リンク不可例題からの距離学習とオブジェクト識別

小山 聡　田中克己　（京都大学）

概要

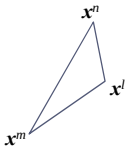
- リンク不可例題からの距離学習
  - 距離学習とは
  - リンク不可例題からの学習の定式化
  - 半正定値性，SVM学習との関係
- オブジェクト識別
  - オブジェクト識別とは
  - リンク不可例題の生成
  - 実験

リンク不可例題からの距離学習

距離(Distance Metric)

- 類似度や距離はデータの検索，ランキング，分類，クラスタリングにおいて重要な構成要素
  - 問題に適した類似度／距離を用いなければ，望ましい結果を得ることは困難
- （擬）距離を用いる利点
  - 三角不等式などの望ましい性質を満たす

$$d(\boldsymbol{x}^m, \boldsymbol{x}^n) \geq 0$$
$$d(\boldsymbol{x}^m, \boldsymbol{x}^n) = d(\boldsymbol{x}^n, \boldsymbol{x}^m)$$
$$d(\boldsymbol{x}^m, \boldsymbol{x}^l) + d(\boldsymbol{x}^l, \boldsymbol{x}^n) \geq d(\boldsymbol{x}^m, \boldsymbol{x}^n)$$



    - 上付の添字l,m,nはデータを表す
  - 既存のクラスタリングアルゴリズムの多くを，そのまま用いることができる．

距離の例

- Euclid距離

$$d_E(\boldsymbol{x}^m, \boldsymbol{x}^n) = \left(\sum_{i=1}^{D} (x_i^m - x_i^n)^2 \right)^{\frac{1}{2}}$$

  - 下付の添字はデータの特徴（単語の出現等）を表す
- 重み付きの距離（$w_i \geq 0$）

$$d_w(\boldsymbol{x}^m, \boldsymbol{x}^n) = \left(\sum_{i=1}^{D} w_i (x_i^m - x_i^n)^2 \right)^{\frac{1}{2}}$$

  - IDF(Inverse Document Frequency)

$$w_i = (\mathrm{IDF})^2 = \left(\frac{N}{\log \mathrm{DF}_i + 1} \right)^2$$

行列を用いた距離の定義

- $D \times D$ 行列$\boldsymbol{A}$を用いて，特徴間の相互作用を表現

$$d_{\boldsymbol{A}}(\boldsymbol{x}^m, \boldsymbol{x}^n) = \left((\boldsymbol{x}^m - \boldsymbol{x}^n)^T \boldsymbol{A} (\boldsymbol{x}^m - \boldsymbol{x}^n) \right)^{\frac{1}{2}}$$
$$= \left(\sum_{i=1}^{D} \sum_{j=1}^{D} a_{i,j} (x_i^m - x_i^n)(x_j^m - x_j^n) \right)^{\frac{1}{2}}$$

- $d_{\boldsymbol{A}}$が擬距離であるための必要十分条件
  - $\boldsymbol{A}$が半正定値（固有値が全て正）な対称行列

距離の教師付き学習

- 人間が一部のデータに対して，制約を与える
  - 類似な（同じクラスタに属するべき）データの対
    - リンク必須例題対（Must-be-linked example pairs）
  - 非類似な（異なるクラスタに属するべき）データの対
    - リンク不可例題対（Cannot-be-linked example pairs）
- 制約が満たされるように，行列Aを学習
- 制約が与えられたデータ以外に対しても，正しいクラスタリングが得られることを期待

Xing *et. al.*の方法

- 以下の最適化問題を解くことで，行列 A を決定

$$\begin{aligned}\min_{A}\ & \sum_{(x^m,x^n)\in\mathcal{S}} d_A^2(x^m,x^n)\\ s.t.\ & \sum_{(x^m,x^n)\in\mathcal{D}} d_A(x^m,x^n)\geq 1\\ & A\succeq 0\end{aligned}$$

$\mathcal{S}$：リンク必須例題対の集合
$\mathcal{D}$：リンク不可例題対の集合

- 問題点
  - 行列 A が半正定であるという条件を満たすため，行列の特異値分解を繰り返し行う必要がある．
  - 訓練集合の作成に人手が必要

リンク不可例題のみを用いた距離学習

- 問題領域によっては，リンク必須な例題とリンク不可な例題を得るのに必要なコストが異なる
- オブジェクト識別問題（後述）
  - 同じオブジェクトであり得ない（同じクラスタに入らない）例は比較的簡単に見つけることができる．
- 非類似なデータしか登録されないデータベース
  - 学術論文，特許，商標
  - 類似度の高いデータは，登録されずに却下される

