2F4-OS-04-2

協調的制約獲得によるクラスタリング

Collaborative Constrained Clustering

窪田 暁[*1] 山田 誠二[*2*1]
Gyo Kubota Seiji Yamada

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

The number of constraints for interactive constrained clustering is limited because the user's cognitive load to give them is significantly high. Also it is difficult to gain the correct classification by assigning the constraint with a single user's limited background knowledge because the big data is huge. In this paper, we propose collaborative constrained clustering by repeating both the assigning many users' constraints and clustering.

1. はじめに

近年，大量のデータがエンドユーザからアクセス可能になっている．この大量に蓄積されたデータから潜在的な知識や法則性を発見するための分析手法として，データマイニングが盛んに研究されている．

その中でもクラスタリングは，一般的にはあらかじめデータにラベルを付与する必要がない教師なし学習であり，類似したデータをグループ化することによりデータを分類する，データマイニングにおいて最も広く普及している手法の 1 つである．しかし，教師なしであるクラスタリングを 1 回行うだけではユーザに望ましい結果を与えることは困難であるため，背景知識を持ったユーザが制約を与え，それを満たすようなクラスタリングを行う制約付きクラスタリング [1] により精度の改善が可能である．それに加えて，ユーザの背景知識を利用するインタラクティブな方法として，ユーザフィードバックによる制約付与と制約付きクラスタリングを交互に繰り返しながら制約を追加していくことで精度を改善していくインタラクティブ制約付きクラスタリング [2] が研究されている．

一方，大量のデータにおいてインタラクティブ制約付きクラスタリングを用いると，制約を与える度にユーザに認知的負荷がかかるため，付与することの出来る制約数がデータに対して少数となり，インタラクティブ制約付きクラスタリングによる精度向上は難しい．また，大量に蓄積されているデータは多様であるため，ユーザ 1 人の背景知識のみで制約を与えることは容易ではない．そこで本研究では，多人数のユーザによる制約付与を導入し，多人数が協調しながら [3]，インタラクティブ制約付きクラスタリングを行う協調制約付きクラスタリングを提案し，そこにおける課題とその解決方法を議論する．

類似した研究としてヒューマンコンピュテーション [4] がある．ヒューマンコンピュテーションとは，コンピュータには難しいが人間には容易に出来ることと人間には難しいがコンピュータには容易に出来ることの双方を組み合わせることによって，片方では解決できなかった問題を効率的に解決することを目指す研究領域である．インタラクティブ制約付きクラスタリングは，制約付与における人間の力とクラスタリング処理におけるコンピュータの力を組み合わせることによって可能となる技術の研究という意味でヒューマンコンピュテーションと類似している．例えば，犬と猫の大量の画像データを 2 クラスに分類したい場合，人間にとって犬と猫の違いを判別することは容易だが，大量のデータを処理することは苦手である．一方で，コンピュータにとって大量のデータを処理することは容易だが，犬と猫の違いを正確に判別することは苦手である．つまり，片方のみを用いるのでなく，双方の特性を生かしたインタラクティブな技術を開発することが出来れば，効率的，知的にデータを分類することが可能になる．

