

ACO を用いたデータクラスタリングにおける最適パラメータの考察

Toward Optimal Parameters for ACO in Data Clustering

中山 貴幸^{*1} 水野 一徳^{*1}
Takayuki Nakayama Kazunori Mizuno

^{*1}拓殖大学工学部情報工学科

Department of Computer Science, Faculty of Engineering, Takushoku University

In data clustering, ACO based methods have been effective for cluster distributions with high accuracy. However, it has been difficult to set optimal parameters for each data set, requiring trial-and-error repetitions of parameter tuning. In this paper, we have analyzed the behavior of the ACO based algorithm, called ESACC, and clarified optimal parameters in ESACC for some benchmark datasets. We also demonstrate that ESACC with our found parameters can be more effective than some clustering method.

1. はじめに

クラスタリング (clustering) とは、分類対象の集合を、内的結合と外的分離が達成されるような部分集合に分割することである。統計解析や多変量解析、データマイニングといった分野において頻繁に利用されている。現在、その効率的な手法について、誠意研究がされている。

群知能の一種である蟻コロニー最適化 (Ant Colony Optimization: ACO) を取り入れたクラスタリングアルゴリズムとして ESACC が挙げられる。ESACC は、適切なパラメータ設定により精度の高いクラスタリングが実現可能であるが、パラメータ設定は非常に困難である。これは、パラメータ設定を複数回の試行や経験に基づいて行う必要があり、データによって適切なパラメータが変化するためである。

そこで、本稿では ESACC の挙動や傾向を解析することで、最適なパラメータの設定を行うにあたり考察を行う。具体的に、クラスタリング対象データの解析によるパラメータの自動設定、パラメータを動的に変更することによる最適化を行うことを視野に入れる。最終的には、パラメータをユーザーが設定する必要なく、高い精度のクラスタリングを行うことを目標とする。

2. 概要

2.1 クラスタリング

クラスタリング (clustering) とは、分類対象の集合を、内的結合と外的分離が達成されるような部分集合に分割することである。統計解析や多変量解析、データマイニングといった分野において頻繁に利用されている。

クラスタリングは、外的基準でなくデータの持つ特徴量を用いて行うため、教師なし学習である。よって、後述の目的関数等を用いて解候補を評価しつつ解探索を行う。

クラスタリングは大きく 2 つに分けることができる。階層型クラスタリングと非階層型クラスタリングである。前者は、データを階層的に分類することを目的とした処理であり、データウォード法、群平均法、最短距離法、最長距離法等、様々な手法が提案されている。後者は、データを非階層的、つまり階

層的な構造を持たず決められた数のクラスタにデータを分類することを目的とした処理である。本稿では後者を対象とし、以降クラスタリングと呼称する。

非階層型クラスタリングでは、 n 次元の特徴量を持つデータを K 個のクラスタに分類する。

各データ x とそれが所属するクラスタ i の中心 c_i との距離 $d(x, c_i)$ を算出し、その合計を目的関数とする。目的関数を小さくすることを目的とした組合せ問題として捉える。

目的関数を式 1 に示す。ただし、クラスタの総数を K 、クラスタ i を C_i 、クラスタ i の中心点を c_i とする。

$$f = \sum_{i=1}^K \sum_{x \in C_i} (d(x, c_i))^2 \quad (1)$$

2.2 ESACC

ACO は組み合わせ最適化問題の近似解を探索するために用いられるメタヒューリスティクスである。ACO では、アリを模したエージェントがフェロモンテーブルをもとに解候補を探索する。フェロモンテーブルにより良い解の情報を蓄積することで、より良い解探索を可能とする手法である。

ACO を用いたクラスタリングアルゴリズムとして ESACC が挙げられる。Algorithm1 に ESACC のアルゴリズムを示す。

ESACC は、SACC[Guluzar:2006] というアルゴリズムに変更を加えたものである。前世代で一番優秀なアリから解候補の一部を引き継ぐことで、形質遺伝といった遺伝的アルゴリズム (Genetic Algorithms:GA) の要素が取り込まれている。