学習の定式化

- リンク不可な例題同士の距離$d_A(x^m, x^n)$ができるだけ遠くになるような行列Aを見つけたい
  - Aを定数倍すれば，任意のデータ間の距離はいくらでも大きくなるが，クラスタリング結果は変わらず意味がない
  - Aをそれ自身のノルムで割ったものを用いる
- 最も近接するリンク不可な例題間の距離を最大化する
  - クラスタリング誤りが起こり易いのは，互いに最も近接するリンク不可例題．これらを遠くに配置することで，誤りのリスクを小さくする
  - SVMのマージン最大化原理と同様

最適化問題

$$\begin{aligned}\max_{A}\ & \min_{(x^m,x^n)\in\mathcal{D}} \frac{d_A^2(x^m,x^n)}{\|A\|_F}\\ \text{s.t.}\ & A\succeq 0\end{aligned}$$

マージン最大化

半正定値条件

$$\|A\|_F=\left(\sum_{i=1}^{D}\sum_{j=1}^{D}a_{i,j}^2\right)^{\frac{1}{2}}$$

等価な問題

$$\begin{aligned}\min_{A}\ & \tfrac{1}{2}\|A\|_F^2 && (1)\\ \text{s.t. } & d_A^2(x^m,x^n)\geq 1 \quad \forall(x^m,x^n)\in\mathcal{D} && (2)\\ & A\succeq 0 && (3)\end{aligned}$$

- (3)の半正定値条件の取り扱いが厄介
- とりあえず，(3)は無視して，(1)と(2)だけからなる問題を解いてみる

制約付き最適化

- ラグランジュ乗数 $\alpha^{(m,n)} \geq 0$ を導入

$$L(\boldsymbol{A},\boldsymbol{\alpha}) = \tfrac{1}{2}\|\boldsymbol{A}\|_F^2 + \sum_{(m,n)} \alpha^{(m,n)}(1 - d_{\boldsymbol{A}}(\boldsymbol{x}^m, \boldsymbol{x}^n))$$
$$= \tfrac{1}{2}\|\boldsymbol{A}\|_F^2 + \sum_{(m,n)} \alpha^{(m,n)}(1 - (\boldsymbol{x}^m - \boldsymbol{x}^n)^T \boldsymbol{A}(\boldsymbol{x}^m - \boldsymbol{x}^n))$$

- 最適性の条件

$$\frac{\partial}{\partial \boldsymbol{A}} L(\boldsymbol{A},\boldsymbol{\alpha}) = 0$$

- 解の形

$$\boldsymbol{A} = \sum_{(m,n)} \alpha^{(m,n)} (\boldsymbol{x}^m - \boldsymbol{x}^n)(\boldsymbol{x}^m - \boldsymbol{x}^n)^T$$

解の半正定値性について

- 行列$\boldsymbol{A}$が半正定であるための必要十分条件

$$\forall \boldsymbol{v} : \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} \geq 0$$

- 実際に計算してみると. . .

$$\boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} = \sum_{(m,n)} \alpha^{(m,n)} \left((\boldsymbol{x}^m - \boldsymbol{x}^n)^T \boldsymbol{v} \right)^2 \geq 0 \quad (\because \alpha^{(m,n)} \geq 0)$$

- 半正定値条件 $\boldsymbol{A} \succeq 0$ は明示的に指定しなくても自動的に満たされる！

双対問題

$$\max \sum_{(m,n)} \alpha^{(m,n)}$$
$$-\tfrac{1}{2} \sum_{(m,n)} \sum_{(m',n')} \left(\alpha^{(m,n)} \alpha^{(m',n')} \left\langle \boldsymbol{x}^m - \boldsymbol{x}^n, \boldsymbol{x}^{m'} - \boldsymbol{x}^{n'} \right\rangle^2 \right)$$
$$\text{s.t. } \alpha^{(m,n)} \geq 0$$

- 線形カーネル以外のカーネルを用いることも可能

SVM学習との関係

- SVM学習

$$\min_{\boldsymbol{w},b} \tfrac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t. } y^m \left(\left\langle \boldsymbol{w}, \boldsymbol{x}^m \right\rangle + b \right) \geq 1 \quad \forall (\boldsymbol{x}^m, y^m) \in \mathcal{T}$$

