# 教師ありクラスタリング
# と
# 絶対/相対クラスタリング

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

2006年情報論的学習理論ワークショップ(IBIS2006)，2006/10/31-11/2

# クラスタリング

クラスタリングとは？

クラスタの良さを類似度・目的関数で定義 ⇔ 困難

↓

教師ありクラスタリング

類似度・目的関数ではなく，教師情報・制約を導入

教師情報・制約に一致するクラスタが良い

↓

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

に分けて考える必要

# 絶対/相対クラスタリング

$\delta(\{\mathbf{x}_i, \mathbf{x}_j\}, \pi(X))$ は分割 $\pi(X)$中で対象 $\mathbf{x}_i$ と $\mathbf{x}_j$ 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**

$$\delta(\{\mathbf{x}_i, \mathbf{x}_j\}, \pi(X)) = \delta(\{\mathbf{x}_i, \mathbf{x}_j\}, \pi(X')),$$
$$^{\forall}\mathbf{x}_j, \mathbf{x}_i \in X \cap X',\ \mathbf{x}_i \neq \mathbf{x}_j,\ ^{\forall}X, X' \subseteq \mathcal{X}$$

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の特徴

$$\delta(\{\mathbf{x}_i, \mathbf{x}_j\}, \pi(X)) = \delta(\{\mathbf{x}_i, \mathbf{x}_j\}, \pi(X')),$$
$$^{\forall}\mathbf{x}_j, \mathbf{x}_i \in X \cap X',\ \mathbf{x}_i \neq \mathbf{x}_j,\ ^{\forall}X, X' \subseteq \mathcal{X}$$

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

$\delta(\{\mathbf{x}_i, \mathbf{x}_j\}, \pi(X)) = \delta(\{\mathbf{x}_i, \mathbf{x}_j\}, \pi(\mathcal{X}))$なので，対象全集合の不変なクラスタ(**絶対クラスタ**$\mathcal{C} \equiv \pi(\mathcal{X})$)が存在

### ② 異なる対象集合間の推移性

$\mathbf{x}_i, \mathbf{x}_j \in X$ と $\mathbf{x}_j, \mathbf{x}_k \in X'$ について $\mathbf{x}_i$と$\mathbf{x}_j$が同じクラスタで，$\mathbf{x}_i$ と $\mathbf{x}_k$ も同じであれば，$\mathbf{x}_j$ と$\mathbf{x}_k$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

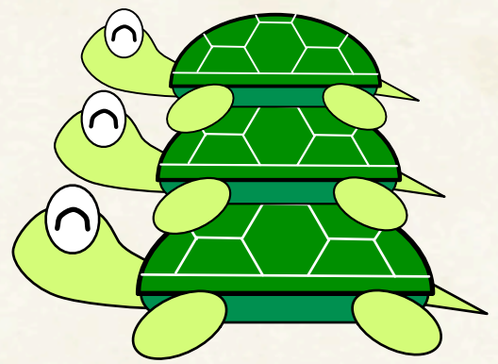
文書中の同じ実体を指し示す名詞句をまとめる問題

“安倍総理” = “安倍晋三” = “首相” = “彼”

A: 親亀 がいる

B: この亀 に 子亀 が乗っている

C: この亀 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

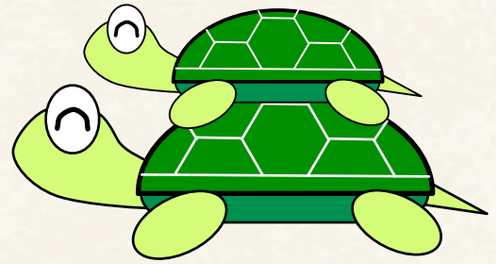
ここで文Bをこの文書から取り除くと……

# 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

“安倍総理” = “安倍晋三” = “首相” = “彼”

**A: 親亀がいる**



**C: この亀 に 孫亀 が乗っている**

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**

# 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，$Y_i$：$X_i$ についての教師情報

任意の $X_{new}$ をクラスタリングする関数を求める

[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
- クラスタ間の類似度の最大値・クラスタ内類似度の最小値

