知 能 と 複 雑 系 128−11
（２００２．５．２３）

波形パターンを分類クラスとするルールの発見支援システムの構成法 —慢性肝炎データセットを対象にして—

畑澤 寛光，　佐藤 芳紀，　山口 高平

静岡大学 情報学部

本稿では，多属性で欠損値を持つ未整備な医療データセットからルール発見支援システムを構築する方法について議論する。ドメイン特有の前処理（例えば異なる表記だが同じものを表すデータを統一する）を行った後，大量の属性の中から出現頻度の高いデータを選び，異なる検査期間を統一し，EM アルゴリズムによってクラスタリングされた時系列データセットを使い，決定木学習によってルールを学習した。学習されたルールを見ると，それらに興味深いルールがいくつかあったとの医者の評価を得ることができた．

Rule Discovery Based on Sequential Pattern Analysis and Mining —In the Case Study of Chronic Hepatitis Datasets—

Hiromitsu Hatazawa,　Yoshinori Sato,　Takahira Yamaguchi
Faculty of Information, Shizuoka University

Here is discussed how to build up a rule discovery support system to ill-defined medical datasets with many attributes and missing values. After having done domain-specific pre-processing such as unifying different names to the same entities, reducing many attributes to relevant ones with high occurrence and unifying different inspection periods, we have got the discretized version of time-series medial datasets by taking EM clustering and learned rules by decision-tree learning. Looking at the learned rules, medical experts have told us that there were some rules that are interesting to them.

1. はじめに

近年，Evidence Based Medicine (EBM)[*1] と呼ばれる医療行為の実践とデータマイニングとの関連に大きな関心が寄せられている．しかしながら，医療現場で構築されるデータベースは，一般的に，時系列・多属性・未整備—など実世界のデータの持つ典型的な特徴を備えているため，知識発見に際しては十分な前処理が要求されるが，現時点では，データマイニングの手法を実際に適用する上でのノウハウの蓄積・方法論の整備は十分とは言えず，依然として試行錯誤的な作業が行われており，その支援環境の開発が望まれている [津本 99]．

以上の背景より，本研究では，千葉大学病院から提供された B 型・C 型ウィルス性慢性肝炎患者の経過追跡データを分析する機会を得たので，上述した問題点を解決するために，進行の度合いを示す血液データ（GPT）と検査データとの相関関係を発見し，検体検査データから予後因子[*2]を同定することを目標として，時系列データに対するデータ前処理／知識発見支援機構の開発，及び実験・評価を行った．

以下，知識発見支援システムの概観，実施した前処理，評価実験について述べる．

2. システム構成

ルールを帰納する機械学習スキームは一般化の対象として，目標概念とその概念を記述する属性からなる事例の集合を入力として受理する．予後の同定を目的とする場合，クラスは検査時よりも将来の症状を表す名義値として与えられる．また，属性はクラス設定時より以前の検査データの傾向を記述するものとして与えられる．クラスには症状の中／長期的トレンドを与えることが直感的である．このような事例は，時系列データから一定の長さのシーケンスを切り出して，クラス設定時を基準に相対化したデータを属性として与えることで構築できる．シーケンスの記述方法には，特徴ベクトルを求める方法や，パターンを求めて離散化する方法などが存在するが，本研究ではパターン化の手法を採用した．

シーケンスを特徴ベクトルで表現する手法では，例えば，一変量データの要約統計量を特徴量として使用できる．名義尺度データについては最頻値や度数分布などの変数，順序・間隔・比例尺度データについては最大値・最小値・中央値・分散・四分偏差・歪度・尖度などの変数が利用可能である．これとは別に，パワースペクトル解析により波形の中に含まれる周期変動成分を分離し，それぞれの強さを属性に与えることもできる．これらの属性はシーケンスそのものを記述するわけではないため可読性の点に問題がある．また，特徴量の選択は発見的にならざるを得ないため，本質的に次元数の増加をもたらす．次元数を圧縮するためには主成分を抽出するなど副次的な作業を実施しなければならない．一方，シーケンスから典型的なパターンを求めてその識別子を名義属性として与える手法では，シーケンスはパターンへと変換されることで一定の抽象化が行われる．抽象化の度合いはパターン数で制御できるため，妥当な数を設定することで過適合を回避できる．また，特徴ベクトルとは異なり，シーケンスの表面的な変化を記述するため直感的で解釈しやすい．

