適応的サンプリングを導入した株価パフォーマンス予測

○瀧川 孝幸† 鈴木 麗璽† 有田 隆也†
†名古屋大学 大学院情報科学研究科

Stock Performance Prediction Using an Adaptive Sampling Method

○ Takayuki Takigawa† Reiji Suzuki† Takaya Arita†
†Graduate School of Information Science, Nagoya University

Abstract: This paper proposes an adaptive sampling method of the input data which provides a good accuracy of the stock performance prediction based on neural networks by evolving the possible combinations of the input data which are used to learn and predict by the networks. We applied this method to the problem in which the network predicts the performance of a certain stock by using the set of previous performances of other stocks and itself. We show that the effective sets of inputs were obtained successfully and their combinations depended on the stock which was predicted. Also, their error rates were 3% better than the cases in which the input data were limited to the previous performances of the stock which was predicted.

1 はじめに

複雑な挙動を示すことが知られる株式価格に関して，その将来の変動予測を行うモデルの開発が，これまで経済学や工学の領域で取り組まれてきた．価格予測の具体的な方策として，かつては価格の時系列それぞれを確率変数の実数値とみなし，その線形和を用いて未知の将来価格に回帰させる線形予測モデル (自己回帰モデル, 自己回帰移動平均モデルなど) が中心的に研究されていたが，その枠組みの中だけでは十分に説明できないより複雑な挙動が価格の時系列に含まれることから，ニューラルネットワークなどの非線形なモデルに基づく予測が現在広く研究されている [1]．

ニューラルネットワークに基づく株価予測では，予測銘柄の時系列と強く関連付けられていると考えられる種々の時系列データを採取して，採取されたデータ系列と予測するべき銘柄の次期の株価との間にある非線形な関係をネットワークに学習させ，そのネットワークに現在のデータを入力することで，次期の株価の予測を行う．その際，ネットワークの予測のために用いられる入出力データとして，S&P500 株価指数の始値，安値，終値を入力として取引の最適なポジショニングを予測するモデル [2, 3] や，流動性比率，株主資本収益率，株価収益率及び株価売上高倍率の 4 つの経営指標を入力として，株価の収益率予測を行うモデル [4] など，与えるデータの種類は研究者ごとにアプリオリに決定されており，そもそもどのようなデータを与えれば予測性能をより改善させられるのかについて統一的な知見は得られていないと言える．

一般に，価格予測に利用可能なデータは，予測する価格の挙動に対して全て均質な意味合いや相関を持っているのではなく，より本質的な意味合いを持つデータとそうではないデータとが混在していると考えられ，価格に対してより本質的に影響を及ぼすデータのみを選択して予測に用いることで，予測性能を改善できる可能性がある．

それにも関わらず，データの与え方自体を評価するための枠組みがこれまで十分に整備されてこなかった理由として，どのようなデータが価格の挙動により本質的な意味を持っているのかの判断が容易でないこと，また，仮にそのような基準があったとしても，考慮するべきデータの組み合わせ数が膨大であるため，現実的な計算時間で最適な組み合わせを求めることが難しいことが考えられる．

そこで，本研究は，ある銘柄の株式価格を予測する際，どのような種類のデータの組み合わせを元に予測するのが望ましいのかを適応的に探索していく，適応的データサンプリング手法を提案し，評価することを目的とする．具体的には，ニューラルネットワークを用いた過去の時系列データに基づく将来の株価予測において，利用する時系列データの種類の組合せを考え，遺伝的アルゴリズムを用いてその最適化を行う．本稿では，本手法の有効性を検証するための具体例として，日次収益率の予測問題を取り上げる．従来，ある銘柄の株価予測において，膨大な量のデータから予測銘柄の次期のデータと関係性の強いデータのみを抽出し，それを予測に役立てるということが十分に検討されてこなかったが，提案手法によって他の銘柄の株式価格の望ましい組合せを探索することで，予測精度の向上が可能であることを示す．

2 複数銘柄の時系列データに基づく日次収益率予測

適応的データサンプリングの有効性を検証するために，過去の時系列データに基づく日次収益率の予測をとりあげる．具体的には，Table 1 に示す株式銘柄群の過去数期分の時系列から，そのうち一つの次期の日次収益率 $\left(\frac{\text{当日の終値}-\text{前日の終値}}{\text{前日の終値}}\right)$ を予測することを考える．このとき，予測銘柄そのものの前期収益率データに加え，Table 1 に含まれる他の銘柄の前期収益率のデータを，予測のために利用するデータの候補とする．その中から，ニューラルネットワークを用いた学習と予測の際に有益となるデータの組合せを，次節で解説する遺伝的アルゴリズムに基づく適応的データサンプリングを用いて探索する．なお，今回は，

ネットワークの学習のために 2002 年 1 月 1 日から 2002 年 12 月 31 日までの各銘柄の日次収益率を訓練データとして用い，予測精度の評価には 2003 年 1 月 1 日から 2003 年 12 月 31 日までの日次収益率をテストデータとして用いるものとする．

3 提案手法

3.1 遺伝的アルゴリズムに基づく適応的データサンプリング

本研究では，予測精度の向上をもたらす入力データの組み合わせを得るために，次のような遺伝的アルゴリズムに基づく探索手法を提案する．

はじめに，個体集団を形成し，各個体が持つ染色体により解候補を表現する．染色体は，入力の候補である各銘柄を入力データとして採用するかどうかを示す複数の遺伝子からなる．各遺伝子座は Table 1 に示すように特定の銘柄と 1 対 1 に対応しており，各値が 1 であれば対応する銘柄の収益率データをニューラルネットワークへの入力データとして採用し，0 であれば破棄する．なお，今回は，予測する銘柄そのものの収益率データは常に利用するものとする．そのため，予測銘柄に対応する遺伝子は常に 1 であるものとする．

