The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

予測因子候補を抽出するための 血液検査データに対する包括的な分類分析

Comprehensive Classification for Blood Test Data to Extract Prediction Factor Candidates

松井 藤五郎 *1*2 Tohgoroh Matsui　永田 夏海 *1 Natsumi Nagata　吉田 智貴 *1 Tomoki Yoshida　平手 裕市 *1 Yuichi Hirate

*1中部大学 生命健康科学部 臨床工学科
Department of Clinical Engineering, College of Life and Health Sciences, Chubu University

*2中部大学 工学部 情報工学科
Department of Computer Science, College of Engineering, Chubu University

The electric medical record of hospitals stores a huge amount of blood test results and it is expected to estimate the effect of treatments and the cause of diseases by analyzing the results. We have proposed a method to analyze blood test data using Linear Discriminant Analysis (LDA) with a variable selection by non-parametric test. But this method is inconsistent because LDA assumes the data comes from a normal distribution and non-parametric test assumes that the data do not come from any distributions. In this paper, we propose a method to comprehensively analyze blood test data using general machine learning methods. This method includes a new sampling method and a new cross-validation method in a stepwise variable selection method without nonparametric test. We show the experimental results for blood test data and confirm the effectiveness.

1. はじめに

病院の電子カルテには、血液検査の結果が記録されており、これを分析することによって治療の効果や疾病の原因を推測ことが期待される。しかしながら、血液検査には膨大な数の検査項目があり、必要に応じて医師が検査項目が選択して実施されるため、多くの検査項目の値が未観測（欠損値）となってしまう。これから新しく検査を実施してデータを収集する前向き研究においては、検査項目を指定しておくことにより欠損値が生じないようにできるが、過去に実施された血液検査の結果を利用する後ろ向き研究では、欠損値の問題や検査項目名の不統一などが問題となる。

これまでに、血液検査データに対して決定木を用いて膵がん患者と非膵がん患者を部類し、予測因子を抽出する研究 [1] や血液検査データから健康状態を把握して将来予測を行う研究 [2] などが行われているが、これらの研究では医学的知識に基づいて説明変数の候補を選択しており、医学的な因果関係があるとは考えられない説明変数が予測因子として抽出されることがない。

そこで我々は、一般病院における通常の治療業務で記録された電子カルテを対象として、医学的知識を用いずにデータ分析だけを用いて電子カルテに記録された血液検査データから医学的研究における予測因子の候補を抽出すること試みた [3]。この研究では、機械学習の手法として線形判別分析だけを用いており、その他の手法については検討されていない。また、前処理としてノンパラメトリック検定による変数選択を行なっているが、線形判別分析はデータが正規分布になっていることを前提とした手法であり、変数選択と分析手法が整合していない。

そこで本論文では、事前にノンパラメトリック検定を用いた変数選択を行わずに、ステップワイズ法を用いて変数を選択し、分類分析に用いられる一般的な機械学習の手法を網羅的に用いて血液検査データを分析する包括的な分析方法を提案する。また、実際の血液検査データを対象とした実験を行い、その有効性を確認する。なお、本研究は名古屋掖済会病院倫理審査委員会の承認を受けて行われたものである (No. 2016-070)。

連絡先: 松井藤五郎, 中部大学, 愛知県春日井市松本町 1200 番地, 0568-51-1111, TohgorohMatsui@tohgoroh.jp

2. 血液検査データ

名古屋市内にある救急救命センターを有する約 600 床の病院において、2016 年の 1 月 1 日から 4 月 18 日までに死亡した 359 名の、死亡日の 1 年前から死亡日までに行われた血液検査の結果を対象とした。