前処理は，一般的に，目標概念を同定する上で十分な属性が与えられている場合と必要な属性が与えられていない場合に分類できる．後者のケースは，当初属性／レコード間に何らかの関係性が規定されているがそれが明示的に与えられていない場合に相当する．本研究が対象とする時系列データでは後者の処理を要求するが，本論文では，その中でも，領域知識を必要と

*1　科学的根拠に基づいた医療
*2　治療がうまく行くかどうかを予想するのに利用される要因

しないパターンの抽出とシーケンスの離散化以降の処理とそれ以前の処理を，それぞれ “高水準の前処理”，“低水準の前処理” と呼び区別する．

図 1 は慢性肝炎データセットに適用した知識発見支援システムの概観である．初期データセットには膨大な数の表記揺れが含まれており，また，非常に多くの検査項目が存在しているので，最初に前処理としてデータの洗浄と検査項目の選択を行う．次に各検査項目の検査周期が一定になるように検査周期の均一化と補完を行う．続いて，時系列データからシーケンスを切り出し，それらにクラスタリング・スキームを適用してパターンへと落とし込むことで離散化する．そして，離散化されたデータにデータマイニング・スキームを適用してルールを発見する．最後に，後処理として，シーケンスを離散化する際に求めたパターンの情報を用いて発見されたルールをグラフに変換し，ユーザに提示する．

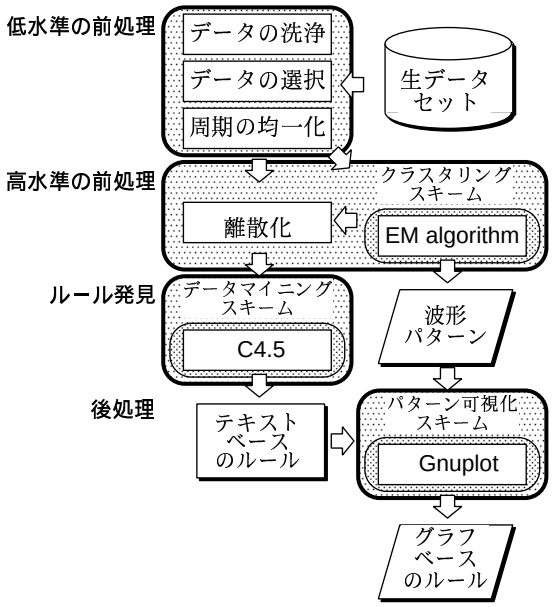

図 1: 知識発見支援システムの概観

3. 前処理

本章では，提供されたデータセットの内容と本データから知識発見を行うために実施した低水準の前処理と高水準の前処理について述べる．

3.1 データセットの概要

全部で 5 種類（pt・laboname・labo・bio・ifn）のデータが提供された．このうち，laboname は検体検査結果情報（labo）のメタデータであるため本研究では分析対象から除外した．

分析対象データの概要は以下の通りである:

- 患者基本情報: pt
  患者の性別・生年月日を提供する．レコード数は 771 で 3 種類の属性から構成されている[*3]．

- 検体検査結果情報（主に血液・尿）: labo
  患者の検体検査の結果情報を提供する．レコード数は 1,597,146．検体検査結果情報は院内データと外注データに分けられ，最大で 11 種類の属性から構成されている[*4]．このうち,「負荷名」「判定結果」「その他」の属性は欠損値が多くマイニングに適さないため,「単位」は知識発見には直接関係しないためそれぞれ削除した．また，院内データの「検査結果」に対して外注データの「検査結果値」を対応付けた．さらに，検査項目のうち，特に「血液型検査」については患者ごとに通常 1 回しか行われず，結果も変化しないことから独立したデータセットとして扱うことにした．

- 肝生検情報: bio
  患者の肝生検[*5]の情報を提供する．レコード数は 960 で 8 種類の属性から構成されている[*6]．このうち,「検体管理番号」は ID なので削除した．また,「採取施設」も知識 発見に寄与しないため削除した．検査結果についての分類法が数年前から変わったため，新しい分類法である「繊維化」,「活動性」の値がほとんど欠損している．相当数の表記揺れが見られる．

