質的データに対する制約条件付クラスタリング手法

栗 田 多喜夫*, 大 津 展 之*

SOME METHODS OF CONSTRAINED CLUSTERING FOR QUALITATIVE DATA

Takio KURITA, Nobuyuki OTSU

This paper is concerned with a problem of constrained clustering for qualitative data. Combinatorial problem of scheduling meetings is a typical example. The desired schedule of meetings is such one that minimizes the total number of man-days.

In those clustering problems, the computational time to find the optimal solution increases exponentially as the number of meetings grows. In order to find a near optimal solution efficiently, we discuss some methods; 1) methods based on quantification methods (Quantification Theory 3, Karhunen-Loève Expansion) and 2) a Bottom-up Merging (Hierarchical Clustering) method which is derived from the inherent criterion of the problem. In method 2), particularly, the representative vector of each cluster and the metric for merging clusters are logically and uniquely determined by the criterion. It is shown by experiments that schedules obtained by these methods are better than those obtained in trial and error.

1. はじめに

多変量解析の手法は、従来、主としてデータの解析のために用いられてきたが、それは、今後急速に需要が高まるであろう複雑な仕事の計算機による自動化のための補助手段としても有効であると考えられる。本稿では、直接最適解を求めるのが難しい問題の準最適解を高速に求めるために多変量解析の手法が有効であることを示す例として、日程表作成問題について考察する。

日程表作成問題とは、例えば、多くの項目とその関係者の間の該当関係(相反表、質的データ)をもとに、以下のような項目の打ち合わせ会議の日程表を作る問題である。会議は時間の制約により一日に何個かずつしか開催できないとする。つまり、全ての項目の打ち合わせ会議を終了するには数日間必要である。各関係者が各々一つの項目のみに関係している場合には問題はないが、多くの項目に同時に関係している場合には会議の日程によって関係者の出席日数が変わる。日程

表の作り方が悪いと関係者の出席日数が多くなり、会議のために無駄にしなければならない日数が増える。特に、関係者が会議に出席するために遠方から出て来なければならないような場合には、出席日数を少なくすることは意味がある。従って、好ましい日程表とは、関係者の延べ出席日数を最小にする日程表である¹⁾。

このような評価基準(延べ出席日数最小)のある場合の質的データの制約条件付きクラスタリング問題に対して、最適解を直接組合せ的に求めようとすると、その計算量は項目の数に関して指数的に増大する。従って、項目の数がある程度以上になると計算不可能になってしまう。例えば、我々が行った計算例では、項目数が13のかなり小さな問題に対しても、VAX-11/780で18時間もの計算が必要であった。

2章において、この問題をもう少し一般的に定式化し、3章において、原データから大まかな情報を抽出することにより、難しい問題を簡単な問題に帰着させて解くという一般的な方策の具体化として、数量化3類あるいはK-L展開(主成分分析)を用いて一次元のスコアを求め、そのスコアをもとに、制約条件の下で、判

* 電子技術総合研究所(Electrotechnical Laboratory)

1) 本稿では、会議の順序については考慮していない

別規準により分割する手法について述べる。4章においては、問題から自然に与えられる評価基準に沿って導出した bottom-up metging (階層的クラスタリング) による手法を提案する。それは問題に即した手法であり、従来の何らかの距離 (メトリック) をトップダウンに仮定する階層的クラスタリング手法と異なり、クラスターの代表ベクトルおよび統合のための測度が論理的に一意に定まる。さらに、クラスター代表ベクトル間の Jaccard の Matching Coefficient をクラスター間の測度とする手法を提案し、それが上記の手法と定性的に似た振舞いをするを示す。また、従来の階層的クラスタリング手法の代表例として Group Average 法との関係についても述べる。5章において、人工データおよび実際の問題に対してこれらの手法を適用し、その有効性を示す。

2. 問題及びその定式化

日程表作成問題は、項目 vs 関係者の相反表から、各クラスター (日に対応) に属する項目数がある値 (1 日の許容会議数) 以下にするという制約条件のもとで、項目をあるクラスター数 (会議の総日数) に分割 (日割り) するクラスタリング問題とみなせる。この時の評価基準は関係者の延べ出席日数を最小とすることである。この問題を以下のように定式化する。

