j, l, m)}{\sum_{i=1}^{n(X)} p(k_j|\mathrm{x}_i) p(k_j) \gamma_2(\mathrm{x}_i, k_j, l, m)} \tag{14}$$

$$\gamma_1(\mathrm{x}_i, k_j, l, m) = \begin{cases} 1(v_i(a_l) = r_m(a_l)) \\ 0(v_i(a_l) \neq r_m(a_l)) \end{cases} \tag{15}$$

$$\gamma_2(\mathrm{x}_i, k_j, l, m) = \begin{cases} 1(v_i(a_l) \neq -) \\ 0(v_i(a_l) = -) \end{cases} \tag{16}$$

で更新される．

なお，検査値を正常と異常に離散化する，$\forall l\ n(a_l) = 2$ の場合，多重サイコロモデルは多重コインモデルに縮退する．実際，我々は単一サイコロモデルの失敗から多重コインモデルを提案し，そして 4.2 節の試行錯誤の末に $n(a_l) > 2$ となる多重サイコロモデルに至った．

[6] 式 (6) から式 (8) では $\prod_{l=1}^{n(A)} v_i(a_l)! = 1$ となるが，式を明確にするためこのように表記する．

4 実験

4.1 単一サイコロモデル

4.1.1 条件

実験には千葉大学附属病院から提供していただいた慢性肝炎データ [2] を用いる．このデータは 58,716 事例，検査数 458 から構成される．3.1 節で述べたように，単一サイコロモデルは類型の表現形式として用いられ，各類型は異常となる検査のパターンを表す．

EM 法は山登り法を用いるため，必ずしも大域的最適解には収束しない．そのため初期値をランダムに与えて EM 法を 100 回行い，式 (2) の ε が最小となる結果を採用する．類型数 $n(K)$ は $2, 3, \cdots, 10$ で行い，ループは各パラメータの変化が 0.01%以下，もしくはループ数が 100 回となった場合に終了する．

なお得られた類型に対し類似度に基づいてクラスタリングも行った．2 つの確率分布間の距離を測るダイバージェンスを用いて，類型 k,l の類似度 $\beta(k, l)$ を求めた．

$$\beta(k, l) = \frac{D(k\|l) + D(l\|k)}{2} \tag{17}$$

$$D(k\|l) = \sum_{i=1}^{c} p_{li} \ln \frac{p_{li}}{p_{ki}} \tag{18}$$

ただし c は k，l の要素数であり $k = (p_{k1}, p_{k2}, \cdots, p_{kc})$, $l = (p_{l1}, p_{l2}, \cdots, p_{lc})$ である．0 による除算を避けるため，$p_{ki} = 0$ のときは $p_{ki} = 1 \times 10^{-100}$ とする．

4.1.2 結果

紙面の制限のため $n(k) = 10$ のときの結果だけ示す．図 1 に作成された 10 個の類型を示す．横軸と縦軸はそれぞれ検査番号，検査値が異常となる確率 [%] を表す．重要と判断される検査項目は相当する類型内に記述されている．類型 k と l は $\beta(k, l) \leq 10$ ならば同一クラスタに所属すると見なすと，クラスタリングの結果は {1,2,3},{4,5,6},7,8,9,10 となった．

図を見ると，類型 10 と 3 は他と大きく異なる．類型 10 では APO[7] 関係の検査が高い確率を示し，一方類型 3 は 2 つの医療検査が他と比べ高い確率を示している．

専門家に類型 10 は脂肪系蛋白に異常が起きている状態をよく表しているとのコメントを頂いた．一方，類型 3 に高確率で所属する大部分の事例は，1 日のうち 2 回目に測定されたものであった．このデータにおいて 1 日のうち 2 回目に測定される医療検査はほぼ固定されており，類型 3 はその傾向を反映すると考えられる．更に，全事例のおよそ半分は高確率で類型 1 に所属し，図からは分からないがそのうち 9,578 事例は異常検査値を持たない．専門家は同じクラスタ内の類型は似すぎているため，もっと多様な類型を見たいとコメントした．これらの結果は，単一サイコロモデルの成功が限定的であることを示している．

4.2 多重コインモデル

4.2.1 条件

前節のモデルでは全てのデータを対象とし，単一サイコロモデルが正常値と未検査を同一視してしまうため，限定的にしか成功しなかった．ウィルスマーカー検査はウィルスと抗体の状態を調べる医療検査である．本稿ではウィルスと抗体の状態の組合せをパターンと呼ぶ．この実験では専門家の意見に従い，B 型肝炎患者のパターンごとに類型を導出する．表 1 に用いた 13 パターンを示す．

