知能と複雑系　132－21
(2003. 3. 15)

時系列医療データにおけるルール発見支援システム

— 慢性肝炎データセットでのケーススタディ —

大崎　美穂[†]　佐藤　芳紀[†]　北口　真也[†]　横井　英人[††]　山口　高平[†]

† 静岡大学情報学部, 静岡大学大学院情報学研究科　〒432–8011 静岡県浜松市城北 3–5–1
†† 千葉大学医学部附属病院医療情報部　〒260–8677 千葉県千葉市中央区亥鼻 1–8–1
E-mail: †{miho,cs8037,cs9026,yamaguti}@cs.inf.shizuoka.ac.jp, ††yokoih@telemed.ho.chiba-u.ac.jp

あらまし　本研究は慢性肝炎の検査データを対象とし，医学的に有用なルールを得ること，前処理，後処理，およびシステムと専門家のインタラクションに関する知見を得ることを試みた．システムには，クラスタリングによるパターン抽出と決定木による分類の組合わせを用い，病状を推定するグラフベースのルールを生成した．そして，専門家とシステムのインタラクションの繰り返しを通し，ルールに基づいて，GPT に関する仮説の形成と検証ができた．最後に，これまでに得た知見に基づき，インタラクションの概念モデルと半自動化への指針を示した．

キーワード　時系列，医療データ，システムと専門家のインタラクション，ルールの洗練化

A Rule Discovery Support System for Sequential Medical Data

— In the Case Study of a Chronic Hepatitis Dataset —

Miho OHSAKI[†], Yoshinori SATO[†], Shinya KITAGUCHI[†], Hideto YOKOI[††], and Takahira YAMAGUCHI[†]

† Faculty of Information, Shizuoka University, Graduate School of Information, Sizuoka University
Johoku 3–5–1, Hamamatsu-shi, Shizuoka, 432–8011 Japan
†† Department of Medical Informatics, Chiba University Hospital
Inohana 1–8–1, Chuo-ku, Chiba-shi, Chiba, 260–0856 Japan
E-mail: †{miho,cs8037,cs9026,yamaguti}@cs.inf.shizuoka.ac.jp, ††yokoih@telemed.ho.chiba-u.ac.jp

Abstract　This research aims to obtain medically vaulable rules and knowledge on pre-/post-processing and the interaction between system and human expert using the data of medical test results on chronic hepatitis. We developed the system based on the combination of pattern extraction with clustering and classification with decision tree and generated graph-based rules to predict prognosis. As the result, the human expert could make and justify the hypothesis on GPT through the iterative interaction. Finally, this paper shows the conceptual model of interaction and discuss on how to systemize the semi-automatic interaction.

Key words　Time Series, Clinical Data, Interaction between System and Human Expert, Rule Refinement

1. まえがき

近年，Evidence Based Medicine (EBM，科学的根拠に基づく医療) へのデータマイニングの貢献に大きな関心が寄せられ，データマイニング技術を医療データに応用した様々な研究が行われている [7], [13]．しかし，EBM のためのデータマイニングには，次の 2 つの問題が残されている．

1 つ目は，データの前処理とルールの後処理に関する問題である．一般に，前処理・後処理は，マイニングスキーム以上にマイニング性能に大きな影響を及ぼす [1], [14]．特に，医療データは構造化が不十分な上に，数値と名義値の混在や検査項目の依存関係といった問題を持ち，領域知識を反映した複雑な前処理を必要とする．また，得られたルールは未知の現象を表すだけでなく，医学的な観点で解釈され得るべきである．したがって，後処理では，ルールの表現方法や他の情報との組合せを工夫し，専門家の思考をサポートする必要がある．しかし，領域

知識を前処理・後処理に反映するための知見や方法については，あまり研究されていない [8], [14], [19].

