High-order Visualization of Medical Test Data with PrototypeLines

Takeshi WATANABE†, Einoshin SUZUKI†, Hideto YOKOI††, and Katsuhiro TAKABAYASHI††

† Faculty of Engineering, Yokohama National University 79-5 Tokiwadai, Hodogaya, Yokohama, 240-8501 Japan
†† Division of Medical Informatics, Chiba University Hospital 1-8-1 Inohana, Chuou-ku, Chiba, 260-8677 Japan
E-mail: †nabekun@slab.dnj.ynu.ac.jp, suzuki@ynu.ac.jp,
††yokoih@telemed.ho.chiba-u.ac.jp, takaba@ho.chiba-u.ac.jp

Abstract In this paper, we propose a method which visualizes irregular multi-dimensional time-series data as a sequence of probabilistic prototypes, and apply it to medical data. Chronic hepatitis data, which represent typical medical test data, pose various challenges to conventional machine learning methods. We have proposed a method which obtains prototypes from the data, and the prototypes were admitted valid and comprehensive by physicians. This paper proposes PrototypeLines which represent a visualization method based on the prototyping method. PrototypeLines consider information criterion and hue recognition, and display medical data of each patient in a comprehensive format. Promising results were obtained by experiments in which we visualized patients based on their degrees of fibrosis.

Key words PrototypeLines, Visualization, Estimation of Mixture Probabilistic Models, Time-Series Data, Medical Test Data

1. はじめに

病院検査データは患者の追跡調査（follow-up study）などで採られるデータであり、高次元（多属性）、不規則な計測、個人差、医師による測定バイアスなどの問題点を従来の機械学習手法に課している[11]。例えば、文献学者が研究費定額領域研究アクティビマーニングプロジェクトの共通データとして用いられている慢性肝炎データは、これらの問題点をすべて含み、ECML/PKDD 2002 Discovery Challenge 国際ワークショップ[8]でのデータマイニングコンテスト[8]でも用いられた。情報処理化、その直観的な理解しやすさのために、ユーザとのインタラクションが必要な応用問題で重宝されており、データマイニングにおいても今後ますます重要となっていくと考えられる[2], [4]。データマイニングにおいて可視化は、データの前処理、知識の発見、および発見結果の提示などに用いる
ことができる。本論文では主にデータの前処理を想定し、病院検査データを理解・整理する方法を提案する。

情報可視化手法は一般に、データを可視化する手法とパターンを可視化する手法に分けられる。データを可視化する手法は、生データを可視化する手法、生データを変換して可視化する手法、および例を可視化する手法に分類できる。本論文では、生データを変換し、人間が分かりやすい形式に変換して可視化する手法を高次可視化手法と呼び、病院検査データに適用する高次可視化手法を提案する。提案手法のブロック図は、われわれが以前に発表した病院検査値の類型化 [12]に基づいている。この手法は、慢性肝炎データから発見された類型の多くが医師によって妥当で理解しやすいと認められた。ブロック図は以下の通りで、データを抽象化し、情報量規律と業界認知を考慮して分かりやすく表示する。

### 2. 慢性肝炎データからの類型導出

#### 2.1 慢性肝炎データ

慢性肝炎は、ウイルス感染により肝細胞に炎症を起こすとし、慢性炎症が起こると肝硬変に移る。肝硬変を示す検査の類型がブロック図を示す。肝硬変を示す検査は、表4及び図1に示すように、細胞膜の障害を示すものである。慢性肝炎は、肝細胞の変性を示すもので、肝機能の低下を示すものである。慢性肝炎の主要な検査は、 GOT、GPT、ZTT、TBL、D-BIL、TBIL、AGB、CHE、T-CHO、PLT、HGB、HCT、MCVを示す。慢性肝炎の検査データの重要性は、以下の通りである。

<table>
<thead>
<tr>
<th>検査項目</th>
<th>直観的説明</th>
<th>カテゴリ</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOT, GPT</td>
<td>増悪の肝細胞の量</td>
<td>N,H,V,H,US</td>
</tr>
<tr>
<td>ZTT, TTT</td>
<td>炎症の程度</td>
<td>N,H,V,UH</td>
</tr>
<tr>
<td>T-BIL, D-BIL, TBIL</td>
<td>免疫球蛋白の異常</td>
<td>N,H,V,UH</td>
</tr>
<tr>
<td>ALB, CHE, T-CHO, TP</td>
<td>肝組織の予備能</td>
<td>RL,H,V,HL</td>
</tr>
<tr>
<td>WBC, PLT</td>
<td>炎症の血液の反応</td>
<td>UL,HV,L,N,H</td>
</tr>
<tr>
<td>RBC, HGB, HCT, MCV</td>
<td>血液検査</td>
<td>L,H</td>
</tr>
</tbody>
</table>

