The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

血液検査データに対する包括的な分類分析のためのデータ変換法

Data Transformation for Comprehensive Classification in Blood Test Analysis

吉田 拓倫 *1　松井 藤五郎 *1 *2
Takumi Yoshida　Tohgoroh Matsui

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

*2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

This paper proposes a method to transform data to better fit a normal distribution in medical record analysis. We proposed a method to comprehensively analyze medical record using various machine learning method. Because some machine learning method assume that the data follow normal distribution, we need to transform the data to fit normal distribution. Ordinary method used Box-Cox transform that is a power transform, but it cannot apply negative values. In this paper, we propose to use Yeo-Johnson transform instead of Box-Cox transform, because it can apply negative values. We also propose one-class power transform that estimates the transform parameter only from one class. We confirmed that our proposed method can transform the data even if some values are negative and the transformed data fit normal distribution better than Yeo-Johnson transform using both classes.

1. はじめに

近年、電子カルテの普及が進みデータは電子データとして蓄積されている。電子データには血液検査の結果が記録されており、データを解析することで病気の原因を推測することが期待される。血液検査には膨大な項目があり、医師のみの判断では全ての検査項目を判断することは難しく見落としが発生し、医療ミスへと繋がることが問題となっている。機械学習で検査項目を網羅的に分析することで見逃しを防ぐことも期待される。

これまでに、我々は、複数の機械学習手法を用いて血液検査データを包括的に分析する方法 [1] を提案した。機械学習手法の中には、データが正規分布に従っていることを仮定しているものがあるため、従来手法では、べき乗変換の一種である Box-Cox 変換を用いてデータが正規分布に近づくように変換する。しかしながら、Box-Cox 変換は正の値しか変換できないため、負の値があると適用できないという問題が生じていた。

そこで本研究では、血液検査データに対する包括的な分類分析において、負の値でも変換可能なべき乗変換を用いることを提案する。また、従来手法と同様に、正常値を多く含むクラスだけからべき乗変換のパラメータを推定することを提案する。

2. 血液検査データ

名古屋市内にある救急救命センターにおいて、2016 年 1 月 1 日から 4 月 18 日までに死亡した 359 名の患者から、死亡日から 1 年前と死亡日までの間に実施された血液検査の結果を対象とした。検査結果を患者ごとに 1 日分をまとめ、4,408 件のデータがある。

血液検査データに対する包括的な分類分析 [] では、検査結果を死亡日直近に行われた結果とそれ以外の結果に分類・予測する。死亡日またはその前日に検査が行われた血液検査結果の値を 1（正例）、それ以外の血液検査結果の値を 0（負例）としている。4,408 件のデータのうち、死亡日直近（正例）は 273 件、それ以外（負例）は 4,135 件である。

連絡先: 吉田拓倫，中部大学 工学部 情報工学科, ep15146-1409@sti.chubu.ac.jp

3. 従来手法

従来手法は、血液検査データに対して複数の機械学習手法を用いて分類分析を行う。機械学習手法の中にはデータが正規分布に従っていることを仮定しているものがあるが、実際のデータは正規分布に従っているとは限らないため、従来手法ではデータの分布を正規分布に近づける 1 クラス Box-Cox 変換という手法を提案している。

Box-Cox 変換は Box と Cox により、1964 年に提案された [2] データ全体を正規分布に従うよう変換する手法である。

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、λ は変換パラメータであり、尤度法によって推定される。

Box-Cox 変換はデータ全体が正の値であることを前提としている為、0 以下の値が含まれる場合には、変換を行う前にバイアスとして定数を加えることでデータ全体を正の値へと変換してから適用する必要がある。

通常の Box-Cox 変換において、λ はデータ全体から推定される。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従うが、臨床的に異常な値は正規分布から大きく外れると考えられる。そこで、負例のみからべき乗変換のパラメータ λ を推定することで異常値が含まれる可能性を軽減し、推定された λ を用いてデータ全体を変換する。この手法を **1 クラス Box-Cox 変換**と呼ぶ。