使用するデータ構造を以下に示す。ただし、クラスタ数を K 、クラスタ対象のデータ数を n とする。

フェロモンテーブル：2 次元配列 (行数:k, 列数:n)

アリ (解候補)：配列 (要素数:n)

フェロモンの更新を行う際、目的関数値 (式 1) の逆数を、アリが持つ解候補をもとにフェロモンテーブル内に追加する。よって、目的関数が小さく優秀なアリは、より多くのフェロモンを堆積する。フェロモンを追加する処理は上位 L のアリだけ行われる。

例えば、データ x_a をクラスタ C_b に属する解候補を持つアリの目的関数値を示す値を f_c であるとする。その場合は、フェロモンテーブル内 a 行 b 列の要素に $1 \div f_c$ の値を追加する。

連絡先: 中山貴幸, 拓殖大学工学部情報工学科, 東京都八王子市館町 815-1, 042-665-0519, takanakahiko@gmail.com

Algorithm 1 ESACC

```

パラメータの設定
フェロモンテーブルの初期化
while 一定回数 do
  複数の解候補をフェロモンテーブルより生成
  解候補の評価
  フェロモンの堆積
  フェロモンの蒸発
end while
評価の一番高い解候補を最終的な解とする

```

フェロモンテーブルの値は毎世代、蒸発率に応じて値を減らされる。この操作を行うことで、フェロモンが飽和してしまうことを防ぐ。なお、フェロモンテーブルは 0 以上 1 未満の値を保持し、範囲を超えた場合は切り捨てを行う。

解候補の生成には、フェロモンテーブルと前世代で一番優秀であった解候補を利用する。解候補は、クラスタリング対象のデータ数の長さを持つ配列に格納される。まず、 E を E_{min} 以上 E_{max} 以下の整数よりランダムで決定する。 E の値は解候補生成時に毎回決め直す。新しく解候補を生成する際、一番優秀なアリが持つ解候補内のランダムな E 個の要素を、次世代のアリの解候補に用いる。それ以外の要素については、フェロモンテーブルより確率的に選択することで解候補を生成する。つまり、フェロモンテーブル内 a 行 b 列の要素の数値を正規化したものを、次の世代のアリがデータ x_a をクラスターに分類する確率と見立てる。

ESACC アルゴリズムでは、以下のパラメータ群を設定し使用する。

世代数

解候補の生成、評価、フェロモンテーブルの更新、といった一連の処理を繰り返す回数。

アリの数

1 世代につき用いる解候補の個数。

蒸発率

1 世代につき行うフェロモンの蒸発処理における蒸発量。

L

1 世代につき、フェロモンにテーブルに値を書き込むアリの匹数。アリの総数に対する割合を考慮し決定する。

E_{min}

解候補生成時に用いる E の最小値。クラスタリング対象データの総数に対する割合を考慮し決定する。

E_{max}

解候補生成時に用いる E の最大値。ラスタリング対象データの総数に対する割合を考慮し決定する。

ESACC は、適切なパラメータ設定により精度の高いクラスタリングが実現可能であるが、パラメータ設定は非常に困難である。

ESACC には、パラメータを決定するための指標が存在しない。よってユーザがパラメータ設定を複数回の試行や経験に基づいて行う必要がある。また、クラスタリング対象データによって適切なパラメータが変化するため、適宜変更する必要がある。

表 1: FM-index で使用するラベル

	flag2=True	flag2=False
Flag1=True	TP	FP
Flag1=False	TN	FN

表 2: 挙動解析実験に用いたパラメータ

	最小	最大	間隔	パターン数
世代数	1000	1000		1
アリの数	20	20		1
蒸発率	0.01	0.30	0.01	30
L	1	30	1	30
E _{min}	0	15	1	16
E _{max}	1	20	1	20