- インタフェロン投与情報: ifn
  インタフェロン[*7]を投与した患者の情報を提供する．レコード数は 198 で 4 種類の属性から構成されている[*8]．このうち,「投与開始日」は今回提供されたデータではすべて “1 回目” なので削除した．

3.2 低水準の前処理

3.21 データの洗浄

提供されたデータには相当数の表記揺れが存在しているため以下の方針でデータの洗浄を行った:

- 検体結果情報について
  - 急性肝炎のデータは分析対象外なので削除する．
  - 検査結果値末尾の記号を除去する．
    e.g. $0.982H \rightarrow 0.982$
  - 検査結果値の表記方法を統一する．
    e.g. $20-30 \rightarrow 25$, $10*4.5 \rightarrow 45$
  - 検査結果の大部分が数値で表現されるものについては名義値を欠損値にする．
    e.g. ヨウケツフカ，ニュウリョクミス，キャンセル，ケンサフカ $\rightarrow$?
  - 検査結果の大部分が名義値で表現されるものについては結果を 2 値化する．
    e.g. { ヨウセイ，+，*etc.*} $\rightarrow +1$, { インセイ，−，*etc.*} $\rightarrow -1$

- 肝生検情報について
  - 生検結果の名称を大枠で統一する．
    e.g. *LC+Hemangioma of the liver* $\rightarrow$ *LC*（肝硬

*3 「患者 ID」「性別」「生年月日」

*4 「患者 ID」「検査日」「検査項目名」「負荷名」「検査結果値」「単位」「判定内容」「その他」「コメント」「結果評価」「結果項目の子コード」

*5 肝臓の組織を採取して顕微鏡で調べる検査

*6 「患者 ID」「検体管理番号」「肝炎型」「生検年月日」「生検結果」「採取施設」「生検情報（繊維化）」「生検情報（活動性）」

*7 ウィルス性肝炎の特効薬

*8 「患者 ID」,「投与回数（N 回目）」,「投与開始日」,「投与終了日」

変）
e.g. *CAH with briding necrosis* → *CAH*（活動性肝炎）
e.g. *CPH+Slcoidosis* → *CPH*（持続性肝炎）
e.g. *AH(Subacute)* → *LH*（急性肝炎）

– 急性肝炎のデータを削除する．
– 活動性の表記を統一する．

- その他データセットについて
  – インタフェロン投与情報で「投与開始日」と「投与終了日」が逆転しているレコードを削除する
  – 検体検査結果情報・肝生検情報・インタフェロン投与情報で値が “0” のみ，あるいは改行コードのみの行を削除する
  – 「生年月日」「検査日」等の年表記を統一する．
    e.g. *“YYMMDD”* → *“YYYYMMDD”*
  – 項目がずれていたり，連結しているレコードを修正する．

3.22 検査項目の選択

初期の検体検査結果情報には 957 種類（15,94,390 レコード）の検査項目が存在する．分析対象外の急性肝炎のデータを取り除いても依然 930 種類（1,549,299 レコード）もの検査項目が存在している．本章では入力事例の次元数—検査項目数—を減少させるための方針を述べる．

ほとんどの機械学習アルゴリズムには属性を選択する能力があるが，現実には冗長，あるいは無関係な属性は学習に深刻な影響を及ぼすことが知られている [John 94]．このため，一般的にはクラスと相関の強い少数の属性を選別するために属性選択を絡めた学習が行われる．不適切な属性を削除してデータの次元数を減らすことで学習アルゴリズムの性能は改善するが属性選択は計算集約的である．本研究では出現頻度に基づいて検査項目を剪定した後，属性選択メソッドを併用してルールの発見を行った．

図 2 は検査項目数あたりのレコード数の累計である．全 930 項目のうち出現頻度の多い上位の約 100 項目で全データの 99% 近くを占め，その後は漸近的に増加している—残りの検査項目はほとんど寄与していない—ことが分かる．最も出現数が多い項目で 46,000 回程度である．