連絡先: 窪田暁，山田誠二，kubota@ntt.dis.titech.ac.jp, seiji@nii.ac.jp

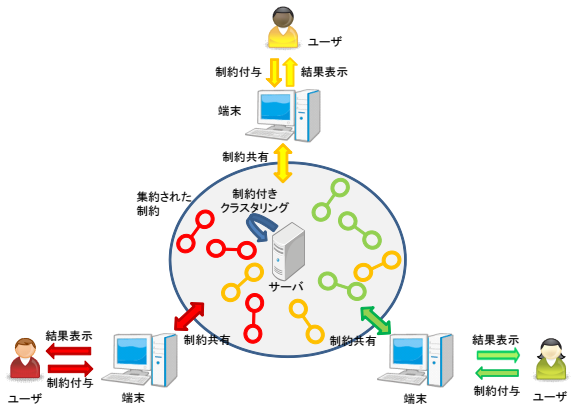

図 1 協調制約付きクラスタリングの概念図

一方で，ヒューマンコンピュテーションのシステムは，人間が行った計算処理をシステムで活用する時に，必ずしも計算処理を行った人間の目的と一致していない場合がある．例えば，ESP ゲーム [5] は 2 人のゲームプレーヤーが表示された画像に対して適切だと思うタグを同時に付与し，同一であった場合，信頼できるタグとして情報検索に活用されるメカニズムとなっている．しかし，ゲームプレーヤーはゲームをプレイすることを目的としていて，情報検索自体とプレーヤーに直接の関係はない．それに対して，インタラクティブ制約付きクラスタリングは，人間が介入しながらその人間の目的に合わせてコンピュータが支援しつつ，片方が欠けてしまうと出来ないような計算を，両方の特性を生かしつつそれらを組み合わせることによって協調的に計算していくという点に特化している．

2. 協調制約付きクラスタリング

図 1 に協調制約付きクラスタリングの概念図を示す．まず最初に，与えられたデータに対して制約なしクラスタリングを

行い，初期クラスタリング結果が得られる．そして，その初期クラスタリング結果をもとに，複数のユーザが非同期／同期してインタラクティブ制約付きクラスタリングを行っていく．この際，ユーザに提示される部分データの決定，ユーザの制約付与の順序付け，多ユーザにより付与された制約の共有方法，を通じて協調制約付きクラスタリングがシステムにより制御される．このような枠組みにより，インタラクティブ制約付きクラスタリングにおいて付与できる制約付与数を増やし，多様なデータに対し様々な背景知識をもつユーザがお互いに協調しながら精度の高いクラスタリングを目指す．

3. 協調制約付きクラスタリングにおける課題

協調制約付きクラスタリングには，主に制約付与の制御に関していくつか重要な課題が考えられる．以下に，それらについて記述する．

3.1 データスコープ，制約スコープ，ユーザ間スコープ

本研究においては，大量のデータ (例えば，数万の画像データ) の分類を可能にすることを念頭に置いているが，分類したいデータや，それに対する制約，参加する人間，それらの情報を全て画面に表示することは難しい．そのような環境において，何を表示するのが適切なのかという問題がある．以下にそれをまとめる．

1. **データスコープ**: 図 2 のような，大規模なデータのどの部分を，ユーザにどのように提示するかの問題がある．UI の物理的なサイズは限られているので，画像などの大規模データをすべて一度にユーザに提示することは難しい．そのため，どうしてもごく一部のデータをユーザに見せることになるが，その際どの部分を提示するかが問題となる．基本的に，ユーザは提示されたデータ間にペアワイズ制約を付与していくので，提示されたデータだけが制約の対象となる．つまり，この問題は能動学習 (active learning)[6] の一種である．
   さらに，複数のユーザが制約付与を行うため，それぞれのユーザについて，このデータスコープの問題が生じる．複数ユーザのデータスコープは，同一であるべきか，積集合がない部分集合であるべきか，また非同期付与の場合にはその順序はなどの問題がある．基本的に能動学習の問題であるため，このデータスコープにより制約付きクラスタリングのパフォーマンスが左右される可能性が高い．
2. **制約スコープ**: 図 3 のような，ユーザにユーザが付与した制約をどのように表示するかの問題がある．制約を多く見せた方が良いのか，少ない方が良いのか．多くのユーザを一画面に集めて制約を付与させた方が良いのか，全画面に均等に制約を付与させた方が良いのか，それともどの部分に制約を与えるかはユーザに全て任せてしまう方が良いのかなどを決定する必要がある．この問題の解決には，他ユーザの制約を見ることは，そのユーザの制約付与にどのような影響を与えるかを解明する必要がある．そのような影響として，「他ユーザの制約からできるだけ離れたデータペアに制約を与える」などのヒューリスティックスが考えられる．
3. **ユーザ間スコープ**: 他ユーザから付与された制約を見せる UI を採用した場合，誰が与えたかわからないようにするのか．また，1 人 1 人区別出来るようにするのかの問題がある．この問題の解決にも，先の制約スコープ同様に，制約付与者が特定できることが，ユーザにどのような影響を与えるのかについて考察する必要がある．

