# クラスター分析： 知っておきたいこと

**András Lebó**

※この投稿は、Cluster Analysis: Everything You Need to Know(https://rapidminer.com/blog/cluster-analysis/)を翻訳したものです。

製品の不良品を最小限に抑えたい工場長や、次のキャンペーンの結果を予測したいマーケティング担当者にとって、必要なデータの扱いが簡単ではない可能性が高いです。

多くの場合は、有用な予測モデルを作成するのに大量の非構造データを取り込む必要があり、これらのデータを使用できる形式に変換するのに多大な作業が必要となります。問題点は、これらのデータセットが多くの列をもつため、どこから手をつけるべきか決められない点です。

そこで登場するのが、クラスタリングです。

## クラスター分析とは？

クラスタリングは教師なし学習の一種で、結果を考慮せずに、似た特徴を持つデータにグループ分けするプロセスのことを言います。典型的なクラスター分析では、データポイントは類似性を基にグループに分けられ、同じグループ内のアイテムとはよく似ていますが、違うグループ内のアイテムとは大きく異なります。

クラスタリングには様々な種類があり、データセットの状態に応じて Mean-Shift、DBSCAN などの複数のアルゴリズムから選択できることを覚えておきましょう。最もよく知られているのは k-means クラスタリングで、データポイントの中心点をランダムに選択し、反復するごとに位置を最適化することでグループを作成します。

また、全てのデータサイエンスプロジェクトにクラスタリングを適用することはできないでしょうが、代わりに、多大な時間とエネルギーを節約できるケースがあります。

## クラスター分析をなぜ使うのか？

この投稿の序章を読んでいる際に察した方もいるかもしれませんが、クラスタリングの最も大きな利点は、一見すると扱いにくい大きなデータセットを機械学習に使いやすいように変換できる点です。その方法をご紹介します。

大量の非構造データを扱う際、人の手で整理することは非効率的です。実世界では、データセットは数百列、数百万行にも及ぶことがあるため、手動でデータを整理し、カテゴライズすることは非効率的(一部分しか効果的でない)時間の使い方です。

クラスタリングは、大きなデータセットを扱いやすい形式に変えることで、初期の分析にかかる時間を大幅に減らすことができます。列を整理して共通の特徴を見つけることで、クラスタリングアルゴリズムはすばやくデータを整理し、さらに探索する価値のある、意味のあるパターンの発見に役立ちます。

## クラスター分析はいつ使うのか？

クラスタリングは主に、グルーピングやパターン認識を行うものであるため、ビジネスの試みの幅広い範囲に対応することができます。ここでは、最も使用されている例をいくつかご紹介します。

**■顧客セグメンテーション**

マーケティングチームは、共通の特徴を基に顧客のセグメントを開発するのに、クラスタリングを活用することができます。これにより、条件にあったメッセージを送り、同じような興味や行動をもつグループそれぞれにオファーを届けることができます。

**■異常検知**

この技術は、データセット内の他のデータ点とは全く異なる特徴をもつものや、一般的なパターンではないアイテムを見つけるのに使用できます。これにより、クレジットカードの不正利用の発見や、機械の部品に修理が必要かどうかの判断など、様々な場面で役立ちます。

**■レコメンドシステム**

レコメンドシステムは共通の特徴を基に、ユーザーグループに関連性の高い提案を行います。Netflix のおすすめにハマったことや、Amazon のおすすめを基に購入したことがあるなら、この仕組みはすでになじみ深いものです。

# RapidMiner Go でのクラスタリングの仕組み

ここまでで、クラスター分析とは何か、いつ使うのかを説明してきました。ここからは、どのように動作するのかを掘り下げていきましょう。最近はビールの話が多かったので、興味が続くようにワインの話をしましょう。

あなたがワイン通であろうと、もしくは時々ただ一杯を楽しむだけであろうと、ワインには味に影響する特性がたくさんあることは知っているでしょう。この章では、RapidMiner Go を用いて基本的なクラスター分析を実行してみましょう。ゴールは、ワインのタイプを区別できるようなデータセット内のパターンを発見することです。

