回帰分析によるストリームデータのクラスタリング

Clustering Stream Data by Regression Analysis

本吉 正博♥　三浦 孝夫♦
塩谷 勇♠

Masahiro MOTOYOSHI　Takao MIURA
Isamu SHIOYA

本稿では，局所的に異なる傾向を持つストリームデータにおけるクラスタリングについて述べる．このようなデータはそれぞれのクラスタが異なる線形関数で回帰すると考えられる．そこで，結合の基準として回帰式の *F* 検定統計値を用いた階層的クラスタリングの差分方法を提案する．本手法により，解釈可能な数のクラスタを抽出でき，かつストリームデータに適応できることを検証する．

In this investigation, we present clustering a collection of data stream where we see several trends within. Such kind of data could be regressed locally by using linear functions. We propose an incremental method of hierarchical clustering based on *F* value of regression. Using our method, we can extract clusters of a moderate number to interpret and to describe to stream data.

1. まえがき

クラスタ分析とは，異質なものが混じっているオブジェクトの中から，似ているものを凝集し，グループ（クラスタ，cluster）分けを行うためのアルゴリズムの総称を指す．すべての研究領域で各研究者が直面している一般的な問題は，観測されたデータをいかに意味のある体系に組織立てるか，すなわち，いかにグループ化を行うかにある．

一方，高速ネットワークとウェブ技術の発達によって，近年，電子商取引やウェブサイト管理など，新しい形態のネットワーク上の応用プログラムが普及し始めている．これらの応用プログラムでは，静的な情報集合ではなく，ネットワーク上を流れ続けるデータストリームの形態をしていることが特徴である．今後，このような大規模データストリームを対象とした，効率のよいクラスタリング手法がますます重要になる[1]．

一般的にクラスタ内の類似性が高く，クラスタ間の類似性は低いほどよいクラスタリングである．このクラスタ化の能力は，類似性の定義とその実行方法に依存する．使用された類似度が実際の類似度，あるいは，分析者にとって意味のある類似度であるかどうかについては何の保証もない．特定の応用に対して正しい方法を選択することは分析者の責任となる．また，隠れたパターンをどれだけ見出せるかが重要である．本研究では，対象として局所的に異なる傾向を持つデータを想定する．これは，局所的な部分線形空間を擁するデータ構造の推定の問題である．著者らは，クラスタを結合する基準として，分散とクラスタ内の回帰分析による*F*検定統計値を使用する方法を提案した[2]．本研究では膨大なストリームデータに対応するために，時間の経過に従ってオブジェクトの重みを指数的に下げることによって有限時間で処理する差分方法を提案する．

次章では，既存の方法で異なる傾向を持つクラスタを分類できないことを論じる．第3章では，いくつかの定義とデータの前処理について述べる．第4章では，クラスタを結合する方法とその基準について述べる．第5章では，ストリームデータへ対応させる方法について述べる．第6章では，実験結果を示して第7章で結びとなる．

2. 局所傾向を持つクラスタ

異なる傾向を持ったオブジェクトが混在しているようなデータのクラスタリングを考える．このようなデータは，局所的にはそれぞれが異なる線形関数で回帰できる．これらのクラスタは，互いに交差することも考えられる．

最も簡単な解決法は，階層的手法の最近隣法を用いることにより，一番距離の近いオブジェクトを探して数珠繋ぎにクラスタを結合することができるが，クラスタが交差するような場合は，交差点でクラスタが分断されてしまう．つまり，異なる傾向を持つデータであっても，距離さえ近ければ誤って合併してしまう可能性がある[3]．オブジェクト集合を重心で代表するk-means法は本来，凸型でないクラスタには向かない．分散の基準に加え，オブジェクトが移動することで双方のクラスタの線形性が向上するような点を探したとしても，そのような点はかならずあるわけではない．クラスタ数を決めなければいけないことも問題である．２つの方法に共通した問題は，クラスタ間の類似度にある．つまり，距離測度，分散の他に部分線形性を考慮した他の類似度を導入する必要がある．

事前に与えられたデータに基づき未来の事象を何らかの関数により予測する多変量解析手法に回帰分析がある．本研究では，回帰式の*F*値をクラスタ結合の類似度基準として導入する．これは，点でクラスタを代表させる手法に対して，線でクラスタを代表させる”線のクラスタリング”に相当する手法といえる．本研究では，距離，分散に加えこの*F*値を類似度の基準とし，徐々に線形クラスタを結合しながら目的のクラスタへと’復元’していく方法をとる．