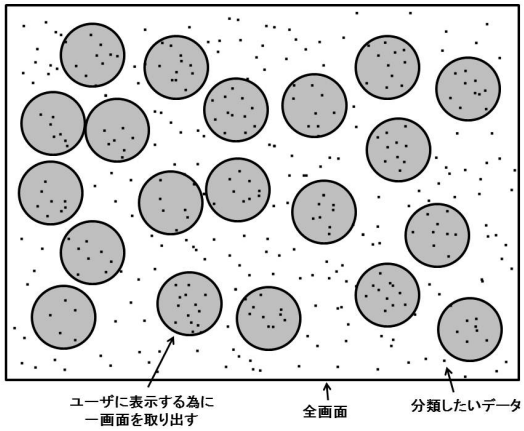

図 2 データスコープ

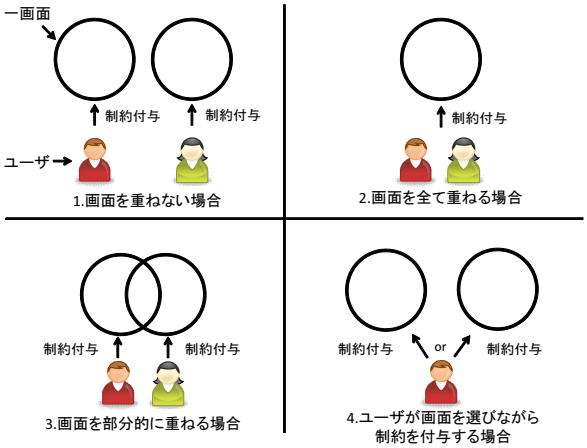

図 3 制約スコープの様々な表示方法

3.2 マルチ視点

大量のデータを分類する場合，ユーザそれぞれによって与える制約の視点が異なり，多人数で制約を与えると分類に統一性がなくなってしまう問題が考えられる．例えば，車とバイクを分類する場合，乗り物として同じグループに分類しようとするユーザがいれば，車とバイクは別々に分類しようとするユーザもいるかもしれない．

本来，1 人で制約を付与する場合においては，どのようなアプローチで制約を付与しても分類全体は一つの制約方法で統一されていたのでこの問題について大きく考える必要はなかったが，多人数で制約を与える場合大きく影響が出ることが考えられる．

4. 課題を解決するための方法

データスコープについては，制約を付与していく人間のモデルを設定し，それを用いてシミュレーションを行うことにより決定していく．なお，シミュレーションを行うにあたっては最初から分類の正解が明らかなデータを使用し，インストラクションによりマルチ視点のような問題は考えなくてもよい場合を想定する．

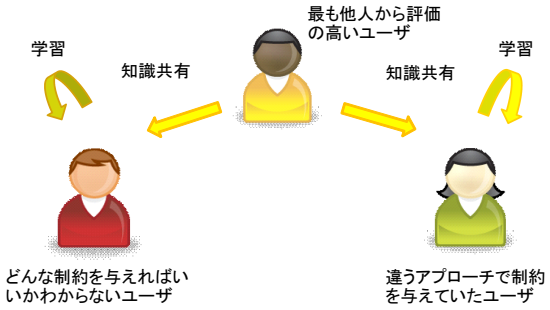

図 4 評価システムのメカニズム

シミュレーションで行う内容を以下に示す.

1. 人間のモデルを，人間の能動学習のための GUI に関する先行研究 [7] に基づき，人間の能動学習の戦略を取り込んで設定する.
2. 設定した人間のモデルの数を 20，1 人につき付与する制約数を 100 と定める.
3. それぞれのモデルが画面を見てデータに制約を与える．その中で，あるモデルが付与した制約数が 100 を超えたら，そのモデルの制約付与を終了する.
4. 全モデルが制約を 100 ずつ付与した時点で終了する.

