カーネルファジィc-回帰法について

On Kernel Fuzzy c-Regression

◯ [1] 大井　祐介,
◯ [1]Yusuke Oi,
◯[1] 筑波大学大学院システム情報工学研究科
◯[1]Department of Risk Engineering, University of Tsukuba

[2] 遠藤　靖典
[2]Yasunori Endo
[2] 筑波大学システム情報系
[2]Faculty of Engineering, Information and Systems, University of Tsukuba

Abstract: The c-regression models are known to be useful for datasets with various correlations. To deal with the nonlinear datasets, the kernel hard c-regression was proposed. However, the method is weak against the datasets which includ noise or the real datasets. Therefore, we introduce fuzziness into the kernel hard c-regression and propose kernel fuzzy c-regression to overcome the above problem. Moreover, we verify the usefulness of the proposed method through numerical examples.

1　はじめに

情報通信技術の発達により，蓄積するデータの大規模・複雑化が進んでいる．このようなデータから有用な知見を発見することは非常に重要であるため，データ分析等の情報技術の重要度が増している．情報技術の一つとしてクラスタリングがある [1]．クラスタリングとは，外的基準なしに自動的にクラスタと呼ばれるデータ集合ごとに分類する方法であり，クラスタリングにファジィ理論を導入したファジィクラスタリングが提案されている [2, 3]．従来のハードクラスタリングでは，クラスタに帰属するか帰属しないかの 2 値で帰属度を定義していたが，クラスタリングにファジィ理論を導入したファジィクラスタリングでは連続値で帰属度を定めることができるようになり，クラスタへの帰属度を柔軟に表現することができるようになった．ファジィクラスタリングは様々なデータ解析に使用されるようになった．例えば，画像処理 [9] やバイオインフォマティクス [11] があげられる．

近年，クラスタリング等のソフトコンピューティング技法が用いられるようになっている分野の一つに発破工学がある [4]．発破の中でも鉱石採掘のための発破において，設計通りに鉱石採掘用の穴を掘ることは必要不可欠である．設計より大きな穴を掘ってしまうと，穴が崩落してしまい人命が危ぶまれる可能性がある．また，設計よりも小さい穴を掘ってしまうと設計通りの穴にするためにさらなるコストがかかってしまう．発破に必要なパラメータは，従来は人間の経験によって設定していたため，設計通りの穴を掘ることが難しいだけでなく，大きく予想を外れてしまうことも少なくはなかった．しかし，発破に用いるパラメータ群が掘る穴の大きさにどの程度貢献しているのか分かれば上記のコストを削減できる．そこで，今までに蓄積した発破の結果データを解析する研究が行われるようになってきている．

この研究において，相関関係ごとにデータを分類することは非常に有益である．クラスタリングにおいては，回帰分析と同時に，クラスタリングを行う方法として，c-回帰法が提案されており，様々な拡張が提案されている [5, 6]．特に，先に挙げたような実データの解析においては，ファジィクラスタリングのようにノイズに頑強である手法が有効と考えられる．しかしながら，非線形 c-回帰法のファジィ化については，従来法よりも高い効果が期待されるにも関わらず，いまだ十分な考察がなされていない．

このような背景から，本稿では実データを相関構造ごとに分類することを目的とし，従来より柔軟に相関構造を把握できると考えられるカーネルファジィc-回帰法を提案する．まず，今までに行われてきた c-回帰周辺の既存研究の紹介と新規手法のアルゴリズムの説明を行い，次に，新規手法の人口データと実データの数値例を示す．最後に結果の考察を行う．