3. 初期クラスタの選定

本稿では，複数の変数を扱う．変数は，すべて環境から与えられる入力であり，分類のための外的基準はないとする．分析者から与えられた回帰分析の際の基準となる１個の変数を基準変数とし，その他の複数個の変数を説明変数と言う．個々のオブジェクトが持つ説明変数$X = x_1, x_2, \ldots, x_m$，及び基準変数 $Y = y$ は列で表現する．オブジェクトはデータ行列の行で表現し，全体として $n \times (m+1)$ の行列を用いて記述する．ここで各変数は標準データ（z-score）に変換されているものとする．初期クラスタは，オブジェクトの集合であ

♥ 学生会員　法政大学大学院工学研究科修士課程
i02r3243@k.hosei.ac.jp
♦ 正会員　法政大学工学部情報電気電子工学科
miurat@k.hosei.ac.jp
♠ 正会員　産能大学経営情報学部情報学科
shioya@mi.sanno.ac.jp

$$(X|Y) = \left(\begin{array}{ccc|c} x_{11} & \cdots & x_{1m} & y_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{k1} & \cdots & x_{km} & y_k \\ \vdots & \ddots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nm} & y_n \end{array}\right) \quad (1)$$

り，各オブジェクトは，初期クラスタに排他的に含まれる．通常，凝集法において，最初の段階では，各オブジェクトが各クラスタを表しており，類似度はオブジェクト間の距離によって定義されるが，本手法では分散を持つ集合でなければならない．この最初の段階で初期クラスタを用いる．オブジェクトを小クラスタに分割し，これを初期クラスタとする．初期クラスタは内積により動的に求める[2].

4. クラスタの結合

この章ではクラスタ間の類似度を定義し，類似度基準が結合とどのように関係するかを述べる．我々は，ここでクラスタ間の類似度を２つの側面から定義する．類似度のひとつはクラスタ間距離である．我々は既にクラスタ重心間のユークリッド距離を用いる方法を提案している．しかしこの方法は重心のみで判断し，変数間の相関から生じるクラスタの偏りを考慮しない．クラスタ i に含まれるベクトルはそのクラスタの他のベクトルに対して役割を果たしていると考える．言い換えれば，分散と共分散に影響を与えている．そこで，距離測度には重心だけでなく分散も考慮するマハラノビス距離を適用する．それぞれ一方のクラスタともう一方のクラスタ重心とのマハラノビス距離を計算し，その平均をクラスタ間距離とする．

$$d^2(i,j)=\frac{(\mu_i-\mu_j)^T C_j^{-1}(\mu_i-\mu_j)+(\mu_j-\mu_i)^T C_i^{-1}(\mu_j-\mu_i)}{2} \quad (2)$$

ただし各クラスタの分散共分散行列を C, その逆行列を C^{-1}とする．これにより非類似度行列$d^2(i,j)(\in R^{n\times n})$を定義することができる．距離が最も小さいクラスタの組み合わせは，結合するクラスタの組み合わせの候補となる．その候補が結合する上で適正かどうかを検査しなければならない．

もうひとつの類似度基準は回帰式の検定統計値である F 値により定義する．F検定は，回帰式が予測に役立つのかを F値を用いて検定するためのものである．F値が F分布の数表に基づいた有意水準以上であれば，回帰式が予測に役立つといえる．

(1)式と同様のデータ行列で与えられるメンバ数 n のクラスタに対して重回帰分析のモデルは

$$y = b_1x_1 + b_2x_2 + \ldots + b_mx_m + e_i \quad (3)$$

である．ここでb_iの最小二乗推定量$\tilde{b}_i$は，

$$B = (\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_m) = (X^TX)^{-1}X^TY \quad (4)$$

であり，これを回帰係数という．実際には，標準データを前提としているため標準化回帰係数である．yは実測値であるのに対して，回帰係数Bによる予測値をYとする．この時，回帰による変動要因について，平方和S_Rと平均平方V_Rは，

$$S_R = \sum_{k=1}^{n}(Y_k - \bar{Y})^2 \quad ; \quad V_R = \frac{S_R}{m} \quad (5)$$

