k-means クラスタリングに用いるマハラノビス距離の非反復的推定法

† NHK 放送技術研究所　‡ 東京大学大学院 情報理工学系研究科　§ マンチェスタ大学

†‡ 小早川 健　† 木下 明徳　† 熊野 正　† 加藤 直人　† 田中 英輝

‡ 松崎 拓也　‡ 宮尾 祐介　‡§ 辻井 潤一

1　はじめに

分類範疇が未知の場合に、複数のデータの中から類似したものを同一クラスタに集める方法はクラスタリング [1] と呼ばれ、言語処理でもよく用いられる [2]。クラスタリングでは、類似したものを同一クラスタに集めるため、類似しているかどうかの判断に、類似度、つまり距離が用いられる。従来から、クラスタリングのための距離には、ユークリッド距離 [3]、または、コサイン距離を用い、距離尺度は固定するのが一般的であった。

分類問題では、古くから、教師あり機械学習法が研究されてきたのに対して、クラスタリングは教師なし手法が確立されていた。しかし、最近になって、クラスタリングにも教師あり手法が用いられるようになり、半教師ありクラスタリング [4] として研究がなされている。

よいクラスタリング性能を得る一つの方法は、よい距離尺度を設定することである。本研究では、k-means クラスタリング [3] において、マハラノビス距離を距離尺度に用いる場合のパラメータを推定することにする。先行研究 [5] では、文書クラスタリング問題で、マハラノビス距離のパラメータ推定を行う反復解法が提案されているが、本稿では、同じ問題設定で、それとは別の新たな非反復解法を提案する。

2　定式化

クラスタリングの対象となるデータは特徴量空間の中の点で表し、ベクトル表記を流用し太字で表す。データの添字の集合を $\mathcal{N}$ とし[*1] 、添字が n に対応するデータの特徴量は $\mathbf{x}_n$ で表す。クラスタの添字の集合は $\mathcal{K}$ とし、k はクラスタの添字とする。データがどのクラスタに属するかは r_{nk} で表し、添字 n のデータが添字 k のクラスタに属する場合は 1, それ以外の場合は 0 という 1-of-k 表記法を用いる。添字 k のクラスタの中心は $\boldsymbol{\mu}_k$ で表す。

クラスタリングは、目的関数を定義し、それを最小化する問題として定式化 [6] することができる。

2.1　データの平均

データの平均に関して、次の 2 つの量を定義する。

- クラスタ平均

$$\text{cluster-mean}(k; \mathbf{x}_n, r_{nk}) = \frac{\sum_{n\in\mathcal{N}} r_{nk}\mathbf{x}_n}{\sum_{n\in\mathcal{N}} r_{nk}} \tag{1}$$

- 大域的平均

$$\text{global-mean}(\mathbf{x}_n) = \frac{\sum_{n\in\mathcal{N}} \mathbf{x}_n}{|\mathcal{N}|} \tag{2}$$

2.2　ユークリッド距離によるクラスタリング [6]

$\mathbf{a}$ と $\mathbf{b}$ の間のユークリッド距離 $E(\mathbf{a}, \mathbf{b})$

$$E(\mathbf{a}, \mathbf{b}) = \sqrt{||\mathbf{a} - \mathbf{b}||^2} \tag{3}$$

を用いた目的関数 $J(r_{nk}, \boldsymbol{\mu}_k)$ を

$$J(r_{nk}, \boldsymbol{\mu}_k) = \sum_n \sum_k r_{nk} E(\mathbf{x}_n, \boldsymbol{\mu}_k)^2 \tag{4}$$

と定義する。与えられた $\mathbf{x}_n$ に対して $J(r_{nk}, \boldsymbol{\mu}_k)$ を最小にする $r_{nk}, \boldsymbol{\mu}_k$ を決定する問題となる。

この最小化は、局所的最小解が必ずしも大域的最小解とならないという問題 [1] を抱えているが、局所的最小化は r_{nk} による最小化と $\boldsymbol{\mu}_k$ による最小化を交互に繰り返し行う方法で解かれる。

[*1] 例えば、N 個のデータが $1, \ldots, N$ という添字を持てば、$\mathcal{N} =$ {1,. . . ,N } となる。

- r_{nk} による最小化
  異なる n に対して $\sum_k r_{nk} E(\mathbf{x}_n, \boldsymbol{\mu}_k)^2$ は独立なので、n 毎に J を最小化すればよい。

$$r_{nk} = \underset{r_{nk}}{\operatorname{argmin}} \sum_k r_{nk} E(\mathbf{x}_n, \boldsymbol{\mu}_k)^2 \tag{5}$$