患者は匿名化されて検査値ごとに 1 レコードとして保存されており、全部で 260,965 件のレコードがある。

2.1 データベースの正規化

まずはじめに、患者 ID と検査日を主キー、検査項目名をフィールドとする表に変換した。

検査項目名と検査値には全角カナで書かれたものと半角カナで書かれたものが混在していたため、文字コードを UTF-8 に変換した上で、NFKC で正規化した。

血液検査は、医師の指示によって行われ、結果が出るとすぐに電子カルテに記録される。ある医師が血液検査を指示した直後に別の医師が別の血液検査を指示した場合には、別のタイムスタンプが記録されて電子カルテに記録される。そこで本論文では、同じ日に行われた（結果が出た）検査を一つの検査として扱い、結果の到着日が同じものを一つのレコードにまとめた。

この正規化によって、621 個の検査項目を持つ 4,408 件のレコードになった。

2.2 ノイズ除去

今回の血液検査データの中には、検査結果の値として、3 つ以上連続したハイフン ---、5 つ以上連続したアスタリスク *****、スラッシュ /、「測定不能」という文字列が含まれている。これらは、検査を実施していないか、検査を実施したが検査値が得られなかったものであると考えられる。

そこで本論文では、これらの値を空欄に置き換え、欠損値として扱った。

2.3 次元削減

実施されることが少ない検査項目は、欠損率が高く、説明変数として利用できない。検査結果が離散値で記録される検査項目は、判別分析など連続値を入力とする機械学習の手法では説明変数として利用できない *a。また、全て同じ値が記録されている検査項目は、目的変数を説明できない。

そこで、欠損率が高い説明変数、連続値でない値を持つ説明変数、全て同じ値が記録されている説明変数を除外する。利用できない説明変数を事前に除外することによって、次元削減の効果がある。本論文では、欠損率の閾値を 90% としてこれらの説明変数を除外し、次元数を 621 から 76 に削減した。

2.4 目的変数

本論文では、アウトカムが明確であり、かつ、サンプルが極端に限定されない死亡に焦点を当て、死亡日直近に行われた血液検査の結果とそれ以外の結果を分類・予測する *b。死亡日またはその前日に行われた（結果が出た）検査のレコードを死亡日直近として値を 1、それ以外の検査結果のレコードを死亡日直近でないとして値を 0 とした。本論文では、これを目的変数として用いる。

このようにして死亡日直近を判定した結果、4,408 個の事例のうち、死亡日直近の事例は 273 件、死亡日直近でない事例が 4,135 件となった。

3. 血液検査データに対する一般的なデータ分析法の問題点

一般的なデータ分析では、欠損値の処理、データを正規分布に近づけるための Box-Cox 変換 [5]、標準化などの前処理を事前に行って機械学習用にデータセットを整え、このデータセットに対してステップワイズ法を用いて変数を選択する。しかしながら、血液検査データに対して一般的なデータ分析の方法を用いるといくつかの問題が生じてしまう。

血液検査は、医師の指示に従って必要最小限の検査だけが実施されており、データに欠損値が非常に多く含まれている。このため、欠損値を含む事例を削除するとほとんどの事例が削除されてしまう。また、このような多くの欠損に対して平均値や中央値を補完すると、分布が大きく変わってしまう。

臨床的イベントの発生をアウトカムとして、その直近の事例とそれ以外を分類する問題を対象とすると、直近の事例よりも直近でない事例の方が多くなる。すなわち、不均衡データとなる。また、臨床的イベント発生直近または直近でないのいずれかの事例しかない患者も存在する。さらに、血液検査の結果には個人差もある。このため、一般的なサンプリング手法を用いることができない。

分類問題における評価方法の一つとして、n 個の訓練事例に対して 1 つの事例を検証データ、それ以外のデータを学習データとしてモデルを学習し、検証データを用いて評価することを n 回繰り返し、その平均を求めるリーブ・ワン・アウト・クロス・バリデーション (LOOCV: Leave-One-Out Cross-Validation) がある。しかしながら、LOOCV を用いると、複数の検査結果がある患者の場合、検証データの患者の別の検査結果が学習データに含まれてしまう。

*a カテゴリカルな説明変数をダミー変数に変換して連続的な説明変数として扱うこともできるが、ダミー変数は正規分布にならないため、本論文ではダミー変数を導入していない。