残差による変動要因について，平方和S_Eと平均平方V_Eは，

$$S_E = \sum_{k=1}^{n}(y_k - Y_k)^2 \quad ; \quad V_E = \frac{S_E}{n-m-1} \quad (6)$$

この時，

$$F_0 = \frac{V_R}{V_E} \quad (7)$$

は第１自由度 m，第２自由度 $n-m-1$ の F分布に従う．

次にメンバ数 a であるクラスタ A とメンバ数 b であるクラスタ Bが得られたとき，結合クラスタ $A\cup B$を考える．データ行列は次のようになる．

$$(X|Y) = \left(\begin{array}{ccc|c} x_{A11} & \cdots & x_{A1m} & y_{A1} \\ \vdots & \ddots & \vdots & \vdots \\ x_{Aa1} & \cdots & x_{Aam} & y_{Aa} \\ x_{B11} & \cdots & x_{B1m} & y_{B1} \\ \vdots & \ddots & \vdots & \vdots \\ x_{Bb1} & \cdots & x_{Bbm} & y_{Bb} \end{array}\right) \quad (8)$$

ただし $n = a + b$ である．この場合も，上記と同様に式(4)より回帰式が，式(7)より F 値を求めることができる．結合前の２つのクラスタから得られる F値と，結合クラスタの F 値の間の性質について下記に述べる．

クラスタA，クラスタB，結合クラスタのF値をそれぞれ F_A，F_B，Fとする．

[性質 1] $F_A > F$, $F_B > F$の場合

お互いに回帰の妨げになっている場合は，傾きが有意に異なると考えられる．ゆえに，クラスタAとBの類似性は低く，クラスタの線形性は減少する．特に，$F_A = F_B$, $F=0$ の場合はAとBの回帰式が重心を交点として直交している．

[性質 2] $F_A \leqq F$, $F_B \leqq F$の場合

２クラスタともF値が向上する場合は，傾きが有意に異ならないと考えられる．ゆえに，クラスタAとBの類似性は高く，クラスタの線形性は向上する．特に，$F_A = F_B$, $F=2F_A$の場合は，クラスタAとBのオブジェクト数と座標が全て一致している．

[性質 3] $F_A \leqq F$,$F_B > F$あるいは$F_B \leqq F$,$F_A > F$の場合

一方が向上し，一方が減少するような場合は，クラスタ A と Bの分散または，F値の差が大きい時に起こる．本来結合すべきである場合も，そうでない場合も存在する．

これらより，F値による類似度基準により結合すべきルールは，性質２の場合のみであることがいえる．マハラノビス距離を用いた非類似度は，局所的な距離以上に離れたクラスタと結合しないようにするための方法である．しかし，本手法では，F値基準が絶対的な基準であるゆえに，F値基準を満足するまで距離基準を下げ続けて結合パターンを探索する．そのため，初期クラスタの不良や，もともと局所的に線形回帰できるクラスタをもたないデータの場合，距離基準をどこまでもさげてしまい，結合すべきでないクラスタが結合する可能性がでてくる．この問題を解決するために，離れてもよい距離の閾値Δを分析者が与えるものとする．F値基準とΔを両方満足する場合，クラスタを結合できる．

Δにより，クラスタの内分散が抑えられるため，距離の遠すぎるクラスタ同士が結合する心配がなくなる．この結合アルゴリズムは，性質２とΔを同時に満足する組が存在する限りクラスタを結合し続ける[2].

5. ストリームデータの表現

通常，ストリームデータは時系列的に配置され順次入力される．しかし局所的に傾向が変化しており，データ発見の研究分野ではこの変動をどう扱うかがポイントとなる[4]. 本章では，これまで述べた局所的な回帰傾向を持つデータのクラスタ化手法が，データストリームに対して有効に作用することを示す.

多くの人々にとっては過去の出来事より最近の出来事の方が興味深いであろうと考えるとき，膨大なストリームデータを扱うためには，過去のデータを記憶し続けることは効果的ではない．本研究では，将来生じるであろうデータを受け取るために，過去のデータを要約する.

