G-017

時系列医療データからの知識発見
Knowledge Discovery in Sequential Medical Data

蔀　亜紀子 † **佐藤　新** † **横塚　志行** †
Akiko Shitomi Arata Sato Shikou Yokozuka

1. はじめに

本研究では，医療データからの知識発見の方法として，時系列で採取された検査値の推移率に着目した解析手法を提案し，実験により有効性を検討する．

医療の分野では，前世紀ごろからカルテなどによって蓄積された患者の過去のデータを解析することによって，疾患のリスク因子を探索し，治療や予防に役立てることに興味が持たれ，ロジスティック回帰 [6] や cox 比例ハザード分析 [7] などの統計学的手法が提案された．これらの手法は，対象とする集団をある一定の期間追跡し，追跡開始時の検査値のデータと疾患が何年後に発症したかという情報を用いて，疾患のリスク因子を算出する．つまり，追跡開始時の一時点のみの検査値の中から疾患リスク因子を算出している．しかしながら，疾患の発症に関して注目すべきであるのは，検査値そのものよりも，むしろ，検査値の時間的推移であると考え，本研究では，検査値の時間ごとの推移率に着目した．

近年の電子カルテシステムの普及など，病院の IT 化により，入院期間中の各検査値を時系列で採取し解析することが可能であることから，医療の現場で本手法は有効であること考える．

2. 提案手法

検査値の推移率を考慮した解析方法として次のような手法を提案する．

時刻 $T+1$ において疾患が発症する患者と，発症しない患者を時刻 1 から時刻 T までに行なわれた検査の値から予測することを考える．

解析対象として n 種類の検査項目

$$\mathbf{x} = (x_1, x_2, \cdots, x_i, \cdots, x_n) \tag{1}$$

を考える．これらの検査の時間 t における値を

$$\mathbf{x_t} = (x_{1t}, x_{2t}, \cdots, x_{it}, \cdots, x_{nt}) \tag{2}$$

で表す．検査の行なわれた時刻 $t = (1, 2, \cdots, j, \cdots, T)$ に対して，

$$x'_i = \frac{1}{T-1} \sum_{j=1}^{T-1} \frac{x_{i(j+1)} - x_{ij}}{x_{ij}} \tag{3}$$

は，検査 x_i の値の時間ごとの推移率の平均を表す．このようにして，新規の因子群

$$\mathbf{x}' = (x'_1, x'_2, \cdots, x'_i, \cdots, x'_n) \tag{4}$$

を構成する．これらの因子群を時刻 $T+1$ に疾患の発症した患者群と，発症しなかった患者群に対して求め，それらを用いて判別分析を行なう．これにより，疾患発症リスクの高い患者と，疾患発症リスクの低い患者を分類する規則が導出される．さらに，変数選択を行なうことによって，推移率の変化に着目すべき検査項目が抽出される．このことによって，より少ない検査で患者を分類することができる．

†株式会社 NTT データ技術開発本部

3. 実験

3.1 慢性ウイルス性肝炎について

本節では，実験で扱う慢性ウイルス性肝炎について説明する．

慢性ウイルス性肝炎は，ウイルスの感染によって肝臓が炎症を起こす疾患である．慢性肝炎の状態では生命に危険を及ぼす事はないが，慢性肝炎が進行し，肝硬変を経て肝臓癌を発症した場合，高い確率で死亡に至る．したがって，慢性肝炎から肝硬変への進行を予防することが治療方針として重要となってくる．

肝硬変は，肝細胞の繊維化が進行することにより起こる．現在，繊維化の進行の度合いを判断するためには，肝臓に針状の採取器具を刺入し，肝組織を直接採取する方法が用いられている．この検査は正確に診断ができる反面，身体に傷を負わせるために，患者への負担が大きい．そのため，比較的容易に実施できる血液検査などの臨床検査の値から肝炎の進行の度合いを判断する方法が求められている．

特に，C 型肝炎では，慢性肝炎の状態のインターフェロン治療が肝硬変に対して高い予防効果を認められている一方で，炎症がさらに進行するなどの副作用も認められている．したがって，肝硬変リスクの高い患者にのみインターフェロン治療を施すことが望ましく，臨床検査の値によって高リスク患者を分類する規則を導出することは意義がある．

3.2 提案手法による慢性ウイルス性肝炎データの解析

本節では，C 型肝炎の患者に対して，将来の肝硬変リスクの高い患者とリスクの低い患者を分類する規則を導出することを目的とし，提案手法による実験を行なう．また，比較のため，一時点における検査値のみを用いて判別分析を行った結果も示す．

解析に用いたデータは，6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02) で公開された医療データである．このデータは，千葉大学病院において 1982 年から 2001 年の間に検査を受けた慢性ウイルス性肝炎の患者を対象としている．今回の実験では C 型肝炎患者に関する臨床検査と肝生検のデータセットを用いた．

