人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第11回)
SIG-AM-11-01

従属クラスタ動的生成機構の導入による Must-Link 制約付き K-means 法の拡張に関する提案

Proposal of Must-Link Constrained K-means with Dynamic Generation of Subordinate Clusters

井本博之 [1]　高間康史 [1]

Hiroyuki Imoto[1], Yasufumi Takama[1]

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

Abstract: This paper proposes to extend must-link constrained K-means clustering by introducing dynamic generation of subordinate clusters. When clustering high-dimensional data there is a case where data which should belong to the same cluster form several distinct groups in a data space. In order to handle such a case without using distance metric learning, the proposed method generates subordinate clusters for each data group, which are merged after finishing K-means clustering. Result of a comparison experiment with a baseline method shows the effectiveness of the proposed method in terms of success rate and NMI (normalized mutual information).

1. はじめに

本稿では, Must-Link 制約を利用して従属クラスタを生成する機構を導入した制約付き K-means クラスタリング手法を提案する．クラスタリングはデータ群を類似した複数のグループに分ける操作であり，データ全体を俯瞰的にみる目的でデータマイニングの初期分析などによく用いられる．一般的なクラスタリングアルゴリズムは正解データを必要としない教師なし機械学習であるが，自動生成されるクラスタではユーザの要求する結果を得られない場合が多く存在する．そのため，近年ではユーザの意思をクラスタリング結果に反映させる目的で，ユーザフィードバックを利用して半教師あり機械学習を行う制約付きクラスタリングが研究されている．

制約付きクラスタリングで一般的に用いられる制約形式の 1 つに対制約があり，データ対が同一のクラスタに属すべきであるという Must-Link 制約と，データ対が異なるクラスタに属するべきであるという Cannot-Link 制約の 2 種類から構成される．対制約は様々なクラスタリング手法に適用可能[1]な他，インタラクティブに効率的な対制約付与を行うシステム[2][3]が提案されている．

対制約を利用した制約付きクラスタリングの手法としては, CCL (Constrained Complete-Link) [4]のような距離ベースのものと，COP K-means [5]のような制約ベースのものが提案されている．距離ベースの手法では, Must-Link 制約を付与されたデータ対は近くあるいはデータ間距離が 0，Cannot-Link 制約を付与されたデータ対は遠くあるいはデータ間距離が∞になるようなデータ空間に写像した後にクラスタリングを行う．距離ベースの手法は，K-means などの従来クラスタリング手法で制約を満たすクラスタを求めることができるが，元の距離空間とは異なる空間におけるクラスタリングとなるため結果の解釈が困難となる場合がある．結果のクラスタがどのような意味を持つかという解釈は実際にデータ分析を行う場合，非常に大きな意味を持つと考えられる．さらに，距離行列を計算するには，データ数 N に対し $O(N^2)$ の計算量が必要となるため大量データの分析には多大な時間的コストがかかることが問題となる．

一方，制約ベースの手法では，与えられた対制約をそのまま満たしながら目的関数を最適化してクラスタリングを行う．代表的手法である COP K-means は，単純かつ高速なクラスタリングアルゴリズムとして実用的に良く用いられる K-means に対制約を導入した手法であり，クラスタ割当て時に Must-Link 制約と Cannot-Link 制約の全てを満たすクラスタの中で，最も距離の近いクラスタにデータを割り当てる手法である．空間の写像などは行われないため，

クラスタリング結果についてデータそのものの属性に基づいた解釈が可能である．また，計算量の観点ではCOP K-meansの計算オーダが $O(N)$ となるため距離ベースの手法より優れている．しかしながら，地理的に離れた場所に同種データが存在する様なデータセットや，文書のように高次元のデータで同一クラスタにしたいデータが空間上の一カ所に集中していない場合などは，同一クラスタにまとめられるべきデータが異なる領域に分かれて存在することが考えられる．そのような場合に，データ空間内で離れた位置にあるデータ間に Must-Link 制約が付与された場合などはCOP K-meansでは良好な結果が得られないことがある．例えば，図 1 に示す様に，両端にあるデータグループ間に Must-Link 制約が付与された場合，COP K-means (K = 2) では図 2 に示した様なクラスタが得られてしまう．