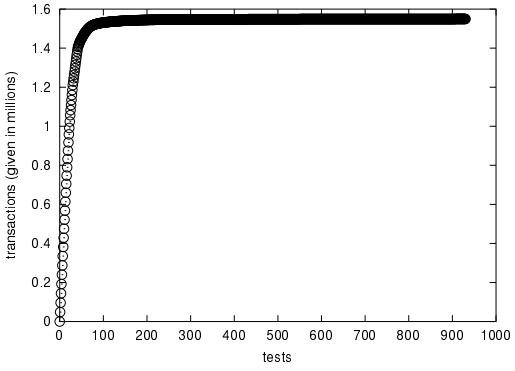

図 2: 検査項目数あたりのレコード数の累計

極端に出現頻度を少ないものを除けば，これらの項目は希少事象を被覆するルールを生成しうるという点で潜在的に有用である．しかしながら，希少事象は頻度の少なさから母集団を正確に代表しにくく過適合を招きやすい上，検査頻度の少ない検査項目は医者によるバイアスを受けていることが多い．したがって，本研究では出現頻度のみを基準に検査項目を選り分けるものとし，背景知識や緻密なデータの分析に基づいた検査項目の選別は実施しない．

実験では簡単に出現頻度が 10000 回未満の検査項目を取り除いた．全 930 種類ある検査項目のうち 888 種類，全データの約 1 割 187,870 レコードがこれにあたる．したがって，残りの 42 種類の検査項目を対象にマイニングを実施した．

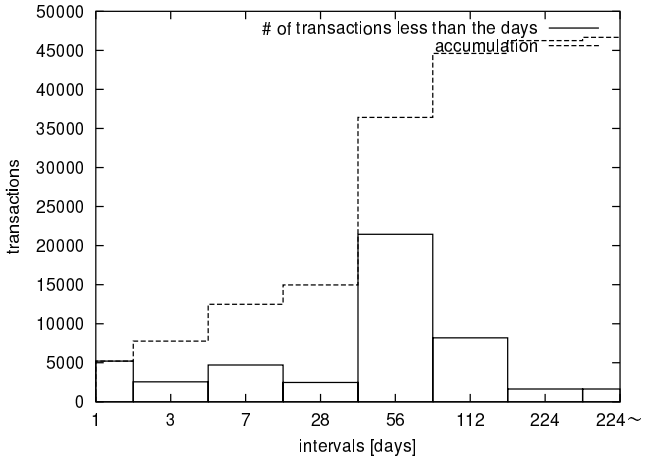

図 3: 検査周期ごとのレコード数

3.23 検査周期の均一化と補完

シーケンスを一般化するためには観測の行われる間隔を一定に保つ必要があるが，その期間の長さをどのように設定するかが問題となる．一度，長さを決定するとそれよりも短い，あるいは長い周期で行われている検査の情報のほとんどは利用できなくなる．

図 3 は一定の検査周期を設けてその頻度をヒストグラムで表したものである．これによると 28 日以上 56 日未満に再検査が行われるケースが最も多い．本研究では事例数を最大化するような検査周期を選択する．したがって，28 日（1 ヶ月）が最も多いため，この周期を採用した．周期が 1 ヶ月未満，あるいは 2 ヶ月以上のデータについては線形補完を行った．

本研究では以下の手順で検査周期を均一化した．まず，患者 ID 毎に検査日でソートされた検査記録を順に読み出して最初の検査から 1 ヶ月未満に行われた検査をすべてマージする．最初の検査から 1 ヶ月以上離れた検査が現れたところでマージ結果から平均値を求めて検査日をその期間の中心（最初の日付から 14 日後）の日付に設定して検査記録を出力する．このとき 2 ヶ月以上離れた検査が現れた場合はその期間に応じて空の検査記録を出力する．あらかじめ指定した補完区間数以上離れているか患者 ID が異なる場合は補完を行わず新しい系列として処理する．

検査周期の均一化が終わった段階ではまだ挿入した検査記録は空なので次のパスで補完処理を行った．まず，各検査項目ごとに検査記録を順に読み込んで欠損している箇所，あるいは区間があればその前後の検査記録の値から移動平均値を挿入する．このときあらかじめ指定した補完区間数以上欠損している区間が続く場合は補完を行わない．これは補完する区間が長くなるほどその精度が悪化するためであり，本研究では最長でも 3 区間（3 ヶ月）の補完に留めた．

