ニューラルネットワークと線形回帰分析のハイブリッド解析法を用いた

データ解析 - Web 版利用による実データ解析法 -

Data Analysis by the Hybrid Approach to Neural Networks and Linear Regression

浅野 美代子 (大東文化大学) ・ 椿 広計 (筑波大学)

Miyoko Asano (Daito Bunka University) and Hiroe Tsubaki (University of Tsukuba)

1. はじめに

ニューラルネットワークモデル解析がすぐれている内容と理由の解明を行い (浅野, 2002, 浅野・椿,2002, Asano et al , 2002), 「解釈可能性(Interpretability)」と予測精度の良さを併せもつデータ解析法としてニューラルネットワークモデルと線形回帰分析のハイブリッド解析法 (浅野・椿, 2003) の提案を行った. 2007年７月にWeb版が完成し、インターネット利用による解析を提供している. この解析法を用いて、モデル選択を行い探索的なデータ解析を行うことができるのでWeb版でも検討することが可能になった. 本稿では、Web版ニューラルネットワークモデルと線形回帰分析のハイブリッド解析法を用いてデータ解析事例を示す.

2. ニューラルネットワークモデルと線形回帰分析のハイブリッド解析 (浅野･椿, 2003) とその解法

x を p-次元入力データ, y をこのデータに対応する目的変数とする. ニューラルネットワークモデルと線形回帰分析のハイブリッド解析に対応するモデルは次式である.

$$y=\beta_0+\sum_{j=1}^{p}\alpha_j x_j+\sum_{i=1}^{q}\beta_i f_i\left(\sum_{j=1}^{p}w_{ij}x_j\right)+\varepsilon,\ \ \varepsilon\sim N(0,\sigma^2), \qquad (1)$$

ここで, $f(t)=\dfrac{1}{1+\exp(-t)}$ は, シグモイド関数.

パラメータの算出は次の3ステップ手順を用いて求める.

Step 1: By analysis of the neural network using Kurita (1990)'s method of minimizing the Akaike Information Criteria (1974), we decide the numbers of the hidden layers (q). The formula of the Akaike Information Criteria (AIC) in our case is as follows:

$$AIC = n(LnRSS/n)+2((p+1)*q+q)$$

Where, RSS is the residual sum of squares. The output values of q units of the hidden layer are added as the input variables to B of model (1).

Step 2: A stepwise analysis of multiple regression is carried out based on F-test analysis using significant level α in the forward selection and backward elimination methods. Here, the p

variables of A of Model (1) are all included. The variable selection process focuses on the hidden layer's output q variables from Step 1 using forward selection. The q' variables are to be selected from the q variables. Therefore, the input variables are now $p + q$'.

Step 3: In order to select the appropriate input variables from the original p input variables, we do an F-test analysis using the usual significant level α in the backward elimination method. All q' variables from Step 2 are included. The model selection process concentrates on the p variables and p' variables are to be selected from the p variables.

ステップが終ると, p'と q' が得られ (1) 式は以下の式になる.

$$y = \beta_0 + \sum_{j=1}^{p'} \alpha_j x_j + \sum_{i=1}^{q'} \beta_i f_i\left(\sum_{j=1}^{p} w_{ij} x_j\right) + \varepsilon$$

3. 実例:決算期変更のある企業財務データの売上高解析

ここでは、松下電器産業 (株) の売上高の解析を行う手順を示す. データは1974年度から2003年度までの29年間で解析を行った. 但し, 1986年11月まで11月決算であり、以降は3月決算に変更があった. 決算期変更が行われた財務データのこれまでの解析では、解析前にデータ加工を行なっていたが、本システムの解析では事前の加工は行わずに、すべてのデータを用いて解析を行う. 解析にあたり、Excel で CSV ファイルを作成する.

本システムでは、ニューラルネットワークモデルと線形回帰分析のハイブリッド法による解析を行えると同時に、線形回帰分析も行うことができる. 解析メニューの中から、ニューラルネットワークモデルと線形回帰分析のハイブリッド法を選択する. CSV ファイルを指定 (図1参照) して、入力変数、目的変数とパラメータを設定する. 設定項目を表1に示す.