これら 13 パターンの全患者数は 263，事例数は 492 である．数が大幅に減少したのは 1 日に 1 回で 4 種類のウィルスマーカー検査を受けた患者を対象とし，これらの患者は非常に少ないためである．また検査項目は主要な 11 種類を用い，類型数 $n(K)$ は 3 とする．

4.2.2 結果

図 2 に実験結果を示す．パターン 2,12,13 は事例数が少ないので省略する．また医学的にはパターン 12 と 13 は起こり得ない．各パターンは 11 種類の検査項目からなり，類型はそれぞれ異なる濃淡

[7] 蛋白質の一種

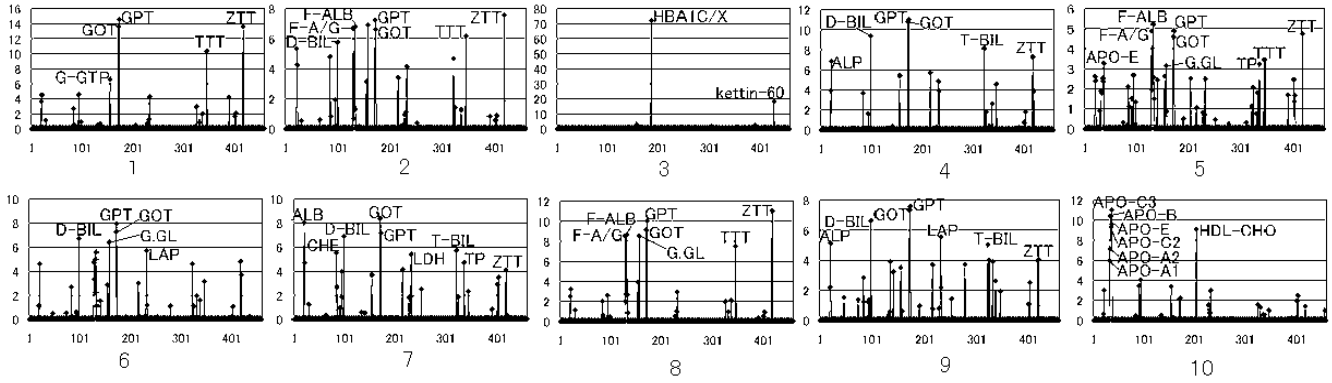

図 1: 単一サイコロモデルにより生成された類型

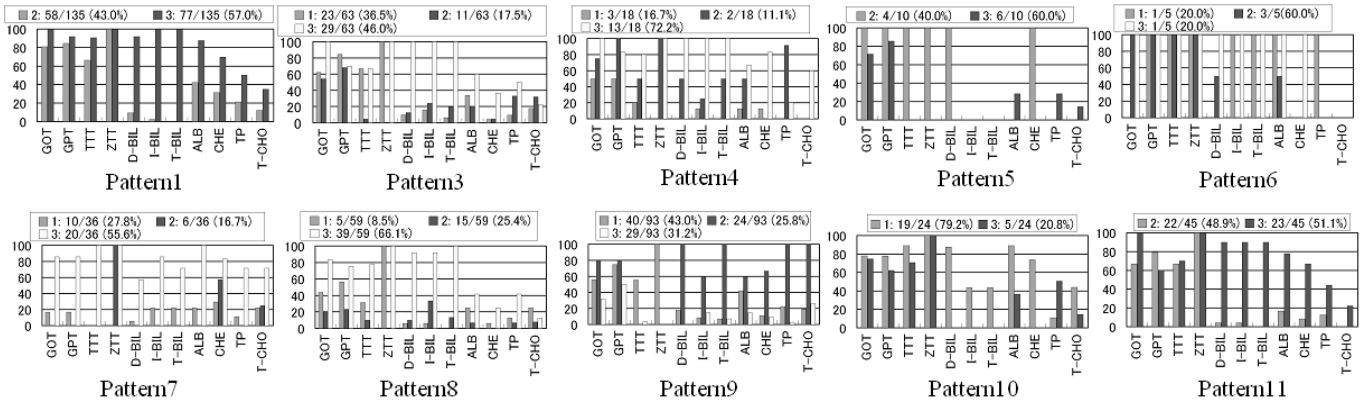

図 2: 多重コインモデルにより生成された類型

表 1: B 型肝炎の進行を示すパターン．#は各パターン内の事例数を表す.