3.3 高水準の前処理

3.31 時系列データの離散化

以下のような手順で時系列データを離散化した．この手順は Das[Das 98] が時系列データからのルール発見に示したフレームワークに則る．

最初に時系列データから切り出すシーケンスの長さを与える．我々はこれをウィンドウ・サイズと呼び，w で表す．シーケンス $s = (x_1, \ldots, x_n)$ が与えられているときウィンドウ・サイズ w で切り出されるサブシーケンスは $s' = (x_i, \ldots, x_{i+w-1})$ となる．シーケンス s の先頭から 1 ずつスライドさせることサブシーケンスの集合 $W(s) = \{s_i | i = 1, \ldots, n - w + 1\}$ を求める．

次にクラスタリング・メソッドを用いて各サブシーケンスを離散化する．本研究では EM アルゴリズムを適用するため，ここでは各サブシーケンスが属する分布（母集団）を考える．全サブシーケンスの確率分布（尤度）を最大化するようなパラメータを推定することでクラスタリングを行い，$W(s)$ をクラスタ集合 $C_1, \ldots, C_k$ に変換する．それぞれのクラスタ C_h に対して記号 a_h を与える．したがって，すべてのサブシーケンスの離散化集合は $D(s) = a_{h(1)}, \ldots, a_{h(n-w+1)}$ となる．a_h はサブシーケンスの時系列変化を表すプリミティヴな形状に対応付けられた名義値であり，目標概念の記述言語に使用される．

時系列データの離散化の手順を図 4 に示す．図 4(a) の破線のグラフはクラスとなるシーケンス，その他のグラフは属性となるシーケンスを表している．サブシーケンスはあらかじめ与えられたウィンドウ・サイズで切り出される．クラスには属性として切り出されたサブシーケンスの直後を起点とするサブシーケンスが与えられる．図 4(b) では (a) で切り出したサブシーケンスをクラスタリングしてパターンに落としている．図 4(c) の表はクラスタリング結果をもとに構築したデータセットを表している．属性値とクラス値にはそのシーケンスの典型的な形状に相当する名義値が与えられている．

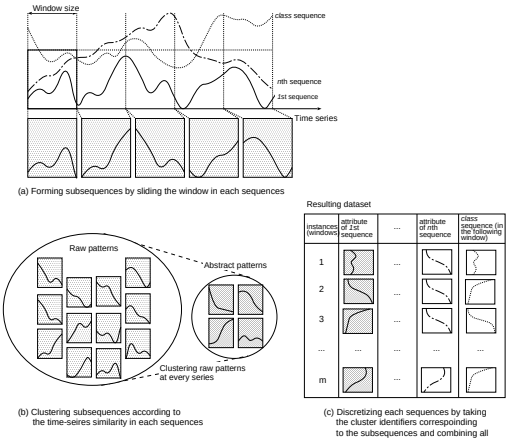

図 4: クラスタリングによる時系列変化の離散化

3.32 クラスタリング・アルゴリズムの選択

本研究ではシーケンスのクラスタリングに不完全データからの学習アルゴリズムである **EM アルゴリズム**を採用した [赤穂 96]．

k-means に代表されるクラスタリング・アルゴリズムは有限個の根拠から事例を特定のクラスタに決定的に配置するため，単純な問題を除けば，過適合しやすく局所的最適解に陥りやすい [Hartigan 75]．また，類似性の指標にはユークリッド距離の他にも音声認識で使用されている DTW[*9] や LCS[*10] など様々なものが提案されているが [Gunopulos 00]，初期決定に解が左右されるというアルゴリズムの欠点を補うものではない．一方で EM アルゴリズムは大域的最適解への収束性に優れていることが経験的に知られており，欠損値やノイズに対しても頑強である．

EM アルゴリズムは正規混合分布モデルに基づいている．このモデルではクラスタは正規分布として表現される．K 個の正規分布 $N(\mu_k, \sigma_k^2), (k = 1, \ldots, K)$ があるとして標本 y はこのうちの 1 つから与えられるものとする．k 番目の正規分布が選ばれる確率は π_k であるとする．標本が与えられたときにパラメータ π, μ, σ を推定することが問題となる．y の分布は次のように表される:

$$g(y|\pi, \mu, \sigma) = \sum_{k=1}^{K} \pi_k \phi(y|\mu_k, \sigma_k),$$
$$\sum_{k=1}^{K} \pi_k = 1$$