2つ目は，システムと専門家とのインタラクションに関する問題である．システムと専門家の情報交換がシステムの改善と興味深いルールの発見に結び付くことは，データベースからの知識発見プロセスとして，よく知られている [2]．特に医学分野では，インタラクションを通して，専門家が持つ高度で莫大な知識をシステムの前処理・後処理に反映する必要がある．しかし，インタラクションが，なぜ，どのように，システムの改善と興味深いルールの発見に寄与するかを調べた研究は少ない．

以上より，データの前処理とルールの後処理，およびシステムと専門家のインタラクションに関する体系的な方法論の確立が必要と考えられる [14]．しかし，ケーススタディの積み重ねなく，これを達成することは困難である．そこで，本研究は最初のステップとして，慢性肝炎の検査履歴のデータを使ったケーススタディを通し，**目的1** 医学的に有用なルールを実際に得ること，**目的2** 前処理・後処理とインタラクションに関する知見を得ることを試みる．そして，最終的に，前処理・後処理の最適化と専門家の創発を支援するインタラクションのモデルと，その半自動化への指針を示す．

本論文では，2. において，本研究の問題解決のアプローチを説明する．3. では我々が提案するシステムの構築について，4. では，慢性肝炎データセットに提案システムを適用した結果を示し，ルールの有用性，今回行った前処理・後処理とインタラクションの妥当性，これらに関して得られた知見を検討する．5. では，得られた知見の体系化を試み，インタラクションについて，概念的レベルでのモデルと半自動化のフレームワークを示す．最後に 6. で，まとめと今後の展開を述べる．

2. 問題解決のアプローチ

2.1 先行研究で得られた成果

我々は，先行研究 [6] で得られた成果に基づき，本研究のコンセプトを決め，システム設計を行った．そこで，まず先行研究の概要を述べる．先行研究では，本研究と同じデータセットとシステムを用い，慢性肝炎の病状把握の主要な指標である，GPT という検査項目の傾向を推定するルールを得た．そして，専門家がルールを評価した結果，いくつかのルールは興味深いと判断された．図1にルールの一例を示す．ルールに対する専門家のコメントは，次の通りである．医師の間では一般に，GPT は微小変化はするがほぼ一定に単調減少する，と言われている．一方，ルールは GPT が約3年周期で変動することを示唆し，これが事実であれば新発見となり得る．そこで，本研究では，この GPT 変動仮説の検証を試みる．

2.2 システムのフレームワークとルールの有用性の定義

まず，ルールの有用性を定義する．得られたルールが有用でなければ，構築したシステム，および，ルール発見の過程で得られた前処理・後処理やインタラクションに関する知見が妥当とは言えない．本研究では，医学分野において興味深いルールを得ることが目的であるため，再現率や精度だけでなく，専門家によるルールの評価結果を有用性の指標とする．そして，有用性が確認された上で，システムと知見の妥当性を議論する．

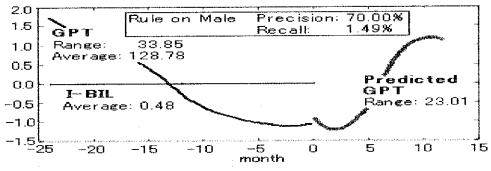

図1 先行研究で高く評価されたルールの一例．
Fig. 1 One of rules though highly in our previous research.

次に，支援システムのフレームワークを決める．なお，本支援システムの詳細については，3. で述べる．我々は，価値があるルールを得るにはインタラクションが重要と考え [8], [19]，これも支援システムの構成要素と考えた．これ以降，前処理・マイニングスキーム・後処理を行う狭義のシステムをマイニングシステム，我々が提案する人間とのインタラクションをも系に含めたシステムを支援システム，と呼ぶ．

本研究では，時系列医療データに適したマイニングシステムを構築して，支援システムに組み込み，マイニングシステムと専門家とのインタラクションを繰り返す．そして，前処理・後処理とインタラクションに関する知見を得るとともに，ルールの洗練化を試みる．先行研究 [6] では，GPT 変動の仮説と前処理の改善点が得られた．本研究では，この成果を支援システムにフィードバックし，2回目のインタラクションを実施する [15].

