北海道大学工学部研究報告
第119号（昭和59年）

Bulletin of the faculty of Engineering,
Hokkaido University, No. 119 (1984)

k—最近隣距離の分布によるクラスター分析法

佐藤　義治　中西　寛子　河口　至商

（昭和58年9月30日受理）

Clustering Method by the Distribution of the Distance of k-th Nearest Neighbour

Yoshiharu SATO, Hiroko NAKANISHI and Michiaki KAWAGUCHI

(Received September 30, 1983)

Abstract

The purpose of this research work is to obtain the algorithm wich can be used in whether the observed data has clusters or not, and by which clusters can be detected automatically. But it is difficult to obtain a generalized algorithm. Thus, in this paper, the clustering method is offered under a condition in which the clusters are considered to be some clumps inconsistent with randomness, which is represented by a multidimensional poisson set. As the merging criterion, the distribution of the distance of k-th nearest neighbour is used. According to the level of the test of randomness, it is shown that this method has a hierarchical property.

1．まえがき

クラスター分析における結果の評価は，データおよびクラスターの定義（アルゴリズムも含めて）の多様性により，すべて同一の評価基準をもってすることは困難である。したがって種々の条件に即した評価方式が必要となる。これらに関してはすでに数多くの研究が行われており，Hartigan[3),4),5)]，Sokal[10)]，Sneath[9)]，Ohsumi[8)]等により従来の研究の概説および新しい方法論が詳しく著述されている。

クラスター分析を行う場合の前提としてつぎの2つのものが考えられる。(1)クラスターの存在は既知である。(2)クラスターが存在するか否かは未知である。前者の場合の手法や有意性の評価に関する研究は数多く見られるが，後者の場合の研究は少ない。本論文では後者の立場からクラスターを構成するとともにその有意性を評価することを目的とする。

ここでは，分類されるべき個体の集合がm次元ユークリッド空間E^mの有限個の点として表される場合のみを扱う。もしその点集合がE^m中にランダムに分布しているならば，このようなデータにはクラスターは存在しないものとし，クラスターとはランダム性から乖離した部分集合であると定義する。このとき，ランダムであるとは、E^m中の点集合が単純ポアソン点過程[1),2)]に従っている(このような点集合を単にポアソン集合[1)]と呼ぶ)と表現する。クラスターを抽出するために，ポアソン集合におけるランダムな点からk番目に近い点（k—最近隣と呼ぶ；k-th Nearest

情報工学専攻　情報数理工学第一講座

Neighbour)までのユークリッド距離の分布[7),11)]を用い，クラスターの有意性をポアソン集合におけるそのクラスターの生起確率によって評価する。

個体間の距離の分布や密度を考慮したクラスター分析の最近の研究として，Törn[12)]および脇本・山本・垂水[13)]によるものがある。これらの研究では，クラスターのシード点の決定や融合距離の閾値を定めるために一様分布の下における2点間のユーリッド距離（Random Distance[13)]）の分布や平均密度が用いられている。

次節においてポアソン集合におけるランダムな点からk—最近隣までのユークリッド距離の分布について論じ，第3節ではクラスター化のアルゴリズムを提示し，このアルゴリズムにより生成されるクラスターは検定基準に関して階層的であることを示す。最後に適用例について述べる。

2．多次元ポアソン集合における k—最近隣距離の分布

m次元ユークリッド空間E^m上のポアソン集合（Feller[1)]，Fisher[2)]参照）において，正の定数λを単位球に含まれる平均個体数とする。このとき，あるランダムな点(個体をE^m中の点と表現する）からk—最近隣までのユークッド距離をr_kとし，$w_k=(r_k)^m$とすると，w_kの累積分布関数は

$$F_{(k)}(w_k)=1-e^{-\lambda w_k}\{1+(\lambda w_k)/1!+(\lambda w_k)^2/2!+\cdots+(\lambda w_k)^{k-1}/(k-1)!\} \tag{2.1}$$

となる。（Thompson[11)]，Morisita[7)]）したがってその確率密度関数はつぎのように得られる。

$$f_{(k)}(w_k)=(\lambda^k w_k^{k-1}e^{-\lambda w_k})/(k-1)! \tag{2.2}$$

すなわち$F_{(k)}(w_k)$はランダムな点を中心とする半径r_kの球が少なくともk個の点を含む確率と考えることができる。

n個の標本から（2．1）を求めるためにはパラメータλを推定しなければならない。一般にw_kの観測値がn個与えられたとき，それらの算術平均を$\bar{w}_k$とすると

$$\hat{\lambda}=(kn-1)/n\overline{w}_k$$

がλの不偏推定量であることが証明される。（証明は略すが，Moore[6)]の$k=1$の場合と同様）しかしこれは標本値の変動に敏感であり，kの決定も困難である。したがってここではn個の標本を含む最小のm次元立方体の体積Vを求め

$$\hat{\lambda}=nS_m/V, \qquad S_m=\pi^{\frac{m}{2}}/\Gamma(\frac{m}{2}+1) \tag{2.3}$$