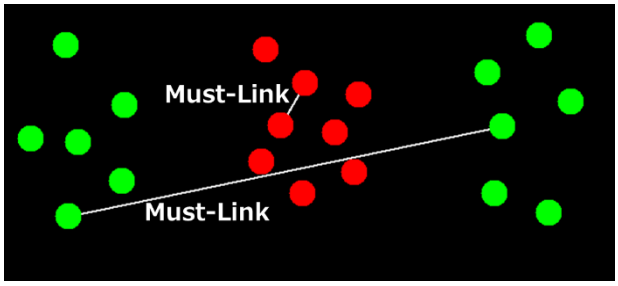


図 1．2 次元人工データセット

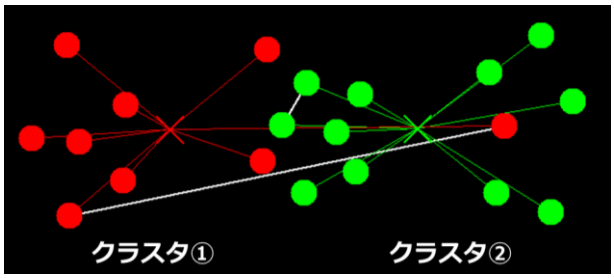


図 2．COP K-means の実行結果

提案手法ではCOP K-meansをベースとし，同一クラスタにまとめられるべきデータが複数領域に分かれて存在する場合には，それぞれの領域に対応した従属クラスタを動的に生成する．クラスタリング終了後，Must-Link によって繋がれた従属クラスタを統合することによって，距離学習せずにクラスタリング結果を求める．Must-Link 制約のみを含む人工データセットを用いて提案手法と COP K-means との比較実験を行った結果，NMI，クラスタリング成功率ともにCOP K-meansよりも良好な結果であることを示す．

2. 関連研究

本節では提案手法のベースとなるCOP K-meansについて述べ，関連研究として距離学習を用いている岡部ら[6]の制約付きグラフカットによる逐次クラスタリング手法を取り上げ，提案手法との違いについて述べる．

2.1. COP K-means クラスタリング

COP K-means は制約ベースの代表的手法であり，対制約をインスタンスレベルで K-means に組み込んでクラスタリングを行う．その基本アルゴリズムは以下の通りとなる．

① k 個のクラスタの初期値を設定する．
② データに付与された対制約を全て満たすクラスタのうち最も近いクラスタに各データを割当てる．
③ 各クラスタの重心位置を計算する．
④ ②，③を繰り返し，②の前後で所属クラスタに変更がなくなった時点で終了とする．

ただし②の処理の時，対制約のペアとなるデータのクラスタ再割り当てが行われていない場合はその制約を無視する．データに対制約が付与されていない場合は K-means と同様に，最も重心との距離が近いクラスタに割当てる．なお，全ての対制約を満たすクラスタが存在しなかった場合，強制終了となる．

COP K-means では制約数が多くなればなるほど計算量が大きくなるものの，計算オーダは $O(N)$であり高速なクラスタリングが期待できる．しかしながら，クラスタリングが正常に終了するかについては順序に大きく依存することや，正しい制約を付与したにも関わらずクラスタリング精度が落ちる場合がある[7]など，いくつか問題が指摘されている．

2.2. 制約付きグラフカットによる逐次クラスタリング

岡部らは，目的関数と制約条件を半正定値計画問題 (SDP：Semi-Definite Programming) で定式化し距離学習を行う手法を提案している[6]．アルゴリズムの概要を以下に示す．

