F-048

回帰診断に基づく医療データの補間の検討

A Study on Clinical Data Interpolation Based on Regression Diagnosis

中瀬 正和 * 大崎 美穂 † 片桐 滋 † 畑田 由香利 †
Masakazu NAKASE Miho OHSAKI Shigeru KATAGIRI Yukari HATADA

1. はじめに

今日，EBM(Evidence-Based Medicine:根拠に基づいた医療) の概念の普及に伴い，医療情報環境の整備が進められている．さらに次の段階として，蓄えられた医療データから病状把握や治療に役立つ知識を発見する試みもなされつつある．慢性病の検査履歴といった医療時系列データからの知識発見には，時系列解析手法の適用が必要である．しかし，検査の間隔は基本的に一定でないため，等間隔標本化を前提とする多くの時系列解析手法をそのまま適用することはできない．そこで本研究では，医療時系列データを補間して標本間隔を一定にする手法を提案する．

2. 関連研究

医療情報は比較的新しい分野であり，数年から十数年の観察を要する医療時系列データを扱えるようになったのは近年である．このため，医療時系列データの補間に関する研究は多く見られないが，一般的な補間手法としては以下がある．

最も簡易なものは一定の幅を持つ時間窓をスライドし，窓内の平均値や回帰直線による推定値を補間値とする手法である [1]．これらには窓幅や重なりの決め方，緩急様々な病状変化に関わらず一定の窓幅で良いのか等の問題がある．一方，データ生成メカニズムを反映させるべく，自己回帰モデルによる補間手法もある．しかし自己回帰モデルのようなパラメトリックモデルの多くは，定常過程への適用が前提となる．従って，病状及び検査頻度の変化に伴い，本質的に非定常かつ不均一標本化時系列である医療時系列データへの適用は困難である．

以上を考慮すると，時系列解析の第 1 のステップとして，欠損値補間による標本間隔の均一化と定常的部分時系列への領域分割とが必須であり，またこの処理が，データの生成過程に関して比較的緩い前提のみを持つ統計量を用いて行われるべきことが導出される．そこで，本研究では回帰診断 [2] に基づく領域分割と領域内の回帰直線当てはめを提案する．本来，回帰診断は外れ値の除去を目的として提案されたが，外れ値で領域分割を行えば，過度な細分化の危険はあるものの，少なくとも分割境界に病状の転換点を含むことができ，一定の病状に対応する定常的データからなる領域への分割が可能となるものと，我々は考えた．

3. 回帰診断に基づく補間の提案

提案手法では，回帰診断と領域分割の再帰適用し，領域ごとに当てはめた回帰直線を用いて医療時系列データを補間する．以下では，まず回帰診断の概要を示し，次に提案手法の詳細を述べる．

回帰診断とは，データに対する回帰モデルの当てはまりの良し悪しを判定する方法である [2]．単回帰の場合，一変量 x_i から一変量 y_i を推定する式 (1) の回帰直線を当てはめる．次に，観測値 y_i と推定値 $\hat{y}_i$ の残差平方和の算出と最小二乗法の適用を経て，係数 β_0，β_1 を得る．なお，回帰直線では，残差項の確率密度関数推定と最尤法の適用でも同様の結果となる．最後に，てこ比や Cook の距離といった測度で回帰直線の当てはまりの良し悪しを判定する．

てこ比は，点 $(x_k,\ y_k)$ が外れ値として回帰結果に与える影響を表し，式 (2) のように，1 点あたりの影響度合い $\frac{1}{n}$ (n はデータ数)，X 方向の外れ度合い $(x_k - \bar{x})^2$ ($\bar{x}$ は平均) で定義される．てこ比が大きい点は外れ値と見なされる．Cook の距離 D_k は，y_k を除いた場合の推定値 $\hat{y}_{sub}$ と y_k を含めた場合の推定値 $\hat{y}$ の差分を分散の推定値 s^2 で正規化したものである．式 (3) の式変形により，Cook の距離は，1 点あたりの影響，X 方向の外れ度合い，Y 方向の外れ度合い $(e_k(s))^2$ ($e_k(s)$ は標準化残差) という 3 つの判定基準から構成される．Cook の距離が大きい，即ち，データ数が少なく，X 方向の外れ度合いが大きく，Y 方向の外れ度合いが大きいと，対象の点は外れ値と見なされる．

