光学 28, 6 (1999) 329-336

研究論文 Received January 11, 1999; Accepted April 8, 1999

大量のデータからの知識抽出を目的とした項目の階層的クラスタリング

西澤 秀和・小尾 高史・山口 雅浩・大山 永昭

東京工業大学像情報工学研究施設 〒226-8503 横浜市緑区長津田町 4259

Hierarchical Clustering of Items for Extraction of Knowledge from a Large Amount of Data

Hidekazu NISHIZAWA, Takashi OBI, Masahiro YAMAGUCHI and Nagaaki OHYAMA

Imaging Science and Engineering Lab., Tokyo Institute of Technology, 4259, Nagatsuta, Midori-ku, Yokohama 226-8503

The information systems introduced in the medical field make it possible to accumulate a large amount of data. Analysis of such data will give useful information for medicine. Especially analysis of health screening data is expected to give new information for health care and may lead to preventive medicine. The health screening data consist of many examination values called items. Investigation of relation between items is important to extract new information. In this paper, we propose a method to classify data into some groups which are formed by strongly related items. The proposed method considers that data is generated from populations and has a hierarchical structure. The likelihood based criterion is defined to estimate populations. We applied our method to both computer-generated data and practical health screening data, and the results show the effectiveness of the proposed method.

1. はじめに

近年の保健医療分野における情報化の進展により，大量のデータを蓄積することが可能になりつつある．このようなデータを医療機関や検診機関の間で相互利用することは，新しい疾病や健康に対する診断学の構築に貢献すると予想される[1]．そのためには，これら蓄積されたデータを構成する身体計測，血圧，尿，血液などの多数の検査項目と，疾病等の要因との相関関係を明らかにすることが重要である．

そこで本論文では，大量に蓄積されたデータについて要因と項目との相関関係を明らかにすることを目的として新たな解析手法の開発を行う．このような解析手法が開発できれば，保健医療の分野だけでなく産業や社会科学，学術等の分野にも応用でき，蓄積されたデータの2次利用が進展すると考えられる．

多数の項目からなるデータを解析し，さまざまな要因と項目との関係を調べる手法としては，多変量解析がよく知られている．多変量解析には，主成分分析のように多数ある項目の中から特徴的な成分を見つけ出す方法と，回帰分析のように複数個の項目間の関係を多項式で表す方法がある．しかしながら主成分分析は，データが多次元空間上で正規分布に従う密度分布をもつ場合には有効であるが，複雑な形状の密度分布をもつデータへの適用では，必ずしも特徴的な成分が得られるとは限らない．また回帰分析は，比較的少数の項目間の関係を表すには有効な手法であるが，データを構成する項目数が膨大になると，解が安定しない場合があることが指摘されている[2]．このため，保健医療データ等のように多数の項目からなり，複雑な密度分布をもつデータの解析では，特定の要因に注目し，正規分布で近似可能なデータのみを取り出して解析を行ったり，あるいは特定の項目に注目し，少数の項目に対してのみ解析が行われている．そのためさまざまな要因と項目の組み合わせの中から試行錯誤的に調査しなければならないのが現状であり，多くの要因と項目との関係を処理するには限界がある．

この問題を解決する方法として，クラスタリングの手法

E-mail: yama_off@isl.titech.ac.jp

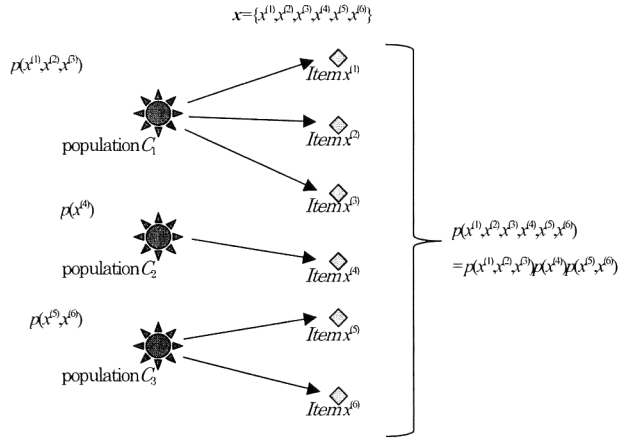

Fig. 1. Ideal image of data generation model. Three populations are assumed as source of items.

を用いてデータの分類を行い，得られたクラスターと要因や項目との関係を調べる方法がある．この手法では特定の要因に注目することなしに，特徴的な要因を自動的に抽出することができる．しかし，一般に用いられるクラスタリング手法[3-5]では，クラスターの個数に対する先験知識を必要とし，またクラスターの密度分布が正規分布で近似可能であるとする条件が必要である．一方，階層的なクラスターリング手法[6]では，データや項目間の類似度を基準にして，2分割によるクラスタリングを階層的に行うことから，クラスターの個数をあらかじめ与える必要がないという特徴をもつ．しかし，類似度としてよく用いられる最近隣距離や最遠隣距離[3]では，検査項目の値の統計的なゆらぎが考慮されていないため，結果がゆらぎの影響を大きく受けてしまうという問題がある．