① データ集合から非類似度に応じた重み付き隣接行列を作成する．
② 隣接行列からグラフラプラシアンを作成し，

最大グラフカット問題を定式化する．

③ 最大グラフカット問題を SDP による緩和問題として再定式化し，対制約条件を組み込む．

④ SDP を解いて得られた解行列を基にデータ集合を 2 分割する．

⑤ 生成されたクラスタのうち最大のデータ数を持つクラスタを選択し，②～④を繰り返して 2 分割操作を行う．

⑥ ⑤の操作をあらかじめ設定したクラスタ数になるまで繰り返す．

なお，岡部らの手法では Cannot-Link 制約は用いず，Must-Link 制約のみを用いている．これは，初回の 2 分割から Cannot-Link 制約も無理に満たそうとするため，Cannot-Link 制約が複数存在するとクラスタリングに悪影響を及ぼす可能性があるためである．Must-Link 制約のみを用いた COP K-means との比較実験を行った結果，NMI の値は複数のデータセットに対して互角もしくは優位な結果であったとしているが，計算時間は COP K-means が圧倒的に良い結果を示している．これは，①における隣接行列の計算に $O(N^2)$ かかることに加えて，SDP の処理にも多くの計算コストがかかるためである．

3. 従属クラスタ生成機構を持つ制約付きクラスタリング

3.1. クラスタリングアルゴリズムの概要

提案手法では COP K-means を拡張し，1 節で示した図 1，2 のように離れたデータ間に張られた Must-Link 制約によってデータが距離の遠いクラスタに割り当てられそうになった場合，動的に従属クラスタを生成する機構を導入する．さらに，クラスタリング終了後，Must-Link で繋がれたデータを含むクラスタ同士を 1 つに統合することにより，複数の領域に存在するクラスタを生成する．

提案手法のフローチャートを図 3 に示す．提案手法では K-means の初期値依存の影響を避けるため，1 回目のクラスタ割当てでは従属クラスタの生成は行わない．また，提案手法及び COP K-means では，ある領域にクラスタが集中した場合，Must-Link 制約により距離の近いクラスタ同士でデータの奪い合いが起こり，クラスタリングが収束しないことがあることを予備実験で確認したため，重心計算とクラスタ割当てのループ回数 *step* が閾値 *L* を超えた場合クラスタリングを終了とする．さらに，Must-Link 制約によるクラスタ統合のみではあらかじめ指定したクラスタ数とならない場合があり，その場合はクラスタ重心の距離が近いクラスタを統合する凝集型クラスタ統合を併用する．従属クラスタ生成機構，Must-Link クラスタ統合，凝集型クラスタ統合の詳細については 3.2，3.3，3.4 節にそれぞれまとめる．

