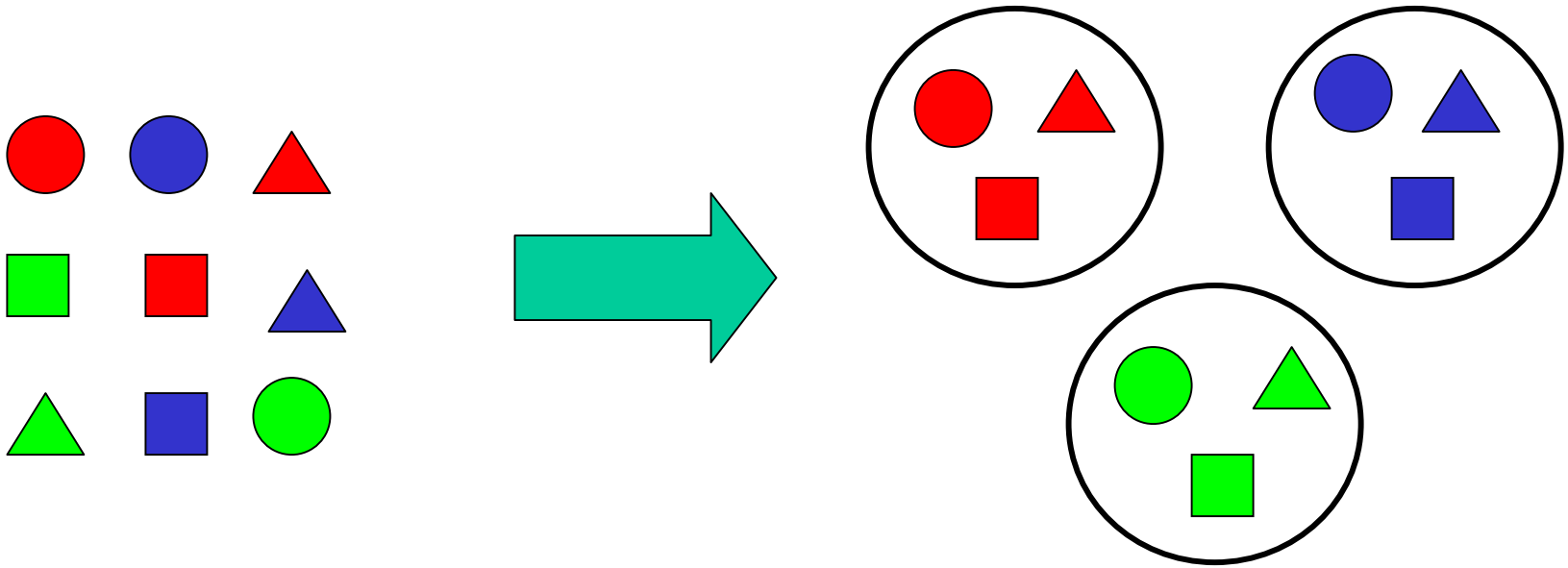
「文書クラスタリングの技法」ゼミ

Ⅰ　はじめに
Ⅱ　文書クラスタリングの特徴と類型
　　A. 文書クラスタリングの一般の特徴
　　A. 文書クラスタリングの技法の類型

新納浩幸

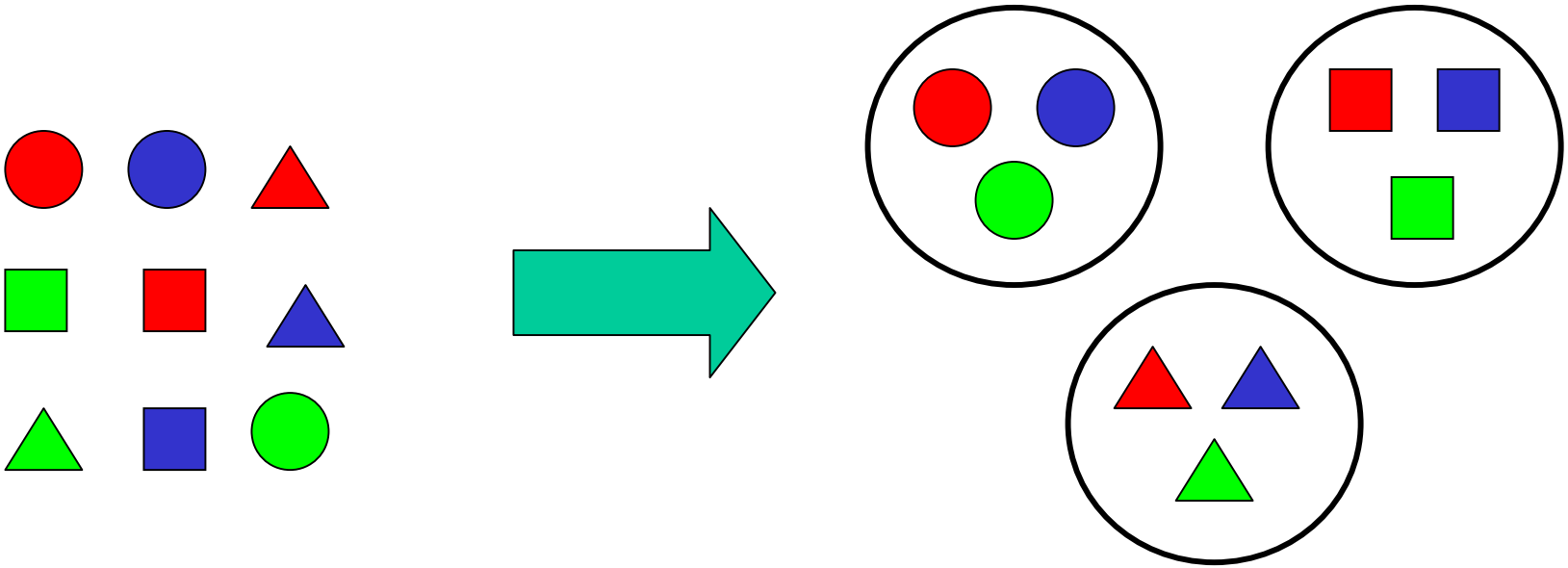
クラスタリングとは

似たものどうしに分類すること



分類の観点

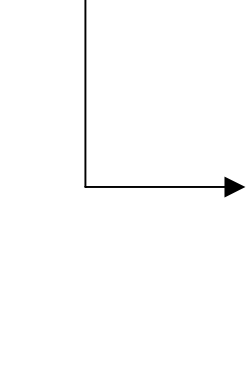
「似ている」にはある観点が必要



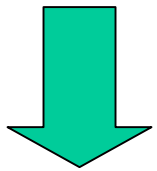
形の類似性

クラスタリングの重要性

データ解析はコンピュータの応用システムの基盤



exploratory（調査）
　　データから法則を予想する
confirmatory（確認）
　　新規のデータに対してある決定を行う



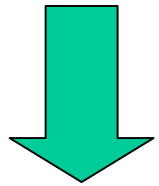
どちらの処理においてもクラスタリングが
本質的な処理となる

クラスタリングの適用分野

パターン分析
grouping
意思決定
機械学習
データマイニング
文書検索
画像分割
パターン分類
…

データには事前の情報がない

例）正規分布など



クラスタリングはデータ間の
関係を調べる基本手法

分類と識別（１）

clustering  discriminate analysis

違いは重要

データを分類する