これまでに筆者らは，複雑な密度分布をもつデータの解析に適した方法として，対数尤度を用いた階層的クラスタリング手法を提案した[7]．この手法はデータの統計的性質に基づく類似度を用いているため，データのゆらぎの影響を受けにくいという特徴がある．またデータの密度分布をParzenの窓関数法[8]により推定するため，クラスターの密度分布に対して正規分布等の仮定を必要とせずに複雑な密度分布をもつデータを分類することが可能となっている．

本論文では，このような対数尤度を用いたデータの階層的クラスタリング手法を基にして，多数の項目からなるデータを統計的に関係をもつ項目のグループに自動的に分類する方法を提案する．まず，複雑な密度分布をもつデータに対しても適用可能とするため，項目間の類似度として密度分布の独立・従属の度合いを対数尤度で評価する関数を定義する．定義した評価関数を基にグループ間で項目の密度分布が独立となるように分類する．これにより，同一グループに属する項目間の関係についてのみ考慮すればよくなるため，効率よく要因と項目との関係を調べることが可能となる．最後に，提案した手法を擬似的に生成したデータおよび健康診断データに適用することで，その有効性を示す．

2. 手　　法

2.1 データ発生のモデル

本論文では項目の分類に，対数尤度を用いた階層的クラスタリング[7]の手法を応用する．文献7）の手法では，まずデータが階層的な構造をもつ複数個の母集団からの無作為標本であるとするモデルを考える．次に対数尤度で与えられる評価関数を基にデータから母集団の階層構造を推定し，データの階層的なクラスタリングを求める．

ここでは上に述べた手法を基にして，項目の分割に応用するためのモデルを定義する．データが N 個の検査項目からなるとき，これを N 個の変量で表す．n 番目の変量を $x^{(n)}$ で表し，N 個の変量の集合 $\{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\}$ をベクトル $\boldsymbol{x}^{(1,2,\cdots,N)}$ で表すことにする．またデータの発生源として，C 個の母集団を考える．i 番目の母集団からの無作為標本が $\{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\}$ の部分集合 $\boldsymbol{\xi}^{(i)}$ からなる変量で構成されるとする．ここで，$\boldsymbol{\xi}^{(i)}$ に属する変量 $x^{(j)}$ について，$\{x^{(j)}|x^{(j)}\in\boldsymbol{\xi}^{(i)}\}$ かつ $\{x^{(j)}|x^{(j)}\in\boldsymbol{\xi}^{(k)}, i\neq k\}$ であるとする．さらに，i 番目の母集団について，$\boldsymbol{\xi}^{(i)}$ に属する変量の同時密度分布を $p(\boldsymbol{\xi}^{(i)})$ で表すことにする．各母集団から1個ずつ無作為に標本を抽出した結果，全部で N 個の変量からなる1つのデータが得られたとするモデルを考えと，N 個の変量の同時密度分布 $p(x^{(1)}, x^{(2)}, \cdots, x^{(N)})$ は，

$$p(x^{(1)}, x^{(2)}, \cdots, x^{(N)}) = \prod_{i=1}^{C} p(\boldsymbol{\xi}^{(i)}) \qquad (1)$$

と表せ，$p(\boldsymbol{\xi}^{(i)})$ は互いに独立となる．一例として Fig. 1 に3個のデータの発生源（母集団）から6個の検査値が得られた場合のモデルの模式図を示す．

本論文では，このようなモデルに従って発生したデータが M 個得られたとし，この M 個のデータから母集団を推定する．しかしデータに含まれるノイズや密度分布の推定誤差のため，独立な母集団から発生したデータであっても，推定した密度分布が厳密に独立になるとは限らない．そこで本論文では次節で述べるように，推定された密度分布の独立性を評価する関数を定義し，最も独立性が高くなるような母集団を求める．ここで母集団の個数は一般に未知であり，また独立性の高い母集団や，独立性の低い母集団の集合が混在していると考えられる．そこで本論文では

最も独立性の高い母集団の集合を抽出し，得られた母集団の個々の集合をさらに複数の母集団の集合に分割することを繰り返すことにより，階層的な構造をもつ母集団として抽出する．ここでは簡単のため2分木で表せる階層構造を考え，評価関数に基づいて任意の母集団の密度分布を2つの密度分布の積に展開する操作を逐次的に繰り返すことで，階層的に母集団の構造を推定する．

2.2 評価関数の定義

ある階層において，母集団から密度分布 $p(x^{(1)}, \cdots, x^{(N)})$ に従う M 個のデータ $X^{(1,2,\cdots,N)}=\{\boldsymbol{x}_1^{(1,2,\cdots,N)}, \boldsymbol{x}_2^{(1,2,\cdots,N)}, \cdots \boldsymbol{x}_M^{(1,2,\cdots,N)}\}$ が得られたとき，母集団を2つのグループに分割することを考える．ここでは，密度分布 $p(x^{(1)}, \cdots, x^{(N)})$ に従う N 個の変量からなる標本を1つのデータとして考え，標本を M 個無作為に抽出した結果，M 個のデータが得られたとする．このとき，$X^{(1,2,\cdots,n)}$ の同時密度分布関数は

$$f(X^{(1,2,\cdots,N)}|p^{(1,2,\cdots,N)})=\prod_{m=1}^{M} p(x_m^{(1)}, x_m^{(2)}, \cdots, x_m^{(N)}) \quad (2)$$

