ARG WI2 No.9, 2016
WI2-2016-29

# 従属クラスタ動的生成機構を導入した Must-Link 制約付き K-means の提案

井本博之 [1]　岡部正幸 [2]　高間康史 [1]

[1] 首都大学東京大学院 システムデザイン研究科, [2] 県立広島大学 経営情報学部

*ytakama@tmu.ac.jp*

**概要**　近年, 計算機能力の向上とエンドユーザーによる情報発信量の増加によって大規模データの活用需要が高まり, 文書や画像などの様々なデータに対するクラスタリング需要が増加している. しかし, クラスタリングを行う際, 多種多様なデータに対する人間と機械との間に存在する意味解釈の違いにより, 同一クラスタに所属させたいデータが空間内で複数のグループに分かれて存在し上手くクラスタリングを行えない場合が考えられる. 提案手法では, クラスタリング中の従属クラスタ生成と、Must-Link を利用したクラスタ統合によりこの問題に対応する. また, 2 種類の人工データを用い, グラフカットに基づく従来手法などと比較実験を行った結果より, 提案手法の有効性を示す.

**キーワード**　クラスタリング, COP　K-means, セマンティックギャップ

## 1 はじめに

本稿では, 従属クラスタ動的生成機構を導入した Must-Link 制約付き K-means を提案する. データマイニングの初期分析などにおいて, 簡単かつ素早くクラスタリングを行う手法として K-means が広く知られている. また, ユーザの意思をクラスタリング結果に反映させるための制約付きクラスタリングも広く研究されている[1]. しかし, 依然として高精度かつ高速なクラスタリングを行うことは困難である. 原因の一つとしてセマンティックギャップと呼ばれる機械と人間の間にある意味解釈の違いがある. 視覚や聴覚によって直接データを認識することができる人間に対し, 数値でしかデータを認識できない機械は人間がデータから得る情報の一部しか扱えないことが多い. そのため, 計算機が扱う特徴空間上でユーザが同一クラスタに所属させたいデータが複数のグループに分かれて存在してしまう場合などが存在する. 提案手法では, 従属クラスタ動的生成機構を制約付き K-means に導入することでこのケースに対応する. また, 他の制約付きクラスタリング手法と比較実験を行うことにより, 提案手法の有効性を示す.

## 2 関連研究

制約付きクラスタリングで一般的に用いられる制約の一つに対制約があり, データ対を同じクラスタに属させる Must-Link とデータ対を異なるクラスタに属させる Cannot-Link の 2 種類がある. 対制約を利用した制約付きクラスタリング手法は様々なものが提案されているが, ここでは COP K-means (CKM), GCUT, BCKM について紹介する.

CKM[2]は付与された制約を常に満たしながら K-means を行う手法である. 制約を組み込むことによって K-means と同じく高速ながらもより高い精度が期待できる. しかしながら, Must-Link と Cannot-Link の両方を同時に満たそうとするため, クラスタ割り当て時に解が存在せず強制終了となる場合がある点や, 正しい制約を付与した場合でもクラスタリング精度が落ちることがある[3]点などの問題がある.

GCUT[4]は最大グラフカット問題に Must-Link の制約条件を組み込み定式化したものを SDP として解き, 得られた解行列を基にデータ群を逐次的に 2 分割していくことでクラスタリングを行う. K-means が同サイズかつ超球状のクラスタを想定しているのに対し, GCUT ではそのような想定をもっていないため, クラスタサイズの異なる結果が想定されるデータ群にも対応可能である. しかし, 距離行列の計算や SDP ソルバーの処理に多くの計算コストがかかるというデメリットがある.

BCKM[5]は, CKM を弱学習器とみなし, ブースティングを行う手法である. BCKM では Adaboost アルゴリズム[6]を用いて各制約の優先度を更新しながら COP K-means を繰り返す. この時優先度の高い制約を張られたデータからクラスタ割り当てを行い, 途中で発生する制約の矛盾を無視することで強制終了が起こらないよう CKM を修正している. 各クラスタリングの結果と誤差率を用いてカーネル行列を作成し, カーネル K-means を行うことで最終的なクラスタリング結果を得る.