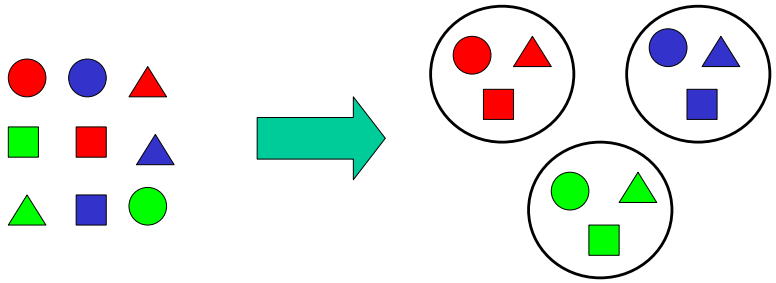
（教師なし学習）

少数の分類されたデータが与えられ，新たなデータをそれらクラスターのどれかに分類する

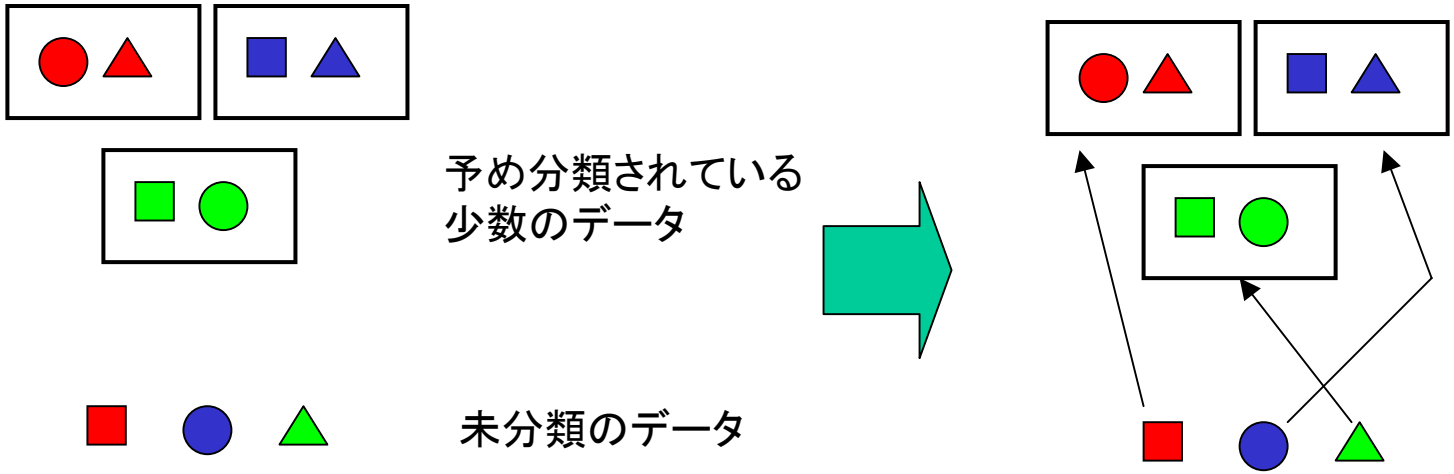
（教師付き学習）

分類と識別（２）

クラスタリング



識別（→帰納学習）



文書クラスタリングの目的

- ● テキストマインイングの要素技術
- ● 検索結果のグルーピング
- ○ 検索性能の向上
- ○ ブラウジングに基づく検索様式

利用する記号（1）

x_{ij} ：文書 d_i における単語 t_j の頻度

n_j ：単語 t_j を含む文書の数

$\tilde{n}_k$ ：クラスター C_k に含まれる文書の数

w_{ij} ：文書 d_i における単語 t_j の重み

利用する記号（2）

$\tilde{w}_{ij}$ ：クラスターベクトル c_k における単語 t_j の重み

$|A|$ ：集合 A の要素数

$\|d_i\|$ ：ベクトル d_i のノルム

（非）排他的クラスタリング

排他的クラスタリングと非排他的クラスタリング

1件の文書が唯一のクラスターに属する

ハードクラスタリング

（一般的）

1件の文書が複数のクラスターに属する

ソフトクラスタリング

文書ベクトル

文書は通常、ベクトル空間モデルにより
ベクトルで表現される

佐々木さんの解説

データがベクトル表現できれば、
一般のクラスタリングの問題となる

文書をベクトル化しない手法

マイナーだけど存在する

