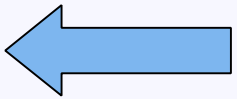
半教師ありクラスタリングに対する半正定値計画問題による解法

新納浩幸

話の出所

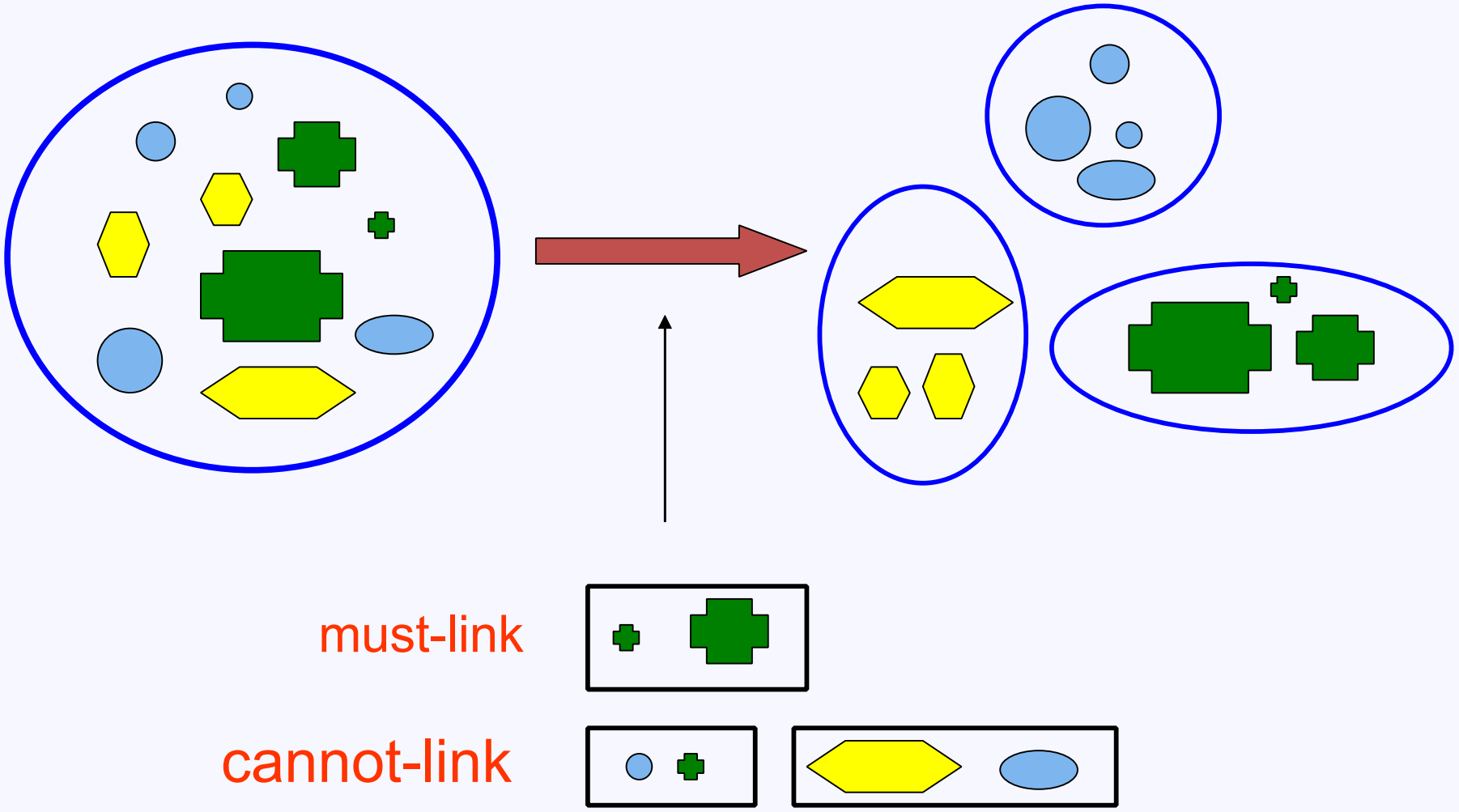
Li,Z., Liu, J. and Tang, X.: "Pairwise Constraint Propagation by Semidefinite Programming for Semi-Supervised Classification", ICML'08 (2008).