パターン	直観的な説明	#
1	未感染	135
2	感染初期	2
3	ウィルス活動中	63
4	活動対処	18
5	活動中存在対処	10
6	活動存在対処	5
7	非活動中存在対処	36
8	ウィルス消滅	59
9	非活動中	93
10	元から活動対処抗体所持	24
11	元から存在対処抗体所持	45
12	あり得ない 1	1
13	あり得ない 2	1

表 2: データ修正後の B 型肝炎の進行を示すパターン

パターン	直観的な説明	#
1	未感染	74
2	感染初期	113
3	ウィルス活動中	3867
4	活動対処	1157
5	ウィルス非活動中	3419
6	存在対処	254
7	ウィルス消滅	368
8	元から存在対処抗体所持	10
9	元から活動対処抗体所持	22

で示される．

専門家には多くの類型は妥当であるが，互いにとても似ているグループが存在するとのコメントを頂いた．これは検査値の分類を正常か異常という粗い離散化で行ったためであり，より良い離散化が必要と考えられた．

4.3 多重サイコロモデル

4.3.1 データ修正

元データではウィルスマーカー検査の多くの値は未検査，もしくは不明確な値[8] であった．専門家によるとこれらの検査の基準は頻繁に変わり，誤判定もよくあるとのことである．

専門家の意見に従ってデータ選択の基準を変更し，ウィルスマーカー検査の未検査部分を教えていただいたルールに従って補完した．この結果，新しいデータは患者 102 人，9,190 事例で構成される．表 2 にデータ修正後のパターンを示す．

4.3.2 結果

3.2 節で述べたように，各検査値は離散化によってラベル付けできる．各検査ごとにラベル集合[9] を定義した．

図 3 に得られた類型を示す．各パターンでは左から右に 3 つの類型が示され，各医療検査において離散化検査値は割合グラフで表される．専門家からは多くの類型は可読性に優れ，医学的な意味が明確であるとのコメントを頂いた．また本手法は，インターフェロンの[10] 効果や肝炎被害の進行の予測問題などの応用に有望と考えられる．パターン 2,3,5,7 は明確に分かれた妥当な類型であるとのコメントを得た．パターン 1 の類型は最初は不適当と思われたが，これは少数の特殊な患者によって構成されるデータから得たためであった．多重サイコロモデルはこのような特殊な患者の発見にも効果的と考えられる．

[8] 主に疑陽性や疑陰性などの不明確な判定

[9] 各ラベルは vL (very Low), L (Low), N (Normal), H (High), vH (very High), uH (ultra High) であり，N に近いほど症状が軽い事を表す．ただし検査によってラベル集合が異なる．

[10] 肝炎ウィルスの特効薬．人によって効果が異なる．

4.3.3 可視化への応用

図 4 は，ある患者について時系列順にパターンと類型の組を示したものであり，この患者の確率的な状態推移を表している．図から ID446 の患者は 1987 年 4 月 20 日において活動中パターンの類型 2 に 100%所属している．図 3 のパターン 3 の類型 2 を見ると，ほとんどの検査値は正常であることを示しており，この患者はその日には活動中パターンでも比較的正常な状態であった．しかし 1990 年 10 月 24 日には，活動対処の類型 3 に所属し，やや悪化している．続く一連の (パターン，類型) の組から，この患者は (活動中，悪い)，(非活動中，良い)，(非活動中，とても良い)，(ウィルス消滅，やや悪い) と推移していく様子が分かる．

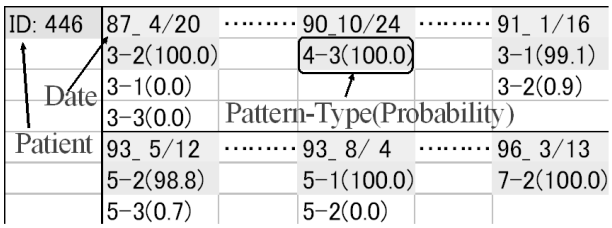

図 4: 患者病態推移の視覚化

多重サイコロモデルを採用した本手法は実用的には見ることができない検査結果集合を可読性に優れた類型に変換し，患者の病態推移を効果的に調べられる．専門的見知からあまり直観的でない類型もあるが，更なる領域知識を用いてパターンを再分類したり，4.3.2 節で述べたような特殊な患者を除外することでより良くなると考えられる．我々は今後も慢性肝炎データのアクティブマイニングを続け，本手法を発展させる予定である．

5 結論

本稿では確率混合モデルに基づいて医療検査値の類型を導出した．専門家の意見に基づいて試行錯誤した末に，類型の表現として多重サイコロモデルを提案し，データの選択・修正の上で成功をおさめた．専門家によると，得られた大部分の類型は可読性に優れ，医学的な意味が明確である．