によってλを推定する。ここで$\Gamma(\cdot)$はガンマ関数を表わす。

3．クラスター化のアルゴリズム

E^m中にN個の点集合$O=\{O_1, O_2, \cdots, O_n\}$が与えられ，それらの座標が$\{X_{ia}; i=1, 2, \cdots, n, a=1, 2, \cdots, m\}$と表わされているものとする。すなわち$x_{ia}$は点$O_i$の第a座標の値を書わす。この$n$個の点集合に以下のアルゴリズムA1～A7を適用してクラスター化を行う。

A1　点集合Oをポアソン集合と見なし，(2.3) によりλを推定する。

A 2　すべてのO_i, O_j ($i \neq j$) についてユーリッド距離の2乗

$$d_{ij}^2=\sum_{a=1}^{m}(x_{ia}-x_{ja})^2$$

を計算する。

A 3　すべてのi ($i=1, 2, \cdots, n$) についてO_iからO_jを除いたk—最近隣 ($k=1, 2, \cdots, n-1$) にある点を求める。

A 4　点O_iからのk—最近隣をO_jとするとき，O_iを中心として半径d_{ij}の球の内部にk個の点 (O_jは含まない) が存在する確率$P_{ij(k)}$を (2.1) より

$$P_{ij(k)}=F_{(k)}(d_{ij}^m)$$

として，すべてのiとkについて計算する。

A 5　ある適当な確率P_o (0.01とか0.05など) を与え，各点O_iについて番号kの集合

$$K_{io}=\{k: P_{ij(k)} \leqq P_o\}$$

を求め，もし$k_{io} \neq \phi$ならば

$$k_{iO} \equiv \max_{k \in K_{io}} k$$

なるk_{io}を決定する。$K_{io}=\phi$ならば$k_{iO} \equiv 0$　(O_iを弧立点として扱うことになる) とおく。

A 6　n個の点集合上につぎのような関係 (Relation) R_oを定義する。

R_o：「同一のクラスターに属す」

(a) $O_i R_o O_i$：O_iはO_iと同一のクラスターに属す。

(b) $O_i R_o O_j$, $j \leqq k_{io}$：O_iとk_{io}番目以内 (1番目, 2番目, …, k_{io}番目) に近い点はO_iと同一のクラスターに属す。

A 7　関係R_oをもとに点集合Oをクラスター化するために，R_oを含む最小の同植関係R_o^*を求めR_o^*により点集合Oを類別し，クラスターを決定する。

このアルゴリズメを適用する場合，λの推定に関してつぎの2つの点に注意しなければならない。その1つは，ポアソン集合においては重複した点が起る確率は無限小であるとするため，データーに全く同一のものがあるときには1個と見なしてλを推定すること。他の1つは，データの本質的な次元が減少していないかどうかの検討が必要なことである。そのためには，Törn[12]も指摘しているように，主成分分析を行ない，次元が減少しているときには適当な部分空間に射影してλを推定し，その部分空間内でこのアルゴリズムを適用すべきである。

ここではアルゴリズムによって得られるクラスターは，つぎのような意味で階層的である。

「$P_o > P$なる任意のPによって得られるクラスターは，P_oで得られるクラスターに等しいか，またはその細分である。」

(証明)

P_oおよびPに対応して，アルゴリズムのA 5で得られる番号kの集合をそれぞれk_{io}, k_iとすると$P_o < P$よりすべてのiについて

$$k_{iO} \supseteq k_i, (i=1.2, \cdots, n)$$

であるから

$$k_{iO} \equiv \max_{k \in kiO} k \leqq \max_{k \in ki} k \equiv k_i \tag{3.1}$$

が成り立つ。一方 k_{io}，k_iによって得られる関係をそれぞれ R_o，Rと表わすと，任意のO_i，O_jに対して（3.1）より

$$O_i R O_j \Rightarrow O_i R_O O_j$$

となる。したがってR_o，Rを含む最小の同値関係をそれぞれ R_o^*，R^*とすると，また任意のO_i，O_jについて

$$O_i R^* O_j \Rightarrow O_i R_O^* O_j$$

が得られる。すなわちR^*はR_o^*より強い同植関係となり，その類別はR_o^*により類別より細かい，つまり細分となっている。すべての i について(3.1)の等号が成り立つときR^*とR_o^*は同等であるからそれらによる類別は等しい。