時間軸上に連続したブロックを想定しこれを処理単位uと呼び，それぞれの処理単位に含まれるオブジェクトに重みを与える．過去に進むにつれ古い処理単位に含まれるオブジェクトの重みは指数的に減少していくものとする．分析者は重みとしてwを与え，新しい順にi番目の処理単位に含まれるすべてのオブジェクトはw_iの重みをかける．また，重みの閾値δを与え，一度に処理する処理単位を処理対象と呼び，その処理単位数をh個とする．ただし$w^{h-1} \leqq \delta$.

h 個の処理単位 $u_0,u_1,\cdots,u_{h-1}$に，新しい処理単位 u_0 を追加することを考える．まず初めに最も古い処理単位u_{h-1}を削除する．この時，他の$u_0,\cdots,u_{h-2}$のオブジェクトは重みを次の段階へ移動する．最後に$u_1,\cdots,u_{h-1}$と新しい処理単位u_0で本手法を実行する.

重みが指数的に減少するため，重みw^iを持ったオブジェクトはw^i個のオブジェクトとして考慮される．ここで重み$w^0,\cdots,w^n$を持つオブジェクト$e_0,\cdots,e_n$の変数Aの値$a_0,\cdots,a_n$を想定する(ただし$w^i > 0.0, i=0,\cdots,n$)．$p_i=w^i/W$, $W=w^0+\cdots+w^n$とすると，$E[A]=p_0\times a_0+\cdots+p_n\times a_n$を重み付き期待値という．また関数$f(X)$がある時，$E[f(A)]$を$p_0\times f(a_0)+\cdots+p_n\times f(a_n)$とおくと，$V[A]=E[(A-E[A])^2]=E[A^2]-E[A]^2$を重み付き分散という．これと同様に，2 つの変数$A,B$の重み付き共分散は$C[AB]=E[AB]-E[A]E[B]$である．マハラノビス距離に基づく非類似度$d^2(i,j)$を始め，$S_R, S_E, V_R, V_E$, そして$F$値はすべて上記の重み付き統計値へと拡張できる.

u_0のすべてのクラスタが既存のクラスタのいずれかに結合できる場合，重心，分散，回帰係数，F値を更新して既存のクラスタを保つ．つまり，既存のクラスタに単純に追加される．結合できないクラスタがある場合は，時間経過による変動が起こったと考えられるので，$u_1,\cdots,u_{h-1}$のすべてのクラスタを初期クラスタまで分解してクラスタを作り直す.

このようにしてクラスタリングと各種統計値を調整した後，u_{h-1}のすべてのオブジェクトを単純に削除する．この方法は合理的であると考えられる．なぜならクラスタリングの質を保つ間，我々はできるだけ多くのオブジェクトを反映させようとする（つまりクラスタリング結果が最新の傾向を反映する）．u_{h-1}の重みは最も小さいのでそのオブジェクトがクラスタへ与える影響は少ない.

以下に**ICFR**(*Incremental Clustering using F-value by Regression analysis*)のアルゴリズムを示す．計算量については再クラスタリングの発生回数に依存するが，通常は少数回の回帰分析の計算量と考えることができる.

1. $u_1,\cdots,u_{h-1}$の処理単位毎に初期クラスタを計算する.
2. 初期クラスタに対して *CFR* を適用する.
3. u_0のオブジェクトで初期クラスタを計算する.
4. 3.の結果を 2 の結果と結合できるか調べる．できないならば$u_0,u_1,\cdots,u_{h-1}$に存在するクラスタを初期クラスタまで分解して再クラスタリング(*CFR*)を実行する.
5. u_{h-1}を捨てて$u_0,\cdots,u_{h-2}$の番号を$u_1,\cdots,u_{h-1}$へと書き換える．それと同時に重みも次の段階へ移動する.
6. 回帰分析とクラスタ間距離の再計算を行い，3.に戻る.

6. 実験

この章では，本手法の有効性を検証するため，気象データを用いた実験を行う．実験に用いたデータは，新潟気象台と稚内気象台の作成した 1997 年の気象観測時別値データで，各気象台からそれぞれ 8736 件を単純に連結した合計 17472 件の気象データである[5]. サイズは 720KB である．これは，もともと局所的に線形回帰できるという初期条件の元で本手法を適用するためである.

データは，1 時間ごとに観測された合計 22 項目の変数を持ち，この実験ではそのなかから数値データでしかも欠損値のない，日（day），時（hour），現地気圧（hPa），海面気圧（hPa），気温（℃），露点温度（℃），蒸気圧（hPa），相対湿度（%）の 8 項目を変数の対象とする．この他に観測地点番号があり，クラスタリングの評価にのみ用いる.