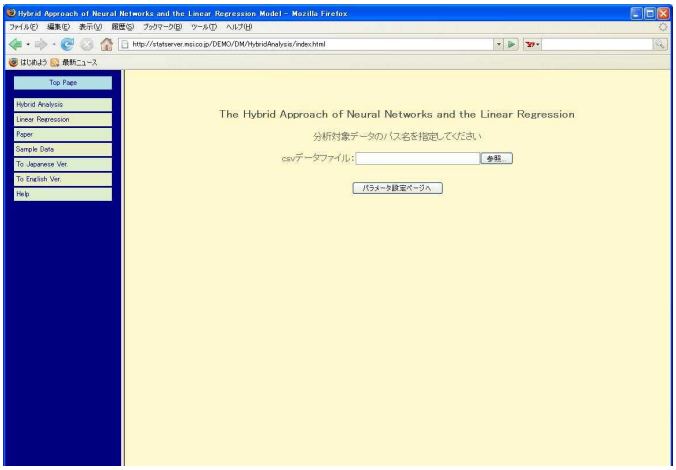


図1. ハイブリッド解析法 分析対象データ入力画面

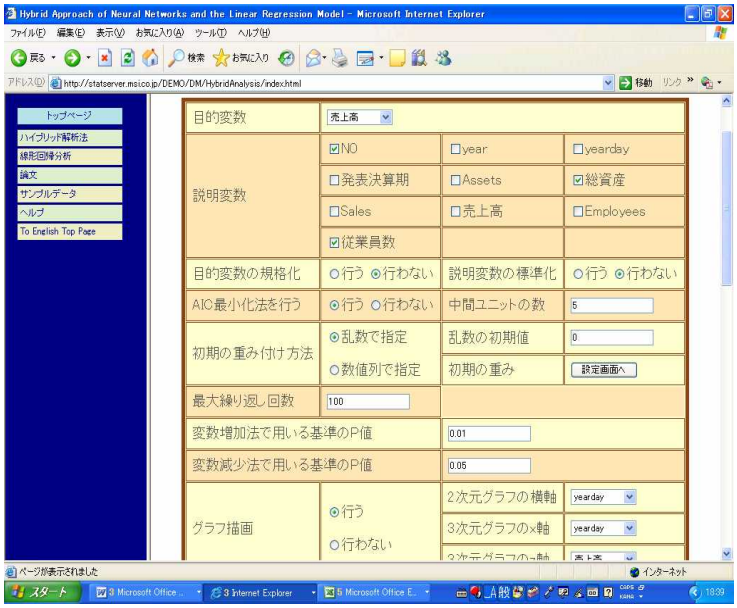
図1の分析対象データ入力画面の操作では CSV ファイルは参照ボタンをクリックしてファイル一覧より指定できる.

現在のWeb版ニューラルネットワークモデルと線形回帰分析のハイブリッド解析法システムは,日本語と英語の２種類の言語構成であるが,将来はその他様々の言語に対応していく予定である.

表1. パラメータ設定画面の入力項目 (使用説明書より引用)