である．ここで，$x_m^{(i)}$ は m 番目のデータについての i 番目の変量である．$X^{(1,2,\cdots,N)}$ を与えられたものとして固定し，$p(x^{(1)}, \cdots, x^{(N)})$ の推定 $\hat{p}(x^{(1)}, \cdots, x^{(N)})$ の関数として対数尤度 l を定義すると，

$$l(X^{(1,2,\cdots,N)}|\hat{p}^{(1,2,\cdots,N)})=\sum_{m=1}^{M}\log\hat{p}(x_m^{(1)}, x_m^{(2)}, \cdots, x_m^{(N)}) \quad (3)$$

となる．式(3)は推定した密度分布 $\hat{p}(x^{(1)}, x^{(2)}, \cdots, x^{(N)})$ が真の密度分布 $p(x^{(1)}, \cdots, x^{(N)})$ にどれだけ近いかを表している．

次に，N 個の変量を2つの集合 $\xi^{(1)}$，$\xi^{(2)}$ に分割したとき，密度分布 $p(x^{(1)}, \cdots, x^{(N)})$ が2つの密度分布，$p(\xi^{(1)})$，$p(\xi^{(2)})$ の積に展開できると仮定する．$p(\xi^{(1)})$，$p(\xi^{(2)})$ の推定をそれぞれ，$\hat{p}(\xi^{(1)})$，$\hat{p}(\xi^{(2)})$ とすると，

$$\hat{p}(x^{(1)}, x^{(2)}, \cdots, x^{(N)})=\hat{p}(\xi^{(1)})\times\hat{p}(\xi^{(2)})\times\Delta(x^{(1)}, x^{(2)}, \cdots, x^{(N)}) \quad (4)$$

と表せる．ここで，Δ は密度分布の推定誤差によって決まる関数であり，推定誤差がない場合には $\Delta=1$ となる．式(4)を式(3)に代入すると，対数尤度は2つの集合 $\xi^{(1)}$，$\xi^{(2)}$ とその密度分布 $\hat{p}^{(1)}$，$\hat{p}^{(2)}$ の関数として，

$$l(X^{(1,2,\cdots,N)}|\hat{p}^{(1)}, \hat{p}^{(2)}, \Delta, \xi^{(1)}, \xi^{(2)})=\sum_m\log\hat{p}(\xi^{(1)})+\sum_m\log\hat{p}(\xi^{(2)})+\sum_m\log\Delta(x_m^{(1)}, \cdots, x_m^{(N)}) \quad (5)$$

と書ける．ここで，密度分布の推定誤差がなく，$\Delta=1$ の場合には，式(5)の第3項は0となることから，第3項を用いて推定誤差の程度を評価することができる．ここで，$l(X^{(1,2,\cdots,N)}|\hat{p}^{(1)}, \hat{p}^{(2)}, \Delta, \xi^{(1)}, \xi^{(2)})=l(X^{(1,2,\cdots,N)}|\hat{p}^{(1,2,\cdots,N)})=\Sigma\log\hat{p}(x_m^{(1)}, x_m^{(2)}, \cdots, x_m^{(N)})$ であることから，推定誤差を最小にするような変量の集合 $\xi^{(1)}$，$\xi^{(2)}$ およびその密度分布 $\hat{p}^{(1)}$，$\hat{p}^{(2)}$ を推定するための評価関数 L として

$$L=\sum_m\log\Delta=\sum_m\log\hat{p}(x_m^{(1)}, \cdots, x_m^{(N)})-\sum_m\log\hat{p}(\xi^{(1)})-\sum_m\log\hat{p}(\xi^{(2)}) \quad (6)$$

を定義する．密度分布の推定誤差は $\hat{p}^{(1)}$ に従う母集団と，$\hat{p}^{(2)}$ に従う母集団が互いに独立であるとき小さくなり，これらが互いに高い従属性をもつとき大きな値となる．そこで，任意の母集団を，評価関数 L が最小となるような2つの密度分布の積に展開することで，階層的な母集団の構造を求めることにする．これにより，従属性が高い変量の集合に分類することが可能であり，結果として項目間の従属性に基づいた分類が得られる．

2.3 密度分布の推定

式(6)の評価関数を求めるには，密度分布を標本から推定する必要がある．ここでは，一般的な場合を考え，母集団から N 個の変量からなる M 個のデータの集合 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_M\}$ が得られたとき，母集団の密度分布を Parzen の窓関数を用いて推定する方法を述べる．

M 個のデータが密度分布 $p(\boldsymbol{x})$ をもつ母集団からの無作為標本であるとすると，密度分布の推定 $\hat{p}(\boldsymbol{x})$ は Parzen の窓関数[8]を用いて，

$$\hat{p}(\boldsymbol{x})=\frac{1}{M}\sum_{m=1}^{M}W(\boldsymbol{x}-\boldsymbol{x}_m) \quad (7)$$

