2G2-OS9-4

少数制約の伝播による類似度学習とクラスタリングへの応用

岡部正幸[*1]　山田誠二[*2]

[*1]豊橋技術科学大学　[*2]国立情報学研究所／総合研究大学院大学

This paper proposes a clustering algorithm based on maximum cut problen and its semidefinite programming relaxation. The algorithm is different from some other clustering algorithms using semidefinite programming in the use of learned similarity matrix and grouping procedure. We verify the effectiveness of the algorithm on some datasets.

1. はじめに

情報検索などインタラクティブな環境においてユーザからアドホックに得られるフィードバック情報はあまり多くは期待できない．このような環境で機械学習を適用するには，半教師あり学習などのアプローチにより訓練データを有効に活用することが必要となる．

本研究では，類似度学習を伴なう制約付きクラスタリングに焦点を当て，制約を活用するための具体的なクラスタリングアルゴリズムを提案する．提案方法は，グラフの最大カット問題に基づく分割型のクラスタリング方法をベースとしており，制約を組み込むため半正定値計画問題（Semidefinite Programming, SDP）を利用している．

半正定値計画問題を利用した類似度行列（またはカーネル行列学習，距離行列）学習を伴なう制約付きクラスタリングはこれまでにいくつか提案されているが [Hoi 07, Li 08]，これらの研究とは学習後の類似度行列を離散化して利用する点や k-means などの既存のアルゴリズムを利用しない点が異なる．

2. 近傍グラフの最大カット問題と SDP 緩和

ここでは，提案手法のベースとなる最大カット問題の SDP 緩和による解法ついて述べる．

最大カット問題とは，無向グラフ $G = (V, E)$（V は頂点集合，E は枝集合，また $|V| = n$ とする）において，V を 2 分割して (V_1, V_2) の部分集合に分けるとき，2 組の頂点集合間に張られたすべての枝の重みの和 $\sum_{i \in V_1, j \in V_2} w_{ij}$（$w_{ij}$ は頂点 i, j 間の枝の重み）を最大化するような分割を求める問題である．これは，各頂点に対応する変数 u_i を導入し，$i \in V_1$ のとき $u_i = 1$，$i \in V_2$ のとき $u_i = -1$ の値をとるとすると以下のように定式化できる．

最大カット問題

$$\begin{aligned} \text{maximize} \quad & \frac{1}{4} \sum_i \sum_j w_{ij}(1 - u_i u_j) \\ \text{subject to} \quad & u_i^2 = 1 \end{aligned}$$

この問題は，変数 $x_{ij} = u_i u_j$ を導入することにより，SDP による緩和問題として解くことができる．W, X をそれぞれ w_{ij}, x_{ij} を要素としてもつ行列，ラプラシアン行列を $L = \frac{1}{4}(diag(We) - W)$（$e$ はすべての要素が 1 のベクトル）とすると，この問題は以下のように記述できる．

最大カット問題の SDP 緩和

$$\begin{aligned} \text{maximize} \quad & L \bullet X \\ \text{subject to} \quad & E_{ii} \bullet X = 1 (1 \leq i \leq n) \\ & X \succeq O \end{aligned}$$

ただし，$L \bullet X = \sum_i \sum_j l_{ij} x_{ij}$，$E_{ii}$ は (i, i) 成分だけが 1，あとの成分は 0 の $n \times n$ 行列とする．

クラスタリング対象となるデータ集合 D から生成される近傍グラフに上記の問題を適用することにより行列 X が求まり，これを利用して 2 分割クラスタリングを行うことができる．実際の分割方法としては，X の各要素がとる-1 から 1 までの実数値を各データ間の類似度とみなし，k-means（2 クラスなので 2-means）などの既存クラスタリングアルゴリズムを適用することも可能であるが，本研究では，既存手法によらない方法を次章で提案する．

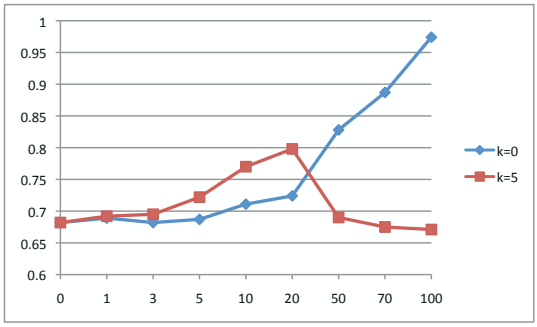
3 クラス以上のクラスタリングを行いたい場合は，2 分割操作を繰り返し行う．また，この問題を制約付きクラスタリングとして考えた場合，与えられた制約はそのまま SDP の制約（データ (i, j) が must-link の場合，$E_{ij} \bullet X = 1$）として組み込むことが可能である．ただし，3 クラス以上の分割を行う場合，cannot-link を適用することはできず，must-link のみ適用可能である．(例えば 3 クラスの分割の場合，2 回目の分割で有効な cannot-link は 1 回目の分割では must-link と見なさなければならないため)

3. クラスタリングアルゴリズム

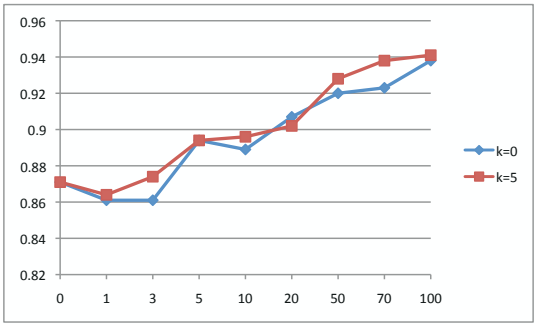
前章で説明した最大カットの SDP 緩和問題を解くことによって-1 から 1 までの実数値を要素として持つ類似度行列 X が求められるが，本来この行列の要素は 1 または-1 の 2 つの値しかもたないことを想定していた．本研究では，SDP を解くことによって実際に得られた行列 X の各要素を 1 と-1 に離散化し，行を交換する（対称行列なので列も同時に交換）ことによって 1 と-1 を集約させ，分割を行う方法を提案する．具体的な手順は以下の通り．