3.1 従来手法の問題点

従来手法の 1 クラス Box-Cox 変換には問題点が 2 つある。

一つ目の問題点は、バイアスを加えてもには Box-Cox 変換では変換が行えないことがあることである。1 クラス Box-Cox 変換ではデータに 0 以下の値が含まれる場合、訓練データか

ら最小値を取り定数を加算することで正の値へと加工する必要がある。テストデータ（未知）の値に訓練データより小さな値が含まれる場合、訓練データから得られた値を加算してもテストデータは正の値に変換されない。また、正例にバイアスより小さな値が含まれる場合においても、正の値に変換されない。Box-Cox 変換は 0 以下の値が含まれる場合には変換ができない為、変換は不可能となる。

二つ目の問題点は、データを加工する際に含まれる定数ことである。定数が大きすぎる場合、データはある値を境に頭打ちになってしまい、最適な定数を定めなければデータは変換出来ない為、定数を定める手順が増えてしまう。

4. 提案手法

4.1 提案手法の概要

本論文では、テストデータに 0 以下の値が含まれている場合においても、正規分布に従っていないデータを正規分布に従わせる変換手法を提案する。提案手法では、Box-Cox 変換の問題を改善した手法である Yeo-Johnson 変換 [3] を用いて、1 クラス Box-Cox 変換と同様に負例のみから λ の推定を行うことで異常値が含まれる可能性を軽減する。本論文において、この手法を 1 クラス Yeo-Johnson 変換ではなく、2 つの手法の総称を用いて **1 クラス Power Trans**（べき乗変換）と呼ぶ。

4.2 Yeo-Johnson 変換

Yeo-Johnson 変換は 2000 年に Yeo と Johnson によって提案された手法で、Box-Cox 変換と同様に、データ全体を正規分布に従うよう変換する手法である。

Yeo-Johnson 変換は次式で表される。

$$x^{(\lambda)} = \begin{cases} \frac{((x+1)^{\lambda}-1)}{\lambda} & \text{if } \lambda \neq 0,\ x \geq 0 \\ \log(x+1) & \text{if } \lambda = 0,\ x \geq 0 \\ \frac{-\left[(-x+1)^{2-\lambda}-1\right]}{(2-\lambda)} & \text{if } \lambda \neq 2,\ x < 0 \\ -\log(-x+1) & \text{if } \lambda = 2,\ x < 0 \end{cases} \tag{2}$$

ここで、λ 変換パラメータであり、尤度法によって推定される。

Yeo-Johnson 変換は Box-Cox 変換を拡張させた手法である。Box-Cox 変換とは異なり、データ全体に 0 以下の値が含まれていた場合においても、データを加工せず変換を行うことができる。

4.3 1 クラス Power Trans

1 クラス Power Trans では、Box-Cox 変換の代わりに Yeo-Johnson 変換を使用する。1 クラス Box-Cox 変換同様に負例のみから λ の推定を行う。1 クラス Power Trans では、未知の値に 0 以下の値が含まれ、かつ、訓練データの最小値より小さい値が入った場合でも使用できる。また、1 クラス Box-Cox 変換とは異なり、定数を設定する必要がない為、データを加工する手順を省くことができる。

5. 評価実験

5.1 実験方法

提案手法の有効性を確認するため、実際の血液検査データに含まれる v52（メトヘモグロビン）を元にテストセットを作成した。v52 の値の平均値を基準に、それぞれの値から平均値を減算し、偏差となるように加工したものをテストセットとした。テストセットの値の分布を図 1 に示す。

テストセットに対して、通常の Power Trans と 1 クラス Power Trans を用いて変換し、正規分布に従っているか比較を行った。正規分布に従っているかの比較には、シャピロ-ウィルク検定 [5][6] を用いた。シャピロ-ウィルク検定は次式で表される。