臨床検査データセットについては，実行頻度が多く，値が連続数値である 29 の検査を解析の対象とした．このうち，提案手法による解析では入院から 8ヶ月までのデータを，判別分析による解析では，8ヶ月目のデータのみを使用した．解析対象のサンプルとしては，上記 29

種類の検査を少なくとも 2 度は受けている者 144 名を使用した．肝硬変の高リスク者と低リスク者の基準に関しては，新犬山分類において，繊維化の度合いが F3〜F4 になる者を高リスク群，F0〜F2 になる者を低リスク群とした．この分類による高リスク者は 44 名，低リスク者は 100 名であった．

判別分析としては Fisher の正準判別分析法，変数選択としてはステップワイズ変数選択法を用いた．

まず，判別分析と提案手法による将来の肝硬変に関する予測精度を表 1 に示す．なお，予測精度の算出には一つ抜き法を用いた．

表 1: 高リスク者と低リスク者の予測精度（全変数使用）

	判別分析	提案手法
予測精度	69.4 %	73.6 %

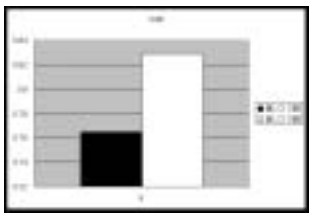
医療データ解析においては，疾患の予測精度は 7 割以上のものが求められる．この基準を満たすという意味で，提案手法により有効な予測規則を導出することができた．

次に，両手法により変数選択を行なった．表 2 に選択された検査項目を寄与率の高い順に 3 つ示す．

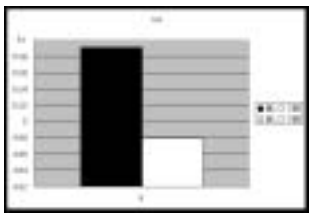
表 2: ステップワイズ変数選択によって選択された検査項目

	判別分析	提案手法
第一変数	CRE	F-A1GL
第二変数	UA	TG
第三変数	ALB	NA

判別分析で選択された検査項目である CRE(クレアチニン)，UA(尿酸)，ALB(アルブミン) はいずれも腎機能に関する検査項目である．一方，提案手法で選択された検査項目は F-A1GL(α 1 グロブリン)，TG(中性脂肪) が肝機能に関する検査項目，NA(血清ナトリウム) が腎機能に関する検査項目である．肝硬変リスクを推測するための検査項目としては，肝機能に関する検査結果を観察することが妥当であると考えると，提案手法によって有効な検査項目が抽出されているといえる．

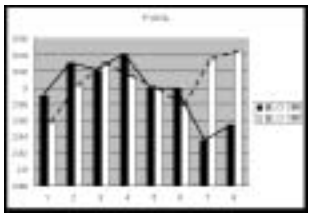


(a)CRE

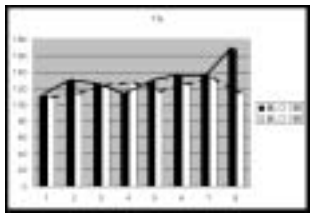


(b)UA

図 1:8ヶ月目の検査における各群の平均値



(a)F-A1GL



(b)TG

図 2：1ヶ月目〜8ヶ月目の検査おける各群の平均値とその推移

4. まとめと今後の課題

本報告では，時系列医療データ解析において，検査値の推移率に着目した解析手法を提案した．さらに，慢性ウイルス性肝炎データへの適用実験を行い，提案手法の有効性を検討した．

今後は，これらの結果の医学的な観点からの妥当性を専門家らと検討していく予定である．また，他のデータに適用した場合の有効性も検討していく．

参考文献

[1] 元田浩，沼尾正行，山口高平，津本周作，”アクティブマイニングの構想と展開”，人工知能学会誌，Vol.17,No.5,pp.615-621(2002)

[2] 横井英人，平野章二，高林克日己，津本周作，里村洋一，”慢性ウイルス性肝炎データに関するアクティブマイニング－病院情報システムにおける知識発見プロセスの実現に向けて－”，人工知能学会誌，Vol.17,No.5,pp.622-628(2002)

[3] 稲田政則，寺野隆雄，”肝機能検査データからの因果モデルの構築”，人工知能学会論文誌,17 巻 6 号 G(2002)，

[4] 鈴木英之進，津本周作，”特集：データマイニングコンテスト”，情報処理,42(5)(2001)

[5] 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)

[6] 市川伸一，岸本淳司，大橋靖雄，浜田知久馬，”SAS によるデータ解析入門　SAS で学ぶ統計的データ解析”，東京大学出版会

[7] 中村剛，”Cox 比例ハザードモデル”，朝倉書店

[8] 浜島信行，”多変量解析による臨床研究”，名古屋大学出版会