このような表現を個体ごとに持つように設定し，以下の処理手順を実行する．

1. 個体集団の初期化
   各個体の染色体をランダムに初期化する．
2. 個体の学習と予測
   各個体が遺伝情報として持っている銘柄の組み合わせに従って生成された訓練データを用いて 3 層パーセプトロンを学習させ，テストデータを用いて予測を行った場合の平均予測二乗誤差の逆数を各個体の適応度として割り振る (次節にて詳細に解説)．
3. 個体集団の進化
   各個体の適応度に応じて，個体集団を上位 5 個体から成るエリート集団とその他の個体から成る非エリート集団に分割する．エリート集団に占められる各個体はそのまま次世代においても保存され，非エリート集団中の個体の子孫は選択，交叉，突然変異の進化操作によって確率的に決定される．その際，個体の選択にはルーレット選択法を，交叉には一点交叉法を，突然変異には各遺伝子について一定確率で行われるビット反転をそれぞれ用いた．
4. 終了条件の判定
   現世代数が指定世代数に達していたら処理を終了し，達していなければ 2 に戻る．

3.2 3 層パーセプトロンによる学習と予測

本稿では，入力素子数が可変であり，中間素子数，出力素子数がそれぞれ 10, 1 である 3 層構造のフィードフォーワード型ニューラルネットワークを採用する (Fig. 1)．入力素子数は予測のためにいくつの銘柄の収益率データを組み入れるかという銘柄数と，1 銘柄当たり過去何期分の収益率データを用いるかの積によって決定され，ネットワークの各入力素子に対して，どの銘柄の何期前データを入力するかを 1 対 1 で対応付ける．また活性化関数として，ゲイン値が 1 に固定されたシグモイド関数を用いる．

各入力素子に対応付けられたデータを入力した際の出力素子の出力を，予測銘柄の次期収益率に関する予測値とみ

Table 1　各遺伝子座が指定する入力候補銘柄

遺伝子座	社名	証券コード	遺伝子座	社名	証券コード
0	ミネビア	6479	11	三洋電機	6764
1	日立製作所	6501	12	アルプス電気	6770
2	東芝	6502	13	パイオニア	6773
3	三菱電機	6503	14	クラリオン	6796
4	富士電機 H	6504	15	横河電機	6841
5	日本電気	6701	16	デンソー	6902
6	富士通	6702	17	カシオ計算機	6952
7	沖電気	6703	18	京セラ	6971
8	松下電器産業	6752	19	松下電工	6991
9	シャープ	6753	20	キャノン	7751
10	TDK	6762			

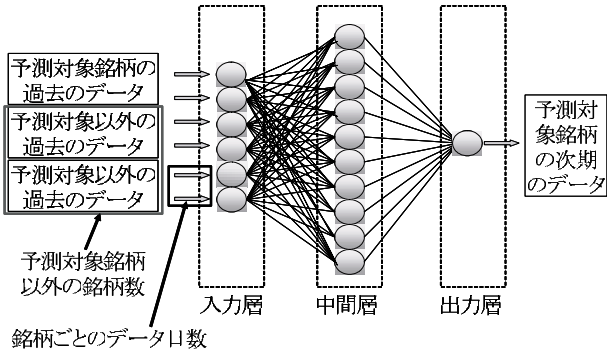


Fig. 1　3 層パーセプトロンの模式図

Table 2　実験におけるパラメタ設定

GA のパラメタ設定		3 層パーセプトロンのパラメタ設定	
個体数	50	中間素子数	10
世代数	100	最大学習反復回数	1000
エリート保存率	0.1	許容絶対誤差	$1.0 \cdot 10^{-8}$
交叉率	0.3	許容相対誤差	$1.0 \cdot 10^{-8}$
突然変異率	0.2		