項目の総数を M , 関係者の総数を N , クラスター数を K , 各クラスターに属する項目数の上限を D , 項目 vs 関係者の相反表を $X = [x_{ij}]$, ($1 \leq i \leq M, 1 \leq j \leq N$) とする。ただし, $x_{ij} \in \{0, 1\}$ であり, $x_{ij} = 1$ は関係者 j が項目 i に関係していることを意味する。また, 各項目に対応する行ベクトル $x_1, \dots, x_M$ を該当ベクトルと呼ぶことにする。この時, 問題は,

$$n_k \leq D \quad (1 \leq k \leq K) \quad (1)$$

なる制約条件のもとで, 関係者の延べ出席日数

$$f(C_1, \dots, C_K) = \sum_{k=1}^K w\left(\bigcup_{i \in C_k} x_i\right) \quad (2)$$

を最小にする日割り $\{C_1, \dots, C_K\}$ を求めることである。ここで, n_k は第 k 日に割り当てられた項目の個数を表し, C_k は第 k 日に割り当てられた項目の集合を表す。また, $w(x)$ はベクトル x の要素の 1 の個数を評価する関数であり, $\bigcup_{i \in C_k} x_i$ は第 k 日に割り当てられたすべての項目に対して, その該当ベクトルの要素ごとの OR (論理和) を取ることを意味する。

3. 数量化によるクラスタリング

3.1. 手 法

問題を解くために必要な情報のうち詳細な情報を無視し大まかな情報のみを抽出することにより, 難しい問題を簡単な問題に帰着させて解く方法は, 難しい問題を解くための一般的な方策である。日程表作成問題において, 関係者の延べ出席日数を最小にするためには類似した該当ベクトルを持つ項目をなるべく同じ日に割り当てればよい。従って, 類似した該当ベクトルに対してはその差が小さくなるような数量化を行い, そのスコアによって日割りを決定すればよい。つまり, 情報を抽出することにより, M 次元の組み合わせ問題を近似的に 1 次元の組み合わせ問題に帰着させることになる。

数量化の方法としては, 数量化 3 類, K-L 展開 (主成分分析) 等が考えられる (例えば, 林 (1970), 駒沢 (1982), 西里 (1982))。数量化 3 類を適用した結果を用いる方法は, 一般の質的データに対するクラスタリング手法として, すでに Jambu (1983) に見られる。また, K-L 展開の結果を用いたクラスタリング手法は, 質的データだけでなく連続データに対しても用いられるよく知られた手法である。ただ, 日程表作成問題に対しては, 制約条件 (1) を満足するような分割が必要となる。3.2. において, 分割のための近似手法として, 平均クラスター内分散を最小とする手法について述べる。

3.2. 判別規準による分割

各項目のスコアを一次元の数直線上にならべると, 数直線上での点の近さは項目間の類似性を表しているから, 類似した該当ベクトルを持つ項目を同じクラスターに割り当てるためには平均クラスター内分散 (級内分散) を小さくすればよい。数直線上での順序を無視してクラスターを割り当てると明らかに級内分散は大きくなる。従って, 数直線上の M 個の点を, 各クラスターの点の数を D 以下にするという判約条件のもとで, 級内分散を最小にするように $K-1$ 個のしきい値によって K 個のクラスターに分割する問題を考えればよいことになる。この問題は本質的には, 大津 (1980) の多値の自動しきい値選定に相当する。しきい値の組は制約条件付き動的計画法 (DP) によって求めることができ (付録参照), その手続きの手間は $O(K * (M - K)^2)$ である。従って, 全体としてクラスタリングに必要な手間は, 数量化の手間 $O(M^3)$ を加えて $O(M^3)$

