絶対クラスタリング
と
相対クラスタリング

神嶌 敏弘
産業技術総合研究所
http://www.kamishima.net/

クラスタリング

クラスタリングとは？

↓

与えられたデータ集合をクラスタに分類

クラスタとは？

↓

まぼろし

- 見ようとすれば見えるが，見なければ見えない
- 類似度・目的関数の定義しだいでどうにでもなる

ではどうするのか？

内的規準： 非均一性などメタなクラスタ妥当性を導入

外的規準： 妥当なクラスタの具体例と比較する

教師ありクラスタリング

具体例があるなら，最初から，
それを利用してクラスタリングすれば良い

クラスタリング問題を
絶対クラスタリングと相対クラスタリング
に分けて考えなければならない

準教師ありクラス分類

クラス分類：対象が分類されるクラスのラベルを予測

準教師ありクラス分類 (ラベルあり・なし混在データからの学習)

ラベルあり事例に加えて，ラベルなしの事例も用いると，より予測精度の高い分類器が獲得できる

ラベルなしデータを扱う点でクラスタリングと似ているが，次のいずれかの条件を満たさない問題はクラスタリングとする

クラス分類問題の条件

- 有限個のラベルの集合が事前に分かっている
- 対象と対応付けたラベルが教師情報

制約付クラスタリング

[Wagstaff 01]のCOP-KMEANS法

mustリンク：結ばれたデータの対は同じクラスタに分類される

cannotリンク：結ばれたデータの対は違うクラスタに分類される

制約付と教師ありクラスタリングの相違点

制約のあるデータ以外にも，制約が一般化されて適用されるなら教師ありクラスタリング，そうでないなら制約付クラスタリング

COP-KMEANSは制約付クラスタリング

完全教師ありクラスタリング

完全教師ありクラスタリングの訓練事例集合

N 個の対象集合それぞれに教師情報を与える

$(X_1, Y_1), (X_2, Y_2), \cdots (X_N, Y_N)$

X_i：対象集合 (クラスタに分類される)

Y_i：対象集合 X_i についての教師情報

[神嶌 95] [神嶌 03a] [Daumé III 05] [Finley 05] など

教師情報の例

- must/cannotリンク
- X_i のクラスタリング結果
- 同じクラスタになるべき対象の集合
- データ点の相対的な類似性の大小関係

準教師ありクラスタリング
transductiveクラスタリング

準教師ありクラスタリングの訓練事例集合

一個の対象集合 X に教師情報 Y を与える

$$(X, Y)$$

[Xing 03] [Klein 02] [Chang 04] [Bar-Hillel 03] など

transductiveクラスタリング

- 準教師ありクラスタリングだが，対象集合 X をクラスタリングすることが目的で，X にない新たな対象の分類は考慮しない
- X の分割結果を与えるような教師情報は無意味

[Bilenko 04] [Yu 04] [McCallum 05] など

※ 以後準教師ありクラスタリングとはnon-transductiveな場合をさす

教師ありクラスタリングの分類

クラス分類：ラベル情報が既知でラベル付けによる教師情報

クラスタリング：ラベル情報が未知

制約付クラスタリング：制約を使うが，その一般化はしない

教師ありクラスタリング：教師情報は他の対象にも一般化される

完全教師ありクラスタリング：複数の対象集合に教師情報

準教師ありクラスタリング：一個の対象集合に教師情報

transductiveクラスタリング：新たな対象の分類はしない

絶対/相対クラスタリング

$\mathrm{isc}(x_i, x_j, \pi)$ 分割 π中で対象 x_iと x_jが同じクラスタなら1，違うなら0

クラスタリング関数 $\pi(X)$は，対象集合 X をクラスタリングして分割を出力

対象全集合 $\mathcal{X}$は，未知のものを含めた全ての対象の集合

教師ありクラスタリングとは，対象集合と教示情報から適切なクラスタリング関数を獲得する問題

獲得すべき真のクラスタリング関数が次の性質をもつなら**絶対クラスタリング**，でなければ**相対クラスタリング**

$$\mathrm{isc}(x_i, x_j, \pi(X_1)) = \mathrm{isc}(x_i, x_j, \pi(X_2)),$$

$$^{\forall}x_i, ^{\forall}x_j \in X_1 \cap X_2,\ x_i \neq x_j,\ ^{\forall}X_1, ^{\forall}X_2 \subseteq \mathcal{X}$$