2.3 FM-index

クラスタリングの精度を計測する指標として FM-index (Fowlkes Mallows index) [Fowlkes:1983] が挙げられる。FM-index は二つのクラスタリング結果の同一性を求めるものであるが、クラスタリング結果と正解ラベルを比較することで、クラスタリングの精度を求めることが可能である。

クラスタリング結果 A とクラスタリング結果 B を比較する際、以下のように FM-index を求めることが可能である。まずデータ 2 つの組合せすべてに対し、以下の真偽値を求め、表 1 のようにラベルを付与する。

flag1: 結果 B にて同じクラスターである

flag2: 結果 B にて同じクラスターである

ラベリングされたデータの数に式 2 に代入することで FM-index を求めることができる。

$$FMI = TP \div \sqrt{(TP + FP) * (TP + FN)} \quad (2)$$

3. 挙動解析実験**3.1 実験内容**

ESACC について、データセット Iris [UCI] に対して表 2 に示すパラメータ群の全ての組合せを試行し、クラスタリングの結果を評価する。

クラスタリング結果は FM-index を用いた評価を行う。

パラメータ群とデータセットの組合せを 1 つ評価する行動を 1 試行とする。1 試行において同じパラメータにおいて実験した 10 回の評価値の平均を算出する。

3.2 結果

フェロモンの蒸発率と L 値の間は、FM-index に対して同じような相関があることが分かった。図 1 に、蒸発率と L 値を変更した組合せによる 81 回の試行について評価値を算出したものを示す。

E_{min} 及び E_{max} が異なる場合でも同様の相関関係を示した。つまり、フェロモンの蒸発率が決定された場合、それに対する最適な L 値が存在すると考えられる。FM-index 値が低い場合、フェロモンの堆積量と蒸発量の釣り合いが適切でない場合が多い。

		蒸発率									
		0.02	0.04	0.06	0.08	0.1	0.12	0.14	0.16	0.18	
L	2	0.83016	0.83103	0.77319	0.77004	0.71564	0.73729	0.694	0.67965	0.66059	
	4	0.75088	0.83677	0.80592	0.79052	0.77546	0.77502	0.74211	0.72765	0.72934	
	6	0.66475	0.81533	0.82325	0.81786	0.80635	0.78479	0.76873	0.76912	0.75869	
	8	0.61577	0.76582	0.81765	0.82195	0.79997	0.81086	0.78557	0.78165	0.7591	
	10	0.55722	0.68149	0.81515	0.81791	0.82157	0.80045	0.79891	0.75017	0.75207	
	12	0.52188	0.65771	0.73169	0.81148	0.81173	0.79966	0.79945	0.76932	0.75639	
	14	0.47242	0.61094	0.69838	0.76425	0.80146	0.77944	0.79561	0.78297	0.75283	
	16	0.41972	0.58838	0.65736	0.72401	0.76231	0.79488	0.75579	0.71994	0.71288	
	18	0.35855	0.46067	0.58149	0.66503	0.70338	0.71997	0.68973	0.67168	0.6422	

図 1: 蒸発率と L 値の組合せによる評価値

4. パラメータ設定とその評価

4.1 パラメータ設定

挙動解析実験の結果をもとに、パラメータの設定を行う。パラメータ設定にあたり、パラメータをタイプ A 及びタイプ B へと分類する。

タイプ A : 世代数, アリの数

タイプ B : 蒸発率, L, Emin, Emax

タイプ A のパラメータは、処理に与えられる時間やマシンスペックによって適切なパラメータを設定しなければならない。これらは、組み合わせ最適化問題を解く際におけるパラメータであり、ESACC 特有のパラメータではない。よって、今回のパラメータ設定の適用外とする。

タイプ B のパラメータは適切な設定をすることで、高い精度を期待することができる。よって、よって、今回のパラメータ設定の対象とする。

タイプ B に関して、図 1 に示した実験結果により、パラメータの設定が容易となったと考えられる。

蒸発率はフェロモンの総量の調整に用いる。目的関数が低くなるデータセットの場合は、フェロモンの堆積量が多くなる。よって、単純なデータセットの場合、蒸発率を高くする必要がある。データを解析し、特徴量の分散度を計算することで、蒸発率を決定することが可能である。