*b 本論文では死亡のみを対象にするが、本論文の提案手法は死亡以外の臨床的イベントの発生にも適用可能である。

Algorithm 1 POStepwize アルゴリズム（変数増加法）

Require: データ集合 D, 候補変数集合 V_{cand}, 使用変数集合 V_{used}, 最小患者数 m
  for all $v \in V_{cand}$ **do**
    $V_v \leftarrow V_{used} \cup \{v\}$
    $D' \leftarrow D$ から V_v が欠損しているデータを削除
    $D' \leftarrow D'$ に対して患者サンプリング
    if $|D'| \geq 2m$ かつ D' における VIF 統計量が全て 10 未満 **then**
      $D' \leftarrow D'$ に対して 1 クラス Box-Cox 変換
      $D' \leftarrow D'$ に対して 1 クラス標準化
      LOPOCV を用いて V_v の評価値 $E(V_v)$ を求める
    end if
  end for
  $v^* \leftarrow \arg\max_{v \in V_{cand}} E(V_v)$
  if $E(V_{v^*}) > E(V_{used})$ **then**
    $V_{cand} \leftarrow V_{cand} \setminus \{v^*\}$, $V_{used} \leftarrow V_{used} \cup \{v^*\}$
    $V_{used} \leftarrow$ POStepwize$(D, V_{cand}, V_{used}, m)$
  end if
  return V_{used}

4. 提案手法

4.1 提案手法の概要

本論文では、前処理をステップワイズ法の前に行うのではなく、ステップワイズ法において新しい変数を選択するたびに前処理を行い、モデルを評価する分析方法を提案する。ステップワイズ法の中で行われる前処理として、患者サンプリング、1 クラス Box-Cox 変換、1 クラス標準化を提案する。また、選択された説明変数の評価方法として、リーブ・ワン・ペイシェント・アウト・クロス・バリデーション (**LOPOCV**) を提案する。本論文では、このステップワイズ法を患者指向ステップワイズ法 (POStepwize: Patient-Oriented Stepwize Method) と呼ぶ。患者指向ステップワイズ法のアルゴリズムを Algorithm 1 に示す。このアルゴリズムは連続値の説明変数を対象とした分類分析のための機械学習法に広く適用できるため、複数の機械学習法を用いた分析を行うことができる。

4.2 患者指向ステップワイズ法

4.2.1 患者サンプリング

本論文では、死亡日直近のレコードと死亡日直近でないレコードが両方とも存在する患者のデータだけを用い、死亡日直近のレコードが複数存在する場合には最後の（死亡日に最も近い）レコード以外を削除し、死亡日直近でないレコードが複数存在する場合には最前の（死亡日から最も遠い）レコード以外を削除した上で学習を行う。患者ごとに一対のサンプルを選ぶため、本論文ではこれを患者サンプリングと呼ぶ。

ただし、患者サンプリングを行うと、事例数が減少してしまう。そこで本論文では、閾値を導入し、説明変数を選択して患者サンプリングを行なったときに患者数が m 未満になってしまう場合は、この変数を採用しない。

4.2.2 1 クラス Box-Cox 変換

機械学習の手法の中には、データが正規分布であることを仮定する手法もあるため、候補となったすべての説明変数に対して Box-Cox 変換 [5] を用いて、正規分布に近づくように変換する。

Box-Cox 変換は次式で表される。

$$x(\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & (\text{if } \lambda \neq 0) \\ \ln x & \text{Otherwise} \end{cases} \quad (1)$$

ここで、パラメーター λ はプロファイル尤度法によって推定さ

れる。ただし、Box-Cox 変換では変換前のデータが全て正であることを仮定しているため、その変数の最小値の絶対値と 1 を加え、すべての値を正に変換してから Box-Cox 変換を行う。