マイニングシステムは，時系列パターンの組合せで将来傾向を示すルールを得るように設計した．理由は以下の通りである．EBM では，検査結果からその後の症状を推定するニーズが高い．病状は時々刻々と変化し，様々な検査が連続的，断続的に行われる．また，検査結果の多くは名義値ではなく数値である．したがって，検査の時系列数値データから予後を推定することが，妥当と考えた．前処理については，データセットの詳細を調べ，時系列医療データとしての性質を検討した．そして，前処理を医学の領域知識に依存する (低レベル前処理) ／しない (高レベル前処理) の2つに明確に分けた．なお，今回は，後処理はグラフ表示によるルールの可視化にとどめた．

3. 支援システムの構築

パターンに基づく時系列データからのマイニング手法は多いが，その基本はクラスタリングによるパターン抽出と分類器による学習の組合せである [3], [16], [17], [20]．この一般的な方法は，次の通りである．最初に，観察したい期間の幅を持つ窓をスライドさせ，元のデータシーケンスからサブシーケンスを切り出す．次に，クラスタリングによってサブシーケンスから代表的なパターンを抽出する．そして，代表パターンに記号を与えて属性やクラスとし，マイニングスキームに入力してルールを得る．最後に，必要であればルールをグラフ形式に可視化する (図2の左側参照).

我々は，これをベースとした，時系列医療データからのルール発見支援システムを構築した．図2の右側に本支援システムの構成を示す．このマイニングプロセスのうち，クラスタリング

には K-means アルゴリズム [5] を，分類器には C5.0(C4.5 [18] の商用版) を用いた．本支援システムは，2.2 で述べたように，医学的な知識発見のため，**機能 1** 時系列医療データに適した 2 レベルの前処理，**機能 2** 専門家とのインタラクションを持つ．

なお，今回は，インタラクションを半自動化する機能は実装していない．これは，現段階では，本支援システム内でマイニングの機能とインタラクションの機能を分け，各機能のルール発見への影響を調べるためである．代わりに，第三者が専門家にルールやデータを提示して評価とコメントを受け取り，これを本支援システムに反映した．この過程で，インタラクションの効果と半自動化に必要な知見を得る．

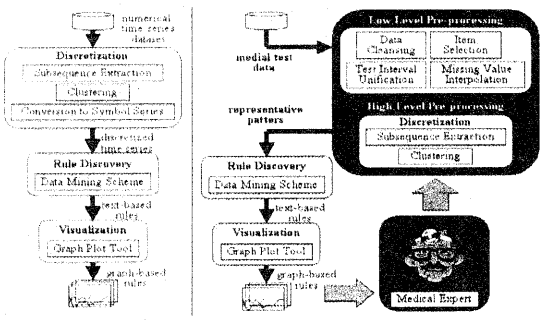

図 2　一般的な時系列データからのマイニング手法 (左)，提案するルール発見支援システム (右)．

Fig. 2　Popular mining method from sequential data (left) and Proposed rule discovery support system (right).

本研究では，千葉大学病院から提供された B 型・C 型ウィルス性肝炎患者の検査結果のデータセットを用いた [4]．我々は，本データセットについて専門家にヒアリングを行い，データセットの性質とそれに合った低レベル前処理を検討した．高レベル前処理については，先行研究 [6] で得た領域知識に基づき，クラスタの最大数を 8 個に，クラスタ内の最小事例数を 10 個に設定した．そして，前処理を行った結果，データセットの検査項目数は 957 から 80 に，患者数は 771 から 448 に，レコード数は約 160 万から 123 万に低減した．なお，データセットの性質，前処理の具体的な内容，これらの知見の一般性については，文献 [15] を参照されたい．

4. 支援システムの適用とルールの評価

4.1 システムの適用条件

先行研究 [6] では，観察期間 (切り出し窓の幅) を 6，12，24ヵ月の 3 種類とし，各期間で代表パターンを抽出した．そして，検査項目のパターンを属性，GPT のパターンをクラスとしてルールを得た．その結果，過去 24ヵ月検査項目のパターン組合せで，将来 6ヵ月の GPT のパターンを推定するルールが高く評価され，GPT が約 3 年周期で変動するという仮説を得た．そこで，本研究では，前処理において観察期間を過去 60ヵ月，将来 6ヵ月に延長し，長期的な視点で GPT 変動の仮説検証を試みることにした．

また，先行研究では，後処理に関して次の知見も得られた．多くの検査項目について正常範囲が知られており，正常範囲内の変動を表すルールは価値が低い．ただし，一般の正常範囲は安全性を確実に保証するものであり，医療現場の実用的な正常範囲よりもかなり狭い．そこで，本研究では，実用的な正常範囲を基準としてルールをフィルタリングした．

