クラスタリング

平成 31 年 1 月 23 日

富山県立大学 電子・情報工学科 情報基盤講座 3 年

沼田　賢一

はじめに

発表の流れ

1. 背景と目的
2. クラスタリングとは
3. クラスタリングの実行
4. まとめ

背景と目的

背景

クラスタリングとは, ビッグデータの分析においてもっとも重要な地位を占めていてよく使われる手法の一つである.

目的

1. クラスタリングの分類
2. 階層的クラスタリングの実行

クラスタリング

クラスタリングとは

クラスタリングとは, 機械学習のうちの代表的な教師なし学習である.
また, データの集合をクラスタという部分集合に分けることである.
クラスタは内的結合と外的分離の性質を持っている.
クラスタリングで分類されたものは教師なし学習の手法であるため, 最適と呼べるクラスタリングが存在するわけではない.

クラスタリング

クラスタリングの分類

クラスタリングは, 階層クラスター分析と非階層クラスター分析がある.

階層クラスター分析

最も似ている組み合わせから順番にまとまりにしていく方法で, 途中経過を階層のようにあらわせる方法.

非階層クラスター分析

階層的な構造を持たず, あらかじめいくつのクラスターに分けるかを決め, 決めた数の塊にサンプルを分割する方法.

階層クラスター分析

階層クラスター分析の分類

まとまりにしていく過程でできた新しいクラスタと他のクラスタの距離を出す手法として, 最短距離法, 最長距離法, 群平均法, ウォード法などがある.

最短距離法

2 つのクラスタの中から, 最も短い要素同士をクラスタ間の距離としたもの.

$$d_{kc} = \min(d_{ka}, d_{kb}) \tag{1}$$

階層クラスター分析

最長距離法

2 つのクラスタの中から, 最も遠い要素同士をクラスタ間の距離としたもの.

$$d_{kc} = \max(d_{ka}, d_{kb}) \tag{2}$$

群平均法

まとまる前のそれぞれのクラスタとの大きさに比例した重みをつけて平均したもの

$$d_{kc} = \frac{|C_a|}{|C_c|} d_{ka} + \frac{|C_b|}{|C_c|} d_{kb} \tag{3}$$

階層クラスター分析

ウォード法

まとまる前後のクラスタの分散の和と差が最小になるものをまとめていく方法

$$d_{kc} = V(C_k \cup C_c) - (V(C_k) + V(C_c)) \tag{4}$$

非階層クラスター分析

非階層クラスター分析

代表的なものとして k-means 法がある.

k-means 法

k 個のサンプル（初期値）を選択して, 全サンプルそれぞれを k 個のサンプルのうち一番近いものとまとめて、k 個の塊の重心を求めてそこを新しい点として繰り返す. 重心が移動しなくなったらおわり.

$$f(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C} (\bar{x}_k - x_i)^2 \tag{5}$$

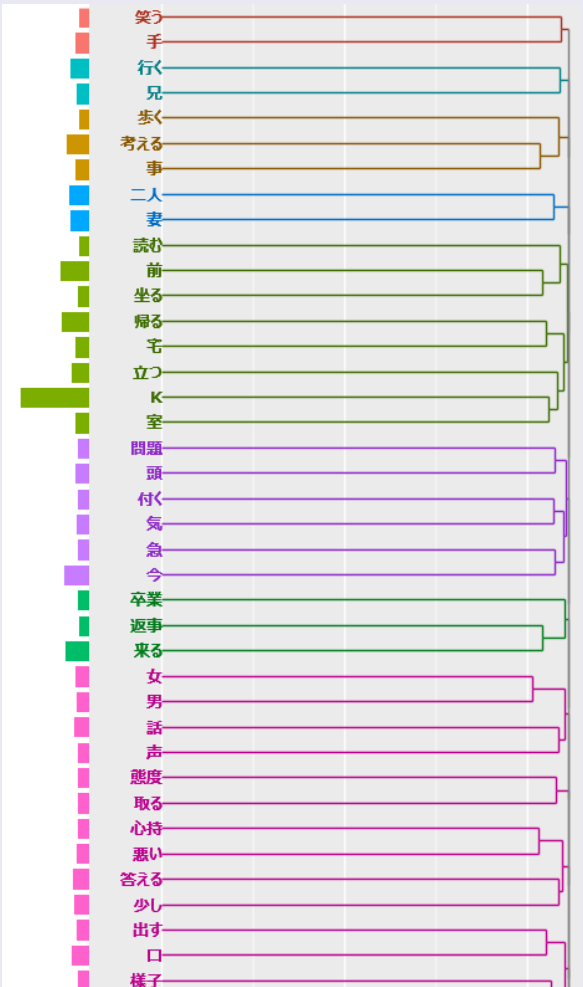
クラスタリングの実行

KH corder3 を使ったクラスタリング

テキスト型（文章型）データを統計的に分析できる KH corder3 を使って, サンプルデータ (夏目漱石のこころ) の語を最長距離法, 群平均法, ウォード法でそれぞれの手法でクラスタリングした.

