制約付き距離学習による文書クラスタリング

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

This paper proposes a method of metric learning for constrained clustering. We modify a technique of metric learning that utilizes graph laplacian and semi-definite programming by adding optional constraints. The preliminary experiments shows that our method is promising.

1. はじめに

近年，機械学習の分野では半教師あり学習に関する研究が精力的に行われている．制約付きクラスタリングも半教師あり学習の一種であり，同一クラスタに属するデータ対または別クラスタに属するデータ対の一部を教師情報として与えてやることにより，学習性能を向上させようとするアプローチである．この教師情報について，前者の同一クラスタに属するデータ対は must-link，後者の別クラスタに属するデータ対は cannot-link とそれぞれ呼ばれる．

この制約の具体的な利用法は主に二つある．一つは各データの配属クラスタを決定する際に用いる場合で，クラスタリングアルゴリズムによって must-link データ対が別クラスタに，または cannot-link データ対が同一クラスタに配属されないための情報として用いる．一方，制約を利用して must-link データ対の距離が小さく，cannot-link データ対の距離が大きくなるように擬似的な距離を計算するアプローチも存在する．現在，様々な方法が提案されているが [Shwartz 04, Davis 07, Tang 07]，本研究ではグラフラプラシアンと半正定値計画問題を利用した方法 [Hoi 07, Li 08] をベースに，制約に関係するデータの近傍データについての制約を追加して距離学習を行う方法を提案する．

この方法では，疑似距離行列を最適化問題の変数として直接求める．具体的には，グラフラプラシアンと疑似距離行列の内積を目的関数として，must-link と cannot-link を制約条件として加えた半正定値計画問題を解く．この問題を解くことで制約に関係するデータ間の距離を変化させることができるが，これにより制約に関係しないデータとの距離がどのように変化するかについては，特に指示しているわけではない．本研究では，制約に関係するデータの近傍データについて距離学習後も近傍に存在させるための制約を明示的に組み込む方法を提案する．

2. 距離学習

データ集合 $P = \{\vec{p}_i \mid \vec{p}_i \in R^m, i = 1, ..., n\}$ に対し，P 内の任意の 2 データ間距離 s_{ij} を要素とした行列 $S \in R^{n \times n}$ を考える．ただし，$0 \leq s_{ij} \leq 1$ である．

また，制約として must-link 集合 M と cannot-link 集合 C が下記のように与えられているとする．

$$M = \{(i,j) | \vec{p}_i と \vec{p}_j は同じクラスタに属する\}$$
$$C = \{(i,j) | \vec{p}_i と \vec{p}_j は異なるクラスタに属する\}$$

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

距離学習の目的は，これらの制約を満たす新たな距離行列 K を求めることである．K の求め方について，Li らが提案したグラフラプラシアンを利用した半正定値計画問題による解法 [Li 08] について以下に説明する．

行列 S に対し $d_{ii} = \sum_{j=1}^{n} s_{ij}$ を対角成分とする対角行列を D とすると，グラフラプラシアン L は $L = D - S$ のように表現される．ベースとなる半正定値計画問題は，最適化関数として正規化したグラフラプラシアン $\bar{L} = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$ を用いて以下のように定式化できる．

$$\begin{aligned} \min_K : \quad & \bar{L} \bullet K \\ s.t. : \quad & k_{ii} = 1, && i = 1, ..., n \\ & k_{ij} = d_{must}, && {}^{\forall}(i,j) \in M, \\ & k_{ij} = d_{cannot}, && {}^{\forall}(i,j) \in C, \\ & K \succeq 0 \end{aligned}$$

$\bar{L} \bullet K$ は行列 $\bar{L}$ と K の内積を表し，$\bar{L} \bullet K = \sum_{i=1}^{n} \sum_{j=1}^{n} \bar{l}_{ij} k_{ij}$ と計算される．また，$K \succeq 0$ は K が半正定値行列であることを示している．この条件は，求めた K の各距離が三角不等式を満たすことを保証するために必要である．なお，d_{must} と d_{cannot} は元の距離行列 S に応じて決まる値である（例えば，ユークリッド距離なら $d_{must} = 0$，コサイン距離なら $d_{must} = 1$ となる）．

3. 制約を利用した距離行列の生成

前章で説明した最適化問題を解くことにより，制約に関係するデータ間の距離を変化させることができるが，その影響で制約に関係しないデータとの距離がどのように変化するかについて上記の定式化では明らかではない．そこで，本研究では図 1 に示すように制約に関係するデータの近傍データを追随して変化させることを制約として前述の最適化問題に明示的に組み込む．図 1 の左半分・右半分の図はそれぞれ，ベースの解法での距離学習，提案方法での距離学習のイメージである．どちらも赤色のデータと青色のデータ間にそれぞれ制約が与えられており，距離学習によってそれぞれの距離が変化している様子を示している．図 1(a) のベース手法ではなるべく元の距離を保存しようとするため，主に赤色のデータを中心とした制約リンクの距離のみが変化している．一方，図 1(b) の提案手法では近傍データである緑色のデータ間の距離も変化させている．具体的には，must-link(または cannot-link)(i,j) があった場合，$\vec{p}_i$ の k-近傍データ $\vec{p}_{r_t^i}$ $(t = 1, ..., k)$ と制約対のもう一方のデータ $\vec{p}_j$ との学習後の距離を $k_{jr_t^i}$ $(t = 1, .., k)$ とすると，以下のような制約を追加する．

$$k_{jr_t^i} \leq -\bar{l}_{jr_t^i}, \quad if\ (i,j) \in M$$
$$k_{jr_t^i} \geq -\bar{l}_{jr_t^i}, \quad if\ (i,j) \in C$$

