解説

最小ユーザフィードバックに基づく制約付きクラスタリング†

岡部 正幸 *1・山田 誠二 *2

1. はじめに

最小ユーザフィードバック (Minimal User Feedback, MUF) [山田 11] は、人間とシステムが協調しながら問題解決を行う知的インタラクティブシステムにおいて、人が行うべきフィードバック作業を最小限に抑えつつ、かつその情報をシステムが最大限に活用するためのアプローチであり、インタラクティブデータマイニング・情報検索システムなどをより知的に利用するために、機械学習を要素技術としたアルゴリズムの構築やインタフェースデザインに関する研究が行われている。

近年、機械学習の分野では、半教師付き学習や能動学習など学習に必要な訓練データに関する現実的な設定を想定した研究が精力的に行われている [Chapelle 06]。これは、訓練データを準備するのは結局のところ人間であり、システムの実用的な運用を考えた場合、訓練データの準備コストは無視できず、システム構築における重要課題としてむしろ積極的に検討すべきとの認識が広がっているためと思われる。これらの研究は、これまで主としてラベル付きデータを利用した分類学習を対象として行われてきたが、近年ではデータへ付与されたラベルの代わりにデータペア間の類似性に関する情報を利用して分類学習を行う制約付きクラスタリング (Constrained Clustering) [Basu 08] へと研究対象が広がっている。最小ユーザフィードバックの観点からみた従来の分類学習と制約付きクラスタリングの違いは、フィードバック対象が個々のデータではなくデータペアとなることであり、フィードバック情報の利用・獲得方法などの扱いがより難しくなっている。本稿では、急速に発展の進むこの制約付きクラスタリングにおける最小ユーザフィードバック関連研究の動向について具体的事例を交えながら解説する。

2. 制約付きクラスタリング

クラスタリング [Han 06] は、教師なし学習と呼ばれるようにデータ集合の特徴空間の構造や分布などを頼りにデータのグループ化を行う方法論であるが、近年データ間の距離や類似性についての情報を取り込み、クラスタリング精度を向上させようとするアプローチに関する研究が行われている。この研究は制約付きクラスタリングと呼ばれ、従来のクラスタリングに対する付加情報として、一般に次のような知識が与えられる。

Must リンク：同じクラスタに属すべきデータペア。
Cannot リンク：異なるクラスタに属すべきデータペア。

これらの知識はクラスタリングを行う際に考慮すべき制約条件 (以下では単に「制約」と呼ぶ) として扱われる。また、分類学習として見た場合、情報が与えられる対象が個々のデータに対してではなく、データペアであるという点が従来の分類学習と大きく異なる。個々のデータにクラスラベルなどの情報が与えられれば、それを利用して Must / Cannot リンクは生成可能なため、フィードバックの点からは汎用性が高い方法であるといえる。制約付きクラスタリングでは、与えられた制約の活用方法とその実装アルゴリズムの設計が主たるテーマとなるが、ここではこれまでに提案されてきた2つのアプローチについて紹介する。

まず一つ目は、Wagstaffらによって提案された COP-Kmeans [Wagstaff 01] である。COP-Kmeans は分割型クラスタリングアルゴリズムの代表的手法である K-means を改良し、制約を考慮した操作を組み込んだアルゴリズムである。K-means の基本的手順は以下の作業の繰り返しである。

1. 初期クラスタ中心の決定
2. 全データのクラスタ中心への割り振り
3. クラスタ中心の更新

COP-Kmeans では、手順 2 のデータ割り振り操作の際に各クラスタ内に制約に違反するデータペアが存在しないように割り振りを行う。ただし、この操作は割り振り順序に依存しており、順序によっては割り振

† Constrained Clustering based on Minimal User Feedback
Masayuki OKABE and Seiji YAMADA

*1 豊橋技術科学大学情報メディア基盤センター
Information and Media Center, Toyohashi University of Technology

*2 国立情報学研究所／総合研究大学院大学／東京工業大学
National Institute of Informatics, SOKENDAI, Tokyo Institute of Technology

り可能なクラスタ中心がなくなることもある。COP-Kmeansでは、そのような場合のバックトラックなどの方法は特に提案されておらず、アルゴリズム自体が停止してしまうという欠点がある。