4.2 ルールの評価結果

GPT をクラスとして，本支援システムを慢性肝炎データセットに適用した結果，33 個のルールが出力された．そして，GPT 値が 100 未満という実用的な正常範囲を用い，異常値を示す 21 個のルールを選出した．これらをグラフとして専門家に提示したところ，3 つのルールに興味を持った．このうち，専門家が価値があると判断した 2 つのルールを図 3 に示す．矛盾を感じると述べた 1 つのルールは，後ほど本文中で示す．図において，横軸は月，縦軸は各検査結果の数値である．グラフは，過去 60ヵ月の検査パターンの組合せが将来 6ヵ月にどんな GPT パターンを引き起こすかを意味する．

図 3 のルール 1，2 の条件部 (過去 60ヵ月) において，GPT の変動パターンは同じである．これは，過去 60ヵ月の GPT 値が大局的に 2 回変動することを表しており，専門家はこれを重要視した．したがって，先行研究で立てた「GPT は約 3 年周期で変動する」という仮説がより支持された．

そこで，これらのルールの客観的な妥当性を検討した．まず，ルール内の GPT の代表パターンを元のサブシーケンスと比較し，代表パターンが本来の傾向を正しく反映するかを調べた．その結果，我々と専門家の目視による判断ではあるが，大きな乖離はなく代表パターンは妥当であった．また，ルール 1，2 の精度，再現率を合わせると，この仮説はデータの持つ傾向をある程度客観的に示すと考えられる．

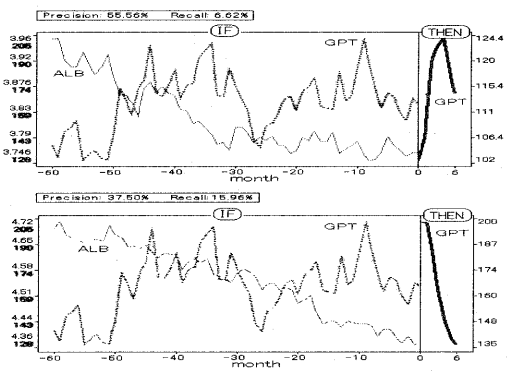

図 3　専門家に仮説を支持すると評価されたルール 1(上)，2(上)．

Fig. 3　Rule 1 and 2 thought highly by a human medical expert due to its support for the hypothesis on GPT.

ルール 1，2 では，条件部の GPT のパターンは同じであるが，条件部の ALB と結論部の GPT のパターンは異なる．我々は，過去の ALB の傾向が将来の GPT 値に影響すると解釈し，専門家に質問したが，ALB 値の差が誤差範囲であり，結論部も観察期間が短すぎるため，この点には興味を持たなかった．

一方，専門家はルール 3 に対して違和感を感じた．ルール

3は，GPT，GOT，TTTという検査項目の組合せから成り，GPTが下降後，GOTが上昇する傾向を示した．専門家は，GPTとGOTは同様の肝細胞酵素なので同期するはずであり，明らかに医学の領域知識と矛盾すると指摘した．そこで，我々は，このルール内のGPT，GOTの代表パターンを元のサブシーケンスと比較した．その結果，元のサブシーケンスからの乖離があり，代表パターンが実際のデータ傾向を反映しないことが分かった．

以上をまとめると，**結果1** ルール1，2より，GPTの約3年の周期変動という仮説がさらに支持され，**結果2** ルール3より，今回のパターン抽出では，代表パターンが元データから乖離する場合がある，という結果が得られた．**結果1** より，GPTに関する仮説の信頼性を高めた点では，本研究の**目的1** 医学的に有用なルールの発見を1つ達成し，本支援システムの有効性を示せたと言える．そこで，次節では，本研究の**目的2** 前処理・後処理とインタラクションに関する知見獲得について検討する．また，**結果2** で見つかった問題点も次節で議論する．