実験では，すべての変数を標準化し，提案するアルゴリズムによって解析した．基準変数は，気温としその他の変数は説明変数とした．初期値決定のための閾値Θは 0.6，クラスタ間距離の閾値Δは 1.0E+6 とした．処理単位は 1 週間，重み w は 0.8，重みの閾値 δ は 0.25 とした．つまり処理対象は 7 ブロックである．1 年分の気象データをストリームデータとみなして逐次処理したところ 47 回中 12 回の割合で再クラスタリングが発生した．9 月 10 日の時点の結果を表 1 に示す.

	内分散	$\mathcal{F}$ 値	所属クラスタ数	新潟	稚内
Cluster1	6.71152	76433.5	5	102	152
Cluster2	4.50243	36137.4	14	653	315
Cluster3	1.67056	12109.1	3	119	47
Cluster4	2.96373	4404.97	1	48	48
Cluster5	0.12095	5528.73	1	29	0
Cluster6	0.44520	826016	7	0	537
Cluster7	0.09501	28841.3	3	216	77

表 1　最終クラスタ
Table 1　Final Clusters

得られた 7 個のクラスタのうち，特にクラスタ 2 と 6 は含まれるオブジェクト数が多いことが分かる．クラスタ 2 では，14 個の初期クラスタが結合された．クラスタ 6 では，7 個の初期クラスタが結合された．一般的に言って，この結果は観測地点による特徴を示している．実際，クラスタ 2 は新潟地点のオブジェクト 1176 個のうち 653 個（55.6%）のオブジェクトを含んでいる．クラスタ 6 は，新潟地点のオブジェクト 1176 個のうち 537 個（45.7%）を含んでいる．このことから，クラスタ 2 は新潟地点，クラスタ 6 は稚内地点固有の傾向を持つオブジェクトをそれぞれよく抽出していることがわかる.

例えば表 2 で示される各クラスタの重心から，クラスタ 2 は他のクラスタと比較して露天温度，蒸気圧，気温が高く，湿度が低い．ゆえに，南の地域，かつ夏に降水量の少ない日本海側の地域で観測されたと推測できる．クラスタ 6 は気圧が高く，露天温度，蒸気圧が低い．また，月が低く真夏に近いが気温は低い．ゆえに，北の地域，かつ高度が低い地域で

観測されたと推測できる.クラスタ1は月,日に特徴がある.月が高く日が低いことから観測地域とは関係ない，両地点に共通した最近(つまり 9 月上旬)の気象状態（季節の変化等）を示すクラスタであると推測できる．実際，クラスタ 1 は気温が低く，新潟地点と稚内地点のオブジェクトをほぼ同数持っている.

	Clust1	Clust2	Clust6
月	1.18569	0.47771	-0.2102
日	-0.4501	-0.0566	-0.2631
時	0.08692	-0.0682	-0.0298
気圧	-0.0020	0.00151	0.40699
海面気圧	0.00152	-0.0075	0.46129
露天温度	0.01840	0.17168	-1.1524
蒸気圧	0.01640	0.13634	-1.1048
相対温度	0.55982	-0.2062	0.57358
気温	-0.2807	0.23217	-1.2369