2　既存研究

本章では，c-回帰とその周辺の既存研究について紹介する．以下，本稿では，説明変数を $x_k \in \mathbb{R}^p$，被説明変数を $y_k \in \mathbb{R}(k = 1, \cdots, n)$ とし，あらかじめ与えられた $c < n$ に対して，各クラスタ $C_i(i = 1\ c)$ が，以下に示す回帰式 $f_i(x_k, \alpha_i)$ $(i = 1, \cdots, c)$ によって表現されているとする．

$$y = f_i(x_k, \alpha_i) + e_i$$

ただし，α_i は回帰パラメータ，e_i は残差項を示す．
$f_i(x_k, \alpha_i)$ に様々な関数をとることが可能であるが，単純な線形回帰式で記述すると

$$f_i(x, \alpha_i) = \sum_{j=1}^{p} \alpha_{ij} x_j + e_i$$

となる．
また，データ x_k が C_i に属することを示す帰属度を u_{ki} とする．

2.1 ハード c-回帰

非階層的クラスタリングの代表的手法である c-平均法の拡張として，ハード c-回帰がある．基本的アルゴリズムを以下に示す.

Algorithm 1 ハード c-回帰

Step1.
初期クラスタをランダムに生成する.

Step2.
データごとにクラスタ (回帰式) との非類似度 $D_{ki}(k = 1, \cdots, n)(i = 1, \cdots, c)$ を計算し，最も D_{ki} が小さいクラスタ (回帰式) への帰属度を $u_{ki} = 1$，それ以外は $u_{ki} = 0$ とする.
ただし，$D_{ki} = \Big(y_k - f_i(x_k; \alpha_i)\Big)^2$

Step3.
α を以下により更新する.

$$\alpha = \arg\min J_{hcr}$$
$$J_{hcr} = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki} D_{ki}$$

Step4.
収束判定を行い，満たしていなければ **Step2** へ戻り，**Step2**, **Step3**, **Step4** を繰り返す.

2.2 ファジィ c-回帰

c-平均法の帰属度を，ファジィの概念を導入することによって連続値に拡張した手法がファジィ c-平均法であるが，同様に，ハード c-回帰にファジィの概念を導入した手法であるファジィ c-回帰法のアルゴリズムを Algorithm 2 に示す.

2.3 カーネル法

本来は線形モデルのクラスタリングをカーネル法を用いることで非線形モデルに拡張することができる [13]．カーネル法とは，カーネル関数を用いて，データを非線形空間に写像し，データ処理を行う総称である．カーネル関数の定義とそのいくつかの種類を以下に示す.

2.3.1 カーネル関数

入力空間であるパターン空間 $\mathbb{R}^p$ から高次元特徴空間 $\mathbb{R}^s$ への写像 Φ: $\mathbb{R}^p \to \mathbb{R}^s$ を考える．ここで，$x \in \mathbb{R}^p$ に対して，$\phi_l(x)(l = 1, \ldots, s)$ を用いて，非線形関数を

Algorithm 2 ファジィ c-回帰

Step1.
初期クラスタをランダムに生成する.

Step2.
データごとにクラスタ (回帰式) との非類似度 $D_{ki}(k = 1, \cdots, n)(i = 1, \cdots, c)$ を計算し，以下により u_{ki} を計算する.

$$u_{ki} = \left[\sum_{j=1}^{c}\left(\frac{D_{ki}}{D_{kj}}\right)^{\frac{1}{m-1}}\right]^{-1}$$

ただし，m はファジィ化パラメータと呼ばれ，$m > 1$ が使われる.
D_{ki} に関してはハード c-回帰と同様である.

Step3.
α を以下により，更新する.

$$\alpha = \arg\min J_{fcr}$$
$$J_{fcr} = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ki})^m D_{ki}$$

Step4.
収束判定を行い，満たしていなければ **Step2** へ戻り，**Step2**, **Step3**, **Step4** を繰り返す.

$\Phi(x) = (\phi_1(x), \ldots, \phi_s(x))$ と定義する．一般的には特徴空間上のベクトルの内積を表すカーネル関数 k を用いて計算を行う.