4.3 得られた知見

- マイニングプロセスについて

本研究では，パターンベースの推定ルールを得るように，本支援システムのマイニングプロセスを設計した．この妥当性を議論する．専門家は，他のマイニング手法のルール評価の経験があり，パターンの組合せルールは比較的理解しやすいと述べた．専門家は医療現場で検査履歴データを目にしており，病状把握の心的イメージがパターンの組合せに近いと推測される．これより，パターンベースのルール表現は妥当と言える．一方，専門家は，推定した将来の病状よりも，パターン形状や検査項目の組合せに着目した．コメントより，将来の観察期間が短すぎ傾向を把握しにくい，将来の病状が興味深い場合でも，過去の病状が複雑なため因果関係の解釈が困難，等の理由が挙がった．

4.2では，**結果2** の代表パターン乖離の問題が現れた．これは，パターンに基づく時系列データからのルール発見に共通する，大きな問題である[3]．今回は，代表パターン抽出のパラメータ，つまり，クラスタ数とクラスタ内事例数は領域知識に合うものとした．しかし，これだけでなく，パターンへの抽象化における粒度調整の重要性が示された．

- 前処理・後処理について

データの前処理については，主に2つの知見が得られた．1つ目は，慢性肝炎データセットの調査による，時系列医療データの性質とそれに適した前処理方法である[15]．2つ目は，上述したように，前処理での粒度調整の重要性である．

粒度の調整手法はすでに提案されているが[3]，妥当な粒度の値を得るための方法はまだ確立していない．これは，データ本来の客観的な傾向と専門家のニーズのバランスを保つ必要があり，困難な問題である．今回は，専門家の粒度に関するコメントから，観察期間は数カ月から長くて2年であると，経験的に知り得た．今後は，観察期間等の粒度に関するパラメータの値を，インタラクションプロセス内で半自動的に得る仕組みが必要である．

ルールの後処理でも，主に2つの知見が得られた．1つ目は，検査値の正常範囲を用いたフィルタリングの効果である．実際，専門家はフィルタリングで削除したルールに興味を示さなかった．これより，正常範囲というおおまかな基準でも，ある程度，不要なルールを削除できると分かった．2つ目は，ルールと元のデータの対提示の重要性である．専門家は，医学分野においては，病状の一般傾向だけでなく個別の症例も重要であり，仮説の検証では必ず元のデータに目を通すと述べた．したがって，後処理では，専門家の要求に合わせて元のデータも提示する機能が必要である．

- インタラクションについて

我々は，先行研究[6]から本研究[15]を通し，インタラクションによって，GPTに関する仮説の形成と検証がなされ，ルールが洗練化されることを示した．そこで，なぜインタラクションが有用なルールの発見に貢献するのかを検討する．この理由としては，専門家から支援システムへ，および支援システムから専門家への相補的な情報交換が考えられる．

専門家から支援システムへの情報として，対象分野でのルールの価値，システムの問題点・問題の原因・問題の箇所，が挙げられる．言い換えると，精度や再現率ではなく，主観のバイアスは含むが領域知識に裏打ちされたルールの価値が分かる．また，対象分野に適するかという観点で，前処理・マイニングスキーム・後処理において，なぜ，どこに，どんな問題があるかが明示される．これらの情報によって，マイニングシステムが適切に改善される．

支援システムから専門家への情報として，専門家の新たな視点や発想の材料，および，考えの客観的妥当性の材料，が挙げられる．つまり，インタラクションの初期段階では，精度が高くないルールでも，専門家はそれをヒントとして自分が持つ高度で莫大な領域知識に結び付け，新しい仮説を思い付く．そして，インタラクションの繰り返しの後半では，ある程度の精度を持つルールから，仮説を客観的に検証できる．

本研究では，第三者を介してインタラクションを行った．しかし，コストや客観性の面から，インタラクションの半自動化が望まれる．それでは，このプロセスのどこをどのように半自動化すれば，システムと専門家から有益な情報を抽出し，円滑な交換と相互の活用が可能になるであろうか．これについては，5.でインタラクションの概念モデルを提案後，詳細に議論する．

5. インタラクションのモデル化

5.1 概念レベルでのモデル

我々は，システムと専門家のインタラクションに関して本研究で得た知見，および，他の研究[8],[19]を参照し，概念レベルでインタラクションの体系化を試みた．図4に，我々が提案するインタラクションの概念モデルを示す．

