情 報 学 基 礎 42－7
（1996．7．26）

考古学データベースからの情報の抽出

宝珍 輝尚　　安達 政伸　　都司 達夫

福井大学 工学部 情報工学科

考古学では、遺跡から発掘された遺物のデータのデータベース化が強く望まれ、そこからの新たな発見の導出や仮説の検証のサポートが望まれている。そこで、考古学の遺物のデータとして、一乗谷朝倉氏遺跡から出土した染付の碗や皿を例に、通常のクラスタ分析を行ったところ、定説とされている分類が得られなかった。これは、定説とされている分類では、一つの項目が複数の分類に出現しているためであった。そこで、このような分類を可能にするように、1回のクラスタ分析で最も類似性の高いクラスタを求め、そのクラスタに属するデータを取り除いて再度クラスタ分析を行うという方法を用いた。これにより、定説に近い分類を得ることが出来た。

Information Discovery from an Archeological Database

Teruhisa Hochin　　Masanobu Adachi　　Tatsuo Tsuji

Dept. of Information Science, Faculy of Eng., Fukui University

Archeologists require to build databases of remains, and to discovery new facts and to test assumptions. The pieces of china dug up from Asakura site have been already categorized by the archeologists. We tried to categorize them through the cluster analysis in order to obtain the same categories as the authorized ones. The ordinary clustering method can not bring us the required categories. This reason is that the authorized categories share the same item. Then, the clustering method is improved to obtain one cluster in one clustering process. This method brings us the categories similar to the authorized one.

1 はじめに

近年、計算機利用分野の拡大に伴い、データベースの応用分野も急速に拡大している。これらの応用分野には、衛星からのイメージデータの管理、コンビニエンスストアの販売品の管理等が含まれている。これらの応用分野では、極めて大量のデータが発生し、それらのデータはデータベースに格納されている。この大量のデータから、価値の高い情報を抽出することが出来れば、データベースの利用価値を更に高めることが出来る。遺跡から発掘された遺物データの管理も上述の応用分野の１つである。遺物データも同様に大量のデータを含んでおり、人手による情報の抽出は困難な作業である。そのため、新たな発見や見識の導出や仮説の検証に計算機を利用することが望まれている。

ここで、遺跡から発掘された陶器の破片からは、陶器全体に関するデータは得られない。例えば、陶器の底のみの破片からはその口縁部に関するデータは得られない。このような、全ての項目に対して値が存在するとは限らないデータ集合に関しては、従来から情報抽出に用いられているクラスタ分析では、うまく分類できないことがある。例えば、染付の碗をクラスタ分析した場合、考古学上明らかになっている分類と異なる項目が数多く見られる。

そこで、本研究では、考古学データベースからの情報の抽出の自動化を目的として、考古学データベースに対して有効な情報抽出法に関する検討を行なう。特に本研究では、全ての項目に値が存在するとは限らない遺物データを対象とし、その自動分類を行なう手法を開発することを目的とする。本論文では、一般的な分類方法であるクラスタ分析を用い、クラスタ分析の際、最も相関の強いクラスタを確定クラスタとし、データ集合全体から確定クラスタを削除し、再度クラスタ分析を行う方法の適用結果について述べる。この方法を漸次クラスタリング法と呼ぶ。一乗谷朝倉氏遺跡から出土した染め付けの皿、碗に対して漸次クラスタリング法を適用した結果、良好な結果が得られることを示す。

以下、次に、2. では、本研究で対象とする考古学データについて述べる。次に、3. では、クラスタリングについて述べる。そして、4. では、漸次クラスタリング法について述べ、5. で、漸次クラスタリング法の適用結果を示す。最後に、6. でまとめを行う。