1. 行列 X の各要素について，$x_{ij} \geq 0$ の場合 1 に，$x_{ij} \leq 0$ の場合-1 に離散化する．
2. 各行における $1 \sim n$ までの列成分を-1, 1 の順序列とみなし，その順序パターンが同じ列は一まとまりにする．(同一パターンについては 1 つの行で代表させる)
3. 出現回数の最も大きいパターンを決め，そのパターンと類似しているもの（ハミルトン距離で計算）順に並べる．

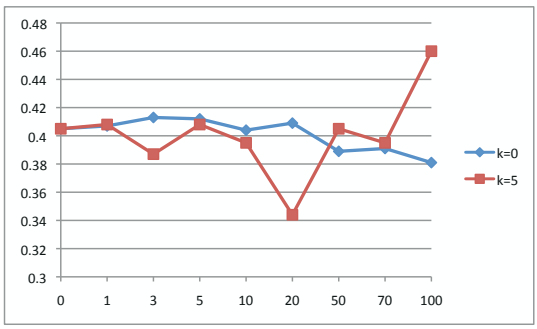
連絡先: 岡部正幸，豊橋技術科学大学情報メディア基盤センター
〒 441-8580 豊橋市天伯町雲雀ヶ丘 1-1
okabe@imc.tut.ac.jp



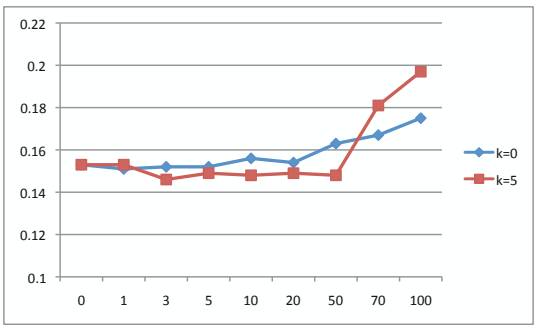
(a) Soybean-small



(b) Iris



(c) Wine



(d) Ionosphere

図 1　UCI Repository における評価 (縦軸:NMI，横軸:制約数)

例えば，$p_0, p_1, ..., p_n$ と並んでいる場合，p_0 が出現回数の最も大きいパターンで，p_1 が p_0 に最も類似しているパターン，p_2 が次に類似しているパターンである．

4. パターン列の分割境界を決めるため，各分割境界候補（p_0 と p_1 の間，p_1 と p_2 の間...）で分割を行った場合について，以下の評価値 F を計算する．

$$F = \sum_{i=1}^{4} -x_{-1}^{i} \log(x_{-1}^{i}) - x_{1}^{i} \log(x_{1}^{i})$$

x_{-1}^{i}, x_{1}^{i} はそれぞれ，行と列をそれぞれ 2 分割することにより生じる 4 つのエリア（1. 左上，2. 右上，3. 左下，4. 右下の順）の-1, 1 の割合である．つまり，-1, 1 の出現確率のエントロピーを計算することにより，-1, 1 の集約度を計算している．

5. F が最も低い値となる候補を分割境界と決定する．

提案手法による，3 クラス以上の場合も含めた一般的なクラスタリング手順についてまとめると以下のようになる．

1. 分割済みクラスタ集合のうちデータ数が最も多い集合 D_i を選択する．
2. must-link 集合の中から，D_i の分割に適用可能な must-link 制約を選択する．
3. D_i について最大カット問題を解き，上記の方法により実際の分割を行う．
4. ターゲットのクラスタ数が k の場合，これを $k-1$ 回繰り返す．

4. 実験

提案手法について，UCI Repository から 4 つのデータセット (Soybean-small, Iris, Wine, Ionosphere) を使用して評価した．図 1 は制約数と評価値 NMI(Normalized Mutual Information) の関係を示したものである．制約は各個数につき異なる 20 パターンを用意し，評価値はその平均値である．$k = 0$ のグラフは与えられた制約のみを利用したもの，$k = 5$ は与えられた制約と [岡部 09] で提案した方法によって選択した擬似制約も利用したもののグラフである．Wine 以外では制約の効果が得られていること．また，Soybean-small, Iris では擬似制約の効果が得られていることが見て取れる．

5. まとめ

本研究では，最大カット問題の SDP 緩和に基づく制約付きクラスタリングアルゴリズムを提案した．実験結果では，制約追加の効果があることを確認したが，より詳細な解析が必要である．また，他のデータでの検証，他の手法との比較なども今後の課題である．

参考文献

[Hoi 07] Hoi, S. C. H., Jin, R. and Lyu, M. R.: “Learning Nonparametric Kernel Matrices from Pairwise Constraints”, In Proc. ICML’07 (2007).

[Li 08] Li, Z., Liu, J. and Tang, X.: “Pairwise Constraint Propagation by Semidefinite Programming for Semi-Supervised Classification”, In Proc. ICML’08 (2008).

[岡部 09] 岡部正幸，山田誠二.:“ 制約付き距離学習による文書クラスタリング ”, 第 23 回人工知能学会全国大会，2B3-03