4.3で述べたインタラクションの意義をまとめると，**意義1** マイニングシステムの問題点・問題の原因・問題の箇所を発見できる，**意義2** 専門家に新しい視点・発想を与え，仮説の形成・検証を促進する，の2つに集約される．本モデルは，これらを実現するインタラクションの構造とプロセスを，概念的に

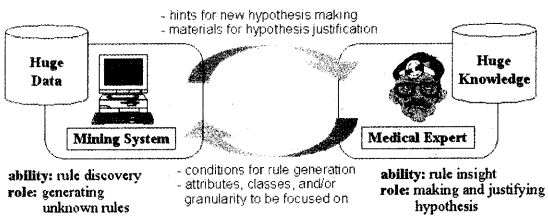

図 4　提案する概念レベルでのシステムと専門家のインタラクションモデル.

Fig. 4　Proposed interaction model between system and human expert at a concept level.

明確化する.

データマイニングの特徴は，仮説を立てず，データに忠実にルール生成することである．これは，未知ルールの発見という利点を持つが，多数の解釈困難なルールを生成する問題を持つ．一方，医学では，仮説の形成と検証が知識発見の基本である．現象を発見しても，その医学的な理由や意味を知る手がかりがなければ，実用的な知識とならないからである．以上から，支援システムに含まれるマイニングシステム，専門家における能力と役割は，次のように定義できる.

マイニングシステムの能力は，莫大なデータに埋もれた未知ルールの発見であり，役割はルール生成である．一方，医療の専門家の能力は，高度で莫大な領域知識と経験に基づくルールの洞察であり，役割は仮説形成・検証である．システムと専門家のインタラクションによって，お互いの能力が相補・活用される.

システムから専門家への情報は，新しい仮説のヒントと仮説の検証材料としてのルールである．専門家からシステムへの情報は，ルールの生成条件や着眼点の指摘，つまり，検査項目の関連性，正常値の範囲，ルールの粒度等である．これらの交換がシステムの改善と専門家の創発を促し，ルールが洗練化され，最終的に医学的に興味深い知識が得られる.

ただし，専門家が提供する情報は，システムが提供する情報より圧倒的に少ない．そこで，システムはマイニング機能だけでなく，専門家からの情報量の不足を補償する，あるいは許容する機能を持つ必要がある．また，認知的観点で専門家が理解しやすく，専門家が新しい視点・発想を生み出すタイミングに応じた形式で，ルールを提示する機能も不可欠である.

5.2　半自動化のためのフレームワーク

提案したモデルに基づきインタラクションを半自動化するには，専門家とシステムの間の情報を計算機上で扱える形式にすること，情報の使用目的・方法と情報の量・質を整合させること，が必要である.

情報の形式について，システムから専門家への情報であるルールは，本来，計算機上で扱える形式を持つ．一方，専門家からシステムへの情報は，言葉によるルールの評価結果やコメントである．これらを計算機上で半自動的に扱うには，次の方法が考えられる．**方法 1** 事前に評価結果やコメントを体系化し，その候補から選択する．**方法 2** 事前に評価やコメントの基準を体系化し，その基準について評価値を与える．**方法 3** これらを体系化せず，総合的な判断で善し悪しの評価値を与える．どれを用いるかは，情報の使用目的・方法に依存する.

情報の使用目的・方法は，マイニングシステム (前処理・マイニングスキーム・後処理の各モジュール) における，処理手法の改善とパラメータ最適化に大別できる．しかし，専門家からの情報は非常に少なく，処理手法そのものを半自動的に改善することは難しい [8], [12]．将来的には，専門家の評価の学習によって情報の欠如を補い [10]，マイニングスキームを自動生成する手法 [9] を応用できるかもしれない．しかし，現時点では困難と思われる.

専門家からの情報を，処理手法のパラメータ最適化に利用することは可能であろう．特に，前処理・後処理のパラメータ最適化には適すると考えられる．なぜなら，これらのパラメータは専門家が直感的に理解しやすい場合が多く，その数も少ないからである．例えば，専門家から粒度についてフィードバックがあれば，前処理の切り出し窓幅というパラメータを最適化することが挙げられる．一方，専門家からの評価情報を，マイニングスキームの主なパラメータであるマイニング基準の最適化に用いて良いかは疑問である．専門家の主観的バイアスの不必要な介入が，マイニングスキーム本来の客観的ルールの生成に悪影響を及ぼす恐れがある.