ここで ϕ は正規分布の密度関数

$$\phi(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

である．もし標本が何番目の正規分布から生成されていたかがすべて分かれば問題は自明となる．EM アルゴリズムを適用するために，その番号 z を含めたものを完全データとし，y を不完全データとみなす．完全データ (y, z) の分布は次のように書ける:

$$f(y, z|\pi, \mu, \sigma) = \pi_z \phi(y|\mu_z, \sigma_z)$$

また，独立にこの分布に従う N 個の完全データ $(y_1, z_1), \ldots, (y_N, z_N)$ が与えられたときの対数尤度は

$$\sum_{i=1}^{N} \log f(y_i, z_i|\pi, \mu, \sigma) = \sum_{i=1}^{N} \log(\pi_{zi} \phi(y_i|\mu_{zi}, \sigma_{zi}))$$

となる．

EM アルゴリズムはパラメータ ξ をある適当な初期値に設定し，**E ステップ** (Expectation step) と **M ステップ** (Maximization step) と呼ばれる二つの手続きを繰り返すことにより ξ の値を逐次更新する方法であり，次のように定式化される:

1. パラメータの初期値を適当な点 $\xi = \xi^{(0)}$ に設定する．
2. $p = 0, 1, 2, \ldots$ に対して次の二つのステップを繰り返す:
   (a) E ステップ: 完全データの対数尤度 $\log f(x|\xi)$ の，データ y とパラメータ $\xi^{(p)}$ に関する条件付き平均を求める．すなわち，
   $$Q(\xi) = E[\log f(x|\xi)|y, \xi^{(p)}]$$
   を計算する．
   (b) M ステップ: $Q(\xi)$ を最大化する ξ を ξ^{p+1} とおく．

---

*9 Dynamic Time Warping
*10 Longest Common Subsequence

3.4 時系列データを除く属性の構築

時系列データを離散化することでクラスと属性を与えた．しかしながら，その他にも患者のプロフィール・肝生検・インタフェロン情報が利用できるため，最終的にこれらの属性に適切な処理を施して追加することで最終的なデータセットを構築した．

まず，患者基本情報から患者の「性別」「年齢」を追加した．「年齢」についてはクラス設定時を基準にして「生年月日」から算出した．

次に，インタフェロン情報からサブシーケンスの時期における患者のインタフェロンの「投与状況」を属性として与えた．この属性には 3 つの値 “未投与” ・“投与中” ・“投与済み” を設定した: サブシーケンスの開始時が「投与開始日」以前であれば “未投与”，サブシーケンスの開始時が「投与開始日」と「投与終了日」に挟まれている場合には “投与中”，サブシーケンスの開始時が「投与終了日」以降であれば “投与済み”，インタフェロン投与情報が無い患者の場合には欠損値を与えた．

最後に，生検結果情報から患者の「生検結果」を追加した．肝生検は血液や尿を調べる検体検査と違って患者一人あたりの検査回数は非常に少ないため，時系列データとしてパターン化することはできないのでクラス設定時の近傍（前後 1 年以内）の生検結果の値を属性値として与え，存在しない場合には欠損値とした．

クラス属性には検体検査項目「GPT」のパターンを与え，それぞれ直前の短・中・長期的の検査データのパターンから予測を行った．

4. 実験

データ前処理手法・属性構築手法を実際に提供されたデータに適用し，実験用のデータセットを作成した．

本研究で時系列データから構築される属性はある期間のデータの特徴を一般化したものである．検査項目ごとに同時期の時系列データから切り出した一定の長さのサブシーケンスを，EM アルゴリズムにより同定した混合分布モデルにしたがってパターンに落とし込んで対応する名義値に置換したものと，そのシーケンスの移動平均値をパターンの基準値として属性に与えた．また，パターンのウィンドウ・サイズには 3 種類の長さ—短期（6ヶ月）・中期（12ヶ月）・長期（24ヶ月）—を与えた．また，全部で 42 種類の検査項目が利用可能であるため，データセットの属性数は最大のもので 240 種類の検査項目属性に肝生検・インタフェロン等の属性を加えた 250 種類程度となった．

クラスには血液データの中から肝機能を推定する上で有用な指標として利用されている GPT を採用し，各検査項目の直前の変化パターン，及び患者基本情報・肝生検情報等から将来の GPT の変化パターンの同定を試みた．今回はパターンを 8 つに設定してそれぞれのパターンを短期・中期・長期の 3 種類のウィンドウ・サイズで同じように同定し，属性とした．

