F-031

音環境理解のための音データ分類に関する一手法
An affinity propagation clustering method of sound data for computational auditory scene analysis

河本 満† 幸島明男† 車谷浩一†
Mitsuru Kawamoto Akio Sashima Koichi Kurumatani

1. はじめに

大量のデータから意味のある情報を抽出し、抽出した情報をサービスなどに活かす試みは近年盛んに行われるようになってきている。一方で、データはあるがどのように分析・解析してよいか困っているという声も多いようである。[4]

本稿では、データ分析に必要になるであろうデータ分類の手法を提案する。提案手法では、ディリクレ混合過程モデルによるデータ分類機能に代表データ(Exemplar)を使うメッセージ交換型の Affinity Propagation（AP 法）を適用し、それらを交互に動作させることによって分類を実行する。AP 法においては、データ間の類似度によって Exemplar に付属するデータが決まるが、Exemplar となるデータはデータ自身の自己相関値と周りのデータの類似度との関係で決まる。このとき、自己相関値の設定は自由度があり、どのような分類結果を求めるかで決まってくる[3]。また、ディリクレ混合過程モデルによる分類に関しては、Exemplar に相当する中心値の個数は、与えられたデータに応じて定まるが、ごくごく小さなクラスタとして判断されるようなデータがあれば、ほぼ必然的にそれをクラスタリングしてしまう傾向にあり[1]、結果として得られるそれぞれのクラスタは、データ間の類似度による構成の観点から見ると、バラバラなクラスタリング結果となってしまう。

本提案手法では、ディリクレ分布により、初期に定められた中心データを AP 法の Exemplar 候補として取扱い、分類し、その結果をディリクレ混合過程モデリングに活かす。この操作を交互に行うことによって、ある程度類似度が共通したデータが集まったクラスタリングを実現する。

提案手法の特徴は、他のクラスタリング手法（k-means 法、k-medoids 法）とも比較することにより示される‡。

2. 提案手法

2.1 ディリクレ混合過程

ノンパラメトリックベイズモデル[5]であるディリクレ過程(Dirichlet Process; DP)では、観測データ xi を生成する要素分布 p(xi|θi)のパラメータθi が xi 毎に対応づけられている。つまり、データ数分の n 個のパラメータを用いた最大 n 混合モデルまで実現できる。しかしながら、1 つのデータに 1 つの要素分布を割り当てるのは非現実的であるので、DP では、データが観測される毎に必要に応じて要素分布数が増える柔軟なデータ生成過程となっている。つまり、θ1,…,θn がすべて異なるのではなく、観測データに適応した異なる K 個のパラメータθ(1),…,θ(K)からなる。このとき、観測データ x1,…,xn は有限個（K 個）のクラスタに分割されることになる。DP の場合、K の値は予め固定ではなく、観測データに応じて定まる。この性質を利用して混合モデルを構成するモデリングがディリクレ混合過程(Dirichlet Process Mixture: DPM)モデルである[2]。従って、観測データを用いて DPM モデルを生成することによって、その観測データに応じたクラスタ分類ができることが分かる。

† 国立研究開発法人産業技術総合研究所人間情報研究部門
† National Institute of Advanced Industrial Science and Technology (AIST), Human Informatics Research Institute

‡ 本稿では、ページ制限の都合上、計算機で作成したデータによる実験結果のみの掲載とする。

2.2 Affinity Propagation

Affinity Propagation（AP）法は、観測データ間の類似度 $s(i,k)$に応じて、以下の式を使って代表データ(exemplar)とそれに類似する観測データを求めることができるクラスタリング法である。ただし、$s(i,k)$は観測データ xi と xk の類似度を表し、類似度が高い程、大きな値を持つように設定される。

$$r(i,k) \leftarrow s(i,k) - \max_{k' s.t. k' \neq k}\{a(i,k') + s(i,k')\} \quad (1)$$

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i' s.t. i' \notin i,k} \max\{0, r(i',k)\}\} \quad (2)$$