Table 3　入力となるデータの日数と予測誤差

社名	1	2	3	4	5	6	7	8	9
ミネビア (6423)	$8.70{\cdot}10^{-4}$	$8.77{\cdot}10^{-4}$	$8.67{\cdot}10^{-4}$	$8.65{\cdot}10^{-4}$	$8.62{\cdot}10^{-4}$	$8.83{\cdot}10^{-4}$	$8.86{\cdot}10^{-4}$	$18.08{\cdot}10^{-4}$	$18.72{\cdot}10^{-4}$
日立製作所 (6501)	$7.59{\cdot}10^{-4}$	$7.57{\cdot}10^{-4}$	$7.57{\cdot}10^{-4}$	$7.60{\cdot}10^{-4}$	$7.59{\cdot}10^{-4}$	$7.77{\cdot}10^{-4}$	$7.78{\cdot}10^{-4}$	$14.23{\cdot}10^{-4}$	$17.84{\cdot}10^{-4}$
東芝 (6502)	$7.68{\cdot}10^{-4}$	$7.61{\cdot}10^{-4}$	$7.60{\cdot}10^{-4}$	$7.65{\cdot}10^{-4}$	$7.63{\cdot}10^{-4}$	$7.71{\cdot}10^{-4}$	$7.80{\cdot}10^{-4}$	$21.13{\cdot}10^{-4}$	$24.72{\cdot}10^{-4}$
三菱電機 (6503)	$7.06{\cdot}10^{-4}$	$7.23{\cdot}10^{-4}$	$7.22{\cdot}10^{-4}$	$7.27{\cdot}10^{-4}$	$7.27{\cdot}10^{-4}$	$8.40{\cdot}10^{-4}$	$10.13{\cdot}10^{-4}$	$11.14{\cdot}10^{-4}$	$11.32{\cdot}10^{-4}$
富士電機 H(6504)	$6.10{\cdot}10^{-4}$	$6.15{\cdot}10^{-4}$	$6.19{\cdot}10^{-4}$	$6.24{\cdot}10^{-4}$	$6.25{\cdot}10^{-4}$	$8.78{\cdot}10^{-4}$	$9.20{\cdot}10^{-4}$	$8.67{\cdot}10^{-4}$	$9.14{\cdot}10^{-4}$
日本電気 (6701)	$8.33{\cdot}10^{-4}$	$8.31{\cdot}10^{-4}$	$8.42{\cdot}10^{-4}$	$8.48{\cdot}10^{-4}$	$11.81{\cdot}10^{-4}$	$12.13{\cdot}10^{-4}$	$25.56{\cdot}10^{-4}$	$28.40{\cdot}10^{-4}$	$11.74{\cdot}10^{-4}$
富士通 (6702)	$9.84{\cdot}10^{-4}$	$10.11{\cdot}10^{-4}$	$10.06{\cdot}10^{-4}$	$10.10{\cdot}10^{-4}$	$12.63{\cdot}10^{-4}$	$14.15{\cdot}10^{-4}$	$22.14{\cdot}10^{-4}$	$25.55{\cdot}10^{-4}$	$19.51{\cdot}10^{-4}$
沖電気工業 (6703)	$10.02{\cdot}10^{-4}$	$9.95{\cdot}10^{-4}$	$10.04{\cdot}10^{-4}$	$11.12{\cdot}10^{-4}$	$11.52{\cdot}10^{-4}$	$11.91{\cdot}10^{-4}$	$11.93{\cdot}10^{-4}$	$15.15{\cdot}10^{-4}$	$23.87{\cdot}10^{-4}$
松下電器産業 (6752)	$4.26{\cdot}10^{-4}$	$4.78{\cdot}10^{-4}$	$4.78{\cdot}10^{-4}$	$4.82{\cdot}10^{-4}$	$4.78{\cdot}10^{-4}$	$4.79{\cdot}10^{-4}$	$4.82{\cdot}10^{-4}$	$4.79{\cdot}10^{-4}$	$4.82{\cdot}10^{-4}$
シャープ (6753)	$3.66{\cdot}10^{-4}$	$3.70{\cdot}10^{-4}$	$3.71{\cdot}10^{-4}$	$3.72{\cdot}10^{-4}$	$3.74{\cdot}10^{-4}$	$3.74{\cdot}10^{-4}$	$3.80{\cdot}10^{-4}$	$3.80{\cdot}10^{-4}$	$3.77{\cdot}10^{-4}$
TDK(6762)	$7.50{\cdot}10^{-4}$	$7.52{\cdot}10^{-4}$	$7.55{\cdot}10^{-4}$	$7.57{\cdot}10^{-4}$	$7.66{\cdot}10^{-4}$	$7.77{\cdot}10^{-4}$	$11.53{\cdot}10^{-4}$	$13.44{\cdot}10^{-4}$	$15.88{\cdot}10^{-4}$
三洋電機 (6764)	$5.96{\cdot}10^{-4}$	$5.99{\cdot}10^{-4}$	$6.03{\cdot}10^{-4}$	$6.09{\cdot}10^{-4}$	$6.13{\cdot}10^{-4}$	$6.11{\cdot}10^{-4}$	$10.57{\cdot}10^{-4}$	$11.09{\cdot}10^{-4}$	$10.13{\cdot}10^{-4}$
アルプス電気 (6770)	$8.37{\cdot}10^{-4}$	$8.47{\cdot}10^{-4}$	$8.51{\cdot}10^{-4}$	$8.54{\cdot}10^{-4}$	$8.61{\cdot}10^{-4}$	$8.65{\cdot}10^{-4}$	$15.35{\cdot}10^{-4}$	$14.93{\cdot}10^{-4}$	$19.23{\cdot}10^{-4}$
パイオニア (6773)	$4.73{\cdot}10^{-4}$	$4.88{\cdot}10^{-4}$	$4.97{\cdot}10^{-4}$	$5.18{\cdot}10^{-4}$	$5.81{\cdot}10^{-4}$	$5.55{\cdot}10^{-4}$	$6.39{\cdot}10^{-4}$	$6.98{\cdot}10^{-4}$	$7.21{\cdot}10^{-4}$
クラリオン (6796)	$22.16{\cdot}10^{-4}$	$48.32{\cdot}10^{-4}$	$45.51{\cdot}10^{-4}$	$106.94{\cdot}10^{-4}$	$143.60{\cdot}10^{-4}$	$318.18{\cdot}10^{-4}$	$278.76{\cdot}10^{-4}$	$203.16{\cdot}10^{-4}$	$166.11{\cdot}10^{-4}$
横河電機 (6841)	$9.86{\cdot}10^{-4}$	$6.86{\cdot}10^{-4}$	$6.90{\cdot}10^{-4}$	$7.03{\cdot}10^{-4}$	$8.20{\cdot}10^{-4}$	$8.57{\cdot}10^{-4}$	$11.49{\cdot}10^{-4}$	$12.58{\cdot}10^{-4}$	$12.13{\cdot}10^{-4}$
デンソー (6902)	$3.04{\cdot}10^{-4}$	$3.03{\cdot}10^{-4}$	$3.03{\cdot}10^{-4}$	$3.02{\cdot}10^{-4}$	$3.06{\cdot}10^{-4}$	$3.08{\cdot}10^{-4}$	$3.10{\cdot}10^{-4}$	$3.12{\cdot}10^{-4}$	$3.16{\cdot}10^{-4}$
カシオ計算機 (6952)	$4.41{\cdot}10^{-4}$	$4.38{\cdot}10^{-4}$	$4.38{\cdot}10^{-4}$	$4.39{\cdot}10^{-4}$	$4.40{\cdot}10^{-4}$	$4.40{\cdot}10^{-4}$	$5.39{\cdot}10^{-4}$	$5.36{\cdot}10^{-4}$	$5.55{\cdot}10^{-4}$
京セラ (6971)	$5.59{\cdot}10^{-4}$	$5.61{\cdot}10^{-4}$	$5.63{\cdot}10^{-4}$	$5.63{\cdot}10^{-4}$	$5.74{\cdot}10^{-4}$	$5.77{\cdot}10^{-4}$	$5.77{\cdot}10^{-4}$	$9.89{\cdot}10^{-4}$	$13.25{\cdot}10^{-4}$
松下電工 (6991)	$2.28{\cdot}10^{-4}$	$2.90{\cdot}10^{-4}$	$2.91{\cdot}10^{-4}$	$2.91{\cdot}10^{-4}$	$2.93{\cdot}10^{-4}$	$2.95{\cdot}10^{-4}$	$2.92{\cdot}10^{-4}$	$2.90{\cdot}10^{-4}$	$2.92{\cdot}10^{-4}$
キャノン (7751)	$4.79{\cdot}10^{-4}$	$4.79{\cdot}10^{-4}$	$4.81{\cdot}10^{-4}$	$4.88{\cdot}10^{-4}$	$4.89{\cdot}10^{-4}$	$4.90{\cdot}10^{-4}$	$4.93{\cdot}10^{-4}$	$7.36{\cdot}10^{-4}$	$7.51{\cdot}10^{-4}$