- リンク不可例題からの距離学習

$$\min_{\boldsymbol{A}} \tfrac{1}{2}\|\boldsymbol{A}\|_F^2$$
$$\text{s.t.} \left\langle \boldsymbol{A}, (\boldsymbol{x}^m - \boldsymbol{x}^n)(\boldsymbol{x}^m - \boldsymbol{x}^n)^T \right\rangle_F \geq 1 \quad \forall (\boldsymbol{x}^m, \boldsymbol{x}^n) \in \mathcal{D}$$
$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \sum_{i=1}^{D} \sum_{j=1}^{D} a_{i,j} b_{i,j}$$

なぜ半正定値問題を解かなくてよいか？

- SVM学習の解

$$\boldsymbol{w} = \sum_m y^m \alpha^m \boldsymbol{x}^m$$

- リンク不可例題からの距離学習の解

$$\boldsymbol{A} = \sum_{(m,n)} \alpha^{(m,n)} (\boldsymbol{x}^m - \boldsymbol{x}^n)(\boldsymbol{x}^m - \boldsymbol{x}^n)^T$$

ソフトマージンの導入

- $\boldsymbol{x}^m - \boldsymbol{x}^n = \boldsymbol{0}$ であるデータ対がなければ，最適化問題は解を持つ
  - $x_i^m - x_i^n \neq 0$ に対応する$a_{i,j}$を十分大きくとれば良い
- しかし，$\boldsymbol{x}^m - \boldsymbol{x}^n \approx \boldsymbol{0}$ なデータのノイズの影響を大きく受けることを避けるため，ソフトマージンSVMと同様にスラック変数を導入

$$\min_{\boldsymbol{A}} \tfrac{1}{2}\|\boldsymbol{A}\|_F^2 + C \sum_{(m,n)} \xi^{m,n}$$
$$\text{s.t. } d_{\boldsymbol{A}}(\boldsymbol{x}^m, \boldsymbol{x}^n) \geq 1 - \xi^{m,n} \quad \forall (\boldsymbol{x}^m, \boldsymbol{x}^n) \in \mathcal{D}$$

- この解も半正定値条件を満たすことが，スラック変数なしの場合と同様に示せる

オブジェクト識別

オブジェクト識別問題

- データベースや文章に含まれる名前（人名や地名）が実世界の同じオブジェクトを参照しているか否かを識別
- 情報検索・情報統合における基本的な問題
- Record Linkage（データベース），Co-reference Resolution（自然言語処理）等，異なる分野において様々な方法で研究が行われてきた

従来のアプローチ

- 名前を含むデータ（データベースのレコードや文書など）をクラスタリングすることで解決
- 適切な類似度や距離を用いることが，識別性能の向上には必要
  - 学習を用いたアプローチ
- 制約の指定に時間や問題領域の知識が必要な場合がある

A. Gupta, V. Harinarayan, D. Quass: Aggregate-Query Processing in Data Warehousing Environments. VLDB 1995: 358-369

Must-be-linked

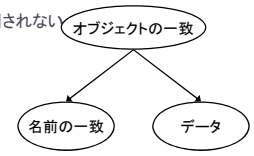
A. Gupta, I. S. Mumick, V. S. Subrahmanian: Maintaining Views Incrementally. SIGMOD Conference 1993: 157-166

Cannot-be-linked

A. Gupta, M. Tambe: Suitability of Message Passing Computers for Implementing Production Systems. AAAI 1988: 687-692

2つの仮定

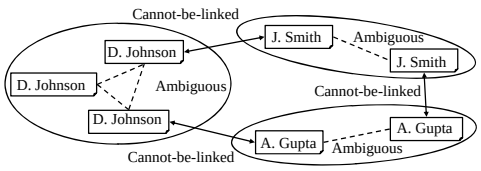
- 1.別名の別オブジェクトへの対応
  - D. Johnsonと J. Smithが同じ人物であることはない
  - 偽名，ペンネーム等は想定しない
- 2.名前とデータの独立性
  - 別名の別オブジェクトのデータ対と，同名の別オブジェクトのデータ対は区別されない
    - 別人のD. Johnson同士の論文同士が，D. Johnsonと J. Smithの論文よりも似ている可能性が高いとは考えない
  - 同じオブジェクトか否かが分かれば，名前の一致とデータの一致は統計的に独立