一方、制約を利用してデータ間の距離を変化させる距離学習[Weinberger 05, Xing 02]と呼ばれるアプローチがある。図1に示すようにMustリンク集合に属するデータペア同士の距離をより近くに、Cannotリンク集合に属するデータペア同士の距離をより遠くに変化させることで従来のクラスタリングアルゴリズムを特に改良することなく精度を向上させることができるという利点がある。また、データ間の距離はクラスタリングのみならず、 k -近傍学習など分類学習全般における基本的な類似性判定尺度であり、広範囲での適用が期待できる。

距離学習では、学習対象として以下で示される行列 A を変数としたマハラノビス距離が用いられることが一般的である[Xiang 08]。

$$d_A(v_i, v_j) = (v_i - v_j)^t A (v_i - v_j)$$

ここで、 $v_i \in \mathbb{R}^d$ はデータ i の特徴ベクトルを表している。学習の目的は、先の制約集合に属するデータペアが満たすべき距離を実現するような半正定値行列 $A \succeq 0$ を求めることである。行列 A の半正定値性は、 $d_A(v_i, v_j)$ が距離として満たすべき性質(非負性、三角不等式など)を持つために必要十分な条件である。マハラノビス距離は特徴空間における任意の線形拡大縮小、回転操作を包含しており、行列 A が一旦求めれば、 $A = W^t W$ と分解することにより、特徴ベクトルに対する射影行列を求めることができる。これにより、新たなデータが追加された場合でもこの射影行列を適用することで距離学習の効果を反映させることができる。また、距離学習の類似研究としてカーネル行列の学習に関する研究[Hoi 07, Li 08]がある。カーネル学習では Mercer の定理[赤穂 08]を満たすカーネル関数を用いることにより、特徴空間の射影先での内積計

算(= 類似度)を簡単に行うことができるが、同じように半正定値性を満たすカーネル行列が求めれば、既存のデータ集合に適用範囲は限定されるものの、カーネルとして学習に利用可能である。距離学習もカーネル行列の学習も解法へのアプローチは類似しているため、以降では距離学習とカーネル行列の学習を特に区別することなく単に「距離学習」と呼ぶ。

以上のように、近年の制約付きクラスタリングに関する研究は距離学習をベースとしたものが多く、研究対象も学習アルゴリズムだけでなく、制約の利用・獲得方法などに広がっている。次章からは、この距離学習に関する研究のうち、最小ユーザフィードバック研究に関連する3つのトピックを取り上げ、それぞれの代表的研究の紹介とともに最小ユーザフィードバックとの関連について述べる。

3. 距離学習における最小ユーザフィードバック

最小ユーザフィードバックの主要課題の一つにユーザフィードバックの負荷の軽減がある。これは更に計算論的コストと認知的コストの議論に展開されるが、後者については[高間 11]を参照頂くとして、本稿では前者の距離学習における最小ユーザフィードバックを実現するための計算論的コストについて考える。ここでいうユーザフィードバックとは、距離学習を行う際に利用する制約集合の獲得を指している。一般に制約数が多いほど距離学習の効果が増し、制約付きクラスタリングの精度も向上すると考えられているが、制約数を増やすとユーザの労力もそれだけ増すため、性能とユーザ負荷のバランスをうまく取るための技術の構築が課題となる。本稿では、このような技術開発を行うため、次の3つのアプローチに着目する。

1. 利用可能な制約数が少数に限定された状態で学習性能を向上させる方法。
2. 効果の高い制約を選択する方法。
3. システムの対話性を高めるためのオンライン型学習アルゴリズム

以降では、上記アプローチそれぞれの代表的研究について順に説明する。

3.1 制約伝播による制約集合の擬似的拡張

制約付きクラスタリングにおける最小ユーザフィードバックを実現する上でまず考えなければならないのは、少ない制約の有効な活用方法である。ここでは、そのような方法の一つとして、制約の振られたデータペアの近傍データに制約関係をうまく伝播させる研究について紹介する。