（証明終り）

4．適　用　例

図1に示した2次元データ（$n=30$）について，前節のアルゴリズムを適用した結果について述べよう。このデータから λ の推定値として0.112が得られた。最初に $P_o=0.01$ としてクラスター化を行うと，つぎのようなクラスターが得られる。

クラスターA：{11, 12, 13, 14, 15, 16, 17, 18, 19}
クラスターB：{2, 3, 4, 5, 6, 7, 8, 9}
他はすべて弧立点

すなわち，データをポアソン集合（ランダムな点集合）と仮定したとき，クラスターAや，クラスターBが生起する確率は0.01以下であるということができる。図1の左下のクラスターBのそばにある弧立点{1}は$P_o=0.045$でBに融合され

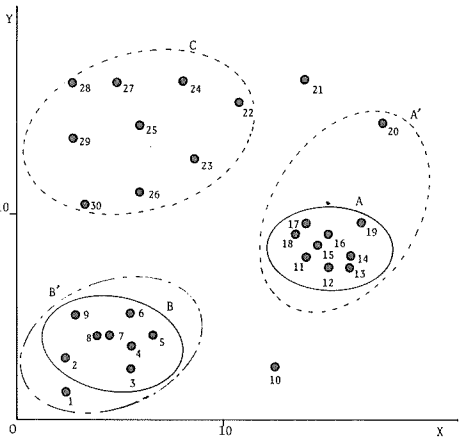
クラスターB′：{1}∪B

を形成する。このときクラスターAと他の弧立点は不変である。一方P_oを0.75まで大きくするとつぎの3つのクラスターと2個の弧立点が得られる。

クラスターA′：{20}∪A
クラスターB′：不変
クラスターC：{22, 23, 24, 25, 26, 27, 28, 29, 30}
弧立点　　：{10}, {21}

これら結果をまとめたものを図1に示してある。

以上から，クラスターの有意性を評価すると，部分集合AおよびB′は〝クラスターである〟ということの危険率が5％以下であるが，部分集合Cをクラスターとする危険率は75％と結論される。



図―1　個体の2次元配置とクラスター化の結果

5．あとがき

2次元平面上に描かれたn個の点の形状を見ながら，互いに近い点を似ているとし，クラスターを抽出するとき，〝クラスターである〟との判断にはその確からしさの程度が含まれていよう。例えば，たぶんこのあたりにクラスターがあるとか，これは確かにクラスターである，ということである。ここで提案したアルゴリズムは，複雑な形をしたクラスターは抽出できないにしても，上で述べた確からしさの程度をより客観的な尺度（ここでは確率）で表現することを目的としたものである。単に2次元上の点集合ならば直観的な判断で十分であるが，一般に多次元のデータを扱う場合には直観（目で見ること）では判断できず，客観的な基準が必要となる。

クラスター分析において，たとえその対象にクラスターの存在が明確な場合でも，観測したデータに関して，それらのクラスターが得られるかどうかを調べようとするときなどにも，ここで述べたアルゴリズムが有効であると思われる。

この手法は，検定基準（P_0）に関して階層的であることを示したが，他の階層的手法との関連や，クラスター数の検定法（例えばHartigan[4]が述でている最短距離法におけるギャップ検定など）との関連は未解決であり，今後の課題として残されている。

引用文献

1）Feller, W. : An Introduction to Probability theory and its Applications, Vol.2, 2nd, ed.,(1971), John Wiley & Sons.
2）Fisher, L. : Stochastic Point Process, (P.A.W.Lewis ed.), (1972), P.468, John Wiley & Sons.
3）Hartigan, J. A. : Clustering Algorithms, (1975), John Willey & Sons.
4）Hartigan, J. A. : Classification and Clustering, (J.Van Ryzin ed.), (1977), p.45, Academic Press.
5）Hartigan, J. A. : Ann. Statist., 6(1978), p.117.
6）Moore, P. G. : Ecology, 35(1954), p. 222.
7）Morishita, M. : Mem. Fac. Sci. Kyushu Univ. Series E, 1(1954), p. 187.
8）Ohsumi, N. : Data Analysis and Informatics, (E.Diday et al. ed.), (1980), p. 509, North-Holland.
9）Sneath, P. H. A. : Data Analysis and Informatics, (E.Diday et al. ed.),(1980), p. 491, North-Holland.
10）Sokal, R. R. : Classification and Clustering, (J.Van Ryzin ed.), (1977),p. 1, Academic Press.
11）Thompson, H. R. : Ecology, 37(1956), p. 391.
12）Törn, A. A. : On Systems, Man and Cybernetics, SMC-7 (1977), p. 610.
13）脇本・山本・垂水：数理科学講究録345（1979），p.1.