k-means法による Generalized Aggregation Operatorを用いたクラスタリング

201011327 辻　陽介

経営工学主専攻　　指導教員：イリチュ（佐藤）美佳　教授

1.　研究目的

データ解析手法の一つであるクラスタリングは、個体間の類似構造に基づき、同じ特徴を持ったデータ同士に分類する手法であるが、二点間の類似性の定義は様々である。本研究で用いたクラスタリング手法の一つである k-means 法では非類似度に基づき分類を行うが、このときの非類似度には一般にユークリッド距離を用いる。

ユークリッド距離は内積を用いて表現可能であるが、この内積構造に対して Generalized Aggregation Operator(GAO)[3] を用いて、多様にある GAO の中で代表的な 3 つに対して結果の傾向を調査することを研究の目的とする。

2.　Aggregation Operator

t ノルム [1],[2],[4],[5] は、三角不等式を確率分布で表現するために導入されたものである。つまり、

$$Pr\{d(p,r) \le x+y\} \le Pr\{d(p,q) \le x,\ d(q,r) \le y\} \quad (1)$$

(1) 式において、$\pi_x = Pr\{d(p,q) \le x\}, \pi_y = Pr\{d(q,r) \le y\}$ とし、$T(\pi_a, \pi_b)$ で表すと、

$$Pr\{d(p,r) \le a+b\} \le T(\pi_a, \pi_b) \quad (2)$$

という関係式が得られるが、(2) 式における T が満たすべき性質を取り出したのが、t ノルムである。

t ノルムは以下のように定義される。

$[0,1] \times [0,1] \to [0,1]$ からなる演算 が以下の条件を満たすとき t-ノルムと呼ばれる。

(T.1)　$T(a,b) = T(b,a)$ (対称性)
(T.2)　$T(a,c) \le T(b,d)$
　　$whenever\ a \le b, c \le d$ (単調性)
(T.3)　$T(a,1) = T(1,a) = a$
　　$T(a,0) = T(0,a) = 0$ (境界条件)
(T.4)　$T(T(a,b),c) = T(a,T(b,c))$ (結合性)

Aggregation Operator(AO) は t ノルムの 4 つの条件のうち、3 つの条件（対称性、単調性、境界条件）を満足するものである。

3.　Generalized Aggregation Operator(GAO)

Generalized Aggregation Operator(GAO)[3] は、AO を線形空間上の直積空間上で定義した演算である。

GAO は以下のように定義される。

線形空間を X とすると、X×X→ $[0,1]$ からなる演算 $\tilde{\rho}$ が以下の条件を満たすとき、Generalized Aggregation Operator と呼ばれる。

($\tilde{A}$.1)　$\tilde{\rho}(\boldsymbol{a},\boldsymbol{b}) = \tilde{\rho}(\boldsymbol{a},\boldsymbol{b})$　　(対称性)
($\tilde{A}$.2)　$\tilde{\rho}(\boldsymbol{a},\boldsymbol{c}) \le \tilde{\rho}(\boldsymbol{b},\boldsymbol{d})$
　　$whenever\ \boldsymbol{a} \le \boldsymbol{b}, \boldsymbol{c} \le \boldsymbol{d}$　　(単調性)
($\tilde{A}$.3)　$\tilde{\rho}(\boldsymbol{a},\boldsymbol{1}) = \tilde{\rho}(\boldsymbol{1},\boldsymbol{a}) = \alpha \quad \alpha \in [0,1]$
　　$\tilde{\rho}(\boldsymbol{a},\boldsymbol{0}) = \tilde{\rho}(\boldsymbol{0},\boldsymbol{a}) = 0$　　(境界条件)

本研究で用いる GAO は以下の三つである。

Generalized Algebraic Product(GAP):

$$\tilde{\rho}(\boldsymbol{a},\boldsymbol{b}) = \boldsymbol{a}\boldsymbol{b}^t$$

Generalized Hamacher Product(GHP):

$$\tilde{\rho}(\boldsymbol{a},\boldsymbol{b}) = \frac{\boldsymbol{a}\boldsymbol{b}^t}{\boldsymbol{a}\boldsymbol{1}^t + \boldsymbol{b}\boldsymbol{1} - \boldsymbol{a}\boldsymbol{b}^t}$$

Generalized Einstein Product(GEP):

$$\tilde{\rho}(\boldsymbol{a},\boldsymbol{b}) = \frac{\boldsymbol{a}\boldsymbol{b}^t}{2K - (\boldsymbol{a}\boldsymbol{1}^t + \boldsymbol{b}\boldsymbol{1} - \boldsymbol{a}\boldsymbol{b}^t)}$$

また、GAO はいずれも以下の条件を満たしている。

$$\sum_{k=1}^{K} a_k = \sum_{k=1}^{K} b_k = \sum_{k=1}^{K} c_k = \sum_{k=1}^{K} d_k = 1, \qquad 2 \le K \le n \quad (3)$$

4.　GAO を用いたクラスタリング

k-means 法では非類似度にユークリッド距離を用いるが、ユークリッド距離は、内積を用いて表現できる。本研究では内積に GAO を用いたクラスタリングを行う。ユークリッド距離を GAO で表現すると、以下のようになる。

$$d(\boldsymbol{x}_i, \bar{\boldsymbol{x}}_c) = \{\tilde{\rho}(\boldsymbol{x}_i,\boldsymbol{x}_i) - 2\tilde{\rho}(\boldsymbol{x}_i,\bar{\boldsymbol{x}}_c) + \tilde{\rho}(\bar{\boldsymbol{x}}_c,\bar{\boldsymbol{x}}_c)\}^{\frac{1}{2}} \quad (4)$$

