研究レポート　No.4　～クラスタ分析によるデータ分類～

2013 年 7 月 25 日　株式会社アイズファクトリー　https://bodais.com/company/

概要

クラスタ分析とは、様々な性質を有したデータの集合から、類似性を持った複数の小集団（クラスタ）に分類分けを行う手法である。この手法を様々な顧客情報（年代・性別・職業など）を含むアンケート結果等に適用すると、クラスタごとに特徴を見出すことができる。例えば、あるクラスタに「20 代」「女性」「既婚」という属性を持つサンプル（顧客）が多く集まっていれば、そのクラスタは、「若い主婦層」といった特徴づけが可能となる。このようなサンプル間の類似性をデータ間の距離や相関係数として定量化し、分散など所定の統計的基準を用いて分類計算を行うのがクラスタ分析である。本レポートでは、クラスタ分析の代表的手法とその用途について説明する。

1.　はじめに

全ての顧客から良い反応が得られる商品やサービスといったものは難しく、趣向の異なる顧客群が複数存在することが通常である。例えば、多数の顧客に向けたテレビ CM などのマス・マーケティングの効果には限界がある。そのため、顧客や市場を細分化（セグメント化）し、特定のセグメントにターゲットを設定して市場戦略を取ることが有効である。これらのことは、コトラーの市場細分化論 [1] に詳細が述べられている。コトラーの市場細分化論に従ってターゲットを設定し、そのターゲットに対する宣伝、キャンペーン、イベント、商品開発やサービス向上を行うと有効な反応が返ってくると予想される。

顧客をセグメント化するには、アンケートなどの顧客情報（行動特性、RFMi などの購買／取引情報／年代・性別・職業など）を活用することができる。顧客情報の母集団を類似の属性や特徴を持った集団に分類すると、行動／反応／嗜好性において共通性を有すると考えられ、セグメントのターゲット設定に利用できる。そして、精度のよいターゲティングを行うには、あらかじめターゲットに応じて分類しておく必要がある。例えば、ある商品にマッチする特徴をもったセグメントがあるとすると、そのセグメントに属する顧客にダイレクトメール（DM）を送付すれば、分類前に比べて反応は高くなると予想される。顧客の絞込が行われているため、DM 経費削減という点でも効果的である。このように、ターゲットを絞った戦略や施策といった面で、顧客のセグメント化を行うクラスタ分析が有効である。

クラスタ分析 [2] を行うためには、顧客の属性を数値化して、ベクトル表現 x_i （特徴ベクトルと呼ぶ）として記述する必要がある。数値データの類似性を特徴ベクトル間の距離として計量化することで、統計分析が可能となる。さらに、クラスタ分析を実施することにより、主観に基づいたセグメント化とは異なり、客観的（統計的）な基準に基づいて顧客を分類することができ、クラスタごとの類似性や特徴を見出すことが可能となる。

クラスタ分析の手法は、いくつかの観点で分類される。その一つとして、ハードクラスタリングとソフトクラスタリングという分類がある。ハードクラスタリングは、全てのサンプルが必ず１つのクラスタに属するという手法である。一方で、ソフトクラスタリング（Fuzzy クラスタリング [3]）は、サンプルが複数のクラスタに属するという手法である。

もう 1 つの観点として、階層的クラスタ分析と非階層的クラスタ分析という分け方がある。階層的クラスタ分析の代表例として、デンドログラム法が挙げられる。それぞれのサンプル（顧客）同士の距離を計算して樹形図を描き、距離の近いサンプルを集約して、クラスタに分類する方法である。一方、非階層的クラスタ分析の代表例としては、k-means 法が挙げられる。k-means 法では、あらかじめ指定したクラスタ数を設定して、各クラスタの中心点を決定し、それぞれのサンプル（顧客）を最も近い中心点のクラスタに分類する手法である。デンドログラム法とともにこれらはハードクラスタリングに属する手法である。

