ニューラルネットワークによる時系列予測における相関係数を用いた学習用類似データ選定方法

下平 悉作士†

ニューラルネットワークによる時系列データの将来値の予測においては，実際の産業への応用を考えると，構成が簡単で学習時間が比較的短い多層フィードフォーワード型のものを用いて，できるだけ精度のよい予測値が得られる方法の開発が望まれる．多層フィードフォーワード型ニューラルネットワークを用いた時系列データの予測手法には，移動窓データ学習法，類似データ選定学習法，全体データ学習法等がある．本論文では，類似データ選定学習法における類似データを選定する際の距離の計算について，相関係数のべき乗の形で重みづけを行う方法（CSDS 法：Correlation Coefficient Based Similar Data Selection Method）を提案するとともに，数値実験によりこれらの方法を用いた場合の予測精度の比較を行っている．数値実験の結果によると，類似データ選定学習法は，カオス的な変わりやすい性質の時系列データの場合によい予測精度が得られることが推測され，類似データを選定する際の距離の計算において，CSDS 法を用いることにより，かなり予測精度が向上することが分かった．これらの結果から，CSDS 法を用いた類似データ選定学習法は，移動窓データ学習法等の有力な代替的方法となり得るものと考えられる．

A Method for Selecting Similar Learning Data Based on Correlation Coefficients in the Prediction of Time Series Using Neural Networks

Hisashi Shimodaira†

In the field of prediction of future values of time series using multi-layer feedforward neural networks, methods have been proposed such as moving window data learning method, similar data selective learning method (SDSL method), and whole data learning method. In the SDSL method, a data group for learning which is similar to a data group for prediction is selected using a distance between the two groups. This paper proposes a new method (CSDS method : Correlation Coefficient Based Similar Data Selection Method) which uses correlation coefficients as weights in the equation for defining the distance. Numerical simulations were performed using the above three methods. According to the results, in the case of time series of which nature is rather chaotic and/or choppy, the prediction accuracy of the SDSL method was remarkably superior to those by the other methods and the CSDS method was considerably effective to improve the accuracy. Thus the SDSL method with the CSDS method is useful as an alternative one to the moving window data learning method.

1. まえがき

時系列データの過去の値に基づいて将来の値を予測する問題の解法には様々な方法があるが，ニューラルネットワークはその有力な解法の一つである．この問題に適用できるニューラルネットワークには，簡単な構成の多層フィードフォーワード型およびその変形型，さらに複雑な構成のリカレント型などがあり，用いるニューラルネットワークの型によって学習時間や予測精度などはかなりの差がある．たとえば，フィードフォーワード型とリカレント型のニューラルネットワークを比較した例[1]によると，フィードフォーワード型ではかなり精度のよい予測値が，また，リカレント型ではきわめてよい精度の予測値が得られているが，リカレント型はフィードフォーワード型に比べて，きわめて多大な学習時間を要している．実際に産業への応用を考えると，構成が簡単で学習時間が比較的短い多層フィードフォーワード型のニューラルネットワークを用いて，できるだけ精度のよい予測値が得られる方法の開発が望まれるのであり，本論文の目的はこの点にある．

多層フィードフォーワード型ニューラルネットワークを用いた，時系列データの予測のための学習データ

† 日本メックス株式会社 研究開発部
Research and Development Department, Nihon MECCS Co. Ltd.

選定法には，予測すべき時点の直近に学習データを選定する一定範囲の窓を設定し，予測すべき時点の移動に応じて，次々にその窓を移動していく方法がある[2]．ここでは，この方法を移動窓データ学習法（MWDL 法：Moving Window Data Learning Method）とよぶ．また，予測時点直近の予測の基礎となるデータと類似のデータを選定して学習する方法がある[3]．ここでは，この方法を類似データ選定学習法（SDSL 法：Similar Data Selective Learning Method）とよぶ．また，予測開始時点より前の過去のすべてのデータを用いて学習することにより，ニューラルネットワークの性能を評価することが行われている[4]．ここでは，この方法を全体データ学習法（WDL 法：Whole Data Learning Method）とよぶ．Peng[3] は提案した SDSL 法は MWDL 法よりも予測精度が優れていると述べているが，具体的な比較は示されていない．筆者の知る限りでは，これらの3方法による予測精度について，具体的に比較した研究はない．

SDSL 法における類似データの選定の際には，予測の基礎となるデータと選定すべき学習用データ間の距離によって類似性の大小を判断している．本論文では，距離の計算において，予測時点との相関係数の大きいデータほどその影響が大きく評価される類似データの選定方法（CSDS 法：Correlation Coefficient Based Similar Data Selection Method）を提案し，数値実験によりその予測精度を調べる．合わせて，提案する方法を用いた SDSL 法，MWDL 法，および WDL 法の予測精度の比較を行い，これらの方法の得失についての知見を報告する．

2章では，フィードフォーワード型ニューラルネットワークによる時系列データの予測方法の概要と提案する類似データの選定方法について述べ，さらに，MWDL 法と WDL 法の概要について述べる．3章では，性質の異なる3種類の時系列データを用いて行った数値実験の結果について述べる．4章では，既往の研究と対比して提案方法の特徴について述べ，5章では，まとめと今後の課題について述べる．

