

メタ学習による肝機能検査データからの因果モデルの構築

稲田 政則*1 寺野 隆雄*2

*1 虎の門病院臨床検体検査部 *2 筑波大学経営システム科学専攻
{inada, terano}@gssm.otsuka.tsukuba.ac.jp

概要: 本論文では、医療におけるアクティブユーザーリアクションの実施例を示す。本論文の目的は、肝機能検査データの因果モデルの構築と予測システムとしての実用化である。因果モデルは、複数のデータマイニング手法によるメタ学習によって知り得た知見と、ユーザの背景知識を基に構築する。メタ学習のフェーズでは、線形回帰分析とニューラルネットワークを学習用データに適用し、 n -fold 学習にてマイニングモデルが得る。得られたマイニングモデルの予測性能は、テストデータにおける予測値の振る舞いの比較によって評価する。ここで得られた知見に基づいて、ユーザは因果モデルを構築する。因果モデル構築のフェーズでは、ユーザはメタ学習で得られた知見と背景知識とに基づき、重要な変数を選択し、肝機能という潜在変数を導入した因果関係を導く。この因果関係は構造方程式モデリングによって表現する。因果モデルの予測性能は、テストデータにおける予測値の振る舞いを、他のマイニングモデルと比較することで評価する。本論文では、このプロセスから線形回帰モデルと同等で、かつユーザが容易に理解できる優れた因果モデルを得ることができた。単一のデータマイニング手法の適用、あるいは構造方程式モデリングだけの適用では、こうしたモデルを簡単に得ることは難しい。

Developing a Causal Model from Liver Function Tests Data via Meta-Learning Methodologies

Masanori Inada*1, Takao Terano*2

*1 Department of Clinical Laboratory, Toranomon Hospital

*2 Graduate School of Business Sciences, University of Tsukuba

In this paper, we present a novel data mining method, which includes a knowledge evaluation framework with meta-learning about the performance of predicted values. The objective of the proposed method is to obtain high credible knowledge to analyze medical data. The method consists of the following phases: model exploration, model implementation, model validation, and model integration. The model exploration is $m \times n$ parallel KDD processes containing m learning programs and n -fold validations. We evaluate each learned model by comparing its predictive values against unknown data. The model implementation generates practical models from both background knowledge and acquired knowledge from the data and verifies the prediction performance by test data. These two phases are for internal validity of the knowledge. In the model integration phase, we compare the prediction performance of practical models learned by independent different data sources in order to generalize the acquired knowledge. This phase is to generate a generalized model integrated from individual models for the external validity. This paper applies the proposed method to practical medical data analysis and discusses the effectiveness.

1. はじめに

近年、機械学習の分野でメタ学習の研究が進んでいる。学習結果を学習するという定義の下で、複数の研究グループから種々の方法論が提案されている[1]。研究グループの多くは、メタ学習の目標に分類精度の向上[2]、適切な学習アルゴリズムの選択[3]、適切な仮説空間の選択[4]を掲げており、知識評価の側面に焦点を当てた報告は少ない。本研究では、確実な知識の発見・獲得が求められる医療分野へのデータマイニング技術の応用に向けて、メタ学習を知識評価の枠組みとしたマイニング戦略の確立を目標とする。

近年の医学界において Evidence-Based Medicine

(EBM)の考え方[5]が広まりつつある。EBMは、病気の診断・治療に科学的な根拠を求め、確実な医療の実現を目指すものである。EBMの実現には Evidence の蓄積が急務となっている。Evidence の効率的な作成に、医学データベースから有用な知識を発見・獲得する Knowledge Discovery in Databases (KDD) あるいはデータマイニングの技術が注目されている[6]。

一方、医学研究の方法においては、個体差の大きな生体を対象とするために、結論を誤らせるバイアスや交絡の制御が要請される。データマイニングの立場は観察的研究であり、これらが十分に制御されないという問題がある。

データマイニングの医療応用においては、知識評価のプロセスを重視し、仮説検証のプロセスを追加することが必要である[7]。これらのプロセスにメタ学習の枠組みを導入し、医療の専門家による実用モデルの構築を支援することで、データマイニング技術が EBM に寄与できるものと期待される。