- 2つの仮定が成り立てば，異なる名前を含むデータの対をリンク不可な例として学習した距離を，同じ名前を持つデータの識別に利用可能

訓練データの生成

- 別名の別オブジェクトを分離する距離は，同名の別オブジェクトも分離できることが期待される
- 異なる名前 (例えば J. Smith と D. Johnson)のデータの組み合わせを，リンク不可例題として使用
- このようなデータ対は機械的に生成可能で，人手による分類を必要としない



データセット

- 多くの著者が混在している，8つのファーストネームのイニシャル+ラストネームの組み合わせを選択
- 各省略名について，対応するフルネームを著者として含む100件のエントリをDBLPから取得
- 異なる省略名のデータを組み合わせることで，訓練集合を作成

Abbreviated name	# of distinct authors	Abbreviated name	# of distinct authors
D. Johnson	17	L. Zhang	31
A. Gupta	23	H. Zhang	26
J. Smith	29	R. Jain	10
R. Johnson	29	J. Mitchell	11

評価指標

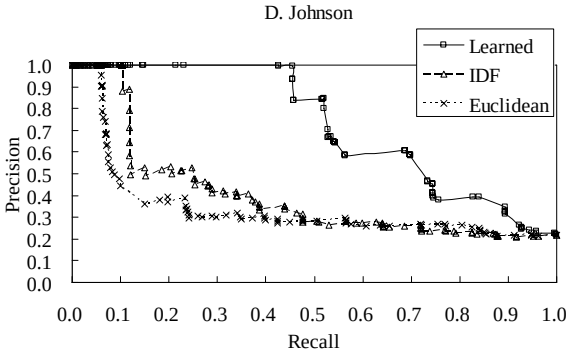
- 学習した距離を用いて，同一省略名のデータをSingle-linkage法でクラスタリング
- クラスタリング結果を，フルネームと比較することで評価（同一フルネームは同一人物と仮定）

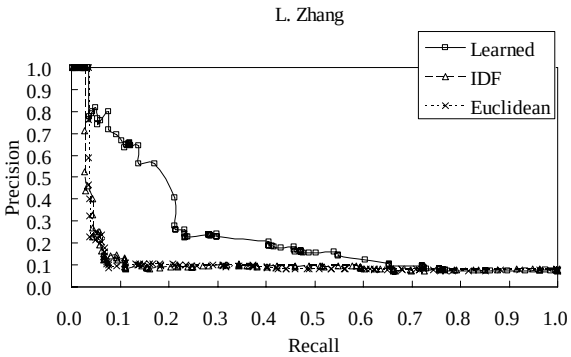
$$\text{Precision} = \frac{|S \cap T|}{|S|} \quad \text{Recall} = \frac{|S \cap T|}{|T|} \quad F = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

S: 同一クラスタに属するデータ対（アルゴリズムが同一のオブジェクトと判定したデータ対）の集合

T: 同一のフルネームを持つデータ対（実際に同一オブジェクトであるデータ対）の集合

- 比較対象
  - ユークリッド距離
  - IDF
    - 文献のエントリでは同一単語が複数現れることはまれなので，単語頻度（TF）による重み付けは行っていない．





*F*の最大値

Abbreviated name	*F*		
	Learned	IDF	Euclidean
D. Johnson	.644	.390	.399
A. Gupta	.490	.170	.169
J. Smith	.417	.270	.292
R. Johnson	.508	.253	.227
L. Zhang	.278	.165	.158
H. Zhang	.423	.226	.226
R. Jain	.709	.569	.552
J. Mitchell	.640	.535	.536

課題

- スケーラビリティの問題
  - 凸2次計画問題は，特異値分解を用いる方法よりは高速に解くことができるが，なお計算負荷は大きい
- 制約の数
  - 全てのリンク不可な対を制約とするとデータ数の2乗
  - 学習に有効なデータ対のみを選択的に用いるSelective Samplingの採用

まとめ

- リンク不可なデータ対のみから距離の学習手法の提案
  - 特異値分解を行わずに，半正定値行列を用いた距離を学習することが可能
- オブジェクト識別問題で，以下の2つの仮定を置いて，別名のデータ対から距離を学習する手法を提案
  1. 別名の別オブジェクトへの対応
  2. 名前とデータの独立性