絶対クラスタリングの特徴

$$\mathrm{isc}(x_i, x_j, \pi(X_1)) = \mathrm{isc}(x_i, x_j, \pi(X_2)),$$
$$^{\forall}x_i, {}^{\forall}x_j \in X_1 \cap X_2,\ x_i \neq x_j,\ {}^{\forall}X_1, {}^{\forall}X_2 \subseteq \mathcal{X}$$

一対の対象が同じクラスタに分類されるかは，クラスタリングする分類対象集合中の他の対象とは独立

絶対クラスタリングでのクラスタリング関数の性質

① 絶対クラスタの存在

$\mathrm{isc}(x_i, x_j, \pi(X_1)) = \mathrm{isc}(x_i, x_j, \pi(\mathcal{X}))$ なので，対象全集合の不変なクラスタ(**絶対クラスタ**$\pi(\mathcal{X})$)が存在

② 異なる対象集合間の推移性

$x_i, x_j \in X_1$と $x_i, x_k \in X_2$ についてx_iとx_jが同じクラスタで，x_iとx_kも同じであれば，x_kとx_jは分類対象集合は異なっていても同じクラスタ

reference matching

論文の参考文献を示す文字列の集合を
同じ文献を引用している文字列ごとにまとめる問題

- 表記の違い：“神嶌敏弘” と “T.Kamishima”
  “ICML” と “Int’l Conf. on Machine Learning”
- 表記順の違い：“著者→題名→…” や “著者→年→…” の順

ある文字列集合中の文字列1と文字列2は同じ文献を表している

↓

文字列3が加わっても，文字列1や2が表す文献は不変

↓

文字列が同じクラスタに分類されるかどうかは，
分類する文字列集合には依存しないので，

reference matching は絶対クラスタリング問題

名詞句のcoreference

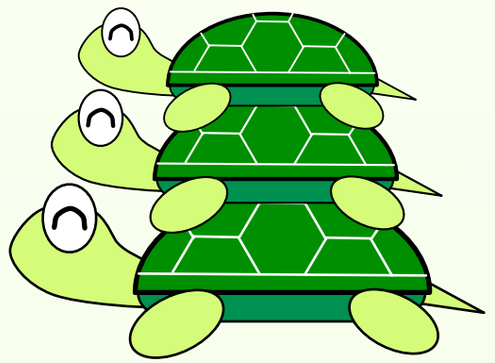
文書中の同じ実体を指し示す名詞句をまとめる問題

“小泉総理” = “小泉純一郎” = “首相” = “彼”

A: 親亀 がいる

B: この亀 の上に 子亀 がいる

C: この亀 の上に 孫亀 がいる



文Aの“親亀”と文Cの“この亀”は違うクラスタ

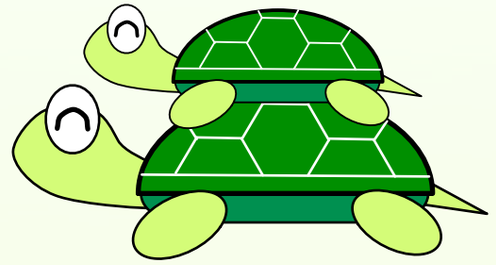
ここで文Bをこの文書から取り除くと……

名詞句のcoreference

文書中の同じ実体を指し示す名詞句をまとめる問題

“小泉総理” = “小泉純一郎” = “首相” = “彼”

A: 親亀がいる



C: この亀の上に孫亀がいる

文Aの“親亀”と文Cの“この亀”は同じクラスタ

文書に含まれる名詞句の構成が変化すると指し示す実体は変化する

名詞句の coreference は相対クラスタリング問題

例題の提示方法 (1)