と表せる．ここで，$W(\boldsymbol{x}-\boldsymbol{x}_m)$ は Parzen の窓関数を表し，本論文では，N 次元正規分布

$$W(\boldsymbol{x}-\boldsymbol{x}_m)=\frac{\sqrt{|A|}}{(\sqrt{2\pi})^N}\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{x}_m)^{\mathrm{T}}A(\boldsymbol{x}-\boldsymbol{x}_m)}{2}\right) \quad (8)$$

を用いることにする．ここで A は $N\times N$ の正定値実対称行列で $|A|$ は行列式である．このとき，A^{-1} は正規分布の共分散行列を表し，各変量のばらつきによって決められるパラメーターとなる．各変量のばらつきの影響が互いに独立と考えると，行列の対角成分以外は0となり，式(8)は

$$W(\boldsymbol{x}-\boldsymbol{x}_m)=\frac{1}{(\sqrt{2\pi})^N\sigma_1\sigma_2\cdots\sigma_N}\exp\left(-\frac{(x_1-x_{m,1})^2}{2\sigma_1^2}-\frac{(x_2-x_{m,2})^2}{2\sigma_2^2}\cdots-\frac{(x_N-x_{mN})^2}{2\sigma_N^2}\right) \quad (9)$$

となる．σ_n は A^{-1} の n 番目の対角成分であり，n 番目の変量の分散に依存するパラメーターである．

2.4 窓関数の分散と評価関数の関係

ここでは分散 σ_n と評価関数との関係について考察する．評価関数は項目間の独立の度合いと窓関数の分散の両方に依存する関数であることから，項目のグループが完全に従属である場合を想定することにより，窓関数の分散のみの

Table 1. The value of the σ_i for window function obtained by proposed method. The V_i shows the variance of data.

	V_i	σ_i
Item 1	10.806	1.0000
Item 2	6.9633	0.5035
Item 3	1.1096	0.1037
Item 4	122.09	10.309

影響を調べる（完全に独立である場合には値は0となる）．以下，簡単のため N 個の変数を $x^{(j)}$ と $\xi_j|x^{(j)}\in\xi_j$ の集合に分割した場合を考える．$x^{(j)}$, ξ_j が完全に従属な場合として，ξ_j が定まれば一意に $x^{(j)}$ が定まる場合，すなわち条件付密度分布 $p(x^{(j)}|\xi_j)$ が $\delta(x^{(j)}-f(\xi_j))$ となる場合を考える．ここで f は ξ_j の任意の関数であり，$\delta(x)$ は Dirac のデルタ関数である．$p(x^{(j)}|\xi_j)=p(x^{(j)},\ \xi_j)/p(\xi_j)$ であることから，これを式（6）に代入すると，

$$L=\sum_m \log\hat{p}(x^{(j)}|\xi_j)-\sum_m \log\hat{p}(x^{(j)}) \tag{10}$$

となる．

ここで，密度分布は Parzen の窓関数により推定することから，窓関数の影響を考慮して $\hat{p}(x^{(j)}|\xi_j)$ を求める．Parzen の窓関数による推定では，密度分布 $p(x)$ に従う標本から推定される密度分布の期待値 $E\{\hat{p}(x)\}$ は，窓関数と $p(x)$ のコンボリューションで求まり[8]，

$$E\{\hat{p}(x)\}=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x-\tau)^2}{2\sigma^2}\right)\times\delta(\tau-f(y))\,\mathrm{d}\tau=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x-f(y))^2}{2\sigma^2}\right) \tag{11}$$

である．よって $\delta(x^{(j)}-f(\xi_j))$ に従う母集団からの M 個のデータから推定した密度分布は

$$E\{\hat{p}(x_m^{(j)}|\xi_j)\}=\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left(\frac{-(x_m^{(j)}-f(\xi_j))^2}{2\sigma_j^2}\right) \tag{12}$$

となるが，$p(x^{(j)}|\xi_j)=\delta(x^{(j)}-f(\xi_j))$ であることから，M 個のデータについて $x_m^{(j)}=f(\xi_j)$ が成り立つ．よってこれを式（12）に代入し，$E\{\hat{p}(x_m^{(j)}|\xi_j)\}=1/\sqrt{2\pi}\sigma_j$ となる．これを式（10）に代入すると，

$$\begin{aligned}E\{L\}&=\sum_m\log\frac{1}{\sqrt{2\pi}\sigma_j}-\sum_m\log\hat{p}(x_m^{(j)})=\sum_m log\frac{1}{\sqrt{2\pi}\sigma_j}\\&\quad-\sum_{m=1}^{M}\log\frac{1}{M}\sum_{l=1}^{M}\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left\{\frac{-(x_m^{(j)}-x_l^{(j)})^2}{2\sigma_j^2}\right\}\\&=M\log M-\sum_{m=1}^{M}\log\sum_{l=1}^{M}\exp\left\{\frac{-(x_m^{(j)}-x_l^{(j)})^2}{2\sigma_j^2}\right\}\end{aligned} \tag{13}$$

