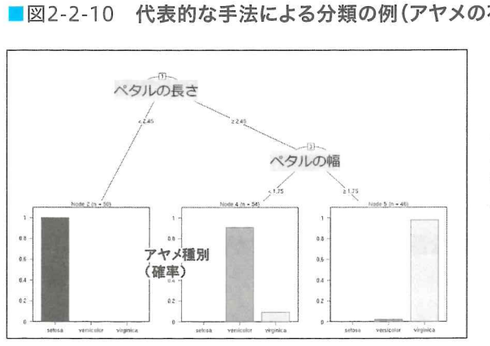
（表2-2-3　続き）

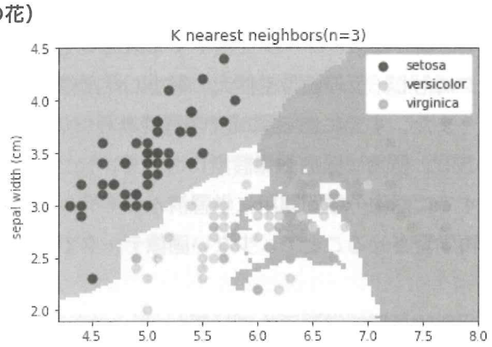
名称	概要	特徴
決定木（Decision Tree）	性質（例：男女）や数値（例：購入数5個以上／未満）に基づき、データを木構造で分類する手法。予測にも使える。	・人が理解しやすい。前処理が少ない。 ・過学習を起こしやすい。
ランダムフォレスト（Random Forest）	ランダムに選んだ学習データと説明変数を用いて決定木群を作成し、その多数決や平均値を結果として出力する。優れた決定木（判別機）を生成するとされる。	・性能の高いモデルを素早く構築できる。 ・中身がブラックボックス化。
k近傍法	入力データに近い方からk個の学習データ（分類ラベル付き）を取得し、多数のものをとって、分類結果とする。	・柔軟にモデルをつくれる。 ・データの量が少ないと効果を発揮しにくい。
主成分分析	多次元のデータに対して正味に効果のあるより少ない成分を抽出（次元の削減）する手法である。	・変数間に相関のないデータに対して有効でない。
k平均法（k-means）	データ点の所属するクラスタを、各データ点からクラスタ重心への距離が最も近いものから選択する。	・手法が理解しやすく、大規模なデータにも適用可能。

図2-2-10に、決定木、k近傍法、SVM、主成分分析の4つの手法によりアヤメの花の分類を行った結果を示す。

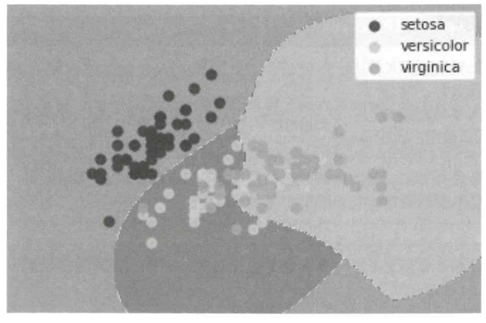
■図2-2-10　代表的な手法による分類の例（アヤメの花）



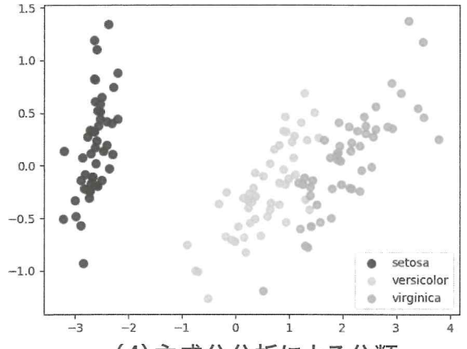
（1）決定木による分類



（2）k近傍法（n=3）による分類



（3）SVMによる分類例



（4）主成分分析による分類

出典：KaggleのWebページ※20を利用して作成

※20 https://www.kaggle.com/

なお、様々な分野で活用され成果を上げている機械学習であるが、あらゆる問題に適用できるわけでもなく、一つの学習済みモデルで様々な用途に対応できるわけでもない。こうした限界を数学的に証明したものに「No Free Lunch定理」がある[1]。本定理は、ある問題に対してアルゴリズムXが別のアルゴリズムYよりも汎化誤差（未知のデータに対する推論の誤り）が小さくても、かならずアルゴリズムXの汎化誤差がYを上回る別の問題が存在することを示す。これにより、あらゆる問題において汎化誤差の良い汎用モデルの存在が否定されることになる。

また、我々が「特徴」と呼んでいるものは、対象が持つ特徴の中で無意識に重視している一部のものだけであり、対象が持つ特徴を全て同等に評価すると識別が困難になる。これを「みにくいアヒルの子定理」と呼ぶ[2]。みにくいアヒルの子は「色」や「大きさ」という特徴が目につくため、普通のアヒルの子との差異が容易に分かるが、くちばしや羽の有無、目の数、足の位置などあらゆる特徴も同じ重みで評価すれば、普通のアヒルの子同士の差異の数と大差がなくなるであろう。機械学習における特徴抽出でも同じで、学習データの選択は、人間が対象を識別するのに適切と考えている特徴を重視して行う必要がある。ネコの画像であれば、耳や目、しっぽなどが写っている画像を学習データとする。このことは、機械学習で使われる「特徴選択」、「次元削減」といった処理が人間の主観的な特徴選択と同様であり、識別にとって本質的なものであることも示している。

(3) 最新技術動向

機械学習に関しては、コストがかかる機械学習システム構築作業の自動化が進められているため、これを紹介する。また、機械学習で使用されている主な手法を紹介する。

学習自動化のための計算機環境整備

機械学習では、人間があらかじめ与えたハイパーパラメーター※21を用いて試行錯誤で行う必要があり、様々な機械学習ツールを使いこなすには、相応の訓練や経験が必要となる。こうした環境が整い始めたのはごく最近であることに加え、大規模な学習環境は相応のコストもかかるため、利用機会も限られている。

問題の技術的解決方法として、機械学習システム構築作業の自動化がある。例えば、様々な応用事例から適切な機械学習手法やアルゴリズム、モデル（ネットワークの構造など）を提案するサービス（DataRobotなど）や特定分野向けの自動設計技術などである。NECは、リレーショナルデータベースでの大規模データ予測分析プロセスを自動設計する技術を開発、dotDataを設立し、機械学習システム構築の自動化に展開した。同社によれば、従来2 ～ 3か月ほど必要だった銀行のデータ分析業務に適用したところ、作業時間が1日に短縮されたという。

また、大規模な機械学習環境をクラウドで提供することで、機械学習システム開発の裾野を広げるような試みも行われている。GoogleやMicrosoftなどは、自社の機械学習システムをサービスとして提供する中で、クラウド側でのトレーニング実行や各種の自動化ツールの提供を行う。また、開発途上の機械学習システムをβ版として無償提供するような試み（GoogleのAutoMLなど）がある。GPUの導入で計算コストが下がったとはいえ、大量の学習データによるトレーニングを行うことは容易ではない。こうしたサービスでは、提供されるニューラルネットワークがすでに利用実績があったり、基本的なトレーニングが行われているなどのメリットがある。

※21 ハイパーパラメーターとは、機械学習においてアルゴリズムを制御するためのパラメーター。例えば、決定木における木の深さ、ディープニューラルネットにおける層の深さや、最適化のアルゴリズムとして何を選択するかなどがある。