4.1 ルールの評価

図 5-図 8 のルールは実験により発見されたルールの一例である．これらのルールについて専門医から評価を得ることができた．

学習システムにより発見されたテキストベースのルールはすべてグラフベースのルールへと変換した．クラスタリング時に得られた各クラスタの中心座標の情報，パターンの乖離（最大／最小値の差），基準値（移動平均値）の情報をすべて単一のグラフにプロットして出力した．グラフ中央付近の x 軸が 0 となる位置が現在，正の方向が未来，負の方向が過去をそれぞれ表している．ルールに含まれているパターンは波形，基準値は y 軸の中心付近を通る x 軸に平行な直線で表現されている．パターンの乖離の大きさ・基準値の値はそれぞれグラフ右上の凡例に記載されている．尚，パターンについては表示の都合上，分散の大きさが 1 になるようなスケーリングを施してある．

まず，図 5 のルールについては次のようなコメントを頂いた:

> I-BIL（ビリルビン）が高い状態は肝硬変の症状が進んでいることを示すものであり，このルールは直前の 24ヶ月で I-BIL が減少すると GPT が上昇に転じる，という意味に解釈できる．

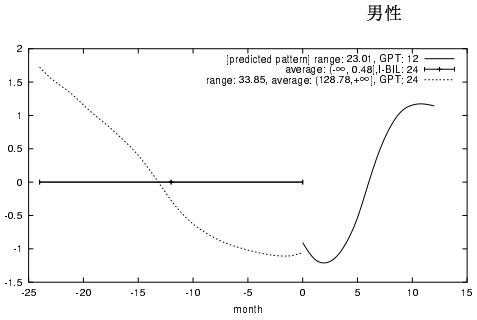

図 5: $precision$: 70.00% (18/25), $recall$: 1.49% (18/1208)

図 6 のルールは「直前 24ヶ月のビリルビンの平均値が高い値を維持し，かつ，TTT が減少すると GPT が減少に転じる」という意味を表す．このルールについては前述のルールと併せて次のようなコメントを頂いた:

> 医者の感覚では，GPT の値は，ほぼ上昇-下降の周期的な変化を繰り返し，多少の上がり下がりはあるものの，ほぼ一定であると理解されてきた．このルールは GPT の値が上昇から下降，あるいは下降から上昇へと転じる状況を説明するものであり大変興味深い．ウィルスの活動・バクテリアの増殖に周期性があるのか，また，その周期性はウィルスの種類により異なるのか，など，5 ～ 10 年の期間で周期性・法則性の示唆ができれば興味深い．

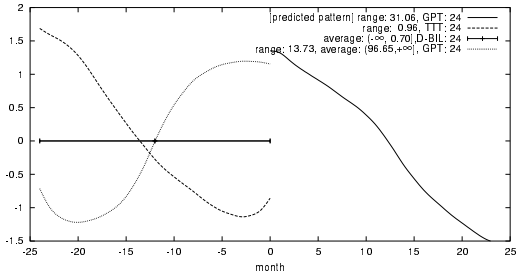

図 6: $precision$: 60.90% (21/34), $recall$: 1.43% (21/1470)

図 7 のルールについては次のようなコメントを頂いた:

> 乳ビはコレステロール，TG（中性脂肪）が高いことに相当するため，このルールは慢性肝炎がコ

レステロールとの関連性が高いことを示唆している．また，乳ビは血液の濁り具合を示す医師による主観的な指標であり，ルールに現れてくるのは興味深い．

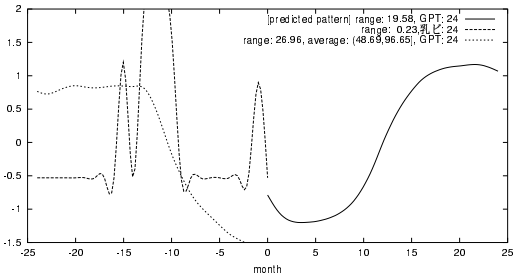

図 7: *precision*： 68.40% (17/24), *recall*： 1.41% (17/1203)

最後に，図 8 のルールについては次のようなコメントを頂いた:

