発見ルールフィルタリングシステムの予備評価
A Preliminary Evaluation of Discovered Rule Filtering System

飯田 暁[1], 北村 泰彦[2], 朴 勤植[3], 辰巳 昭治[1]
Akira Iida, Yasuhiko Kitamura, Keunsik Park, Shoji Tatsumi
[1]大阪市立大学大学院工学研究科 [2]関西学院大学理工学部 [3]大阪市立大学大学院医学研究科
[1] Graduate School of Engineering, Osaka City University
[2] School of Science and Technology, Kwansei Gakuin University
[3] Graduate School of Medicine, Osaka City University

Abstract: A data mining system semi-automatically discover knowledge by mining a large volume of data, but the discovered knowledge is not always novel to the user. We discuss a discovered rule filtering method to filter rules discovered by a data mining system into ones that are novel to the user by using information retrieval results from the Internet. We have two approaches; the micro view and the macro view approaches, to achieve discovered rule filtering. In the micro view approach, we extract keywords from a discovered rule and rank the rule referring to the number of hits when we submit the keywords to the MEDLINE database. Unfortunately, this approach does not work well because the retrieved result contains a number of noisy documents and because the number of hits depends much on the number of keywords but on the novelty of rule. In the micro view approach, we first retrieve documents by submitting every pair of the extracted keywords and then make the results gather in clusters. As a preliminary evaluation of the macro view approach, we made a survey on a relation between a set of outputs of the system and judgments made by medical experts. We found a tendency that keywords in a known rule form a single cluster.

1 はじめに

アクティブマイニングは情報収集，データマイニング,ユーザリアクションの技術を融合することにより，ユーザの目的にあった質の高い知識の効率的な発見を目指すデータマイニングの新しいアプローチである[1]. 本稿ではインターネット上からの文献情報検索結果に基づきデータマイニング結果をフィルタリングする発見ルールフィルタリング手法[4,5]とその予備評価について述べる.

データマイニングは大量のデータを機械処理することにより，ユーザにとって有用な知識を自動的に発見しようとする手法である．一般的には与えられたデータに含まれる属性間の関係から統計的に意味のある関係を発見する手法がとられている．しかしながら単に統計的な特徴だけでデータマイニングを行うなら，(1)ユーザが扱いきれないほどの大量の知識が得られる，(2)ユーザにとって既知の知識が得られる，といった問題が生じる．そこで本研究ではルール形式で得られる大量の発見知識の中から新規なもののみにフィルタリングする発見ルールフィルタリングの開発を行っている．この発見ルールフィルタリングを実現するには発見されたルールが新規かどうかを判定する必要があるが，これは当然のことながらマイニングの対象となる情報源以外の外部情報源を参照する必要がある．そこで，本研究でははインターネット上の情報源を用いて，発見ルールの新規性判定を行おうとしている．

本論文では以下，2 章において発見ルールフィルタリングの概要について述べる．発見ルールフィルタリングにはミクロビューアプローチとマクロビューアプローチの二つが考えられるが，3 章ではミクロビューアプローチとその予備評価について述べる．4 章ではマクロビューアプローチとその予備評価について述べる．最後に，5 章でまとめと今後の展望について述べる．

2 発見ルールフィルタリング

発見ルールフィルタリング[4,5]ではデータマイニングシステムにより発見されたルール形式の知識に対して，それに関連する情報をインターネット上から検索し，その結果に基づき，発見ルールのフィルタリングを行う．ここでの議論をより具体的なものとするために，肝炎データからのマイニングを知識発見の対象領域として議論を進める．発見ルールフィルタリングの具体的な手順は以下の通りである．

2.1 発見ルールの獲得

データマイニングシステムを利用してルール形式の知識を得る．静岡大学の山口グループでは，千葉大学医学部より提供された肝炎データから肝炎の進行具合を示す血液データ(GPT)と他の検査データとの時系列的な相関関係に着目し，様々なルール形式の知識発見を行っている[2]．発見されたルールの一例を図 1 に示す．