# 準教師ありクラスタリング

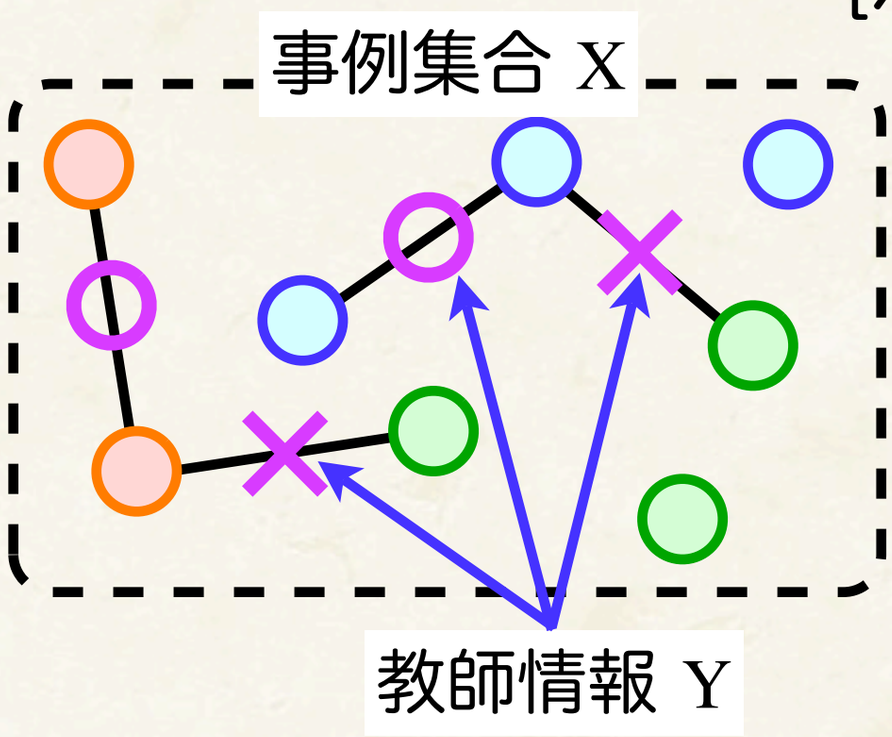
## 準教師ありクラスタリング

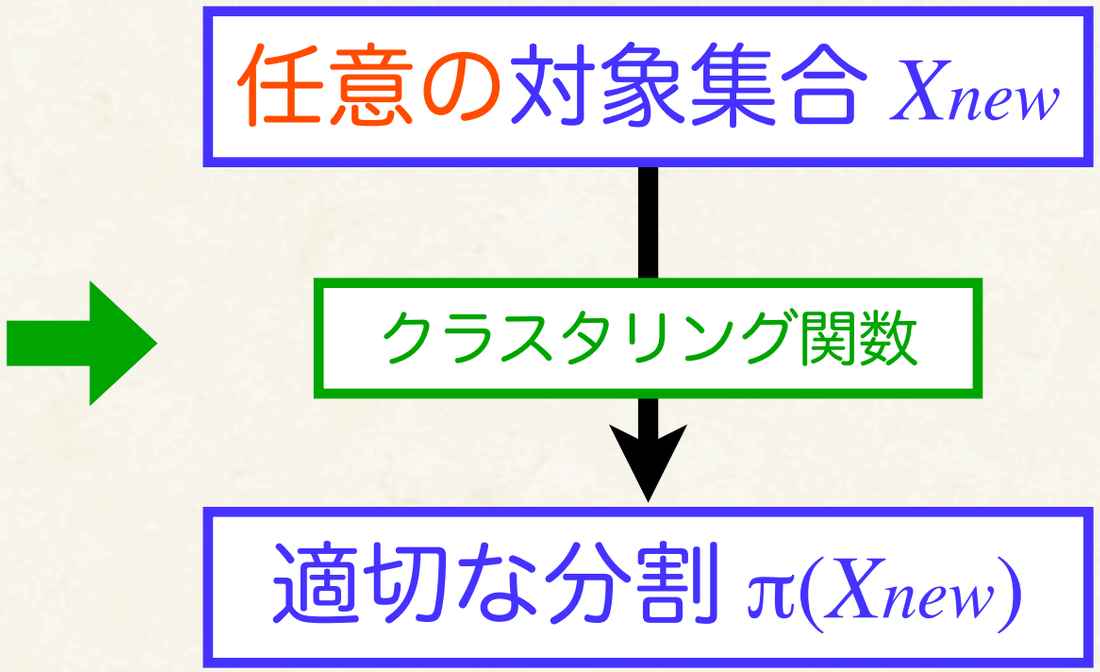
一個の対象集合 $X$ に教師情報 $Y$ を与える

$$(X, Y)$$

- 学習後は $X$ に含まれない未知の事例も分類可能
- 制約のない対象の属性値などは参照しない

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



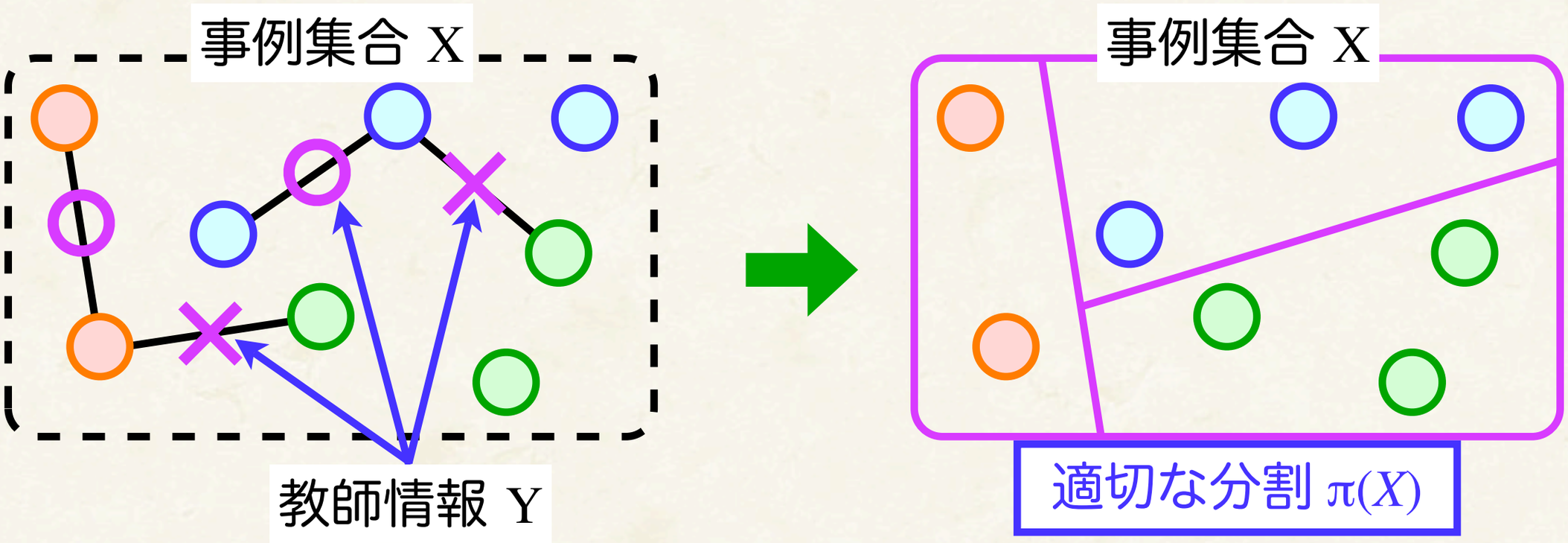
# transductiveクラスタリング

## transductiveクラスタリング

準教師ありクラスタリングと同じ教師情報の形式

- $X$ 中の対象だけを分類することが目的で，$X$ に含まれない対象の分類は考慮しない
- 制約・教師情報のない対象の属性値・位置情報も参照

[Kulis 05] [Yu 04] [McCallum 05] など



# 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

# 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**：未知の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

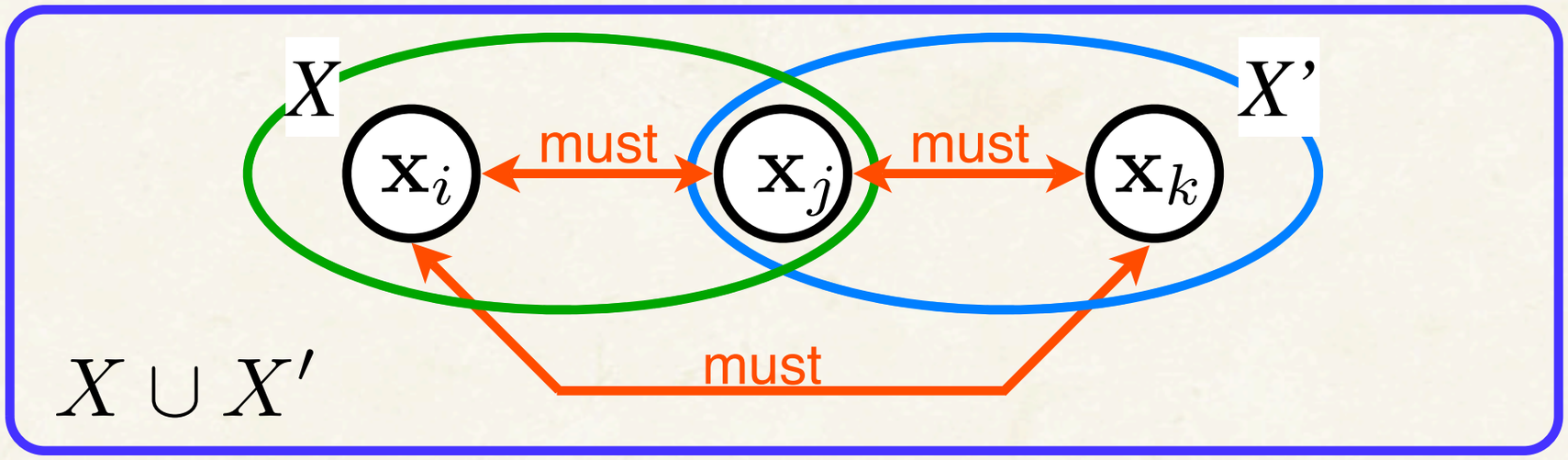
