グラフィクスとCAD 37-4
(1989. 2. 22)

データ解析のための特徴空間の次元縮小の一方法

A method of dimensionality reduction for data structure analysis

森　薫　　市野　学　　矢口博之

Kaoru Mori　　Manabu ICHINO　　Hiroyuki YAGUCHI

東京電機大学大学院　　東京電機大学　　東京電機大学

Graduate School of Tokyo Denki University　　Tokyo Denki University　　Tokyo Denki University

あらまし　データ解析における方法の1つとして、データを構成するサンプル群の相対的構造をできるだけ保存しながら、各サンプルを2次元平面に写像するという方法がいくつか研究されているが、まだその方法は確立されていない。そこで本研究では以下のような非線形の写像法を提案する。それは、データに対して構成された最小全域木を用いてサンプル群をいくつかのクラスターに分け、そのクラスターどうしの相対的位置関係としてデータの大局的構造を平面に表現していく。次に、各クラスターごとに局所的構造を表現していく。この方法においては、写像の結果全体における誤差ができるだけ小さくなるように、サンプルの写像の順序に関して新たな試みを行なった。

Abstract　The purpose of this paper is to present a nonlinear mapping algorithm which gives a kind of trde-off between local structure preservation and global strusture preservation. Our algorithm is composed of three steps; 1)We divide the data set into several clusters by using minimal spanning tree; 2)Then, we find a two-dimensional space which preserves inter-cluster distances; 3)Finally, we embed intra-cluster strctures into the two dimensional space obtained in step2) so as to preserve intersample distances in each cluster as possible.

1. まえがき

データを解析する方法の1つとしてデータ内の幾何学的構造を把握するという方法がある。その幾何学的構造とは，データを構成するサンプル群の相対的位置関係としてとらえられる。データが2つの量的特徴で記述されている場合は，データに対して直接散布図を構成できデータ内の幾何学的構造を把握することができる。しかし，データを記述するために用いた特徴が3個以上の場合，データ内のサンプル間の相対的関係を保存するような何らかの写像が望まれる[1]。

このような方法としてもっとも有名なものは主成分分析（principal component analysis)である。それは，いくつかの点ですぐれているが，以下のような欠点をもつ[1]。

①データを記述する特徴の数が多くなるほど共分散行列の固有ベクトルを見つけるのが困難であるという欠点をもつ。

②非数量的データを扱えない。

1.1 Sammonによる非線形写像

主成分分析は線形写像によるものだが，それに対して非線形写像による方法もいくつか研究されている。非線形写像による方法でもっとも有名なものは，Sammon による方法である[2]。

この方法は，データ空間におけるすべての距離ができるだけ保存されるように，２次元平面上で，各サンプルに対応する点の座標を最適化していくというものである[3]．その際，どちらかというと局所的構造が保存されるように評価関数が構成されている．この方法では，すべてのサンプル間の距離を用いて座標の最適化を行なっているため，誤差がサンプル群全体にわたって平均化されてしまうという欠点がある．また，計算量も当然多い．

また，この計算量を減らすためＳammonの評価関数を速く収束させるアルゴリズムに関する研究が行われているが，それらによるアルゴリズムでは，評価関数の収束そのものあるいは得られる結果に関してまだ問題がある[2]．

１．２　三角法（triangulation method）

Ｓammonはすべてのサンプル間の距離を用いたアルゴリズムを用いているが，それに対してサンプル間の距離の一部を用いる方法がある．そのような方法として，Ｒ．Ｃ．Ｔ．Ｌeeらによる三角法（triangulation method）がある[1]．これは、三角不等式の考え方を応用して、各サンプルについて（ただし、１番めと２番めに写像されるサンプルを除いて）２つずつの距離を保存していくものである。それは，以下の４つのステップからなる．

１）与えられたデータを構成するサンプル群に対して最小全域木（minimal spanning tree,以下ＭＳＴ）を構成する．すると，ＭＳＴ上にはそれぞれのサンプルに関してもっとも近いサンプル，およびそれらの間の距離に関する情報が表される．