2. 医療データ解析とデータマイニング

医療データの解析には、古くから統計学的手法が利用されてきた。医療データの統計解析はこれまで、仮説を検証する立場で利用されてきた。複雑な知識構造を探るという動機付けは弱かったように思われる。データマイニングツールの商用化が進むにつれ、探索型のデータ解析に対する関心が高まりつつある。本章では、統計学的手法を適用する立場とデータマイニングの立場の違いについて論じた後、メタ学習の必要性について言及する。

統計学的手法の利用には、手法を適用する以前に、データの取り方について綿密な計画性が求められてきた。データ獲得後に、計画に示されていない手法を適用することは、後知恵解析として統計学的手法の誤用と批判されてきた[8]。この批判は、人命を扱う医療の分野において、解析結果の盲信や解析者の意図的な情報操作を回避する意味で重要である。

しかし一方、獲得した貴重なデータに対して、一通りの解析に終始することは、データからより多くの知識を抽出する機会を狭めているとの指摘ができる。近年、統計学の分野から派生した「データの科学[9]」の中では、精緻に理論化された統計学を反省し、データによる現象の理解を目指す動きが現れている。さらにデータマイニングの立場では、データの獲得に関して強い要求を設けず、既に存在する大量データを有効に利用することを前提に、属性選択やデータ前処理に多くの方法が提案されている。メタ学習の枠組みでは、同一データに対して異なる学習アルゴリズムを適用し、多面的に得られる知識を有効に活用する試みが行なわれている。データへの理解を深め、より複雑な知識構造を広く探るためには、ここに挙げたようなラフな探索型データ解析が必要である。

しかし、多くの疫学者が指摘する通り、計画性の欠

如した研究において、バイアスや交絡因子によって不適切な結論が導かれることも事実である[10]。また、どのようにデザインされた研究であっても、完全にバイアス・交絡を回避できないことも事実である。故にメタ分析[11]が EBM の重要な方法論となっている。データマイニングにおいてもこの問題への対応が求められる。

データマイニングで得られた知識の確実性を高める方法の一つは、メタ学習の枠組みを導入し、多くの学習結果を比較評価することである。メタ分析が個々の研究結果に示される統計量を統合して全体の方向性を導くように、メタ学習にはデータマイニングで得られたモデルを比較統合することで、知識の一般化を図る役割が期待される。次の章では、メタ学習を知識評価の枠組みとしたマイニング戦略を提案する。

3. マイニング戦略の提案

提案するマイニング戦略は、Fayyad の示した KDD プロセス[12]を拡張し、モデル探索、モデル構築、モデル検証、モデル統合の段階から構成するものである (Fig. 1)。本マイニング戦略には、Chan らが提唱する予測値の性能評価を主体とするメタ学習[2]の枠組みを導入する。適用範囲は、教師付き学習問題すなわち回帰問題と分類クラスに限定される。

3.1 モデル探索

モデル探索の段階は、複数の学習アルゴリズムの適用と n -fold 学習から構成されるメタ学習の機構である。学習アルゴリズムは、解析の目的に適合するものを複数選択する。 n -fold 学習では、所与の学習用データセットを n 個のサブセットに分割した後、 $n-1$ 個のサブセットで再構成される n 個の新たな学習用データセットを各学習アルゴリズムに供する。すなわち、学習アルゴリズムの数を m とすると、KDD プロセスを $m \times n$ 個並列化することを意味する。次に、テストデータを用いて $m \times n$ 個のマイニングモデルの予測性能を比較することで、各モデルの同質性を把握するとともに、学習データへの過学習を評価する。また、各モデルに表現された知識構造を吟味し、実用モデル構築のための知見とする。

3.2 モデル構築

ここでは、モデル探索で得られた知見と解析者の背景知識に基づき、できるだけ理解可能な簡潔な実用モデルを選択あるいは作成する。実用モデルには、モデル探索で得られたマイニングモデルを再利用するか、あるいは、別の検証型モデリング手法（例えば構造化方程式モデリング[13] [14]）で作成する。新たにモデルを作成する場合は、学習データから n-fold 学習によりパラメータを推定する。