ここで、$r(i,k)$は exemplar 候補観測データ xk が観測データ xi の exemplar になるのがどのくらい適切かを示すもので、観測データ xi から exemplar 候補観測データ xk に送られるメッセージと考えている。また、$a(i,k)$は、観測データ xi が exemplar 候補観測データ xk のクラスタのメンバーになるのがどれくらい適切かを表すもので、exemplar 候補観測データ xk からクラスタメンバーとなる見込みの観測データ xi に送られるメッセージとなる。このとき、$a(k,k)$に関しては、以下の式を用いて更新する。

$$a(k,k) \leftarrow \sum_{i' s.t. i' \notin i,k} \max\{0, r(i',k)\}\} \quad (3)$$

つまり、AP 法は、各観測データ間でメッセージの交換を繰り返して最終的に $r(i,k)+ a(i,k)$の値で評価し、exemplar を決定、クラスタを生成するメッセージ交換型クラスタリング手法である。また、AP 法においても、クラスタ数は、与えられた観測データに応じて定まる。ここで、$a(i,k)$の初期値はゼロに設定される。

2.3 DPM+AP

AP 法では自己相関にあたる $s(k,k)$の値によってクラスタ数が変化することが知られている[3]。適切なクラスタ数は観測データの種類によって異なる。本研究では、DPM モデルを生成する過程において、観測データの分布に合った要素分布のパラメータθ(k)を用いて、$s(k,k)$を決め、AP 法を動作させ、その結果、得られた exemplar 情報を要素分布のパラメータθ(k)を更新する情報に用いて、更新結果を DPM モデルのモデリングにフィードバックする。これを繰り返すことにより、できるだけ観測データの分布を顧慮したクラスタリングを実現するクラスタリング手法を提案する。

Copyright © 2015 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

3. 実験結果

提案手法の有効性をいくつかのクラスタリング手法と比較することによって示す。クラスタリングを行うデータは、[-5,5]の範囲の一様分布からランダムに出力される二次元平面上の値(x,y)を中心に 10 点のデータを 1 つのクラスタとし、そのクラスタを 30 個分布させた二次元平面のデータを正解のクラスタデータとした（図 1 参照）。

3.1 DPM モデルとの比較

DPM モデルのみで図 1 のデータをクラスタリングした結果を図 2 に示す。図 2 では、正解のクラスタ数よりも少ないクラスタ数(19 個)となってしまい、クラスタリングがうまく実行されていないことが分かる。しかしながら、提案手法では、正解のクラスタ数に近いクラスタリング結果(クラスタ数：29 個)となっており、正解クラスタ分布ともほぼ類似したクラスタリング結果となっている（図 3 参照）。このことから、本提案手法(DPM+AP)は、DPM モデルの特徴である観測データの分布に合った要素分布のパラメータθ(k)を見つける能力を AP 法のクラスタリング能力を使うことによって最大限に活かせる手法になっていることが分かる。ここで、この傾向は、正解分布データをいくつか変更したとしても変わらないことを確認している。