となる。

4. bottom-up merging によるクラスタリング

4.1. 評価基準に沿って導出した bottom-up merging による手法

次に、各項目を一つのクラスターに対応させた M 個のクラスターから始めて、クラスター数が K になるまで、次々と二つのクラスターを統合し、新しいクラスターをつくる bottom-up merging による階層的な手法を問題の評価基準に沿って導出する。

今、問題の評価基準 (2) は、項目を $\{C_1^{(p)}, \dots, C_p^{(p)}\}$ の p ($K \leq p \leq M$) 個のクラスターに分割する場合、

$$f(C_1^{(p)}, \dots, C_p^{(p)}) = \sum_{k=1}^p w\left(\bigcup_{i \in C_k^{(p)}} x_i\right) \quad (3)$$

と書ける。ここで、 $C_k^{(p)}$ は、 p 個に分割したときの第 k ($1 \leq k \leq p$) 番目のクラスターに属する項目の集合を表す。この時、第 k クラスターの評価値は第 k クラスターに属する項目の該当ベクトルの要素ごとの OR を取ったベクトルの 1 の個数によって決まる。従って、第 k クラスターを代表するベクトルを

$$r_k^{(p)} = \bigcup_{i \in C_k^{(p)}} x_i \quad (4)$$

とすると、(3) 式の右辺は

$$\sum_{k=1}^p w(r_k^{(p)}) \quad (3')$$

と書ける。このクラスター代表ベクトルは、会議の関係者がそのクラスターに関係しているかどうかを示す該当ベクトルである。

最終的な評価値 (2) を小さくするためには、クラスターを統合する時、統合後の評価値を最小にする、つまり、統合の前と後の評価値の差を最大にするクラスターの対を統合すればよい。

統合前のクラスターを $\{C_i^{(p)}, \dots, C_j^{(p)}\}$ ($K \leq p \leq M$) とし、クラスター $C_i^{(p)}$ と $C_j^{(p)}$ ($i < j$) を統合する場合を考える。統合前の評価値は (3') で与えられ、統合後の評価値は、

$$\sum_{k=1}^{i-1} w(r_k^{(p)}) + \sum_{k=i+1}^{j-1} w(r_k^{(p)}) + \sum_{k=j+1}^p w(r_k^{(p)}) + w(r_i^{(p)} \cup r_j^{(p)}) \quad (5)$$

と書ける。ただし、 $x \cup y$ はベクトル x と y の要素ごとの OR を取ることを表す。

従って、統合前と統合後の評価値の差は、簡単に

$$\begin{aligned} & w(r_i^{(p)}) + w(r_j^{(p)}) - w(r_i^{(p)} \cup r_j^{(p)}) \\ & = w(r_i^{(p)} \cap r_j^{(p)}) \end{aligned} \quad (6)$$

となることが分かる。ただし、 $x \cap y$ はベクトル x と y の要素ごとの AND (論理積) を取ることを表す。従って、二つのクラスター間の評価測度として、

$$m(r_i^{(p)}, r_j^{(p)}) = w(r_i^{(p)} \cap r_j^{(p)}) \quad (7)$$

が導かれる。この測度 m は Russell と Rao の Matching Coefficient R (Anderberg (1973)) と

$$R = m/N \quad (8)$$

なる関係がある。従って、 m を最大とするクラスター対を統合することはクラスター代表ベクトル間の Russell と Rao の Matching Coefficient R を最大とするクラスター対を統合することと等価になる。ただし、評価基準から自然に導かれた測度 m には両方のクラスターにまたがって関係する人の数という論理的な意味があり、問題とのつながりが明白である。

また、 $C_i^{(p)}$ と $C_j^{(p)}$ の統合によって新しく作られるクラスターの代表ベクトル $r_{i+j}^{(p-1)}$ は、上述の議論から、

$$r_{i+j}^{(p-1)} = r_i^{(p)} \cup r_j^{(p)} \quad (9)$$

となり、統合前のクラスターの代表ベクトルの要素ごとの OR を取ったベクトルによって簡単に表せる。それ以外のクラスターの代表ベクトルは変わらず、

$$r_s^{(p-1)} = r_s^{(p)} \quad (s \neq i, j) \quad (10)$$

である。

階層的クラスタリング手法には、二つのクラスター間の測度をどう定めるかによって多くの変種が存在する (Anderberg (1973))。例えば、Single Linkage 法では、最も似た二つのベクトルを各クラスターの代表ベクトルとして、それらの間に定義された測度を二つのクラスター間の測度と定義している。これに対し、ここで提案した手法では、評価基準から各クラスターの代表ベクトルおよびクラスター間の測度 m が一意に定まり、それぞれ明確な論理的意义を持っている。