## 3 提案手法

提案手法では制約付き K-means を行う際, 破壊的クラスタ割り当てを検知した場合に従属クラスタを動的に生成する. 破壊的クラスタ割り当ては, 内部結合を達成

するため重心位置の近いクラスタにデータを割り当てるはずが，Must-Link によってデータを距離の遠いクラスタに割り当ててしまうことを指す．このような場合，内部結合度が下がるため良いクラスタリングにならないと考えられる．そこで提案手法では 閾値 *th* を用いて図 1 のように破壊的クラスタ割り当てを検出した場合，データ $x$ はクラスタ $c_1$ には割り当てず，新たに生成した $x$ と同座標にクラスタ重心を持つ従属クラスタ $c_s$ に割り当てる．

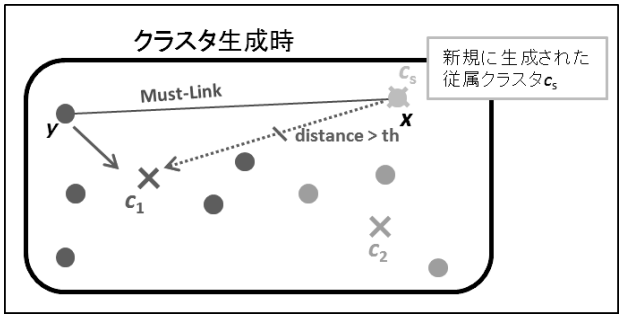
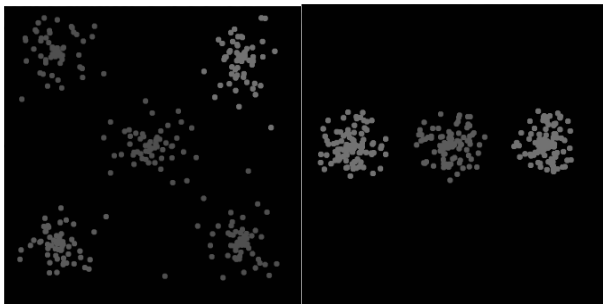



図 1. 従属クラスタ生成時の様子

クラスタリング終了後，Must-Link で繋がったデータが所属するクラスタ同士を統合する．しかし，Must-Link 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，その場合は凝集型クラスタ統合を補足的に行う．

## 4 比較実験

図 2 に示すデータ数 300，X，Y 軸の値の範囲[0, 700] のデータセット A，B に対して，2 節で述べた BCKM, GCUT, CKM(COP K-means), および提案手法を適用し比較実験を行った．図において，同一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 NMI を用いた．

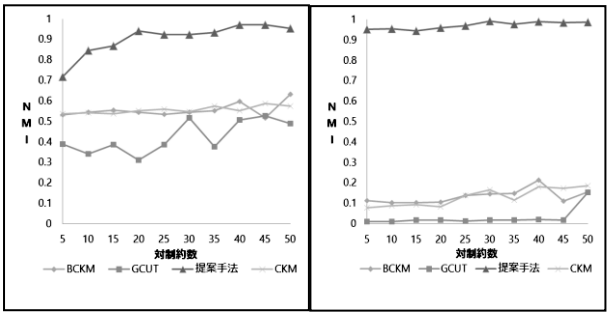


a. データセット A　　b. データセット B

図 2. 実験に使用した 2 次元人工データセット

対制約数は全手法とも 5, 10, ..., 50 の 10 パターン用意し，各対制約数におけるデータ対パターン及び順序は手法間で全て統一した．なお，提案手法とCKM は初期クラスタ依存性を考慮し 10000 回の平均 NMI を用い，提案手法の閾値 $th^2$=140000 とした．なお，$th^2$ は予備実験を行った際，最も良い結果となった閾値を採用した．

図 3 に対制約数を増やしていった場合の NMI の推移を示す．提案手法が他の手法に比べて最も良好な結果を示していることわかる．特に正解クラスタが不連続に存在するデータセットB において他手法との差が大きく，不連続なクラスタを持つデータ群に対して提案手法が特に有効といえる．また，少ない対制約数でも高い NMI を示しており，高い効果が期待できる．




(a) データセット A　　(b) データセット B

図 3. NMI 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CKM を拡張した手法を提案した．2 次元人工データセット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異なる領域に分散して存在するような場合に，CKM，GCUT，BCKM よりも NMI の高い結果が得られることを示した．提案手法の適切な閾値はデータの範囲や次元数，クラスタサイズなどデータセットのパラメータによって変化すると考えられるため，今後はデータセット毎に適切な閾値を決定する方法について検討する予定である．

## 参考文献

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

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

[3] I. Davidson, K. L. Wagstaff, S. Basu: Measuring constraint-set utility for partitional clustering algorithms, Proc. Knowledge Discovery in Databases(PKDD-2006), pp. 115–126, 2006.

[4] 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.

[5] M. Okabe, S. Yamada: Clustering by Learning Constraints Priorities, Proc. the IEEE International Conference on Data Mining (ICDM 2012), pp.1050-1055, 2012

[6] Y. Freund, R. E. Schapire: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, Journal of Computer and System Sciences (JCSS-1997), vol.55, pp. 119–139, 1997.