2. 予測方法と学習データの選定方法

2.1 予測方法の概要

ここでは，問題としている時系列データのみを用いてその将来の値を予測する1変量時系列モデルを扱う．各時点における時系列データを x_i $(i=1,\cdots,t,\cdots,\infty)$ とする．x_{t-1} までの値が観測されているものとし，1時点先の x_t を予測する場合を扱う．x_t の直近の d 個のデータ（予測用データグループとよぶ）に基づいて x_t を予測するものとすると，

$$(x_{t-d},\cdots,x_{t-2},x_{t-1})\rightarrow X_t \qquad (1)$$

ここで，X_t は x_t の予測値である．予測時には，入力層のノードに式(1)の左辺のデータが入力され，出力層のノードに X_t が出力される．学習時には，次式のように，教師データ $x_{t'}$ の直近の d 個のデータ（学習用データグループとよぶ）を用いて，$X_{t'}$ を出力する．

$$(x_{t'-d},\cdots,x_{t'-2},x_{t'-1})\rightarrow X_{t'} \qquad (2)$$

誤差逆伝播法[5] により，n 個の学習用データグループを用いて，出力値と観測値の差の2乗和の 1/2 が最小になるように，繰り返し計算によって重みの値を求める．予測値の計算は，このようにして求めた重みの値を用いて行う．

予測値の精度は，学習用データを選定する範囲，学習用データグループのデータ個数 d，学習用データグループ数 n に依存する．また，適切な隠れ層のノード数 h は，データの性質や d，n によって異なる．d，n，h の最適値は，対象とする時系列データの性質によって異なり，理論的には求められないので，既存の観測値を用いて，数値実験において試行錯誤により定める．

カオス理論では，d は埋め込み次元と呼ばれている．カオス時系列では，適切な埋め込み次元を用いることにより，1次元の時系列データから元のカオスの性質を再構成できるということが，理論的に証明されている[6]．

2.2 類似データ選定学習法（SDSL 法）

この方法では，予測用データグループに類似したデータグループを，学習データ選定範囲内から類似度が高い順に選定して学習する．ここでは，予測用と学習用のデータグループを d 次元ユークリッド空間における点と考えて，その間の距離を計算し，距離が小さいほど類似度が高いと考える．CSDS 法では，次式によって計算した距離が小さい順に n 個の学習データグループを選定する．

予測用と学習用のデータグループの間の重みつきマンハッタン距離は，次式で表される．

$$D=\sum_{i=1}^{d}|\rho_i|^m|x_{t-i}-x_{t'-i}| \qquad (3)$$

また，重みつきユークリッド距離は次式で表される．

$$D=\sqrt{\sum_{i=1}^{d}|\rho_i|^m(x_{t-i}-x_{t'-i})^2} \qquad (4)$$

ここで，ρ_i $(0 \leqq \rho_i \leqq 1)$ は，x_t と x_{t-i} の間の自己相関係数である．自己相関係数を m 乗しているのは，距離の計算においてその値の大小を考慮する度合を調整するためであり，その望ましい値は数値実験において試行錯誤により定める．

自己相関係数 ρ_i は，x_t と x_{t-i} が依存する度合を表しており，その値が大きいほど x_t は x_{t-i} の影響を強く受ける．式(3)と(4)は，その距離の計算において，自己相関係数が大きい座標軸ほどその座標値を大きく扱うことを表している．これは，データグループの類似度の計算において，先験的知識に基づいて，より重視するデータの座標値により大きな重みづけを行うテクニック[7]と同じ考え方である．これにより，予測すべき値により強い影響を持つデータを重視した距離が計算される．

学習データ選定範囲の始点を x_s とすれば，ρ_i は，厳密には，(x_s, x_{s+i}), (x_{s+1}, x_{s+i+1}), …, (x_{t-i-1}, x_{t-1}), (x_{t-i}, x_t) のデータの組によって計算すべきであるが，x_t は未知であるから，(x_{t-i}, x_t) を除いたデータの組によって計算する．

式(3)と(4)では，x_t と x_{t-i} および $x_{t'}$ と $x_{t'-i}$ についての自己相関係数が同一であることを前提としている．このような前提がほぼ成立する場合には，x_s を固定し，観測されたデータが増えるに従って，学習データ選定範囲を拡大していくことにより，過去のデータを有効に活用できる．この方法を，学習データ選定範囲拡大法とよぶ．上記の前提が成立しない場合には，自己相関係数が大幅に異ならない範囲で学習データ選定範囲を設定し，かつ，予測すべき時点に応じて，順次これを移動していく（x_s を前方にずらしていく）方法を考える．これを，学習データ選定範囲移動法とよぶ．

式(3)のマッハッタン距離では，座標値の差そのものが考慮されるのに対し，式(4)のユークリッド距離では，その2乗が考慮される．したがって，ユークリッド距離ではマンハッタン距離に比べて，座標値の差の大きなものがより大きな影響を与えることとなる．用いる距離の定義が予測精度に及ぼす影響を，数値実験により調べることとする．