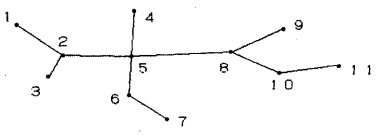


図１　ＭＳＴの例

２）ＭＳＴを有向木とみなし，サンプルを平面に写像する過程を，ＭＳＴを探索する過程とみなしてサンプルを写像する順序を決める．なぜならば，ＭＳＴ上に表現されている距離が正確に平面上で再現されるためである．（各サンプルを平面に写像する際その座標は，それまでに写像されたサンプルに対応する点の座標に依存する．）

３）ＭＳＴ上に辺（edge）として表されている距離は（サンプル数－１）個あるが，それだけでは実際に各サンプルの平面における座標を決めることはできない．そこで，ある１つのサンプルあるいはデータ外の１点を参照点（reference point）として決め，それから各サンプルへの距離を補足情報として利用する．すなわち，各サンプル（ただし１番めと２番めに写像されるサンプルを除く）を平面に写像する際，ＭＳＴ上で隣りにあるサンプルとの間の距離，および参照点との間の距離といった２つの距離が正確に保存される．

また，この参照点との間の距離を用いて各サンプルの平面における座標を決めることにより，サンプルの座標が，ある大局的な制約をうけて決められる．

さらに，参照点を変えることにより，データに関してさまざまな写像を得ることができる．参照点の近くのデータの構造ほど正確に平面上で表現されると考えられる．

４）実際に平面上における点の座標を決める際は２つの円を利用する．ここで，以下のように記号の定義を行う．

＜記号＞元の多次元のデータ空間においてｋ番めのサンプルをP_k^*で表す．また，それに対応して平面上に写像された点をP_kで表す．さらに，元のデータ空間において，P_i^*とP_k^*との間の距離を$d(P_i^*, P_k^*)$で表す．また，写像先の平面上においてP_iとP_kとの間の距離を$d(P_i, P_k)$で表す．

点P_iとP_jが，サンプルP_i^*とP_j^*にそれぞれ対応して平面に写像されているとする．このとき，あるサンプルP_k^*をP_i，P_jの座標に基づいて平面に写像するには以下のようにして行う．なお，P_i^*とP_j^*のうち一方はP_k^*にもっとも近い点で，他方は参照点であるとする．まず，P_iを中心として半径$d(P_i^*, P_k^*)$の円を，次にP_jを中心にして$d(P_j^*, P_k^*)$の円をそれぞれ書く．その２円は２点で交わるかあるいは１点で接する．なぜならば，３つの距離$d(P_i^*, P_k^*)$，$d(P_j^*, P_k^*)$，そして

$d(P_i, P_j)$の間に三角不等式が成立するからである．なぜならば，ここで各サンプルと参照点の間の距離は平面上で正確に保存されているから，

$$d(P_i, P_j) = d(P_i^*, P_j^*)$$

となる．よって，

$$d(P_i^*, P_k^*) + d(P_j^*, P_k^*) \geqq d(P_i, P_j)$$

となるからである．

さて，１点で接する場合はその点がP_kとなる．

また，２点で交わる場合については次のように考えている．その２交点をP_k^1，P_k^2とし，また，すでに写像された平面上の点のうちP_k^1にもっとも近い点をQ^1，P_k^2にもっとも近い点をQ^2とする．このとき，P_k^*をP_k^1に写像したときのQ^1に対する誤差すなわち

$$|d(Q^1, P_k^*) - d(Q^1, P_k^1)|$$

と，P_k^*をP_k^2に写像したときのQ^2に対する誤差すなわち

$$|d(Q^2, P_k^*) - d(Q^2, P_k^2)|$$

とを比較し，小さい誤差を与えるほうのP_kに座標を定める（図２）．

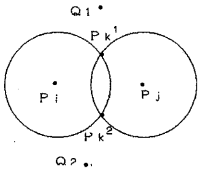


図２　三角法で２円の交点から座標を決める方法

１．３　三角法に対する考察

ＭＳＴはデータのもつ情報を圧縮したものだと考えることができる[4]．この三角法はそのような意味でＭＳＴを用いたことが注目に値する．ただし，ＭＳＴはデータのもつ局所的構造を表すことはできても，大局的構造を表すことはできない．この三角法では，ＭＳＴ上に辺として表されている距離は平面上で正確に保たれるので，局所的構造はある程度保存されるが，大局的構造についてみれば参照点を用いただけでは十分とはいえない．