  つまり、$E(\mathbf{x}_n, \boldsymbol{\mu}_k)^2$ を最小にする k に対して、$r_{nk} = 1$ とする。
- $\boldsymbol{\mu}_k$ による最小化
  J を $\boldsymbol{\mu}_k$ で微分すると

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = 2 \sum_n r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{6}$$

  となるが、これを 0 とおいて、

$$\boldsymbol{\mu}_k = \text{cluster-mean}(k; \mathbf{x}_n, r_{nk}) \tag{7}$$

  を得る。

2.3 マハラノビス距離によるクラスタリング

$\mathbf{a}$ と $\mathbf{b}$ との間のマハラノビス距離 $M(\mathbf{a}, \mathbf{b})$ は、対称正定値行列 A を用いて

$$M(\mathbf{a}, \mathbf{b}; A) = \sqrt{(\mathbf{a} - \mathbf{b})^T A (\mathbf{a} - \mathbf{b})} \tag{8}$$

で定義される距離である。目的関数は

$$J(r_{nk}, \boldsymbol{\mu}_k) = \sum_n \sum_k r_{nk} M(\mathbf{x}_n, \boldsymbol{\mu}_k; A)^2 \tag{9}$$

であり、2.2 節と同様に繰り返し法で最小化する。

- r_{nk} による最小化
  異なる n に対して $\sum_k r_{nk} M(\mathbf{x}_n, \boldsymbol{\mu}_k; A)^2$ は独立なので、n 毎に J を最小化すればよい。

$$r_{nk} = \underset{r_{nk}}{\operatorname{argmin}} \sum_k r_{nk} M(\mathbf{x}_n, \boldsymbol{\mu}_k; A)^2 \tag{10}$$

  つまり、$M(\mathbf{x}_n, \boldsymbol{\mu}_k; A)^2$ を最小にする k に対して $r_{nk} = 1$ とする。
- $\boldsymbol{\mu}_k$ による最小化
  J を $\boldsymbol{\mu}_k$ で微分すると

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = \sum_n r_{nk} (A + A^T)(\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{11}$$

$$= 2 \sum_n r_{nk} A (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{12}$$

  となるが、これを 0 とおいて、

$$\boldsymbol{\mu}_k = \text{cluster-mean}(k; \mathbf{x}_n, \boldsymbol{\mu}_k) \tag{13}$$

  を得る。A に逆行列がある場合は必要十分条件に、A に逆行列がない場合でも十分条件になっている。

2.4 マハラノビス距離の学習

目的関数は同じであるが、A と $\boldsymbol{\mu}_k$ の関数と見なし、

$$J(A, \boldsymbol{\mu}_k) = \sum_n \sum_k r_{nk} M(\mathbf{x}_n, \boldsymbol{\mu}_k; A)^2 \tag{14}$$

与えられたデータ $\mathbf{x}_n$ に対して、J を最小にする A と $\boldsymbol{\mu}_k$ を求める問題と見ると、学習データを最もよくクラスタリングするマハラノビス距離の学習と捉えることができる。

$J(A, \boldsymbol{\mu}_k)$ を A によって最小化することを考えると、A で偏微分して

$$\frac{\partial J(A, \boldsymbol{\mu}_k)}{\partial A} = \sum_n \sum_k r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \tag{15}$$

となる。右辺は、ベクトルの内積ではなく、共分散行列であることに注意する。対角成分について考えると、$\mathbf{x}_n = \boldsymbol{\mu}_k$ でない限り正であり、増大し続けるため、最小解が存在しない。いわゆる、不良設定問題 (ill-posed problem) になっている。

そこで、A の大きさを制限する拘束条件

$$\text{tr}(A^T A) = 1 \tag{16}$$

を導入し、マハラノビス距離の学習を定式化し直す。

マハラノビス距離学習の定式化

与えられた $\mathbf{x}_n, r_{nk}$ に対して、

$$\text{tr}(A^T A) = 1 \tag{17}$$

の下で、

$$J(A, \boldsymbol{\mu}_k) = \sum_n \sum_k r_{nk} M(\mathbf{x}_n, \boldsymbol{\mu}_k; A)^2 \tag{18}$$