以上から，専門家とシステムのインタラクションにおいて半自動化が可能な機能は，専門家からの情報に基づく前処理・後処理のパラメータ最適化，と結論つけられる．そこで，我々は，これを実現し得るインタラクションのフレームワークを検討した (図 5 参照).

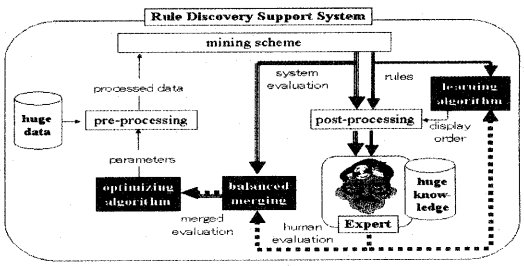

図 5　提案するインタラクションの半自動化フレームワーク.

Fig. 5　Proposed framework for semi-automatic interaction.

図 5 の黒色の項目は，半自動的なインタラクションに寄与するモジュールである．まず，前処理のパラメータ最適化について述べる．これには，専門家，マイニングシステムのそれぞれがルールに与える評価情報を統合し，利用する．専門家の評価情報は，主観的バイアスを含むが領域知識を反映し，マイニングシステムの評価情報 (正答率，再現率等) は，領域知識は反映しないが客観的である．これらを，$\alpha \times F_{human} + (1-\alpha) \times F_{system}$ として，重み付け平均し統合することで，両者のバランスを調整する.

最適化アルゴリズムは，統合された評価情報に基づき，マイニングシステムの前処理パラメータを設定する．設定されたマ

イニングシステムは新たなルールを生成し，専門家に提示する．そして，専門家とマイニングシステムは，新たな評価情報を最適化アルゴリズムにフィードバックする．この繰り返しによって，前処理パラメータが最適化される．

なお，繰り返しの初期段階ではαを大きく設定し，精度は高くなくても，専門家の新たな視点・発想を促し仮説形成につながるルールを生成する．そして，だんだんとαを小さく設定し，客観的な仮説検証の材料となるようにルールの性質を変化させる．また，インタラクション過程において，仮説の形成，検証のどちらを行いたいかに応じ，専門家がαを自由に変更できるようにする．

後処理については，専門家とマイニングシステムで統合した評価情報ではなく，専門家からの評価情報を用いる．なぜなら，後処理の主な目的は，認知的な観点で専門家が理解しやすいように，ルールや付随する情報を提示することだからである．この方法として，ルールと専門家の評価の対応関係を学習アルゴリズムに学習させ，その結果を，ルールのソート提示やフィルタリングのパラメータ最適化に使用することが挙げられる．

現在，我々は，このフレームワークのうち，前処理パラメータの最適化にはインタラクティブ進化計算[8],[12]を，後処理パラメータの最適化には線形モデル，あるいは階層型ニューラルネットワーク[11]の適用を考えている．したがって，専門家からの情報は**方法 3** を用いて形式化する予定である．なお，前処理については文献[8],[19]，後処理については文献[10]に同様の研究があり，その成果が示されている．ここで，特に関連が深いと考えられる文献[8],[19]の研究と，本研究の違いについて述べておく．

文献[8],[19]のシステムは，インタラクティブ進化計算の一種である摸擬育種と通常のGAの2段階で構成され，前処理の属性選択を試みている．なお，摸擬育種では，ルールに対する専門家の主観的な評価値，通常GAでは，ルールの精度やコンパクトさを重み付け加算した客観的な評価関数が使われている．一方，我々が提案するフレームワークは，インタラクティブ進化計算の1段階のみを用い，仮説の形成から検証の段階に合わせて，専門家とシステムの評価バランスを制御することを意図する．また，後処理についても，ユーザインタフェースの向上を目的とし，専門家の評価の学習機能を組み込もうとしている．

6. むすび

本研究は，慢性肝炎の検査データを対象とし，システムと専門家のインタラクションを繰り返した．そして，医学的に有用なルール，および前処理，後処理，インタラクションに関する知見の獲得を試みた．まず，クラスタリングによるパターン抽出と決定木による分類に基づくルール発見支援システムを提案，構築した．次に，GPTという検査項目をクラスとし，将来の病状を推定するパターンベースのルールを生成した．生成と専門家による評価を2回繰り返した結果，GPTの周期変動に関する仮説を検証でき，システムの有効性が示された．最後に，得られた知見に基づき，概念レベルでインタラクションをモデル化し，半自動化のフレームワークを提案した．今後は，このフレームワークの実現を試る予定である．

文　献

[1] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi, "Discovering Data Mining — From Concept to Implementation —," International Business Machines Corporation, 1998.

