論文メモ

文献番号	0009
日付	2021 年 12 月 05 日
名前	武川海斗

文献情報

著者	Anil K. Jain
英文タイトル	Data clustering: 50 years beyond K-means
和文タイトル	データクラスタリング: 50 年後の K-means
書誌情報	Pattern Recognition Letters, ELSEVIER, Vol. 31, pp. 651–pp. 666, 2010
キーワード	Data clustering,User’s dilemma ,Historical developments, Perspectives on clustering

1　論文の要約

本論文では，クラスタリングの歴史，手法，問題点，動向について俯瞰的にまとめた論文である．そして最後の結びとして，クラスタリングの注目すべき問題と研究の方向性について示している．データが複雑化する現代では，クラスタリングの需要が高まることは驚くべきことではない．ここで重要となるのは，クラスタリングは利用者に対して，仮説を考えるためのヒントを与えるモノでしかないということである．また，完璧なクラスタリング結果を求めるための最良なクラスタリングアルゴリズムは存在しない．それは，クラスタリングアルゴリズムは，クラスタ構造を求めるためのものであり，そこに価値を見出すのは我々だからである．クラスタリング結果をどう表現し，把握するのかは重要なトピックとなる．

ここで注意すべきことは，本論文は 2010 年時点でのクラスタリングの研究動向，歴史をまとめたものであるということである．クラスタリングの研究手法は年々増えており，自分自身で最新の研究手法を追いかけることが必要となる．

2　手法の動向

2.1　半教師付きクラスタリング

クラスタリングとは本質的に，与えられたデータのみの情報からクラスタ分割を行う．そこで，半教師付きクラスタリングでは，側面的な情報を与えることで，より正確なクラスタリングを行う．ここで，「側面的な情報」として，ペアワイズ制約が一般的である．この制約知識にはデー タペアが同じクラスに属するか否かという情報が用いられ，前者は must-link，後者を cannot-link と呼ばれる．実際の半教師付きクラスタリングのアプローチも，既存のクラスタリングの目的関数にペアワイズ制約を加えたものが大半である．

2.2　クラスタリングアンサンブル

教師あり学習における「アンサンブル学習」をクラスタリングに適応した手法である．具体的には，同一のデータに対して，クラスタ数 K などの初期値やアルゴリズムを変えた手法をクラスタリングした結果を組み合わせること (多数決) で，良好な分割結果を得られるというものである．ただし，これにはクラスタ結果を評価するための妥当性基準が必要である．

3　今後の課題

この節では，クラスタリングにおける問題点，今後の課題についていくつか取り上げてまとめる．

3.1　分割結果の可視化

クラスタリング結果を表現することは難しい．例えば，四次元のデータをクラスタリングした場合，クラスタリング結果をどのように表現すれば良いだろうか？ 我々は三次元データまでしか認識することはできず，不可能な問題である．そのため，クラスタリング結果は二次元でプロットされることが多い．そのため，特徴量の選択はユーザーの選択に依存することになる．これがユーザーにとって難しく，データに対するドメイン知識 (専門知識) が必要になるため，問題となっているのである．

3.2　クラスタ数

クラスタ数の自動推定は，最も困難な問題の一つである．K-Means などの手法では，クラスタ数を事前に与え，パラメータとしてクラスタリングを行う．しかし，未知のデータをクラスタリングを行う場合，事前にクラスタ数がわからない場合はほとんどである．そこで，クラスタ数を自動に推定することが重要となる．

3.3　クラスタの妥当性基準

クラスタの妥当性とは，クラスタリングの結果を定量的かつ客観的に評価する指標のことである．一般にクラスタリング結果の有効性を確かめることは難しい．教師あり学習と異なり，正解ラベルが存在しないからである．

クラスタの妥当性基準は内部，相対，外部の三つの異なる基準に基づいている．内部基準に基づく妥当性基準はクラスタ構造とデータの関係 (分散や類似度など) を用いてクラスタリングがうまく行えているかを確かめるものである．次に，相対基準とは，異なるアルゴリズム間を比較し，どういったデータでは優れているのかを分類する．最後に，外部基準とは正解ラベル等を照合することで性能を測定する．しかし，正解ラベルを得られるのであれば，クラスタリングなど必要ないのではないかといった主張を著者はしている．妥当性基準を用いて，クロスバリエーションを行うことで，クラスタ数の推定を行うことができる．そのため，良質な妥当性基準が開発されることはクラスタ数決定の問題を解決することにも繋がるため，重要なトピックであると言える．