機械学習による芸人の特徴分析と距離の定義

2017SS002 安藤芳慧
指導教員：小市俊悟

1　はじめに

M-1 グランプリは漫才の大会であり，審査員の採点により，決勝戦への勝ち上がりや順位が決定される．視聴者にとっては，誰が優勝するかを予測することが，一つの楽しみになっている．このような予測について，俗説と呼べるものがしばしばあるが，M-1 グランプリについても存在している．本研究の 1 つ目の目的は，そのような俗説を機械学習を用いて検証することである．そのために，芸人の様々なデータを収集し，テレビ放送の時点で残っている 10 組から決勝に進む 3 組となるのに, 芸人が持つどのような属性が大きな影響を与えているのかを決定木分析やランダムフォレストを用いて分析する．2 つ目の目的は，決定木分析により得られた決定木を用いて，芸人間の距離を定めることである．決定木の木構造に基づいて，芸人のような定量化しにくいと思われるようなデータに対して，距離を定める方法を提案する．さらに得られた距離を用いてクラスタリングを行った結果を示す．このような決定木を用いた距離の定義と, クラスタリングを組み合わせる方法は, データの新たな分類方法になると考える．

2　使用するデータ

M-1 グランプリの第 2 回から第 15 回までに登場した各芸人について，下記に挙げる属性に関してデータを集めた．[1]

①審査委員の採点による得点の分散, ②結成年数, ③ネタ披露の順番, ④敗者復活戦からの出場, ⑤出身地 (関西, 関東，その他), ⑥平均年齢, ⑦血液型 (A 型，B 型，O 型，AB 型), ⑧所属事務所, ⑨ダークホース (無名からの出場), ⑩ラストイヤー (結成年数が出場資格年数制限と同じ)

このうち，②の結成年数は，大会出場の時点での結成年数であり，さらに，参加可能条件として第 2 から 10 回は結成 10 年まで，第 11 から 15 回は結成 15 年までというものがあったので，第 2 から 10 回までの参加芸人については，結成年数を 1.5 倍することにした．④，⑨，⑩に関しては，該当する場合は 1，そうでない場合は 0 で表現する．また，⑤と⑦については，グループの人数が異なるので，カッコ内に示した項目をそれぞれ属性とする形で細分し，該当者がいる場合は 1，そうでない場合は 0 で表した．⑧に関しても，所属事務所それぞれを属性として列挙し，所属しているか否かを，0 と 1 で表した．

3　分析方法と結果

分析方法として，決定木分析とランダムフォレストを用いる．どちらも，決勝に進む 3 組に入るか否かを，他の属性で判定できるようにすることを目的とする．

3.1　決定木分析

決定木分析により作成される決定木とは，データに対する条件の集まりであり，条件を適用する順番が，木構造により表現され，それに従って条件を適用したときに正しく判定できれば，一般には，それは良い決定木となる．本研究では, Python における scikit-learn パッケージを用いた．本研究では，決定木を作成する際に，収集したすべてのデータを利用するのではなく，3/5 程度をランダムに抽出して決定木を作成し，残り 2/5 程度で，その精度を検証することを繰り返した．作成される決定木は異なることもあった．図 1 は得られた決定木の一例である．図 1 にある

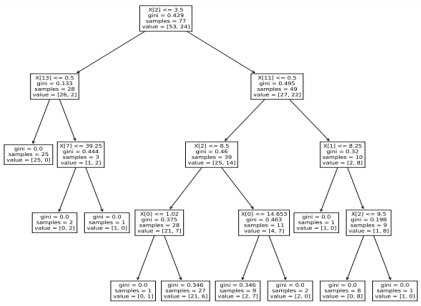

図 1　得られた決定木の例

ような決定木の上位 2 層に現れる属性はデータを大別するので重要であると考える．繰り返しの中で複数得られた異なる決定木，それぞれについて，最上位層に現れた属性と，第 2 層に現れた 2 つの属性を集計した

3.2　ランダムフォレスト

ランダムフォレストは，決定木分析を拡張したもので，使用するデータや属性をランダムに抽出して決定木を作成するということを自動で繰り返し，複数の異なる決定木を作成した上で，最終判定を，作成された決定木の多数決で決める．ランダム性を取り入れることで，外れ値のようなデータや本質的でない属性の影響が軽減されると考えられる．ランダムフォレストでも複数の決定木を作成するため，決定木において重要な属性を選定したのと同様の考えにより，各属性の重要度を算出することができる．ランダムフォレスト自体を繰り返し，それぞれの出力において，重要度が高いとされた上位 3 つの属性を集計した．

4　決定木分析とランダムフォレストの結果の考察

決定木分析とランダムフォレストから得られた重要属性で共通したのは，「結成年数」,「ネタ順」,「分散」,「A 型」であった．このうち,「結成年数」と「ネタ順」は，視聴者