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

また，通常 ϕ は明示的に定義せず，カーネル関数 k を設定し，特徴空間への写像を行う．カーネル関数 k はいくつか存在するが，そのうち代表的なものを以下に示す [7].

$$\begin{aligned} k(x, x') &= x^T y \\ k(x, x') &= (1 + x^T x')^p \\ k(x, x') &= \exp(-\gamma \|x - x'\|^2) \end{aligned} \tag{1}$$

この中でも最も代表的なカーネルが式 (1) のガウシアンカーネルであり，本研究においてもガウシアンカーネルを使用する.

2.4 カーネルハード c-回帰

カーネル法を用いたハード c-回帰法であるカーネルハード c-回帰を以下で示す [8]．一般に，カーネル法を用いて回帰分析を行う際には，目的関数に罰則項を設けたリッジ回帰を導入するため，以下の目的関数 J_{khcr} を用いる：,

$$\begin{aligned} J_{khcr} &= \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}D_{ki} + \frac{1}{2}\sum_{i=1}^{c}\lambda_i\alpha_i^T K\alpha_i \\ &= \sum_{i=1}^{c}(y-K\alpha_i)^T U_i y - K\alpha_i) + \frac{1}{2}\sum_{i=1}^{c}\lambda_i\alpha_i^T K\alpha_i \end{aligned} \tag{2}$$

この目的関数を最小化するカーネルハード c-回帰アルゴリズムを示す．ただし，K はカーネル関数 $k(x,x')$ を用いて，以下の様に定義する．

$$K = \begin{pmatrix} k(x_1,x_1) & k(x_2,x_1) & \cdots & k(x_n,x_1) \\ k(x_1,x_2) & k(x_2,x_2) & \cdots & k(x_n,x_2) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_1,x_n) & k(x_2,x_n) & \cdots & k(x_n,x_n) \end{pmatrix}$$

Algorithm 3 カーネルハード c-回帰

Step1.
初期クラスタをランダムに生成する．

Step2.
データごとにクラスタ (回帰式) との非類似度 D_{ki} を計算し，D_{ki} が最も小さいときの i,k に対し，$u_{ki}=1$ とし，それ以外の場合 $u_{ki}=0$ とする．

$$D_{ki} = \Big(y_k - f_i(x_k,\alpha_i)\Big)^2$$

Step3.
α を以下により更新する．

$$\alpha_i = \left(\frac{1}{2}\lambda_i + U_i K\right)^{-1}(U_i y)$$

Step4.
収束判定を行い，満たしていなければ **Step2** へ戻り，**Step2**, **Step3**, **Step4** を繰り返す．

3 提案法

本研究における新規手法であるカーネルファジィ c-回帰のアルゴリズムを以下に示す．カーネルファジィ c-回帰は前章で紹介したカーネルハード c-回帰をファジィ化するものである．ファジィ化することで，カーネル c-回帰よりも柔軟なクラスタリングができると考えられる．この手法を実データに用いることで，線形・非線形構造をより詳細に解析することが可能になると考えられる．

3.1 アルゴリズム

目的関数 J_{kfcr} は

$$\begin{aligned} J_{kfcr} &= \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}^m(y_k - f_i(x_k,\alpha_i))^2 + \frac{1}{2}\sum_{i=1}^{c}\lambda_i\alpha_i^T K\alpha_i \\ &= \sum_{i=1}^{c}(y-K\alpha_i)^T U_i(y-K\alpha_i) + \frac{1}{2}\sum_{i=1}^{c}\lambda_i\alpha_i^T K\alpha_i \end{aligned} \tag{3}$$