また、実験結果より L は蒸発率によって決定することが可能である。よって、蒸発率を適切に設定することで、適切な L を設定可能である。

Emin 及び Emax に関しては、前途の実験では関係性が明らかにならなかった。よって、グリッドサーチで決定するものとする。

4.2 評価実験概要

上記に示したパラメータ設定を用い、他のデータセットにおいても有効であるか検証を行う。

ここでは、UCI Machine Learning Repository[UCI] のデータセットである、Iris, Wine, Synthetic Control Chart Time Series Data Set(Sccts) を用いた。

Iris

アヤメのデータを品種毎にラベリングしたデータセットである。がく片の長さ、がく片の幅、花弁の長さ、花弁の幅、といった 4 つの特徴量を持つ 150 個のデータで構成されている。これらを 3 つの品種に分類を行う。特徴

表 3: 実験に使用したデータセット

	データ数	属性数	クラス数
Iris	150	4	3
Wine[UCI]	178	13	3
Sccts[UCI]	600	60	6

表 4: ESACC に用いたパラメータ

	蒸発率	L	Emin	Emax
Iris	0.1	10	2	6
Wine	0.1	16	5	10
Sccts	0.1	30	10	20

量、クラス数とともに少数であるため、比較的クラスターリングが容易なデータセットである。

Wine

白ワインのデータを品種毎にラベリングしたデータセットである。アルコール、リンゴ酸、灰、灰のアルカリ性、マグネシウム、フェノール類全量、フラバノイド、非フラバノイドフェノール類、プロアントシアニン、色彩強度、色調、蒸留ワインの OD280/OD315、プロリンといった 13 つの特徴量を持つ 178 個のデータで構成されている。これらを 3 つの品種に分類を行う。

Sccts

連続した数値のデータを傾向毎にラベリングしたデータセットである。有効数字 6 桁かつ 100 以下の有理数を 60 個持つ 600 個のデータで構成されている。これらを 6 つの傾向 (通常, 循環, 上昇, 下降, 上方移行, 下方移行) に分類を行う。

データセットの詳細を Table3 に示す。

実験に用いるプログラムは、以下の環境で実装を行なった。

OS : ubuntu16.04 LTS 32bit

CPU : Intel®Core™i7-4790 CPU

メモリ : 15.7GiB

プログラミング言語 : Python2.7

また、実装にあたり numpy ライブラリを使用し、並列処理や遅延評価を取り入れた効率的な計算を行った。

以下の 3 つのアルゴリズムに対して、それぞれ評価を行う。挙動解析実験と同様に FM-index を用いた評価を行い、1 試行において 10 回の評価値の平均を算出する。

1. ESACC:提案された論文 [Liu:2010] によるパラメータを用いた ESACC
2. ESACC-OPT:前章で示したパラメータ設定を適用した ESACC
3. k-means:クラスターリングにおける代表的なアルゴリズム

ESACC に用いたパラメータの一部を表 4 に示す。ESACC には、ESACC が提案された論文のパラメータをそのまま使用する。

表 5: ESACC-OPT に用いたパラメータ

	蒸発率	L	Emin	Emax
Iris	0.04	4	2	2
Wine	0.1	5	2	2
Sccts	0.09	17	10	20

表 6: 共通で用いたパラメータ

	世代数	アリの数
Iris	1000	20
Wine	1000	20
Sccts	1000	60

ESACC-OPT に用いたパラメータの一部を表 5 に示す。ESACC-OPT には、前章で示したパラメータ設定を施したパラメータを設定する。

なお、ESACC 及び ESACC-OPT に用いた共通のパラメータを 6 に示す。

4.3 結果

評価結果を図 2, 図 3, 図 4 に示す。

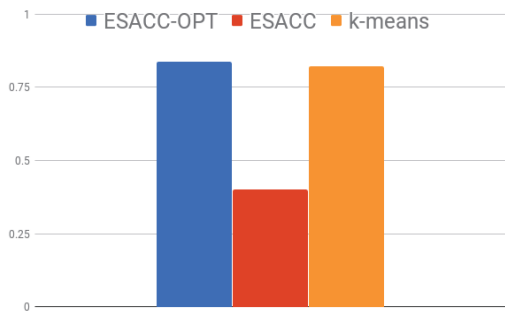


図 2: 各アルゴリズムの FM-index 値 (Iris)