2 対象とするデータ

本研究で対象とするデータは、福井県一乗谷朝倉氏遺跡から出土した染付の碗と皿である。

各部の名称には、口縁部、胴部、腰部、見込み、高台といった名称が付けられている (図 1)。

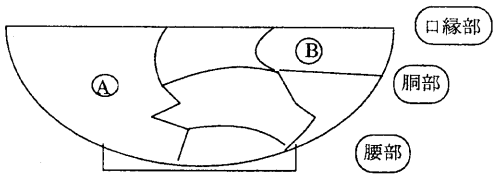

図 1: 皿の部分の名称

例えば図 2の様な破片の場合、破片は碗の底の部分であり、鳥の絵が描かれている。また、図 2からは分からないが、裏側の高台の中に「天下太平」と書かれている。また、重さ、高台の高さ、半径などを測定することができる。これらから、「40-82569 外面腰部:界線１本, 見込み:鳥, 高台内:天下大平, 重さ:44.1g, 高台半径:2.5cm, 高台高さ:9.5mm, 分類:マントーシン型碗」という測定データが得られる。

このような測定データから、分析用のデータを次のようにして得る。まず、全てのテキストデータから含まれる単語を全て取り出して項目とする。次に、各破片の分析用データを作成する。測定データ中に項目 X があれば、項目 X の値を'1' とし、なければ'0' とする。例を図 3に示す。図 3では、「碗」であり、「皿」ではないので「碗」の項目を'1' とし,「皿」の項目を'0' とする。また、「界線」と「鳥」の模様があるので、それぞれの項目を'1' とする。その他の項目は'0' である。

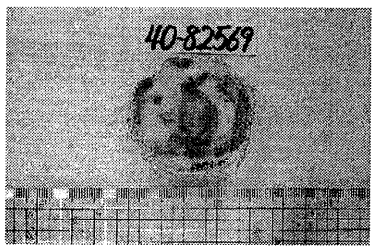

図 2: 破片の例

対象とする皿と碗は、各々、表1と表2のように分類されている [3]。表では1つの分類が1列で表されている。例えば、表1のC群に属する皿には、外面口縁部に波涛文帯、外面胴部に芭蕉葉文、外面腰部に界線といった模様があることを表している。

また、項目数は、同一の項目を1つにまとめてある。例えば、「界線」は複数の分類に現れるが、1つと数えている。また、「なし」は項目数には数えない。さらに、「ごけ底で内湾」は「ごけ底」と「内湾」を分けている。

表 2: 染付の碗の分類 (抜粋)[3]

項目数		5	5	7	6
分類		C群	E群1	E群2	B群
レンツー碗 マントーシン型		◎	◎	◎	端反り
外面	口縁部	波涛文帯	界線	界線	雷文
	胴部	芭蕉葉文	山水人物	花	唐草
	腰部	界線	界線	界線	界線
内面	口縁部	界線	界線	四方	四方
	胴部	なし	なし	なし	なし
見込み		法螺貝	瑞果	牡丹・唐草	花
高台内		なし	文字	文字	なし

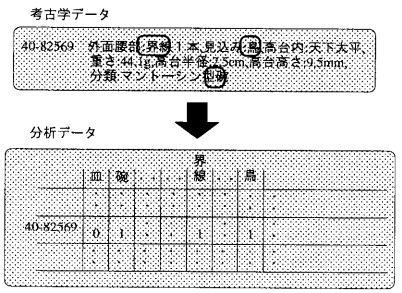

図 3: 分析用のデータ

表 1: 染付の皿の分類 (抜粋)[3]

項目数		6	5	4
分類		C群	D群1	D群2
ごけ底	内湾 端反り	◎		
高台付	内湾 端反り		◎	◎
外面	口縁部	波濤文帯	界線	界線
	胴部	芭蕉葉文	牡丹唐草	なし
	腰部	界線	界線	界線
内面	口縁部	界線	界線	四方襷文
	胴部	なし	なし	なし
見込み		花鳥	玉取獅子	文字
高台内		なし	なし	文字

3 クラスタリング法について

3.1 クラスタ分析

クラスタ分析とは、個体間の類似性をもとにそれらの個体を、幾つかのグループに分類することである [1]。クラスタ分析には、階層的方法と非階層的方法がある。階層的方法は、似たもの同士を併合していくつかのグループにまとめてゆく方法である。非階層的方法は、グループ数を与え、似たものが同じグループに入るように集団の分割と個体の入れ替えを行う方法である。ここでは、前もってグループ数を指定しなくても良い階層的方法を使用する。