図1は、慢性肝炎データを示す。図1は、慢性肝炎データを示す。図1は、慢性肝炎データを示す。図1は、慢性肝炎データを示す。図1は、慢性肝炎データを示す。図1は、慢性肝炎データを示す。図1は、慢性肝炎データを示す。慢性肝炎データの類型は、以下の通りである。

#### 2.2 EM法に基づく確率的類型の導出

図1の慢性肝炎データの類型は、以下の通りである。

$$ p(x_i) = \sum_{j=1}^{n(K)} p(x_i|k_j)p(k_j) $$

(1)

ただし、$p(k_j)$は類型$k_j$の生起確率を表し、$p(x_i|k_j)$は$k_j$のときに$x_i$が起きる条件付き生起確率を表す。

(注1): データ集合[1]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
われわれは類型の推定に、EM 法 [3] を用いた。この手法は山登り法によって \( p(k_j) \) と \( p(x_i|k_j) \) の最尤値を推定する。これらの最尤値は負の対数尤度 \( \varepsilon \) を最小化する値として定義できる。

\[
\varepsilon = - \sum_{i=1}^{n} \ln \left( \sum_{j=1}^{n} p(x_i|k_j)p(k_j) \right)
\]

2.3 多重サイコロモデル
われわれは専門家を交えた反復的なデータ解析の結果、次に示す多重サイコロモデルを考察した。検査 \( a_i \) の値は、

\[ R(a_i) = (r_1(a_i), r_2(a_i), \ldots, r_n(a_i)) \]

で離散化されるとする。ただし、\( n(a_i) \) と \( r_m(a_i) \) は離散化区間数と \( m \) 番目の区間を表す。検査値を区別するため、\( r_i(a_i) \) は右辺の \( i \) の値をとることができる。

\[
v_i(a_i) = \begin{cases} 
1 & (r_1(a_i)) \\
2 & (r_2(a_i)) \\
\vdots \\
n(a_i) & (r_n(a_i)), \quad \text{未検査}
\end{cases}
\]

検査 \( a_i \) は \( n(a_i) \) 面のサイコロを投げる行為としてモデル化されるため、\( n(A) \) 個の検査をモデル化するために \( n(A) \) 個のサイコロを用いる。式 (3) は、サイコロを投げない場合があるサイコロ投げを表現する。

類型 \( k_j \) は次式で表される。

\[
k_j = (k_{j_1}, k_{j_2}, \ldots, k_{j_{n(A)}})
\]

ただし \( k_{j_i} = (p(r_1(a_i)|k_j), p(r_2(a_i)|k_j), \ldots, p(r_n(a_i)|k_j)) \)

\[
p(r_m(a_i)|k_j) \quad \text{は、類型} \ k_j \text{において、検査} \ a_i \text{の値が未検査でな}
\]

い場合 \( r_m(a_i) \) となる確率を表す。この類型を用いた確率集合モデルを、多重サイコロモデルと呼ぶ。EM 法において、条件つき確率 \( p(x_i|k_j) \) は次式で与えられる。

\[
p(x_i|k_j) = \prod_{l=1}^{n(a_i)} p(v_l(a_i)|k_j)
\]

ただし \( v_l(a_i) = - \) の場合は \( p(v_l(a_i)|k_j) = 1 \) とする。

条件つき確率 \( p(x_i|k_j) \) は、次式を用いて更新される。

\[
p^{new}(r_m(a_i)|k_j) = \frac{\sum_{i=1}^{X} p(k_j|x_i)p(k_j)\gamma_1(x_i, k_j, l, m)}{\sum_{i=1}^{X} p(k_j|x_i)p(k_j)\gamma_2(x_i, k_j, l, m)}
\]

ただし

\[
\gamma_1(x_i, k_j, l, m) = \begin{cases} 
1 & (v_l(a_i) = r_m(a_i)) \\
0 & (v_l(a_i) \neq r_m(a_i))
\end{cases}
\]

\[
\gamma_2(x_i, k_j, l, m) = \begin{cases} 
1 & (v_l(a_i) = -) \\
0 & (v_l(a_i) = -)
\end{cases}
\]

実験では、B 型慢性肝炎患者を抗原と抗体の状態で分けて類型を導出した。発見された類型はほぼ全て、医師らによって説明性に優れた「妥当であることが認められた。さらに医師らは、この手法は患者の病態変化を把握することと例外的な患者を発見することに関してきわめて有効であることをコメントした。