となる．

式（13）は $x^{(j)}$ と分散 σ_j のみに依存する関数となっていることから，式（13）を用いることでデータのスケールに合わせて分散 σ_j を最適に定められる可能性がある．例えば，定数 L_{const} に対して，任意の項目 $x^{(n)}$ について，

$$M\log M-\sum_{m=1}^{M}\log\sum_{l=1}^{M}\exp\left\{\frac{-(x_m^{(n)}-x_l^{(n)})^2}{2\sigma_n^2}\right\}=L_{\mathrm{const}} \tag{14}$$

となるように σ_n を定める等の方法が考えられる．ここで L_{const} の決定方法や，理論的な妥当性の検討は今後の課題であるが，以降の章で述べるように，擬似データや健康診断データに対する適用結果においては，式（14）によるデータのスケールの補正方法の効果を確認している．

Table 2. The simulation result. The value of cost function giving by eq. (8), (a) between two items for each possible pairs, (b) after merging item 1 2.

(a)

Combination of items	Value of cost function
Item 1-Item 2	495.9
Item 1-Item 3	162.3
Item 1-Item 4	160.9
Item 2-Item 3	162.4
Item 2-Item 4	161.2
Item 3-Item 4	435.2

(b)

Combination of items	Value of cost function
Group A-Item 3	162.3
Group A-Item 4	161.1
Item 3-Item 4	435.2

3. 擬似データに対する適用結果

本章では，提案する手法を擬似的に生成したデータに適用した結果を示す．4個の検査項目 Item 1，Item 2，Item 3，Item 4 からなる4次元データ $\boldsymbol{x}=(x^{(1)},\ x^{(2)},\ x^{(3)},\ x^{(4)})$ を考える．ここで，Item 1 と Item 2，Item 3 と Item 4 はそれぞれ従属で，データの密度分布は以下に従うとする．

$$p(x^{(1)},x^{(2)},x^{(3)},x^{(4)})=p(x^{(1)})\,p(x^{(2)}|x^{(1)})\,p(x^{(3)})\,p(x^{(4)}|x^{(3)}) \tag{15}$$

さらに

$$\begin{aligned}p(x^{(1)})&=\frac{1}{\sqrt{2\pi\times10}}\exp\left(-\frac{(x^{(1)})^2}{2\times10^2}\right),\\p(x^{(2)}|x^{(1)})&=\frac{1}{\sqrt{2\pi\times4}}\exp\left(-\frac{(x^{(2)}-x^{(1)}\times5/10)^2}{2\times4^2}\right)\\p(x^{(3)})&=\frac{1}{\sqrt{2\pi\times1}}\exp\left(-\frac{(x^{(3)})^2}{2\times1^2}\right),\\p(x^{(4)}|x^{(3)})&=\frac{1}{\sqrt{2\pi\times50}}\exp\left(-\frac{(x^{(4)}-x^{(3)}\times100/1)^2}{2\times50^2}\right)\end{aligned} \tag{16}$$

とする．

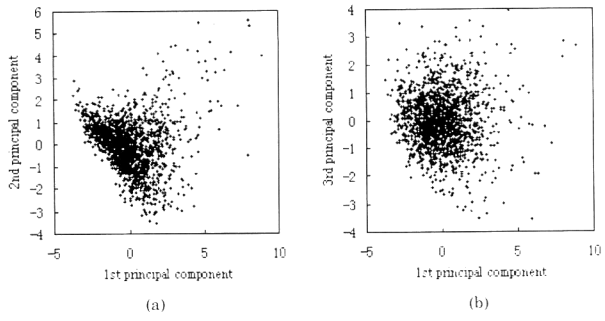

Fig. 2. Results of the principal component analysis, (a) the distribution of health screening data projected on two dimensional plane using first and second components, (b) projection of first and third components.

Table 3. The principal components of health screening data. Table is given up to third order components.

	1st	2nd	3rd
TG	0.340	−0.046	−0.502
SBP	0.350	−0.517	0.275
HDLC	−0.210	0.152	0.714
DBP	0.355	−0.503	0.300
GGT	0.378	0.228	0.078
HbA1c	0.136	−0.195	−0.089
UA	0.272	−0.033	−0.071
GPT	0.424	0.413	0.070
GOT	0.417	0.436	0.216

Table 4. Table shows variances, average of health screening data, and σ_i for window function.