絶対/相対クラスタリングの区別は，分割する対象集合が変化する場合にのみ生じる

↓

transductiveクラスタリング：新たな対象の分類はしない

対象集合の変化を考えないtransductiveクラスタリングは無関係

相対クラスタリング問題

対象のクラスタへの帰属は分類する対象集合に依存

↓

教師情報は，それが付加されている対象集合に依存しているので，対象集合を一つにまとめたり，変えたりすると教師情報は無効

完全教師ありクラスタリング：複数の対象集合に教師情報

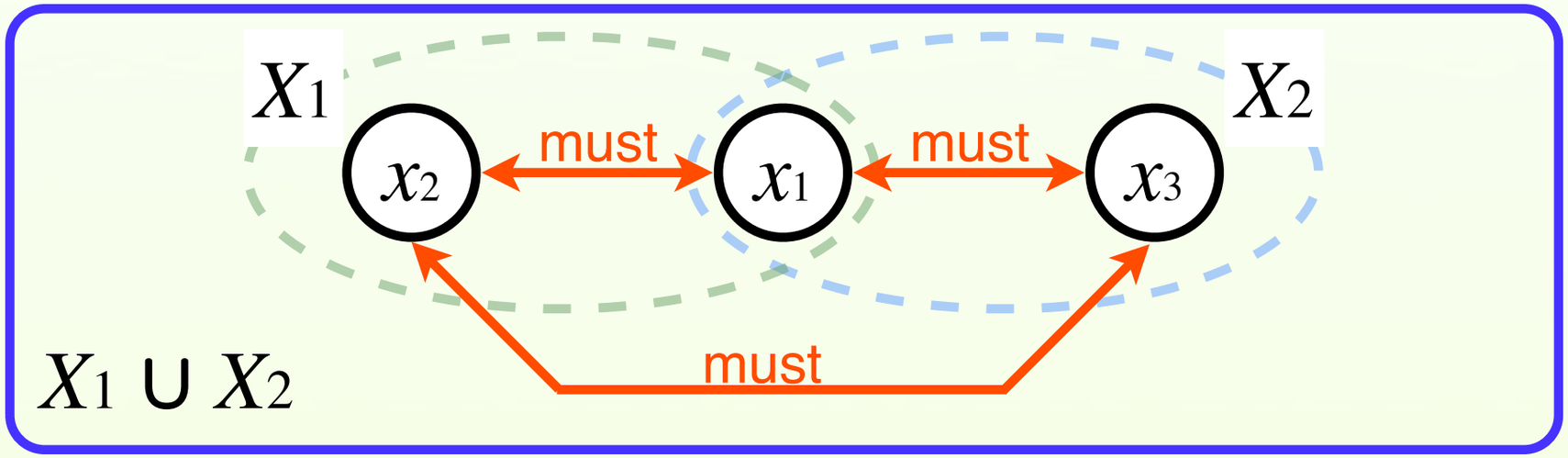
相対クラスタリング問題は完全教師ありクラスタリングの枠組みで解かなければならない

例題の提示方法 (2)

絶対クラスタリング問題

対象のクラスタへの帰属は分類する対象集合とは独立

対象集合を一つにまとめることで，推移性からより多くの教師情報を利用できる



準教師ありクラスタリング：一個の対象集合に教師情報

絶対クラスタリング問題は準教師ありクラスタリングの枠組みで解くべき

必要な特徴量

絶対クラスタリング問題

絶対クラスタが存在

対象を絶対クラスタと対応付け

↓

各対象を記述する属性があれば十分

相対クラスタリング問題

対象集合中の他の対象との関連を考慮して対象を分類

対象間の関連を示した特徴が必要

例：名詞句のcoreference問題での名詞句対の属性

- 受けることのできる代名詞か？ (人を「これ」で受けるのは不正)
- 同義語かどうか？

まとめ

まとめ

- 教師ありクラスタリング手法を整理・分類
- 絶対/相対クラスタリングの概念の提案
  - 絶対クラスタリング問題は，各対象を属性で記述し，完全教師ありクラスタリングの枠組みで解く
  - 相対クラスタリング問題は，各対象に加えて，対象の間の関係を記述する属性も必要で，準教師ありクラスタリングの枠組みで解く

追加情報

ホームページ：http://www.kamishima.net/

おまけ：朱鷺の杜Wiki (機械学習について書き込んでください)

http://www.neurosci.aist.go.jp/ibisforest/