階層的方法におけるクラスタの合併時の距離を求める方法の主なものには、最近隣法、最遠隣法、メディアン法、群平均法、重心法がある。最近隣法は、クラスタ間の距離を求める際、各クラスタ内部の点の中で、クラスタ間の距離を最小に取るような点をそのクラスタの代表とする方法である。最遠隣法は、最近隣法と逆で、クラスタ間の距離を最大に取るような点を代表とする方法である。メディアン法は、クラスタ c_i とクラスタ c_j が併合して成立するクラスタ c_{ij} と、クラスタ c_k の間の距離を、クラスタ c_k の重心から、クラスタ c_i とクラスタ c_j の中央点とする方法である。群平均法は、クラスタ間の距離をあるクラスタ中の個体と、もう一つのクラスタ中の全ての個体間の全ての距離の平均をもって定義する方法である。重心法は、2つのクラスタを結合する際、各クラスタの重心をもって、そのクラスタの代表とする方法である。

また、本研究では、クラスタ分析のための項目間の類似性を、分析データの項目間の相関係数 $r(0 \leq r \leq 1)$ として求め、さらに、$ds = 2 - 2 * r$ として距離 (非類似度) に変換する。

一般に、n 個のデータがあった時に、$n-1$ 回クラスタの合併を行なうと1個のクラスタにまとまってしまう。従って、どこでクラスタの合併を止めるかが重要である。ここでは、2つ以上の破片を含むクラスタ同士が合併するときクラスタの合併を停止するという方法を採用する。

3.2 単純なクラスタリング法の問題点

1回で全クラスタを求める通常のクラスタリング法(以下、単純なクラスタリング法と記す)の問題点は、全クラスタの中で同じ項目が高々1回しか出現しないことである。すなわち、1つの項目は、複数のクラスタに属することはないということである。本研究で対象とするデータは、以下に示すように、1つの項目が複数のクラスタに属することがない場合、考古学で定説とされている分類を導き出せない。

例えば、表1の染め付けの皿の分類の表をみてみよう。この分類の表において、「界線」は全ての分類に現れている。仮に、C群とD群1の破片を含むデータがあり、さらにC群がD群1よりも相関が強いとした場合、初めに、C群の(ごけ底、内湾、波涛文、芭蕉葉文、界線)がクラスタにまとまると、次にD群1(端反り、界線、牡丹、唐草、玉取り獅子)がクラスタにまとまるはずであっても、D群1は(端反り、牡丹、唐草、玉取り獅子)の様に「界線」が除外されたクラスタとなってしまう。

4 漸次クラスタリング法

前述の問題点は、1回のクラスタリングで全ての分類を行なうことにある。そこで、相関の強いクラスタを確定クラスタとし、そのクラスタをデータ集合から削除し、残りのデータに対して再度クラスタ分析をし直すという方法をとる。単純なクラスタリング法では、多くの分類に重なっていて隠れていた情報が表に出て来ると考えられる。

ここで、問題となるのは、破片には、細かい破片から、大きな破片まであるということである。図1のAのように、大きな破片であれば、口縁部、胴部、腰部の特徴を含む。しかし、図1のBのように小さな破片では、口縁部の特徴しか含まない。すなわち、小さな破片では、全てのデータが得られない。そのため、削除したいクラスタの、全ての項目を含むデータをそのクラスタに属するとすると、小さな破片は削除されなくなる。一方、1項目でも含むデータを、そのクラスタに属するとすると、他のクラスタに属しているデータまで削除してしまう。従って、「クラスタを削除する」ことは、正確には出来ない。そこで、ここでは、そのクラスタの項目の半分以上を含むデータを「そのクラスタのデータ」とした。例えば、(玉取獅子, 界線, 端反り, 唐草, 牡丹)であれば、その中の3つ以上の項目を含むデータはそのクラスタに含まれることになる。