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

$$h_{kk} = \frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}} \tag{2}$$

$$D_k = \frac{(\hat{y}_{sub} - \hat{y})'(\hat{y}_{sub} - \hat{y})}{2s^2} = \frac{1}{2}\frac{h_{kk}}{1 - h_{kk}}(e_k(s))^2 \tag{3}$$

提案手法の処理手続きを図 1 に示す．提案手法では，図中の **(a)** のように，まず時系列データのある区間に回帰直線を当てはめる．次に，Cook の距離の 3 つの判定基準に関して，全てが閾値を超えた場合に対象点を外れ値と判定し，**(b)** 最も外れ度合いが大きい点を見つけ出す．そして，**(c)** この点を境界として元の区間を分割し，**(d)** 同じ手続きを新たな区間に再帰的に適用する．以上により，領域分割と回帰直線当てはめがなされる．上述の手続き後，領域ごとに回帰直線で欠損値を推定すれば補間可能である．

4. 評価実験

4.1 実験条件

提案手法の有効性を検証するため，提案手法，および，従来手法 (時間窓内の平均値で補間する手法) を実際の医療時系列データに適用し，両者の性能を比較する実験

*同志社大学大学院工学研究科知識工学専攻
†同志社大学理工学部情報システムデザイン学科

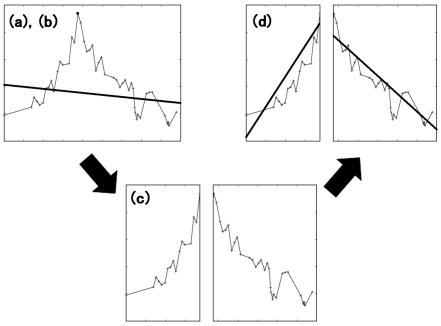

図 1: 提案手法の処理手続き

を行った．入力には，慢性肝炎の検査履歴データセットを用いた [3]．データセットは 607 名の患者ごとに得られた様々な検査値の時系列から成るが，病状把握に重要な検査値 G-GTP を選定した．今回は提案手法を試す第一歩として，病状の転換点が明確に分かる患者 14 名の G-GTP 時系列を用いることにした．

提案手法では，3 つの判定基準の閾値をあらかじめ設定する必要がある．今回は，以下のように様々な閾値で性能を探ることにした．1 点あたりの影響度合いの閾値 $Tn = 1/3$．X 方向の外れ度合いの閾値 Tx, $0.01Rx \leq Tx \leq 0.25Rx$ (Rx は X のレンジ)．Y 方向の外れ度合いの閾値 Ty, $0.01s \leq |Ty| \leq 0.8s$ (s は標準化残差)．従来手法では時間窓の幅と重なりの事前設定が必要である．28 日以上 56 日未満の検査間隔が最も多い性質に基づき [1]，窓幅 56 日，重なり 14 日とした．

性能の評価基準は次のように行った．本データセットの検査間隔の頻度を調べたところ，28 日以上 56 日未満を中心に大きなピークが，7 日未満に小さなピークが現れた．これは，小康状態では 1〜2ヶ月に 1 回の定期的な検査を行い，病状が悪化した時は，経過観察のため 1ヶ月に数回の臨時の検査を行うことを意味すると考えられる．そこで検査間隔が 7 日未満であり，かつそれが 1ヶ月継続する場合 (7 日未満の検査間隔が 3 回以上続く場合)，その始点を病状の転換点と見なした．そして，提案手法，あるいは従来手法がこの転換点で領域を分割できたかを評価基準とした．