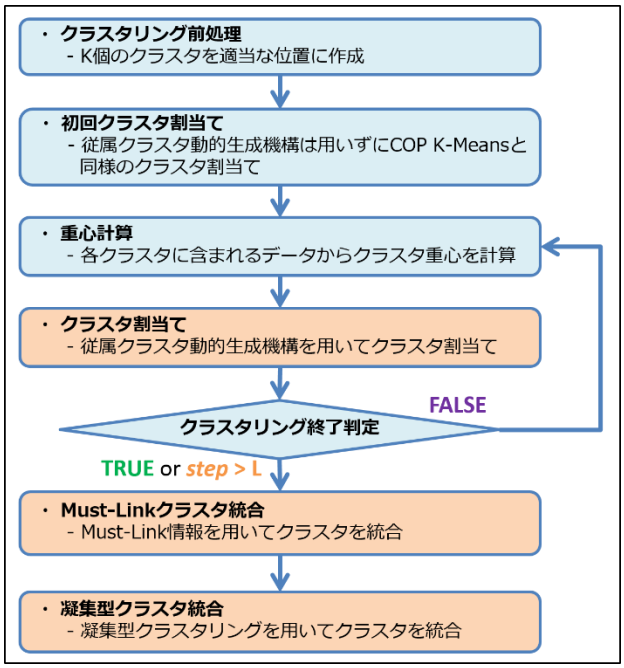


図 3. 提案手法のフローチャート

3.2. 従属クラスタ動的生成機構

提案手法では，図 4 のようにデータ x が Must-Link 制約によって距離の遠いクラスタ c_1 に割当てられそうになった場合，x の位置にクラスタ c_s を生成し，その後 x は c_s に固定割当てとする．距離の近い/遠いの判定には閾値 *th* を用い，距離が *th* よりも大きい場合にクラスタ生成を行う．ただし，同じ位置におけるクラスタ生成は行うべきではないという考えから，1 つのデータが行えるクラスタ生成は 1 回までに制限する．また，クラスタ生成を行うと次のクラスタ割当てによって各クラスタの位置が大きく変動することが考えられるため，1 回のクラスタ割当て時に行えるクラスタ生成も 1 回に制限する．

従属クラスタ動的生成機構を用いてデータ x のクラスタを決定する手続きを図 5 に示す．各データ x について，x に対する従属クラスタ生成が行われてなく，かつ Must-Link 制約が付与されている場合にクラスタ生成を行う（8 行目以降）．SEARCH_CANDIDATECLUSTER(x, C) により既存クラスタから割当て候補クラスタ集合 *CC* を求め（8

行目)，その中で最も近いクラスタに x を割り当てる（9 行目）．ただし，そのクラスタと x の距離が閾値以上の場合は初回クラスタ割当て時を除き従属クラスタを生成する (13 行目)．また，効率の良い探索を行うため，クラスタ割当て時に Must-Link 制約をクラスタに登録する (3，15 行目)．

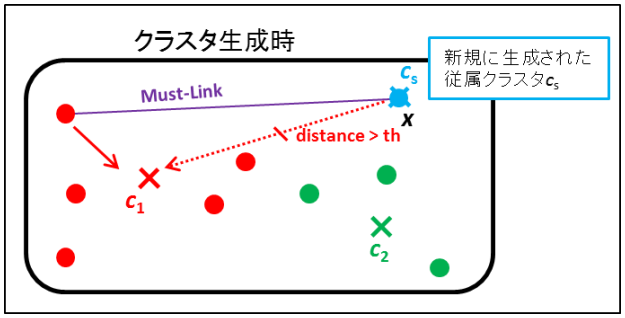


図 4．従属クラスタ生成時の様子

```
1  if x.fixed == TRUE {
2    x.curc = x.prec;
3    x.curc.REGISTER_MUSTLINK(x.mlg);
4  else{
5    if x.mlg = NULL{
6      x.curc = MOST_NEARCLUSTER(x, C);
7    }else{
8      CC = SEARCH_CANDIDATECLUSTER(x, C);
9      x.curc = MOST_NEARCLUSTER(x, CC);
10     if DIS(x, x.curc) > th
11       & step > 1
12       & crflg == FALSE{
13       x.curc = CREATE_CLUSTER(x);
14     }
15     x.curc.REGISTER_MUSTLINK(x.mlg);
16   }
17 }
```

図 5．従属クラスタ動的生成機構を用いたクラスタ割当て

SEARCH_CANDIDATECLUSTER(x, C) では，図 6 のようなクラスタ間の Must-Link による接続関係を利用して割当て候補クラスタ集合を求める．各 *mlg* は Must-Link 制約によって直接繋がったデータ集合を表しており，例えばデータ a と b の間に Must-Link 制約があり，a，b をそれぞれ含むクラスタ $c1$, $c2$ がある場合，a，b を含む *mlg* と $c1$，$c2$ が接続される．図 6 の例では，x の所属する *mlg*1 にはクラスタ $c1$, $c2$ に割り当てられたデータが所属しており，$c1$ に割り当てられたデータには *mlg*1，*mlg*2，*mlg*3 に所属するものが存在している．このような接続関係を x の所属する *mlg*1 を起点として全探索し，得られたクラスタ集合を候補クラスタ集合として出力する．