血液検査データにおいては、臨床的に正常な値は正規分布になるが臨床的に異常がある値は正規分布から大きく外れていると考えられる。そこで、本研究では、死亡日直近でないデータだけを使って Box-Cox 変換のパラメーター λ を推定し、推定された λ を使ってすべてのデータを変換する。本論文では、これを 1 クラス Box-Cox 変換と呼ぶ。

4.2.3 1 クラス標準化

標準化では、変数間の比較を行いやすいように、次式を用いて z スコアを求め、値が平均 0、標準偏差 1 となるよう値を変換する。

$$z(\lambda) = \frac{x(\lambda) - \mu}{\sigma} \tag{2}$$

ここで、μ は $x(\lambda)$ の平均、σ は $x(\lambda)$ の不偏標準偏差を表す。

1 クラス Box-Cox 変換と同様に、臨床的に正常な値は正規分布になるが臨床的に異常がある値は正規分布から大きく外れていると考え、死亡日直近でないデータだけから平均 μ と不偏標準偏差 σ を推定し、推定された μ と σ を使って全てのデータを標準化する。本論文では、これを 1 クラス標準化と呼ぶ。

4.2.4 リーブ・ワン・ペイシェント・アウト・クロス・バリデーション (LOPOCV)

m 人の患者のデータに対し患者ごとにデータ集合を分割し、そのうちの 1 つを検証データ、それ以外を学習データとしてモデルを学習し、検証データを用いて評価することを m 回繰り返す。本論文ではこれをリーブ・ワン・ペイシェント・アウト・クロス・バリデーション (**LOPOCV: Leave-One-Patient-Out Cross-Validation**) と呼ぶ。

こうすることによって、検証データに含まれる患者の別のデータが学習データに含まれないようにすることができる。

4.3 分類分析の手法

本論文では、Python の機械学習ライブラリーである scikit-learn [6] に含まれている次の 13 種類の分類分析法を用いた。

- 線形判別分析 (LDA) [4]
- 二次判別分析 (QDA) [7]
- 二項ロジスティック回帰 (LR) [8]
- 確率的勾配降下法 (SGD) [9]
- 決定木 (DT) [10]
- ランダム・フォレスト (RF) [11]
- 勾配ブースティング (GB) [12]
- AdaBoost (AB) [13]
- サポート・ベクター・マシン (SVM) [14]
- 人工ニューラル・ネットワーク (ANN) [15]
- 単純ベイズ法 (NB) [16]
- ガウス過程 (GP) [17]
- 近傍法 (kNN) [18]

5. 評価実験

5.1 実験方法

患者指向ステップワイズ法によって選択された説明変数を用いたときの LOPOCV の Accuracy（分類精度）の平均を求め、評価した。また、選択された説明変数を全て用いて患者サンプリングを行い、サンプリングされた患者以外のデータをテスト・データとして Precision（適合率、陽性的中率）、Recall（再現率、感度）、その調和平均である F-measure（F-値）を求めた。

本論文では、患者サンプリングの患者数を $m \geq 30$ とし、すべての機械学習手法において scikit-learn のデフォルトの学習パラメーターを用いた。

5.2 実験結果

実験結果を表 1 に示す。LOPOCV は検証データに対する評価値、Precision、Recall、F-measure はテスト・データに対する評価値である。選択された説明変数は、選択された順に並んでおり、3 つ以上の手法で選択されたものを太字で示している。また、説明変数に対応する検査項目の名前を表 2 に示す。

LOPOCV における平均 Accuracy は、最も悪い手法でも 0.849 であり、全ての分類分析法で高い精度が得られた。テスト・データに対しては、Precision が 0.324 から 0.535 だったのに対し、Recall は 0.663 から 0.814 と適合率に比べて高かった。

患者指向ステップワイズ法で選択された検査項目では、カリウム（血液ガス）、pCO_2、アルブミン、CRP、γGPT が 3 つ以上の手法で選択された。

5.3 考察

テスト・データに対する評価では、全ての手法において Precision よりも Recall が高く、誤って死亡日直近と予測してしまう傾向がある。これは、患者サンプリングと LOPOCV によって作られた学習データと検証データのクラスが 1 対 1 で均衡しているのに対し、テスト・データは死亡日直近でない方が非常に多い不均衡データであるためだと考えられる。