今後は予測問題への応用が有望と考えられる．成功への鍵は更なるデータ前処理とクラス情報の

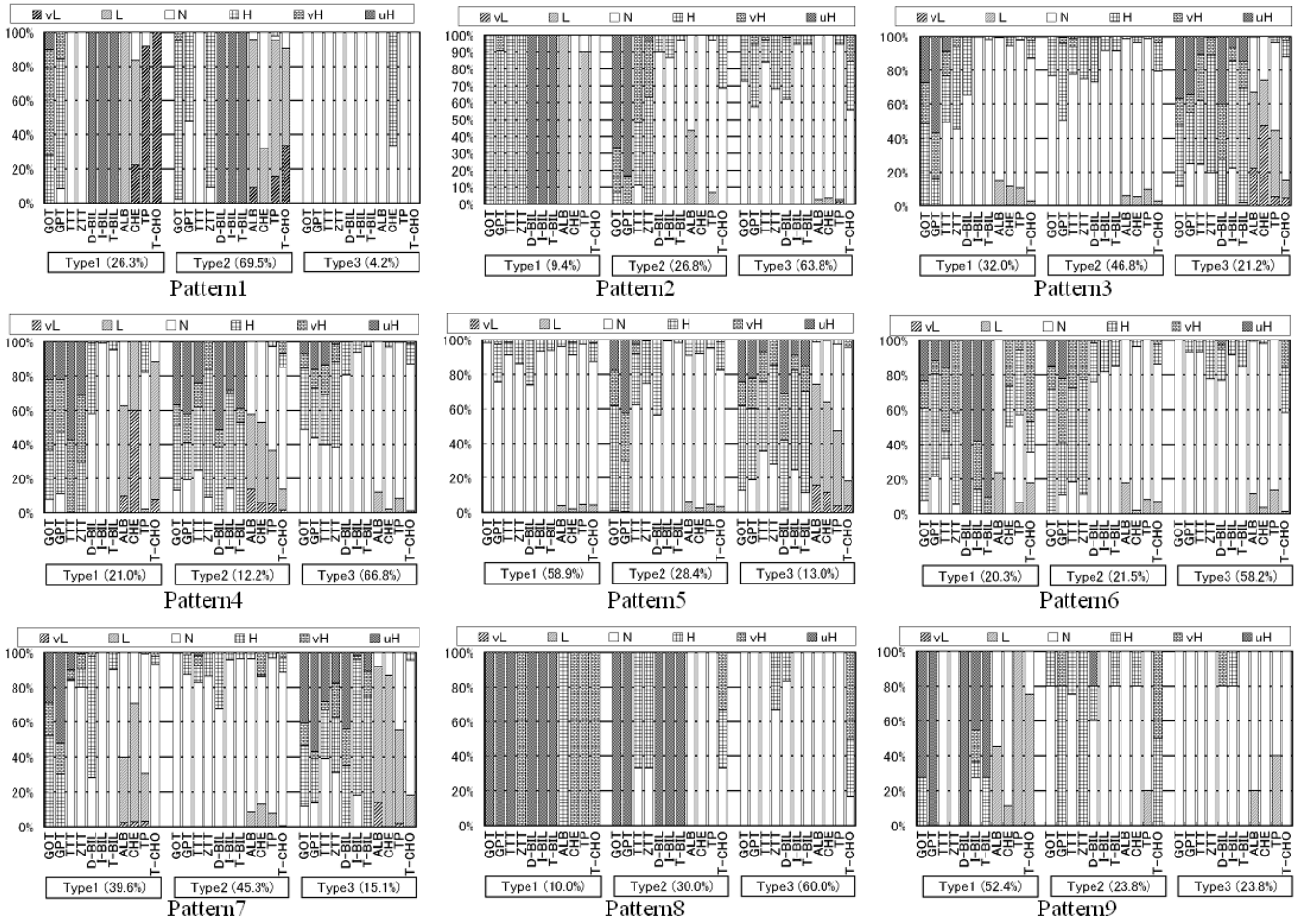

図 3: 多重サイコロモデルによって生成された類型

利用法となるであろうと考える.

参考文献

[1] R. Agrawal et al. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, Menlo Park., Calif., 1996.

[2] P. Berka. ECML/PKDD 2002 discovery challenge, download data about hepatitis. http://lisp.vse.cz/challenge/ecmlpkdd2002/, 2002. (current September 28th, 2002).

[3] I. V. Cadez, P. Smyth, and H. Mannila. Probabilistic modeling of transaction data with applications to profiling, visualization, and prediction. In *Proc. Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 37–46, 2001.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38, January 1977.

[5] W. Feller. *An Introduction to Probability Theory and Its Applications, Volume 1.* John Wiley & Sons, 1957.

[6] S. Russell and P. Norvig. *Artificial Intelligence, A Modern Approach*. Prentice Hall, Upper Saddle River, N. J., 1995.