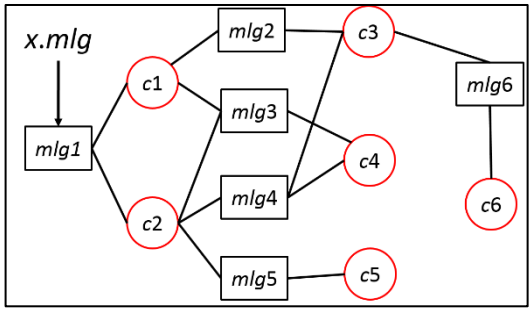
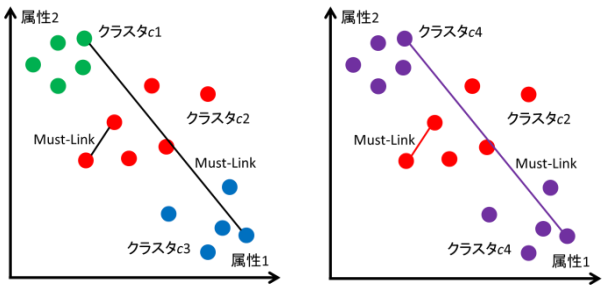


図 6．データ x と各クラスタとの接続関係

3.3. Must-Link クラスタ統合

3.1 節で述べたように提案手法では，クラスタリング終了後に Must-Link クラスタ統合を行う．クラスタ $c1$ 内のデータが他のクラスタ $c2$ 内のデータと Must-Link 制約で繋がっている場合，$c1$ と $c2$ を統合する．例えば，クラスタリング終了時の状態が図 7 (a) の様であった場合，クラスタ $c1$ とクラスタ $c3$ の間に Must-Link 制約が張られているため両者は統合され，図 7 (b)に示すクラスタ $c4$ が形成される．このクラスタは，「属性 1 あるいは属性 2 が排他的に大きい」という特徴を持つクラスタであると解釈できる．



(a) 統合前　　(b) 統合後

図 7．Must-Link クラスタ統合の例

なお，Must-Link クラスタ統合で統合されるクラスタ集合は，3.2 節の図 6 で示した様な接続関係にあるクラスタの集合であり，図 5 に示された SEARCH_CANDIDATECLUSTER() によって求められる．

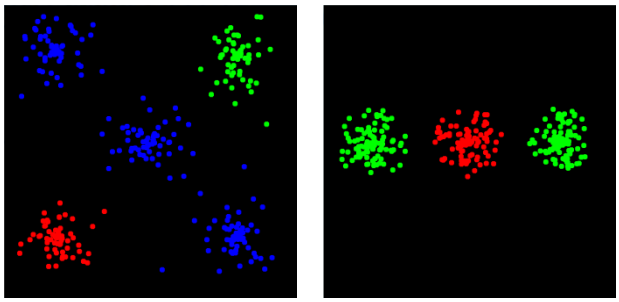
3.4. 凝集型クラスタ統合

3.1 節で述べたように，3.3 節の Must-Link クラスタ統合のみでは初期設定したクラスタ数 K 以上のクラスタ数となってしまう場合がある，これを補うた

め凝集型クラスタ統合を行う．凝集型クラスタ統合では Must-Link クラスタ統合後の各クラスタ中心をデータとし，凝集型クラスタリング (AHC) の最短距離法を適用してクラスタ数が K となるまで統合する[8]．凝集型クラスタ統合は，クラスタ生成過多により，本来は一つにまとめられるべきデータ集合が複数のクラスタに分割されてしまっている状態を修正することを目的とするため，鎖効果を期待して最短距離法を採用する．

