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26回)
SIG-AM-26-08

Must-Link 制約付き K-means における情報量規準に基づく従属クラスタ動的生成機構の提案

Proposal of Dynamic Generation of Subordinate Clusters Based on Information Criterion for Must-Link Constrained K-means

坂谷内 駿 [1]　柴田 祐樹 [1]　高間 康史 [1]
Shun Sakayauchi[1]　Hiroki Shibata[1]　Yasufumi Takama[1]

[1] 東京都立大学 システムデザイン学部
[1] Faculty of System Design, Tokyo Metropolitan University

Abstract: 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 を拡張し情報量規準に基づき従属クラスタを生成する手法を提案する. 従属クラスタを生成することで遠い位置にあるデータ点間の対制約を処理する従来手法では, 対制約の長さに関する閾値を手動で設定する必要があった. この課題に対し提案手法では, 情報量規準に基づき従属クラスタの生成を判断する. 対数尤度の計算手法などが異なる複数の手法を提案し, 比較実験によってその有効性や特性を検証する.

1 はじめに

本稿では, 従属クラスタ動的生成機構を組み込んだ Must-Link 制約付き K-means を拡張し, 情報量規準に基づき従属クラスタを生成する手法を提案する.

データマイニング手法の一つであるクラスタリングは, 似ているデータ同士をまとめることで複数のグループにデータを分割する. また, 教師無し学習であるクラスタリングに分析者によるフィードバックを導入した制約付きクラスタリングが提案されており, 代表的なものに K-means に対制約を導入した COP K-means[1] がある. この手法の課題の一つとして, データ空間上で遠く離れたデータ間に Must-Link 制約を付与した場合に, クラスタリング結果に大きな影響を与える点がある. そのような対制約に基づくクラスタ形成を破壊的クラスタ割り当てと定義し, 従属クラスタ動的生成機構を導入することで解決する手法 [2] が提案されている. 評価実験によりその有効性が確認されているが, 破壊的クラスタ割り当てを判別する閾値を手動で設定する必要があるため, ユーザの負担を増やす要因となっている.

本稿では, 上述の従属クラスタ動的生成機構を閾値の設定なしに適用可能とする拡張手法を提案する. 具体的には, 各クラスタに所属するデータの生起確率を, クラスタ中心からの距離に従う確率密度関数を定義して求め, ベイズ情報量規準 (Beyesian Information Criterion, BIC)[3] を用いてクラスタリング結果を評価する. クラスタから遠い位置にあるデータを追加することによるクラスタ内データの対数尤度低下と, クラスタを新規生成することによるモデルの複雑化のトレードオフに基づき, 従属クラスタを生成するか否かを判断する. これにより, 従来手法で必要であった閾値が不要となる.

BIC で用いる確率密度関数を複数用意し, 標本分散と不偏分散の違いなども考慮したいくつかの対数尤度計算法を実装して従来手法との比較実験を行い, 提案手法の有効性を評価する.

2 関連研究

2.1 K-means

クラスタリングの代表的手法である K-means は, ランダムに設定した初期クラスタ中心から式 (1) に示す評価関数を最小化するクラスタ中心及びクラスタ割り当てを反復計算により求めることで, データ集合 N を設定した k 個のクラスタ $C=\{c_1,\ldots,c_K\}$ に分割する.

K-means で得られる解はランダムに設定されるクラスタ中心の初期値に依存することが知られており [4], その改善手法についても研究されている [5].

$$f(C)=\sum_{x_j\in N}\min_{i\in C}\|x_j-c_i\|^2 \tag{1}$$

2.2 COP K-means

COP K-means[1] は, ユーザの背景知識を K-means のクラスタリングアルゴリズムにフィードバックすることで, 期待されたクラスタリング結果に近づける手法である. ユーザから与えられるデータ分割に関する情報を制約と呼び, 制約を優先的にクラスタ割り当てに反映させることでユーザが期待するクラスタリング結果に近づけることを目指している. 制約は対制約 [1] と呼ばれ, 以下の 2 種類が存在する.