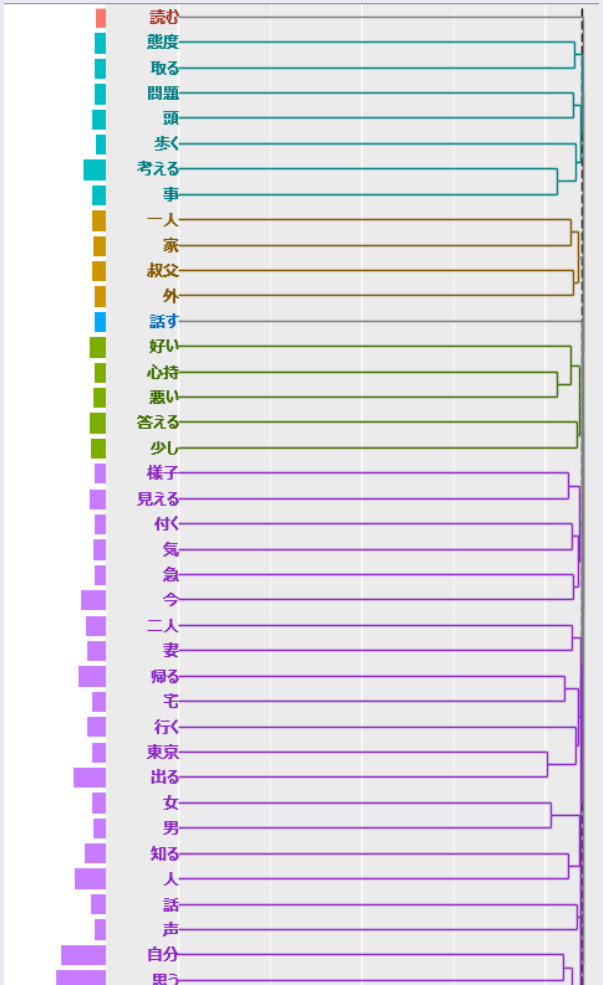
クラスタリングの実行

最長距離法を使ったクラスタリングの結果



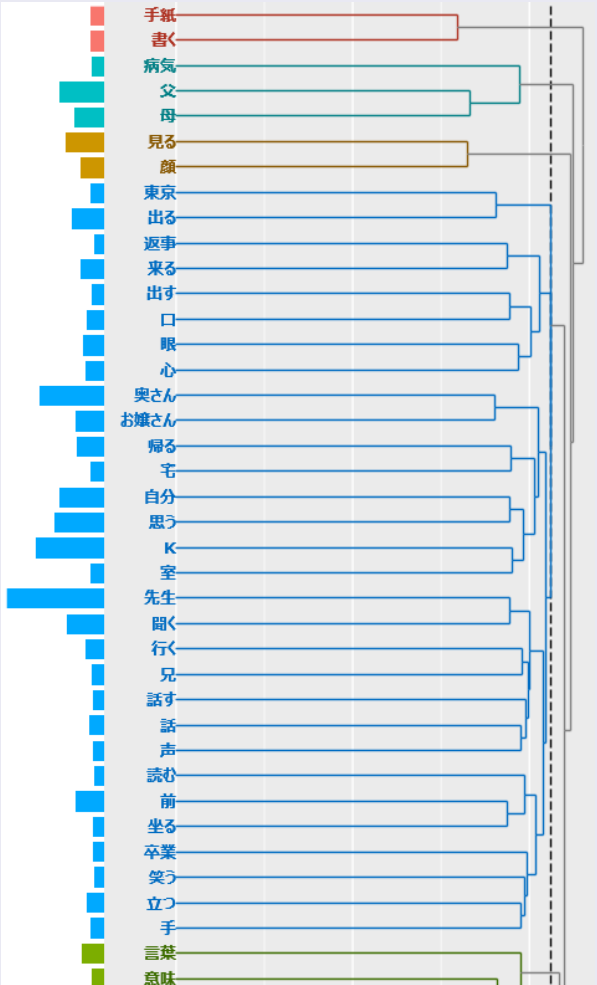
クラスタリングの実行

群平均法を使ったクラスタリングの結果



クラスタリングの実行

ウォード法を使ったクラスタリングの結果



まとめ

学んだこと

1. クラスタリングについて分かった.
2. クラスタリングの手法によってまとまり方がかなり違っていた.