選択された検査項目では、v56 のカリウム（血液ガス）が 10 個の手法で選択され、そのうちの 7 つで最初に選択されたことから、死亡に対する予測因子の候補として最も有力であると考えられる。実際に、カリウム（血液ガス）が実施された検査は、全体では死亡日直近の割合が 39.4% であるのに対し、値が 6 以上の事例では 89.0% が死亡日直近のものだった。臨床的には、血液中のカリウムは 3.6 から 5.0 程度の狭い範囲で維持されており、3 以下または 7 以上では致命的になり得る。このことは、提案手法が医学的な知識を用いずに臨床的に意義のある予測因子候補を抽出できる可能性があることを示唆している。

6. 結論

本論文では、医学的知識を用いずに血液検査データに対して一般的に用いられている機械学習の手法を用いて包括的な分類分析を行う方法を提案した。提案手法の一部である患者指向ステップワイズ法は、ステップワイズ法において新しい変数を選択するたびに前処理として患者サンプリング、1 クラス Box-Cox 変換、1 クラス標準化を行い、LOPOCV を用いて説明変数を評価する。これにより、欠損値が多い血液検査データに対して一般的な分類分析のための機械学習手法を網羅的に適用できる。

実際の血液検査データを用いた実験の結果、提案手法は分類精度が高いモデルを学習し、予測因子の候補を抽出できることを確認した。ただし、予測因子の候補として抽出された検査項目の臨床的意義については今後検討する必要がある。

本論文では血液検査データを対象としたデータ分析の可能性を示すためにアウトカムが明確である死亡を対象としたが、提案手法は死亡以外の臨床的イベントに対しても同じように適用できる。今後は、特定の疾病に関連する臨床的イベントを対象として提案手法の有効性を確認したい。

参考文献

[1] 田中, 佐藤, 湊, 松村: データマイニングを利用した血液検査項目からの膵がん診断支援の検討, 信学技報, 111(234):13–17, 2011

表 1　実験結果

Classifier	LOPOCV	Precision	Recall	F-measure	Used variables
LDA	0.894	0.450	0.663	0.536	**v56**, **v42**, **v33**, v18, v37
QDA	0.871	0.466	0.782	0.584	v38, **v63**, v47, v24
LR	0.903	0.411	0.674	0.511	**v56**, **v42**, **v33**, v18, v13
SGD	**1.000**	0.394	0.771	0.522	**v56**, v32, v76, **v40**, v05, v09, v64, v44
DT	0.897	0.383	0.779	0.513	**v56**, **v42**, **v63**, v04, v35, v21
RF	0.929	0.409	0.798	0.540	**v56**, **v63**, v25, v01, v36, v43, v26, v41, v50
GB	0.849	0.395	0.815	0.532	**v56**, **v40**, v62, v43
AB	0.870	0.341	0.677	0.453	**v40**, **v56**, v04, v13
SVM	0.866	0.324	0.763	0.455	**v40**, **v56**, v39
ANN	0.969	0.374	**0.814**	0.513	**v56**, **v42**, **v33**, v23, v37, v17, v51, v64
NB	0.864	**0.535**	0.764	**0.630**	v38, **v63**, v47, v01, v34
GP	0.915	0.369	0.711	0.486	**v40**, **v56**, **v63**, v22, v58
*k*NN	0.875	0.449	0.717	0.552	**v40**, v06, v35, **v63**