- Must-Link : 指定したデータペアを同じクラスタに割り当てる.
- Cannot-Link : 指定したデータペアを異なるクラスタに割り当てる.

与えられた対制約は, 各データのクラスタ割り当てにおいて考慮される.

2.3 従属クラスタ導入手法

COP K-means では, データ空間上で遠く離れたデータ間に Must-Link 制約を付与した場合, クラスタリング結果に大きな影響を与える問題がある. この問題を解決するため, 不連続領域からなるクラスタに対し従属クラスタを生成することで対処する, 従属クラスタ動的生成機構が提案されている [2]. 従属クラスタ動的生成機構を導入した COP K-means の具体的な手順を以下に示す. ただし, この手法では Must-Link 制約のみを対象としている.

1. 初期値設定
   クラスタ中心の初期値 (クラスタ初期値) をランダムに決定する.
2. 初回クラスタ割り当て
   COP K-means と同様に割り当てる.
3. クラスタ割り当て
   Must-Link 制約が付与されたデータオブジェクトについて, 割り当て先となるクラスタ中心とのユークリッド距離が閾値を超えた場合, 破壊的クラスタ割り当てと判断する.
   破壊的クラスタ割り当てである場合 :
   そのデータオブジェクトを要素とした従属クラスタを生成する. クラスタ初期値はそのデータオブジェクトの座標となる. このデータオブジェクトは以降も別クラスタへの割り当てを禁止する. また, 従属クラスタと元のクラスタとの関係を保持しておく.

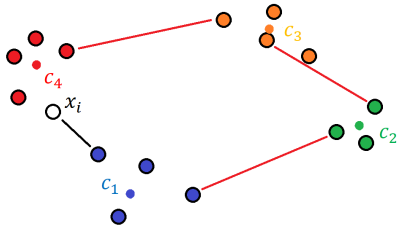


図 1: クラスタ統合後を考慮したクラスタ割り当ての例

   破壊的クラスタ割り当てでない場合 :
   COP K-means と同様に Must-Link 制約を適用する. このとき, 本来の割り当て先となるクラスタだけでなく, その従属クラスタを含めた中で最も近い位置にあるクラスタを求め, 割り当てる. 図 1 の例では, データオブジェクト x_i がクラスタ c_1 に属するデータオブジェクトと Must-Link 制約関係にある. このとき, c_1 は従属クラスタ c_2, c_3, c_4 を持つため, x_i はそれらの中の最近傍クラスタである c_4 に割り当てられる.
4. 重心計算
   クラスタ中心を K-means と同様に更新する. 従属クラスタにおいても同様に更新する.
5. 収束判定
   クラスタ割り当てに変化がないか, 収束しない状況が検出された場合は反復計算が 100 回を超えた時点で終了する.
6. Must-Link クラスタ統合
   Must-Link 制約に基づいて従属クラスタと親元のクラスタの統合を行う. この処理のみで指定したクラスタ数にならなかった場合 7 を実行する.
7. 凝集型クラスタ統合
   階層的クラスタリングを適用し, 指定したクラスタ数となるまでクラスタを統合する.

3 提案方法

3.1 ベイズ情報量規準に基づく従属クラスタ生成

従属クラスタ動的生成機構では, 破壊的クラスタ割り当てを検出する閾値をデータセットごとに手動で設定する必要がある. この問題を解決するため, 本稿では情報量規準に基づく閾値が不要な手法を提案する. ベイズ情報量規準 (BIC)[3] に基づく, 式 (2) に示すモデル M の評価値に関して, 従属クラスタを形成した場合としない場

合のモデルを比較し値が大きい方を採用する.

$$\begin{aligned} BIC(M) &= \log \prod_{x_i \in X} p(x_i; M) - \frac{q}{2} \log(|X|) \\ &= \sum_{x_i \in X} \log p(x_i; M) - \frac{q}{2} \log(|X|) \\ &= L(X) - \frac{q}{2} \log(|X|) \end{aligned} \tag{2}$$

式 (2) において X は対象となるクラスタに割り当てられたデータオブジェクト x_i の集合である. q は M におけるパラメータの個数であり, $p(x_i; M)$ は M における x_i の生起確率である. 本稿では 3.2 節に示す様に, 複数の確率密度関数についてその有効性や特性を調査する.