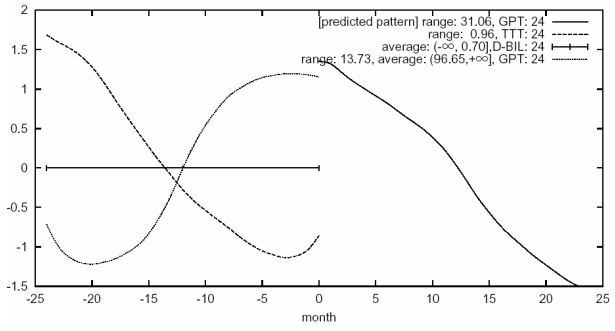

図 1 ：発見ルールの一例

これは GPT（グルタミン酸ピルビン酸トランスアミナーゼ）と TTT（チモール混濁試験），D-BIL（直接ビリルビン）の関係を図的に表すものであるが，「24 ヶ月間，D-BIL が一定で，TTT が減少し，GPT が増加しているなら，その後 24 ヶ月間の GPT は減少する」というルール表記を行うことも可能である．

2.2 発見ルール駆動情報検索

発見されたルールに関連する情報を，インターネットを利用して収集する．肝炎に関連する一般の Web 情報源はあまりにも雑多な情報が含まれているので，本研究では医学・生物学関係の文献データベースである MEDLINE を用いている．MEDLINE（MEDlars on LINE）は，米国をはじめ 70 カ国で出版された 4000 誌を超える医学・生物学系学術雑誌からのアブストラクトを含む書誌情報データベースであり，1966 年以降の 1100 万件以上のデータが蓄積されている．PubMed [3] (http://www.ncbi.nlm.nih.gov/entrez/query. fcgi)は NCBI (National Center for Biotechnology Information)によりインターネット上に無料提供されている MEDLINE の検索サービスであり，一般の検索エンジン同様，キーワードを入力することにより，MEDLINE 文献の検索を行うことができる．

発見ルールに関連する情報を MEDLINE データベースから検索するためには，それに関連するキーワードを獲得する必要がある．このキーワードは発見ルール，マイニング領域から導出している．例えば，図 1 の例では GPT, TTT, D-BIL が発見ルールより抽出される．またマイニング領域は肝炎であるので，hepatitis が領域キーワードとなる．これらのキーワードを組み合わせたものが MEDLINE データベースの検索に用いられる．

図 1 で示すルール自体は各検査値の変動に関する情報も含んでいるが，現時点ではそのような情報を文献検索に反映することは困難であり，これは今後の課題となる．

2.3 発見ルールフィルタリング

MEDLINE 文献検索の結果に応じて発見ルールのフィルタリングを行う．その手法としては，文献検索結果に応じてルールのランキングを行ったり，ある閾値を設定して，それ以下のものを排除したりすることである．文献検索結果とランキングをどのように対応付けるかは本研究の中核となる重要な研究課題であるが，ミクロビューとマクロビューの二つのアプローチが考えられる[4]．次章以下ではそれぞれのアプローチの詳細と予備評価について述べる．

3 ミクロビューアプローチとその評価

3.1 ミクロビューアプローチの仮説

ミクロビューアプローチは文献検索のヒット数に応じて発見ルールのランク付けをするアプローチである．ミクロビューアプローチでは以下の仮定をおいた．

【仮定】（ミクロビューアプローチ）

1. 既知のルールのヒット数は多い．
2. 未知のルールのヒット数は少ない．
3. ゴミのルールのヒット数はゼロである．

この仮定では既知のルールはよく研究がなされており，それに関連する文献は多く出版されていることを前提としている．ヒット数が少なければ，それは十分な研究がなされているとは限らないので，未知ルールに分類することができると考えられる．一方で，ルールが無意味なものに関しては誰も研究しないのでゼロであると仮定している．

