離散最適化手法による変量のクラスタリング

針谷　尚幸
(東京大学大学院情報理工学系研究科数理情報学専攻　現所属・㈱日本総合研究所)
指導教官　岩田　覚 助教授

1. はじめに

近年の情報技術の進歩とともに，人間が取り扱う情報量は飛躍的に増加している．それに伴い，莫大な情報を整理し，有用な形に加工，活用する技術が求められてきている．象徴的な例を挙げれば，データマイニングはITのキーワードとして頻繁に紹介されており，学術的な専門用語から一般用語になりつつある．しかし，用語が広く認知される一方，その語の定義は曖昧模糊となり，乱用されているように映る．未加工のデータから情報を掘り起こす手法は様々であり，どの手法で分析するかという点においては恣意が入るが，それらの手法は極力，主観，先入観といった人為的な偏見を排し，客観的であるべきだと私は考える．本論文では，データマイニングの一手法である変量のクラスター分析について論じる．最尤推定法や情報量規準(AIC，BIC)に基づく規準を提示し，それらの規準を最適化する分割を求めるのに，対称劣モジュラ関数最小化を利用したり，線形計算に工夫を凝らしたりした離散最適化技法によるアプローチについて考察する．そして，それらのアルゴリズムを実装，比較検討を行い，新しい多変量解析手法を提案する．

2. クラスター分析のアルゴリズム

ます，変量の全体集合を V，変量数を n，変量の部分集合であるクラスターを C，クラスターの集合である変量の分割を $\mathcal{P}$ と表記することにする．また，標本相関行列 R から C に対応する行と列を取り出した主小行列を $R[C]$ とし，標本数を m として，

$$f_{\mathrm{ML}}(\mathcal{P})=\sum_{C_j\in\mathcal{P}}\log\det R[C_j], \tag{1}$$

$$f_{\mathrm{IC}}(\mathcal{P})=\sum_{C_j\in\mathcal{P}}(m\log\det R[C_j]+x|C_j|^2) \tag{2}$$

と定義し，これらを最小化する分割を採用することにする．これは，正規分布を仮定した時，それぞれ尤度を最大にし，情報量規準を最小にする変量の分割となっているためである．この分割は，真の分布が正規分布でなくとも，分割の一つの有力な候補になっていると考えられる．なお，ここで x は定数で，$x=1$ ならAIC，$x=\log m$ ならBICとなる．また，相関行列は変量のスケールの変換に不変であるため，式(1)や式(2)で分割を求める手法は，ロバスト性を有しているといえる．

ここでは，変量の分割を，葉に変量を対応させた階層的木構造で表現する．木構造の構成法には，葉を出発点とする集積型と根を出発点とする分割型が考えられる．

集積型アルゴリズムは，$\mathcal{P}_t=\{C_1,\cdots,C_t\}$ が得られているとき，式(1)や式(2)を最小にするように，2クラスターを併合して $\mathcal{P}_{t-1}$ を作ることを考える．このとき，単純に新たに登場する合併候補を全列挙し，LU分解などにより関数値を計算しても，$O(n^5)$ で木構造を得られるが，重複する主小行列式の計算を省略することによって，$O(n^4)$ で木構造が求められる．

分割型アルゴリズムは，$\mathcal{P}_t=\{C_1,\cdots,C_t\}$ が得られているとき，クラスター $C_j\in\mathcal{P}_t$ を X と $C_j\backslash X$ に分割して，$\mathcal{P}_{t+1}$ を作ることを考える．関数値をすべて求めて比較しようとすると分割は指数個存在するため工夫が必要となる．式(1)を最小化する分割問題の目的関数は対称劣モジュラ性を有しており，Queyranneのアルゴリズム[2]を用いることにより，多項式時間で最適解と最適値を求めることができ，$O(n^7)$ で木構造を求められる．さらに，重複する主小行列式の計算を工夫することで，木構造を得るまでの計算量を $O(n^6)$ とすることができる．これに加え，サイズの大きい主小行列の計算を保存することで，木構造を得るまでの計算量を $O(n^{5.5})$ とすることができる．式(2)を最小化する分割問題の目的関数は劣モジュラ関数とは限らず，別のアプローチか必要である．そこで，半正定値計画緩和と局所近傍探索のアプローチについて考察する．しかし，前者については未完だった．