[2] P. Adriaans and D. Zantinge, "Data Mining," Addison-Wesley, 1996.

[3] G. Das, L. King-Ip, M. Heikki, G. Renganathan, and P. Smyth, "Rule Discovery from Time Series," Proc. of Int'l Conf. on Knowledge Discovery and Data Mining (KDD-98), New York, USA, pp.16–22, 1998.

[4] "Hepatitis Dataset for Discovery Challenge," European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'02), Helsinki, Finland, http://lisp.vse.cz/challenge/ecmlpkdd2002/ (Aug., 2002) .

[5] J. A. Hartigan, "Clustering Algorithms," Wiley Publishers, New York, USA, 1975.

[6] 畑澤寛光, 佐藤芳紀, 山口高平, "シーケンシャルパターン分析に基づくルール発見支援システム－慢性肝炎データセットを対象にして－," 人工知能学会, 第56回知識ベースシステム研究会講演論文集 (SIG-KBS-A201), pp.55–60, 2002.

[7] S. Hirano, and S. Tsumoto, "Mining Similar Temporal Patterns in Long Time-Series Data and Its Application to Medicine," Proc. of IEEE Int'l Conf. on Data Mining (ICDM'02), Maebashi, Japan, pp.219–226, 2002.

[8] 北野宏明 (編), "遺伝的アルゴリズム 4," 産業図書, 2000.

[9] 酢山明弘, 山口高平, "オントロジーを利用した帰納アプリケーションの自動構成," 人工知能学会誌, vol.15, no.1, pp.155–161, 2000.

[10] 大崎美穂, 高木英行, "対話型EC操作者の負担低減, — 評価値予測による提示インターフェイスの改善 —," 人工知能学会誌, vol.13, no.5, pp.712–719, 1998.

[11] F. Rosenblatt, "Principles of Neurodynamics, — Perceptrons and the Theory of Brain Mechanisms," Spartan, 1961.

[12] Hideyuki Takagi, "Interactive Evolutionary Computation: Fusion of the Capacities of EC Optimization and Human Evaluation," Proc. of the IEEE, vol.89, no.9, pp.1275–1296, 2001.

[13] 松田喬, 元田浩, 鷲尾隆, "一般グラフ構造データに対するGraph-Based Induction とその応用," 人工知能学会誌, vol.16, pp.363–372, 2001.

[14] H. Motoda (ed.), "Active Mining," IOS Press, 2002.

[15] M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi, "A Rule Discovery Support System for Sequential Medical Data, – In the Case Study of a Chronic Hepatitis Dataset –," Int'l Workshop on Active Mining (AM-2002) in the IEEE Int'l Conf. on Data Mining (ICDM'02), Maebashi, Japan, pp.97–102, 2002.

[16] P. Patel, M. Keogh, J. Lin, S. Lonardi, "Mining Motifs in Masssive Time Series Databases," Proc. of IEEE Int'l Conf. on Data Mining (ICDM'02), Maebashi, Japan pp.370–377, 2002.

[17] J. Pei, G. Dong, W. Zou, and J. Han, "On Computing Condensed Frequent Pattern Bases," Proc. of IEEE Int'l Conf. on Data Mining (ICDM'02), Maebashi, Japan pp.378–385, 2002.

[18] J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, San Fransisco, USA, 1993.

[19] 寺野隆雄, 稲田政則, "対話型進化計算を利用した医療データからの知識発見," 人工知能学会, 第42回知識ベースシステム研究会講演論文集 (SIG-KBS-9802), pp.13–18, 1999.

[20] "Time Series Data Mining Archive," The Webpage of Dept. of Computer Science and Engineering, Univ. of California, http://www.cs.ucr.edu/ eamonn/TSDMA/, 2003.