図1 距離学習のイメージ

距離学習はそもそもそれ自体が限られた制約関係をデータ集合全体に暗黙的に伝播させていると捉えることもできるが、筆者らのグループでは、より明示的に伝播させる方法を提案した[Okabe 10]。詳しい説明は、[山田 11]を参照して頂きたいが、この方法ではクラスタ仮説(cluster assumption)[Chapelle 05]に基づき、データペア (i, j) における制約関係をデータ i とデータ j の近傍データおよびデータ j とデータ i の近傍データへと明示的に伝播させている。クラスタ仮説は近接するデータ同士は同クラスに属する可能性が高いとする仮説であり、実際のデータにおいてもよく当てはまることが多い。伝播させた制約は人手による判定が行われていない擬似的なものであるため、間違った制約を与えてしまうこともあるが、制約を伝播させる近傍データ数をうまく調節することで性能を向上させることができることを示した。

制約の伝播に関する他の研究として、Liuらの研究[Liu 10]がある。彼らは、距離学習の前処理として既存の制約を以下の式に基づき近傍に拡散させる方法を提案した。

$$W^{t+1} = (1 - \alpha)W^0 + \alpha PW^t$$

ここで、 W はデータ間の近接度を表す行列であり、 $W_{ii}^0 = 1$ かつ W_{ij}^0 はデータ (i, j) がMustリンクであれば+1、Cannotリンクであれば-1、その他の場合は0である。また、 α はトレードオフを決定するパラメータであり、 P は以下の要素からなる行列である。

$$P_{i,j} = \begin{cases} 1/k & j \in \mathcal{N}_i \\ 0 & \text{otherwise} \end{cases}$$

$\mathcal{N}_i$ はデータ i の近傍データ集合でこれはユークリッド距離を基に求める。この方法により、与えられた制約に関する情報が制約の振られていないデータペア間へと拡散していき、すべてのデータペア間に擬似的な制約が振られた状態 W^* を作り出すことができる。

$$W^* = (1 - \alpha)(I - \alpha P)^{-1}W^0$$

この前処理単体の効果についての結果は記載されていないが、彼らが提案した距離学習アルゴリズムは従来手法より優れていることが報告されている。

3.2 能動学習による効率的な制約選択

3章の冒頭でも述べたように、一般に制約付きクラスタリングの精度は利用する制約数の増加とともに向上する傾向にあるが、その過程は単調ではなく図2に示すように不安定な傾向を示すことが多くの研究から明らかになっている[Klein 02, Tang 07]。これは、

図2 制約数とクラスタリング精度の関係

個々の制約の効果にバラつきがあることを示している。ランダム選択では制約数が増えると効果の高い制約も取り込まれるが、同時に悪影響を与える制約が取り込まれる可能性もある。このため、効果のある制約のみを選別する技術の構築が必要とされている。このように学習効果の見込まれるデータを選別するアプローチは能動学習と呼ばれ分類学習では古くから研究されてきた。ここでは、従来の能動学習の考え方を距離学習に応用したHoiらの研究[Hoi 08]について紹介する。

Hoiらは、下記に示す半正定値計画問題をベースとしたカーネル行列の学習問題に有効な制約選択方法を考えた。

$$\begin{aligned} \min_{Z, \varepsilon} \quad & \text{tr}(LZ) + \frac{c}{2} \sum_{(i,j) \in T} \varepsilon_{i,j}^2 \\ \text{s.t.} \quad & T_{i,j} Z_{i,j} \geq 1 - \varepsilon_{i,j}, \varepsilon_{i,j} > 0, \forall (i,j) \in T \\ & Z \geq 0 \end{aligned}$$

ここで、 Z は求めるべきカーネル行列、 L はデータ集合から生成した近傍グラフのグラフラプリアンである。また、 $T_{i,j}$ はデータ i とデータ j がMustリンクの場合は+1、Cannotリンクの場合は-1をとる変数である。