データオブジェクト x を, Must-Link 制約に基づき従属クラスタを生成せずに既存クラスタ C に追加した場合のモデル M_1 の BIC は, C に x を追加後のクラスタを C' $(= C \cup \{x\})$ とすると式 (3) で表される.

$$BIC(M_1) = L(C') - \frac{q}{2} \log(|C| + 1) \tag{3}$$

従属クラスタを生成した場合のモデル M_2 の BIC は式 (4) で表される. データオブジェクトの生成確率はクラスタにより異なるため, 既存クラスタ C と従属クラスタに分けて対数尤度を計算する. 既存クラスタの対数尤度 $L(C)$ は変化せず, x_i のみからなる従属クラスタの対数尤度は 0 とする. クラスタが 1 つ増えるため M_2 のパラメータ数は M_1 の 2 倍となる.

$$BIC(M_2) = L(C) - q \log(|C| + 1) \tag{4}$$

$BIC(M_2) - BIC(M_1)$ を求め, 正の値であれば従属クラスタの生成, 負の値であれば既存クラスタへの割り当てを実行する.

3.2 予備実験

既存のクラスタに対するデータオブジェクトの追加を, その座標を変化させながら行うことで, 選択されるモデルの変化を調べる.

データオブジェクトを平均値 0, 分散 0.5, 0.8, 1.0 の正規分布に従い, ランダムに 100 個配置したデータセットをそれぞれ既存クラスタとして用いる. 対数尤度の計算に用いる確率密度関数として以下の 6 手法を比較する.

手法 1：正規分布, C' の対数尤度の計算には分散にデータ追加後の標本分散を利用
手法 2：正規分布, C' の対数尤度の計算にはデータ追加後の不偏分散を利用
手法 3：正規分布, C' の対数尤度の計算にはデータ追加前の標本分散を利用
手法 4：ウィグナー半円分布
手法 5：正規分布, C' の対数尤度の計算にはデータ追加前の不偏分散を利用
手法 6：ロジスティック分布

手法 2, 手法 5 に関して, 不偏分散は標本分散よりも大きな値となるため, クラスタ中心から離れた位置にあるデータオブジェクトの生成確率が大きくなる.

手法 4 で用いるウィグナー半円分布は母数 R を用いて式 (5) のように定義される. その分布は正規分布, ロジスティック分布と比べて尖度が低い. 本稿では, C, C' においてそれぞれ属するデータオブジェクトとクラスタ中心との最大距離を R とし, 追加するデータオブジェクトとクラスタ中心との距離を x とする.

$$f(x) = \begin{cases} \frac{2}{\pi R^2} \sqrt{R^2 - x^2} & (|x| < R) \\ 0 & (\text{otherwise}) \end{cases} \tag{5}$$

手法 6 で用いるロジスティック分布は平均値を μ, 分散の調整を行う尺度パラメータを s とすると, 以下の式 (6) で表される. その分布は正規分布と同様釣り鐘型だが, 裾野が広いのが特徴である. 本稿では, データオブジェクトとクラスタ中心との距離を x として, それらの平均値を μ, s をクラスタ中心と所属するデータオブジェクトの最大距離とする.

$$f(x; \mu, s) = \frac{\exp(-(x - \mu)/s)}{s(1 + \exp(-(x - \mu)/s))^2} \tag{6}$$

図 2-7 は各手法の実験結果であり, それぞれ左から順に, 既存クラスタの分散が 0.5, 0.8, 1.0 である場合の実験結果である. 各図において, 既存クラスタのデータオブジェクトは黒, 新規データオブジェクトは赤, 青で表され, 青は既存クラスタ割り当て選択, 赤は従属クラスタ生成選択を示している.