測度 m は、離散的な値を取るため値が等しい (タイ) の場合の処理方法を考えることは重要である。トップダウンに何等かの距離 (メトリック) を導入した場合には、それは非常に難しい問題であるが、日程表作成問題では、題意から簡単な heuristics が考えられる。延べ出席日数を最小とするためには、各クラスターに関係する人の数を少なくする必要がある。従って、タイの場合には、統合後のクラスターに関係する

人の数の少ないクラスター対を統合すればよい。統合後のクラスター対に関係する人の数は、二つのクラスター代表ベクトルの要素ごとの OR を取ったベクトルの 1 の個数に他ならない。それは、測度 m の計算と同時に計算できる。従って、タイの場合の処理を加えても計算量はほとんど変わらない。

実際のアルゴリズムでは、クラスターの統合を定数時間で実行するため、各クラスターをそのクラスターに属する項目のダブルリンクドリストで表現している。そうすることにより、このアルゴリズムの手間は $O(K * M^2)$ となる。

4.2. Jaccard の Matching Coefficient を用いた手法の導出

二つのクラスター代表ベクトルの対 $(x_1, x_2), (y_1, y_2)$ に対して、測度 m がタイの場合、上述の手法では、統合後のクラスターに関係する人の数 $w(x_1 \cup x_2)$ あるいは $w(y_1 \cup y_2)$ の小さいクラスター対が統合される。従って、 a, b をクラスター代表ベクトルとし、新しく測度 J を、

$$J(a, b) = \frac{m(a, b)}{w(a \cup b)} = \frac{w(a \cap b)}{w(a \cup b)} \quad (11)$$

と定義すると、 J はタイの場合には上述の手法と同じクラスター対を選ぶ。 m が等しくない場合に対しても、クラスター代表ベクトルの各要素の 0 および 1 がランダムに現れると考えると、 $w(a \cup b)$ は平均的には $(N - m(a, b))/2$ となる。従って、 $m(x_1, x_2) > m(y_1, y_2)$ ならば、平均的に

$$\begin{aligned} J(x_1, x_2) &= \frac{2m(x_1, x_2)}{N - m(x_1, x_2)} \\ &> J(y_1, y_2) = \frac{2m(y_1, y_2)}{N - m(y_1, y_2)} \end{aligned} \quad (12)$$

となる。こうして、クラスター代表ベクトルをクラスターに属する該当ベクトルの要素ごとの OR を取ったベクトルとし、クラスター統合のための測度を J とした手法は、前述のタイの場合の処理を含んだ手法と定性的に似た振舞いをする事が分かる。

ここで導出した測度 J は、Jaccard の Matching Coefficient (Anderberg (1973)) と呼ばれている。

4.3. 従来の階層的クラスタリング手法との比較

上述の bootom-up merging によるクラスタリングでは、問題から自然に与えられる評価基準から、クラスターの統合のための測度および新しく作られるクラスターの代表ベクトルが論理的に一意に定まった。し

かも、得られた測度およびクラスター代表ベクトルは、該当ベクトルの要素ごとの論理演算を含んでいる。二値（質的）データに対して論理演算を施すことは自然であり、データの意味を考慮することにより、測度およびクラスター代表ベクトルの意味が自然に理解できる。さらに、論理演算中心の上述の手法は、計算機との相性（ビット演算）がよく、従来の手法に比べて大きなサイズの問題に対しても適用可能である。