なし，訓練データの各期の各銘柄の収益率を入力したときの予測値と正しい値との平均二乗誤差を小さくするように BFGS(Broyden-Fletcher-Goldfarb-Shanno) 準ニュートン法を用いて学習を行う．これは，学習過程において，次のような近似方法で結合荷重及び閾値の更新値を求める方法である．いま，関数 $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ に対し，点 $\boldsymbol{x}_0$ の周りで二次の項までテイラー展開した $f(\boldsymbol{x}+\varepsilon) = \boldsymbol{x}_0 - \nabla f(\boldsymbol{x})^{-1}\varepsilon + \frac{1}{2}\varepsilon^T \boldsymbol{H}(\boldsymbol{x})\varepsilon$ の式において，$\frac{\partial f(\boldsymbol{x})}{\partial \varepsilon} = 0$ と置くことで求められる関係

$$\varepsilon = -\boldsymbol{H}(\boldsymbol{x})^{-1} \cdot \nabla f(\boldsymbol{x}) \tag{1}$$

を用いて

$$\boldsymbol{x} = \boldsymbol{x}_0 - \nabla f(\boldsymbol{x})^{-1}\varepsilon \tag{2}$$

によって値更新を行うことを考える．このとき，式 (1) 中のヘシアン $\boldsymbol{H}(\boldsymbol{x})$ が実際の計算の中で直接求まらない場合もあるため，このヘシアンを近似するための何らかの関数が必要となるが，これに対して $\boldsymbol{s}_t = \boldsymbol{x}_{t+1} - \boldsymbol{x}_t$，$\boldsymbol{y}_t = \nabla f(\boldsymbol{x}_{t+1}) - \nabla f(\boldsymbol{x}_t)$ と置いて，

$$\boldsymbol{H}_{t+1} = \boldsymbol{H}_t - \frac{\boldsymbol{H}_t \boldsymbol{s}_t (\boldsymbol{H}_t \boldsymbol{s}_t)^T}{\boldsymbol{s}_t^T \boldsymbol{H}_t \boldsymbol{s}_t} + \frac{\boldsymbol{y}_t \boldsymbol{y}_t^T}{\boldsymbol{s}_t^T \boldsymbol{y}_t} \tag{3}$$

としてヘシアンを近似することによって，更新値を求めるものである．

学習の終了条件に関して，許容絶対誤差，許容相対誤差，最大学習反復回数を設定した．許容絶対誤差 Δx とは，3 層パーセプトロンの出力値 x_0 と正しい値 x との差分の絶対値 ($\Delta x = |x_0 - x|$) であり，許容相対誤差 δx とは絶対誤差を正しい値で正規化した値 ($\delta x = \frac{\Delta x}{x} = \left|\frac{x_0 - x}{x}\right|$) である．学習過程における許容絶対誤差と許容相対誤差との値が，予め定められたそれぞれの設定値以下になった場合に学習を終了させる．また，学習のための損失関数として学習誤差の平均平方和に関する最小化関数を用いた．なお，本稿では，学習を行う際の結合荷重及び閾値の初期値は，各結合荷重及び閾値についてあらかじめランダムに生成して用意した値を常に用いるものとした．

学習後のネットワークを用いた予測は，ネットワークにテストデータの各期の各銘柄の日次収益率を入力することで行われる．その際のネットワークの出力を予測値とみなし，実際の予測銘柄の収益率との平均二乗誤差を予測誤差とする．