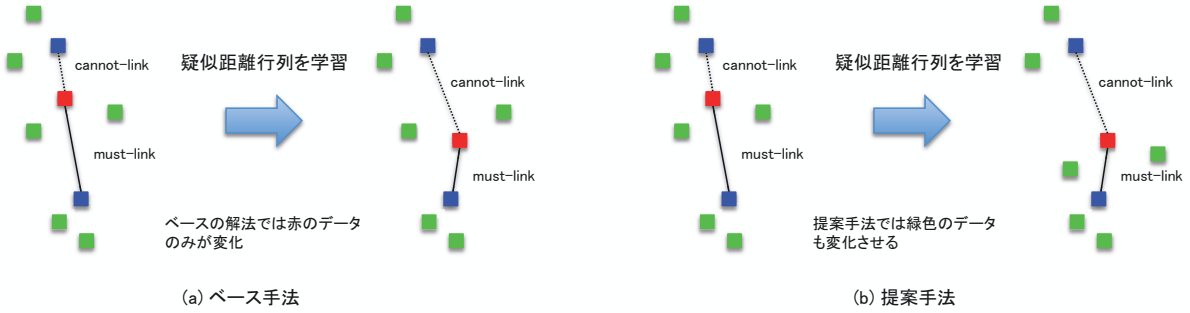

図 1: 距離学習による近傍データの変化の比較

同様に，$\vec{p}_j$ についても以下の制約を加える．

$$k_{ir_t^j} \leq -\bar{l}_{ir_t^j}, \quad if\ (i,j) \in M$$
$$k_{ir_t^j} \geq -\bar{l}_{ir_t^j}, \quad if\ (i,j) \in C$$

ここで，$-\bar{l}_{ij}$ は正規化後の s_{ij} の値である．つまり，must-link の場合は近傍データも同様に近づく（少なくとも学習前より遠ざからない），cannot-link の場合は近傍データも同様に遠ざかる（少なくとも学習前より近づかない）ことを制約として明示している．以上の制約を追加した最適化問題は以下のようになる．

$$\begin{aligned}
\min_K : \quad & \bar{L} \bullet K \\
s.t. : \quad & k_{ii} = 1, && i = 1, ..., n \\
& k_{ij} = d_{must}, && {}^\forall(i,j) \in M, \\
& k_{ij} = d_{cannot}, && {}^\forall(i,j) \in C, \\
& k_{ir_t^j} \leq -\bar{l}_{ir_t^j},\ k_{jr_t^i} \leq -\bar{l}_{jr_t^i}, && {}^\forall(i,j) \in M \\
& k_{ir_t^j} \geq -\bar{l}_{ir_t^j},\ k_{jr_t^i} \geq -\bar{l}_{jr_t^i}, && {}^\forall(i,j) \in C \\
& K \succeq 0
\end{aligned}$$

4. 実験

本章では，これまで説明した距離学習の効果を調べるために行った予備実験について述べる．タスクは文書データ (Reuters21578) のクラスタリングである．3 つのクラス (acq,earn,money-fx) からそれぞれ先頭の 30 データを抽出しクラスタリングを行う．クラスタリングアルゴリズムには k-means(medoid 版) を用いた．データ間距離にはコサイン距離を用いた．また最適化問題は SDPA パッケージ*1を用いて解いた．処理手順を以下に示す．

1. データから距離行列 S を計算し，SDPA を用いて疑似距離行列 K を求める．
2. 疑似距離行列 K を用いてクラスタリングを行う．

提案手法の効果を調べるため，以下の 2 つの距離行列を用いて実験を行った．

- must-link と cannot-link のみを制約としたもの．(KK-MEANS)
- must-link と cannot-link に加え，近傍データに関する制約も加えたもの (NKK-MEANS)．本実験で考慮する近傍データ数は 1 とした．

*1 http://sdpa.indsys.chuo-u.ac.jp/sdpa/

表 1: 実験結果

制約数	KK-MEANS	NKK-MENAS
1	0.268	0.285
3	0.304	0.331
5	0.323	0.361
10	0.340	0.374

クラスタリングの性能評価には NMI(Normalized Mutual Information) を用いた．表 1 は実験結果である．各制約数につき，制約ペアをランダムに 100 回発生させた．また発生させた制約集合ごとに k-means を 100 回走らせた．よって各値は合計 10000 万回の平均値である．詳細な検討が必要であるが一定の効果があるのが見て取れる．

5. まとめ

本研究では，クラスタリング性能を向上させるため，制約を利用して距離学習を行う方法として，グラフラプラシアンを用いた方法に制約データと近傍データに関する制約を追加する方法を提案した．今後はデータを増やし，より詳細な解析を行うとともに近傍データに関する制約の与え方について改良を行っていく予定である．

参考文献

[Shwartz 04] Shwartz, S., Singer, Y. and Ng, A. Y. : “Online and Batch Learning of Pseudo-Metrics”, In Proc. ICML’04 (2004).

[Davis 07] Davis, J., Kullis, B. and et.al. : “Information-Theoretic Metric Learning”, In Proc. ICML’07 (2007).

[Tang 07] Tang, W., Xiong, H., Zhong, S. and Wu, J. : “Enhancing Semi-Supervised Clustering: A Feature Projection Perspective”, In Proc. KDD’07 (2007).

[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).