を最小にする $A, \boldsymbol{\mu}_k$ を決定する。

これは拘束条件付最小化問題であるから、ラグランジュの未定乗数 α を導入して、

$$J'(A, \boldsymbol{\mu}_k) = J(A, \boldsymbol{\mu}_k) - \alpha(\text{tr}(A^T A) - 1) \tag{19}$$

を最小化する。

- $\boldsymbol{\mu}_k$ による最小化
  J' の $\boldsymbol{\mu}_k$ による偏微分

$$\frac{\partial J'}{\partial \boldsymbol{\mu}_k} = -(A + A^T) \sum_n r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{20}$$

  を 0 とおくと、

$$\boldsymbol{\mu}_k = \text{cluster-mean}(k; \mathbf{x}_n, r_{nk}) \tag{21}$$

  となる。

- A による最小化

$$\frac{\partial J'}{\partial A} = \sum_n \sum_k r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T - 2\alpha A \tag{22}$$

を 0 とおくと、

$$A = \frac{1}{2\alpha} \sum_n \sum_k r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \tag{23}$$

α は拘束条件から決定される。拘束条件より、

$$1 = \mathrm{tr}(A^T A) = \frac{1}{4\alpha^2} \mathrm{tr} \sum_{n,k} \sum_{n',k'} r_{nk} r_{n'k'} \cdot (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T (\mathbf{x}_{n'} - \boldsymbol{\mu}_{k'})(\mathbf{x}_{n'} - \boldsymbol{\mu}_{k'})^T \tag{24}$$

を得るが、これより、

$$2\alpha = \mathrm{tr}\Big(\sum_{n,k} \sum_{n',k'} r_{nk} r_{n'k'} \cdot (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T (\mathbf{x}_{n'} - \boldsymbol{\mu}_{k'})(\mathbf{x}_{n'} - \boldsymbol{\mu}_{k'})^T \Big)^{1/2} \tag{25}$$

を得る。

この方法では、マハラノビス距離に用いる行列 A の計算は、反復することなく求められることを指摘しておきたい。

3 文書クラスタリング実験

文書クラスタリングの実験を行い、提案法の性能を評価する。文書を形態素解析し、その出現 (bag-of-words) を特徴量に用いる。放送局に寄せられる意見を模擬的に集めたコーパスに対して、その意見の種類を類似した範疇に人手で分類したものを正解とする。

3.1 クラスタリングされた正解データ

実験用に、表 1 にある放送番組を 125 人の方に視聴していただき、自由記述で 1 件あたり 1〜2 文、複数件の感想を頂戴した。

	教育テレビ (ETV)
番組名	「知るを楽しむ」この人この世界 命の水を求めて
放送日時	2006.7.24 22:25〜22:50

表 1　反響の対象とした番組

集まった総反響文数は 915 文になり、人手によってクラスタリングされた。人手によるクラスタリングでは、表 2 にある 8 つのクラスタに分類され、それぞれのクラスタは範疇の欄にある特徴を持った文である。

範疇	全体	アンバランス
肯定的な意見	384	36
否定的な意見	35	7
番組を見て知ったこと	69	6
番組を見て考えたこと	260	34
番組への要望	95	11
番組への質問	15	2
その他の意見	22	4
意見でないもの	35	
合計	915	100

表 2　範疇と反響文数

意見でないものを除いてランダムに 100 文選んだものをアンバランスデータとする。クラスタ毎の反響文数の偏りが無いように、全範疇から同数の 15 文ずつ合計 120 文を選んだものをバランスデータとする。

3.2 クラスタリング性能の評価

クラスタリングの性能評価には様々な方法 [7] がある。精度 (accuracy) の測定には、すべてのクラスタ対応 (クラスタ数の階乗通りの可能性) を考慮する必要があるため、クラスタ数が多い場合は現実的でない [8]。純度 (purity) などは、その困難さを回避するために用いられている評価法であると考えられる。本実験では、クラスタ数が多くないため、厳密な精度の測定を採用する。すべてのクラスタ対応の中から、最良の精度を与える数値を精度とする。

3.3 マハラノビス距離の推定時におけるクラスタ中心

3.4 節で実験結果を示すが、定式化どおりの推定法が必ずしもよい結果をもたらさない。そのため、クラスタ中心を求める方法を必ずしも (21) 式の通りとはせず、下記の 3 通りの方法を比較する。