ここで、$\boldsymbol{x}_i$ は個体 i を表すベクトル、$\bar{\boldsymbol{x}}_c$ はクラスター c の重心を表している。(4) 式に対して、の GAO(GAP,GHP,GEP) を用いて、分類の結果を調査した。

5. 数値例

数値実験では人工データを使用した。まず、1 番目から 25 番目までの 2 変数のデータを平均 (4,6)、分散 1 の正規乱数で発生させ、これをクラスター 1 とした。次に、クラスター 2 として 26 番目から 50 番目までの 2 変数データを平均 (8,2)、分散 0.8 の正規乱数により作成した。この時 1 番目から 50 番目までのデータを $\boldsymbol{x}_1, ..., \boldsymbol{x}_{50}$ とし、i 番目 (i=1,...,50) のデータを $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ と表すとき、各データが変数について和が 1 であるという条件を満たすように次の変換を行う。

$$\tilde{\boldsymbol{x}}_i = (\tilde{\boldsymbol{x}}_{i1}, \tilde{\boldsymbol{x}}_{i2}), \quad \tilde{\boldsymbol{x}}_{ia} \equiv \frac{x_{ia}}{\sum_{b=1}^{2} x_{ib}}, \quad a = 1, 2 \qquad (5)$$

以上の変換を行ったデータに対して、GAO を用いたクラスタリングで解析すると、それぞれ図 1、図 2、図 3 の結果となった。この時一つの GAO に対する図の数は重心が変化しなくなるまでの更新回数を示している。

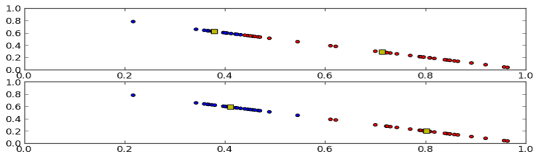

図 1　GAP によるクラスタリング

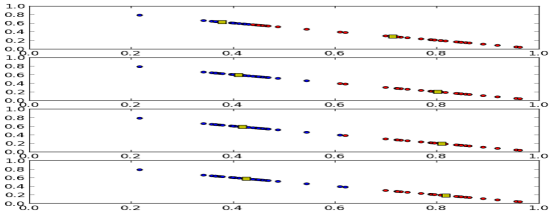

図 2　GHP によるクラスタリング

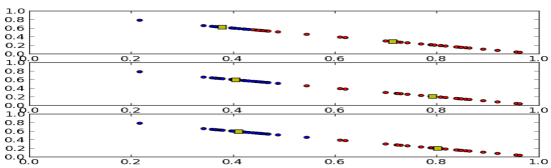

図 3　GEP によるクラスタリング

図 1、図 2、図 3 より、GHP、GEP、GAP の順番に重心の更新回数が多いことが分かり、さらに GHP はほかの二つとは違う分類の結果を示した。一方 GEP は GAP とは最終的に同じ分類の結果を示した。

次に、1 番目から 1000 番目までの 2 変数のデータを平均 (2.5,7.5)、分散 0.8 の正規乱数で発生させ、これをクラスター 1 とした。また、クラスター 2 として 1001 番目から 2000 番目までの 2 変数データを平均 (7.4,2.6)、分散 0.7 の正規乱数により作成した。初めの人工データと同様、(5) 式による処理を行い、GAO を用いたクラスタリングを行った。

結果は図 4 のように得られ、すべて同じ結果が得られた。

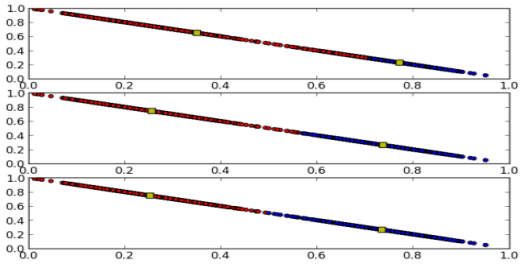

図 4　GAO によるクラスタリング

6. 結論

本論文では GAO を用いたクラスタリングについての研究を行った。k-means 法で用いられている非類似度で最も一般的なユークリッド距離を内積で表現し、GAO を用いることでそれぞれの GAO(GAP、GEP、GHP) の分類の相違性を調査したものである。

本研究の数値実験より、分類構造が明確なデータの場合は、GAP、GEP、GHP による結果に違いが生じるが、分類構造が不明確なデータの場合は、違いが生じないということがわかった。

参考文献

[1] E. P. Klement., R. Mesiar., E. Pap., *Triangular Norms*, Kluwer Academic, 2000

[2] K. Menger.,*Statistical Metrics*, Proc. Nat, Acad. Sci.,USA, vol.28, pp. 535-537, 1942.

[3] M. Sato-Ilic, Generalized Aggregation Operator Based Nonlinear Fuzzy Clustering Model, *Engineering Systems through Artificial Neural Networks*, vol. 20, pp. 493-500, 2010.

[4] B. Schweizer., A. Sklar., *Probablistic Metric Space*, Dover Publications, 2005

[5] 中島信之, t ーノルム, t コノルム通覧 (1)-概説と歴史, 日本ファジィ学会誌 Vol. 11, No. 4, pp. 561-576, 1999