	Average	Variance	σ_i
TG	138	100	162
SBP	125	17	31
HDLC	54	14	25
DBP	77	11	21
GGT	36	39	59
HbA1c	51	7	12
UA	56	12	22
GPT	25	20	28
GOT	23	10	15

式(16)の密度分布に従う4項目からなるデータを200個発生させた．また，発生させたデータから求めた各項目の分散をTable 1に示す．はじめに式（14）を適用し，各項目のスケールの違いを補正した．$x^{(1)}$に対する窓関数の分散を1とし，これを基準としてニュートン法により数値計算により窓関数を求めた結果をTable 1に示す．この表から，各項目の分散にほぼ比例した窓関数の分散が得られていることが確認できる．

次に，式（6）の評価関数に従い項目の階層的な分類を行った．実際の計算では，2つの項目のグループを統合したときの評価関数が最大となるような項目の集合を順に結合していくことで，階層の下のほうからツリーを求める方法を用いた．Table 2（a）は，項目の組み合わせについての評価関数の値であり，Item 1とItem 2の組み合わせが最も従属性が高いことが確認できる．次に，Item 1とItem 2を結合してグループAとし，グループA，Item 3，Item 4の3個の項目のグループについて，各組み合わせについての評価関数の値をTable 2 (b) に示す．表から，Item 3とItem 4の従属性が最も高いことが確認できる．そこで，Item 3とItem 4を結合してグループBとする．この結果，最終的に得られるツリーでは，第1階層でItem 1とItem 2およびItem 3とItem 4の2つのグループA，Bに分類され，第2階層でそれぞれのグループがさらに個々の項目に分類されることになり，もとの密度分布の式（17）に合致した分類結果が得られる．以上の結果から，仮定した密度分布に基づいて生成したデータに対し，項目間の独立・従属性に基づく分類が可能であることが示せた．

4. 健康診断のデータに対する適用結果

次に，提案する手法を1985年4月から1997年春までに得られた実際の健康診断データに適用した例について述べる．解析に用いた検査項目は，最大血圧（systolic blood pressure: SBP），最小血圧（diastolic blood pressure: DBP），中性脂肪(triglyceride: TG)，HDL-コレステロール（high density lipopropein cholesterol: HDLC），glutamic oxaloacetic transaminase（GOT），glutamic pyruvic transaminase（GPT），γ-glutamy transaminase (GGT)，グリコヘモグロビン(glycohemoglobin: HbA1c)，尿酸（uric acid: UA）の9項目である．全データのうち，9項目すべてについて検査結果がある人を対象とし，また同一人物で検診日の異なるものは，別々の標本として取り扱った．データ数は10,000以上になるが，計算の都合上，無作為に抽出した1,589個の標本に対して解析を行った．内訳は健康な人の標本が1,250を占め，残りは疾患をもつ人の標本であった．

はじめにこのデータに対し，主成分分析を適用した結果を示す．スケールの補正には，従来法として平均0，分散1になるように各項目に対して標準化を行う手法を用いた．主成分分析により得られた第1基底，第2基底にデータを投影した図，および，第1・第3基底に投影した図をFig. 2に示す．また各基底の成分をTable 3に示す．これらの結果から，データに直接主成分分析を適用しても，項目間の関係等の情報が得られないことがわかる．

次にこのデータに対して本手法を適用し，項目の分類を行った．まず，2.4節に示した方法を用いて窓関数の分散を求めた結果をTable 4に示す．表から，各検査項目の分

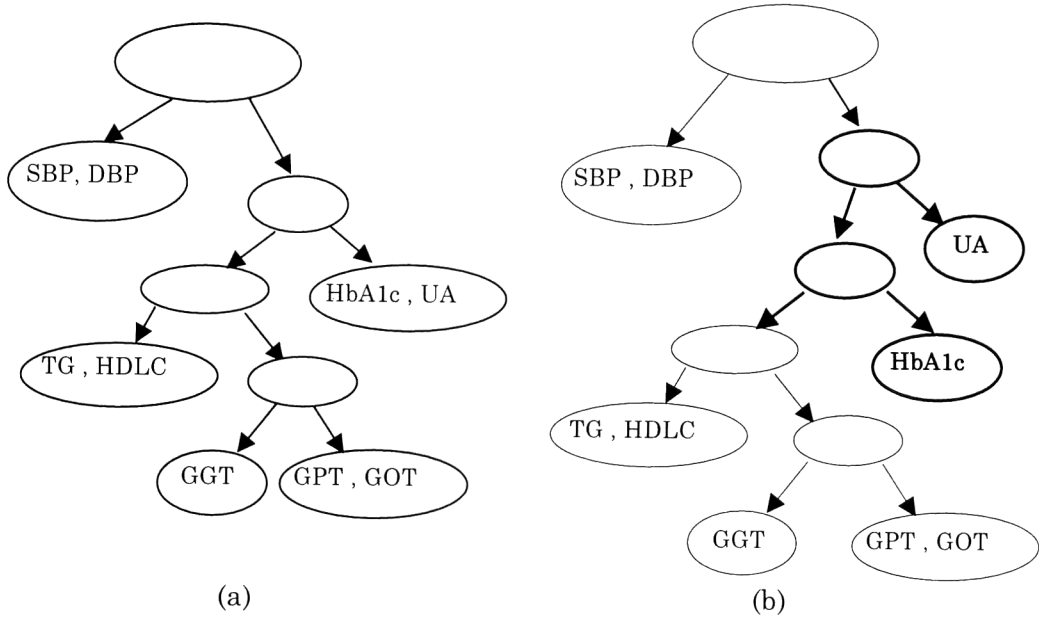

Fig. 3. Results of the classification of health screening data, (a) including healthy and diseased persons, (b) including only healthy persons.