図 2 は手法 1 での実験結果であり, 既存クラスタの分散が増加するにつれて既存クラスタに割り当てる範囲が狭まっている. 正規分布の性質として, 分散が大きくなるとクラスタ中心から離れた位置にあるデータオブジェクトの生成確率が大きくなる. 手法 1 での実験結果をもとに, 確率上昇が見られたデータオブジェクトの個数を分散, 追加するデータオブジェクトごとに表 1 にまとめると, 追加するデータオブジェクトがクラスタ中心から遠ざかるにつれて確率が増加するオブジェクトが減少する傾向が見られ, 分散が大きくなるにつれて減少していくことも確認できる. このことから, 分散が小さいほど生成確率の上昇が大きく, 既存クラスタへの割り当てが選択されやすくなっていると解釈できる. しかし, まとま

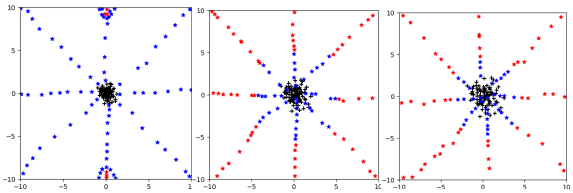


図 2: 手法 1 の実験結果

りの良い (分散の小さい) クラスタほど, 遠くにあるデータオブジェクトを追加しやすいというのは直感に反するといえる.

図 3 は手法 2 での実験結果である. 不偏分散を用いても, 図 2 と同様の結果となっていることがわかる.

図 4 は手法 3 での実験結果であり, 既存クラスタの分散が大きくなるほど既存クラスタに割り当てられる範囲が広がる傾向がある. 手法 3 では既存データオブジェクトの生成確率が変化しないため, 追加するデータオブジェクトの生成確率とモデルの複雑さ増加のバランスのみでモデルの選択が決まる. 既存クラスタの分散が大きいほど, 追加するデータオブジェクトの生成確率が大きくなるため, 既存クラスタへの割り当てが選択されやすくなる.

図 5 に示す手法 4 での実験結果では, ウィグナー半円分布を用いているため, 新規データオブジェクトの追加による分散の増加によって, 既存クラスタ割り当てのデータオブジェクトの生成確率が減少するため, 従属クラスタを生成しやすくなっている. また, 既存クラスタの分散が大きいほど, 新規データオブジェクトの追加による生成確率の減少が少ないため, 既存クラスタへの割り当てが選択されやすくなる.

図 6 に示す手法 5 での実験結果では, 図 4 の実験結果に近い分布を示している. 手法 1, 2 の場合と同様, 標本分散と不偏分散による違いは小さいことがわかる.

図 7 に示す手法 6 での実験結果は, 手法 4 と同様の結果となっている. ロジスティック分布も正規分布と同様の釣り鐘型関数であるため, 新規データオブジェクトの追加によって既存クラスタに所属するデータオブジェクトの生成確率が上昇することがあり得るが, 正規分布よりも裾野が広いため, 生起確率の上昇が起こりにくく, 手法 1, 2 とは反対の傾向となっている.

表 1: 手法 1 における生起確率が増加したデータオブジェクトの個数

追加データ座標		分散		
x	y	0.5	0.8	1.0
1	1	74	61	56
2	2	69	57	53
3	3	64	53	48
4	4	61	51	46