3.2 仮説の検証と考察

以上の仮説の妥当性を検討するために，医学専門家へのアンケートに基づく予備評価を行った．

アンケートの質問文を図 2 に示す．問 1 の質問項目には，静岡大学のグループにより得られた 30 個のルールの中からランダムに 20 個選択してそこに含まれる属性名を使用した．アンケートの対象は，医師国家試験を直前に控えた医学部の 6 回生，４７人である．国家試験を直前にひかえた医学生であることから，医学的

問１：下のキーワードの組み合わせで MEDLINE を検索するとしたときに，検索結果がどのようになると思いますか？　次の三つの中から選んでください．

１．　当然（どのような内容の文献が見つかるか，ある程度予想できる）
２．　おもしろい（自分が知らないことが載っている文献が見つかるかもしれない）
３．　無意味（ほとんどヒットしないだろう）

（１）ALT　TTT　（1.当然、２．おもしろい、３．無意味）
（２）TTT　直接ビリルビン　ALT　（1.当然、２．おもしろい、３．無意味）
（３）ALT　総コレステロール　C 型肝炎　（1.当然、２．おもしろい、３．無意味）
・・・

問２：次の検査項目同士の関係について，次の四つの中から適切だと思うものを選んでください．

１．強い関係がある．
２．関係がある．
３．あまり関係は無い．
４．関係ない．

（１）ALT　総ビリルビン　（　１、　２、　３、　４　）
（２）ALT　総コレステロール　（　１、　２、　３、　４　）
（３）ALB　総コレステロール　（　１、　２、　３、　４　）
・・・

図２：アンケートの質問文

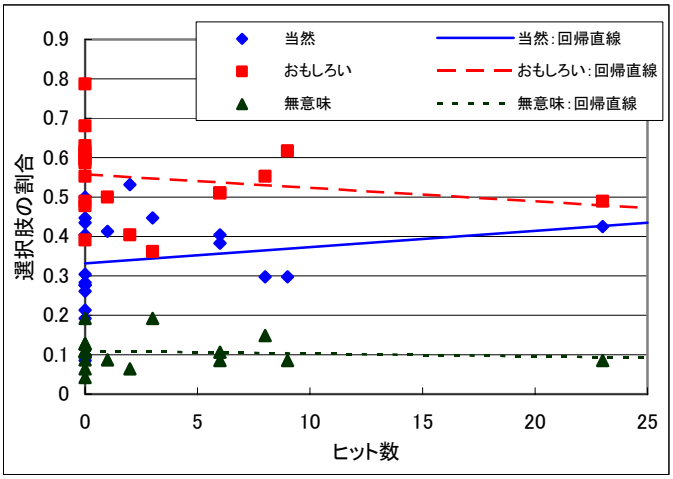

図３：選択肢の割合とヒット数の関係

な常識とも言える教科的な知識についてはかなり詳しいと考えられる．

図３に問１のアンケート結果と文献検索結果をまとめたものを示す．それぞれの発見ルールに対して，三つの選択肢が選ばれた割合と関連文献検索したときのヒット数の関係をプロットしたものである．またそれぞれの選択肢について求めた回帰直線も表示している．

それぞれの選択肢について，その回帰直線の傾きを有意水準 95%で t 検定[6]した．自由度 18（データ数 20 から２を引いたもの）での 2.5%点の値（2.101）を，それぞれの t 値の絶対値が上回っていれば，ヒット数は選択肢の割合に影響を与えていると言える．

それぞれの選択肢について t 値を求めると，「当然」では 0.887，「おもしろい」では-0.812，「無意味」では -0.441 となり，アンケート結果とヒット数の間には有意な相関を見出すことはできなかった．