3.3 モデル検証

ここでは、テストデータを用いて実用モデルの予測性能を評価する。他の学習アルゴリズムとの比較、および、実用化に関して要求される予測精度を考慮して、実用モデルの採否を決定する。

3.4 モデル統合

モデルの探索から検証までは単一研究者内に完結するプロセスである。テストデータによって予測性能を調査したとしても、同一データソース内で有効な実用モデルが得られたに過ぎず、知識の内的妥当性が評価されたレベルにある。データソース自身にバイアスが存在することは十分に想定される。より普遍的なモデルを得るには、データソースの異なる独立した研究者

間で実用モデルの性能を比較する必要がある。ここでは、複数の実用モデルの予測性能についてメタ学習を行なうことで、知識の外的妥当性が評価される。このプロセスを経て、実用モデルの選択あるいは統合によって普遍モデルを構築する。

3.5 予測性能の比較によるメタ学習

モデルの予測性能は、テストデータに対する各モデルの予測値の振る舞いによって評価する。予測性能として主に、予測精度とモデル間の同質性を評価する。

予測精度は、目標値に対して各モデルの予測値がどれだけ一致しているかの評価である。分類問題ではクロス集計表に整理し正しく分類された割合を求める。回帰問題では相関分析を行ない、決定係数を求める。

モデル間の同質性は、各モデルの予測値の対比較およびクラスタ分析で評価する。分類問題ではクロス集計表による整理、回帰問題では相関行列を基に決定係数の比較を行なう。クラスタ分析では変数のクラスタ化により各モデルの同質性を評価する。

その他の評価として、データセットの個々のケースに関する予測値の同質性評価を行なうこともできる。Chan は分類問題におけるメタ学習手法の効果を見積もるために様々な指標を提案している[15]。特に情報量

