1F5-05

ILPを用いた時系列データからの知識発見

Time-series Analysis of Sequential Data using ILP

佐藤 慶宜[*1] 市瀬 龍太郎[*2] 沼尾 正行[*3]
Yoshinori Sato Ryutaro Ichise Masayuki Numao

[*1]東京工業大学大学院情報理工学研究科
Department of Computer Science, Tokyo Institute of Technology

[*2]国立情報学研究所知能システム研究系
National Institute of Informatics

[*3]大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

The most important point of treating medical data is that it contains time-series data with irregularities. Since Inductive Logic Programming (ILP) has more expressive representation than propositional logic, we propose to handle this kind of data employing ILP system. Experimental results show that ILP system is suitable to acquire time-series knowledge from medical data.

1. はじめに

医療分野では，近年注目されている Evidence Based Medicine（科学的根拠に基づいた医療）の一環として，これまでに病院やその他の関連施設で蓄積されたデータから，有用な情報の抽出することが望まれている．

現在筆者等は，医療データに関するアクティブマイニングプロジェクト[*4]に参加している．蓄積された医療データを解析することで，これまでの医療現場では見逃されていた有用な知識の発見することが，このプロジェクトの目的の１つである．

医療データはその特徴として，属性数の多さや属性間の関係の複雑さ，時系列の重要性などがある．そのような性質上，医療データの解析では時系列や属性間の関係を無視することはできない．

これまでにも帰納論理プログラミング（ILP）によるデータ解析は行なわれてきたが，時系列データを扱った例は少ない [Rodriguez 00][Zem 98]．しかし，ILP は述語論理を使用することで表現力が豊かであり，また背景知識として医療情報を組み込むこともできるので，より精度の高い解析が行えると考えられる．そこで，代表的な ILP システムの１つである Progol[Muggleton 95] を利用し，背景知識として時系列を扱う述語を与えて医療データの解析を行なった．

2. 解析目標

肝炎の１つとして血液を介して感染する C 型肝炎が存在する．C 型肝炎の多くは慢性化し，肝硬変や肝ガンに進行する事も多いが，この C 型肝炎に有効な治療法として，肝炎ウイルスの増殖を抑える物質であるインターフェロン (IFN) を投与する IFN 療法がある．しかし，IFN 療法は全ての患者に有効とは限らず，副作用も存在し，またコスト的な問題など様々なリスクを抱えている．もし，比較的検査の容易な血液検査などの検査履歴から IFN 投与の有効性を予測することができれば，このようなリスクを避ける事ができ，患者の負担を減らす事になる．そこで，IFN 投与以前の血液検査データを使用し，IFN 療法の有効・無効性を分類するルールの作成を目標とした．

連絡先: 佐藤慶宜，東京工業大学大学院情報理工学研究科計算工学専攻，〒 152-8550 東京都目黒区大岡山 2-12-1, sato@nm.cs.titech.ac.jp

*4 http://www.ar.sanken.osaka-u.ac.jp/activemining/

3. 手法

3.1 ILP 使用の利点

ILP の特徴として，述語論理により表現力が豊かであることと背景知識を組み込めることが挙げられる．医療データという時系列データを ILP で扱うことに関して，次のような利点が考えられる．

述語論理による利点に関しては，例えば，ある検査項目の２点間の変化を表わす述語を１つ作成すれば，その述語の連言により，ある検査項目の一連の時系列が表現可能である．また連言が複数の検査項目に関する述語で構成されていれば，それは検査項目間の共起性を表現していることになる．すなわち，時系列と属性間の関係が同時に表現することが可能であるということである．

また，背景知識に関する利点としては，検査間の関係や検査値の動向などで，既知の情報をあらかじめ背景知識として組み込むことで，解析精度を上げられると期待される．

3.2 作成述語

解析目標と ILP の利点を踏まえ、今回作成した述語は主に以下の４つである．

ifn_effect(Mid)
患者 Mid はインターフェロン効果が著効 or 有効である．

test_result(Mid,Date,Tlist)
患者 Mid の検査日 Date の検査結果一覧は Tlist である．

move(Mid,Date1,Date2,Test,Val1,Val2)
患者 Mid は検査日 Date1 から Date2 の間に検査項目 Test の値が Val1 から Val2 に変化した．

relation(A,>,B)
A は B より大きい．

“ ifn_effect/1 ”は目標概念を表わす述語でありルールのヘッドに使用され，ルールのボディは主に“ move/6 ”の連言として構成されることになる．“ test_result/2 ”には患者の検査データが格納され，“ move/6 ”の作成に使用される．