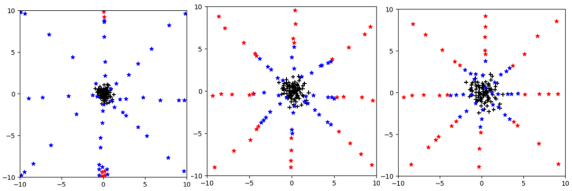


図 3: 手法 2 の実験結果

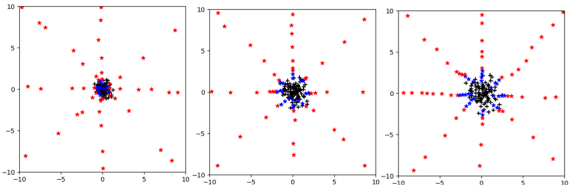


図 4: 手法 3 の実験結果

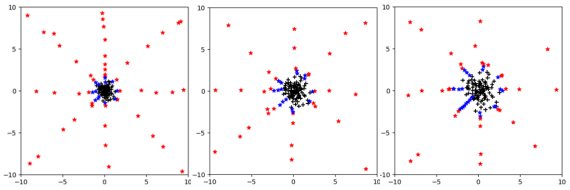


図 5: 手法 4 の実験結果

4 評価実験

4.1 実験の概要

3.2 節で提案した 6 手法を, 従属クラスタ動的生成機構を導入した COP K-means に組み込んで, クラスタリング精度の評価・比較を行う. 閾値を導入した従来の手法 [2] とも比較を行う. 扱うデータセットは先行研究 [2] で用いられた, 図 8 に示す 2 種類である. 対制約は先行研究と同じものを用いており, データセットごとにそれぞれ

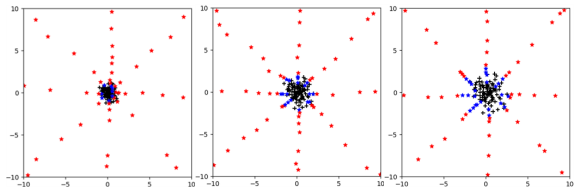


図 6: 手法 5 の実験結果

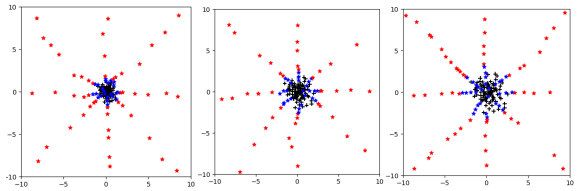


図 7: 手法 6 の実験結果

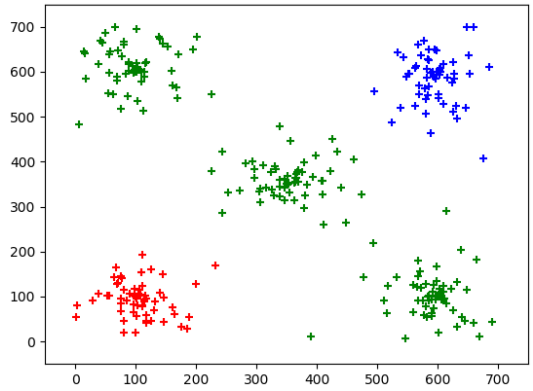
5, 10, 15, 20, 25, 30, 35, 40, 45, 50 対の Must-Link 制約で構成されている. 正規化相互情報量 (Normalized Mutual Inforamtion, NMI)[6] で評価を行い, クラスタリング結果と正解データの一致度を評価する. COP K-means の初期値依存性を考慮し, ランダムな初期値で 10,000 回実行し, NMI の平均, 分散を求める. また, クラスタ割り当てプロセス終了時点の従属クラスタ総数の平均をまとめ, 手法ごとにその差異を調査する.

4.2 実験結果

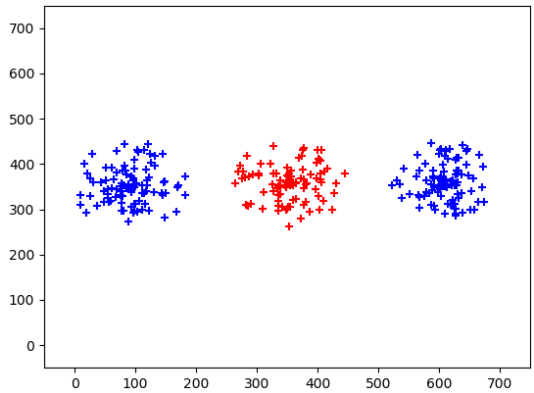
NMI の平均を表 2 に示す. 各行において最も良い結果を太字で示している. データセット B における 30 対の場合の提案手法の結果が, 従来手法と比較して悪くなっている. この結果を除けば, データセット A, B ともに提案手法, 従来手法ともにほとんどの場合で 0.9 以上と, 良好な結果が得られている.