この方法を、フロチャートの形で、図4に示す。

ここで、フロチャートの「クラスタに含まれる項目の数の入力」のクラスタに含まれる項目の数とは、そのクラスタの項目数の上限と下限のことである。考古学上の分類と比較して、項目数が少なすぎても、多すぎても意味があるデータが抽出できないからである。

ここで述べた方法を、漸次クラスタリング法と呼ぶことにする。

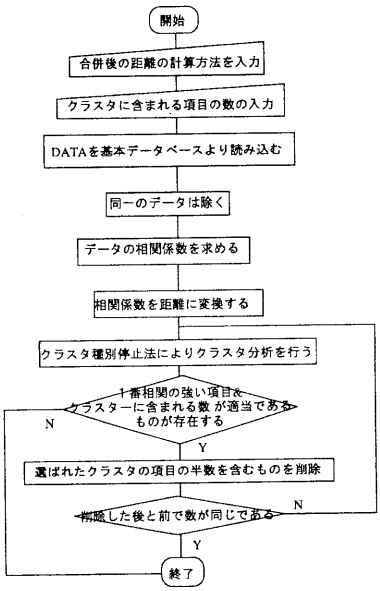

図 4: 漸次クラスタリング法のアルゴリズム

5 評価

5.1 評価方法

単純なクラスタリングによる分析結果と漸次クラスタリングによる分析結果を比較することで評価を行なう。評価項目は、次の3項目である。

A 抽出されたクラスタの数

B $\frac{\text{抽出されたクラスタの項目の数}}{\text{考古学上の分類の表の項目の数}}$ の平均

C 考古学上の分類と比較して誤っている項目の数の合計

ただし、「抽出されたクラスタの数」は、重複しているクラスタがある場合は、それを1つと数える。

例えば、染付の皿を群平均法で合併した場合、以下のような分類が得られる。ただし、誤った分類は□で囲んで示している。

1. (花, 内湾, ごけ底, 芭蕉, 端涛文)
2. (玉取獅子, 界線, 端反り, 唐草, 牡丹)
3. (四方, 植物)

この場合、A、B、Cは以下となる。

$A = 3$

$B = \frac{\frac{5}{5}+\frac{5}{5}+\frac{1}{4}}{3} = 69\%$

$C = 1$

5.2 評価対象

一乗谷朝倉氏遺跡から出土した染付の皿と碗のデータを対象とし評価する。前掲の表1が考古学研究で明らかにされている染付の皿の分類であり、表2が碗の分類である。評価対象とした皿のデータの数は458であり、碗の数は122である。

5.3 評価結果

5.3.1 染付の皿を対象とした場合

単純なクラスタリングの結果を表3に示し、漸次クラスタリングの結果を表4に示す。表中のA，B，Cは評価項目を表す。Bの()内は重複したクラスタを1つのクラスタと考えたものである。

漸次クラスタリング法では、重複したクラスタを一つのクラスタとすると、Bが向上していることが分かる。また、誤りもなくなっている。

5.3.2 染付の碗を対象とした場合

単純なクラスタリングの結果を表5に示し、漸次クラスタリングの結果を表6に示す。Bの()内は重複したクラスタを1つのクラスタと考えたものである。

ここでも、誤りはなくなっている。

表3: 単純なクラスタリング法による染付の皿の評価

クラスタの合併方法	A	B	C
最近隣法	1	80%	0
最遠隣法	3	64%	2
メディアン法	1	80%	0
群平均法	3	69%	1
重心法	1	100%	0

表4: 漸次クラスタリング法による染付の皿の評価

クラスタの合併方法	A	B	C
最近隣法	1(重複あり)	70%(100%)	0
最遠隣法	3(重複あり)	69%(92%)	0
メディアン法	1	80%	0
群平均法	2(重複あり)	72%(92%)	0
重心法	1	100%	0