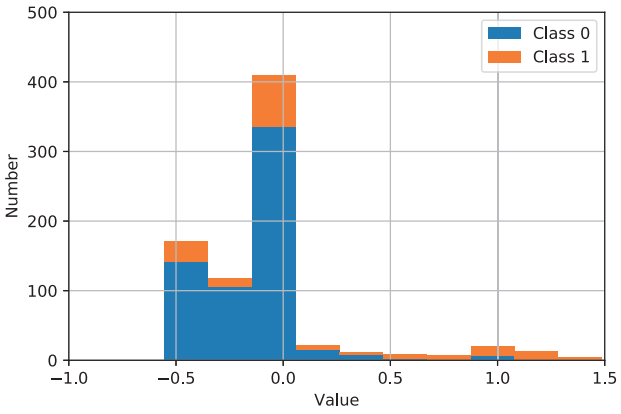

図 1　変換前のテストセット

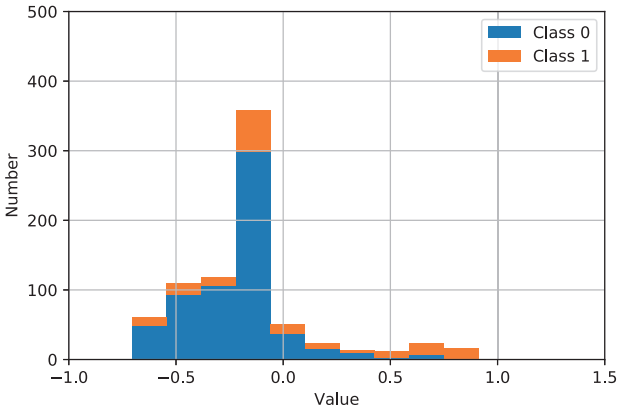

図 2　通常の Power Trans で変換後

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3}$$

$x_{(i)}$ は i 番目の順序統計量、x_i は i 番目の標本、a_i は有意水準であり平均から得られた定数を表す。

シャピロ-ウィルク検定で得られる W は $0 \leq W \leq 1$ で 1 に近いほど正規分布に近い。p-value が 0.05 未満の場合、5% 有意水準での統計的有意性を持つ。

5.2 実験結果

テストセットのヒストグラムを図 1、通常の Power Trans で変換したヒストグラムを図 2 に、1 クラス Power Trans で変換したヒストグラムを図 3 に示す。シャピロ-ウィルク検定による検定結果を表 1 に示す。

表 1 を見ると、通常の Power Trans と 1 クラス Power Trans の p-value は 0.05 より小さい為、統計的有意性がある。1 クラス Power Trans は $W = 0.87$ と、通常の Power Trans の $W = 0.85$ に比べて W が 0.02 大きく、正規分布に近い。

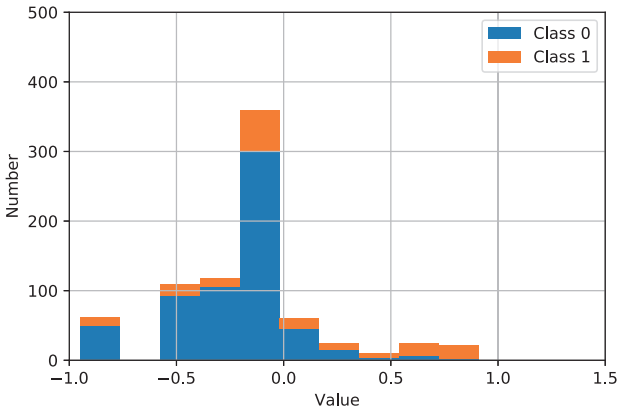

図 3　1 クラス Power Trans で変換後

表 1　シャピロ-ウィルク検定

	W	p-value
Origin Data	0.73	1.26391×10^{-33}
Normal PT	0.85	2.88913×10^{-37}
1-Class PT	0.87	5.26289×10^{-33}