NMI の分散に関しては, データセット A では従来手法の方が小さい傾向にあり, データセット B では提案手法で 0 となる場合が多く, 安定した結果が得られている. 平均値が低くなっていた, データセット B における 30 対の制約対セットでの実験結果では, 提案手法の分散が他の場合よりもかなり大きくなっていた. NMI が特に低い手法 3 と手法 5 は, 他の提案手法よりも低い分散となっているため, 初期値によらず結果が悪かったといえる.

表 3 は生成された従属クラスタの個数の平均であり, 従来手法より多数の従属クラスタを生成する傾向にあることがわかる. 提案手法間の差は小さく, 手法 3, 手法 5 で NMI の大きな低下がみられた場合も, 生成された従属クラスタ数の違いが影響したものではないと考えられる.



(a) データセット A



(b) データセット B

図 8: 実験に用いるデータセット

5 考察

手法 1,2 に関しては, 3.2 節で示した予備実験で他の提案手法とは異なるモデル選択傾向を持つことが確認されていたが, 両データセットにおいて他の手法と同程度の結果が得られている.

図 9 は手法 3 でのデータセット B, 対制約数 30 におけるクラスタリング結果の 1 例であり, 図 8 の正解データと大きく異なっている. 手法 3, 手法 5 に共通する点として, 新規データ追加前の分散を対数尤度計算に用いており, これが結果に影響したことが考えられる.

データセット B の制約対数 30 においては全ての提案手法の平均が 0.8 未満であり, 分散も大きいことから, 従来手法に劣る結果といえるが, それ以外では従来手法と同等かそれ以上の結果が閾値を設定することなく得られていることから, 提案手法の有効性が示されたといえる.

表 2: 各手法における制約数ごとの NMI 平均

(a) データセット A

対制約数	手法 1	手法 2	手法 3	手法 4	手法 5	手法 6	従来手法
5	0.933	0.933	0.933	0.932	0.933	0.933	**0.985**
10	0.867	0.865	0.872	0.864	0.869	0.864	**0.985**
15	**0.923**	0.922	0.922	0.922	0.922	**0.923**	0.846
20	0.983	0.982	0.956	0.984	0.955	0.983	**0.995**
25	0.899	0.899	0.901	0.898	0.903	0.898	**0.954**
30	0.973	0.973	0.972	0.974	0.972	0.975	**0.985**
35	**0.974**	0.973	0.970	**0.974**	0.971	0.973	0.953
40	0.963	0.963	0.957	0.964	0.958	0.963	**0.999**
45	0.901	0.901	0.900	0.902	0.899	0.899	**0.919**
50	0.909	0.909	0.915	0.912	0.912	0.911	**0.937**

(b) データセット B

対制約数	手法 1	手法 2	手法 3	手法 4	手法 5	手法 6	従来手法
5	0.978	0.977	0.977	0.978	0.978	0.977	**1.000**
10	**0.999**	**0.999**	0.982	**0.999**	0.979	**0.999**	0.988
15	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.985	0.994
20	**1.000**	**1.000**	0.997	**1.000**	0.996	**1.000**	0.981
25	0.999	0.999	0.999	0.999	0.999	0.999	**1.000**
30	0.797	0.795	0.244	0.791	0.247	0.794	**0.978**
35	**1.000**	**1.000**	0.984	**1.000**	0.984	**1.000**	0.922
40	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.987
45	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.985
50	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.986