ここで、さらに、従来の階層的クラスタリング手法の代表例として、Group Average 法との関係について調べてみる。各項目間の類似度を s_{ij} とすると、Group Average 法でのクラスター統合のための測度 g は、二つのクラスター C_p, C_q に対して、

$$g(C_p, C_q) = \sum_{i \in C_p, j \in C_q} s_{ij} / N_p N_q \quad (13)$$

のように、二つのクラスター間のペアごとの類似度の平均で定義される (Anderberg (1973))。ここで、 N_p, N_q はそれぞれクラスター C_p, C_q に属する項目数である。

今、項目間の類似度を

$$s_{ij} = w(x_i \cap x_j) = \sum_{k=1}^N x_{ik} x_{jk} \quad (14)$$

のように、該当ベクトルの要素ごとの AND を取ったベクトルの 1 の個数とすると、 g は、

$$\begin{aligned} g(C_p, C_q) &= \sum_{i \in C_p, j \in C_q} \left(\sum_{k=1}^N x_{ik} x_{jk} \right) / N_p N_q \\ &= \sum_{k=1}^N \left(\sum_{i \in C_p, j \in C_q} x_{ik} x_{jk} \right) / N_p N_q \\ &= \sum_{k=1}^N \left(\frac{1}{N_p} \sum_{i \in C_p} x_{ik} \right) \left(\frac{1}{N_q} \sum_{j \in C_q} x_{jk} \right) \\ &= \left(\frac{1}{N_p} \sum_{i \in C_p} x_i \right) \cdot \left(\frac{1}{N_q} \sum_{j \in C_q} x_j \right) \end{aligned} \quad (15)$$

のように、各クラスターの平均該当ベクトル間の内積となる。

一方、4.1 で導出したクラスター統合のための測度 m は、(4), (7) から分かるように、

$$m(C_p, C_q) = \left(\bigcup_{i \in C_p} x_i \right) \cdot \left(\bigcup_{j \in C_q} x_j \right) \quad (16)$$

と書ける。従って、二つの項目間の類似度を該当ベクトル間の要素ごとの AND を取ったベクトルの 1 の個数とした場合の Group Average 法と、4.1 で導出した手法との本質的な差は、クラスター間の類似度を、平均該当ベクトル間の内積とするか、要素ごとの OR を取ったベクトル間の内積とするかであることが分かる。

5. 実験

5.1. 人工的に作ったデータによる実験

まず、最適解と比較するため、全ての場合を調べ上げられる比較的小さなデータを人工的に作り、実験を行った。表1が項目 vs 関係者の相反表である。この表をもとに、1日の許容会議数を4とし、4日間の日程表を作るとする。表2および表3が最適解および各手法の結果である。

K-L展開による手法および数量化3類による手法は、表2では全く同じ結果になっている。実際の分割は、表3a)およびb)のように多少異なっているが、両者はかなり似た傾向にある。

タイの処理をしない測度 m を用いる手法は、K-L展開による手法および数量化3類による手法よりも良い日程表を作っている。タイの処理をすることによって、結果がさらに改善され、最適解が得られている。また、Jaccard の Matching Coefficient を用いる手法も最適解を与えている。

図1は延べ出席日数と可能な日程表の個数の関係を示したヒストグラムである。ただし、日程表の個数は、

表1. 人工的に作った相反表

関係者 項目	1	2	3	4	5	6	7	8	9	10
1				●						
2	●	●	●							
3			●	●	●					
4	●			●	●	●				
5		●	●	●	●	●	●			
6		●	●	●	●	●	●			
7		●	●	●	●	●	●	●		
8				●	●	●	●		●	
9				●		●	●	●		
10				●	●		●	●	●	●
11						●	●	●	●	●
12						●	●		●	●
13									●	●

表2. 人工データに対するクラスタリングの結果

a) K-L展開による手法

b) 数量化 III 類による手法

出席日数 (日)	人 数 (人)	延べ出席日数 (人日)
1	2	2
2	3	6
3	2	6
4	3	12
計		26

c) タイの処理をしない測度 m を用いる手法

出席日数 (日)	人 数 (人)	延べ出席日数 (人日)
1	0	0
2	6	12
3	4	12
4	0	0
計		24

d) タイの処理をした測度 m を用いる手法

e) Jaccard の Matching Coefficient を用いる手法

f) 最適解

出席日数 (日)	人 数 (人)	延べ出席日数 (人日)
1	1	1
2	6	12
3	3	9
4	0	0
計		22

a) K-L 展開による手法で作られた日程表

[illegible][illegible][illegible]

f) 最適な日程表

[illegible]

度の日程表ができることになる.