3. プロトタイプライン
3.1 動 機
2.1 節と 2.2 節で述べた動機に基づき、病院検査データを整理・理解しやすいように患者毎に可視化することが考えた。可視化が適切であれば、患者の病態変化を把握しやすく、学習・発現アルゴリズムに入力するデータを準備しやすくなると期待される。

例えば慢性肝炎データには、問題を限定しても数百人の患者がおり、主要な検査値だけでも約 17 個ある。図 1 の形式では、A4 用紙 1 枚に数人しか示せず、多数の患者を同時に分析することは難しい。この問題を解決するため、生データを 2.3 節で紹介した確率的類型を用いて変換し、事例を式 (1) の形式で表示することにした。これは、検査結果を確率的類型の線形和として表現することに相当する。

3.2 類型への色調り当て
前節の形式でデータを表示する場合、類型を分かりやすく表示することが重要である。データを直接可視化する手法では通常、色、形、および大きさなど様々な情報媒体を利用する [2], [4]。このような手法の表示結果は、観察者には分かりやすいと考えられるが、重要な情報を見落としてしまう場合も多いと考えられる。本論文では病院検査データの複雑さとユーザへの分かりやすさを考慮し、類型を色相 (hue) [5] だけで表示することにした。類型は応用分野における尺度、例えば慢性肝炎データでは検査結果の善し悪しに応じて整列され、紫から赤の色相を割り当てられよう。

類型を整列するためには、各検査の異なるカテゴリを評価する単一の指標が必要である。慢性肝炎データにおいては表 1 に示すように検査値にカテゴリが存在し、VL は L よりも悪さなどの情報が関与している。さらに、例えば GOT の検査値は比較的 H や VH の分類になりやすいが、T-BII は多様になるなど、検査結果のカテゴリは検査法によって重要度が異なると考えられる。

これらの問題を解決するために、われわれは類型を平均類型からの逸脱度合いで表現する情報量規準を提案する。この規準において検査項目 \( a_i \) のカテゴリ \( r_m(a_i) \) の重要度は、情報量 \( -\log_2 p(r_m(a_i)) \) で表される。われわれが実験した条件下では、N 以外のカテゴリの生起確率は N から遠いほど小さくなると、\( p(r_m(a_i)) \) が小さいほど情報量 \( -\log_2 p(r_m(a_i)) \) が大きいことから、上記の要求に満たされる。類型 \( k_j \) の悪さを表す情報量規準 \( I(k_j) \) は、次式で与えられる。

\[
I(k_j) = \sum_{i=1}^{n(a_i)} \sum_{m=1}^{r_{m(a_i)}} -p(r_m(a_i)|k_j)\log_2 p(r_m(a_i))
\]

\( k_j \) においてカテゴリ N を無視するのは、N は普通であるため。
に注意されず情報量が0と見なせるため、Nの生起確率がHやLの生起確率より小さい場合があるためである。

HSV 表色系は、人間の色表現特性を考慮したカラーパレットの表現方法であり、色相、彩度、および明度を用いる。色相は、赤→黄→緑→青→紫を基調色相とし、これらを中間色と共に円周上に並べた色相環を分ける [5]。われわれは、暗色で寒冷な色よりも日立つことを考慮し、類型を悪性順から赤(0度)、橙(30度)、黄(60度)、黄緑(90度)、緑(120度)、青緑(150度)、青(210度)、紫(240度)に対応させることにした。これららの色相の選択は、主に類型の区別を容易にとし、類型が8種類よりも少ない場合には、適宜見分けにくい色相から省くこととする。

3. 3 表示画面の設計
表示画面の設計の当たる、多様な患者の時系列的な変化が明らかになるように配置することを最重視し、重要情報を表示し表示位置の図面化を目的とした。プロトタイププランでは、患者を行として表示し、検査結果を変数類の線形図を時間軸に沿って配置することにした。患者毎にあまりにも時間幅が異なるため、患者毎の横幅を固定し、時間幅を変化させた。この表示法により、最初の目標が達成されたと考えた。次の2点に関しては、肝生検とIFN情
報を示し、1 検査結果の幅を 5 ピクセルとして実際を実施した。このピクセル数は、類型の見やすさと検査結果の重なりのトレードオフを考慮して決定した。

4. 慢性肝炎データの可視化

4. 1 表示方法
慢性肝炎の性質および、肝生検の前500日後500日に線形化程度の変化の反対である。したがって、この期間においては、各患者の線形化程度が比較的確実に把握できると考えた。肝生検を受けて線形化程度が分かっている全患者について、この期間の検査結果から2節の手法で類型を導出した。検査項目は上表の17種のEM法のランダムリスタート数は100とした。類型数は、類型間の相関度と類型の生起確率を考慮し、3.8とした。なお2.3節の手法を変更し、多数の検査を受けた患者の影響を減らす工夫を行った。