http://videolectures.net/icml08_li_pcp/



発表時の スライドと ビデオを
みることができる

半教師ありクラスタリング



代表的アプローチ

この論文はこっち

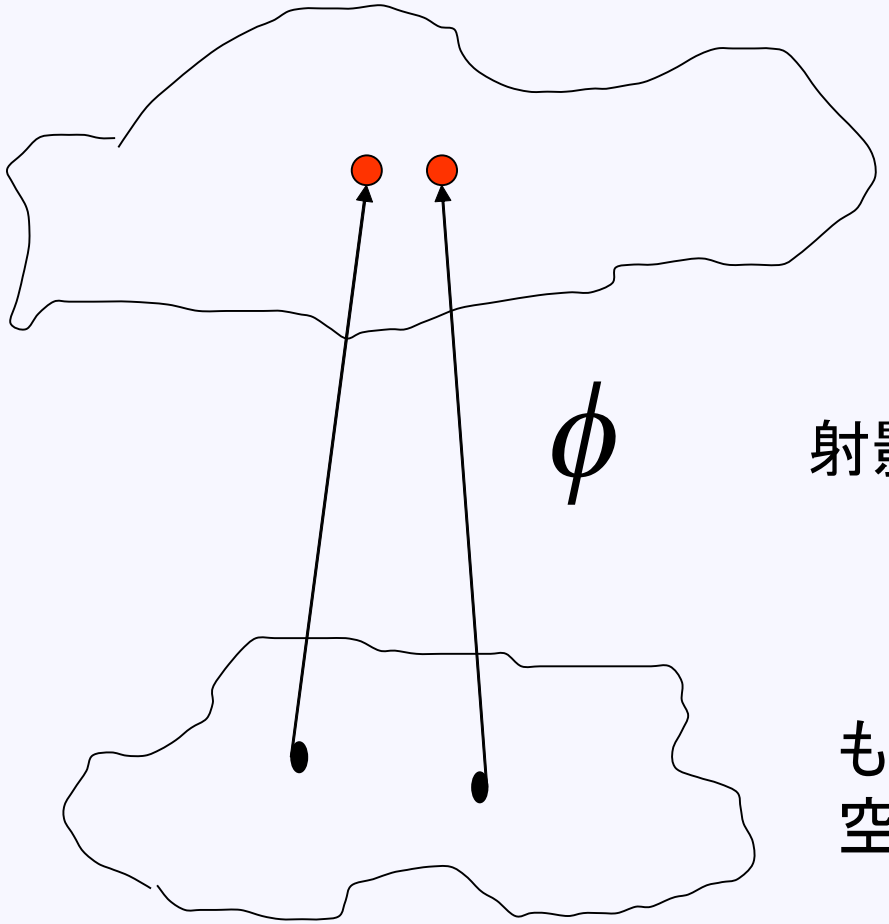
制約ベースの手法

クラスタリングの評価関数の方に
工夫。制約を満たさないと、悪い
値になるように設定。

距離ベースの手法

データ間の距離を制約によって、
適切に設定し直す。
直した距離で通常のクラスタリング

高次元への射影



高次元の
空間

ϕ

射影してからクラスタリング

もとの次元の
空間

高次元での制約

$$\| \phi(x_i) - \phi(x_j) \| < \varepsilon \quad (x_i, x_j) \in M$$

$$\| \phi(x_i) - \phi(x_j) \| > \delta \quad (x_i, x_j) \in C$$

上記を満たすような ϕ を探せばよい、、、たくさんある

$S(\phi)$ ：スムーズネスの尺度

近い点は近くに移す

これを最小にするようなϕを探す

高次元の単位球面上に射影

単位球面上なので、長さは1

$$< \phi(x_i), \phi(x_i) > = 1$$

制約は以下の形になる

$$< \phi(x_i), \phi(x_j) > = 1 \quad (x_i, x_j) \in M$$

$$< \phi(x_i), \phi(x_j) > = 0 \quad (x_i, x_j) \in C$$

同じクラスターのものは一致し、
異なるクラスターのものは直交する

行列による表記

W : データ間の類似度行列

データ数が n だと、n 行 n 列
対称行列、半正定値行列

D: 対角行列、対角要素はd_{ii}

$d_{ii} = \sum_{j} w_{ij}$　各データ i に対する類似度の和

ラプラシアン

グラフのラプラシアン

$$L = D - W$$

$$\overline{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$$

参考）

D^{-1} : Dの逆行列

$D^{-1/2}$: 2乗するとD^{-1} となる行列

一般に求めるのは困難だけど、この場合
D は対角行列なので、すぐ求まる

スムーズネスの関数

$$S(\phi) = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij} \| \frac{\phi(x_i)}{\sqrt{d_{ii}}} - \frac{\phi(x_j)}{\sqrt{d_{jj}}} \|^2$$

.....

K: 内積の行列

$$k_{ij} = < \phi(x_i), \phi(x_j) >$$

目的関数の変形

$$S(\phi) = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left(\frac{1}{d_{ii}}k_{ii} + \frac{1}{d_{jj}}k_{jj} - 2\frac{1}{\sqrt{d_{ii}d_{jj}}}k_{ij}\right)$$

$$= \sum_{i=1}^{n} k_{ii} - \sum_{i,j=1}^{n} \frac{w_{ij}}{\sqrt{d_{ii}d_{jj}}} k_{ij}$$

$$= I \bullet K - (D^{-1/2} W D^{-1/2}) \bullet K$$

$$= (I - D^{-1/2} W D^{-1/2}) \bullet K = \overline{L} \bullet K$$

$$A \bullet B = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} b_{ij}$$

半正定値計画問題

$$\min_{K} \ \overline{L} \bullet K$$

制約条件

$$\begin{cases} E_{ii} \bullet K = 1 \\ E_{ij} \bullet K = 1 \quad \forall (x_i, x_j) \in M \\ E_{ij} \bullet K = 0 \quad \forall (x_i, x_j) \in C \\ K \succeq 0 \end{cases}$$

小まとめ

タスクは半教師ありクラスタリング

制約を考慮して、類似度行列をチューン

チューンするために、
問題を半正定値計画問題に変形

最終的にはチューンされた類似度行列を使ってクラスタリング（論文では kernel k-means）

（実験結果と考察は省略します）

参考）ソフトウエア

半正定値計画問題を解くフリーソフト

SDPA

http://www.me.titech.ac.jp/~nakata/software.html

インストールは少し面倒、誰かやってください

論文では以下を使っていた

https://projects.coin-or.org/Csdp/