○ 共引用や書誌結合で類似度を測る

Small(1997,1999)

○ WWW のリンク構造の利用

Wang, Kitsuregawa (2002)

索引語文書行列

文書ベクトルを縦ベクトルにして、
それを横に並べて作られた行列 Ｍ × Ｎ 行列

Ｍ： 異なり単語数

Ｎ： 文書数

ＭとＮが巨大であり、行列がスパース

類似度の計算

データ間の類似度はベクトル間の距離で
測るのが基本

文書ベクトルでは余弦（コサイン）で測るのが一般的

文書ベクトルの大きさが1で正規化されていれば、
距離と余弦による順位付けは同等

文書クラスタリングの類型

階層的クラスタリング

単連結法、完全連結法、群平均法

処理時間から文書クラスタリングには不適

非階層的クラスタリング

単一パス・アルゴリズム

k-means

単一パス・アルゴリズム

文書ファイルの１度の走査でクラスタリングを行う

⟶ 計算時間 O(N)

Dattola の方法、Rocchio の方法

階層的クラスタリング

凝集型

各データをクラスターに対応させ、
類似度行列を使って、目的のクラスター数
になるまでクラスターを合併してゆく

分割型

全データからなる1つのクラスターを作り、
類似度行列を使って、目的のクラスター数
になるまでクラスターを分割してゆく

その他の手法

次元縮約法の利用

特異値分解で文書ベクトルの次元を縮約し、
縮約したベクトルに対してクラスタリング

← LSI

確率モデルに基づく方法

クラスターに属する確率を求める
（ソフトクラスタリング）、EMアルゴリズム

← pLSI