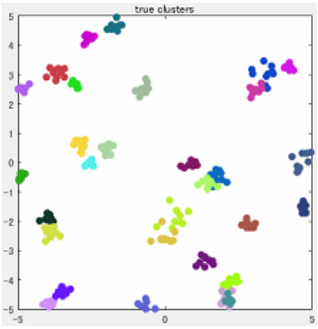

図 1 正解クラスタ分布

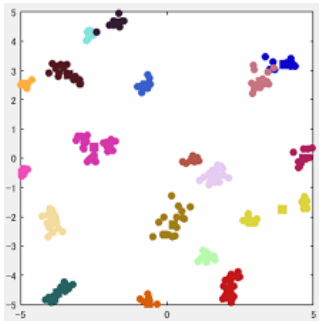
図 2 DPM でのクラスタリング

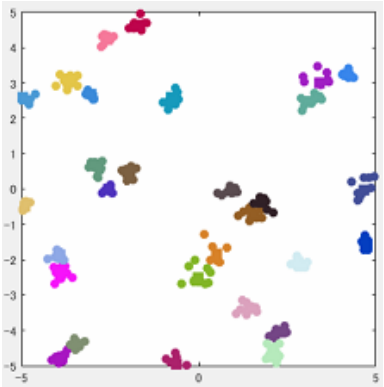
図 3 DPM+AP 法でのクラスタリング

3.2 k-means, k-medoids 法との比較

k-means 法、k-medoids 法は共に、クラスタ数を予め与えることによってクラスタリングが実行される方法である。このことから、両手法のクラスタリング結果は、本提案手法で推定したクラスタ数を両手法に与えることによって出力した。図４、５にそれぞれの結果を示している。図３，４，５を比較してみると、本提案手法は k-medoids 法と似たようなクラスタリングが実現できることが分かる。k-means 法は、クラスタ内のデータと代表点との距離の平均値を算出した場合、k-means 法の方が長いことから、本提案手法と比較して効率の良いクラスタリングをしていないことが分かる（表 1 参照）。従って、提案手法は k-medoids 法と似たようなクラスタリング能力を持っており、k-means 法よりは効率の良いクラスタリングが実現できるという特徴を持っていることがいえる。ここで、本提案手法は、クラスタ数が未知であってもクラスタリングが実行可能であることから、正解クラスタ数が未知な実データに対して有効であるということに注意されたい。

表 1 それぞれのクラスタの代表点からのクラスタ内のデータ点までの距離の平均値の一例

本提案手法(DPM+AP)	k-means 法	k-medoids 法
0.04447	0.060944	0.046031

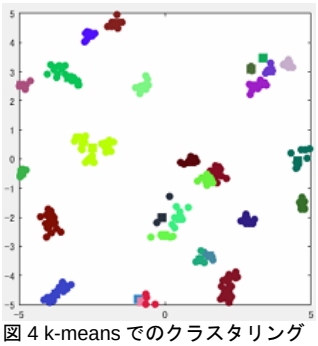
図 4 k-means でのクラスタリング

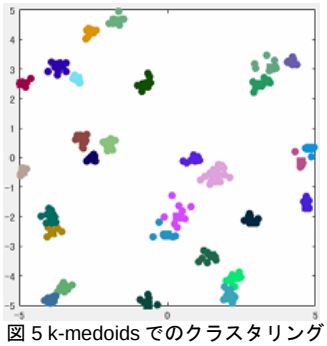
図 5 k-medoids でのクラスタリング

4. おわりに

ディリクレ混合過程モデルと Affinity Propagatoin 法を組み合わせたクラスタリング手法を提案した。本提案手法では、ディリクレ混合過程モデル単独で実行するよりも正解クラスタ数を推定する能力があり、k-means 法より効率の良いクラスタリングを実現し、k-medoids 法と類似のクラスタリング能力を持っていることが分かった。

今後は、実データ(環境音データ)に対してクラスタリングを実行し、その結果を音環境理解に用いるモデル作成に利活用することを考えている。

謝辞

本研究は、科研費(＃25330379)の援助を受けた。

参考文献

[1] J.W.Miller and M. T. Harrison, “A simple example of Dirichlet process mixture inconsistency for the number of components”, NIPS2013, (2013), http://papers.nips.cc/paper/4880-a-simple-example-of-dirichlet-process-mixture-inconsistency-for-the-number-of-components.

[2] C.E.Antoniak, “Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems”, Annals of Statistics, 2, 1152-1172, (1974).

[3] B.J.Frey and D.Dueck,”Clustering by passing message between data points”, Science, 315, 972-976, (2007).

[4] “平成 25 年度我が国経済社会の情報化・サービス化に係る基盤整備(地域ビジネスの振興支援に資するデータプラットフォーム構築とベンチマーキング手法開発に関する調査研究事業)”, 成果報告書 http://www.meti.go.jp/meti_lib/report/2014fy/E004195.pdf.

[5] 上田修功, 山田武士,”ノンパラメトリックベイズモデル”, 日本応用数理学会, Vol.17, No.3, pp.196-214, 2007.

Copyright © 2015 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.