表 2　選択された説明変数に対応する検査項目

変数名	検査項目名	選択された回数
v56	カリウム（血液ガス）	**10**
v63	pCO_2 *c	**6**
v40	アルブミン	**4**
v42	CRP *d	**4**
v33	γGPT *e	**3**
v01	eGFR *f	2
v04	尿素窒素	2
v13	白血球数	2
v18	好塩基球数	2
v35	γGT/ALT 比 *g*h	2
v37	血糖値	2
v38	総タンパク	2
v43	クレアチンキナーゼ	2
v47	アミラーゼ	2
v64	pH	2
v05	塩素	1
v06	カリウム	1
v09	MCHC *i	1
v17	RDW *j	1
v21	リンパ球率	1
v22	リンパ球数	1
v23	好中球数	1
v24	好中球率	1
v25	好酸球率	1
v26	MPV *k	1
v32	LD/AST 比 *l*m	1
v34	AST	1
v36	総ビリルビン	1
v39	浸透圧	1
v41	A/G 比 *n	1
v44	尿酸	1
v50	一酸化ヘモグロビン	1
v51	ヘモグロビン（血液ガス）	1
v58	pO_2 *o	1
v62	HCO_3^- *p	1
v76	血糖値（血液ガス）	1

[2] 高橋: 人間ドックの一次予防および二次予防のための判定・教育支援システムの開発に関する研究, 健康医学, 18(1):49–56, 2003

[3] 永田, 松井, 平手: 血液検査データに対する判別分析を用いた生命予後予測因子の抽出, 第 37 回医療情報学連合大会, 3-L-3-PP9-3, 2017

[4] R. A. Fisher: The use of multiple measurements in taxonomic problems, *Annals of Human Genetics*, 7(2):179–188, 1936

[5] G. E. P. Box and D. R. Cox: An analysis of transformations, *Journal of the Royal Statistical Society*, B, 26(2):211–252, 1964

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, et al.: Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, 12:2825–2830, 2011

[7] T. Hastie, R. Tibshirani, J. Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Ed., Springer-Verlag New York, 2009

[8] D. R. Cox: The regression analysis of binary sequences, *Journal of the Royal Statistical Society*, B, 20(2):215–242, 1958

[9] B. Zadrozny and C. Elkan: Transforming classifier scores into accurate multiclass probability estimates, *Proc. of KDD 2002*, 694–699, 2002

[10] J. R. Quinlan: Induction of decision trees, *Machine Learning*, 1(1):81–106, 1986

[11] L. Breiman: Random forests, *Machine Learning*, 45(1):5–32, 2001

[12] J. H. Friedman: Greedy boosting approximation: A gradient boosting machine, *Annals of Statistics*, 29(5):1189–1232, 2001

[13] Y. Freund and R. E. Schapire: A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, 55:119–139, 1997

[14] V. Vapnik, and A. Lerner: Pattern recognition using generalized portrait method, *Automation and Remote Control*, 24:774–780, 1963

[15] F. Rosenblatt: *Principles of Neurodynamics: Perceptrons and The Theory of Brain Mechanisms*, Spartan Books, 1961

[16] T. F. Chan, G. H. Golub, R. J. LeVeque: Updating formulae and a pairwise algorithm for computing sample variances, *Proc. of COMPSTAT 1982*, 30–41, 1982

[17] C. E. Rasmussen and C. Williams: *Gaussian Processes for Machine Learning*, The MIT Press, 2006

[18] N. S. Altman: An introduction to kernel and nearest-neighbor nonparametric regression, *The American Statistican*, 46(3):175–185, 1992

∗c　pCO_2: 二酸化炭素分圧
∗d　CRP: C 反応性タンパク
∗e　γGPT: ガンマ・グルタミン酸ピルビン酸トランスアミナーゼ
∗f　eGFR: 推定糸球体濾過量
∗g　γGT: ガンマ・グルタミール・トランスペプチターゼ
∗h　ALT: アラニン・アミノトランスフェラーゼ
∗i　MCHC:平均赤血球血色素濃度
∗j　RDW: 赤血球粒度分布幅
∗k　MPV: 平均血小板容積
∗l　LD: 乳酸脱水素酵素
∗m　AST: アスパラギン酸アミノトランスフェラーゼ
∗n　A/G 比: アルブミン/グロブリン比
∗o　pO_2: 酸素分圧
∗p　HCO_3^-: 重炭酸イオン