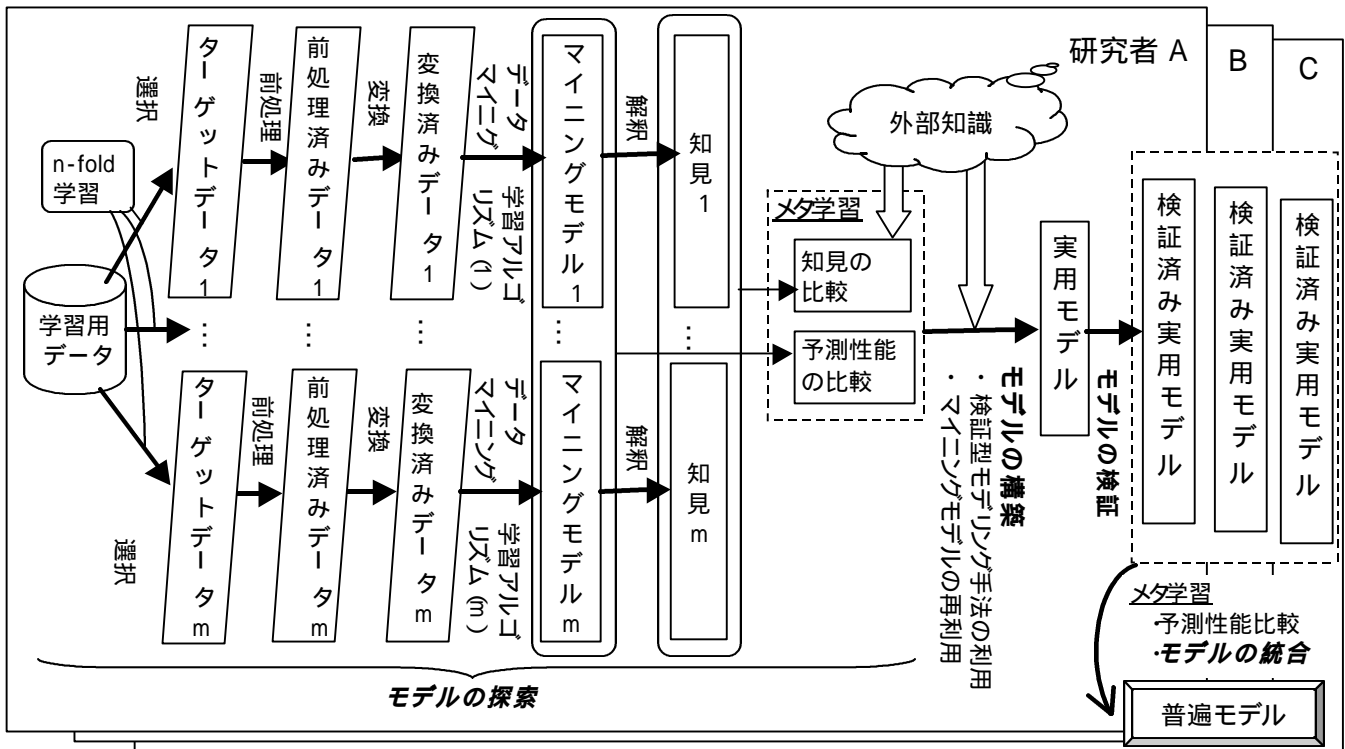


Fig.1 マイニング戦略

理論を用いた指標は、ケース毎に予測値のばらつきを調査しており、ケースの同質性の評価に有効である。回帰問題ではケース毎の予測値の分散が同様の指標となる。ケースについてのクラスタ分析も有効である。

4. 適用事例

本章では、3章で提案したマイニング戦略を実際の医療データ解析に適用した事例を示す。

4.1 対象問題

肝臓の異物排泄機能検査である ICG 検査（インドシアニンググリーン試験）を取り上げ、この ICG 値を他の臨床検査データから予測する問題を対象とする。

4.2 データセット

データセットは、当院臨床検査データベースから ICG 検査の依頼のあったレコードを抽出したものであり、内容は以下の通りである。

目的変数：ICG 値（数量データ，対数処理済）

説明変数：生化学的検査 15 項目（数量データ）

TP, ALB, TTT*, ZTT*, TBIL, DBIL*, AST*, ALT*, LD, ALP, GGT, LAP*, CHE, TCH, A/G*

*の付いている変数には欠測値が存在する。LAP は 95%，それ以外は 8%以下の欠測値を含んでいる。

166 件を学習用データ、74 件をテストデータとした。

4.3 解析ツールと手法

学習アルゴリズムには線形回帰分析とニューラルネットワークを使用した。ここでは SPSS 社の Clementine 6.0.2 を利用した。メタ学習における予測値の比較とクラスタ分析には SPSS 社の SPSS 9.0J を、実用モデルの構築では構造方程式モデリングに Small Waters 社の AMOS 4 を利用した。

4.4 マイニング戦略の結果

本事例は未だ単一研究者による解析に終わっている。よってここでは、モデル検証までの過程を示す。

4.4.1 モデル探索

学習用データを 10 分割し 10-fold 学習用のデータセットを作成し、線形回帰分析とニューラルネットワークを適用した。欠測値は解析対象から除外した。

線形回帰分析では、ステップワイズ変数選択を指定した。得られた 10 個のモデルにおいて、ALB と ZTT が全モデルで、CHE が 9 個のモデルで、DBIL は 7 個のモ

デルで選択された。

ニューラルネットワークでは、トポロジーについて Clementine に組み込まれた経験則による決定機能を利用し、所与のデータをすべて学習に用いるよう設定した。その他の条件は Clementine の標準的な設定とした。10 個のモデルはすべて、入力層 15、隠れ層 4、出力層 1 のトポロジーで学習された。