Algorithm 4 カーネルファジィ c-回帰

Step1.
初期クラスタをランダムに生成する．

Step2.
u_{ki} を以下により更新する．

$$u_{ki} = \frac{(1/D_{ki})^{\frac{1}{m-1}}}{\sum_l^c (1/D_{lk})^{\frac{1}{m-1}}}$$

$$D_{ki} = \Big(y_k - f_i(x_k,\alpha_i)\Big)^2$$

Step3.
α を以下により更新する．

$$\alpha_i = \left(\frac{1}{2}\lambda_i + U_i K\right)^{-1}(U_i y)$$

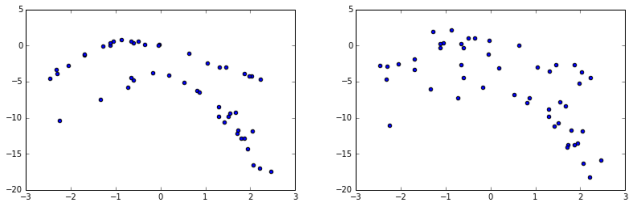
Step4.
収束判定を行い，満たしていなければ **Step2** へ戻り，**Step2**, **Step3**, **Step4** を繰り返す．

3.2 数値例

数値例を用いて既存手法 (KHCR) と本提案手法 (KFCR) を比較することでの特徴を評価する．評価手法として，ARI [12] を用いる．なお，各個体はメンバーシップ値の高いクラスタに割り当てることとする．

3.2.1 人工データ

図 1〜図 7 に数値例に用いる人工データを示す．図 1–

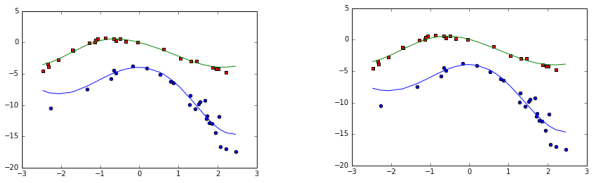


(a) ノイズを含まないデータ　(b) ノイズを含むデータ

図 1: 人工データ

(a) はノイズを含まないデータで図 1–(b) はノイズを含むデータである．

以上のデータセットを用いて，本提案手法と既存手法である KHCR を比較することで，本提案手法の特徴を考察する．クラスタリングの結果については，図 2,5 に，カーネルパラメータ γ 及び，回帰パラメータ λ に対する ARI の挙動については図 4,3,6,7 にそれぞれ示す．



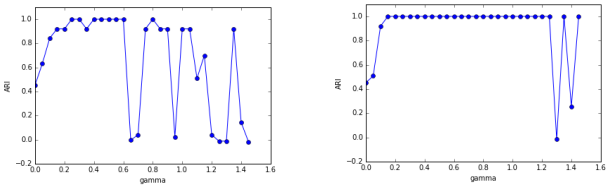
(a)KHCR
$\lambda = 0.70, \gamma = 0.45,$
ARI= 1.00

(b)KFCR
$m = 1.50, \lambda = 0.70,$
$\gamma = 0.35$, ARI= 1.00

図 2: ノイズを含まないデータに対する最も ARI の高い結果

3.2.2 実データ

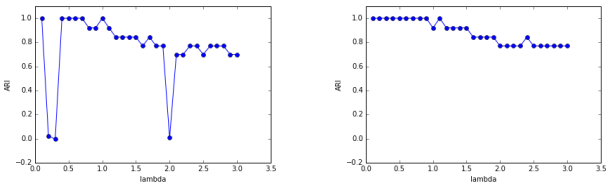
4ヶ国の GDP データのクラスタリング結果を図 8 に示す．KHCR,KFCR の結果とそれぞれのカーネルパラメータ γ に対する ARI の挙動，正則化パラメータ λ に対する ARI の挙動を，図 9, 10 に示す．