4.2 結果と考察

実験結果の一例を図 2 に示す．図 2 の上左図は，提案手法により病状が異なる領域を分割できた結果，上右図は同じ条件における従来手法の結果である．また下左図と下右図は各々，領域分割の成否を詳しく調べるために上左図，上右図の一部を拡大表示したものである．

図の下側において，灰色の領域の境界線は病状の転換点を表す．下左図より，判定基準の閾値が適切であれば，提案手法は灰色の領域の開始点と終了点を境界として明確に見出し，異なる病状の領域ごとに回帰直線を当てはめていると言える．一方下右図のように従来法ではこれらの転換点そのものを出せないことがある．

今回扱った 14 名の時系列のなかで，病状の転換点は 26 箇所見られた．そのうち提案手法では 23 箇所，従来手法では 21 箇所において領域分割がなされた．ドメイン知識を活用した従来手法に比べると，閾値の設定がアドホックにとどまっていた提案手法においてもやや優れた結果が得られたことから，提案手法の基本的な有効性を確認できるものと考えられる．また，図 2 に示したデータにおいては，従来法が提案法の 5 倍程度という大きな水準で過剰な分割，即ちフォールスアラームを発生させており，この点も提案法の有効性を支持している．

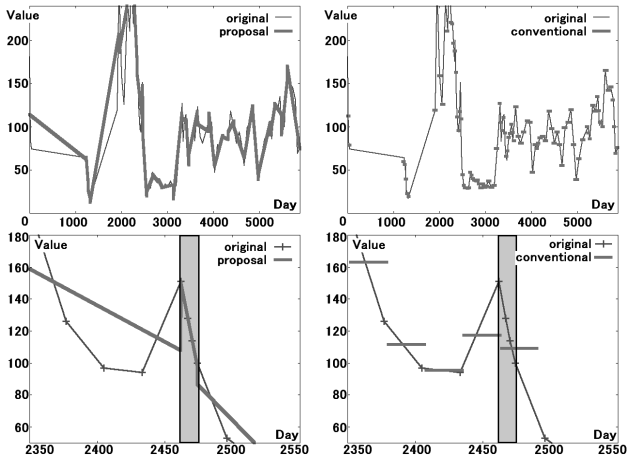

図 2: G-GTP 時系列の補間結果 (左側:提案手法による補間，右側:従来手法による補間，上側:時系列全体，下側:時系列の一部を拡大表示したもの，また灰色の領域の境界線が病状の転換点である)

以上より，提案手法の可能性が示唆された．一方で，提案手法の性能は閾値に大きく依存するという問題が見られた．今後，閾値の自動設定の枠組み，あるいは，3 つの判定基準を個別に扱わず Cook の距離の大小のみで判定する方法を考える必要がある．また，今回は性能の定量的な評価に至らなかったため，性能を数値で見積もる必要がある．これらを今後の課題としたい．

5. まとめ

極端な不等間隔性を持つ医療時系列データに対処するため，回帰診断と領域分割の再帰適用に基づく補間手法を提案した．提案手法および，時間窓内の平均で補間する従来手法を肝炎の検査履歴に適用した結果，提案手法の基本的な有効性が示された．提案手法と従来手法の定量的な性能比較等が今後の課題である．

参考文献

[1] M.Ohsaki et al.: A Rule Discovery Support System for Sequential Medical Data - In the Case Study of a Chronic Hepatitis Dataset -, ECML/PKDD-2003 WS on Discovery Challenge, pp.154-165 (2003).

[2] R. D. Cook, and S. Weisberg: Diagnostics for heteroscedasticity in regression, Biometrika, Vol. 70, No. 1, pp.1-10 (1983).

[3] S. Tsumoto: Hepatitis Dataset for Discovery Challenge, http://lisp.vse.cz/challenge/ecmlpkdd2002/ (2002).