クラスタリング

富山県立大学電子・情報工学科　　指導教員：奥原浩之
1615033　沼田賢一

1　はじめに

1.1　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. クラスター（cluster）とは、英語で「群れ」のことで, 似たものがたくさん集まっている様子を表す.「クラスター分析」とは, 異なる性質のものが混ざり合った集団から, 互いに似た性質を持つものを集め, クラスターを作る方法だ. 対象となるサンプル（人, 行）や変数（項目, 列）をいくつかのグループに分ける, 簡単にいえば「似たものを集めの手法」である.

1.2　目的

今回の目的は, クラスタリングについて理解し, 実際にクラスタリングを実行してみることである.

2　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.

3　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は, 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である.
非階層クラスター分析は, 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である. 階層クラスター分析は,, 最短距離法, 最長距離法, 群平均法, ウォード法などの手法がある. 非階層クラスター分析は,k-means 法という手法がある.
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

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

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.

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

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.

$$d_{kc} = \max(d_{ka}, d_{kb}) \tag{3}$$

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

$$d_{kc} = \frac{|C_a|}{|C_c|} d_{ka} + \frac{|C_b|}{|C_c|} d_{kb} \tag{4}$$

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

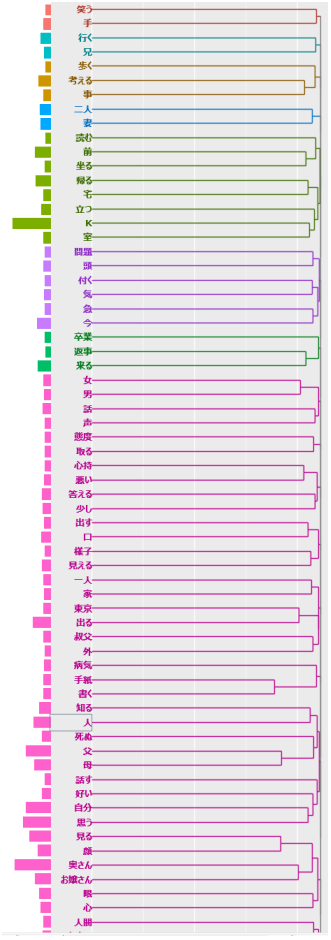
$$d_{kc} = \max(d_{ka}, d_{kb}) \tag{5}$$

最長距離法
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.

$$f(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C} (\bar{x}_k - x_i)^2 \tag{6}$$

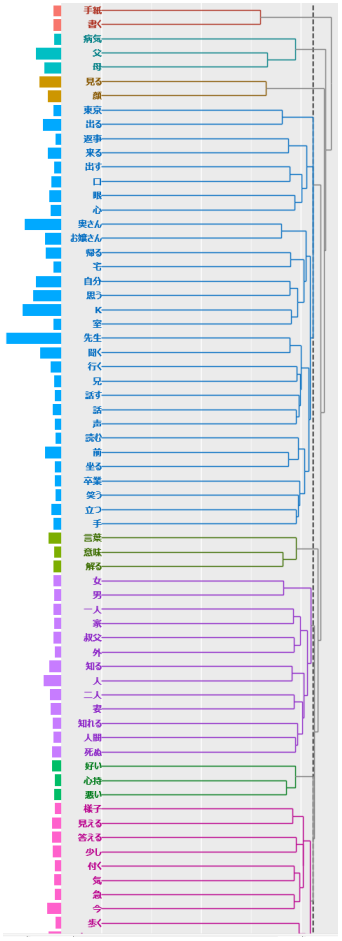
4　クラスタリングの実行

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.



最長距離法でのクラスタリング結果

群平均法でのクラスタリング結果



ウォード法でのクラスタリング結果

5 まとめ

クラスタリングについて理解できた. クラスタリングの手法が分かった. クラスタリングの手法によって, クラスタのまとまり方が大きく異なっていた.

参考文献

[1] 神嶌 敏弘,“クラスタリング (クラスター分析)” http://www.kamishima.net/jp/clustering/