3E2-2

C-means クラスタリングアルゴリズムの現状とリスク解析への応用

Algorithms of c-means and applications to risk analysis problems

宮本定明[*1] 井口亮[*2] 水谷清隆[*3]
Sadaaki Miyamoto Ryo Inokuchi Kiyotaka Mizutani

[*1]筑波大学 University of Tsukuba
[*2]筑波大学 University of Tsukuba
[*3]筑波大学 University of Tsukuba

Clustering algorithms of c-means and competitive learning are described and two applications concerning risk analysis are considered. Hard and fuzzy c-means algorithms are formualted as an alternate optimization problem of a family of objective functions, while clustering using competitive learning uses a simple updating of reference vectors. Kernelized algorithms are also mentioned. First application uses a method of fuzzy c-regression, while second application employes a kernelized competitive learning algorithm.

1. はじめに

最近，様々な分野でクラスタリング技法が利用されるようになってきている．本稿では，c-means と競合学習によるクラスタリング，更にカーネル関数を利用したアルゴリズムについて簡単に述べ，リスク解析にかかわる応用について紹介する．

2. c-means と競合学習クラスタリング

まず，c-means[宮本 99] について簡単に説明しよう．クラスタリングされる対象集合（あるいは特徴ベクトル）を $x_1, \ldots, x_n \in \boldsymbol{R}^p$ とし，これらを c 個のクラスターに分けるものとする．各クラスターのプロトタイプを v_i とし，x_k と v_i との非類似性を測る測度を $D_{ki} = D(x_k, v_i)$ とする．D_{ki} の具体的な形は後で述べる．また，x_k のクラスター i に対する帰属度を u_{ki} と書く．古典的な場合，$u_{ki} = 0$ OR $u_{ki} = 1$ であり，その場合には，アルゴリズムを hard c-means と呼ぶが，やはり頻繁に使用される fuzzy c-means では，$0 \le u_{ki} \le 1$ である．

Hard/fuzzy c-means ともに，v_i と u_{ki} を繰り返し計算するアルゴリズムであるが，ある目的関数の交互最適化として表すことができる．便宜的に $U = (u_{ki})$, $V = (v_i)$ と書き，

$$J(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ki})^m D_{ki}. \tag{1}$$

ここで，m は $m \ge 1$ を満たす定数であり，$m = 1$ のとき hard c-means，$m > 1$ のとき fuzzy c-means に対応する．アルゴリズムは，適当な初期値を定め，V を直前の解に固定して $J(U,V)$ を U に関して最適化するステップと U を直前の解に固定して $J(U,V)$ を V に関して最適化するステップと収束するまで繰り返す．一般に，V についての最適化では制約条件はないが，U についての最適化では，制約条件として $0 \le u_{ki} \le 1$, all k, i, $\sum_{i=1}^{c} u_{ki} = 1$, all k, が課せられる．具体的な解の形式は省略するが，hard c-means において v_i がクラスターの中心を表す場合，v_i はクラスター内重心ベクトルとなり，u_{ki} は最近プロトタイプ分類となる．また，c-means には様々な変形があり，後に示す第一の応用例では，fuzzy c-regression が用いられている．その場合，プロトタイプは，そのクラスターに対応する回帰直線等であり，$D_{ki} = D(x_k, v_i)$ は，従属変数の値と，回帰から予測される値との二乗誤差が用いられる．なお，後の例では，エントロピー関数を利用した fuzzy c-regression を用いていることも付け加えておく．