> 医者の考えと似ているが 6 ヶ月も遅れて GPT が下がるのは意外である．医者の感覚では TTT が下がってから 1, 2 ヶ月後くらいで下がり始めるのが妥当であるが，このようなケースがあってもおかしくはない．もう少し，再現性が高ければ妥当な結果といえるかもしれない．

11 歳以下

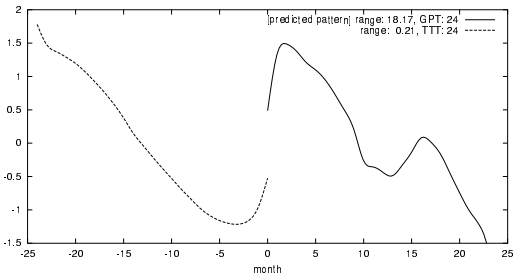

図 8: *precision*： 85.70% (5/5), *recall*： 1.06% (5/470)

時間的制約から，本実験で発見されたルールのうち，実際に医師による評価を得ることができたものは極一部であったが，医学的常識から外れるようなルールは比較的少なく，どちらかといえば医師の経験的知識に近い結果が得られたようである．

5. おわりに

本研究では，ウィルス性慢性肝炎データを対象に，時系列データに対するデータ前処理／知識発見支援機構の開発，及び実験・評価を行った．特に，検査データから予後（GPT）の推定を試みた実験では，肝炎を引き起こすウィルスの周期性について専門医の仮説生成を支援する上で有用な知見を得ることができた．

今後は今回の実験で得られた評価から，肝炎の周期性，及び法則性について有益な示唆をもたらすルールの発見を目指し，手法の改良・実験を行っていく予定である．時系列データはパターンを求めて離散化することでシーケンスデータへと一般化することが可能であるが，これによりシーケンスデータを対象にした様々な属性構築手法を時系列データにも適用することができるようになる．これらの手法を取り入れることも有用であろう．

また，医師から提示されたターゲット—インタフェロン治療例を類型化し，各群を特徴付けるデータの発見する—についても研究を開始する予定である．今回の実験では，主にルーチン検査と呼ばれる比較的出現頻度の多い検査項目のみを属性として使用したが，実験結果の評価の過程で医師からルーチン検査以外の検査項目を使用した実験についても強い関心が示された．この点についても，今後の研究で分析を重ねて，カバーしていきたいと考えている．

6. 謝辞

本研究において実験データを提供，ならびに実験結果を評価していただいた千葉大学病院医療情報部 高林克日己・横井英人 医師に深く感謝します．

また，本研究は，文部科学省平成 13 年度科学研究費特定領域研究 B(13131205) によるものである．

参考文献

[Cheeseman 88] Cheeseman, P., et al.: “A bayesian classification system”, In Proceedings of the Fifth International Conference on Machine Learning. Morgan Kaufmann (1988).

[Das 98] Das, G., et al.: “Rule Discovery from Time Series”, Proceedings of KDD-98 (1998).

[Gunopulos 00] Gunopulos, D, et al.: “Time Series Similarity Measures”, KDD 2000 tutorial (2000).

[Hartigan 75] Hartigan, J.A.: “Clustering algorithms”, New York: John Wiley (1975).

[Ian 00] Ian, H.W., et al.: “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”, Morgan Kaufmann Publishers (2000).

[John 94] John, G.H., et al.: “Irrelevant features and the subset selection problem”, In Hirsh, H., and W. Cohen, editors, *Proc. Eleventh International Conference on Machine Learning*, pp.121-129. New Brunswick, NJ. San Francisco: Morgan Kaufmann (1994).

[Quinlan 93] Quinlan, J,R.: “C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers (1993).

[Smyth 92] Smyth, P., et al.: “An Information Theoretic Approach to Rule Induction from Databases”, IEEE Transactions on Knowledge and Data Engineering, 4 (Aug. 1992), 301-316 (1992).

[赤穂 96] 赤穂 昭太郎: “EM アルゴリズムの幾何学”, 情報処理 Vol.37 No.1, 情報処理学会 (1996).

[津本 99] 津本周作: “科学的データベースからの知識発見”, チュートリアル JSAI ’99, pp.21-38 (1999).

[津本 00] 津本周作: “知識発見手法の比較と評価のための共通データ”, 人工知能学会誌 Vol.15, No.5, pp.751-758 (2000).