１．４　本研究の目的

本研究では，多次元データの各サンプルを２次元平面に写像する際，従来の方法と比べて局所的構造と大局的構造の両者を表現する際のトレードオフにすぐれた写像法の研究を目的とする．

2．今回提案する写像法の概要

本研究では，ＭＳＴの利用のしかたを見直し，ＭＳＴを１種のクラスタリングに利用することにする．その結果得られるクラスターどうしの相対的位置関係としてデータの大局的構造をとらえる．また，データの局所的構造に関しては各クラスターごとに表現していく．また，データが本来もっている大局的構造と局所的構造の両者を平面上で表現する際のトレードオフをはかるため，また，結果における誤差をできるだけ少なくするため，サンプルを平面に写像する順序を工夫した．

今回提案する写像法のおおまかな手順を以下に示す．

①与えられたデータを構成するサンプル群に対して，各サンプル間の距離を計算する．
②距離を辺の重みとしてＭＳＴをつくる．
③ＭＳＴ上の辺のうち，ある大きさ以上の辺を切り離すことによって，サンプル群をおおまかにクラスターに分ける（ＭＳＴ上の辺の大きさに関するヒストグラムをつくり，それを見ながらしきい値を入力する，そのしきい値以上の大きさの辺は切り離す）．
④各クラスターにおいて，もっとも分布の中心にあるとみられるサンプルを抽出する．
⑤クラスターを写像する順序を，それに含まれるサンプル数の多い順とする．そして，その順に，④で求めた各クラスターの中心的サンプルを平面に写像する際の２つの"参照点"（６章で述べる）を求める．また，それら２つの参照点とその中心的サンプルとの間の距離をそれぞれ求める．
⑥各クラスター内部のサンプル（ただしそのクラスターの中心的サンプルは除く）を写像していく順序を決める．その順序は，中心的サンプルから近い順とする．次に，それぞれのサンプルに関して，２つの参照点およびそれらとの間の距離を求める．
⑦⑤と⑥で求められたサンプルを写像する順ならびに参照点にしたがって，各サンプルの２次元平面における座標を計算する．

3. 最小全域木（ＭＳＴ）

ＭＳＴは全域木（spanning tree）の特殊なものである．まず，全域木とは以下の条件を満たすグラフである[1]．

①すべての点はグラフ上にある．

②グラフにはループがない．

ＭＳＴとは，その辺の重みの合計が最小となるグラフのことである．このことより，ＭＳＴを点と点との間の距離を重みとして構成したとすると，ＭＳＴには一般に以下のような特性がある[1]．

①もし，点Ｐｉが点Ｐｊにもっとも近い近隣ならば，ＭＳＴ上ではＰｉはＰｊに接続されている．

②もし，ＰｉとＰｊとの間の辺を切り離し，点の集合Ｓ1とＳ2をつくるならば，Ｓ1に含まれる点とＳ2に含まれる点との間の距離で，ＰｉとＰｊとの間の距離よりも近くなるものはない．

ここでは，この②の特性に着目する．この特性を利用すれば，与えられたデータに対してつくられたＭＳＴ上の辺のうち，ある大きさ以上の辺を切り離すことにより，サンプル群をクラスターに分けることができる（図３）．

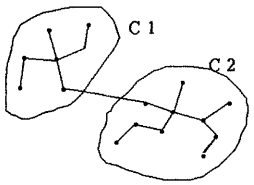


図３　ＭＳＴによるクラスタリング

4. 今回提案する写像法における考え方

データ内の相対的構造をできるだけ保存しながら，データを構成する各サンプルを２次元平面（３次元空間のこともある）に写像することを，次元縮小（dimensionality reduction）という．さて，次元縮小を行う場合，その結果において本来の構造に対してどうしても誤差が生じてしまうと考えられる．特に，元のデータの次元が高い場合ほど誤差が小さい結果は得にくい．そこで問題は，次元縮小を行う際その誤差をどのようにして小さくするかということである．この問題に対して，本方法ではサンプルを平面に写像する順序を工夫することによって対処している．その全体的なレベルでの基本的な考え方は，"大局から局所へ"というものである．

①大局的構造についてはまず，ＭＳＴの辺のうちある大きさ以上の辺を切り離すことにより，サンプル群をいくつかのクラスターに分割する．そして，そのクラスターどうしの位置関係として大局的構造をとらえる（図４）．より具体的には，それぞれのクラスターの中心的なサンプルを求め（その求め方は後述），それらの間の相対的位置関係として大局的構造をとらえる．

