F-049

最適木構造クラスタリングにおけるクラスタアンサンブルの適用
Applying Cluster Ensemble to Adaptive Tree Structured Clustering

山口 崇志† 野口 裕貴† 市村 匠‡ マッキン ケネス J. †
Takashi Yamaguchi Yuki Noguchi Takumi Ichimura Kenneth J. Mackin

1. はじめに

近年、ネットワーク技術の発展や記憶媒体の低コスト化に伴い、大規模なデータを蓄積することが可能になった。これら大量のデータを知識として利用する為、機械的にデータから知識を抽出する技術が注目されている。このような技術の一つで、特にデータ解析の足がかりとして用いられるクラスタリングは、既知の分類カテゴリ等を用いず、データの特徴に基づき分割する手法である。

Kohonen によって提案された自己組織化マップ(SOM) [1]はデータの件数に依存しない逐次的な学習が可能で、可視化による直感的な結果提示に優れたクラスタリング手法の一種である。一方で、SOM の提示するクラスタリング結果は人間の視覚による判断に依存しており、そのクラスタの境界が曖昧である。

過去の研究において SOM の結果の曖昧性を解決する為、SOM と凝集型階層クラスタリング(AHCA)を応用して再帰的にデータを2分割する最適木構造クラスタリング(ATSC) [2]を提案した。ATSC はデータの件数に関わらず逐次的にデータを分割することが可能である一方、パラメータや初期値、クラス間距離関数により分割結果にばらつきが生じる問題が確認された[3]。

本研究では、ATSC の2分割毎に A. Strehl らによって提案されたクラスタアンサンブル[4]を応用し分割結果の安定性を改善する。また提案手法を用いることにより、2分割毎に最適なクラス間距離関数を用いる為、従来の階層クラスタリングでは得られなかった分割結果が期待される。

2. 最適木構造クラスタリング

ATSC は逐次処理が可能な分割型階層クラスタリング(DHCA)の枠組みである。ATSC では図1のように入力データを ATSC ノードによって再帰的に 2 分割することで最終的なクラスタリング結果と木構造を得る。ATSC ノードは(1) SOM による量子化、(2) 量子化されたデータの AHCA によるクラスタリング、(3) 子ノード生成の3つの処理ステップからなる。ステップ(1)と(2)により ATSC ノード $NODE$ へ入力された n 個の m 次元ベクトルの集合 $A = \{ \boldsymbol{x}_i \mid \boldsymbol{x}_i \in \mathrm{R}^m \}$, $i = 1, 2, \ldots, n$ を直和な部分集合 A_0 と A_1 に分割する。次に子ノード生成ステップにおいて分散の減少 $\Delta E(A)$ が閾値 θ より大きい時（式 1）、二つの子ノード $NODE_0$ と $NODE_1$ を生成し $NODE_0$ の入力に A_0、$NODE_1$ の入力に A_1 を与える。

$$\Delta E > \theta \qquad (1)$$

$$\Delta E = E(A) - E(A_0) - E(A_1) \qquad (2)$$

$$E(A) = \sum_{\mathbf{x}_i=1}^{n} \left\| \bar{\mathbf{x}} - \mathbf{x}_i \right\| \qquad (3)$$

なお $\bar{\mathbf{x}}$ は A の重心である。つまり ATSC は分散の減少が閾値 θ 以下になるまでデータの分割を繰り返し、最終的なクラスタと木構造を得る。

† 東京情報大学 情報システム学科
‡ 広島市立大学大学院 情報科学研究科

ATSC は SOM の分類性能を損なうことなくクラスタを明確化しデータ間の潜在的な階層関係を木構造として抽出する。また、抽出される木構造はノードでの処理における AHCA のクラス間距離関数に依存し、一般的な AHCA での空間圧縮や空間拡散といった特性をそのまま継承する。一方で、パラメータや初期値、クラス間距離関数により分割結果にばらつきが生じる問題が確認されている。初期の分割における誤差が以降の分割において悪影響を及ぼす問題は、ATSC に限らずDHCA全般における本質的な問題点である。