表 1: 検査値の離散化指標

GOT					N <=	40	< H <=	100	< VH <=	200	< UH	
GPT					N <=	40	< H <=	100	< VH <=	200	< UH	
TTT					N <=	5	< H <=	10	< VH <=	15	< UH	
ZTT					N <=	12	< H <=	24	< VH <=	36	< UH	
T-BIL					N <=	1.2	< H <=	2.4	< VH <=	3.6	< UH	
ALB	VL <=	3	< L <=	3.9	< N <=	5.1	< H <=	6	< VH			
TP	VL <=	5.5	< L <=	6.5	< N <=	8.2	< H <=	9.2	< VH			
T-CHO	VL <=	90	< L <=	125	< N <=	220	< H <=	255	< VH			

4. 前処理

今回実験で使用するデータは，1 章で述べたアクティブマイニングプロジェクトのために千葉大学医学部付属病院の医療情報部及び第一内科から提供された，肝炎・肝硬変患者のデータである．

このデータの中から解析目標に合わせて，患者番号・検査日・各種検査項目・IFN 効果を属性として選択する事にした．以下で，各属性の前処理について述べる．

4.1 検査日

使用するデータの中には 10 年以上に及ぶ検査履歴を持つ患者も存在するが，古いデータを元に予測するのは現実的ではない．そこで，IFN 投与開始日以前 5 年間 (1826 日) のデータのみを使用することにした．また，今回は同一患者の 2 点間の変化について注目したいのだが，検査日は患者毎に疎らであり，日単位で扱うと検査日の組み合わせがほとんど一致せず，多くの患者をカバーするルールがあまり生成されないことが予想される．さらに肝炎は比較的緩やかに進行する疾病であるので，その変化についても巨視的に捕らえることが望ましい．これらの理由により，IFN 投与開始日からの相対日数を 4 週間（28 日間隔）で月単位に離散化することにした (1826 日間→ 66ヶ月間)．

4.2 検査項目

解析対象の医療データには多くの検査項目があるが，今回は，一般に重要と考えられ，またデータの出現頻度の高い以下の検査項目を選択した．

GOT,GPT,TTT,ZTT,T-BIL,ALB,TP,T-CHO

これらの検査値は，実数値のまま扱うと探索空間が非常に大きくなり，また得られるルールの内容が非常に限定的になり，実用的でないことが予想されるので，医師の作成した離散化指標 (表 1 *5) を元に離散化することにした．ここで，各離散化指標の意味は以下の通りであり，使用に際して離散値は順序関係を保持しながらコード化することにした．

VL	=	very low	: 1
L	=	low	: 2
N	=	normal	: 3
H	=	high	: 4
VH	=	very high	: 5
UH	=	ultra high	: 6

4.3 インターフェロン効果

IFN の効果判定の指標としては一般に血液検査項目の 1 つである GPT の値が使用されている．今回使用した GPT の値による IFN 治療の効果判定基準 [飯野 99] を表 2 に示す．図 1 はある患者の GPT グラフであるが，表 2 の効果判定基準により，この患者は著効ということになる．

表 2: C 型肝炎に対する IFN 治療の効果判定基準

著効	IFN 投与終了後，6 ヶ月以内に GPT が正常化し，その後 6 ヶ月以上正常値が持続した例
有効	IFN 投与終了後，6 ヶ月以内に GPT が正常上限値の 2 倍以下に改善し，その後 6 ヶ月間以上正常上限値の 2 倍以下を持続した例
悪化	IFN 投与終了後，6 ヶ月間の経過で，投与前に比して，GPT が明らかに憎悪した例
不変	上記のいずれにも属さない例

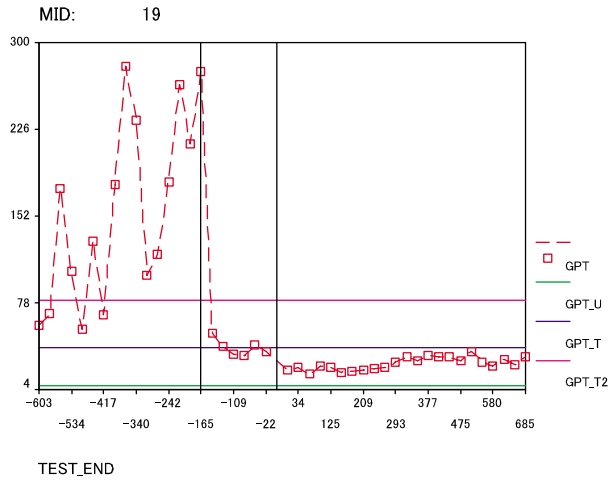

図 1: 著効な患者の GPT：縦軸は GPT の値 [IU/L]，横軸は IFN 投与終了日からの相対日を表わす．2 本の縦線の間が IFN 投与期間であり，GPT_U,GPT_T,GPT_T2 はそれぞれ正常下限 (7)，正常上限 (40)，正常上限の 2 倍 (80) である．