本方法の基本的な考え方は，各サンプルを平面に写像する際は，クラスターごとに行なっていくというものである．そこでまず，各クラスターの位置を平面上に決めなければならない．それは，上で求めた各クラスターの中心的サンプルを平面に写像することによる．その際，それらの中心的サンプルを平面に写像する順序が重要である．それは，クラスターに含まれるサンプルが多い順とする．なぜならば，②で詳しくふれるが，写像の結果における誤差を少しでも減らすためである．要するに，実際に大局的構造を平面に表現するには，各クラスターの中心的サンプルを，クラスターを写像する順に平面に写像するのである．

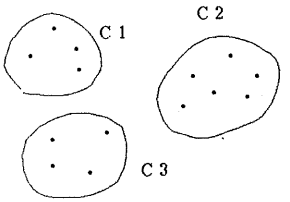


図４　クラスターどうしの位置関係として大局的構造をとらえる

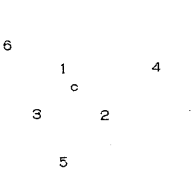
②局所的構造に関しては，各クラスターごとに表現していく．より具体的には，あるクラスターにおいて，大局的構造表現のために平面に先に写像されたそのクラスターの中心的サンプルから近い順に，（そのクラスターに含まれる）その他のサンプルを写像していく．なぜこうするかといえば，サンプルの平面に

おける具体的な座標を決める際は，（７章で詳しく述べるが）それまでに平面に写像されたサンプルのうち２つの点（それぞれサンプルに対応する）の座標が情報として必要となるからである．

一方，前にも述べたように，次元縮小のための写像法においてはどうしても本来の構造に対する誤差が問題となる．誤差とはより詳しくいえば平面において，あるサンプルの座標の，他のサンプルの座標に対する相対的な"ずれ"を意味する．本方法においては特に，後のほうで写像されるサンプルほど誤差を含んでしまう．なぜならば，後のほうで写像されるサンプルの座標は，それより前に写像されたサンプル群の座標に対して決められるので，前に写像されたサンプルの座標がもつ誤差の影響を受けてしまうからである．すなわち，ここでの誤差はサンプルが写像されていくにつれて蓄積されていくものなのである．

さてそれでは，そういった誤差が蓄積される場合，どのように蓄積されれば写像全体の結果に与える影響がもっとも少なくてすむだろうか．それには，まず元のデータにおいてサンプルの粗密を考える．すなわち，後のほうで写像されるサンプルの座標ほど誤差が蓄積されるのだから，サンプルの分布が密な部分にあるサンプルをより先に写像すればよいと考えられる．なぜならば上で述べた理由より，後のほうで写像されるサンプルの座標ほど誤差を含んだものとなる．そこで，より局所的構造が重要となる，分布が密な部分にあるサンプル群を先に写像すれば誤差の影響を最小限に押さえることができると考えられるからである．これによって，誤差は分布が疎な部分に相対的に集まる．以上のことより，実際には写像全体の誤差ができるだけ少なくなるように，各クラスターにおいて分布の中心から外側へ写像していく．まとめると，局所的構造を平面上で表現する際の考え方は"密から疎へ"である（図５）．

また，上の①でクラスターを写像する順序はその要素数の順だと述べたが，これも同じで，写像の結果における誤差をできるだけ減らそうという考え方に基づく．まず，サンプル数の多いクラスターの中心サンプルは，その座標の誤差ができるだけ少ないことが望まれる．なぜならば，あるクラスターにおいて，中心サンプル以外のサンプルの座標は中心サンプルの座標に基づいて決められるので，その中心サンプルの誤差が大きいとそのクラスターの他のサンプルの座標の誤差も大きくなってしまうからである．またすでに述べたように，後のほうで写像されるサンプルの座標ほど相対的に誤差を含んでしまう．そこで，全体の誤差をできるだけ少なくするため，クラスターの中心サンプルは，クラスターの要素数の順に写像していくのである．