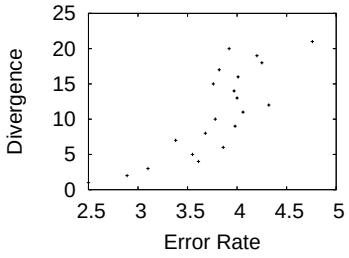


Fig. 2　予測誤差とダイバージェンスの対応関係

4　評価

4.1　予測銘柄の収益率のみを入力データとした場合

はじめに，提案手法の予測精度を検証するための比較基準として，予測銘柄に関するデータのみを入力として与えた場合の予測誤差を計測した．なお，評価実験における各パラメタの値は Table 2 のように設定した．

具体的には，Table 1 に示される 21 銘柄それぞれを予測対象として設定し，前期 1 日分の収益率データのみを与える場合から，過去 9 日分のデータ全てを与える場合まで，予測のために利用するデータの日数を逐次 1 日分ずつ増やしていき，それぞれの場合の予測誤差を計測した．そして，全てのデータ日数に対して予測を行った後に，最も予測誤差が低かったものを，予測銘柄の収益率データのみを与えた場合に達する最良の予測精度であると定義した．

Table 4　銘柄ごとの最良予測誤差と Kullback-Leibler のダイバージェンス

社名	予測誤差	Divergence	順位 (予測誤差)	順位 (Divergence)
ミネビア (6423)	$8.62 \cdot 10^{-4}$	4.25	18	19
日立製作所 (6501)	$7.57 \cdot 10^{-4}$	3.97	14	13
東芝 (6502)	$7.60 \cdot 10^{-4}$	3.76	15	8
三菱電機 (6503)	$7.06 \cdot 10^{-4}$	4.32	12	20
富士電機 H(6504)	$6.01 \cdot 10^{-4}$	3.78	10	9
日本電気 (6701)	$8.31 \cdot 10^{-4}$	4.01	16	16
富士通 (6702)	$9.84 \cdot 10^{-4}$	4.20	19	18
沖電気工業 (6703)	$9.95 \cdot 10^{-4}$	3.92	20	12
松下電器産業 (6752)	$4.26 \cdot 10^{-4}$	3.61	4	6
シャープ (6753)	$3.65 \cdot 10^{-4}$	3.10	3	3
TDK(6762)	$7.50 \cdot 10^{-4}$	4.00	13	15
三洋電機 (6764)	$5.96 \cdot 10^{-4}$	3.98	9	14
アルプス電気 (6770)	$8.37 \cdot 10^{-4}$	3.82	17	10
パイオニア (6773)	$4.73 \cdot 10^{-4}$	3.86	6	11
クラリオン (6796)	$22.16 \cdot 10^{-4}$	4.76	21	21
横河電機 (6841)	$6.86 \cdot 10^{-4}$	4.06	11	17
デンソー (6902)	$3.02 \cdot 10^{-4}$	2.89	2	2
カシオ計算機 (6952)	$4.38 \cdot 10^{-4}$	3.55	5	5
京セラ (6791)	$5.59 \cdot 10^{-4}$	3.68	8	7
松下電工 (6991)	$2.28 \cdot 10^{-4}$	2.50	1	1
キャノン (7751)	$4.79 \cdot 10^{-4}$	3.38	7	4

各銘柄を対象とした実験における使用データ日数とそのときに観測された予測誤差を Table 3 に示す．Table 3 の網掛け部分はその日数で最良誤差を記録したことを表している．Table 3 は，最良となる予測誤差の水準が，銘柄間で異なっていることを表している．例えば，クラリオンの次期収益率を予測対象とした場合には $22.16 \cdot 10^{-4}$ 以下に予測誤差を下げることができない一方で，松下電工の次期収益率を予測対象とした場合には予測誤差を $2.28 \cdot 10^{-4}$ まで抑制できることを示している．また，パイオニアの次期収益率を予測対象とした場合の最良予測誤差である $4.73 \cdot 10^{-4}$ と，キャノンの次期収益率を予測対象とした場合の最良予測誤差である $4.79 \cdot 10^{-4}$ との差が $6 \cdot 10^{-6}$ と非常に小さい値となっている一方で，松下電工の次期収益率を予測対象とした場合の最良予測誤差である $2.28 \cdot 10^{-4}$ と，クラリオンの次期収益率を予測対象とした場合の最良予測誤差である $22.16 \cdot 10^{-4}$ との間の差は $19.88 \cdot 10^{-4}$ と非常に大きな値となっている．

そこで，銘柄間で最良となる予測誤差の水準に差を生じさせている要因を明らかにするために，-0.1 から 0.1 までの範囲を $4 \cdot 10^{-4}$ 刻みに分割した区間 (区間数: 500) を考え，各区間の生起確率に基づいて，訓練データに対するテストデータの Kullback-Leibler のダイバージェンス (2 つの分布のずれ幅を表す指標) を計算し，その値と，その銘柄について最良となる予測誤差との比較を行った．Fig. 2 は，この Kullback-Leibler のダイバージェンスと予測誤差との対応関係を示したものであり，両者はほぼ線形な関係となっていることがわかる．さらに Table 4 は Table 1 の 21 銘柄に関して，最良予測誤差と，学習データに対するテストデータの Kullback-Leibler のダイバージェンスをそれぞれ記した表である．この予測誤差と Kullback-Leibler のダイバージェンスの数値と順位それぞれに対して Peason の相関係数，Spearman の順位相関係数を計測した結果，それぞれ 0.695，0.614 となり，両者の間に正の相関が見られることが示された．Kullback-Liebler のダイバージェンスが低いということは訓練データとテストデータとで分布特性の差異が小さいということを示しており，この結果は分布特性の差が小さいほど予測誤差が低くなることを示唆している．したがって，収益率分布の長期的な安定性は高い予測精度が生じるひとつの条件となっていると言うことができる．