予測性能についてテストデータを用いて比較した。予測精度は、線形回帰モデルで決定係数 0.54 ~ 0.58 (平均 0.56) と安定した予測精度が確認された一方、ニューラルネットモデルでは、0.29 ~ 0.57 (平均 0.46) と不安定な予測精度であった (Fig. 3-a 参照)。モデル間の同質性については、線形回帰モデル群内では決定係数が 0.92 ~ 0.99 と同質性が認められた。ニューラルネットモデルの群内では 0.53 ~ 0.98 と変動が大きく、過学習の可能性が示唆された。線形回帰モデル群とニューラルネットモデル群の間では、0.56 ~ 0.94 と予測値に類似性が認められた (Fig. 3-a 参照)。クラスタ分析では、凝集法によりモデル間の同質性を調べた。デンドログラムを Fig. 4-a に示す。線形回帰モデルの群内で高い同質性が認められ、ニューラルネットモデルとは明かに性質が異なることが示された。

4.4.2 モデル構築

モデル探索にて、線形近似が有効であることと、線形回帰分析の結果から予測における重要な変数に関して知見を得た。ここではさらに、背景知識を加味して実用モデルを作成した。

線形回帰モデルは、ICG 値を他の臨床検査値で説明する構造を表している。しかし、背景知識を基に因果関係を考慮すると、肝機能という因子が検査データに反映していると考えた方が自然である。そこで、Fig. 5 に示す構造方程式モデルを実用モデルとして構築した。実用モデルには、臨床的な意義を考慮して ALB, CHE, DBIL の各変数を導入した。

Fig. 5 のモデルは、各検査データが肝機能という一つの潜在因子（観測されていない変数）と各独自因子で説明される単回帰モデルから構成される一因子モデルである。ICG 値の単回帰式の係数に 1、切片を 0 に指定することで、肝機能の尺度は ICG 値の同じ尺度とな

る。モデルのパラメータは、最尤法を指定し 10-fold 及びニューラルネットモデルと比較した。

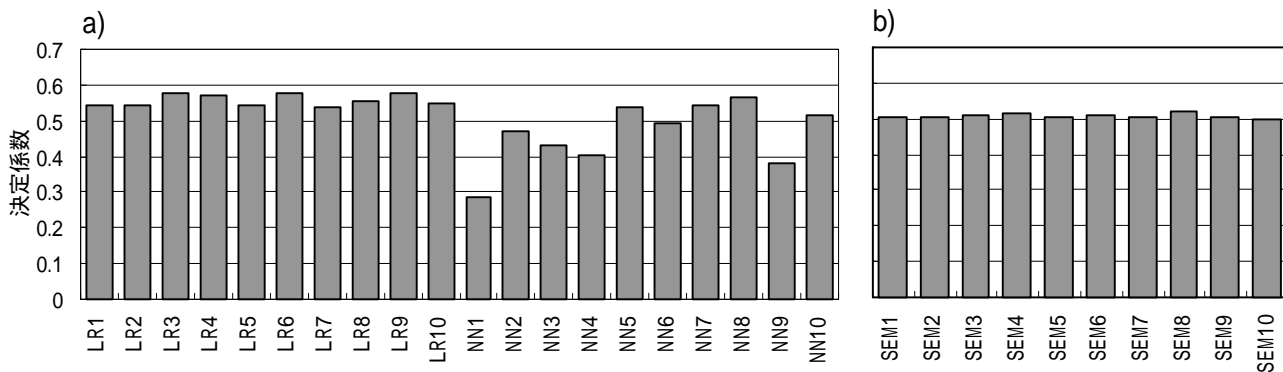


Fig. 2 予測精度の比較 (LR : Linear Regression, NN : Neural Network , SEM : Structural Equation Modeling)