ミクロビューアプローチによる検索結果と専門家との評価の間に相関が得られなかった理由としては，検索ヒット数は検索に用いるキーワード数に依存することが考えられる．図４はキーワード数とヒット数の関係と，そのキーワード数ごとの平均値をプロットしたものである．この図より，キーワード数が増えると，ヒット数は減少する傾向にあることがわかる．よって，文献検索ヒット数は，キーワード数に依存しているために，抽出されたキーワードをそのまま文献検索に用いる手法は有効でないことが明らかになった．

ただし，キーワード数が少ない場合は専門家の評価とヒット数に相関があることが予想できる．そこで，図２の問２に示されるアンケート調査を実施した．問２は，発見ルールに含まれるキーワード対の組み合わせ

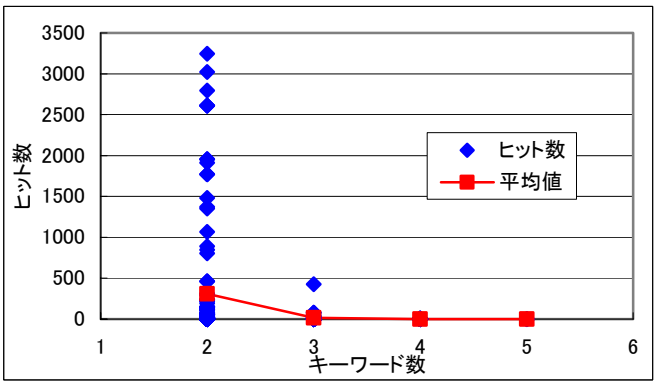

図４：キーワード数とヒット数の関係

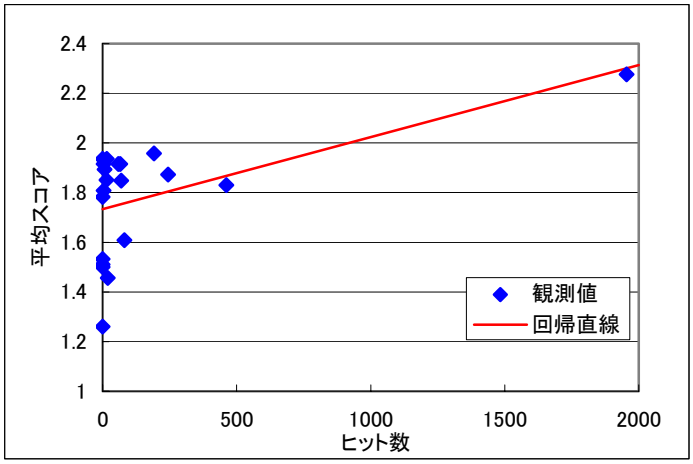

図５：ヒット数と平均スコアの関係

の中から 20 組を選択したものを用いた．

図５に問２の結果を示す．「強い関係がある」，「関係がある」，「あまり関係はない」，「無関係」のそれぞれに 3，2，1，0 のスコアをつけ，その平均と MEDLINE 検索によるヒット数をプロットしたものである．平均スコアとヒット数との間の相関係数は 0.54 であり，ヒット数と専門家の評価には相関があることがわかった．したがって文献検索に用いるキーワード数を２に限定すれば，検索結果と専門家の評価には相関がある．この結果は次章のマクロビューアプローチへつながってゆく．

4 マクロビューアプローチとその評価

4.1 マクロビューアプローチの仮説

ミクロビューアプローチはキーワード検索を用いて発見ルールに直接関連する文献を検索しようとするアプローチであったが，現状では十分なフィルタリング能力を発揮することはできなかった．そこでマクロビューアプローチでは発見ルールから抽出されるキーワードの全ての対に関する共起文献数を参照することで，発見ルールに関連する研究動向を推測しようとするものである．

