Technical Reports on Information and
Computer Science from Kochi
Vol. 1 (2009), No. 10

医療データ解析におけるデータ変換方法

A Method of Data Conversion for Medical Data Analysis

出口 将史　　山下 真　　村岡 道明
高知大学理学部　情報科学コース　村岡研究室

1. まえがき

日常診療から得られる、膨大な医療データをデータベースに記録し、医療データに隠された自明でないルールやパターンを発見する研究は、診断精度の向上や医療費の抑制に貢献する可能性があり、近年、医学分野では重要視されている。しかし、単に長期間の日常の診療データを蓄積しただけのデータベースでは、検査方法の変更や検査値の急激な変化の影響で、直接的に解析を行うことは困難である。

2. 研究概要

本研究では、高知大学医学部に長期間蓄積されている医療データベースを利用して、決定木分析や回帰木分析などのデータマイニングの手法を適用するため、検査データ正規化アルゴリズムを提案する。内容としては、データのクレンジング技術の確立を目的としている。本医療データベースには、1981 年から医療データが蓄積され、現在に至るまでの 28 年間に、検査方法の変わった項目(例：ALP など)が存在する。そのため、検査値の値が過去と現在では全く違う値になっているので、本研究では、血液検査データの正規化アルゴリズムを提案する。

3. 提案手法

本研究では、変換をする方法として医療データの正規化アルゴリズムを提案する。医療データの正規化アルゴリズムは、以下に示すプロセスにより構成されている。

（ｉ）正規分布へ変換（ii）トランケーション（iii）変換式導出の流れになっています。以下に詳しく説明します。

（ｉ）正規分布へ変換

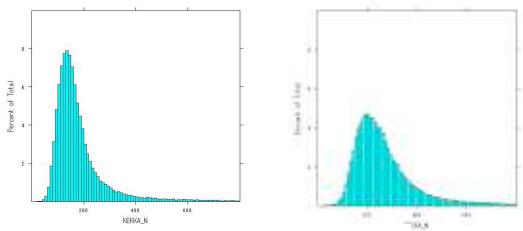

図 1 .ALP ヒストグラム

図 1 に示すヒストグラムは、左図が 2001 年以前の ALP のヒストグラムで、右図が 2001 年から現在に至るまでの ALP のヒストグラムです。この図を見ると、両者で分布が違っていることが確認できます。そのためこのままでは、変換式を導出するための代表値を比較できません。したがって、一度正規分布に変換をして代表値を選出します。

正規分布への変換方法として Box-Cox 変換があります。データの値を x とすると、以下の式であらわあされます。

$$f_\lambda(x) = \int_1^x t^{\lambda-1} dt + 1 = \begin{cases} \log x + 1 & (\lambda = 0) \\ \dfrac{x^\lambda - 1}{\lambda} + 1 & (\lambda \neq 0) \end{cases}$$

λは、一般的に知られているシンプレックス法[1]により求めています。

（ii）トランケーション

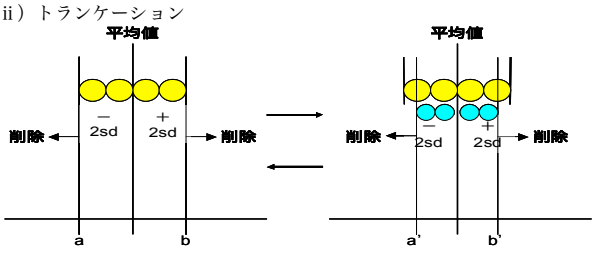

図 2 .トランケーション

標準偏差を sd とします。また、平均値より標準偏差二つ分離れている値を±2sd とします。トランケーションの方法は、先ず正規分布に変換後のデータの−2sd の点を求め、その点より外側を削除します。次に残ったデータで再度±2sd の点を求め、その点より外側を削除します。この作業を±2sd が収束するまで繰り返します。収束するまでトランケーションを繰り返すことで、外れ値が完全に削除されると考えているからです。

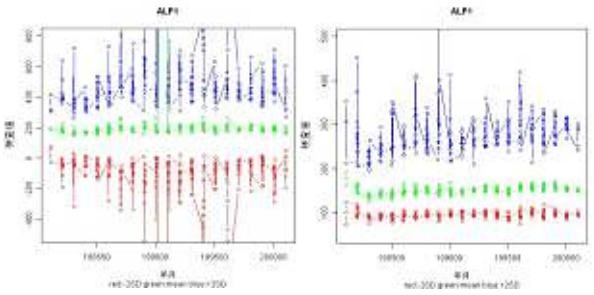

図 3 .トレンドグラフ

左の図が、過去のトランケーション前の ALP、右の図が現在のトランケーションを行った後の ALP のグラフです。この図の縦軸はデータの値、横軸が西暦を示しています。青でプロットされている点が+2sd の値、緑でプロットされている点が平均値の値、
赤でプロットされている点が-2sd の値です。この図で、一列にそれぞれ 12 点ずつ月単位での平均をプロットしています。月単位で見ると、左の図は、縦にプロットされている点が動いているのが分かります。しかし、トランケーションを行った右の図を見ると、月毎のプロットされている点が平均値と-2sd の 2 点は動いていないことが分かります。したがって、このグラフより平均値と-2sd の 2 点を用いることが良いと判断しました。

4. 評価

以上の血液検査データの正規化アルゴリズムより求めた変換式と本学医学部附属病院検査部で従来使われている変換式 を示します。本学医学部附属病院検査部では、検査方法が変わる期間に、同じ患者に 2 種類の検査方法で同時測定を一定期間行っており、その 2 群のデータから算出した変換式を、本研究のリファレンスデータとして比較に用いました。表に示す許容誤差は、トンクスの許容誤差を用いて算出したものです。誤差率は、リファレンスデータと本手法で変換を行ったデータを比較し算出したものです。これを見るとすべての項目において許容誤差範囲以内におさまっているので本手法の有効性が示せました。

表 1．変換式

項目名	本手法算出	リファレンスデータ(*)	誤差率	許容誤差
ALP	y = 1.710499 x − 17.80462	y = 1.46 x + 0	9.4925	13.63637
CK(CPK)	y = 1.147215 x − 7.086342	y = 1.17 x + 0	12.70272	21.66667
LD	y = 0.5511727 x + 8.586403	y = 0.55 x + 0	7.600721	19.44445
AST(GOT)	y = 0.9980327 x + 1.048810	y = 1.09 x + 0	4.181683	13.88889
ALT(GPT)	y = 0.9624585 x + 1.243785	y = 1.1 x + 0	4.869246	8.333335
γ-GT	y = 1.576737 x + 0.8635788	y = 1.77 x + 0	11.6047	14.53489
AMY	y = 0.5616468 x − 1.117373	y = 0.4975 x + 0	11.30881	12.09678
Ca	y = 1.95482 x + 0.05497319	y = 2.0 x + 0	1.658986	3.26087
P	y = 1.637389 x + 0.1653821	y = 1.72 x + 0	0.112605	7.317075

5. まとめ

本研究では、血液検査値データの正規化アルゴリズムを提案した。正規化アルゴリズムを用いて算出された変換式と、現在医学部で使用されている変換式をトンクスの許容誤差を用いて比較した結果、十分な精度を持っていることが示された。本手法を他の医療機関の医療データベースに適用することで、検査値の統一が可能であり、各データマイニングの手法と適用することが可能となります。そのことにより、医療データ解析の発展が望めます。

6. 参考文献

[1] 国沢信, “線形計画と経済”, ダイヤモンド社