ステップ 3 については，それぞれのモデルがどの画面に制約を付与するかは，固定，一定数与えたら移動する，など何らかの規則性を定めることにより行っていく．また，あるユーザの画面が他のユーザの画面と重なっているかなど，様々な条件を想定することにより，シミュレーションを行っていく.

これに加え，制約スコープについても実験を行っていく．これはシミュレーションによる検討だけでなく，参加者実験も実施する．参加者実験においては，制約を与える画面に別のユーザの制約が見えるケースと，ユーザはそれぞれ別の画面に制約を与え，別のユーザの制約が見えないケースを比較し，どのような影響が出るか考察していく．この際，ユーザ 1 人 1 人に別のユーザが与えた制約を見せることで精度が下がるのであればなんらかの問題を引き起こしていることが考えられ，逆に，ユーザが別のユーザの制約を見ることで精度が高くなるのであれば，ユーザ間でなんらかの協調効果を見込めると考えられる.

次に，ユーザ間スコープとマルチ視点問題への対応としては，図 4 のような評価システムを導入する．与えられた制約を見て，良い制約を与えているユーザに対して良い評価を与えることが出来る．逆に，悪い制約を与えているユーザに対して悪い評価を与えることも出来る．このように，どんな制約を与えればいいのかわからないユーザや，違うアプローチで制約を与えていたユーザなど，様々なユーザに良い評価のユーザを意識してもらうようなシステムを導入することで，それぞれのユーザ自身に制約の方向性をコントロールしてもらうような環境で分類全体を統一させていくようにする．一方で，良い評価を与えられたユーザの制約を区別して表示するのか，悪い評価を与えられたユーザの制約を別のユーザに表示しない方が良いのかなどの設定についても考える必要があり，この設定がどの程度ユーザに影響するのかについて考察する必要がある.

5. 評価実験

シミュレーションで行った結果をもとに，協調制約付きクラスタリングにおいて適切なモデルを構築し，参加者実験を行うことによってそれぞれのスコープ問題やマルチ視点に対応していないモデルと比較し，どれだけ協調制約付きクラスタリングの結果に影響が出たかを評価していく.

6. まとめ

本論文では，従来行われていたユーザ 1 人による限られた背景知識による制約付与と，インタラクティブ制約付きクラスタリングで付与出来る制約数の有限性に対応する為に，多人数のユーザが協調しながらインタラクティブ制約付きクラスタリングを行う，協調制約付きクラスタリングを提案した．また，大量のデータを想定し協調制約付きクラスタリングを導入した場合に起こる問題について検討し，それらの解決方法について議論した.

今後の方向性として，実際にシミュレーションと参加者実験を行うことによりどのような制約付与方法が適切かどうか検討していく．また，データに対してそれぞれのユーザが異なった視点で制約を付与してしまい制約付与の統一性がなくなってしまうマルチ視点問題に対応するための評価システムを導入する．その後，参加者実験とその評価を行う予定である.

参考文献

[1] S. Basu, I. Davidson, and K. Wagstaff, Eds: Constrained Clustering: Advances in Algorithms, Theory, and Applications. Chapman & Hall (2008).

[2] Y. Sato and M. Iwayama: Interactive Constrained Clustering for Patent Document Set, In Proceedings of the 2nd international workshop on Patent information retrieval, pages 17-20 (2009).

[3] D. C. Anastasiu, D. Buttler, and B. J. Gao: A framework for personalized and collaborative clustering of search results, In Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11), pages 573-582 (2011).

[4] E. Law, and L. V. Ahn: Human computation. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool (2011).

[5] L. V. Ahn, and L. Dabbish: Designing games with a purpose, Communications of the ACM, 51(8), pages 58-67 (2008).

[6] D. D. Lewis, and W.A. Gale: A sequential algorithm for training text classifiers, In Proceeding of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94), pages 3 - 12 (1994) .

[7] S. Yamada, J. Mizukami and M. Okabe: Designing GUI for Human Active Learning in Constrained Clustering, In Proceedings of IUI 2013 Workshop on Interactive Machine Learning (IUI '13), 4 pages (2013).