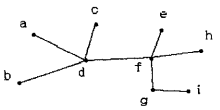


左図でｃはクラスターの中心サンプルを意味する．また，数字はサンプルを写像した順序を意味する．

図５　局所的構造を平面で表現する際の考え方

5. クラスターの中心サンプルの求め方

ここでＭＳＴといった場合はすでに，サンプル群をクラスターに分けるために，ある大きさ以上の辺を切り離したものをさすとする．前述したＭＳＴの特性より，もしサンプルＰ$_i$*がＰ$_j$*にもっとも近いならば，ＭＳＴ上でそれぞれに対応する節点Ｐ$_i$とＰ$_j$は接続されている．ここで，ＭＳＴの各節点（node）の次数（degree）について考えると，もっとも次数が高いものほどそのクラスターにおいて分布の中心に



上図の場合，サンプルｄとｆの次数が４でもっとも高い．そこで、それぞれについて、そのクラスターに属する他のサンプルとの距離の総和を求める．たとえば、ｄについては
d(a,d)+d(b,d)+d(c,d)+
d(d,e)+d(d,f)+d(d,g)+
d(d,h)+d(d,i)　を求める．

図６　クラスターの中心サンプルの求め方

近いと考えることができる．そこで，これをそのクラスターの中心サンプルと呼ぶことにする．

また，もし次数がもっとも高いものが複数あったならば，そのそれぞれに関してそのクラスターの他のサンプルとの間の距離の総和を求め，それがもっとも小さいものを中心サンプルとする（図６）．

6. 参照点 (reference point)

さて，サンプルを写像する順序については，４章でその基本的考え方を述べた．それは，まず各クラスターの中心サンプルを平面に写像し，続いてクラスターごとに中心サンプル以外のサンプルを，中心サンプルに近い順に平面に写像していくというものである．実際に，その点の座標を決めるにはこの場合，ある２つの点とそれからの距離とが必要である．その２つの点を簡単のため参照点と呼ぶことにする（もちろん，この名称はＬｅｅらの論文から借りた）．

なお，実際のアルゴリズムでは，その構成を簡単にするため，２章で述べているように，参照点やそれらとの間の距離を求める部分と，実際に平面上における座標を求める部分とは分けている（アルゴリズムの概要の⑤，⑥と⑦）．

6．1　各クラスターの中心サンプルを写像する場合

まず，最初に写像する中心サンプルをＰ1*とすると，それはｘｙ平面上の原点に写像するものとする．本方法では，（サンプルに対応する）点の相対的な座標が問題なので，特にそうする制約はないが，便宜上そう定める．

また，２番めに写像するクラスターの中心サンプルをP_2*とすると，点P_2の座標は

（０，ｄ（P_1*,P_2*））

と定める．このように，１番めと２番めのクラスターの中心サンプルの座標を決める際は参照点を用いない．

さて，参照点およびそれとの間の距離が必要となるのは，３番め以降に写像するクラスターの中心サンプルおよび，中心サンプル以外のサンプルである．まず，３番めに写像するクラスターの中心サンプルをP_3とする．P_3の座標は，それまでに写像された２つの点P_1とP_2の座標に基づいて決める（座標の決め方については７章で述べる）．

４番め以降のクラスターの中心サンプルについては，次のようにして写像する．これから写像しようとするサンプルをP_kとする．P_kの座標を決めるときは，それまでに平面に写像された点のうち，元のデータにおいてP_kにもっとも近いサンプルに対応するものと，もっとも遠いサンプルに対応するものとの２つを参照点として決める．なぜこのようにするかというと，元のデータと比較して，平面上で（それぞれサンプルに対応する）点の分布が適正に広がることを期待しているからである．

6．2　各クラスターごとにサンプルを写像する場合

これから写像しようというサンプルをP_mとする．この場合は参照点の一方をそのクラスターの中心サンプルに対応する点とする．なぜこうするかというと，各サンプルを平面に写像するときそれぞれに対応する点の座標が，データの大局的構造をある程度反映して決められるようにするためである．

また，もう一方の参照点を，それまでに写像されたサンプルのうち，元のデータにおいてP_mにもっとも近いサンプルに対応する点とする．なぜならば，元のデータにおいてもっとも近いサンプルとの間の距離が，平面上においても保存されるようにするためである．