以上，1変量モデルの場合について述べたが，上記の方法は，問題とする目的変数のほかに説明変数を入力とする多変量モデルの場合にも適用することができる．この場合には，距離の計算を意味のあるものにするために，それぞれの変数について，たとえば，次式を用いてデータを無次元化し，基準化する必要がある．

$$x_{is} = (x_{ir} - \bar{m}) / \sigma \qquad (5)$$

ここで，x_{ir} は観測値であり，x_{is} はその基準化した値である．$\bar{m}$ は当該変数の $[x_s, x_{t-1}]$ の範囲内のデータについての平均値，σ は標準偏差である．式(3)と(4)において，説明変数についての座標値の差の項を追加して計算する．このときの ρ_i は，目的変数と説明変数の間の相関係数を用いる．

2.3 移動窓データ学習法（MWDL 法）と全体データ学習法（WDL 法）

MWDL 法では，予測時点 t の直近に設けた移動窓内のデータを，式（2）のような関係でグループ化して，すべて学習する．本論文では，学習用データグループ数 n によって移動窓の大きさを定義する．

WDL 法では，予測開始時点より前のすべてのデータを，式(2)のような関係でグループ化してすべて学習した後，その重みの値を用いて以後のすべての時点の予測を行う．このとき，予測開始時点以降の予測用データは，予測値ではなく，観測値を用いる．

3. 数値実験

3.1 数値実験の方法

本章では，実際に観測された2種類の時系列データ（ニューヨーク市のはしかの患者数と水ぼうそうの患者数[8]）および人工的に生成したカオス・周期的時系列データ[4]を用いて，SDSL 法，MWDL 法および WDL 法により数値実験を行った結果について述べる．前2者は，時系列の新しい予測方法の検証のための標準的データとして，しばしば用いられるようになってきている[4),9)]．

用いたニューラルネットワークは，3層のフィードフォーワード型のものである．入力層のノード数は，学習（予測）用のデータグループのデータ個数 d に等しい．隠れ層と出力層の活性化関数として，ロジスティック関数 $(g(x) = 1/(1+e^{-x}))$ を用いた．

活性化関数として，ロジスティック関数を用いているため，入力データ（学習用データグループおよび予測用データグループのデータ）を $[0.2, 0.8]$ の範囲内の値になるようにスケーリングした．SDSL 法と MWDL 法では，予測時点ごとに扱うデータの数値の大きさの程度が変わってくるが，このようなスケーリングを行うことによって，扱う数値の大きさの程度をそろえることができる．

誤差逆伝播法における学習率は 0.1，慣性項の係数は 0.9 とした．収束条件は，出力ノードの誤差の 2 乗和の 1/2 が 0.001 になったときとした．

SDSL 法では，学習データ選定範囲の始点を全データの先頭とし，全観測データ数の 1/2 の次の時点から，順次，学習と予測計算を行った．距離の計算方法として式(3)と(4)を用いて，相関係数を考慮しない場合，および m を 1～4 の範囲で変化させて相関係数を考慮した場合について，学習（予測）用データグループのデータ数 d，学習用データグループ数 n，および隠れ層のノード数 h を少しずつ変化させて，予測精度が最もよいと思われる組合せを探した．

MWDL 法では，全観測データ数の 1/2 の次の時点から，順次，学習と予測計算を行った．SDSL 法と同様にして，予測値が最もよいと思われる d, n, h の組合せを探した．

WDL 法では，全データの前半の 1/2 を学習し，後半の 1/2 について予測計算を行った．SDSL 法と同様にして，予測値が最もよいと思われる d と h の組合せを探した．

予測精度の評価のための尺度として，観測値と予測値の間の相関係数 CRC，絶対誤差 ABE の平均値 ($mean$) と標準偏差 (σ)，相対誤差 RLE の平均値 ($mean$) と標準偏差 (σ) を計算した．観測値と予測値の間の相関係数は，個々の値が合っているかどうかよりも，全体としての変化の傾向がどの程度類似しているかを表すものである．絶対誤差は，次式で計算した．

$$ABE=|X_i-x_i|\times 100/x_{max} \qquad (6)$$

ここで，x_{max} は予測したデータの範囲内における観測値の最大値である．x_{max} に対する比を求めているのは，観測値の最大値に比べて，個々の誤差がどの程度であるかをみるためである．相対誤差は，次式で計算した．

$$RLE=|X_i-x_i|\times 100/x_i \qquad (7)$$

予測値が時系列データの下限に近い小さい値であるときは，上式の分母と分子が同程度の大きさになることがあるため，相対誤差の平均値と標準偏差はかなり大きな値になる場合があることに注意すべきである．

3.2 はしかの患者数の予測

ニューヨーク市において 1928 年 1 月から 1963 年 12 月までの間に観測された，月ごとのはしかの患者数のデータ[8]を用いた．データの総数は，432 個である．**図 1** に，データの時間的な変化の状況を示す．データの性質を調べるために，全データを用いて自己相関係数 ρ を計算した．**図 2** に，時差 (τ) と ρ の