目的は既存の制約データペア集合に新たに追加すべきデータペアを選択することである。Hoiらは、まず能動学習においてしばしば用いられるUncertaintyサンプリング[Lewis 94]を試した。これは、分類学習においてラベルが未知のデータ集合のうち、判別結果の確信度が最も低いものを判定すべきとする選択基準であり、先のカーネル行列の学習問題に当てはめると、 $Z_{i,j}$ が最も0に近いものを選択することになる。しかしながら、予備実験による評価ではこの方法はうまく機能しないことが分かった。そこで彼らは新たに以下のような方法を提案した。彼らは未知ラベルのデータペアに制約を付与した場合の最適値 $\omega((k, l), t)$ に着目し、Must/Cannotリンクのそれぞれを付与した場

合の大きい方の値が最小となる, つまり, 以下を満たすデータペア (k, l) を選択する min-max アプローチによる方法を提案した.

$$(k, l)^* = \operatorname{argmin}_{(k, l) \notin T} \max_{t \in \{+1, -1\}} \omega((k, l), t)$$

$\max_{t \in \{+1, -1\}} \omega((k, l), t)$ の部分については, 以下の最適化問題を解くことと等価であることが示されるが,

$$\begin{aligned} \min_{Z, \varepsilon, (k, l) \notin T} \quad & \operatorname{tr}(LZ) + \frac{c}{2} \sum_{(i, j) \in T} \varepsilon_{i,j}^2 + \frac{c}{2} \varepsilon_{k,l}^2 \\ \text{s.t.} \quad & T_{i,j} Z_{i,j} \geq 1 - \varepsilon_{i,j}, \varepsilon_{i,j} > 0, \forall (i, j) \in T \\ & \varepsilon_{k,l} \geq 1 + |Z|, Z \geq 0 \end{aligned}$$

実際に最適なデータペア $(k, l)^*$ を求めるには, すべてのデータペアについて上記の問題を解く必要がある. ただし, これは現実的ではないため, すべてのデータペアについてそのペアが選択される確率を表す変数を導入することで, すべてのデータペアの組み合わせを試すことなく, 最適化問題を一度解くだけで選択すべきデータペアが求まる方法を提案した. この変数の導入により, 非凸な最適化問題への対処など, 計算上の工夫が種々必要となるが, 解法の詳細については文献を参照頂きたい. 実験結果では, 提案手法がランダム選択や Uncertainty サンプリングを上回る性能を示したことが報告されている.

距離学習またはカーネル行列の学習における能動学習は, 分類学習と比べてデータペアが対象となるため選択範囲が広く難しい. しかしながら, 制約選択の問題は距離学習の性能を大きく左右する可能性もあり, 今後の研究が期待される分野である.

3.3 オンライン型距離学習

距離学習に関する研究の中で提案されてきた手法の多くは, 固有値分解, 半正定値計画問題, 2次計画問題など計算コストが大きな解法をベースとしている. 計算コストの削減はデータのスケール化に必要な課題であるが, システムの応答速度が求められる情報検索などのインタラクティブシステムにおいても重要な課題である. 例えば, 図3に示すような画像のクラスタリングをインタラクティブに行う場合, 2つの画像が似ているかどうかの判定はテキスト情報を判定するよりも一般的に楽な作業であると考えられ, アルゴリズムの精度に計算コストをかけるよりも, ある程度の精度をもつアルゴリズムを用いてシステムの応答性を高めることにより, ユーザからフィードバック回数を増やすほうが, インタラクション全体として高い精度を達成できる場合もある. 最小ユーザフィード

画像クラスタリングなどのタスクでは
システムの応答速度が重要になる

図3 インタラクティブクラスタリング

バックにおいて, ユーザとシステム間のスムーズなインタラクティブ性は重要な研究課題であり, 計算コストを軽減する要素技術が望まれている [Fails 03]. ここでは, 距離学習における計算コスト削減のためのアプローチとして, Jainらによって提案されたオンライン型の距離学習アルゴリズム [Jain 08] について紹介する.