4. 実験

図 8 に示すデータ数 300 のデータセット A，B に対して COP K-means 及び提案手法を用いてクラスタリングを行い，比較実験を行った．図において，同じクラスタに属するデータは同じ色としている．評価指標には正規化相互情報量 (NMI : normalized mutual information) を用い，NMI = 1.0 となる場合をクラスタリング成功とし, その割合を成功率とした．



a. データセット A　　b. データセット B

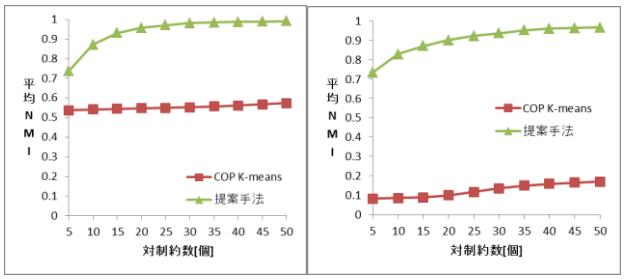
図 8. 実験に使用した 2 次元人工データセット

対制約は両手法ともに Must-Link 制約のみを用い，正解データペアに対して，5, 10, 15, 20, 25, 30, 35, 40, 45, 50 個をランダムに付与し，各 10000 回クラスタリングを行い平均値及び正解率を算出した．なお，データの範囲は各次元 [0，700] とし，提案手法における従属クラスタ動的生成機構の閾値 th は距離の 2 乗値に対して設定するため，データセット A に対しては th = 60000，データセット B に対しては th = 30000 とした．この値は予備実験の結果，各データセットにおいて良い結果が得られたものを選択している. なお, 全手法に対して最大ループ回数 L は 100 回と設定した．

提案手法と COP K-means について，図 9 に平均 NMI，図 10 に成功率，図 11 に平均実行時間を比較した結果をそれぞれ示す．また，図 12 に提案手法における平均最終クラスタ数の推移を示す．データセット A，B に対して，平均 NMI，成功率共に，COP K-means よりも提案手法の方が高い値を示している．特にデータセット B においてその傾向はより顕著であり，線形分離可能でない場合に有効性が期待できると考える．

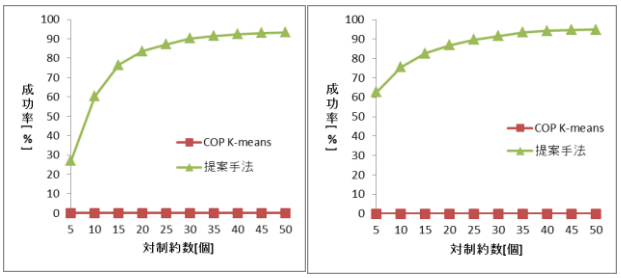
平均最終クラスタ数は対制約数の増加に伴い，データセット A，B ともにわずかな上昇を示しているが，収束の傾向もみられる．この現象に対する検証にはより大きなデータセットにおける実験が必要であり，また閾値の設定とも関連すると考える．

平均実行時間に関して，提案手法では対制約数の 2 乗オーダで増加している．このように計算量が増加するのは，対制約数の増加により生成される従属クラスタ数および mlg の数が増えた結果, 3.2 節に示した SEARCH_CANDIDATECLUSTER()などの計算時間が増加することが原因と考えられ，今後検証を行う予定である．しかしながら，制約付きクラスタリングにおける制約はユーザに付与されることが想定されており，大量に付与されるケースは少ないため，大きな問題とならないと考える．ただし，高間ら[3]のように複数の対制約を自動生成するインタフェースと組み合わせる場合には，制約生成数を抑制するなどの対策が必要と考える．