パラメータ名	説明
目的変数	目的変数として用いる列を指定
説明変数	説明変数として用いる列を指定
目的変数の規格化	目的変数を最大値 1,最小値 0 で規格化を行うか指定
説明変数の標準化	説明変数を平均 0,標準偏差 1 で標準化を行うか指定
AIC 最小化法を行う	複数 NNET モデルを作成して、その中で AIC がもっとも小さいものを選択するか、中間層ユニット数を指定するか指定
中間ユニット数	AIC 最小化法を「行う」とした場合、中間ユニット数 1, 2, ・・・ 指定した数までの NNET モデルを作成し「行わない」とした場合、指定した数を中間ユニット数とした NNET モデルを作成, デフォルト値は 2
初期の重み付け方法	「乱数で指定」とした場合、初期の重み付けは乱数を発生させて初期の重み付けを行い、「数値列で指定」とした場合、初期の重みを数値列で指定
乱数の初期値	初期の重み付けに用いる乱数のシードです。0～ 1023 までの整数を半角で入力、デフォルト値は 55
初期の重み	ボタン「設定画面へ」をクリックすると、初期の重み設定のウィンドウが開く, 開いたウィンドウに初期の重みを 「(半角実数),(半角実数),(半角実数),・・・」のフォーマットで入力、また、「説明変数」、「中間ユニットの数」を変更すると設定した内容が削除される
最大繰り返し回数	NNET モデルを作成するときの最大繰り返し回数, 1 以上の整数を半角で入力、デフォルト値は 100
変数増加法で用いる基準の P 値	変数増加法を行うときに用いる基準の P 値: 0 以上の実数を入力, デフォルト値は 0.01
変数減少法で用いる基準の P 値	変数減少法を行うときに用いる基準の P 値: 0 以上の実数を入力, デフォルト値は 0.05
グラフ描画	「行う」を指定した場合、グラフとテーブル結果を出力し,「行わない」を指定した場合、テーブル結果のみ出力
2 次元グラフの横軸	2 次元グラフの横軸に用いる列を指定
3 次元グラフの x 軸	3 次元グラフの x 軸に用いる列を指定
3 次元グラフの z 軸	3 次元グラフの z 軸に用いる列を指定

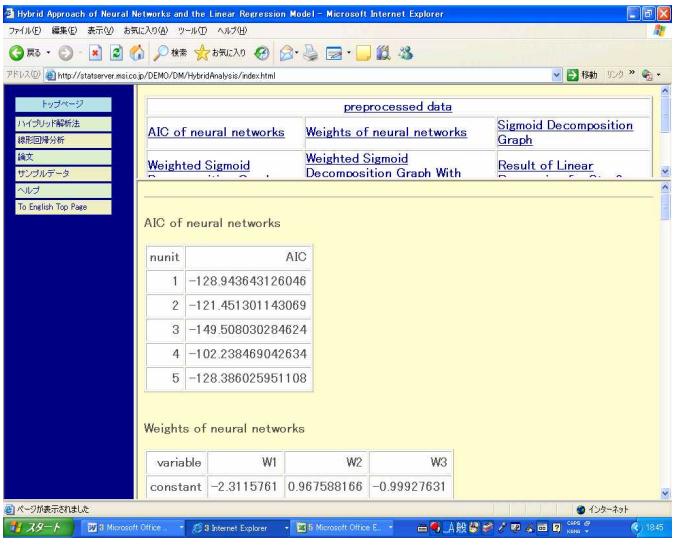
パラメータ設定画面の入力において重み初期値は、2 通りの方法で指定可能である. また、Step2 と Step3 の回帰分析でのモデル選択の有意水準が指定可能である. 実際のパラメータ設定画面を図 2 に示す. AIC が最小である中間層のユニット数を用いて解析を行う. Step2 と Step3 の回帰分析モデル選択の有意水準はそれぞれデフォルト値の 0.05 と 0.01 を用いた.



目的変数の売上高をフィールド名リストから選ぶ．但し，フィールド名リストはＣＳＶファイルの1行目の変数名で、変数名には空白や記号は許されていない．

次に，説明変数として，ＮＯ（通番）と総資産と従業員数の3 変数(p = 3)を選択．中間層のユニット数決定のＡＩＣ最小値法を用いるため,ユニットの最大数5個を指定して,ウエイトの初期値は0を指定している．

図2．松下電器産業（株）の売上高の解析設定画面



中間層ユニット数3個(q = 3)がＡＩＣ値最小なので選択され（図3参照），続いてウエイトの一覧表とStep2 と Step3 の結果が出力される．同時にシグモイド分解図、ウエイテッド･シグモイド分解図、続いて、目的変数とウエイテッド･シグモイド分解図が出力され，最後に，2次元および 3 次元の Asano-Bhattacharyya Graph（2006）が，ユーザ指定によって出力される．