表 2　クラスタ重心
Table 2　Center of Gravity for Clusters

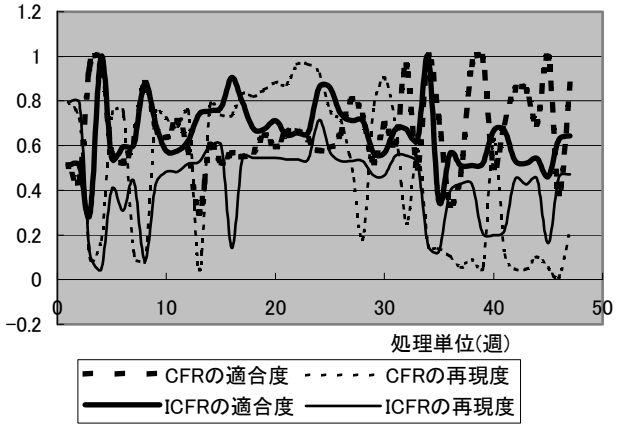

図 1　新潟地点の適合度と再現度
Fig.1　Precision/Recall Factors at Niigata

同じデータに対して差分計算を行わない *CFR* で比較実験を行った．新しい入力のたびにそれを含めて過去 7 週分をひとつの処理単位として再クラスタリングした．各クラスタに対して各地点のオブジェクトの再現度と適合度を計算する．例えば，*A* クラスタに対する新潟地点の再現度はすべての新潟オブジェクトにおける *A* に含まれる新潟オブジェクトの割合で，適合度は *A* のすべてのオブジェクトにおける *A* に含まれる新潟オブジェクトの割合である．もし再現度で最も高いのが *A* クラスタに対する新潟地点の再現度ならば，*A* を新潟クラスタとみなす．この時，*A* クラスタを除く他のクラスタで稚内地点の再現度の最も高いクラスタを稚内クラスタとする．図 1 に新潟地点に対する再現度と適合度を示す.

再現度について，*ICFR* は *CFR* に比べ 20%ほど低下しているのが分かる．これは差分計算から生じた誤差によるものと考えられる．適合度についてはこの結果からは有意な差は認められない．つまり，本差分的手法を用いることで，初めからクラスタ作る場合とほぼ同様の精度を得られる．稚内地点もほぼ同様の結果であった．また *CFR* では，過去のクラスタリング結果と対応がないが，*ICFR* では再クラスタリングが起こらない限りクラスタに大きな変化がないと判断できる．本実験では，平均して 4 処理目に再クラスタリングが発生しているため，大きな気象変化は 1 ヶ月毎に発生していると推測できる．図 2 は合計処理時間の推移である．差分計算を行うことでおおよそ 2 割程度にまで処理時間が減少し，明白な効果を生んでいる.

以上から，本実験によりデータが初期条件を満足することを示せた．処理時間については I/O 処理を効率化することでさらに向上すると考えられる.

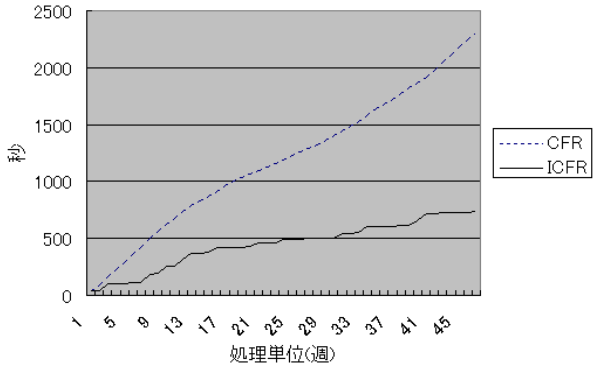

図 2　合計処理時間の推移
Fig.2　Total Processing Time

7. 結論

本稿では，異なる局所的な傾向を持つオブジェクトが混在したストリームデータに対して，有効なデータクラスタリングの新たな手法を提案した．クラスタの傾向を回帰分析を用いて抽出し，回帰式の *F* 値をクラスタの類似度として定義した．クラスタの精度を保つためのクラスタ間距離の閾値を定義した．これらを用いて，ストリームデータに対する差分方法を提案し，実験によりストリームデータに対する本手法の有効性を示した.

[謝辞]

本研究の一部は文部科学省科学研究費補助金
(課題番号 14580392)の支援による.

[文献]

[1] 本吉正博，三浦孝夫，塩谷勇.: "時系列データからの時制クラスの発見", 情報処理学会論文誌:データベース, Vol44 SIG12 (TOD19), pp.44-50 (2003).
[2] Motoyoshi,M., Miura,T., Shioya,I.: "Clustering by Regression Analysis", proc. *Conf. on Data Warehousing and Knowledge Discovery* (DaWaK), pp.202-211 (2003).
[3] Jain, A.K., Murty, M.N. and Flynn, P.J.: "Data Clutering -- A Review", ACM Computing Surveys, Vol. 31-3, pp.264-323 (1999).
[4] J. Han and M. Kamber, Data Mining - Concepts and Techniques -, Morgan Kaufmann (2000).
[5] 日本気象協会編: 気象データひまわり，丸善 (1998).

本吉 正博　Masahiro MOTOYOSHI
法政大学大学院工学研究科電気工学専攻修士課程在学中.
三浦 孝夫　Takao MIURA
法政大学工学部情報電気電子工学科教授.
塩谷 勇　Isamu SHIOYA
産能大学経営情報学部情報学科教授.