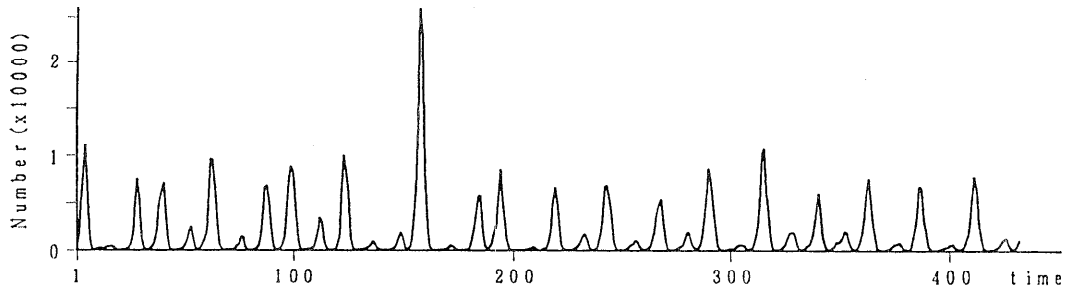


図 1 はしかの患者数についての時系列データ
Fig. 1 Time series for numbers of measles cases.

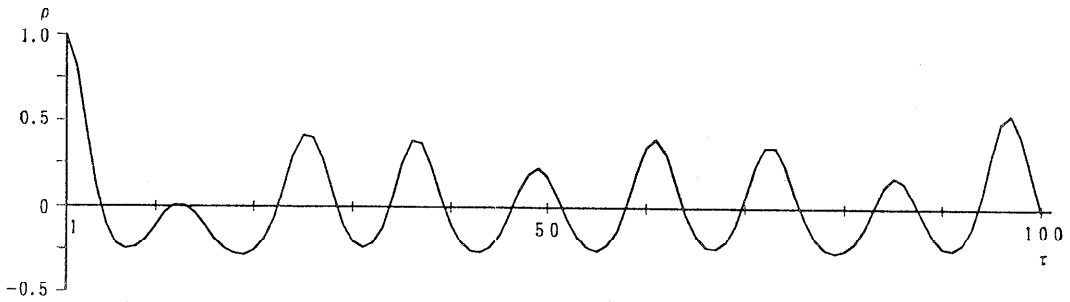


図 2 はしかの患者数についての時系列データのコレログラム
Fig. 2 Correlogram for time series on numbers of measles cases.

関係を表すコレログラムを示す．自己相関係数は，時系列データの本質的な滑らかさ（essential smoothness）を表している[10]．滑らかな（smooth）時系列データでは，τ が大きくなっても大きな自己相関係数を呈する．変わりやすい（choppy）な時系列データでは，τ が大きくなると自己相関係数はほとんどゼロに近い値になる．図2から分かるように，自己相関係数は周期的なピークを呈しているが，その値は大きくはなく，どちらかといえばこのデータは変わりやすい性質を持っているといえよう．このデータは，長期予測不可能であり，カオス的な性質を持っていることが指摘されている[9]．

SDSL 法については，予備実験によると，学習データ選定範囲拡大法に比べて，学習データ選定範囲移動法の方がやや予測精度がよいことが分かったので，後者を用いることとした．**表1**に，最もよい予測精度が得られた場合のパラメータの値を示す．表中でたとえば 8 4-1 は，入力層，隠れ層，出力層のノード数が，それぞれ8，4，1であることを示す．**表2**に，このようなパラメータ値を用いた場合の，SDSL 法による予測結果を示す．表中の記号で，$DST=1$ はマンハッタン距離を，$DST=2$ はユークリッド距離を表す．また，$COR=0$ は式(3)と(4)の距離の計算で相関係数を考慮しないことを，$COR=1$ は考慮することを表す．以下の表においても，同様である．

この結果から，距離の計算において相関係数による重みづけをしない場合に比べて，CSDS 法により適切な次数 m によりこれを考慮すれば，かなり予測精度が向上することが分かる．適切な m は，マッハッタン距離については2，ユークリッド距離については3である．距離の定義による差はあまり大きくはないが，マンハッタン距離で $m=2$ とした場合の予測精度が最もよい．この場合，相関係数を考慮しない場合に比べて，観測値と予測値の間の相関係数は 0.8% 大きく，絶対誤差の平均値と標準偏差はそれぞれ 8.5% と 11.9% 小さく，また，相対誤差の平均値と標準偏差はそれぞれ 39.2% と 76.5% 小さくなっている．

表3に，3種類の方法について，最適のパラメータ値を用いた場合の予測結果（SDSL 法では，$DST=1$, $COR=1$, $m=2$ の場合の値）を示す．SDSL 法の予測精度が最もよく，ついで MWDL 法，WDL 法の順になっている．CSDS 法を用いた SDSL 法は MWDL 法に比べて，相関係数は 3.0% 大きく，絶対誤差の平均値と標準偏差はそれぞれ 28.4% と 18.2% 小さく，相対誤差の平均値と標準偏差はそれぞれ 54.0% と 82.3% 小さくなっている．