図ではほとんど読み取れないが、最適解（タイの処理をした測度 m を用いる手法、Jaccard の *Matching Coefficient* を用いる手法によって得られた日程表）と

[illegible]

図1. 延べ出席日数と可能な日程表の個数の関係

表5. 実際のデータに対するクラスタリングの結果

a) K-L展開による手法

出席日数 (日)	人 数 (人)	延べ出席日数 (人日)
1	4	4
2	5	10
3	5	15
4	1	4
計		33

b) 数量化 III 類による手法

出席日数 (日)	人 数 (人)	延べ出席日数 (人日)
1	5	5
2	4	8
3	4	12
4	2	8
計		33

c) タイの処理をしない測度 m を用いる手法

出席日数 (日)	人 数 (人)	延べ出席日数 (人日)
1	5	5
2	8	16
3	2	6
4	0	0
計		27

d) タイの処理をした測度 m を用いる手法

e) JaccardのMatching Coefficientを用いる手法

出席日数 (日)	人 数 (人)	延べ出席日数 (人日)
1	6	6
2	7	14
3	2	6
4	0	0
計		26

f) 人間が試行錯誤的に作った日程表

出席日数 (日)	人 数 (人)	延べ出席日数 (人日)
1	4	4
2	6	12
3	3	9
4	2	8
計		33

同等な日程表の個数は24個である。つまり、各クラスターを第何日に割り当てるかによって生じる日程表の差を無視すると、最適解は一通りである。タイの処理をしない測度 m を用いる手法により得られた日程表と同等な日程表の個数は4152個であり、K-L展開による手法および数量化3類による手法で得られた日程表と同等な日程表の個数は66648個である。これに対して、延べ出席日数32日の日程表の個数は3241512個である。このことから、タイの処理をした測度 m を用いる手法およびJaccardのMatching Coefficientを用いる手法は、非常に優れた手法であることが分かる。また、K-L展開による手法あるいは数量化3類による手法で作られた日程表もまずまずの日程表であること

表 6. 実際のデータに対して作られた日程表

a) K-L 展開による手法

日 程	項	目
第 1 日	1 2 3 31 32	
第 2 日	4 10 11 13 32 34 35 36 37	
第 3 日	5 18 21 22 24 25 30 38 39	
第 4 日	6 7 8 9 12 14 15 23 26	
第 5 日	16 17 19 20 27 28 29	

b) 数量化 III 類による手法

日 程	項	目
第 1 日	1 2 3 31	
第 2 日	4 10 11 32 33 34 35 36 37	
第 3 日	5 21 22 24 25 26 30 38 39	
第 4 日	6 7 8 9 12 13 14 15 23	
第 5 日	16 17 18 19 20 27 28 29	

c) タイの処理をしない測度 m を用いる手法

日 程	項	目
第 1 日	1 2 3 23 27 28 29 31 33	
第 2 日	4 22 32 34 35 36 37 38 39	
第 3 日	5 6 7 16 17 18 19 20 30	
第 4 日	8 9 10 11 12 13 14 15 26	
第 5 日	21 24 25	

d) タイの処理をした測度 m を用いる手法

日 程	項	目
第 1 日	1 2 3 31 32 33 34 36 37	
第 2 日	4 22 35 38 39	
第 3 日	5 6 7 8 9 10 11	
第 4 日	12 13 14 15 21 23 24 25 26	
第 5 日	16 17 18 19 20 27 28 29 30	

e) Jacard の Matching Coefficient を用いる手法

日 程	項	目
第 1 日	1 2 3 31 32 33 34 36 37	
第 2 日	4 5 6 7 8 9 10 11 35	
第 3 日	12 13 14 15 21 23 24 25 26	
第 4 日	16 17 18 19 20 27 28 29 30	
第 5 日	22 38 39	

f) 人間が試行錯誤的に作った日程表

日 程	項	目
第 1 日	1 2 3 4 5 6	
第 2 日	7 8 9 10 11 12 13 14 15	
第 3 日	16 17 18 19 20 21	
第 4 日	22 23 24 25 26 27 28 29 30	
第 5 日	31 32 33 34 35 36 37 38 39	