# 例題の提示方法 (2)

**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

# 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/

# 参考文献


- A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. ICML2003, pp.11-18 (2003)
- H. Daumé III and D. Marcu. A Bayesian model for supervised clustering with the dirichlet process prior. JMLR, Vol.6, pp.1551-1577 (2005)
- T. Finley and T. Joachims. Supervised clustering with support vector machines. ICML2005, pp.217-224 (2005)
- 神嶌 敏弘, 美濃 導彦, 池田 克夫, "帰納学習を用いた図面部品の抽出と分類のための規則の形成", 情報処理学会論文誌, vol.36, no.3, pp.614-626 (1995)
- T. Kamishima and F. Motoyoshi, "Learning from Cluster Examples", Machine Learning, vol.53, pp.199-233 (2003)
- D. Klein, S. D. Kamvar, and C. D. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. ICML2002, pp.307-314 (2002)
- B. Kulis, S. Basu, I. Dhillon, and R. Mooney. Semi-supervised graph clustering: A kernel approach. ICML2005, pp.457-464 (2005)
- A. McCallum and B. Wellner. Conditional models of identity uncertainty with application to noun coreference. NIPS 17, pp.905-912 (2005)
- E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. NIPS 15, pp. 521–528 (2003)
- S. X. Yu and J. Shi. Segmentation given partial grouping constraints. IEEE Trans. on PAMI, Vol.26, No.2, pp. 173-183 (2004)