7. 具体的に平面上での座標を決める方法

本方法においては，サンプルの平面における座標を決める際，Ｌｅｅらが用いた方法と同様に２つの円を用いている．これは，２つの点からそれぞれある距離だけ離れた点の座標を決める方法としてはきわめて自然なものといえよう．さて，Ｌｅｅらの方法では，前述したようにそれらの円は必ず，２点で交わるか，あるいは１点で接する．そして，細部を順次的に写像していくことによって，サンプル全体を写像しようというものである．

一方，本方法はその基本的考え方がＬｅｅらの三角法とは本質的に異なり，細部を順次的に写像していくというのではなく，まず，大局的構造の表現から始めるというものである．その次に各クラスターごとに細部の構造を表現していくのである．４章と６章で述べたことより，各サンプルを平面に写像する際は，平面において

その間の距離が正確に保たれていない2つの点を参照点とすることがある．

すなわち，いま，そのような2つの参照点をP_i，P_jとする．また，これから写像しようとしているサンプルをP_kとする．P_iとP_jに関してみれば，それらの座標はデータ本来の構造と比べて相対的にずれているということになる．本方法では基本的には，P_i，P_jを中心にそれぞれ半径が$d(P_i*, P_k*)$，$d(P_j*, P_k*)$の円を書いてP_kの座標をきめる．しかし，上で述べたことより，その2円が交わったり，あるいは接したりしない場合が考えられる．もちろん，サンプル群を写像していく過程でこのような場合が生じることは少ないほうが望ましい．

さて，一般に2つの円の位置関係には，その中心間の距離，および2つの円の半径の大きさの関係から5通りの場合が考えられる．我々の方法において問題なのは上で述べた，2円が交わらない場合である．5つのそれぞれの場合に関して，より近いサンプル間の距離が尊重されるように，以下のように座標を決める．

1）2円が外側から1点で接する場合

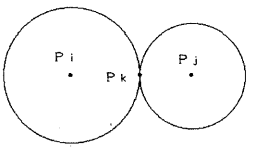


図7　2円が外側から1点で接する場合

2）一方が他方を含む形で接する場合

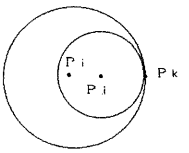


図8　一方が他方を含む形で接する場合

3）2円が2点で交わる場合

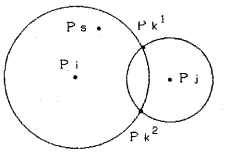


図9　2円が2点で交わる場合

元のデータ空間において，すでに平面に写像されたサンプルのうちP_k*に2番めに近いサンプルをP_s*とする．それに対応する平面上の点P_sと，P_kの候補となるP_k^1およびP_k^2との間の距離，すなわち$d(P_s, P_k^1)$と$d(P_s, P_k^2)$とをそれぞれ求める．また，元のデータ空間において，P_s*とP_k*の間の距離，すなわち$d(P_s*, P_k*)$を求める．次に以下の比較を行う．

$$|d(P_s, P_k^1) - d(P_s*, P_k*)| \leqq |d(P_s, P_k^2) - d(P_s*, P_k*)|$$

が成り立てば，サンプルP_k*をP_k^1に写像する．成り立たなければ，P_k^2に写像する．

4）2円のうち一方が他方を含んでかつ接しない場合

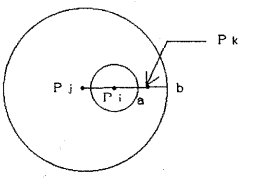


図10　2円のうち一方が他方を含んでしまう場合

上図のように，2つの参照点P_iとP_jを結んだ直線を考える．その直線と2円の円周の交点をどれぞれ上図のようにa，bとする．

さて，P_kはどこに位置させるのがもっとも適当だろうか．それは，線分abを

$$d(P_i*, P_k*) : d(P_j*, P_k*)$$

の比に内分する点がもっとも適当だろうと思われる．なぜならば，できる限りその3つのサンプルの間の相対的関係を忠実に再現したいからである．これは，もちろん写像の結果全体における誤差を減らすためである．

5）2円が離れている場合

この場合も4）の場合と同様に点a，bを定義し同様に考えて，P_kは，線分abを

$$d(P_i*, P_k*) : d(P_j*, P_k*)$$

に内分する点に位置させる．

なお，この4）と5）の場合は生じないほうが望ましい．

8. 例題と考察

以上の議論から明かなように，この方法はサンプル間の距離が求められれば，どんなデータにも適用できる．よって，この方法は距離の種類による影響をうけない．ここでは，市野による"一般化されたミンコースキー距離"[5]を用いている．この距離は以下のような長所をもっている．