Jainらは, マハラノビス距離におけるパラメータ行列 A を制約の追加ごとに更新可能なオンライン型のアルゴリズムを提案した. まず彼らは時刻 t に制約の一つを受け取るという前提で, 時刻 $t+1$ における行列 A_{t+1} を求める問題を以下のように定式化した.

$$A_{t+1} = \operatorname{argmin}_{A \succ 0} D(A, A_t) + \eta \ell(d_A(u_t, v_t), y_t)$$

ここで, $D(A, A_t)$ は更新前の行列 A_t を利用した正則化関数, $\ell(d_A(u_t, v_t), y_t)$ は制約の与えられたデータペア (u_t, v_t) 間のマハラノビス距離と満たすべき距離 y_t との損失関数である. また, η は学習係数であり, 行列の保存と制約の充足のどちらを重要視するかを決定するパラメータである. 正則化関数には LogDet ダイバージェンスと呼ばれる以下の関数を用いている.

$$D(A, A_t) = \operatorname{tr}(AA_t^{-1}) - \log \det(AA_t^{-1}) - n$$

この関数は, 行列 A と A_t の相違を測る尺度として利用可能であり, 実際に行列 A^{-1} と A_t^{-1} をそれぞれ共分散行列にもつ2つの多変量ガウス分布の間の Kullback-Leibler ダイバージェンスと等価であることが示されている [Davis 07]. つまり, 上記の更新式は, 追加された制約を満たす行列のうちできるだけ更新前の行列 A_t を保存するものを計算せよということを意味している.

更新後の行列 A_{t+1} を求めるため, A に関する勾配を0とすると, 以下のような更新式を得ることができる.

$$A_{t+1} = A_t - \frac{\eta(\bar{y} - y_t)A_t z_t z_t^T A_t}{1 + \eta(\bar{y} - y_t)z_t^T A_t z_t}$$

$$\bar{y} = z_t^T A_{t+1} z_t = \frac{\eta y_t \hat{y}_t - 1 + \sqrt{(\eta y_t \hat{y}_t - 1)^2 + 4\eta \hat{y}_t^2}}{2\eta \hat{y}_t}$$

ただし, $z_t = u_t - v_t$, $\hat{y}_t = z_t^T A_t z_t$ である.

このアルゴリズムの大きな利点として, 更新後の行列 A_t の半正定値性について特別な検証作業を行う必要がないという点が挙げられる. 半正定値行列であることの必要十分条件は任意の行列 z_t に対して, $z_t^T A_{t+1} z_t \geq 0$ となることであるが, 上式より,

$$\begin{aligned} z_t^T A_{t+1} z_t &= \frac{\eta y_t \hat{y}_t - 1 + \sqrt{(\eta y_t \hat{y}_t - 1)^2 + 4\eta \hat{y}_t^2}}{2\eta \hat{y}_t} \\ &\geq \frac{\eta y_t \hat{y}_t - 1 + |\eta y_t \hat{y}_t - 1|}{2\eta \hat{y}_t} \geq 0 \end{aligned}$$

であることが明らかである.

実験では, 他のオンライン型学習アルゴリズムのほか, 計算コストの高い最適化手法を用いた方法と比較して遜色の無い性能を示したことが報告されている.

4. まとめ

本稿では, 最小ユーザフィードバックを実現する上で問題となる計算論的コストに関する課題を解決するアプローチについて, 特に距離学習をベースとした制約付きクラスタリングにおけるこれまでの研究事例を紹介した. これらの研究は, 最小ユーザフィードバックに基づくシステムの構築に利用可能である一方, 今後の研究発展の手がかりともなりうる. 特に能動学習に基づく制約の選別方法に関しては, その重要性の割に研究が進んでおらず今後の発展が期待される.

参考文献

- [Basu08] S. Basu, I. Davidson, K. Wagstaff, Constrained Clustering: Advances in Algorithms, Theory, and Applications, Chapman & Hall, 2008.
- [Chapelle 05] O. Chapelle and A. Zien, Semi-Supervised Classification by Low Density Separation, In Proc. AISTATS'05, pp.57-64, 2005.
- [Chapelle 06] O. Chapelle, Semi-supervised Learning, The MIT Press, 2006.
- [Davis 07] J. V. Davis, B. Kulis, P. Jain, S. Sra and I. S. Dhillon, Information-Theoretic Metric Learning, In Proc. ICML'07, 2007.
- [Fails 03] J. A. Fails and R. Olsen, Interactive Machine Learning, In Proc. IUI'03, pp.39-45, 2003.
- [Han 06] J. Han and M. Kamber, DataMining: Concepts and Techniques, 2nd ed., Morgan Kaufmann Publishers, 2006.
- [Hoi 07] S. C. H. Hoi, R. Jin, and M. R. Lyu, Learning Non-

parametric Kernel Matrices from Pairwise Constraints, In Proc. ICML'07, 2007.