さらに、オフラインとオンラインという観点での分類もある。オンラインでは、時間的に逐次データが更新し続けるストリームデータを走査しながら分類木を構築するクラスタ分析手法 BIRCH も提案されている [4]。

本レポートでは、2 章において階層的クラスタ分析のデンドログラム法を、3 章において非階層的クラスタ分析の k-means 法をそれぞれ紹介する。

2.　階層的クラスタ分析

デンドログラムとは、分析の対象となるサンプルが類似性により統合されていく様子を樹形図で表したものである。サンプル間距離（非類似度）の近いもの同士で統合して小クラスタを作り、次に小クラスタ間距離が近いもの同士で順次統合して上位層のクラスタを作るという手法である。

このときクラスタ間の距離はいろいろな定義の仕方があり、データの性質に応じて適切な手法を選択する必要がある。代表的なものを 2 例挙げておく。

「群平均法」

・ふたつのクラスタのそれぞれからひとつサンプルを取り出してサンプル間距離を計算し、全ての組合せについてサンプル間距離の平均値を、クラスタ間距離とする方法。

「ウォード法」[5]

・2 つのクラスタを併合する際に、クラスタ内の平方和の増分を最小にするようにクラスタを併合していく方法。他の手法に比べて比較的安定した解が得られる。

他にも「最短距離法」「最長距離法」「重心法」による距離の定義もある。

これらの定義を用いて、距離の近いサンプルの統合をして 1 つクラスタを作り、同じく距離の近い別のクラスタと統合をして上位層のクラスタを作る。このように、ある階層のクラスタを一つ上の階層のクラスタへと統合して、さらに上の階層のクラスタへと統合するというボトムアップで階層的なアプローチでサンプル間の類似関係を構造化することになる。これらの手順を樹形図（デンドログラム）で描いたものを図 1 に示す。図 1 では、樹形図にある高さの切断線を設定し、切断線の直下に位置する分岐点でサンプルをクラスタ分類した。「高さ」（サンプル間の距離）が高いほど類似していないことになり、切断線より下では類

似度が高いもの同士で統合されているのが分かる。

デンドログラム法では、統合に必要となるサンプル間距離を全サンプル対について計算することになるため、サンプル数を N とすると $O(N^2)$ の処理時間となり、サンプル数が増大すると計算負荷が大きくなる。

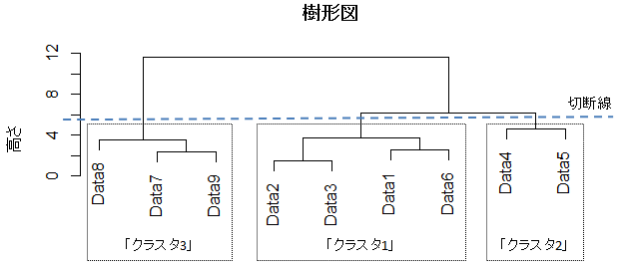

図 1：デンドログラム

3. 非階層的クラスタ分析

k-means 法 [6] は、事前にクラスタ数として指定した中心点を、ランダムに発生し、最も近い中心点にサンプルを統合していく手法である。

サンプルの数を n、クラスタの数を k とすると、

$$\sum_{c=1}^{k}\sum_{i=1}^{n}\min\{u_{ic}\,||x_i - v_c||^2\} \qquad (1)$$

x_i：サンプル i の特徴ベクトル
v_c：クラスタ c の中心
u_{ic} ：サンプル i がクラスタ c に属すると 1、属しないと 0 とする

として、この式（1）を最小化するクラスタ中心を見つけて、サンプルを k 個のクラスタに分類する。

式(1)の最適化問題のアルゴリズムは、以下の通り。

Step1.「初期点の設定：u_{ic}を生成」
ランダムに各サンプル x_i をクラスタに割り振る

Step2.「u_{ic}を固定し、v_cについて最適化」
割り振った各クラスタについて、クラスタの重心 v_c $(c = 1,\cdots,k)$ を計算する