①データが量質混在の特徴で記述されていても適用できる．

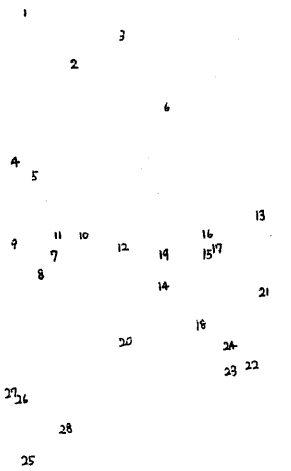
②各特徴のとる値が単に１つの数値や名義的記号（質的特徴の実現値の１つである）ばかりでなく，閉区間や有限集合であってもよい．

また，この距離は０以上１以下の値をとる．

データには１９８８年のトヨタ車のデータを用いた[6]．このデータは以下のような１５種類の量的あるいは質的特徴で記述されている．なお，各サンプルはすべてのグレードに対するものなので，ほとんどの量的特徴は範囲をもって記述されており，また質的特徴は和集合で記述されている：｛全長，全幅，全高，車両重量，エンジン配置，<u>弁配置</u>，総排気量，最高出力，最高出力時回転数，最大トルク，最大トルク時回転数，１０モード燃費，<u>ブレーキ前</u>，<u>ブレーキ後</u>，価格｝（下線のついている特徴は質的特徴である，他は量的特徴である）．

また，ＭＳＴ上のサンプル群をおおまかにクラスターに分けるためのしきい値としては0.15を入力した．その結果サンプル群は７つのクラスターに分かれた．さらに，この方法では２円を用いて点の座標を決めているが，７章で示した４）と５）の場合は生じないほうが誤差少ない写像が得られる．いくつかのしきい値に関してプログラムを動かして，７章に示した４）と５）の手続きの実行回数を調べたところ，しきい値0.15のときがそのような場合は１回（手続き４）でもっとも少なかった．図11に各サンプルを２次元平面に写像した結果を示す．

この結果をみてわかることは，ソアラ，スープラ，クラウン，クレスタといった高級車のクラスターと，ハイエースなどのワゴン車のクラスターがはっきり分かれているということである．また，その他の部分をみてもたとえば，ビスタ，カムリ，コロナ，カリーナＥＤのように似た車どうしが近くに位置しているということがわかる．この結果は我々の感覚に照らしてみて，かなり良好なものといえる．今後の課題は，写像の結果に対する定量的評価法の確立と，もっとも良いしきい値を自動的に発見することである．



1.ソアラ 2.スープラ 3.クラウン 4.マーク2 5.チェイサー 6.クレスタ 7.ビスタ 8.カムリ 9.セリカ 10.コロナ 11.カリーナED 12.カリーナ 13.MR2 14.レビン 15.カローラFX 16.カローラ 17.トレノ 18.スプリンター 19.シエロ 20.カリブ 21.カローラ2 22.ターセル 23.コルサ 24.スターレット 25.ハイエース 26.タウンエース 27.マスターエース 28.ライトエース

図11 トヨタ車のデータに対して次元縮小を行い散布図を書いた例

<参考文献>

[1] R.C.T.Lee, J.R.Slagle and H.Blum, "A triangulation method for the sequential mapping of points from N-space to two-space," IEEE Trans. Computers, C-26, pp.288-292, 1979.

[2] W.Siedlecki, K.Siedlecka and J.Sklansky, "An Overview of mapping techniques for exploratory pattern analysis," Focused Research Program", TP-87-5, 1987.

[3] J.W.Sammon, Jr., "A nonlinear mapping for data structer analysis," IEEE Trans. Computers, C-19, pp.826-829, 1970.

[4] C.T.Zahn, "Graph-theoretical methods for detecting gestalt clusters," IEEE Trans. Computers, C-20, pp.68-86, 1971.

[5] 市野,矢口,"量質混在の記述を許す一般化されたミンコースキー距離,"電子情報通信学会論文誌(A), Feb, 1989.

[6] 八重洲出版, driver 11-20号 -1988