- 原点法
  すべてのクラスタの中心を原点とおく。つまり、

$$\boldsymbol{\mu}_k = 0 \tag{26}$$

- 大域的平均法
  すべてのクラスタの中心を教師データの大域的平均とおく。つまり、

$$\boldsymbol{\mu}_k = \text{global-mean}(\mathbf{x}_n) \tag{27}$$

- クラスタ平均法
  クラスタ毎に個別の中心を持つ。つまり、(21) 式。

3.4 定量的評価

従来法として、ユークリッド距離による k-means クラスタリングを行った。ユークリッド距離には学習すべきパラメータがないため、教師なしクラスタリングである。提案法として、学習データによってマハラノビス距離の推定を行い、そのマハラノビス距離による k-means クラスタリングを行った。マハラノビス距離の推定には、3.3 節の 3 つの方法を試した。なお、実験は学習データと評価データが同一である closed な結果である。

2.2 節で述べたように局所最適解に陥る可能性があるため、10 回の実験を行い、その平均値、標準偏差、最高値を求めた (表 3)。

データ種	バランス	アンバランス
従来法	28.33(± 2.27)	27.80(± 1.83)
	max. 32.50	max. 31.00
原点法	31.33(± 2.53)	30.70(± 1.90)
	max. 35.83	max. 34.00
大域的平均法	30.92(± 1.46)	29.80(± 1.72)
	max. 33.33	max. 32.00
クラスタ平均法	25.33(± 1.91)	27.90(± 2.26)
	max. 28.33	max. 31.00

表 3　クラスタリング精度 (%) とその誤差 (%)
上段最初の数字が平均値、括弧内が標準偏差、下段が最高値。

原点法でいずれも 2.9〜3.0 ポイントの性能改善が見られ、次いで大域的平均法で 2.0〜2.6 ポイントの改善が見られた。一方、クラスタ平均法では改善が見られなかったり、逆に劣化が見られた。

クラスタリングの収束に要する繰り返し回数は、原点法で最も多く、次に大域的平均法、そしてクラスタ平均法であり、クラスタ平均法と従来法でほぼ同数だった。しかし、繰り返し回数の差は有意ではなかった。

4 考察

クラスタ平均法が定式化に忠実でもっとも自然な距離学習に思える。しかし、実験によると、クラスタ平均法では性能改善がみられず、原点法が最善、大域的平均法が次善であるという結果が得られている。いずれの距離学習法も、見方を変えると特徴量選択の手法 [2] と似ているようにも思える。今後は、その関係を明かにしていきたい。

また、Bag-of-words 特徴量の共分散行列の計算は頻度に基づいている。今回は closed な実験であるため、頻度の計算でスムージングを行うことの効果は見られなかったが、今後 open 実験を行うときには、スムージングも併用して効果を検証したい。

半教師ありクラスタリングにも、マハラノビス距離を反復法で学習する方法 [5] が提案されている。そこでは、マハラノビス距離のパラメータを収束するまで最急降下法によるパラメータ更新が行われている。その手法との比較検討も行っていきたい。

5 おわりに

クラスタリングに用いられるマハラノビス距離を正解データから学習する方法を提案した。文書クラスタリングの実験によると、従来のユークリッド距離によるクラスタリングよりも性能改善する場合が見られた。最も性能改善するのは、クラスタ中心をデータの原点に固定する場合であった。また、提案法は、従来のマハラノビス距離の学習法と違って非反復的であることを特徴とする。

謝辞　正解データを作成して頂いた村上留美さんに感謝の意を表します。

参考文献

[1] A.K.Jain, M.N.Murty and P.J.Flynn: “Data clustering: A review”, ACM Computing Surveys, **31**, 3 (1999).
[2] M. W.Berry Ed.: “Survey of Text Mining: Clustering, Classification, and Retrieval”, Springer (2004).
[3] J. Mac ueen: “Some methods for classification and analysis of multivariate observations”, Proc. Fifth Berkeley Symp. on Math. Statis. and Prob., Vol. 1, Univ. of Calif. Press (1967).
[4] S. Basu, M. Bilenko and R. J.Mooney: “A probabilistic framework for semi-supervised clustering”, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 59–68 (2004).
[5] E. P.Xing, A. Y.Ng, M. I.Jordan and S. Russell: “Distance metric learning, with application to clustering with side-information”, Neural Information Processing System (NIPS) (2003).
[6] C.M.Bishop: “パターン認識と機械学習”, Springer (2006).
[7] C. D.Manning, P. Raghavan and H. Schütze: “Introduction to Information Retrieval”, chapter 16, Cambridge University Press (2008).
[8] 新納：“R で学ぶクラスタ解析”, 第 2 章, オーム社 (2007).