第 3 章

決定木

3.1　目的

決定木による識別器の学習アルゴリズムでの数値特徴の扱いや、枝刈りを使った木の大きさの制御について学びます。

3.2　Weka J48

3.2.1　数値特徴の扱いと過学習

決定木の学習アルゴリズムで数値特徴を扱う場合、図 3.1 に示すように、特徴空間を縦横に区切って、同じクラスのものが集まる区画を求めていることになります。

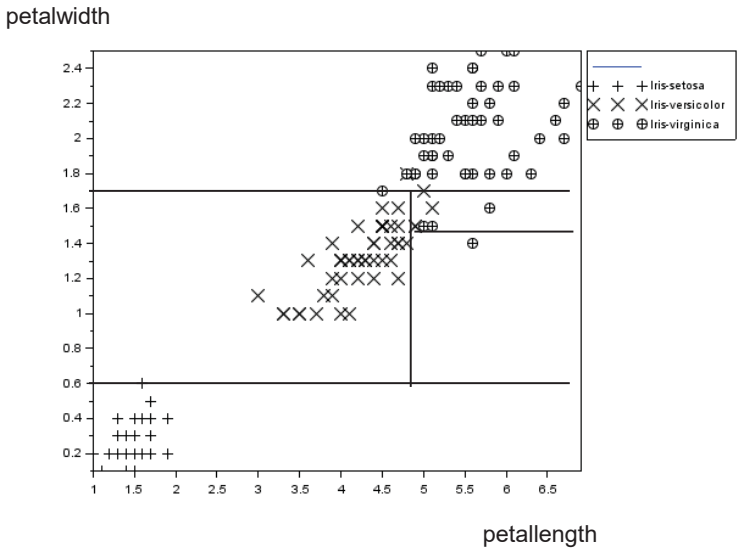

図 3.1　iris.2D データの決定木による識別

この処理を最後まで続けると、学習データに対する識別率はほぼ 100% になる過学習が起きます。そこで決定木学習では、予めデータを学習用と評価用に分けておき、学習用で過学習させたあと、評価用で枝刈りをする処理をします（教科書 p.60）。

実践演習 3-1

iris データから決定木を作成し、Use training set でなるべく高い識別率を実現せよ。その際、調節するパラメータは以下のものとせよ。

- minNumObj: リーフとなる事例数の最小数
- unpruned: 枝刈り処理の有無

実践演習 3-2

iris データから決定木を作成し、10-fold CV でなるべく高い識別率を実現せよ。

実践演習 3-3

credit-g データ[*1] から決定木を作成せよ。その際、識別性能と識別結果の説明可能性（木の単純さ）とのバランスを考えてパラメータを決定せよ。

3.3　sklearn DecisionTreeClassifier

実践演習 3-4

sklearn.datasets にある iris データから決定木を作成し、交差確認法で性能を評価せよ。また、決定木をより単純にしたときの性能の変化を調べよ。

[*1] credit-g データは、金融機関が個人顧客に融資をするかどうか（good or bad）を判定したデータです。特徴は 20 種類で、カテゴリ属性として checking_status（当座預金残高）、saving_status（普通預金残高）、purpose（目的）など、数値属性として age（年齢）、duration（貸出月数）などがあります。