**■分析**

教師なし学習の代表的なチャレンジは以下の二つです。

- うまく分割するクラスターの発見
- クラスター内の主な特徴の理解

クラスタリングアルゴリズムの結果を評価することになると、最初のステップはたいてい、クラスターをプロットすることです。実世界では、データセットは数百もの列があり、最も良く分割できる列を選択することは大変で、時間がかかります。

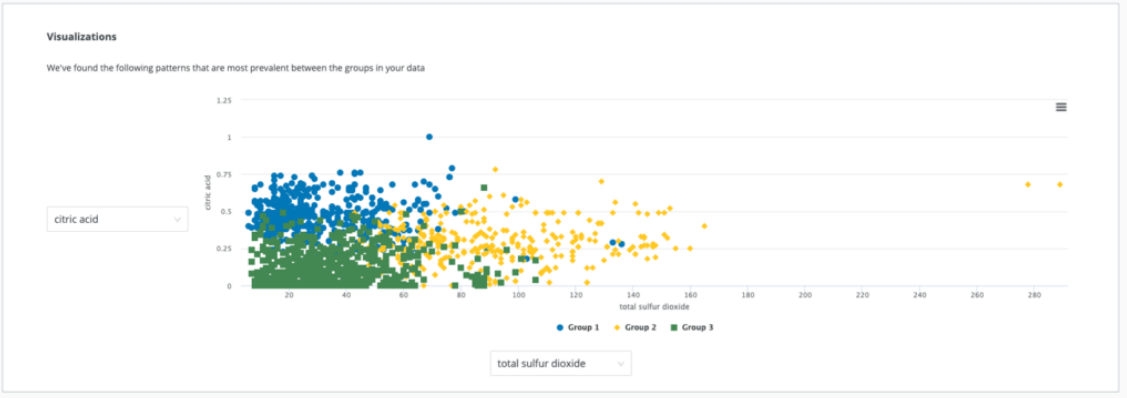
この問題を解決できるかもしれない対処法は、各列のユニークな値の出現回数の平均値を計算して比較することです。RapidMiner Go でパターンを発見する際は、各クラスターグループを特徴づける、最も重要な因子のテーブルが提供されるので、適切な列を選び、期待できるクラスターがあるかを確認できます。

今回のサンプルデータセットの場合は、“driving factors”テーブルは、各グループにプラスかマイナスに寄与する各列の値を示します。カテゴリ値の場合は、緑か赤色のバーで示され、与えられた値がグローバル平均より多いか少ないかを示します。

**■結果**

それでは、結果を見ていきましょう。“driving factors”テーブルが示しているように、グループ 1 のワインは酸が強く、タンニンが多いことから、これらは若いシラーのタイプであると推測されます。対照的に、グループ 2 は酸が平均に近く、タンニンが少ないです。これらより、ピノ・ノワールかジンファンデルと思われます。グループ 3 は酸が少なく、タンニンが多いため、カベルネでしょう。

最も特徴的な二列を軸に選択すると(citric acid と total sulfur dioxide)、三つのクラスターをうまくプロットすることができます(多少は重なる部分があることに注意してください)。




これよりさらに踏み込みたい場合は、実はこのデータセットを使用して予測モデルを作成することができ(目的変数は“Groups”)、モデルシミュレータを用いて自動で将来の測定値を推測することができます。

数回クリックするだけで、様々な精度の 9 つのモデルを作成できます。Deep Learning モデルが最も正確(93%)なようなので、シミュレータにはこれを使ってみましょう。

## まとめ

クラスター分析は、大きな非構造データを用いる際に、データ内の暗黙的な構造を見つけることで、時間や労力の大部分を減らせるかもしれないテクニックです。RapidMiner Go は、各グループにプラス/マイナスに寄与する要因を素早く特定することでさらにデータに踏み込み、データセットの正しい部分を探索するのを可能にし、期待できるクラスターがあるかどうかを確認することができます。

クラスタリングアルゴリズムは、共通の特徴の発見や、データをより使いやすいフォーマットに変換することで、他の方法では見逃していたパターンを特定し、より早く知見を得るのに役立つでしょう。

この投稿で学んだクラスター分析を実際に使ってみたい場合は、RapidMiner Go を無料でお試しください。