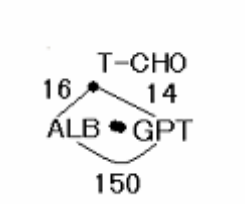

図６：GPT, ALB, T-CHO 間の共起関係

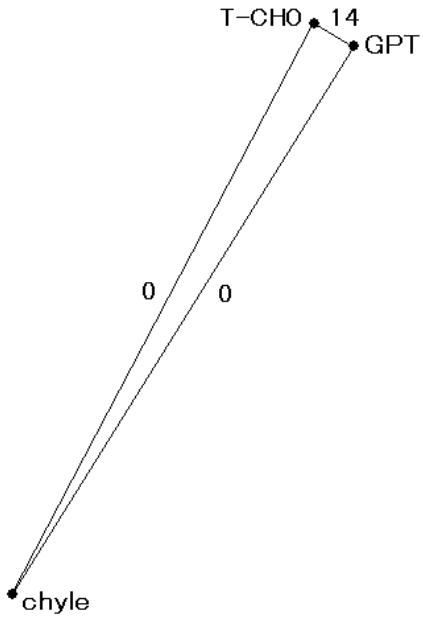

図７：GPT, T-CHO, chyle 間の共起関係

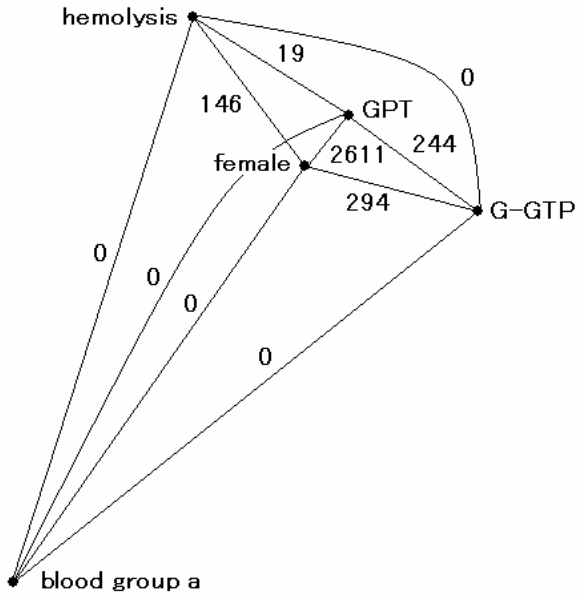

図８：GPT, G-GTP, hemolysis, blood group a, female 間の共起関係

図６，７，８は３つのルールでの共起文献数をグラフとして表したものである．グラフ中でのノードはそれぞれのキーワードを表し，エッジの長さは二つのキーワードでの共起数の逆数，添字は共起数を表す．

図６のルールでは，ALB，GPT，T-CHO ともにそれ

ぞれ関係があるといえる．図７のルールでは，T-CHO, GPT の間には関係があるが chyle との間には関係がないことが分かる．図８のルールでは，GPT, female, G-GTP 間には相互に関係があるが，hemolysis と G-GTP 間と blood group a とその他のキーワードとの間には関係がないことが分かる．

それぞれのルールを明確に特徴付けるために，階層クラスタ分析法[7]を用いてキーワードのクラスタリングを行った．階層クラスタ分析とは，グラフ中のノードをいくつかのグループに分けることであるが，そのクラスタを生成する手法としては共起数の逆数を距離として，最長距離法（CLINK：complete linkage clustering method）を用いた．この手法はまず，最も距離の近いノード同士を統合しクラスタを生成し，新たに統合されたクラスタと他のクラスタとの距離にはそれぞれのクラスタに属するノードのうちで最も遠いもの同士の距離として定め，ある閾値以内にあるノードはクラスタに加えてゆく．こうしてできたそれぞれのクラスタに含まれるキーワード同士は関係が深く，それらに関する知識はすでに知られていると見なすことができる．そこから，生成されたクラスタの数により，ルールを以下のように分類できると仮定する．

