2F4-OS-04-3

グルーピング操作に基づくインタラクティブな対制約生成手法の考察

Consideration on Interactive Pairwise Constraint Generation Based on Grouping Operation

高間 康史 *1　　三宅 遼祐 *1
Yasufumi Takama　Ryosuke Miyake

*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

This paper studies the method for generating pairwise constraints from the sequence of user's grouping operations. Constrained clustering has been studied as one of promising approaches of interactive data mining. However, when it 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 has been proposed. This paper implements two types of constraint generation methods, conservative and sensitive methods, which are based on the sequence of user's past grouping operations. The effect of constraint generation method on user's behavior is examined through experiments with test participants.

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．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り，半教師あり学習の一つである制約付きクラスタリング [Basu04, Wagstaff01] などの活用が期待されている．

制約付き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 [三宅 11] が提案されている．制約の一括生成はユーザの負荷軽減に効果が期待できる反面，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．この問題に対処するため，一回の操作結果だけでなく，反復的なグルーピング作業を想定しその操作履歴を考慮して must-link 制約を生成する手法が提案されている [三宅 12]．本稿では，文献 [三宅 12] で提案されている積極的，保守的の二つの制約生成戦略をプロトタイプインタフェースとして実装し，制約生成手法の違いがインタラクティブクラスタリングにおけるユーザの振る舞いに与える影響について，ユーザ実験により考察する．

2. 関連研究

2.1 制約付きクラスタリング

クラスタリングが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．制約付き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 [Basu04]，must-link，cannot-link の 2 種類の対制約が一般に用いられる [Wagstaff01]．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，Tel/Fax. 042-585-8629，ytakama@sd.tmu.ac.jp

2.2 クラスタ単位での制約生成

制約付き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，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，ユーザの負荷が問題となる．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．ユーザが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し must-link が生成される．

この手法では，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．そこで，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，一連のグルーピング操作を考慮して対制約を生成する手法が提案されている [三宅 12]．

オブジェクトの集合を $O = \{x_1, \ldots, x_n\}$ とする．過去のグルーピング情報を $c(x_i, x_j, t)$ とし，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．

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

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，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．

$$f(x_i, x_j, T) = \sum_{0 \le k < t} c(x_i, x_j, t-k) \quad (2)$$

グルーピング履歴に基づく対制約生成戦略として，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する保守的戦略と，頻度の低いオブジェクト対ほど優先的に制約を付与する積極的戦略を考える．前者の場合，$f(x_i, x_j, T)^2$ の値に基づくルーレット選択で一定数の対を制約として選択する．後者の場合，$\{T - f(x_i, x_j, T) + 1\}^2$ の値に基づくルーレット選択を行う．ただし，どちらも現ステップ (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．

3. 評価実験

保守的，積極的の 2 戦略を実装したプロトタイプインタフェースを用いて，工学系大学・大学院生 10 名にグルーピング作業を行ってもらった．実験には，関西大学ビジネスマイニング研究センターで公開されていた焼肉店の売上に関する実データから，以下に示す二つのデータセットを作成して用いた．

[**D1**] アイテム数 206，トランザクション数 64

[**D2**] アイテム数 213，トランザクション数 73

実験協力者には制約なしでのクラスタリング結果を提示し，それを修正しながら D1 の場合は 3 つ，D2 の場合は 2 つのグループを作成してもらった．作成後にはどのようなアイテムを含むグループを作ろうとしたのかをアンケートに記入してもらった．ユーザのグルーピング操作に基づき対制約を自動生成して再クラスタリングを行った結果を提示し，上記作業を 10 回反復してもらった．制約付きクラスタリングには COP-KMEANS[Wagstaff01] を用いている．式 (2) における T の値は 4，生成する対制約数はグループサイズに比例し最大 5 としている．

制約付きクラスタリングにおいてユーザの意図が反映されたクラスタが得られたかどうかを評価するために，本稿ではリフト値を指標に用いた．リフト値は相関ルールの評価指標に用いられており，本稿では以下の様に定義する．

$$\mathrm{lift}(c, g) = p(g|c)/p(g) \tag{3}$$

ここで，g はユーザがグループ化したいアイテムの集合，c は制約付きクラスタリングにより生成されたクラスタを表し，$p(g)$ はデータセット中で g に属するアイテムを含むトランザクションの割合，$p(g|c)$ は c 中で g に属するアイテムを含むトランザクションの割合とする．これにより，リフト値が大きいほど，ユーザが着目した種類のアイテムを含むトランザクションがそのクラスタ内に集まったことを表す．この値を用いて，ユーザが意図したグループ毎に，最大のリフト値をとるクラスタが異なれば，ユーザの意図するアイテムが集められたクラスタがそれぞれ生成されたと考える．また，意図したグループが一つしかない場合には，手動でのグルーピング結果におけるリフト値の最大値よりも再クラスタリング結果の最大値の方が高い場合に，ユーザの意図が反映されたと考える．

実験協力者 10 人を 5 人ずつ，データセットとアプローチの組み合わせに関し表 1 に示す A，B の二グループに分けて実験を行った．実験順序は共に D1，D2 の順としている．各グループについて，グルーピング作業を行った回数，そのうちユーザ意図が反映されたと見なせる回数およびその割合を表 1 に示す．本来グルーピング作業回数は全て 50（5 人 ×10 回）となるはずだが，システムの不具合により中断した場合があったため 50 回に満たない場合がある．

表 1: ユーザ意図を反映したクラスタリング結果の割合

グループ	条件	作業回数	反映	割合
A	D1・保守的	45	29	0.65
	D2・積極的	50	35	0.7
B	D1・積極的	46	21	0.45
	D2・保守的	50	22	0.44

表より，保守的・積極的の両戦略について，グループ A の方がユーザ意図を反映した結果が得られていることがわかる．両グループの違いとして，実験後のアンケートで制約生成戦略の違いを意識して作業を行ったかどうかを質問したところ，グループ A は全員が意識したと回答したのに対し，グループ B では全員が意識しなかったと回答している．戦略を意識したかどうかが実際の作業にどのように影響したかの分析を今後行う必要があると考えるが，システム側の振る舞いを意識することが計算機と人間の協調作業においては重要であることが示唆される結果と言える．

実験においては，肉を含むグループ，野菜を含むグループ，ドリンクを含むグループなどのおおまかなくくりによるものが多く見られた他，キムチとウーロン茶を含むグループなど，より具体的なグルーピングを試みた協力者もいた．また，グループ A では 10 回全てで同一の意図に基づきグルーピングを行う場合が比較的多かったのに対し，グループ B では途中で変更する場合が A よりも多く見られた．この点も，戦略に対する意識に関係する可能性もあり，今後分析を行う必要があると考える．

4. おわりに

本稿では，ユーザが反復的に行うグルーピング作業から対制約を一括生成するアプローチについて，制約生成戦略の違いがユーザの振る舞いに与える影響について考察した．実験の結果，戦略を意識した実験協力者の方が意図を反映したクラスタリング結果が得られる傾向にあることを示した．今後は作業ログをより詳細に分析することにより，インタラクティブクラスタリングにおける対制約一括生成手法の有効性や，適切な戦略についての検討を進める予定である．

参考文献

[Basu04] 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.

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

[三宅 11] 三宅, 山田, 岡部, 高間, 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, JSAI2011, 1J1-OS9-3, 2011.

[三宅 12] 三宅, 山田, 高間, 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, JSAI2012, 4F1-OS-5-5, 2012.