この効果判定基準を基に、IFN 効果が著効または有効である患者のデータを正例，それ以外の患者のデータを負例として扱うことにした．データ構成を表 3 に示す．

*5 この指標は実験後に修正されており最新版ではない

表 3: データ構成

	患者数	レコード数
正例	113	3179
負例	82	2456
合計	195	5635

5. 実験

CProgol(Version5.0)[*6] を使用して，時系列と属性間の関係に注目しながら IFN の効果が著効 (response) または有効 (partial response) となるルールを求めた．またその際，CProgol のモード宣言により，ルールに使用される述語を，“ ifn_effect/1 ”，“ move/6 ”，“ relation/3 ”に限定した．

5.1 実験結果

以下は結果の一部である．実行時間は 3555.5 秒[*7] であった．

```
ifn_effect(A) :-
    move(A,55,52,gpt,B,C), move(A,55,52,t_bil,D,D),
    relation(B,>,D).

ifn_effect(A) :-
    move(A,2,1,got,B,B), move(A,6,5,got,C,C),
    move(A,12,9,got,D,C), relation(D,>,B).

ifn_effect(A) :-
    move(A,2,1,got,B,C), move(A,2,1,got,B,B),
    move(A,6,5,got,D,D), relation(B,>,C).

ifn_effect(A) :-
    move(A,2,1,got,B,B), move(A,2,1,got,C,B),
    relation(C,>,B).
```

これらのルールのサポートは，全て 6 %以下と低いものであった．また実際には，

```
ifn_effect(id71).
```

のように，ルールに全く含まれない正例も存在した．

5.2 結果の評価と考察

得られたルールの中には２点間の変化と共起性が同時に読み取れるものがあったが，そのようなルールの数は少なかった．また投与直前 1〜2ヶ月の GOT に関する変化を扱うルールが多く得られた．

これらの結果を医師に提示し意見を求めたところ，属性間の大小関係などは決定木では得られなかったルールである点で，評価を頂いた．しかし，サポートの低さについて離散化による情報量の低下を指摘された．

サポートの低さに関しては，今回の述語では日付を月単位に離散化して扱っているものの，２点間の日付の組み合わせ自体が少なく，それが原因で事例がうまく拾えていないことが考えられる．そこで，例えば「１ヶ月前」「２ヶ月前」「３ヶ月前」を「およそ２ヶ月前」として扱うといったような述語を背景知識に組み込むことで，今後サポート数が増えるのではないかと予想される．

また，離散化については、医療データは専門性の高いデータである事から，医療に関する知識の乏しい筆者等が独断で離散化を行うのではなく，医療データの専門家である医師との調整が求められている．同様に表 2 の判定基準のみによる効果の判定では，例えば“ 悪化 ”の「明らかに憎悪」など曖昧な表現が多く判定が不十分であるので，今後は他の検査項目も考慮した医師による総合的な判断が必要とされる．

6. まとめ

今回の実験により，ILP によるデータ解析では，これまで使用されてきた手法では発見が困難な形式のルールを得ることや時系列データを扱うことが可能であることが確認された．また，実行時間も実用性に耐えられるものであることが実験により証明された．以上より，ILP システムを使用した属性間の関係や時系列を扱ったデータ解析が有効であることが分った．

しかし現時点では，得られるルールのサポート数の少なさが目立つ．今後，例えば前章で述べたような述語を背景知識として組み込んだり，データの離散化を医師の意見を元に工夫するなどして，サポート数を増加することが望まれる．さらに、今回は２点間の変化について扱う述語を使用したが、これを３点間または４点間を扱う述語に拡張する事に関して、その影響について検討していきたい．同時に，ILP システムと医療データの相性の良し悪しという問題もあるので，ILP システムを医療データ解析に特化したものに拡張する事も考えられる．

謝辞

本研究にあたり，データの提供及び医学的見地からの御助言を頂いた，横井英人氏をはじめ千葉大学医学部付属病院医療情報部の関係者の皆様方に深く感謝の意を表する．

参考文献

[Rodriguez 00] Juan J. Rodríguez, Carlos J. Alonso, and Henrik Boström: Learning first order logic time series classifiers, In James Cussens and Alan Frisch, editors, *Inductive Logic Programming: 10th International Conference, ILP2000. Work-in-Progress Reports*, pp. 260–275, London, UK, (2000)

[Zem 98] Stefan Zemke: ILP via GA for time series prediction, Dept. of Computer and System Sciences, KTH, report 99-006, (1998)

[Muggleton 95] S. Muggleton: Inverse entailment and progol, *New Generation Computing*, Vol. 13, pp. 245–286, (1995)

[飯野 99] 飯野四郎: C 型肝炎のインターフェロン療法の実際, *Nedical Practice*, Vol. 16, no. 9, pp. 1485–1490, 文光堂, (1999)

∗6 http://www.doc.ic.ac.uk/~shm/

∗7 CPU：Pentium4 2.0GHz，RAM：1024MB