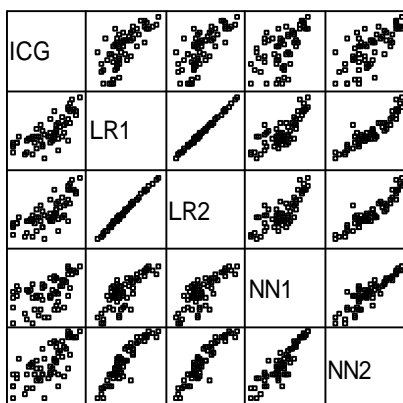


Fig. 3-a) 目標値と予測値の散布図

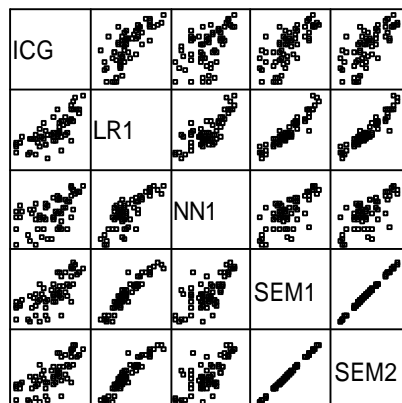


Fig. 3- b) 目標値と予測値の散布図

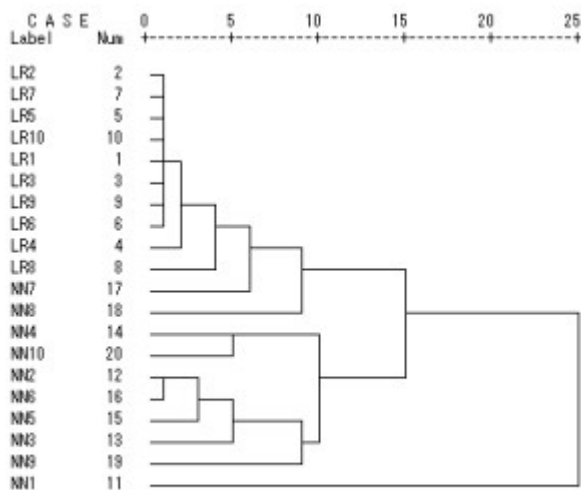


Fig 4-a) 予測値のデンドログラム

学習にて推定した。

4.4.3 モデル検証

パラメータが推定された構造方程式モデルでは、ALB, CHE, DBIL から ICG 値が推定できる。そこで、テストデータにて予測値を計算し、その予測性能を線形回帰お

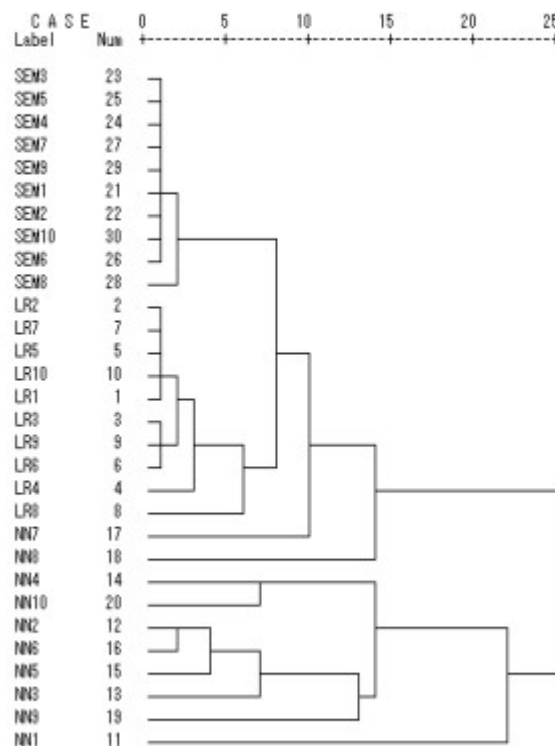


Fig 4-b) 予測値のデンドログラム

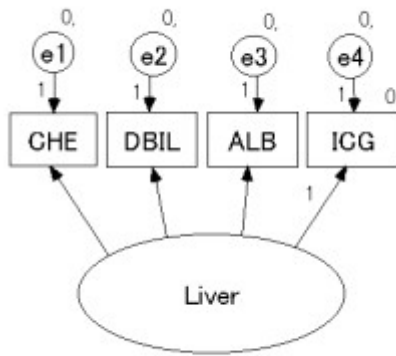


Fig 5) 構造方程式モデル

予測精度は、決定係数 0.49 ~ 0.52 (平均 0.51) と安定しており、線形回帰モデルと同程度であった (Fig. 2-b)。モデル間の同質性は、構造方程式モデルの群内では決定係数がすべて 0.99 と極めて同質であった。線形回帰モデル群との間には 0.81 ~ 0.96 と高い同質性が認められ、ニューラルネットモデル群との間には 0.49 ~ 0.89 と類似性が認められた (Fig. 3-b)。クラスタ分析の結果では、構造方程式モデル群内で高い同質性が認められた (Fig. 4-b)。