5.3　考察

今回使用した血液検査データにおいて、0 以下の値を持つデータは v59（アニオンギャップ）と v60（ベースエクセス）のみであった。v59 は $W = 0.92$、p-value $= 1.96 \times 10^{-19}$ であり、v60 では $W = 0.96$、p-value $= 2.93 \times 10^{-14}$ と、いずれも正規分布に近いため、変換を行う必要はなかった。そのため、本論文では正規分布に従っていない変数を加工することによって 0 以下の値を含む他のテストセットを作成し、提案手法の有効性を確認した。しかし、一般的には 0 以下の値を含む正規分布に従わないデータが存在するため、Yeo-Johnson 変換を用いることは有効であると考えられる。

通常の Power Trans で変換した値は、最も頻度の高い範囲が 0 付近に移動しているが、全体的な分布の形は元データと変わっていない。そのため、シャピロ-ウィルク検定の W がほとんど変わっていない。これに対し、1 クラス Power Trans で変換した値は、山を低くしてその分を左に移動させており、より正規分布に近い形となっている。

1 クラス Box-Cox 変換では、図 4 のように、検証用データや予測対象のデータに元データの最小値より小さい負の値が含まれる場合、変換を行うことが出来ない。図 4 に示した例の場合、学習用データの最小値 −31.1 の絶対値に 1 を足した値をバイアスとして足しても、検証用データの最小値は正の値にならず、Box-Cox 変換が適用できない。提案手法では、Yeo-Johnson 変換を用いているため、このような場合でも変換でき、バイアスも必要ない。

6.　結論

従来の血液検査データに対する包括的な分類分析では、正規分布を前提とした機械学習手法を用いるために、Box-Cox 変換をベースとしたデータ変換法を用いていたが、Box-Cox 変換は負の値に対応していないため、データが変換できないことがあった。本論文では、Box-Cox 変換の代わりに Yeo-Johnson 変換を用いることによって、未知の値が既知の値より小さい値においても変換でき、広範囲のデータを変換できる手法を提案した。Box-Cox 変換ではバイアスを指定する必要があるが、提案手法はバイアスを指定する必要がないため、データを本手法に入力するのみで変換を行うことができる。

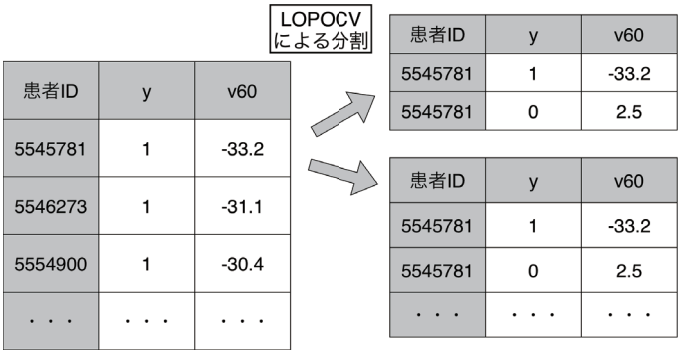

図 4　1 クラス Box-Cox 変換でエラーとなるデータ

正例の中の正常値や負例の中の異常値に関しては処理を施していない為、異常値検出と組み合わせることでより正規分布に近づける為の使用可能な値は残されていると考えられる。現在は、データ全体に対して 1 クラス Power Trans を施している為、シャピロ-ウィルク検定などを用いて正規分布に従っていない場合のみ変換を行った場合と比較したい。

今後は、提案手法を用いたデータ変換が機械学習の予測精度に与える影響について評価したい。

参考文献

[1] 松井, 永田, 吉田, 平手: 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, 第 32 回人工知能学会全国大会, 4C1-OS-27a-04, 2018

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

[3] In-Kwon Yeo and Richard A. Johnson: A new family of power transformations to improve normality or symmetry, *Journal of Biometrika*, 87:954–959, 2000

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

[5] S. S. Shapiro and M. B. Wilk: An analysis of variance test for normality (complete samples), *Biometrika*, 52():591–611, 1965

[6] S. S. Shapiro, M. B. Wilk, and H. J. Chen: A Comparative Study of Various Tests for Normality, *Journal of the American Statistical Association*, 63(324):1343–1372, 1968