【仮定】（マクロビューアプローチ）

1. 既知のルールのクラスタ数は 1 である．
2. 未知のルールのクラスタ数は 2 である．
3. ゴミのルールのクラスタ数は 3 以上である．

クラスタ数が 1 の場合は，すべてのキーワード対に関する研究が活発な場合であるので，そのようなキーワードを含むルールは既知であると考えられる．クラスタ数が 2 の場合は，キーワード集合が二つのクラスタにグループ化される場合である．それぞれのグループに関する研究は行われていても，グループ間の研究は十分ではないと考えられるので，それに関するルールは未知であると判定できる．しかし，クラスタ数が 3 以上の場合は，ゴミルールとした．これは未知の関係があるグループが複数存在する場合には，ルールとして複雑すぎると考えられるからである．

例えば，クラスタリングの閾値を距離１（キーワード間の共起数が１件のもの）とすると，図６のルールでは，全てのキーワードが１つのクラスタに統合さる．図７のルールでは，GPT と T-CHO は統合されるが chyle は統合されず，クラスタ数は２となる．図８のルールでは GPT, G-GTP と female が一つのクラスタに統合され，hemolysis と blood group a は統合されないので，クラスタ数は３となる．

4.2 仮説の検証と考察

以上の仮説の妥当性を検討するために，3 章で行ったアンケート調査（問１）の結果を用いた．表１にその結果を集計したものを示し，図９にクラスタ数と選択肢の割合の関係をグラフにして示す．それぞれの選択肢が選ばれた割合と，階層クラスタ分析の結果のクラスタ数との関係をプロットしたものである．なお，クラスタリングの閾値は，距離１とした．

仮説の検証を行うために，χ^2 検定[8]を用いて，クラスタ数の変化に伴い，それぞれの選択肢の割合に有意な変化があるかを検証した．有意水準としては 95%を用いた．

その結果，「当然」の割合は，クラスタ数１と２，１と３の間に有意な差があることが認められた．「おもしろい」の割合にも，クラスタ数１と２，１と３の間に有意な差があることが認められた．「無意味」の割合は，クラスタ数の変化による差は認められなかった．よって，クラスタ数が１から複数になると「当然」の割合は減少し，「おもしろい」の割合は増加する，と言える．また，「無意味」の割合はクラスタ数の変化と関係があるとは言えない．

このことから，クラスタ数が１のルールは「当然」なものであり，複数個あるルールは「おもしろい」ものであると考えられる．

また，与えられたルールに「無意味」と回答する割合は少ない．したがって，ルールをフィルタリングしてしまうよりは，クラスタ数の多い順にランキングする方が適切であると考えられる．

表１：問１の結果

クラスタ数	当然	おもしろい	無意味	無回答	計
1（10 例）	189	233	46	2	470
2（8 例）	110	219	43	4	376
3（2 例）	22	60	12	0	94
計	321	512	101	6	940