Step3.「v_cを固定、u_{ic}について最適化」
各サンプルx_iと、各クラスタの重心v_cとの距離を求め、x_i を重心が最も近いクラスタに再割り当てする

Step4.「終了条件の判定」
前述の Step3 の処理で、全てのサンプルのクラスタの割り当てが変化しなかった場合、アルゴリズムを終了する。そうでなければ、Step3 に戻る。

k-means 法では、あらかじめ定めた中心点に周辺サンプルを統合するので、トップダウンで非階層的なアプローチとなる。アルゴリズムの特徴として、クラスタ数を事前に設定する必要があり、その後、Step1. のランダムに割り振られたクラスタ中心初期値から最適化を行っていく。

デンドログラム法と k-means 法の違いを表 1 にまとめた。この表にも示すように、k-means 法では、最適化を繰り返してクラスタ分析を行っているが、クラスタ中心の初期点の取り方に依るものが大きく、中心点の取り方によってクラスタリングの結果が異なるのが通常である。また、クラスタ数が事前に分かっていない場合は、クラスタ数を変更して k-means を行い、適切なクラスタ数を探す必要がある。

一方、デンドログラム法は、サンプル同士の距離を算出してから距離の近いものを集約してクラスタを作るので、クラスタ構成は、一意に決まる。また、切断線の設定を変更すれば、クラスタ数の変更も容易である。しかし、サンプルやクラスタ間の距離を計算する必要があるために、サンプル数が多いデータの場合には、計算量が膨大になり処理時間の問題が発生する。

項目	デンドログラム法	k-means 法
結果	再現性あり	再現性なし
計算コスト	$O(N^2)$	$O(kN)$
グラフ化	可能	不可能
アプローチ	ボトムアップ	トップダウン
クラスタ数の設定	事前設定不要	事前設定必要

表 1：デンドログラム法と k-means 法の違い

4. まとめ

本レポートでは、クラスタ分析の代表的な手法である階層的クラスタ分析（デンドログラム法）と非階層的クラスタ分析（k-means 法）について紹介した。3 章でも述べたように、両者の手法に違いがあるために、データのサイズや目的、用途によって適切な手法を選択することが望ましい。

また、上記以外のクラスタ分析方法として、競合学習を用いた自己組織化写像、データ分布密度に基づいた density based methods、格子として分割した Grid-based methods などがあるが、本稿では割愛する。

本稿では、顧客／市場のセグメント化にクラスタ分析を利用する視点で解説したが、クラスタ分析は様々な分野で利用されている。例えば、生物学的な遺伝子種別の分類でのクラスタ分析 [7]、文脈などのクラスタ分析から文書をセグメント化して情報検索の効率を高める方法[8]、立地条件や環境などによる地理的な位置のグループ化 [9]、などの様々な応用事例がある。

今後さらなるデータの複雑化、大規模化が進むにつれて、データ整理や分類、圧縮などが求められるようになる。その際に、データを様々な特徴や性質を持ったクラスタに分類する手法は、さらに重要となってくるであろう。

5. 参考文献

[1] フィリップ・コトラー, ケビン・ケラー著 恩藏直人監修,「コトラー&ケラーのマーケティング・マネジメント 第12版」ピアソン・エデュケーション(2008)
[2] 神嶌敏弘, 人工知能学会誌 Vol.18, pp.59-65 (2003)
[3] 宮本定明, 知能と情報 Vol. 21, pp.1008-1017 (2009)
[4] 神嶌敏弘, 人工知能学会誌 Vol.18, pp.170-177 (2003)
[5] Joe H. Ward, Journal of the American Statistical Association, Vol. 58, pp.236-244 (1963)
[6] J. MacQueen, Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1, pp. 281-297, (1967)
[7] 堀本勝久・藤博幸, 生物物理 Vol.42, pp.110-115 (2002)
[8] 林祐平, 品川徳秀, 第19 回データ工学ワークショップ(DEWS2008), No. B5-6 (2008)
[9] 小川亮・石田貴士, 産開研論 25 巻, pp.13-22 (2013)