(c)KHCR
$\lambda = 0.7$

(d)KFCR
$m = 1.5, \lambda = 0.7$

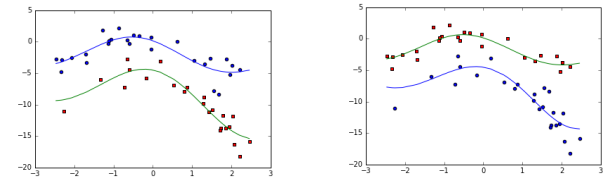
図 3: ノイズを含まないデータにおける γ に対する ARI の推移



(a)KHCR
$\gamma = 0.50$

(b)KFCR
$m = 1.5, \gamma = 0.50$

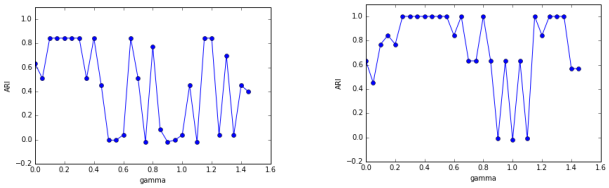
図 4: ノイズを含まないデータにおける λ に対する ARI の推移



(a)KHCR
$\lambda = 0.50, \gamma = 0.35$,ARI= 0.84

(b)KFCR
$m = 1.5, \lambda = 0.70, \gamma =$ 0.30,ARI= 1.00

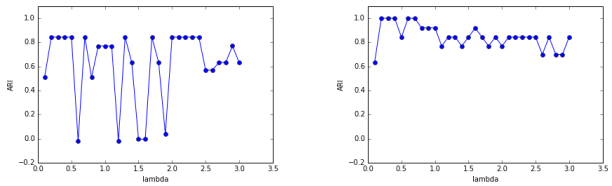
図 5: ノイズを含むデータにおける KHCR と KFCR の最も ARI の高いクラスタリング結果



(a)KHCR
$\lambda = 0.7$

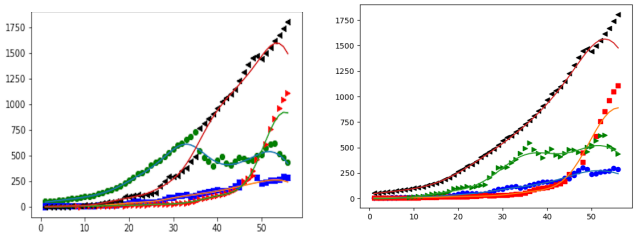
(b)KFCR
$m = 1.5, \lambda = 0.7$

図 6: ノイズを含むデータにおける γ に対する ARI の推移



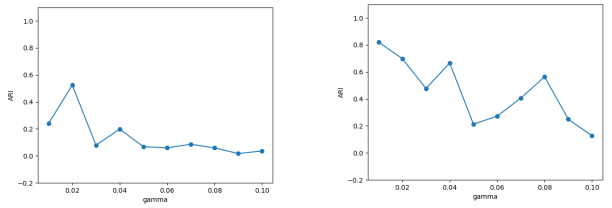
(a)KHCR $\gamma = 0.30$ (b)KFCR $m = 1.5, \gamma = 0.30$

図 7: ノイズを含むデータにおける λ に対する ARI の推移



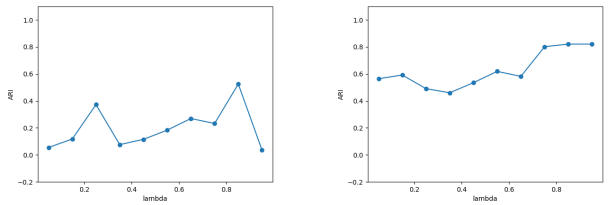
(a)KHCR $\lambda = 0.85, \gamma = 0.02$, ARI= 0.52 (b)KFCR $m = 1.50$, $\lambda = 0.85, \gamma = 0.01$, ARI= 0.82

図 8: KHCR と KFCR の GDP データでの最も ARI の高い結果



(c)KHCR $\lambda = 0.85$ (d)KFCR $m = 1.5, \lambda = 0.85$

図 9: カーネルパラメータに対する ARI の挙動



(a)KHCR $\gamma = 0.02$ (b)KFCR $m = 1.5, \gamma = 0.01$

図 10: 正則化パラメータに対する ARI の挙動

4 考察

人工データでは，曲線が交わらず比較的容易に分類できると予想されるデータと，ノイズを追加したデータに対して，KHCR と KFCR でクラスタリングを行なった．前者では，両者ともパラメータを調整することで ARI が 1 となる分類結果を得ることができた．しかし，後者では，KHCR では ARI が 1 となる結果を得ることはできなかった．これは，ファジィ化したことにより，モデルがノイズに対して頑強性を増したことが要因であると考えられる．さらに，正則化パラメータとカーネルパラメータのうち，1 つを固定し，もう片方のパラメータを変化させたクラスタリング結果の ARI を算出した．この結果を見ると，KFCR の方が KHCR よりも ARI のが安定していることが分かる．このことは，KFCR が比較的パラメータに対してロバストにクラスタリングすることができることを意味している．