各カテゴリの平均値（average）と学習された8個の類型（1, 2, 4, −, 8）を図2に示す。8個の類型は左から右に、良好かつ悪い順に並列されている。図において平均値と各類型は17種類の検査項目で構成されており、各検査項目は縦方向の割合グラフで示されている。類型の後には求心性の割合を、類型の生起確率である。各カテゴリの平均値よりGOT, GPT, TTG, ZTT, ALB, CHE, PLT, および MCV と多くの検査項目が異常なカテゴリーになりやすい。これに比較すると、類型6は一見異常なカテゴリーより少なく見えるが、D-BIL, L-BIL, および T-BIL が異なる点が目立つ。類型7は血清関値の検査項目が、類型8はほぼ全ての検査項目が異なる点が目立つ。図3に、F4患者の肝生検の前後各500日におけるプロトタイプレビューを示す。図では1行が1患者、1本の棒が類型の線形図を示される1回の検査結果を表す。図より、各患者の類型の変化が検査時期と共に経日に把握できることが分かる。プロトタイププランでは、全体的に黒化する患者に関して検査値のグラフを検査

した所、検査結果のカテゴリーや検査値が総合的に判断されて可視化されていることが分かった。プロトタイププランは特に、先入観や不注意による見落としを気づかせ、一見異なって見える類似患者を特定するのに有効であることが分かった。

4. 2 多角的な分析
プロトタイププランでは、既に発見された類型を用いて、任意の期間のデータを表示することが可能である。例えば図4は、図2の類型を用いた患者602の全検査結果に関するプロトタイプビューを示す。この図より、図3においては問題が少ないと推定される患者602の画像は、肝生検の約2年後から急速に悪化したかことが分かる。

全ての検査項目を用いて類型を発見する方法は、多数の患者を同時に見られるが、検査項目をより細かく分けて多角的に分析したいと要求がある。この要求に対応するために、図5, 6に示すように検査グループ毎に類型を求める患者を表示する。マルチプロタイププランも構築した。図6より、患者763は検査項目TTTとZTTが特に悪く、ALB, CHE, TP, およびT-CHOなども悪いことが分かる。

5. 関連研究
これまでの可視化手法は、大量のデータを表示するために多種の情報表示を用いたものに分類される。プロトタイププランは、学習された類型を用いて情報抽象を要約することにより、この問題に対処する新しいアプローチに属する。その設計動機は、ユーザの負担を最小限に取るために、学習手法によって得られた本質的な情報だけを、主に色相という情報媒体だけを用いて示すことにある。


図3 肝生検の前後各500日におけるF4患者のプロトタイプライン。元のカラー画像は[9]

Fig. 3 Prototype Lines of F4-patients' data before 500 days and after 500 days from biopsy. The original figure in color can be found in [9].

するアプローチが有望であると考えた。プロトタイプラインは、情報の抽象化をEM法に基づく確率的クラスタリングで行っている。このことに加えて、型への適切な色相の割り当てや表示法、および医療データの解析に試行錯誤的に取り組んだ結果、4.2節に示す慢性肝炎データにおける成功が示されたと考える。

データマイニングにおいて、パターンの抽象度は事例の情報を省略する度合いを表すと見なすことができる。直観的には、パターンの抽象度が高ければ、そのパターンはより多くの事例を説明する。例えば属性数が少ないルールは、パターンの抽象度が高い。津本と鈴木らは事例発見[7]などを通じて、専門家が理解しやすいパターンの抽象度は、属性数が少ないルールよりも低い場合があることに気づいた。さらに慢性肝炎データに関する解析では、医師はルールに比較して、グラフなど該
図4 患者602の全検査結果に関するプロトタイプライン。元のカラ－画像は[9]
Fig. 4 Prototypelines of whole test results of the patient 602. The original figure in color can be found in [9]

図5 検査グループ毎の類型。元のカラ－画像は[9]
Fig. 5 Prototypes of grouped tests. The original figure in color can be found in [9]

図6 患者763に関するマルチプロトタイプライン。元のカラ－画像は[9]
Fig. 6 Multi-prototypelines of the patient 763. The original figure in color can be found in [9]

謝辞

本研究の一部は、文部科学省科学研究費特定領域研究「アクティブマイニング」の援助を受けている。

文献

当する患者が分かりやすい表現を好むことが観察された。プロトタイプラインは、ルールよりもパターンの抽象度が低いため該当する患者が分かりやすく、グラフよりもパターンの抽象度が高いため多くの情報を一度に見られるという利点がある。今後は、領域知識を併用した可視化、可視化手法の定量的評価、およびプロトタイプラインの他の問題への適用などに取り組んで行きたい。