にも重要と感じられている属性であり，勝敗と関係があると噂されている．一方，得点の「分散」は，勝因になるとは予想していなかったが，詳しく分析すると，分散が小さい芸人が勝ち進むようである．「分散」が小さければ万人受けするネタと考えれば，万人受けのネタを持つ芸人が上位になることを示しているのかもしれない．「A 型」に関して，上位入賞への因果関係は考えにくいが，他の血液型と比較しても上位進出者が多いことから，相関関係の一種と考えられる．

5　決定木を用いたデータ間の距離の定義

5.1　距離の定義方法

まず始めに，用意したデータからランダムに抽出したデータのサブセットを複数用意し，それぞれに対して，決定木を作成する．作成される決定木は同一であるとは限らない．作成された各決定木 T は用意したデータそれぞれを T の末端のノードのいずれかに分類する．例えばデータ a は図 2 のノード A に，データ b はノード B に分類されたとする．この時，T によって定まる a と b の距離 $d_T(a,b)$ を T において，ノード A から B に移る際に通過する枝の本数と定める．図 2 の場合であれば，$d_T(a,b)=3$ である．

決定木は同じようなデータであれば，基本的には同じ

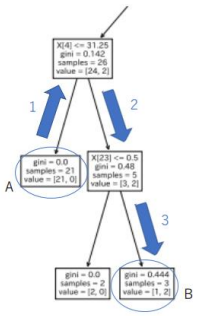

図 2　距離の定義

ノードに分類するので，そのようなデータ a と b の間の距離 $d_T(a,b)$ は小さくなると考えられる．決定木は $T_1, T_2, ..., T_k$ と複数用意しているので，データ a と b の間の最終的な距離を，それらを利用して，

$$d(a,b)=\sum_{i=1}^{k} d_{T_i}(a,b)$$

と定める．複数の決定木を用いたのは，特定のデータに依存することを防ぐためである．

5.2　クラスタリングの適用

決定木を用いて得られた距離から階層クラスタリングを用いて，デンドログラムを作成する．適当な閾値によってデンドログラムを水平方向に切断したとき，依然としてつながっているデータがクラスタとなる．各クラスタは，クラスタに属すデータについて，何かしらの共通する性質を持っていることが期待される．

6　階層クラスタリングの結果

図 3 は，100 個の決定木を作成することで求めた距離を用いて階層クラスタリングを行った時に出力されたデンドログラムである．

定量化しにくいデータの分類のため，複数回デンドログラムの出力を行うと，グループ分けにもばらつきが出てしまうのではないかと考えていたが，決定木を 100 個使用して作成したことで同様のクラスタを与えるデンドログラムを得ることができた．グループごとの特徴は「最近人気があるグループ」「個性派」「関西での活動が多い」「漫才一筋」「その他」に分けられるのではないかと考えられるが，すべてがあてはまるとも思えないところもある．しかし，手法の有用性は十分に感じられる．

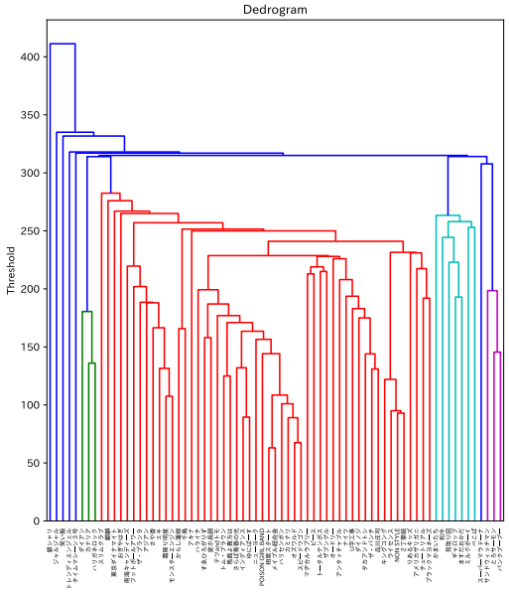

図 3　得られたデンドログラムの例

7　おわりに

本研究では，M-1 グランプリにおいて上位になる芸人の重要要素の解明を目指して決定木分析とランダムフォレストを芸人のデータを適用し，さらに芸人のデータのような定量化にしにくいデータに対して決定木を用いた距離の定義を与え，階層クラスタリングと組み合わせることで，そのようなデータも分類可能な手法の提案を行った．重要要素として選ばれたのは，予想していた「結成年数」・「ネタ披露の順番」の他にも「得点の分散」や，上位入賞との因果関係は考えにくい「血液型 (A 型)」であった．俗説の確認と，思いがけない相関関係を発見できたと考える．本研究が提案する決定木を用いた距離の定義は，カテゴリデータにも適用可能である．必要となる決定木もデータから作成されるので，データが持つ特徴を反映した距離となることが期待できる．

参考文献

[1] M-1 グランプリ公式サイト (2020/08/04 アクセス)
https://www.m-1gp.com/