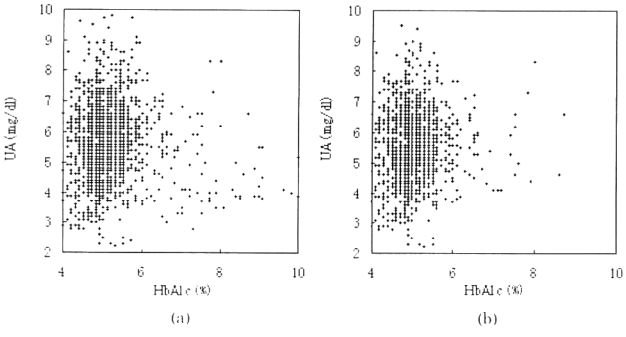

Fig. 4. Plots of items HbA1c and UA, (a) the distribution of samples, which contains healthy diseased persons, (b) the distribution of samples, which contains only healthy persons.

散にほぼ比例した窓の幅が求められていることが確認できる．次に項目の分類により得られた階層を Fig. 3 に示す．疾病をもつ人のデータと健康な人のデータで，項目間の関係が異なる可能性があることから，1,589 個の標本すべてに対して行った分類結果を Fig. 3 (a) に，健康な人の標本に対してのみ行った分類結果を Fig. 3 (b) に示した．Fig. 3 (a) から，SBP と DBP，TG と HDLC，GPT と GOT，HbA1c と UA それぞれのペアは従属性の高いグループとして分類されていることがわかる．また Fig. 3 (b) から，健康者のみの標本では HbA1c と UA の間に従属性がないことがわかる．

HbA1c と UA における標本の分布の様子は Fig. 4 に示すようになっている．健康な人のみの標本では，Fig. 4(b) に示すようにほぼ 2 次元正規分布をしているのに対し，疾

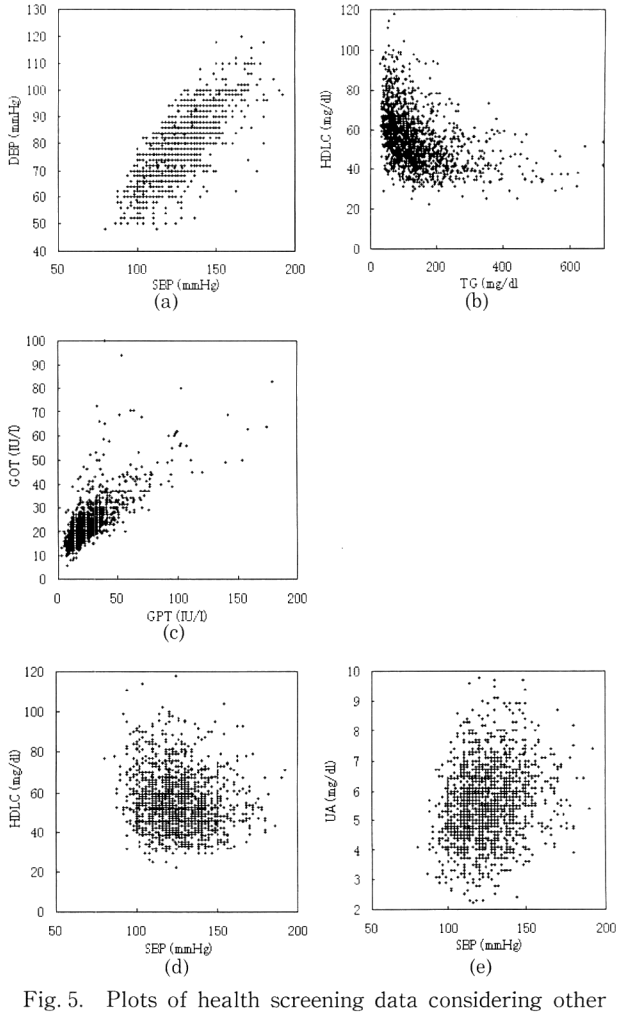

Fig. 5. Plots of health screening data considering other items, (a) for SBP and DBP, (b) for TG and HDLC, (c) for GPT and GOT, (d) for SBP and HDLC, (e) for SBP and UA.