- [Hoi 08] S. C. H. Hoi and R. Jin, Active Kernel Learning, pp.400-407, 2008.
- [Jain 08] P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman, Online Metric Learning and Fast Similarity Search, In NIPS'08, pp.761-768, 2008.
- [Klein 02] D. Klein, S. D. Kamvar and C. D. Manning, From Instance-level Constraints to Space-level Constraints: Making the Most of Prior Knowledge in Data Clustering, In Proc. ICML'02, pp.307-314, 2002.
- [Lewis 94] D. D. Lewis and W. A. Gale, A Sequential Algorithm for Training Text Classifiers, In Proc. SIGIR'94, pp.3-12, 1994.
- [Li 08] Z. Li, J. Liu and Z. Tang, Pairwise Constraint Propagation by Semidefinite Programming for Semi-Supervised Classification, In Proc. ICML'08, 2008.
- [Liu 10] W. Liu, S. Ma, D. Tao, J. Liu and P. Liu, Semi-Supervised Sparse Metric Learning using Alternating Linearization Optimization, pp.1139-1147, In Proc. KDD'10, 2010.
- [Okabe10] M. Okabe and S. Yamada.: Learning Similarity Matrix from Constraints of Relational Neighbors, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.14, No.4, pp.402-407, 2010.
- [Tang 07] W. Tang, H. Xiong, S. Zhong and J. Wu, Enhancing Semi-Supervised Clustering: A Feature Projection Perspective, In Proc. KDD'07, pp.707-716, 2006.
- [Wagstaff 01] K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl, Constrained K-means Clustering with Background Knowledge, pp.577-584, In Proc. of ICML'01, 2001.
- [Weinberger 05] K. Q. Weinberger, J. Blitzer and L. K. Saul, Distance Metric Learning for Large Margin Nearest Neighbor Classification, In NIPS'05, 2005.
- [Xiang 08] S. Xiang, F. Nie and C. Zhang, Learning a Mahalanobis distance metric for data clustering and classification, Pattern Recognition, Vol.41, No.12, 2008.
- [Xing 02] E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell, Distance metric learning with application to clustering with side information, In NIPS'02, 2002.
- [赤穂 08] 赤穂, カーネル多変量解析, 岩波書店, 2008.
- [高間 11] 高間, 陳, 山田: 最小ユーザフィードバックのためのインタフェースデザイン, 知能と情報, Vol.23, No.1, pp.86-92, 2011.
- [山田 11] 山田, 岡部, 高間, 小野田, 最少ユーザフィードバックの枠組みとその要素技術, 知能と情報, Vol.23, No.1, pp.80-85, 2011.

(2011年1月14日 受付)

[問い合わせ先]

〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘1-1
豊橋技術科学大学情報メディア基盤センター
岡部 正幸

TEL: 0532-44-6639

FAX: 0532-44-6639

E-mail: okabe@imc.tut.ac.jp

著者紹介

おかべ まさひろ
岡部 正幸 [非会員]

1996年創価大学工学部情報システム学科卒業。2001年東京工業大学大学院総合理工学研究科知能システム科学専攻博士課程修了。博士(工学)。同年 科学技術振興機構(CREST)研究員。2003年豊橋技術科学大学情報メディア基盤センター助教。現在に至る。知的情報検索の研究に従事。人工知能学会会員。

やまだ せいじ
山田 誠二 [非会員]

1984年大阪大学基礎工学部卒業。1989年大阪大学大学院基礎工学研究科博士課程修了。工学博士。1989年大阪大学基礎工学部助手。1991年大阪大学産業科学研究所講師。1996年東京工業大学大学院総合理工学研究科助教授。2002年国立情報学研究所教授。現在にいたる。HAI, 知的インタラクティブシステムに興味をもつ。