3.3 カオス・周期的時系列の予測

次式[4]により生成した 500 個のデータを用いた．

$$x_{t+1}=3.8\,x_t(1-x_t) \quad \text{for even } t \qquad (8)$$

$$x_{t+1}=(1-x_t) \quad \text{for odd } t \qquad (9)$$

式(8)は，カオスの一つであるロジスティック写像である．式(9)は，1時点とびに同一の値になる周期関数である．初期値は，偶数 t については 0.1，奇数 t については 0.3 とした．**図3**に，データの時間的な変

表1 はしかの患者数の予測時の最適パラメータ値
Table 1 Optimum parameter values for time series of measles cases.

方法	ネットワークの構成	n
SDSL	8-4-1	16
MWDL	30-3-1	50
WDL	8-2-1	(209)

表2 SDSL 法によるはしかの患者数の予測結果
Table 2 Prediction results for numbers of measles cases with SDSL method.

DST	COR	m	CRC	ABE mean	ABE σ	RLE mean	RLE σ
1	0	—	0.953	3.31	5.36	73.5	162.
	1	1	0.955	3.19	5.20	54.1	70.3
		2	0.961	3.03	4.72	44.7	38.0
		3	0.959	3.12	4.73	51.8	69.7
		4	0.952	3.62	4.98	79.6	210.
2	0	—	0.955	3.29	5.11	76.1	156.
	1	1	0.952	3.34	5.34	60.6	83.6
		2	0.956	3.37	4.97	65.8	121.
		3	0.960	3.10	4.78	49.2	58.7
		4	0.957	3.25	4.91	58.8	105.

表3 学習データ選定法とはしかの患者数の予測結果
Table 3 Learning data selection methods vs. prediction results for numbers of measles cases.

方法	CRC	ABE mean	ABE σ	RLE mean	RLE σ
SDSL	0.961	3.03	4.72	44.7	38.0
MWDL	0.933	4.23	5.77	97.2	215.
WDL	0.925	5.08	5.60	123.	144.

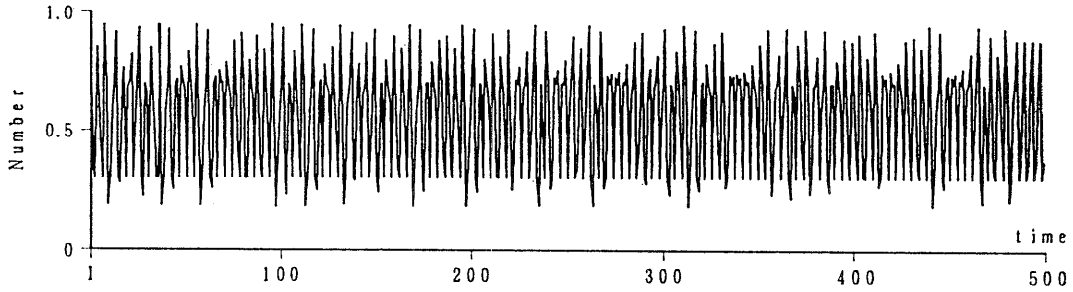


図 3　カオス・周期的時系列データ
Fig. 3　Chaotic and periodic time series.

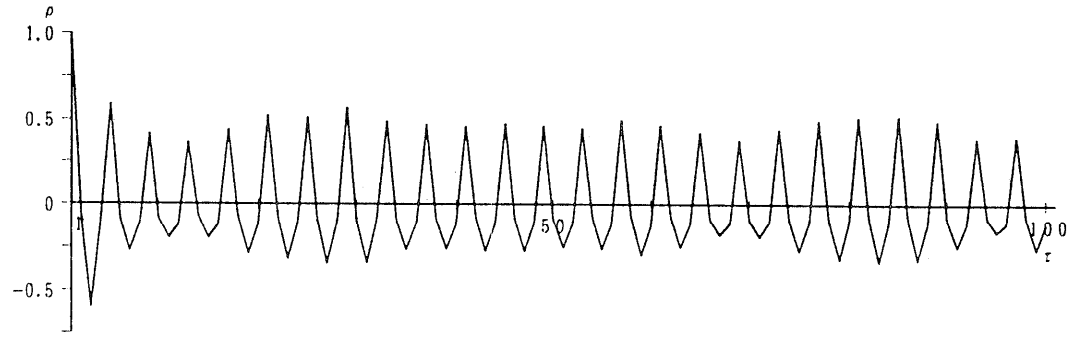


図 4　カオス・周期的時系列データのコレログラム
Fig. 4　Correlogram for chaotic and periodic time series.

表 4　カオス・周期的時系列データの予測時の最適パラメータ値
Table 4　Optimum parameter values for chaotic and periodic time series.