3 つすべてのデータセットについて、ESACC と比較し ESACC-OPT が精度の向上を確認することができた。また、Iris と Wine のデータセットに関して、ESACC-OPT が k-means を上回る精度を得られた。

上記のパラメータ設定が、クラスタリング精度の向上に有効であることが確認できた。

4.4 考察

ESACC に比べ ESACC-OPT の精度がすべてのデータセットにおいて高かったという実験結果により、パラメータ設定がより正しい値を設定できたことがわかった。このことにより、パラメータ設定が適切であることが考えられる。

また、k-means と比較した結果においては、Sccts において精度を上回ることができなかった。属性数及びクラスタ数が多い場合を想定し、蒸発率や L の範囲を広げた挙動解析実験を行うことで、Sccts に対応したパラメータ設定を行うことができると考えられる。

5. 今後の展望

本稿では、挙動解析に基づいたパラメータ設定の有効性の検証を行った。しかし、最適なパラメータを自動的に設定可能

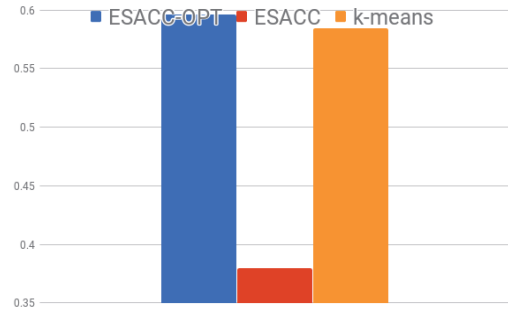


図 3: 各アルゴリズムの FM-index 値 (Wine)

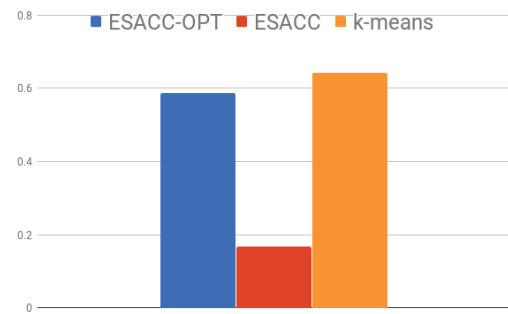


図 4: 各アルゴリズムの FM-index 値 (Sccts)

とする仕組みを確立するに至らなかった。

今後の研究において、以下の方法を検討し、正確かつ自動的なパラメータ決定を行う手法を検討する。

- パラメータ設定の数式化
- E_{min} 及び E_{max} の自動的な決定

さらに、手書き数字画像等、他のデータセットに利用することで、様々なデータのクラスタリングに有効であることを示すとともに、実際のシステムで適用可能であることを示す。

参考文献

- [Guluzar:2006] KEKE, Guluzar, N. Yumusak, and Numan ELEB.: Data Mining and Clustering With Ant Colony., Proceedings of 5th International Symposium on Intelligent Manufacturing Systems. (2006).
- [Liu:2010] Liu, Xiaoyong, and Hui Fu.: "An Effective Clustering Algorithm With Ant Colony." JCP 5.4 (2010): 598-605.
- [Fowlkes:1983] Fowlkes, Edward B., and Colin L. Mallows.: "A method for comparing two hierarchical clusterings.", Journal of the American statistical association 78.383 (1983): 553-569.
- [UCI] UCI Repository for Machine Learning Databases , <http://archive.ics.uci.edu/ml/>