4F1-OS-5-5

インタラクティブクラスタリングにおける対制約生成手法の検討

An Approach to Pairwise Constraint Generation for Interactive Clustering

三宅　遼祐[*1]　山田　誠二[*2]　高間　康史[*1]
Ryosuke Miyake　Seiji Yamada　Yasufumi Takama

[*1]首都大学東京大学院　システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

[*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

In modern society, it is important to get desired information from large-scale information source. Although recent advance of information processing such as information retrieval and recommendation has provided us with information access support, automatic information processing only is not enough. From this viewpoint, semi-supervised clustering has been studied as one of promising ways of collaboration between humans and computer systems. When constrained clustering is applied to actual tasks, it is important to reduce user's cost of specifying constraints. As one of the solutions, the approach of automatic constraint generation from user's grouping operations have been proposed. This paper extends the existing approach by considering the log of user's past grouping operations. The proposed approach covers both of conservative and sensitive approaches based on the same information. A small example is shown to explain the proposed approach.

1. はじめに

現代社会において情報の大規模化が急速に進行しており，情報が持つ価値，重要性が非常に高くなっている．特に，大量の情報の中から自分にとって有益な情報をいかに取りだすかが重要であり，データマイニングなどの計算機による支援技術が研究されている．しかし，大規模情報に対して，計算機による自動処理だけで適切な結果を得ることは困難であるため，人間と計算機の協調による半教師あり学習として制約クラスタリングの研究が行われている．

制約クラスタリングをインタラクティブシステムで利用することを考えた場合，制約生成に要するユーザの負荷を軽減することが重要となる．この問題に対し，ユーザがオブジェクトグループを作成する操作から複数の must-link を一括生成するアプローチが提案されている [三宅 11]．制約の一括生成はユーザの負荷軽減に効果があることが期待できる反面，ユーザの意図と異なる制約を付与した場合には作業の妨害になる可能性もある．そこで，本稿では既存研究で提案された手法を拡張し，過去の操作履歴も考慮した，より一般的な枠組みから must-link 生成手法について検討する．

2. 関連研究

2.1 制約クラスタリング

クラスタリングとは，データ解析方法の一つであり，与えられた類似度を基準にデータを分類する教師なし学習である．対象とするデータは一般に，多様な観点から分類することが可能な場合が多いが，各クラスタリングアルゴリズムはある特定の観点に従い分類するため，ユーザの求める結果になるとは限らない．制約クラスタリングは，対象データの分類に関して，人間の持つ常識や背景知識などに基づき人手で制約を与える半教師あり学習であり，クラスタリングアルゴリズムは与えられた制約を満たすようにデータを分割する．代表的な制約表現に対制約があり [Basu 04]，must-link，cannot-link の 2 種類の対制約が一般に用いられる [寺見 10，wagstaff 01]．must-link を付与されたオブジェクト対は同じクラスタに，cannot-link を付与されたオブジェクト対は異なるクラスタに分類されるべきであることを意味する．

連絡先: 高間康史，首都大学東京大学院システムデザイン研究科，〒 191 － 0065　東京都日野市旭が丘　6 － 6，ytakama@sd.tmu.jp

2.2 クラスタ単位での制約生成

制約クラスタリングは多様なデータ分析に適用可能なアプローチであるが，多数の制約を付与しなくてはならない場合にユーザの負荷が問題となる．すなわち，対制約を与えるオブジェクト対を一つずつ選択しなくてはならないのでは，多数の制約を付与することは困難である．

この問題を解決するために，オブジェクト単位ではなくクラスタ単位でのインタラクションに基づき，複数制約を一括して指定可能なアプローチが提案されている [三宅 11]．この研究では，円として描画されたクラスタに対し分解（break），結合（merge）といった操作をインタラクティブに行うことができる．この研究における，インタラクティブなクラスタリングの作業フローを図 1 に示す．ユーザが制約クラスタリングを再実行したい場合，ユーザが新規に作成したクラスタ内の全オブジェクト対に must-link が生成される．提示されたクラスタリング結果に対しユーザがグルーピング作業を行い制約クラスタリングを実行するまでの一連のプロセスを本稿ではステップと呼ぶ．

3. 過去のグルーピング情報に基づく制約生成

2.2 節に示した手法では，ユーザがオブジェクトレベルでなくクラスタレベルで行なった操作結果に基づき制約を一括生成するため，効率よい制約付与が期待できる反面，不要なオブジェクト対に制約が付与されてしまう可能性もある．そのため，制約クラスタリングを反復的に実行しながら調整を行い，最終的に満足するクラスタを得ることを想定している．そこ

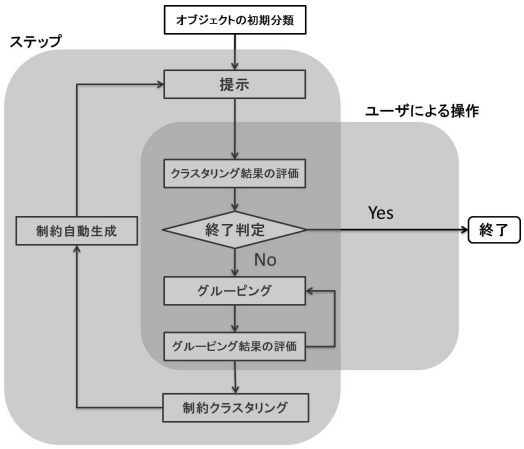

図 1: インタラクティブクラスタリングの作業フロー

で，本稿では，一回の操作結果だけでなく，過去のグルーピング操作も考慮して対制約を生成する手法について検討する．

オブジェクトの集合を $O = x_1,\ldots, x_n$ とし，過去のグルーピング情報を $c(x_i, x_j, t)$ とする．このとき $x_i, x_j \in O$ であり，ステップ t においてオブジェクト x_i, x_j が同一クラスタにまとめられたかどうかを管理する．

$$c(x_i, x_j, t) = \begin{cases} 1 & (x_i \text{と } x_j \text{が同一クラスタ}) \\ 0 & (x_i \text{と } x_j \text{が異なるクラスタ}) \end{cases} \quad (1)$$