表5: 単純なクラスタリング法による染付の碗の評価

クラスタの合併方法	A	B	C
最近隣法	3	57%	0
最遠隣法	2	40%	2
メディアン法	0	0%	0
群平均法	4	50%	2
重心法	1	60%	0

表6: 漸次クラスタリング法による染付の碗の評価

クラスタの合併方法	A	B	C
最近隣法	2(重複あり)	60%(60%)	0
最遠隣法	1	60%	0
メディアン法	0	0%	0
群平均法	3	48%	0
重心法	1	60%	0

5.4 考察

初めに、染付の皿に対する評価結果(表3、表4)について考察する。「抽出されたクラスタの数」は、重複を除くと変化しない。次に、「$\frac{\text{抽出されたクラスタの項目の数}}{\text{分類の表の項目の数}}$」は、重複したクラスタを1つとして考えると、漸次クラスタリング法の方が良い結果が得られている。これは、漸次クラスタリング法が成功していることを示す。単純なクラスタリング法では、1つの分類にしか現れない「界線」が、漸次クラスタリング法では、多くの分類に現れたためである。また、漸次クラスタリング法では「間違った項目」がなくなっている。それは、漸次クラスタリング法が、一回毎に、最も相関の強いクラスタを選んでいるために、最も相関が強いクラスタではない間違った項目を含む分類が抽出されなかったためである。

次に、染付の碗に対する評価結果(表5と表6)について考察する。染付の碗のデータの特徴として、「界線」が全分類の中でほぼ均等に現れる。そのため、漸次クラスタリング法において、最も相関の強いクラスタを選んでも、「界線」が現れない。そのため、単純なクラスタリング法と漸次クラスタリング法を比較しても、結果はあまり変わらない。また、「抽出されたクラスタの数」が減っているのは、間違った項目を含むクラスタを抽出しなくなったためである。ただし、最近隣法の場合は違い、目的とするクラスタが2番目に相関の強いクラスタになり、うまく抽出されなかったためである。

染付の皿のように、最も相関の強いクラスタに、複数の分類にまたがっている項目(「界線」等)を含んでいるものは、漸次クラスタリング法を施すと、良好な結果が得られる。しかし、染付の碗の様に、最も相関の強いクラスタに含まれないものは、どちらの方法でもあまり結果は変わらない。また、漸次クラスタリング法は最も相関の強いクラスタを選んでいるため、間違った項目のような、あまり相関の強くないものは削除される。

6 おわりに

本研究では、一乗谷朝倉氏遺跡から出土した遺物データを対象として、考古学データベースからの情報の抽出方法を検討した。

本研究で対象とした一乗谷朝倉氏遺跡から出土した染付の皿と碗では、複数の分類に同じ項目が存在する。そのため、単純に1回のクラスタ分析では、必要な情報が得られない。そこで、クラスタ分析の結果、最も相関の強いクラスタを削除し、再度クラスタ分析を行なうという漸次クラスタリング法を用いた。最も相関の強いクラスタの削除にあたり、クラスタの項目の半数以上の項目を含むデータをそのクラスタとした。評価の結果、染め付けの皿と碗について良好な結果が得られた。

今後は、多くの種類のデータに対する漸次クラスタリングの適用、破片上の場所に関する情報の考慮、大小様々な大きさの破片への対応が課題である。

謝辞

データの収集、分類作業等でお世話になっている福井県立一乗谷朝倉氏遺跡資料館の岩田 隆 氏、水村 伸行 氏に深謝いたします。

参考文献

[1] 柳井晴夫, 高木廣文, 市川雅教, 服部芳明, 佐藤俊哉, 丸井英：多変量解析ハンドブック, 現代数学社 (1986).

[2] 田中豊, 脇本和昌：多変量統計解析法, 現代数学社 (1983).

[3] 小野正敏：15〜16 世紀の染め付け碗・皿の分類と年代, 貿易陶磁器研究, Vol. 2, pp. 74-76(1982).

[4] 中田充, 宝珍輝尚, 都司達夫：サイエンティフィックデータベースのための一次データベースの一管理法, 情処第 52 回全国大会, 4-265(1996).

[5] 中田充, 宝珍輝尚, 都司達夫：サイエンティフィックデータモデルの一評価, 信学技報 DE95-46, pp. 113-120(1995)