実データでは，曲線が交わる，比較的分類の難しいデータに対して KHCR と KFCR でクラスタリングを行なった．結果として，両手法で AIR が 1 になる結果を得ることはできなかった． しかし，図 8 に示したように，KFCR では自然な分類結果を得ることができ，KHCR では KFCR と比較的不自然な結果となった．先に述べたパラメータへの頑強性に関しても同様の結果を得ることができたが，人工データよりも頑強性は低くなっている．

5 おわりに

本稿では，提案手法のアルゴリズム，人工データ・実データでの実行，結果に対する考察を行なった．数値実験では，ARI の観点で KFCR の性能が KHCR を上回る結果を得ることができた．さらに，パラメータに対する頑強性も高くなることが確認できた．今後の課題として，ファジィ化の方法等を変えた場合についての考察，非類似度の算出方法の変更が挙げられる．

参考文献

[1] 宮本定明: クラスター分析入門, 森北出版 (1999).

[2] Joseph C. Dunn : “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.”, Jounal of Cybernetics, pp.32–57 (1973).

[3] James C. Bezdek : “Pattern recognition with fuzzy objective function algorithms.”, Springer Science & Business Media (2013).

[4] Jang, Hyong Doo. : “Unplanned dilution and ore-loss optimisation in underground mines via cooperativeneuro-fuzzy network”, http://espace.library.curtin.edu.au/cgi-bin/espace.pdf?file=/2014/12/02/file_1/204927

[5] Yasunori Endo, Kouta Kurihara, Sadaaki Miyamoto, and Yukihiro Hamasuna : “Hard and Fuzzy c-Regression models for datasets with Tolerance in Independent Dependent Variables”, Fuzzy Syetems(FUZZ), 2010 IEEE International Conference on IEEE (2010).

[6] Sadaaki Miyamoto and Kenta Arai: “Different Sequential Clustering Algorithms and Sequential Regression Models”, Proceedings of 2009 IEEE International Conference on Fuzzy System (FUZZ-IEEE2009), pp.1107–1112 (2009).

[7] 前田英作, 村瀬洋: “カーネル非線形部分空間法によるパターン認識”, 電子情報通信学会誌 D, Vol.J82-D2, No.4, pp.600–612 (1999.4).

[8] Hengjin Tang, Sadaaki Miyamoto, and Yasunori Endo: “Semi-Supervised Sequential Kernel Regression Models with Penalty Functions”, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.19, No.1, pp.51–57 (2015.1).

[9] 原田洋子, 宮腰政明, 新保勝: “カラー画像セグメンテーションのためのファジィ・クラスタリング手法”, 日本ファジィ学会, Vol6, No.5, pp.1021–1036 (1994)

[10] Lawrence Hubert and Arabie Phipps. “Comparing partitions.” Journal of classification, Vol.2, pp.193–218 (1985).

[11] 有馬, 千夏:“DNA マイクロアレイデータ解析におけるファジィクラスタリングのクラスタ数推定法の開発”, 九州大学システム生命科学博士論文 (2008).

[12] Lawrence Hubert and Arabie Phipps. “Comparing partitions.” , Journal of classification, Vol.2, pp.193-218 (1985).

[13] Vladimir Vapnik. : “The nature of statistical learning theory.”, Springer science & business media (2013).

連絡先
大井祐介
筑波大学大学院システム情報工学研究科リスク工学専攻
遠藤研究室
E-mail: s1620558@u.tsukuba.ac.jp