Table 5　予測銘柄のみにおける最良誤差と適応的データサンプリングによる複数銘柄入力における最良誤差

社名	予測銘柄のみ	複数入力	改善率
ミネビア (6479)	$8.62 \cdot 10^{-4}$	$8.27 \cdot 10^{-4}$	4.23%
日立製作所 (6501)	$7.57 \cdot 10^{-4}$	$7.22 \cdot 10^{-4}$	4.85%
東芝 (6502)	$7.60 \cdot 10^{-4}$	$7.39 \cdot 10^{-4}$	2.84%
三菱電機 (6503)	$7.06 \cdot 10^{-4}$	$6.99 \cdot 10^{-4}$	1.00%
富士電機 H(6504)	$6.01 \cdot 10^{-4}$	$5.84 \cdot 10^{-4}$	2.91%
日本電気 (6701)	$8.31 \cdot 10^{-4}$	$7.90 \cdot 10^{-4}$	5.19%
富士通 (6702)	$9.84 \cdot 10^{-4}$	$9.75 \cdot 10^{-4}$	0.92%
沖電気工業 (6703)	$9.95 \cdot 10^{-4}$	$9.45 \cdot 10^{-4}$	5.29%
松下電器産業 (6752)	$4.26 \cdot 10^{-4}$	$4.50 \cdot 10^{-4}$	-5.33%
シャープ (6753)	$3.65 \cdot 10^{-4}$	$3.60 \cdot 10^{-4}$	1.39%
TDK(6762)	$7.50 \cdot 10^{-4}$	$7.15 \cdot 10^{-4}$	4.90%
三洋電機 (6764)	$5.96 \cdot 10^{-4}$	$5.60 \cdot 10^{-4}$	6.43%
アルプス電気 (6770)	$8.37 \cdot 10^{-4}$	$8.00 \cdot 10^{-4}$	4.62%
パイオニア (6773)	$4.73 \cdot 10^{-4}$	$4.52 \cdot 10^{-4}$	4.65%
クラリオン (6796)	$22.16 \cdot 10^{-4}$	$21.67 \cdot 10^{-4}$	2.26%
横河電機 (6841)	$6.86 \cdot 10^{-4}$	$6.56 \cdot 10^{-4}$	4.57%
デンソー (6902)	$3.02 \cdot 10^{-4}$	$2.92 \cdot 10^{-4}$	3.42%
カシオ計算機 (6952)	$4.38 \cdot 10^{-4}$	$4.21 \cdot 10^{-4}$	4.04%
京セラ (6971)	$5.59 \cdot 10^{-4}$	$5.38 \cdot 10^{-4}$	3.90%
松下電工 (6991)	$2.28 \cdot 10^{-4}$	$4.21 \cdot 10^{-4}$	-45.84%
キャノン (7751)	$4.79 \cdot 10^{-4}$	$4.67 \cdot 10^{-4}$	2.57%
平均	$7.07 \cdot 10^{-4}$	$6.94 \cdot 10^{-4}$	0.90%

4.2　適応的データサンプリングを用いた場合

次に，適応的データサンプリングを導入することで，前節の予測銘柄のデータのみを入力として与えた実験で最良となっていた値以下に，予測誤差を抑制することができるかどうかを検証する．評価実験では任意の銘柄に対して前期 1 日分のみの収益率データを採取することにして，それぞれのパラメタは Table 2 のように設定した．

Table 1 に示されるそれぞれの銘柄を予測対象とした場合に，予測銘柄のみの場合と適応的データサンプリングを用いた場合の最終世代における誤差比較を Table 5 に示す．この Table 5 中の網掛け部分は予測誤差が低いことを表している．同図から，松下電器産業と松下電工を除く 19 銘柄において，適応的データサンプリングを導入したもとでの予測誤差が，導入しない場合の全てのデータ日数における予測誤差よりも低い値に抑制されていることが見て取れる．実際に，この 2 つの系列に対して，有意水準 5% で Wilcoxon の符号付順位和検定を行ったところ，“2 群間で母集団の中央値に差は無い” という帰無仮説は棄却され，提案手法の導入は優位な差を生み出していることが判明した．

実際の平均改善率は，松下電工を予測対象とした場合に-45.84% という特異的な値が観測されたため，その値に引きづられて 0.90% と比較的低い値に抑えられる結果となった．しかし，松下電工の時系列は訓練データとテストデータの分布特性が特に小さく (Table 4),，はじめから予

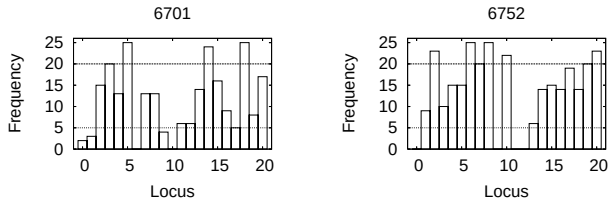
測誤差が極めて小さいために，他のどのような銘柄の時系列を組み込んでもこの分布間の特性の差異を大きくするノイズとしてしか機能しなくなる特殊なものであると考えられ，この松下電工の改善率を除いた場合の平均値を算出したところ，3.23% の改善率であった．Table 1 における 21 種のテスト期間における日次収益率の平均値が-0.01 であるため，松下電工を除いた場合の改善率は価格に対して，$0.03 \cdot 0.01 \approx 0.0004$ 程度の改善率を示していることとなり，これはキャノンの 2003 年 12 月 30 日終値である 4990 円を基準として考えると約 2 円程度の改善であることがわかる．