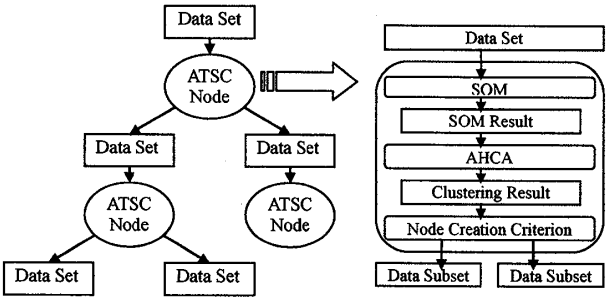

図 1 ATSC モデル図(左) ATSC ノードモデル図(右)

3. クラスタアンサンブル

クラスタアンサンブルは異なるクラスタリング結果（弱クラスタ）からより良いクラスタリング結果（強クラスタ）を抽出する手法の非常に抽象的な枠組みである[4][5]。基本的なクラスタアンサブルでは、r 個のアンサンブル $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}_p\}$, $p = 1, 2, \ldots, r$ を用いて入力データ $X = \{x_1, x_2, \ldots, x_n\}$ をクラスタの集合族 $C = \{C^p\}$ に分割する。この時、p 番目のアンサブルのクラスタリング結果は k^p 個の要素を持つ弱クラスタの集合 $C^p = \{C_l^p\}$, $l = 1, 2, \ldots, k^p$ もしくはラベルベクトル $\boldsymbol{\lambda}^p \in \mathrm{N}^n$ と表される。最終的に Consensus 関数 Γ によりクラスタリング結果 $\boldsymbol{\lambda}^p$ を統合し単一のクラスタラベル $\boldsymbol{\lambda}$ を得ることで強クラスタを得る。

クラスタアンサンブルは各クラスタリングアルゴリズム固有の入力データやパラメータへの依存性から発生する不安定性を解決すると共に、入力に含まれるノイズやはずれ値の検出、強クラスタの新規性、分散処理時における入力データと弱クラスタ統合処理の独立性等、多くの利点がある。本研究では安定性の向上と得られる強クラスタの独創性に注目し、ATSC への適用方法について検討した。

4. アンサンブル ATSC

本論文では各 ATSC ノードにおける 2 分割処理ついて、アンサンブルを適用する。各 ATSC ノードは ht 種類の距離関数の異なる AHCA を持つアンサンブルそれぞれ hm 個、計 $r = hm \times ht$ 個のアンサンブルを持ち、それぞれ異なる SOM の初期マップ、データの入力順を持つ。

Consensus 関数 Γ は次式を用いる。

$$\Gamma(C) = C^b \qquad (4)$$

$$b = \arg\max_{p} \Delta E^{p} \qquad (5)$$

ここで ΔE^{p} は p 番目のアンサンブルにおける分散の減少(式2)である。通常クラスタアンサンブルにおける Consensus 関数 Γ は、アンサンブル毎の入力データと弱クラスタ統合処理の依存性を排除する為に、距離関数等データに依存する値を用いない。しかしながら、本提案では ATSC の分割アルゴリズムとの統合性を考慮し、分散の減少を用いた。

提案するアンサンブル適用手法により、オンライン学習型 SOM における初期値およびデータ依存性が改善されると共に、各分割において最適なクラス間距離関数を用いる為、従来の階層クラスタリングでは得られなかった分割結果が期待される。

5. 実験

実験では提案するアンサンブル ATSC 手法を、分類問題において代表的なベンチマークデータセットである Iris データおよび Wine データ[6]、より複雑な実医療に基づくデータである CHD_DB[7]の Train_A、計3つのデータに適用し、分類性能について検証を行った。

データセットのクラス数、件数、クラスの比率、属性の数と種類を表1に示す。Iris は件数および属性値が少なくクラスの比率が等しい為、基本的な性能の検証が可能なデータである。対して Wine は属性値の数が Iris データと比較して多く、クラスの比率も異なる為、若干複雑なデータである。CHD_DB は曖昧な情報を含む実医療データであり、件数が非常に多く、離散値と連続値の属性値を持つ為、非常に複雑なデータである。

表 1　データセットの特徴比較

データ名	クラス	件数	クラス比率	属性数 連続/離散
Iris	3	150	1:1:1	4 / 0
Wine	3	177	59:71:48	13 / 0
CHD_DB Train A	2	13000	1:1	4 / 4