が分かる。

さらに、計算時間の点でも、直接最適解を求めるために、VAX-11/780 で 18 時間近く計算が必要であるのに比べ、本稿で述べた手法の計算時間は 2 秒以下であった。

5.2. 実際のデータによる実験

表 4 が実際のデータである。1 日の許容会議数は 9 であり、5 日間の日程表を作るとする。表 5, 6 がその結果である。ここで、(f) は人間が試行錯誤的に作った日程表に対する結果である。

ここでも、K-L 展開による手法と数量化 3 類による手法は似た傾向にあることが分かる。測度 m を用いる手法は、タイの処理により、結果が改善されているこ

とが分かる。また、タイの処理をした測度 m を用いる手法と Jaccard の Matching Coefficient を用いる手法とが似た結果を与えていることも分かる。

5.3. 考 察

K-L 展開および数量化 3 類による手法は、大まかな情報を抽出することにより、複雑な問題を簡単な問題に帰着させるという複雑な問題を解くための一般的な方策に従って、多次元の組み合わせ問題を動的計画法によって多項式時間で解が得られる一次元の組み合わせ問題に帰着させて解く。実験では、両手法ともまずの日程表を作っている。このことは、他の複雑な問題を解く場合にも、K-L 展開あるいは数量化 3 類等の情報抽出手法が有効な補助手段となり得ることを示

唆していると考えられる。

一方、評価基準に沿って導出した測度 m を用いる手法は、上述の一般的な二つの手法と異なり、問題に即した手法であり、上述の手法ほどの一般性はないが、本稿で考えた問題に対して良い結果を与えることが分かる。階層的クラスティングには、クラスター間の類似度（距離）をどう定めるかによって多くの手法が存在するが、日程表作成問題の場合、クラスター代表ベクトルおよびクラスター統合のための測度が評価基準から論理的に一意に定まり、しかも、原データに対する論理演算により計算できる点が興味深い。これは、離散的な評価基準を持つ二値（質的）データのクラスティングの問題に対する自然な結果と考えられる。

4.2において、JaccardのMatching Coefficientを用いる手法が、定性的にタイの処理をした測度 m を用いる手法と似た振舞いをすることを示したが、実験的にも、その傾向が分かる。さらに詳細な比較のためには、多くの例についての実験が必要であろう。

また、4.3において、項目間の類似度を該当ベクトル間の要素ごとのANDを取ったベクトルの1の個数としたGroup Average法と測度 m を用いる手法との関係について述べたが、それらに対する実験および詳細な比較等については今後の課題としたい。

謝 辞

日頃御討論頂くソフトウェア部数理情報研究室の皆様、研究の機会を与えて下さった田村前室長、棟上部長に感謝いたします。また、審査員の方々からは有益なご教示をいただきました。ここに記して謝意を表します。

参 考 文 献

- Anderberg, M.R. (1973) Cluster Analysis for Applications, Academic Press.
 Hartigan, J.A., 西田他訳 (1983) クラスター分析, マイクロソフトウェア.
 林知己夫 (1970) 情報処理と統計数理, 産業図書.
 Jambu, M. and Lebeaux, M-O. (1983) Cluster Analysis and Data Analysis. North-Holland.
 駒沢 勉 (1982) 数量化論理とデータ処理, 朝倉書店.
 西里 静彦 (1982) 質的データの数量化, 朝倉書店.
 大津展之 (1980) 判別及び最小2乗法に基づく自動しきい値選定法, 電子通信学会論文誌, D, j63-4, 349-356.
 大津展之 (1981) パターン認識における特徴抽出に関する数理的研究, 電子技術総合研究所研究報告.