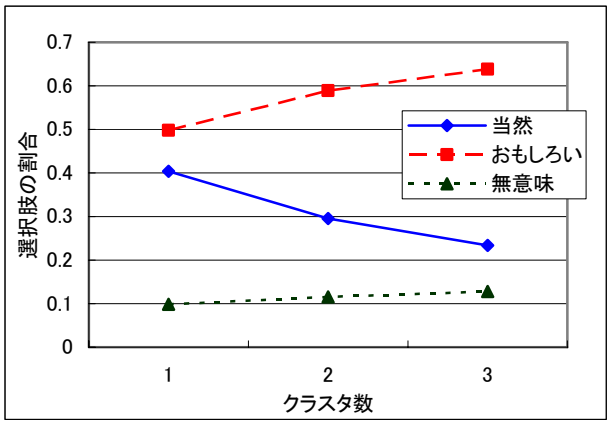

図９：選択肢の割合とクラスタ数の関係

なお,今回の調査ではクラスタ数は最大で 3 であるので，クラスタ数が 4 以上のルールに対する検証は今後の課題である．

5 まとめ

MEDLINE情報検索の結果を用いてデータマイニングにより得られた発見ルールを，利用者にとって未知なものにフィルタリングする発見ルールフィルタリングついて述べ，二つのアプローチとその予備評価について述べた．ミクロビューアプローチで行った文献検索は単純なキーワード検索に基づく手法であり，十分な精度を得るようなフィルタリングを行うことが困難であった．一方，共起関係を用いたキーワードクラスタリングによるマクロビューアプローチによる結果は医学専門家との判断とも相関があることがわかった．ただし，医学専門家は発見ルールに対して「無意味」のラベルを貼ることには慎重であるので，発見ルールをフィルタリングしてしまうことは問題があるかもしれない．むしろ，発見ルールをその重要性に応じてランキングする手法をとることが望ましい．

今後の課題としては以下のものが挙げられる．

(1)情報検索の精度の向上．マクロビューアプローチにおいても，キーワードの共起関係を求めるために検索を行っているが，その精度は必ずしも高いとはいえない．文献検索の精度の向上を行えば，さらに正確なフィルタリングが可能になると考えられる．一つの手法は得られた文献のアブストラクトを自然言語処理の手法を用いて解析することである[9]．例えば，発見ルールに含まれる属性が文献アブストラクト中で離れた場所に存在するとするならばそれらの属性の関連を述べた文献である可能性は低いかもしれない．したがってアブストラクトの構文解析を行うことにより，属性キーワードが同一文中に現れるかどうかを確認できればフィルタリングの精度は向上すると考えられる．さらに属性キーワードが結果や結論に関連する文の中に現れるかどうか，属性キーワード間の関係を修飾する文節は何か，などということが明らかになればフィルタリングの精度はさらに向上すると考えられる．

また先にも述べたが本研究での文献検索は単純なキーワード検索によっているので，属性間の関係や時系列的な変化を反映したものではない．これに関しても自然言語処理の手法を応用することで対処できるかもしれない．

(2)発見ルールフィルタリングの効果の実証．開発した手法を肝炎データマイニング分野に応用し，利用者の支援にどの程度役に立つかを明らかにする．これに関しては静岡大学などで行われている肝炎データマイニングの結果を用いてその検証を行う予定である．

謝辞 : 本研究にあたり，肝炎データベースからの発見ルールを提供していただいた静岡大学山口高平教授，自然言語処理に関して議論いただいた奈良先端大学院大学の松本裕二教授に感謝の意を表します．

参 考 文 献

[1] H. Motoda (Ed.), Active Mining: New Directions of Data Mining, IOS Press, Amsterdam, 2002.

[2] 畑澤寛光，佐藤芳紀，山口高平，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－，人工知能学会研究会資料, SIG-KBS-A201-11, 2002.

[3] 懸俊彦，PubMed 活用マニュアル，南江堂，東京，2000.

[4] Y. Kitamura, K. Park, A. Iida, and S. Tatsumi. Discovered Rule Filtering Using Information Retrieval Technique. Proceedings of International Workshop on Active Mining, pp. 80-84, 2002.

[5] Y. Kitamura, A. Iida, K. Park, and S. Tatsumi. Micro View and Macro View Approaches to Discovered Rule Filtering. Proceedings of 2nd International Workshop on Active Mining, pp.14-21, 2003.

[6] 岸野洋久，社会現象の統計学，朝倉出版，1992.

[7] 岡太彬訓，今泉忠，パソコン多次元尺度構成法，共立出版，1994.

[8] 津村善郎，渕脇学，築村昭明，社会統計入門[第２版]－経済学を学ぶ人のために－, 東京大学出版会, 1988.

[9] T. Yamasaki, M. Shimbo, and Y. Matsumoto: A MEDLINE document search system using section information, 人工知能学会研究会資料，SIG-KBS-A301-05, 2003.