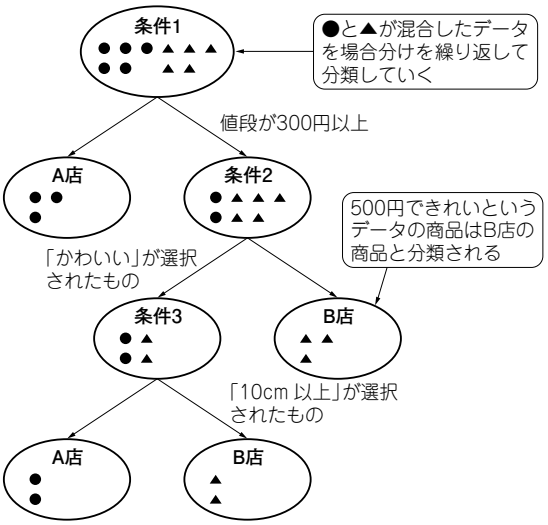
わりとよく使われるタイプは動かしてガッテン！

ダウンロード・データあります

人工知能アルゴリズム探検隊

第30回 決定木の拡張版「ランダムフォレスト」

牧野 浩二



例としてこのようにルールを作れば商品を「値段が300円以下」なら左に分類，「かわいい」ならば右に分類，「10cm未満」なら左に分類．A店で買ったかB店で買ったかを分類できる

図1　ベースになる場合分けを繰り返して分類する手法「決定木」

今回は決定木を拡張した方法である「ランダムフォレスト」を紹介します．決定木は，名前に「木」が入っているように，木構造を持つ分類方法です．ランダムフォレストは，日本語に無理やり訳すと，「でたらめな森」となります．森を構成する木には決定木を使い，決定木を作るときはデータをランダムに選びます．

ランダムフォレストは，普通の決定木に比べて多くのデータが必要となります．そこで，今回は，IRIS（アイリス）データという，データの分類でよく使われる「あやめ」データと，Boston（ボストン）データという「ボストンの家の価格」データを使うことにしました．この2つのデータや統計分析ソフト「R」を使って決定木とランダムフォレストで分類問題と回帰問題を解いてみます．

決定木がベース

● データの混ざり具合を減らすように分類する

決定木は，**図1**のように場合分けを繰り返して分類する手法です．根元が幹で，徐々に枝分かれしていくようにデータの分類を行います．分類していく様子が木のように見えるため木構造と呼ばれています．

● 混ざり具合を示すジニ係数について

決定木の原理は，「データがどれだけ混ざっているかを計算し，その混ざり具合ができるだけ減るように分ける」ということです．ここで混ざり具合という言葉が出てきましたが，これは計算で求めることができます．混ざり方の計算方法には幾つかありますが，ここではジニ係数を紹介します．

ジニ係数は，次に示す計算式で求められる値です．

$$I_G = 1 - \sum_{i=1}^{c} \left(\frac{n_i}{N}\right)^2 = 1 - \left(\frac{n_1}{N}\right)^2 - \left(\frac{n_2}{N}\right)^2 \cdots - \left(\frac{n_t}{N}\right)^2$$

この式の意味を解説します．**図2**のように，●が3個，▲が2個ある場合のジニ係数は，図中の式のように求めます．また，●が3個，■が4個，▲が2個の場合のジニ係数は図中の式のように求めます．次に，**図2**を2つに分けてみます．

ここでは，3種類のある条件を設定すると，**図3**の分け方になるとします．この中でどれが最も良い分け方かを考えてみます．それぞれのジニ係数は，おのおのの図にある通りで，その合計も図中にあります．決定木は，分けた後のジニ係数が減る分け方が，最も良い条件となります．従って**図3**(c)の条件が採用されます．

● できること

▶その1：分類

分類問題とは，例えるとデータを入れるとどの店の商品かを当てるような問題です．決定木は，**図1**のように，多くのデータから分類するためのルールを作ります．分類のルールは，例えば**図1**では値段が「300

第3回　ニューラル・ネットの基本学習法「バックプロパゲーション」(2016年11月号)
第4回　パンダ/トラ/ゾウ…パラメータから分類「主成分分析」(2016年12月号)
第5回　市場調査に使われる多数データのグループ分け「クラスタ分析」(2017年2月号)