図3．ニューラルネットワークモデルの AIC

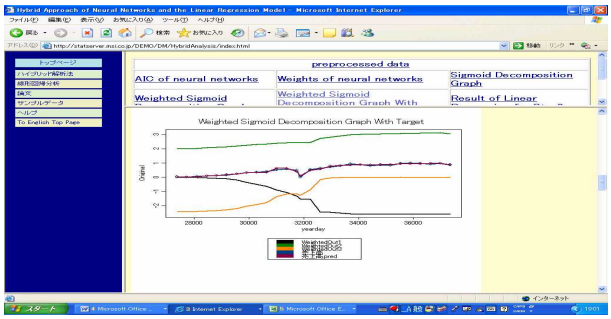


図 3 より中間ユニット数は3個となった目的変数とウエイテッド･シグモイド分解図を図4に示す．図4でドットつきグラフが目的変数の売上高で，この下方には，ユニット1とユニット3のウエイトを乗じた出力値グラフが描かれている．この2つの出力値の和が決算期変更に対応している．

図4．目的変数とウエイテッド･シグモイド分解図

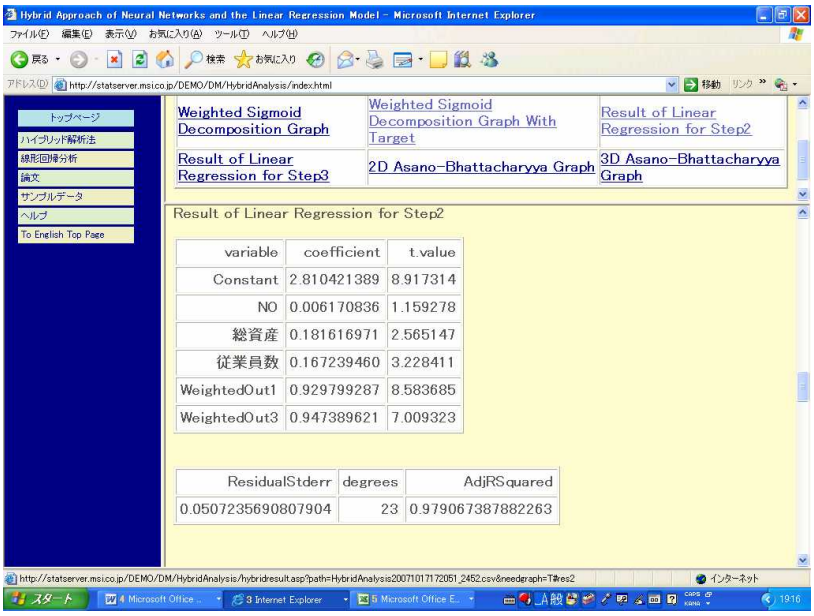


図5. Step2 の結果画面

また，Step2 の結果を図5，Step3 の結果を図6に示す．Step2 では 2 つの中間層出力値（ユニット 1 とユニット 3）が取り込まれたので，回帰分析の説明変数はＮＯ（通番）と総資産と従業員数の3変数（$p = 3$ ）と組み込まれた中間層ユニットの出力値 2 変数（$q' = 2$ ）である．