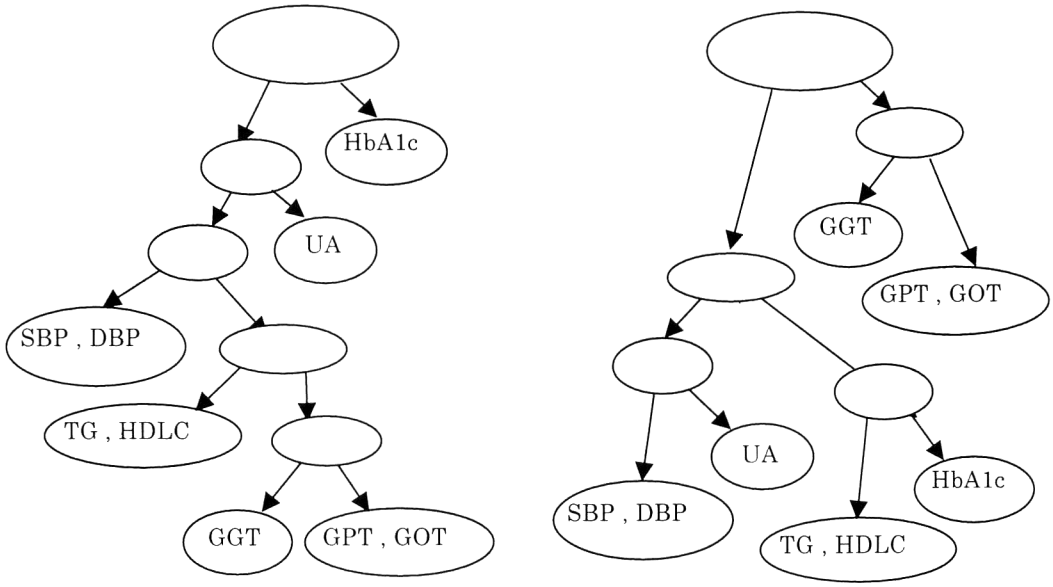

Fig. 6. Results of the classification of health screening data obtained by conventional hierarchical clustering method, (a) using minimum-method, (b) using maximum-method.

患を含む標本では，Fig. 4 (a) に示すようにL字型をした特異な分布をしていることがわかる．このようなHbA1cとUAの間の関係は，主成分分析の結果からは抽出できなかったにもかかわらず，本手法による分類結果では抽出されており，このことから本手法の有効性が確認できる．

他の項目についても，同一グループに属する項目間で実際に従属性が高いことを確認するため，SBPとDBP，TGとHDLC，GPTとGOTそれぞれのペアについての散布図をFig. 5 (a)～(c) に示す．さらに，異なるグループ間のペアとして，SBPとHDLC，SBPとUAそれぞれのペアについての散布図をFig. 5 (d)，(e) に示す．これらの図から，Fig. 3 (a) に示したツリーが項目間の関係を正しく表していることが確認できる．

比較のため，従来の階層的クラスタリング[4]による分類結果をFig. 6に示す．Fig. 6 (a) は最近隣距離を用い，(b) は最遠隣距離を用いた場合の結果を示す．いずれの場合も項目間の距離行列として相関係数を用いている．従来方法の場合，ともにHbA1cとUAが異なるグループに分類されており，本手法で抽出されたHbA1cとUAの関係が抽出されていない．

以上の結果から，本手法を用いることで，健康診断データのような複雑な密度分布をもつデータに対して検査項目間の関係を調べることが可能となることが確認された．

5. ま と め

本論文では，大量に蓄積されたデータの解析を目的として，項目間の関係を調べる方法を提案した．提案した手法は，複雑な形状の密度分布をもつデータを少数の項目からなるグループに分類することで，効率よく項目間の関係を調べられるようにするものであり，具体的には項目間の従属性を統計的な評価関数により評価し，評価関数が最大となるように分類している．そして，提案した手法を擬似的に生成したデータに適用することで，項目の分割の妥当性を確認した．さらに本手法を実際の健康診断データに適用し，主成分分析をそのまま適用した場合には抽出できなかったデータの特徴が，本手法により抽出できることを確認した．

本論文で提案した手法をさまざまな臨床データに適用することで，従来見落とされていた新たな知識が得られる可能性がある．また近年，数値による定量的な診断を目的として，さまざまな生体情報を計測する機器が開発されているが，これら機器によって得られるデータに本手法を適用することで，生体情報と疾病との解明に役立つと期待できる．一方，健康に対する診断学の構築では，得られたデータを基にして個人の健康状態を診断することが重要になる．そのため，今後は，個人差や経年的な変化を考慮した解析手法の開発がより重要な課題になっていくと考えられる．

健康診断データの提供およびその使用を許可してくださ

いました生存科学研究所，日本原子力発電所に感謝いたします．また本研究を進めるにあたり，ご尽力・貴重なご意見をいただきました，バイオコミュニケーションズ株式会社の佐々木敏雄氏に感謝いたします．

文　献

1) 大山永昭："健康社会システムと情報処理技術"，機械振興，**25** (1992) 57-61.
2) 坂本慶行：情報量統計学（共立出版，1989) pp. 138-142.
3) K. Rose, E. Gurewitz and G. Fox: "A deterministic annealing approach to clustering," Pattern Recognit. Lett., **11** (1990) 589-594.
4) K. Rose: "Constrained clustering as an optimization method," IEEE Trans. Pattern Anal. Mach. Intell., **15** (1993) 785-794.
5) X. Lisa and G. Beni: "A validity measure for fuzzy clustering," IEEE Trans. Pattern Anal. Mach. Intell., **13** (1991) 841-847.
6) G. J. McLachlan: "Cluster analysis and related techniques in medical research," Stat. Method Med. Res., **1** (1992) 27-48.
7) H. Nishizawa, T. Obi, M. Yamaguchi and N. Ohyama: "Hierarchical clustering method for extraction of knowledge from data," Opt. Rev., **6** (in press).
8) K. Fukunaga: *Introduction to Statistical Pattern Recognition* (Academic Press, New York, 1972) pp. 166-177.