ところで，予測銘柄のみを入力とした場合では，次期の収益率を精度良く予測するために求められる条件として，収益率が長期に渡って安定的な挙動を示し，テストデータと訓練データとの間の分布特性の差が小さいことが挙げられた．そして，このような特徴は，複数の種類の銘柄を入力として予測を行う場合にも，高い予測精度のためには基本的に満たされているべきであると考えられる．このように考えると，例えば，複数の銘柄の前期収益率を入力とした場合に，ある銘柄の次期収益率が高い精度で予測されるためには，入力データとして組み入れられる銘柄の前期収益率と予測対象として設定される銘柄の次期収益率との関係が長期に渡って安定しているという予測銘柄と組み入れ銘柄との相互関係が重要だと考えられる．

小さな予測誤差をもたらす銘柄同士の関係性について一般的な知見を得るために，まず，高い予測精度を達成しているエリート個体群がどのような銘柄を入力として採用する傾向があるかに注目した．Table 1 で 0.1 として設定されていたエリート保存率を 0.5 まで増やし，世代数を 200 まで増やした上で，具体例として日本電気 (6701) と松下電器産業 (6752) を予測対象としてとりあげ評価実験を行い，最終世代において各エリート個体の染色体中に，21 銘柄それぞれがどのような頻度で現れるかを調べた．その結果をあらわしたのが Fig. 3 であり，横軸は各銘柄に対応する遺伝子座 (Table 1) を表し，縦軸は上位 25 個体のうち対応する銘柄が何個体から選ばれたかという頻度を表している．同図左から，日本電気を予測銘柄とする場合には，エリート個体群内で日本電気，クラリオン，京セラが高い頻度で選択されているが，ミネビア，日立製作所，シャープはほとんど選択されておらず，選択される銘柄に偏りがあることがわかる．同時に，同図右の松下電器産業の場合と比較すると，その偏りは予測銘柄ごとに異なることがわかる．また，解析の結果，同図において，25 個体中 20 個体以上 (75% 以上の個体) の染色体に組み入れられる銘柄は，最良個体の染色体に必ず組み入れられ，25 個体中 5 個体以下 (25% 以下の個体) の染色体のみにしか組み入れられない銘柄は，最良個体の染色体に必ず組み入れられない傾向が見られることがわかり，最良個体の染色体はエリート個体全体の選択傾向をよく反映していることも明らかになった．

そこで，21 銘柄それぞれの次期収益率を予測対象とした時に，どのような銘柄の前期収益率が最良個体の染色体の中に組み入れられるかについて調べた．Table 6 は，Table 2 の設定のもとで，縦軸に予測銘柄，横軸に評価実験において最終世代で得られる最良個体の染色体ビット列として，各予測銘柄に依存して選択される銘柄の組合せを表したものである．たとえば，2 行 3 列のビットが 0 であるということは，日立製作所の次期収益率を予測対象とした場合には，東芝の前期収益率が入力データとして組み入れられないことを表している．

また，この Table 6 における右端の段と下端の段 (合計) の値はそれぞれ次のような銘柄同士の関係性を示しているといえる．



日本電気 (6701)　　松下電器産業 (6752)

Fig. 3　日本電気 (6701) と松下電器産業 (6752) を予測対象とした場合の上位 25 個体の各銘柄の組み入れに関するヒストグラム

右端の段　ある銘柄の次期収益率を予測する 1 回の実験において，どの程度の銘柄の前期収益率が入力データとして組み入れられるか．

下端の段　全ての銘柄に対して次期収益率を予測させる 21 回の実験において，ある銘柄の前期収益率はどの程度の回数，入力として組み入れられることになるか．

前者について最大値をとるのは松下電工であり，自身を含む 13 銘柄を予測データとして利用している一方で，最小値をとるのはシャープと TDK であり，5 銘柄のみを利用していることがわかる．後者について最大値をとるのは富士通であり，自身を含む 16 銘柄から予測データとして利用されているが，日立製作所とアルプス電気はわずか 5 銘柄からしか利用されていないことがわかる．また，この両者の平均がともに 9.05 であるのに対して，前者の標準偏差が 2.33，後者の標準偏差が 3.17 となり，後者の方により強い分布の偏りが見られることもわかった．

この両者の分布に偏りを生じさせている要因として，銘柄間の関係性が非対称であることが挙げられる．例えば，Table 6 において，ミネビア (6479) を予測対象とした場合には，最良個体の組合せでは東芝 (6503) が入力データとして組み入れられるが，逆に東芝を予測対象とした場合には，最良個体の組合せではミネビアは入力データとして組み入れられない．ここで，ある銘柄の予測に他の銘柄のデータを組み入れた際に誤差が小さくなることが，組み入れた銘柄から予測銘柄に対して，価格の変動について強い影響があることを示すものと仮定すると，上記の結果は，東芝の前期収益率はミネビアの次期収益率に影響を及ぼしているが，ミネビアの前期収益率は東芝の次期収益率に影響を及ぼしていないということを意味していると考えることができ，東芝からミネビアへの一方向な影響力が存在するといえる．

また，ある銘柄の次期収益率がどれ位の銘柄の前期収益率から影響を受けているかについては銘柄間で大きな差はないが，ある銘柄の前期収益率がどれ位の銘柄の次期収益率に影響を及ぼしているかについては銘柄ごとに大きな偏りがあるというこの結果は，他の銘柄の次期収益率に影響を及ぼしている “限られた少数の” 支配的な銘柄が存在している一方で，その銘柄の影響を受ける対象となる従属的な銘柄に関しては “一様に” 分布しているということを示すものである．