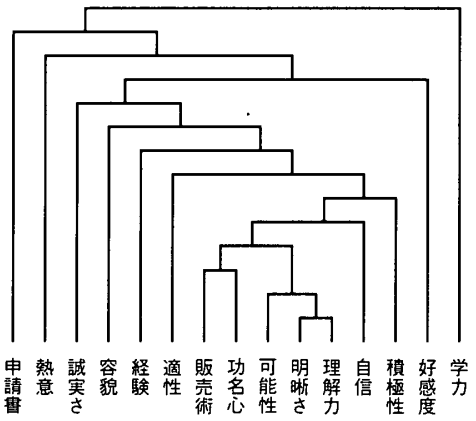

図1　最尤推定に基づくアルコリズムの出力

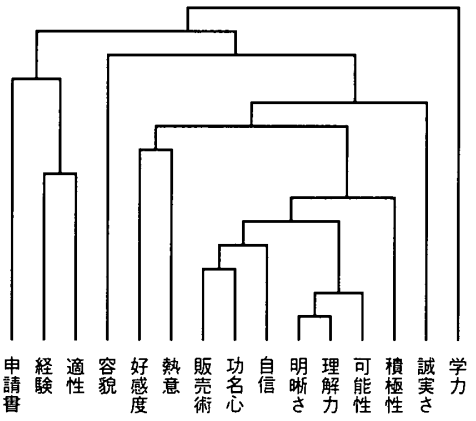

図2　AICに基つくアルコリズムの出力

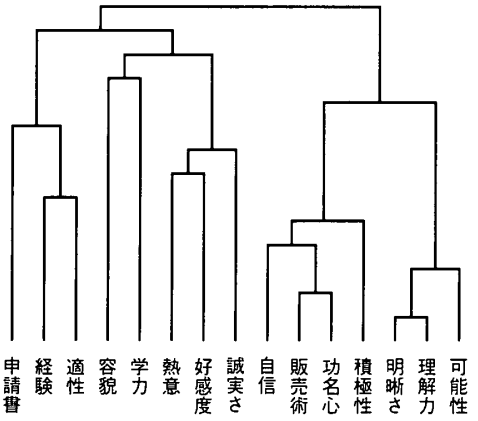

図3　BICに基つくアルコリズムの出力

3.　計算機実験

人工的に作成したデータおよび様々な実データを各アルゴリズムに入力し，アルゴリズムの性質の考察を行った．ここでは，その1例として，文献[1]に引用されている面接スコアの，評価項目の分析を紹介する．これは，48人の応募者を15個の評価項目（「申請書」「容貌」「学力」「好感度」「自信」「明晰さ」「誠実さ」「販売術」「経験」「積極性」「功名心」「理解力」「可能性」「入社の熱意」「適性」）を10点満点で評価したデータである．集積型アルゴリズムが出力した木構造を図1~3に示す．これらの木を比較すると最尤推定を基にしたアルゴリズムでは鎖効果が見られ分析の意味が薄れてしまった．また，AICが出力した木は鎖効果と思われる枝分かれが減少した．さらに，BICを基にしたアルゴリズムでは，大きさが平均的になるようにクラスターが生じやすいと観察された．

4.　まとめと今後の課題

本論文では，最尤推定やAIC，BICに基づく新しいクラスター分析法を提案した．また，最尤推定に基づく分割型アルゴリズムでは対称劣モジュラ関数最小化のアルゴリズムを適用できることを説明し，さらにその計算時間における2段階の改良を施したクラスタリング・アルゴリズムを提案した．さらに，情報量規準に基づく分割型アルゴリズムについて，半正定値計画緩和に基づく近似アルゴリズムのアイデアを提案し，局所近傍探索を用いたメタヒューリスティックスを構築した．最後に，これらのアルゴリズムを実装し，人工データ，実データで比較実験を行った．

未解決の課題としては，情報量規準に基づく分割型アルゴリズムが挙げられる．これに関連して，情報量規準に基づく分割型アルゴリズムに出てくる分割問題がNP困難なのかどうかも判定できていない．このことも残された課題であり，アルゴリズム構築のためには重要なポイントであると思われる．

参考文献

[1] M G ケンドール（奥野忠一，大橋靖雄共訳）：多変量解析　培風館，東京，1988

[2] M Queyranne　Minimizing symmetric submodular functions Mathematical Programming, Vol 82, pp 3-12, 1998