各データに対し表2に示すように異なる 3 つの実験手法 Ensemble 1、Ensemble 2、Non-Ensemble を適用し、それぞれ分類正解率の平均と分散、クラスタ数の平均と分散について比較を行った。Ensemble 1、Ensemble 2 は提案するアンサンブル ATSC である。Ensemble 1 は 4 種類の異なる距離関数からなるアンサンブルをそれぞれ 10 個、Ensemble 2 は異なる距離関数を持つアンサンブル 4 個のみを持つ。Ensemble 1 および Ensemble 2 で用いた距離関数は Single Linkage、Complete Linkage、Group Average、Ward 法の 4 つである。Non-Ensemble はアンサンブルを用いない ATSC であり、距離関数は Group Average を用いた。その他 ATSC のパラメータは共通のパラメータを用いた。

表 2　実験手法の比較

手法名	r	hm	ht	距離関数
Ensemble 1	40	10	4	4 Distance Functions
Ensemble 2	4	1	4	4 Distance Functions
Non-Ensemble	1	1	1	Group Average

表3に実験結果を示す。分類正解率およびクラスタ数において Ensemble 1が平均、分散共に最も良い値を得た。得に分散について大幅な数値の減少が見られており、課題であった ATSC の安定性が大幅に改善したと言える。また、Iris および Wine の結果では、単一ノードでの安定性の向上により、ATSC 全体の分類精度が大きく向上していることが、分類正解率の平均から見て取れる。しかしながら、CHD_DB では分類正解率に大きな向上は見られなかった。この原因については現在検討中であるが、件数が増えた影響かデータそのものの特性によると考えられ、今後追加の実験と検証が必要である。

表 3　分類正解率およびクラスタ数の比較

	分類正解率			クラスタ数	
	平均	分散	最高	平均	分散
Iris					
Ensemble 1	0.934	0.000637	0.967	12.0	0.00
Ensemble 2	0.887	0.000948	0.933	11.0	1.11
Non-Ensemble	0.877	0.002866	0.953	9.4	2.48
Wine					
Ensemble 1	0.953	0.000321	0.977	13.8	2.18
Ensemble 2	0.941	0.001037	0.977	14.0	3.56
Non-Ensemble	0.926	0.001043	0.960	14.4	4.27
CHD_DB Train_A					
Ensemble 1	0.654	0.000027	0.666	30.4	3.37
Ensemble 2	0.653	0.000144	0.667	9.7	46.67
Non-Ensemble	0.650	0.000088	0.664	28.0	16.89

6. まとめ

本研究では、ATSC の安定性向上を目的とし、各分割にクラスタアンサンブルを適用するアンサンブル ATSC を提案した。3種類のデータセットに対して提案するアンサンブル ATSC を適用し、全てのデータセットにおいて大幅な安定性向上が得られることを確認した。また、各分割における安定性向上に伴い、分類精度にも向上が見られ、提案手法が非常に有効である事を示した。

なお、本実験では得られたクラスタや木構造について、新規性や妥当性について定量的な検証を行っていない。これに関しては、今後検証方法も含め検討を行っていく予定である。

最後に、アンサンブルを適用することによる大幅な実行時間増加の解決が課題である。これについは今後、分散処理への適応を検討する。ATSC は階層化によってノード毎に処理を分散可能であり、さらに本稿で提案したアンサンブル適用手法においてはアンサンブル毎に処理の分散が可能である為、合理的な手法である。

参考文献

[1] T. Kohonen,Self-organizing maps, Berlin, Springer, 1995
[2] Takashi Yamaguchi, Takumi Ichimura, Kenneth J. Mackin, Adaptive Tree Structured Clustering Method using Self-Organizing Map, Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on advanced Intelligent Systems, 2008
[3] Takashi Yamaguchi, Takumi Ichimura, Kenneth J. Mackin, Analysis using adaptive tree structured clustering method for medical data of patients with coronary heart disease, The 4th International Workshop on Computational Intelligence and Applications, 2008
[4] A. Strehl, J. Ghosh, Cluster ensembles - a knowledge reuse framework for combining multiple partitions, Journal of Machine Learning Research 3, 583-618, 2002
[5] X. Hu, I. Yoo, Cluster ensemble and its applications in gene expression analysis, in: Y.-P.P. Chen (Ed.), Proc. 2nd Asia Pacific Bioinformatics Conference, Dunedin, New Zealand, 297-302, 2004
[6] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/
[7] Machi Suka, Takumi Ichimura, Katsumi Yoshida, Development of Coronary Heart Disease Databases, Proc. of the 8th International Conference on Knowledge-Based Intelligent Information & Engineering Systems, Vol.2, 1081-1088, 2004