また，must-link 制約の有無として $M(x_i, x_j)$ を定義し，次回クラスタリング実行時に，1 で制約が付与され 0 で制約が付与されないものとする．

ステップ t 終了後の制約 $M(x_i, x_j)$ を求めるために，まず過去 T ステップにおける x_i, x_j の同一クラスタへのグルーピング回数を式 (2) により求める．制約生成には，式 (2) の値を正規化した値を用いる．

$$f(x_i, x_j, T) = \sum_{1 \leqq k \leqq T} c(x_i, x_j, t-k) \quad (2)$$

グルーピング履歴と生成すべき制約の間には，複数の関係性が考えられる．ユーザが何度も同じクラスタにまとめるオブジェクト対は，同一クラスタにまとめるべきというユーザの意図を反映していると解釈することもできる．反対の考え方として，初めて同一クラスタにまとめられたオブジェクト対は，ユーザの新たな要望を表していると捉えることもできる．そこで本稿では，保守的，積極的の二通りのアプローチを検討する．

- 保守的 ： 同一クラスタへのグルーピング頻度の高いオブジェクト対ほど優先的に制約を付与
- 積極的 ： 同一クラスタへのグルーピング頻度の低いオブジェクト対ほど優先的に制約を付与

前者の場合，式 (1) の値が大きいオブジェクト対ほど制約として選ばれやすく，後者の場合には小さいほど選ばれやすくすれば実現できる．頻度上位のものから確定的に選ぶ他，ルーレット選択などで確率的に選択する手法も可能である．ただし，いずれの場合でも現ステップ (t) において $c(x_i, x_j, t) = 1$ となるオブジェクト対に限定することで，ユーザの現在の操作に反しない制約生成とする．

式 (2) において，T の値を変え，利用する過去のグルーピング情報量を増減させることで，生成される対制約の量も変化し，様々な分類結果を取得することができる．$T = 0$ とした場合は既存手法と同様，現在のステップのみを考慮することになり，$T = \infty$ とすれば過去全てのグルーピング履歴を考慮することとなる．パラメータ T が制約生成，およびユーザの作業に与える影響については今後ユーザ実験により検討する予定である．

制約生成例として，$O = x_1,\ldots, x_4$ ，$T = 2$，$c(x_i, x_j, t)$ が図 2 に示す値の場合を考える．

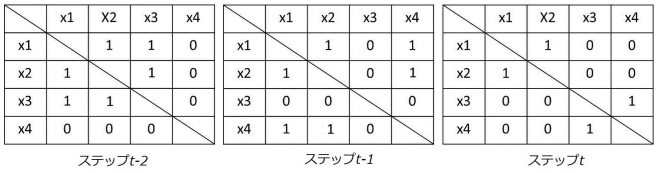

図 2: $c(x_i, x_j, t)$ の例

この時，ステップ t で同じクラスタに入っているオブジェクト対 $(x_1, x_2), (x_3, x_4)$ が must-link 付与の対象となる．保守的なアプローチでは，その中で式 (2) の値が最大となる (x_1, x_2) に制約が付与される可能性が高くなる．積極的なアプローチでは，過去 T ステップで同一クラスタにまとめられたことのない (x_3, x_4) に制約が付与される可能性が高くなる．

4. おわりに

本論文では，インタラクティブクラスタリングにおける対制約生成手法を提案した．提案手法は，考慮する過去ステップ数や積極・保守的アプローチの切り替えができる一般的な枠組みである．今後は，提案した対制約生成手法を実装し，インタラクティブクラスタリングを支援するユーザインタフェースと組み合わせユーザ実験を行うことで，制約生成方法の違いがユーザ作業に与える影響などについて研究を進める．

参考文献

[Basu 04] S.Basu, A.Banerjee., R.J.Mooney: Active Semi-Supervision for Pairwise Constrained Clustering, Proc. of the SIAM International Conference on Data Mining, (SDM-2004), pp.333-344, 2004.

[三宅 11] 三宅遼祐，山田誠二，岡部正幸，高間康史: インタラクティブクラスタリングのためのマルチタッチインタフェースの提案, 第 25 回人工知能学会全国大会，1J1-OS9-3，2011.

[寺見 10] 寺見明久，宮本定明: 階層的クラスタリングにおける対制約の導入のための二つのアプローチ, 第 26 回ファジィシステムシンポジウム，pp13-15，2010.

[Wagstaff01] K.Wagstaff, C.Cardie, S.Rogers,S.Schroedl: Constrained K-means Clustering with Background Knowledge, Proc. 18^{th} International Conf. on Machine Learning, pp. 577-584, 2001.