方法	ネットワークの構成	n
SDSL	5-2-1	2
MWDL	40-9-1	120
WDL	20-9-1	(231)

化の状況を示す．**図 4** に，コレログラムを示す．自己相関係数は周期的にピークを呈しているが，その値はそれほど大きくはない．

計算する範囲によって自己相関係数は変化しないので，学習データ選定範囲拡大法を用いた．**表 4** に，最もよい予測精度が得られた場合のパラメータの値を示す．**表 5** に，このようなパラメータ値を用いた場合の SDSL 法による予測結果を示す．この場合にも，距離の計算において相関係数による重みづけをしない場合に比べて，CSDS 法により適切な次数によりこれを考慮すれば，かなり予測精度が向上することが分かる．適切な次数は，はしかの患者数の場合と同様である．

表 5　SDSL 法によるカオス・周期的時系列データの予測結果
Table 5　Prediction results for chaotic and periodic time series with SDSL method.

DST	*COR*	*m*	*CRC*	*ABE mean*	*ABE σ*	*RLE mean*	*RLE σ*
1	0	—	0.995	1.38	1.96	2.62	4.03
	1	1	0.995	1.51	1.89	2.84	3.96
		2	0.997	1.36	1.21	2.51	2.43
		3	0.995	1.72	1.68	3.04	2.91
		4	0.995	1.84	1.58	3.28	2.89
2	0	—	0.996	1.40	1.78	2.73	3.97
	1	1	0.995	1.47	1.94	2.82	4.12
		2	0.996	1.43	1.59	2.80	3.56
		3	0.997	1.30	1.17	2.50	2.75
		4	0.997	1.51	1.35	2.80	2.80

この場合には，ユークリッド距離を用いて $m=3$ とした場合の予測精度が最もよい．相関係数を考慮しない場合に比べて，観測値と予測値の間の相関係数はわず

かに大きく，絶対誤差の平均値と標準偏差はそれぞれ 7.1% と 34.3% 小さく，相対誤差の平均値と標準偏差はそれぞれ 8.4% と 30.7% 小さくなっている．

表 6 に，3種類の方法について，最適のパラメータ値を用いた場合の予測結果（SDSL 法では，$DST=2$, $COR=1$, $m=3$ の場合の値）を示す．SDSL 法の予測精度が最もよく，ついで WDL 法，MWDL 法の順になっている．CSDS 法を用いた SDSL 法は MWDL 法に比べて，相関係数は 3.2% 大きく，絶対誤差の平均値と標準偏差はそれぞれ 67.6% と 76.0% 小さく，相対誤差の平均値と標準偏差はそれぞれ 68.0% と 78.5% 小さくなっている．

表 6 学習データ選定法とカオス・周期的時系列データの予測結果
Table 6 Learning data selection methods vs. prediction results for chaotic and periodic time series.

方法	*CRC*	*ABE mean*	*ABE σ*	*RLE mean*	*RLE σ*
SDSL	0.997	1.30	1.17	2.50	2.75
MWDL	0.966	4.01	4.87	7.82	12.8
WDL	0.975	3.62	4.07	7.16	10.1

3.4 水ぼうそうの患者数の予測

ニューヨーク市において 1928年1月から 1972年6月までの間に観測された，月ごとの水ぼうそうの患者数のデータ[8]を用いた．データの総数は，534 個である．**図 5** に，データの時間的な変化の状況を示す．**図 6** に，コレログラムを示す．自己相関係数は，周期的に大きなピークを呈しており，τ が大きくなってもピークの値はほぼ同じである．したがって，このデータは，強い周期性を持つ滑らかな時系列データであることが分かる．

SDSL 法については，予備実験によると，学習データ選定範囲拡大法に比べて，学習データ選定範囲移動法の方がやや予測精度がよいことが分かったので，後者を用いることとした．**表 7** に，最もよい予測精度が得られた場合のパラメータの値を示す．**表 8** に，このようなパラメータ値を用いた場合の，SDSL 法による予測結果を示す．この場合には，CSDS 法により距離の計算において相関係数を考慮しても予測精度はほとんど向上せず，むしろ若干悪くなっている．最も予測

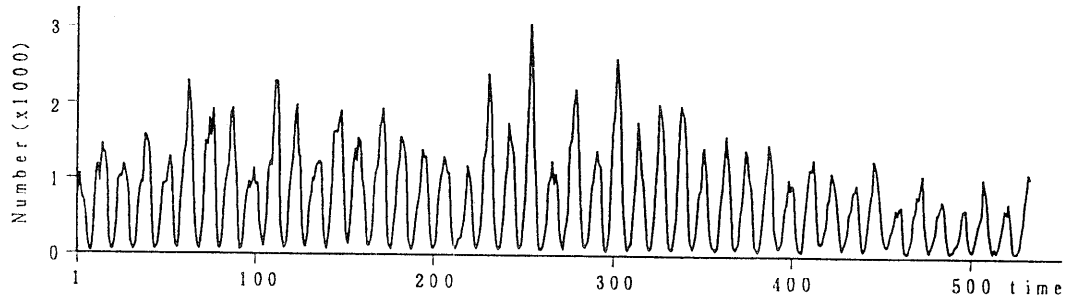


図 5 水ぼうそうの患者数についての時系列データ
Fig. 5 Time series for numbers of chickenpox cases.