以上の結果を踏まえると，本提案手法が予測精度を向上させた理由として，このような支配的な銘柄を適応的に発見し，その銘柄と，その銘柄が影響を及ぼす対象である予測対象銘柄との間にある安定的な収益関係を予測に取り込むことで，長期的な分布特性の差異をより小さくすることができたため，予測銘柄のみのデータを入力としていた場合以上に予測誤差が改善されるようになったと考えることができる．

Table 6　各銘柄における最良個体のビット列

社名	6479	6501	6502	6503	6504	6701	6702	6703	6752	6753	6762	6764	6770	6773	6796	6841	6902	6952	6971	6991	7751	合計
ミネビア (6479)	1	0	1	1	0	1	1	0	0	1	0	1	1	0	0	1	1	1	0	0	1	12
日立製作所 (6501)	1	1	0	0	0	0	1	1	1	0	0	0	1	0	1	1	0	0	0	0	1	9
東芝 (6502)	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0	1	1	1	0	0	8
三菱電機 (6503)	0	0	0	1	1	0	0	1	0	0	0	0	0	1	1	1	0	1	0	0	1	8
富士電機 H(6504)	1	0	0	1	1	0	1	0	0	0	1	0	0	0	1	1	1	1	0	0	0	9
日本電気 (6701)	0	0	0	1	1	1	0	1	1	0	0	0	0	1	1	1	1	0	1	0	1	11
富士通 (6702)	0	1	0	1	1	0	1	1	0	0	0	1	0	1	1	1	0	1	1	0	1	12
沖電気工業 (6703)	0	0	0	1	0	0	1	1	0	0	0	0	0	1	1	1	0	1	0	0	1	8
松下電器産業 (6752)	0	0	1	0	1	1	1	1	1	0	1	0	0	1	0	1	0	1	0	1	1	12
シャープ (6753)	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0	0	5
TDK(6762)	0	0	0	0	0	1	0	0	1	0	1	1	0	1	0	0	0	0	0	0	0	5
三洋電機 (6764)	0	1	1	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	7
アルプス電気 (6770)	0	0	0	1	1	0	1	1	0	1	0	0	1	1	0	0	0	1	1	0	0	9
パイオニア (6773)	0	0	1	0	1	0	1	0	1	0	1	1	0	1	1	1	1	0	0	1	0	11
クラリオン (6796)	0	0	1	0	0	0	1	0	0	1	0	1	0	1	1	1	1	0	1	1	1	11
横河電機 (6841)	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	1	1	1	0	1	0	8
デンソー (6902)	0	0	1	1	1	0	0	0	0	0	1	0	0	0	0	1	1	0	1	1	0	8
カシオ計算機 (6952)	0	0	1	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1	0	1	0	6
京セラ (6971)	0	1	1	0	0	1	1	0	1	0	1	0	0	0	0	1	0	0	1	1	0	9
松下電工 (6991)	1	1	1	1	0	1	0	0	0	1	1	1	0	1	0	0	1	1	0	1	1	13
キャノン (7751)	1	0	1	0	1	0	0	1	0	1	0	1	0	1	1	0	0	0	1	1	1	11
計	6	5	12	10	10	6	16	8	8	8	7	8	5	10	9	15	9	11	7	9	11	-

5　おわりに

従来の株価時系列予測では，学習モデルの改良に焦点を合わせることが多く，予測精度を高めるという観点から入力データの与え方自体についての検討は十分に行われていなかった．本稿では，ある銘柄の収益率を複数の銘柄の収益率データから予測する問題において，遺伝的アルゴリズムに基づく適応的データサンプリングを導入することによって，予測にとって望ましい入力の組み合わせを選択し，予測銘柄のデータのみを入力としていた場合と比べて，より良い予測精度を実現できることを示した．

また，最良の予測誤差をもたらす組み合わせを解析したところ，本提案手法を導入することによって，約 3% の予測精度の向上が見られた背景には，本提案手法が，長期に渡って安定的に作用を及ぼしている銘柄の組み合わせを発見し，そのような銘柄の収益率データを入力データに組み入れながら予測を行うことで，長期的な分布特性の差異をより小さくしていることがあると考えられた．

今回，適応的データサンプリング手法を株価の日次収益率を予測する問題に対してのみ用いたが，今後はより幅広い時系列予測の問題に対して，本提案手法を導入していく必要があると考える．また，予測誤差を改善するために組み入れられる銘柄同士にどのような関係性が存在するのかを本質的に理解するための枠組みの構築が必要であると言える．

参考文献

[1] Zekic, M.: Neural Network Application in Stock Market Predictions - A Methodology Analysis, The 9th International Conference on Information and Intelligent Systems '98, pp. 225-263 (1998).

[2] Trippi, R. R., DeSieno, D.: Trading Equity Index Futures with a Neural Network, The Journal of Portfolio Management, Vol. 19, No. 1, pp. 27-33 (1992).

[3] Kryzanowsky, L., Galler, M., Wright, D. W.: Using Artificial Networks to Pick Stocks, Financial Analyst Journal, Vol. 49, No. 4, pp. 21-27 (1993).

[4] Yoon, Y., Guimaraes, T., Swales, G.: Integrating Artificial Neural Networks with Rule-Based Expert Systems, Dicision Support Systems, Vol. 11, No. 5, pp. 497-507 (1994).