4.5 マイニング戦略の効果

本事例では、線形回帰分析とニューラルネットによるモデル探索を行ない、結果的に線形および非線形近似による予測性能を比較するメタ学習となった。結果として、ニューラルネットが過学習した可能性と、線形近似の有効性が判明した。さらに、線形回帰分析での変数選択の結果と背景知識を基に重要な変数を選択し、因果関係に関する考察を経て、構造方程式モデリングによる実用モデルの構築に至った。実用モデルの予測性能を他のモデルと比較することで、その相対的な優位性を評価することが可能となった。

はじめから構造方程式モデリングを利用すべきであるとの批判があるかもしれない。しかし、背景知識のみでは結局、モデル構造の探索が必要となる。また、構造方程式モデリングに備わるモデル検定は、データへの適合性を評価しており、多くの実用モデルは棄却される傾向にある [14]。本事例の実用モデルも、例えばカイ二乗検定で棄却される。実際的な予測モデルの構築に、モデル検定の各指標は必ずしも向いていない。予測値の性能比較を主体としたメタ学習は、実際的なモデル評価の指針を与えた。

5. まとめ

本研究では、データマイニングの医療応用に向けて、メタ学習を知識評価として位置付けたマイニング戦略を提案した。ICG 予測問題へ適用した結果、メタ学習により各モデルの予測値の振る舞いを把握した上で実用モデルを構築し、その予測性能の相対的な優位性を評価することができた。これらは、単一の手法の適用だけでは得られない知見である。検証された実用モデルは、内的妥当性が評価されたものと考えられる。

予測値の評価を主体とするメタ学習は、分類問題にも適用可能である。今後の研究では、分類問題の事例への適用と、外的妥当性の評価について実用モデルのメタ学習に関する事例解析を重ねたい。

参考文献

- [1] R.Vilalta and Y. Drissi: Research Directions in Meta-Learning, Proceedings of the 2001 International Conference on Artificial Intelligence, Ed. H. R. Arabnia, Las Vegas, Nevada, (2001)
- [2] P. Chan and S. Stolfo: On the Accuracy of Meta-learning for Scalable Data Mining, J. Intelligent Information Systems, 8, 5/28 (1997)
- [3] METAL Project homepage, <http://www.metal-kdd.org/>
- [4] M. DesJardins and D. Gordon: Special issue on Bias evaluation and selection. Machine Learning, 20, 1-2 (1995)
- [5] D.L. Sackett, W.M.C. Rosenberg, J.A. M. Gray, R. B. Haynes, and W.S. Richardson: Evidence-based Medicine: what it is and what it isn't. (ed.) BMJ, 312, 71/72 (1996)
- [6] 津本周作: 医学における知識発見手法の可能性, 情報処理, 42, 5, 472/477 (2001)
- [7] 稲田政則, 寺野隆雄: EBMとデータマイニング, SICEシステム・情報部門学術講演会2001論文集 (2001)
- [8] 橋広計, 藤田利治, 佐藤俊哉編: これからの臨床試験. 朝倉書店 (1999)
- [9] 林知己夫: データの科学, 朝倉書店(2001)
- [10] Science: Epidemiology Faces Its Limits, 269, 14, 164/169 (1995)
- [11] B. Mullen: Advanced BASIC Meta-Analysis, Lawrence Erlbaum Associates (1989)
- [12] U. M. Fayyad, et al. eds.: Advances in Knowledge Discovery and Data Mining, AAAI Press (1996)
- [13] K. A. Bollen: Structural Equations with Latent Variables John Willy & Sons (2001)
- [14] 豊田秀樹: 共分散構造分析 [入門編], 朝倉書店 (1998)
- [15] P. Chan: An Extensible Meta-Learning Approach for Scalable and Accurate Inductive Learning, PhD Thesis, Graduate School of Arts and Sciences at Columbia University (1996)
- [16] 豊田秀樹: 共分散構造分析 [事例編]. 北大路書房 (1998)