—1985年12月初稿受付, 1986年3月改稿受付—

付 録

平均クラスター内分散を最小とする分割

スコアを小さい方から番号を付けて, $y_i (1 \leq i \leq M)$ とする。また、そのスコアに対応する項目も番号 i で参照する。

スコア y_i の全平均 μ_T および全分散 σ^2_T は、それぞれ

$$\mu_T = \frac{1}{M} \sum_{i=1}^M y_i \quad (\text{付1})$$

$$\sigma^2_T = \frac{1}{M} \sum_{i=1}^M (y_i - \mu_T)^2 \quad (\text{付2})$$

となる。

M の項目 $\{1, \dots, M\}$ を $K-1$ のしきい値

$$1 \leq k_1 \leq \dots \leq k_{K-1} < M \quad (\text{付3})$$

で、 K 個のクラスター $C_1 = \{1, \dots, k_1\}, \dots, C_K = \{k_{K-1} + 1, \dots, M\}$ に分割したとする。この時、クラスター $C_p (1 \leq p \leq K)$ でのスコアの平均 μ_p および分散 σ^2_p は、

$$\mu_p = \frac{1}{n_p} \sum_{i=k_{p-1}+1}^{k_p} y_i \quad (\text{付4})$$

$$\sigma^2_p = \frac{1}{n_p} \sum_{i=k_{p-1}+1}^{k_p} (y_i - \mu_p)^2 \quad (\text{付5})$$

となる。ただし、 $n_p = k_p - k_{p-1}$, $k_0 = 1$, $k_M = M$ とする。平均クラスター内分散 σ^2_w および平均クラスター間分散 σ^2_B は、

$$\sigma^2_w = \sum_{p=1}^K \frac{n_p}{M} \sigma^2_p \quad (\text{付6})$$

$$\sigma^2_B = \sum_{p=1}^K \frac{n_p}{M} (\mu_p - \mu_T)^2 \quad (\text{付7})$$

となる。この時、基本的関係

$$\sigma^2_w + \sigma^2_B = \sigma^2_T \quad (\text{付8})$$

が常に成立する。 σ^2_T は分割に無関係であるから、 σ^2_w を最小にすることは σ^2_B を最大にすることと同値である。ここでは、 σ^2_B を最大にする基準を採用することにする。 σ^2_B はさらに、

$$\sigma^2_B = \frac{1}{M} \sum_{p=1}^K \frac{1}{n_p} \left(\sum_{i=k_{p-1}+1}^{k_p} y_i \right)^2 - \mu_T^2 \quad (\text{付9})$$

と書ける。 μ_T は分割に無関係であるから、結局、目的関数は $K-1$ 個のしきい値の関数として、

$$p(k_1, \dots, k_{K-1}) = \sum_{p=1}^K g(k_{p-1}, k_p) \quad (\text{付10})$$

となる. ただし, $g(k_{p-1}, k_p) = \frac{1}{n_p} \left(\sum_{i=k_{p-1}+1}^{k_p} y_i \right)^2$ である.

最適しい値の組は, 条件 (付3) および制約条件

$$n_p \leq D \quad (1 \leq p \leq K) \quad (\text{付11})$$

のもとで, p を最大とするものとして, 制約条件付き動

的計画法 (DP) によって求めることができる.

$$\rho_p^*(k_p) = \begin{cases} \max_{p-1 \leq k_{p-1} \leq k_{p-1}} \{ \rho_{p-1}^*(k_{p-1}) + g(k_p, k_{p-1}) \} & \text{if } \rho_{p-1}^*(k_{p-1}) \neq -1 \text{ \& } n_p \leq D \\ -1 & \text{それ以外のとき} \end{cases}$$

$$\rho_1^*(k_1) = \begin{cases} g(k_0, k_1) & \text{if } n_1 \leq D \\ -1 & \text{それ以外のとき} \end{cases}$$

とすると, $p_M^*(K)$ が p の最大値を与える.