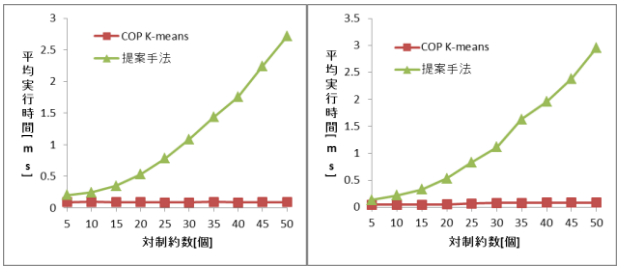
(a) データセット A　　(b) データセット B

図 9. 平均 NMI の推移



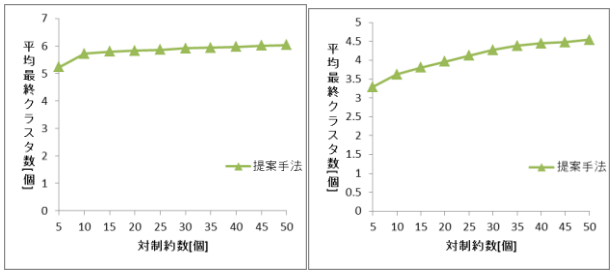
(a) データセット A　　(b) データセット B

図 10. 成功率の推移



(a) データセット A　　(b) データセット B

図 11. 平均実行時間の推移



(a) データセット A　　(b) データセット B

図 12. 平均最終クラスタ数の推移

5. まとめ

本稿では，同一クラスタにまとめられるべきデータが空間上の複数の領域に分かれて存在するケースにも対応できるように，COP K-means に対し Must-Link 制約を基にした従属クラスタ動的生成機構を導入した拡張手法を提案した．2 次元人工データを用いた比較実験により，同一クラスタに属するデータが平面上の異なる領域に分散して存在するような場合に，COP K-means よりも NMI, 成功率共に良好な結果が得られることを示した．また，計算量に関しても対制約数の 2 乗オーダで上昇してしまうものの，データ数 *N* に対しては K-means と同様であるため，距離学習を利用したクラスタリングなどに比べ，高速なクラスタリングが期待できる．距離学習を用いた場合とのクラスタリング精度の比較は今後行う予定である．また，提案手法の特徴は，得られたクラスタの解釈が元の空間上で行えることであり，その利点についてもユーザ実験により検証する予定である．

提案手法では従属クラスタ動的生成機構に対して閾値 *th* を指定する必要があり，適切な閾値をいかに決定するかが今後の課題となる．また，Cannot-Link 制約も利用可能なように拡張することも検討している．

参考文献

[1] 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, FSS2010, MD2-4，2010.

[2] 山田誠二，水上淳貴，岡部正幸：インタラクティブ制約付きクラスタリングにおける制約選択を支援するインタラクションデザイン，人工知能学会論文誌 Vol.29 No.2，pp. 259-267，2014.

[3] 高間康史，三宅遼祐：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，第 27 回人工知能学会全国大会，2F4-OS-04-31，2013.

[4] D.Klein ， S.D.Kamvar ， C.D.Manning ：“ From Instance-level Constraints to Space-level Constraints : Making the Most of Prior Knowledge in Data Clustering,” in Proc. International Conference on Machine Learning (ICML-2002)，pp. 307-314，2002.

[5] K.Wagstaf, C.Cardie, S.Rogers, S.Schroedl:“Constrained K-means Clustering with Background Knowledge,” in Proc． International Conference on Machine Learning (ICML-2001)，pp. 577–584，2001.

[6] 岡部正幸，山田誠二：制約付きグラフカットによる逐次クラスタリング，人工知能学会論文誌 Vol.27, No.3，pp. 193-203， 2012.

[7] I.Davidson ， K.Wagstaff ， S.Basu ：“ Measuring Constraint-Set Utility for Partitional Clustering Algorithms,” in Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD2015)，pp. 115–126，2006.

[8] 宮本定明：クラスター分析入門 ファジィクラスタリングの理論と応用，森北出版，pp. 88-105，1999.