競合学習によるクラスタリングでは，このような最適化を用いるのではないが，対象 x_k が中心 v_i で代表されるクラスターに最も近い場合，

$$v_i(t+1) = v_i(t) + \alpha(t)(x_k - v_i(t))$$

と学習パラメータ $\alpha(t)$ を用いて学習させ，同時に x_k をクラスター i に所属させることを繰り返す．

最近，サポートベクターマシン [Vapnik 98] における諸議論の影響もあって，カーネル関数を利用した c-means および競合学習クラスタリングが考察されている [Mizutani 05]．この技法では，高次元空間 S への写像 $\Phi(x)\colon \boldsymbol{R}^p \to S$ を用い，非類似度として $D(x_k, v_i) = \|\Phi(x_k) - v_i\|_S^2$ を利用する．カーネル関数による方法では，$\Phi(x)$ 自身は一般に未知で，代わりに内積 $K(x,y) = \Phi(x)\cdot\Phi(y)$ が知られていると仮定する．後に示す第二の応用例では，ガウス型カーネル $K(x,y) = \exp(-\lambda\|x - y\|^2)$ が用いられている．カーネルを用いる場合，v_i を明示的に求めることはできないが，$D(x_k, v_i)$ が $\Phi(x_k)$ に関する同次の 2 次式となることを利用すれば，u_{ki} の解と $D(x_k, v_i)$ の更新式が得られることがわかり，これらの計算を繰り返すことで，カーネルを利用した c-means のアルゴリズムが導出される [Miyamoto 02]．また，競合学習によるクラスタリングについては D_{ki} の更新は次式で行われる [Mizutani 05].

$$\begin{aligned} D_{ki}(t+1) = {} & (1-\alpha)D_{ki}(t) - \alpha(1-\alpha)d_{li}(t) \\ & + \alpha(t)(K(x_k,x_k) - 2K(x_k,x_l) + K(x_k,x_l)) \end{aligned} \tag{2}$$

詳しい議論は紙数の都合により省略する．

3. 喫煙とがんに関するデータ

図 1 は 1960 年のアメリカ 43 州とコロンビア特別区のがんに関する死亡者数と喫煙量のデータである．このデータを c-regression(c-回帰) を用いたクラスタリングで分析する．

図中の×，□，●，△印はそれぞれ，膀胱がん，腎臓がん，白血病，肺がんを表わしており，縦軸はがんによる死亡者数 (10 万人)，横軸は一人当たりの喫煙数 (100 本) を表わしている．

図 2 はエントロピーを用いた技法の一種である KL 情報量正則化ファジィ c-回帰法 (**K-FCR**) [宮岸 01] を適用した結果である．なお，他の c-回帰法ではこの結果は得られなかった．

この結果は図 1 と概ね一致しており，その他の手法よりもよい分類結果が得られたといえる．一般に，喫煙は肺がんのリ

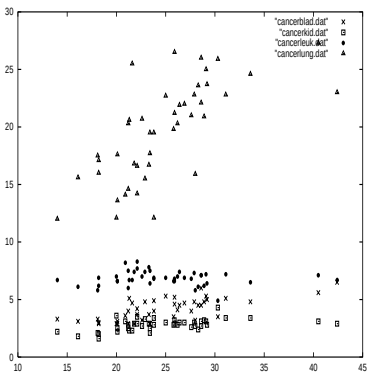

図 1: 喫煙データ

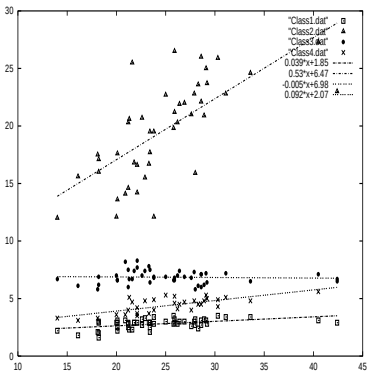

図 2: **K-FCR** の結果 ($\lambda = 35$)

スクを増大させることが知られているが，図の下部の 3 つのクラスのうち，中間に位置するクラスについては喫煙量との若干の相関を示しており，喫煙が肺がん以外のがんにも相関がある可能性を示している．

一般に，ファジィクラスタリングは，違うクラスター同士が密集した部分ではハード技法よりも相互の影響を受けず，安定した結果が得られるといえる．

4. 軍事データの解析

21 世紀に入り，国際社会の流れがさらにスピードアップしグローバル化しつつある中，世界は国際テロや大量破壊兵器，弾道ミサイルの拡散など，予測困難でグローバルな脅威にさらされるようになってきた．このような情勢の中，各国の軍事力の大きさを定量的に解析することは極めて重要である．