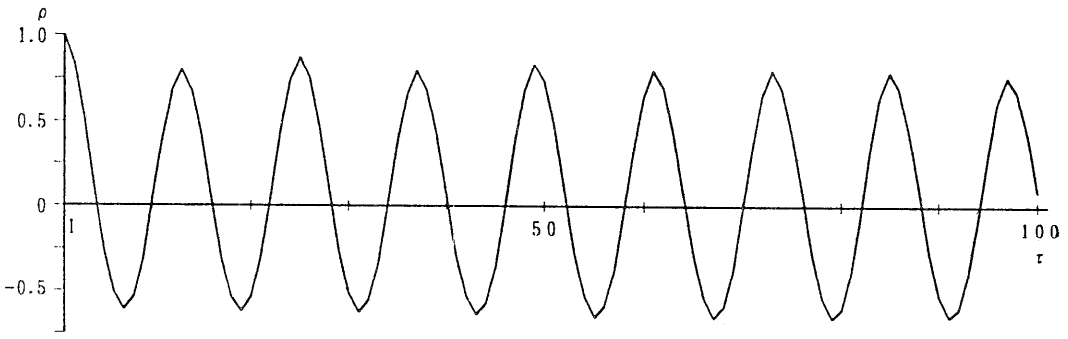


図 6 水ぼうそうの患者数についての時系列データのコレログラム
Fig. 6 Correlogram for time series on numbers of chickenpox cases.

表 7 水ぼうそうの患者数の予測時の最適パラメータ値
Table 7 Optimum parameter values for time series of chickenpox cases.

方法	ネットワークの構成	n
SDSL	32-3-1	35
MWDL	30-3-1	150
WDL	31-4-1	(237)

表 8 SDSL 法による水ぼうそうの患者数の予測結果
Table 8 Prediction results for numbers of chickenpox cases with SDSL method.

DST	*COR*	m	*CRC*	*ABE mean*	*ABE σ*	*RLE mean*	*RLE σ*
1	0	—	0.956	4.62	3.71	43.4	51.2
	1	1	0.959	4.59	3.48	45.1	58.7
		2	0.958	4.72	3.60	44.9	50.8
		3	0.958	4.69	3.58	44.7	49.6
		4	0.958	4.80	3.58	45.1	50.3
2	0	—	0.959	4.49	3.71	41.8	52.0
	1	1	0.959	4.63	3.55	46.4	59.0
		2	0.957	4.75	3.65	48.5	60.6
		3	0.956	4.71	3.73	47.5	62.9
		4	0.953	4.94	3.74	51.8	72.1

表 9 学習データ選定法と水ぼうそうの患者数の予測結果
Table 9 Learning data selection methods vs. prediction results for numbers of chickenpox cases.

方法	*CRC*	*ABE mean*	*ABE σ*	*RLE meam*	*RLE σ*
SDSL	0.959	4.49	3.71	41.8	52.0
MWDL	0.966	4.03	3.51	31.8	37.0
WDL	0.964	4.19	3.72	27.4	29.1

精度がよいのは，ユークリッド距離を用いて相関係数を考慮しない場合である．

表 9 に，3種類の方法について，最適のパラメータ値を用いた場合の予測結果（SDSL 法では，$DST=2$, $COR=0$ の場合の値）を示す．MWDL 法と WDL 法の予測精度はほぼ同等であり，SDSL 法はこれよりも若干劣っている．

3.5 結果の考察

以上の数値実験の結果によると，3種類の方法による予測精度は，対象とする時系列データの性質に依存していることが分かる．カオス的な変わりやすい性質を持ったはしかの患者数のデータの場合には，CSDS 法を用いた SDSL 法が他の方法よりもかなりよい結果を示している．カオス的性質と周期的性質を合わせ持った人工的に生成したデータの場合には，CSDS 法を用いた SDSL 法が他の方法よりも極めてよい結果を示している．ことに，SDSL 法は他の方法に比べて，ニューラルネットワークの規模が極めて小さく，学習データグループ数が極めて少なくてすることが注目される．また，このようなデータの発生機構が複数ある場合にも，よい結果が得られていることも注目に値する．強い周期性と滑らかな性質を持った水ぼうそうの患者数のデータの場合には，SDSL 法は他の方法よりも若干よくない結果を示している．

カオス的な変わりやすい性質の時系列データの場合に，SDSL 法の予測精度が MWDL 法よりよいのは，データの性質や構造をとらえるためには，予測時点近傍のデータのみでは十分ではなく，過去の広い範囲における変化の状況が類似したデータを必要とすることを意味しているものと考えられる．

SDSL 法を適用するのに適した時系列データの場合には，類似データ選定のための距離の計算において CSDS 法により相関係数を適切な次数で考慮することは，予測精度を向上させるのにかなり有効であることが分かる．用いる距離の定義による予測精度の差異はあまり大きくはなく，また，対象とする時系列データの性質に依存するようである．相関係数を考慮する次数は，マンハッタン距離の場合には2，ユークリッド距離の場合には3が適切のようである．