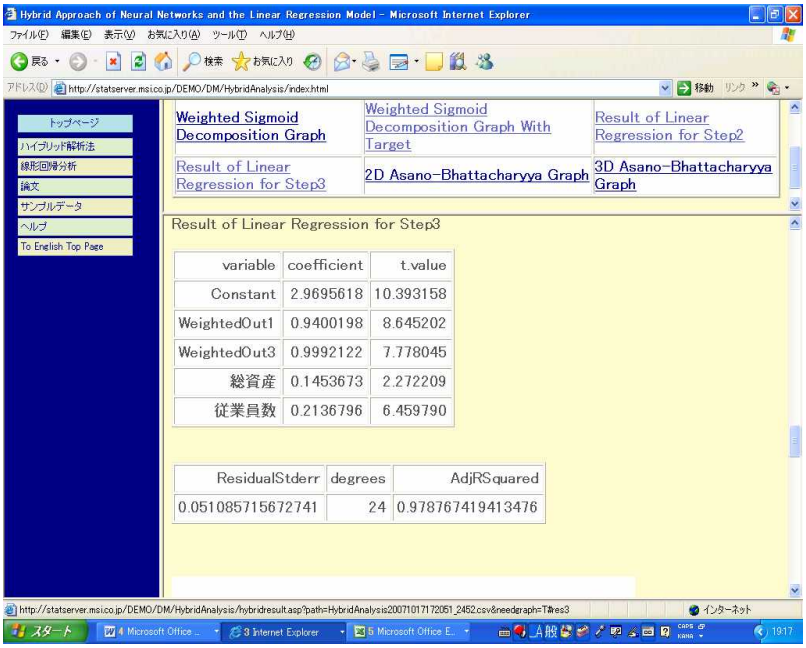


図6. Step3 の結果画面

Step3 では NO（通番）が削除された（$p' = 2$）．Step2 で2つの中間層出力値の和が図4の目的とウエイテッド・シグモイド分解図より決算期変更に伴う売上高変化に対応していることがわかる．入力変数である従業員数のt-値が Step2 の 3.22 から Step3 では 6.46 になった．

これらの結果より、入力変数の検討、中間層のユニット数の検討などを行うことによって、探索的にモデル選択の検討が行うことができる．

ビル建築モデルのニューラルネットワークモデルと線形回帰分析のハイブリッド解析事例として Asano and Yu (2007)を参照ください．

4. まとめ

「解釈可能性(Interpretability)」と予測精度の良さを併せもつデータ解析法であるニューラルネットワークモデルと線形回帰分析のハイブリッド解析法（浅野・椿, 2003）を、企業財務データ解析事例を用いて解説を行った.

Web 版が完成したことにより、インターネット利用による解析が可能になったので、Web 版での利用方法も示した. この解析法を用いて、モデル構造を検討することができるので、様々の分野のデータ解析での利用を希望している.

Web 版ニューラルネットワークモデルと線形回帰分析のハイブリッド解析法 URL を示す. 利用法についてのご質問とデータ等のコメントを、著者の浅野（email: asano@ic.daito.ac.jp）までご連絡ください.

http://statserver.msi.co.jp/DEMO/DM/HybridAnalysis/index.html

参考資料として,8月にポルトガルで開催された ISI2007での発表概要書: Miyoko Asano, Pijush K Bhattacharyya, Hiroe Tsubaki, Marco K. W. Yu 共著を次頁に添付致します.

参考文献

浅野美代子(2001),「ニューラルネットワークを用いた層別因子を含む回帰構造の解析」, 計算機統計学, 第 14 巻第 2 号, 123-138.

浅野美代子・椿広計（2002）「回帰分析の諸問題 ---計算機統計学の貢献とニューラルネットワークの特徴---」シンフォニカ研究叢書,『家計のミクロ統計分析』第2章, pp. 45 -56, 統計情報研究開発センター.

浅野美代子・椿広計(2003), 「ニューラルネットワークと線形回帰分析のハイブリッド解析法―東京都 23 区の給水量予測問題への適用」, 応用統計学, 第 31 巻 3 号, 227-238.

Asano, M., Tsubaki, H., Yoshizawa, T.(2002) Effectiveness of neural networks to regression with structural changes, *Applied Stochastic Models in Business and Industry*, Volume 18, Number 3, 189-195.

Asano, M. and Yu, M. K.W. (2007) "An Introduction to the Hybrid Approach of Neural Networks and the Linear Regression Model: An Illustration in the Hedonic Pricing Model of Building Costs", 大東文化大学紀要＜社会科学＞第四十五号, pp. 1-14.

―――――――――

当研究は以下の科研費の助成を受けたものです。

平成19年度科学研究費 （基盤研究（C））課題番号 19500241

研究代表者 浅野美代子、共同研究者 椿 広計

研究課題「ニューラルネットワークと線形回帰分析のハイブリッド解析法Webシステム」