ここでは、2003 年の軍事データを用いてクラスタリングを行い，世界 157ヶ国をそれぞれの軍事力の大きさに応じて，5 つのグループに分類を行った例について示す．

クラスタリングに用いたデータは，各国の国防支出費並びに兵力，各種兵器の数などの 17 次元のデータを用いて解析を行った．また，クラスタリングアルゴリズムには，初期値依存性が少なくかつ変数間の単位が異なっていても全ての変数を考慮したクラスタリングを行うことができる，カーネル・競合学習クラスタリングアルゴリズムを用いた．

図 3 は，クラスタリング結果を国別に各クラスターごと色分けして，世界地図にプロットした図である．また，各クラスターの特徴を詳細に分析してみると，以下のことが判明した．

Cluster 1 ··· 軍事的能力が高い国
Cluster 2 ··· 軍事的能力が比較的高い国
Cluster 3 ··· 中間国
Cluster 4 ··· 軍事的能力が比較的低い国
Cluster 5 ··· 軍事的能力が低い国

すなわち，我が国の周辺地域（東アジア地域）は，Cluster 1 あるいは Cluster 2 に分類されている国が多く，世界的に見ても軍事的能力が高い国が集まっている地域であることが分かる．一方で，アフリカ諸国は軍事的能力が低い国である Cluster 5 あるいは Cluster 4 に分類されている国が多く，世界的に見て軍事的能力が低い国が集まっている地域であることが分かる．

このように，通常の軍事解析のようにただ単にデータを比較して分析するのではなく，クラスタリングを用いてそのデータの持つ特徴を見ていくことにより，各国の軍事力の大きさを定量的に解析できることが分かる．

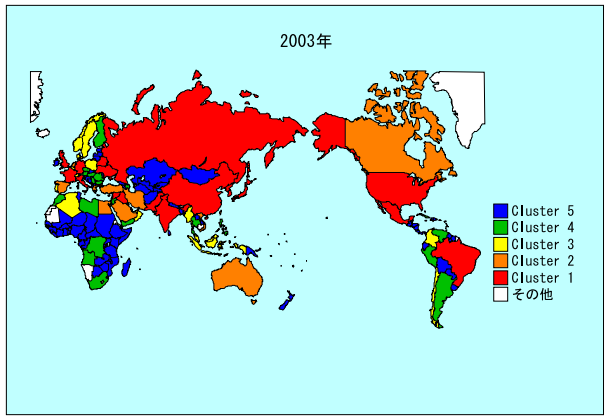

図 3: クラスターの世界地図上へのプロット

5. おわりに

ここでは，最近考察されている c-means と競合学習によるクラスタリングアルゴリズムを紹介し，リスク解析にかかわる応用例を示した．ガウス型カーネルを利用した技法は，クラスター分離性能が良く，注目されている．一方で，応用にかかわる考察はまだ不十分であり，他のデータ解析の諸技法との比較検討が必要である．

なお，本研究には，科学研究費（16300065）の補助を受けた．

参考文献

[宮岸 01] 宮岸，市橋，本多，日本ファジィ学会誌，Vol.13(4) pp. 406–417 (2001).

[宮本 99] 宮本定明: クラスター分析入門，森北書店 (1999).

[Miyamoto 02] S. Miyamoto, D. Suizu, *Proc. of FSKD'02*, Nov. 18-22, 2002, Singapore, Vol.2, pp. 656–660.

[Miyamoto 05] Miyamoto, S., T. Yasukochi, R. Inokuchi: *Proc. of 2005 IEEE International Conference on Systems, Man, and Cybernetics*, Waikoloa, Hawaii, Oct. 10-12, 2005, pp. 3221–3225.

[Mizutani 05] Mizutani, K., S. Miyamoto: *Proc. of FUZZ-IEEE2005*, May 22-25, 2005, Reno, Nevada, USA, pp. 636–639.

[Vapnik 98] V.N. Vapnik, *Statistical Learning Theory*, Wiley (1998).