特　集「編集委員2007年の抱負」

まだまだあるクラスタリングの研究

神嶌　敏弘　　産業技術総合研究所

全然，「今年の抱負」になっていないが，私の研究分野であるクラスタリングについての話題を勝手に紹介する．クラスタリングは，数世紀前からあるよう気がするが，それほど歴史はない．最も著名な k-means 法 [McQueen 67] でさえ40年程度である．そのため，基本的なことでも興味深い話題が残っており，それらをいくつか紹介する．なお，クラスタリングの基本事項を前提にして述べるので，これらについては拙著 [神嶌 03] などを参照されたい．

最初は，Nevanlinna 賞を受賞*1した Kleinberg の「クラスタリングの不可能性定理（An Impossibility Theorem for Clustering）」[Kleinberg 03] である．これは，対象間の類似度が与えられた場合のクラスタリングを対象とした定理で，次の三つの，望ましいと考えられる性質を同時に備えたクラスタリングアルゴリズム f は存在しないというものである．

(1) スケール不変性（scale invariance）：すべての距離を α 倍しても，f により得られる分割は不変．

(2) richness：対象間の距離を適切に設定すれば，f によって，集合 S の任意の可能な分割を導くことができる．

(3) 無矛盾性（consistency）：f によってある分割を得る．この分割の同じクラスタ内の対象を近づけ，違うクラスタの間の対象を遠ざけるように，対象間の類似度を変換する．この変換後の類似度から f によって，変換前と同じ分割が得られる．

これら三つの条件のうち，二つを満たすアルゴリズムは存在する．具体的に，条件 (1) ～ (3) をそれぞれ満たさなくてもよい方法は，単リンク法によるクラスタリングでそれぞれ次の停止条件を採用すればよい．

(1) 併合時のクラスタ間の距離がしきい値以下

(2) クラスタ数がしきい値以下

(3) 対象間の最大距離 ρ^* について，併合時のクラスタ間の距離がたかだか $\alpha\rho^*$

2003年の KDD で，H. Mannila も自身の受賞講演の中で興味深い研究として取り上げていた．着想のオリジナリティがすばらしいと思う．

次に取り上げるのは，単リンク法などの階層的クラスタリングについてである．これらの手法は，一見すると確率的な生成モデルとは関係がなさそうだが，実はそうではないという論文 [Kamvar 02] である．

既存の階層的クラスタリングの中でも，Ward 法では生成モデルとの関連は以前から指摘されていた．具体的には，クラスタ C_j のセントロイドを $\bar{\boldsymbol{x}}_j$ とし，このクラスタの要素 $\boldsymbol{x}_i \in C_j$ が，正規分布 $N(\bar{\boldsymbol{x}}_j, \sigma \boldsymbol{I})$ に従って生成されるモデルを考える．すべてのデータ点について，この確率密度の積をとり，これをデータ全体の生成確率 $\Pr(\Pi)$ とする．ここで，分割 Π 中クラスタ C_a と C_b とを併合して，新たなクラスタ $C_a \cup C_b$ をつくり，併合後の分割を Π' と記す．このときのデータ集合の生成確率の比 $\Delta\Pr(\Pi, \Pi') = \Pr(\Pi')/\Pr(\Pi)$ を最大にするような C_a と C_b を逐次的に併合することで，階層的クラスタリングを実現する．ここで，$\Delta\Pr(\Pi, \Pi')$ の対数をとると，定数項を除いて Ward 法の併合規準

$d_{\mathrm{Ward}}(C_a, C_b) = \mathrm{ESS}(C_a \cup C_b) - \mathrm{ESS}(C_a) - \mathrm{ESS}(C_b)$

が得られる．ただし，$\mathrm{ESS}(C)$ はクラスタ内分散．

この論文ではさらに進めて，Ward 法以外にも，単リンク法，完全リンク法，群平均法についても対応する生成モデルを示している．例えば，完全リンク法では，上記の正規分布の代わりに，超球内での均一分布を用いた生成モデルに対応することを示した．これにより，今まで経験的に知られていた，同じ大きさの球状のクラスタが得られやすいという，完全リンク法の特徴が，この生成モデルにより説明できる．同様に，単リンク法が縦長のクラスタを生成しやすいことや，群平均法が単リンク法と完全リンク法の中間的な結果を導きやすいことなども説明されている．経験則として知られていた事柄に，合理的な説明を与えた点はとても興味深くはないだろうか？

[神嶌 03] 神嶌敏弘：データマイニング分野のクラスタリング手法（1）—クラスタリングを使ってみよう！—，人工知能学会誌，Vol. 18, No. 1, pp. 59-65（2003）

[Kamvar 02] Kamvar, S. D., Klein, D. and Manning, C. D.: Interpreting and extending classical agglomerative clustering algorithms using a model-based approach, *Proc. 19th Int'l Conf. on Machine Learning*, pp. 283-290（2002）

[Kleinberg 03] Kleinberg, J.: An impossibility theorem for clustering, *Advances in Neural Information Processing Systems*, Vol. 15, pp. 463-470（2003）

[McQueen 67] McQueen, J.: Some methods for classification and analysis of multivariate observations, *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297（1967）

*1 受賞理由は別の業績