4. 関連する研究と提案方法の特徴

Farmer ら[6]は，カオス時系列の予測において，埋め込み次元の考え方に基づいてデータをグループ化し，予測用データグループとの距離が近いデータブループを選定し，最小二乗法により1次多項式を当てはめる方法を提案している．距離の計算方法についての詳細には言及していない．

Peng ら[8]は，ニューラルネットワークを用いた電力の需要予測において，Farmer らと同様な方法を用いている．距離の計算において，入力変数による出力変数の偏微分値を重みとして考慮している．重みづけ

を考慮した場合の予測精度を，重みづけを考慮しない場合と比べると，通常のフィードフォーワード型のニューラルネットワークではむしろ悪くなっているが，線形計算を行うネットワークを付加した型のニューラルネットワークでは，相対誤差の平均値と標準偏差がそれぞれ 5.4% と 17.1% 小さくなっている．しかしながら，入力変数による出力変数の微分値は，変数の変動量を関係づける値であり，一般的には，ある時点の入力値そのものが予測時点の出力値に及ぼす影響を直接的に表すものではない．

本論文で提案した CSDS 法の新規性は，類似データを選定する距離の計算において，2時点間のデータの依存関係を表す量として，理論的に明確な相関係数を用いて重みづけを行う点にある．さらに，相関係数のべき乗の形で考慮することにより，係数値の大小による重みづけの度合いを調整できるように工夫している．また，性質の異なったデータを用いて，CSDS 法を用いた SDSL 法を含めて，MWDL 法，WDL 法により数値実験を行い，これらの方法の得失についての新しい知見を提供している．

5. むすび

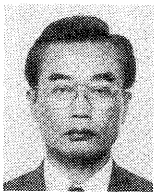
数値実験の結果によると，3種類の学習データ選定方法による予測精度は，対象とする時系列データの性質に依存することが分かった．ここで示した例のみでは断定できないが，SDSL 法は，カオス的な変わりやすい性質の時系列データに対して有効であることが推測される．SDSL 法を適用するのに適した時系列データの場合には，CSDS 法により，類似データ選定における距離の計算において，相関係数のべき乗の形で重みづけを行うことにより，重みづけを行わない場合に比べて，かなり予測精度が向上することが分かった．これらの結果により，提案した CSDS 法を用いた SDSL 法は，MWDL 法，WDL 法の有力な代替的方法となり得るものと考えられる．

今後は，多変量の場合も含めて，様々な観測データについて提案方法を適用し，その有効性を広く検証したいと考えている．

参考文献

1) Lambert, J. et al.: Application of Feedforward and Recurrent Neural Networks to Chemical Plant Predictive Modeling, *Proc. IJCNN*, Vol. 1, pp. 373-378 (1991).
2) Onoda, T.: Next Day Peak Load Forecasting Using an Artificial Neural Network, *Proc. IJCNN*, Vol. 2, pp. 2029-2032 (1993).
3) Peng, T. M. et al.: Advancement in the Application of Neural Networks for Short-term Load Forecasting, *IEEE Trans. on Power Systems*, Vol. 7, No. 1, pp. 250-257 (1992).
4) Wolpert, D. M. et al.: Detecting Chaos with Neural Networks, *Proc. R. Soc. Lond. B*, Vol. 242, pp. 82-86 (1990).
5) Rumelhart, D. E. et al.: *Parallel Distributed Processing*, Vol. 1, Chapter 8, MIT Press (1986).
6) Farmer, J. D. et al.: Predicting Chaotic Time Series, *Physical Review Letters*, Vol. 59, No. 8, pp. 845-848 (1987).
7) Kaufman, L. et al.: *Finding Groups in Data*, p. 13, John Wiley & Sons, Inc. (1990).
8) Yorke, J. A. et al.: Recurrent Outbreaks of Measles, Chickenpox and Mumps, *American J. of Epidemiology*, Vol. 98, No. 6, pp. 469-482 (1973).
9) Sugihara, G. et al.: Nonlinear Forecasting as a Way of Distinguishing Chaos from Measurement Error in Time Series, *Nature*, Vol. 344-19, pp. 734-741 (1990).
10) Shumway, R. H.: *Applied Statistical Time Series Analysis*, p. 17, Prentice Hall (1988).

(平成6年9月26日受付)
(平成6年11月17日採録)



下平丞作士（正会員）

昭和44年東京都立大学工学部建築工学科卒業．昭和46年同大学院修士課程修了．同年日本電信電話公社(現株式会社)入社．武蔵野電気通信研究所および建築部建築技術開発室において，構造工学，数値解析，CAD，データベース，AI 等の研究とシステム開発に従事．平成4年から日本メックス(株)において，建物・設備管理へのコンピュータの応用に関する研究・開発に従事．データ工学，データベース，知識工学，エキスパートシステム，ニューラルネットワーク，GA，図形処理，CG 等に興味を持っている．工学博士．電子情報通信学会，人工知能学会，日本建築学会各会員．