表 3: 各手法における従属クラスタ平均数

(a) データセット A

対制約数	手法 1	手法 2	手法 3	手法 4	手法 5	手法 6	従来手法
5	10.00	10.00	10.00	10.00	10.00	10.00	5.83
10	20.00	20.00	20.00	20.00	20.00	20.00	9.62
15	27.00	27.00	27.00	27.00	27.00	27.00	12.35
20	34.00	34.00	33.78	34.00	33.78	34.00	15.20
25	47.00	47.00	47.00	47.00	46.96	47.00	13.59
30	55.00	55.00	55.00	55.00	55.00	55.00	18.51
35	67.00	67.00	67.00	67.00	67.00	67.00	22.56
40	69.00	69.00	69.00	69.00	69.00	69.00	16.75
45	70.00	70.00	69.97	70.00	69.97	70.00	25.21
50	84.00	84.00	84.00	84.00	84.00	84.00	29.12

(b) データセット B

対制約数	手法 1	手法 2	手法 3	手法 4	手法 5	手法 6	従来手法
5	10.00	10.00	10.00	10.00	10.00	10.00	4.00
10	20.00	20.00	20.00	20.00	20.00	20.00	12.00
15	29.00	29.00	29.00	29.00	29.00	29.00	13.00
20	40.00	40.00	40.00	40.00	40.00	40.00	10.00
25	45.00	45.00	45.00	45.00	45.00	45.00	14.00
30	54.00	54.00	54.00	54.00	54.00	54.00	28.97
35	63.00	63.00	63.00	63.00	63.00	63.00	28.00
40	74.00	74.00	74.00	74.00	74.00	74.00	41.00
45	73.00	73.00	72.59	73.00	72.68	73.00	28.00
50	85.00	85.00	85.00	85.00	84.99	85.00	32.00

6 まとめ

本稿では, 従属クラスタ動的生成機構を導入した拡張型 COP K-means 手法に関して, BIC に基づき従属クラスタ生成を判断する手法を提案した. 対数尤度の計算に用いる確率密度関数等が異なる 6 手法を提案し, 手動による閾値設定に基づく従来手法と比較した結果, 多くの場合で従来手法と同等以上の NMI を達成した. 提案手法は対象データセットに依存したパラメータ調整が不要であるため, 多様な用途での活用が期待できる. 今後は, 実データ含む様々なデータセットに適用し, 提案手法の有効性や特性について明らかにしていく予定である.

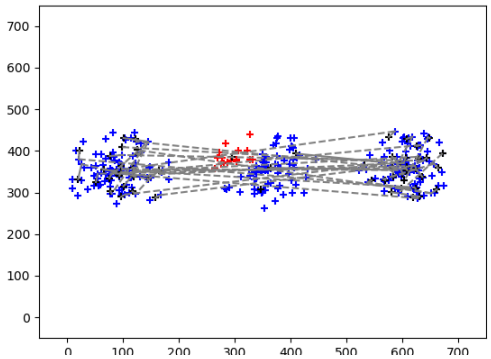


図 9: 手法 3, データセット B, 対制約数 30 におけるクラスタリング結果の例

参考文献

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

[2] 井本 博之, 高間康史, 従属クラスタ動的生成機構の導入による Must-Link 付き K-means 法の拡張に関する提案, 第 11 回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, SIGAM1101, pp. 1-6, (2015)

[3] Schwarz, G. :Estimating the Dimension of a Model, *Annals of Statistics*, Vol. 6, No. 2, pp. 461-464, (1978)

[4] 小野田 崇, 坂井 美帆, 山田 誠二, k-means 法の様々な初期値設定によるクラスタリング結果の実験的比較, 第 25 回人工知能学会全国大会, 1J1-OS9-1, pp. 1-4, (2011)

[5] Arthur, D. and Vassilvitskii, S. :k-means++: The Advantage of Careful Seeding, *Proc. of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithm*, pp. 1027-1035, (2007)

[6] Strehl, A. and Ghosh, J. :Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions, *Journal of Machine Learning Research*, Vol. 3, pp. 583-617, (2002)