バッギング手法による
分類システムの性能の向上

西田健一郎
北陸先端科学技術大学院大学　知識科学研究科
2000 年 3 月

キーワード：機械学習，分類システム，バッギング，ブートストラップ，KDD.

本研究の目的はバッギング手法を応用した分類システムを作成し、バッギング手法の有効性を確かめることである。

今日コンピュータの発展により、複雑で膨大なデータが存在する。分類システムはこれらのデータから知識を抽出する手法の一つである。分類システムの研究の目的は大きく 2 つある。1 つは正確な予想を得る事である。もう 1 つは予想の構造を解明することある。よい分類は適切な予想を与える。本研究では分類システムについて予想の正確さのみに注目する。このとき分類システムを予想機(predictor)と見る。こうすることによって種々の分類システムを予想の正確さによって比較することが可能となる。

本研究では次の 3 つの分類システムを予想機として扱う。Nearest Neighbor（近隣法）による分類システム（以下 NNC)、Naive Bayes による分類システム（以下 NB)、C4.5(Quinlan,1993)、である。NNC ではまず距離の概念を定義する。そして予想が必要な状態からの距離が近い学習用サンプルの要素をいくつか選ぶ。それら要素のクラスのうち最大数を占めるクラスを予想とする。NB は確率論に基づく予想を計算する。C4.5 は仮説空間において木による探索を行う。この探索の評価関数は GainRatio とよばれ、C4.5 を特徴付けている。

Breiman の提案したバッギング手法は予想機の性能を向上するといわれている。このとき性能とは予想の正確さである。Breiman は彼の CART といくつかの予想機においてバッギング手法の有効性を経験的に示した。そして NNC については有効でないことも実験的に示した(1996)。Quinlan は C4.5 でバッギングが有効であることを実験的に示した(1998)。

バッギング予想機には主にブートストラップ（Bootstrap）と投票（Vote）という手順がある。1つの学習用サンプルからブートストラップにより複数の学習用サンプルを複製する。そのそれぞれから予想機による予想を行う。そして投票によって集めた1つの予想を行う。

バッギング手法はどのような予想機でも正確さを上げることができるわけではない。Breiman(1996)は予想機φが不安定性（Unstability）を持てばバッギング手法は有効であるとしている。この不安定という概念は経験則である。したがって理論的根拠はない。Domingos(1999)は「バッギングが有効であるかどうかは予想機の特性からは判断できない」としている。NB にバッギングを応用した例はない。したがってNBについてはバッギング手法が有効であるかどうかは判断できない。

バッギング手法は予想機の評価法の研究で発見された。くわえて本研究の主目的はエラー率の推定にある。真のエラー率は観念的なもので実際には計算ができないような定義がなされている。したがって予想機の推定エラー率を計算することになる。本研究では交差検定（CrossValidation 以下CV）を使用する。その理由は次の２点からである。１．CVは限られたサンプルでも有効な推定ができるとされている。２．現在予想機の評価法として広く使われている。

CVはサンプルをある数ｎ個に分ける。１個はテスト用サンプルとして使う。ｎ-1個は学習用サンプルとして使う。分けられたサンプルはすべて１度だけテスト用サンプルに選ばれる。そしてｎ個のエラー率の平均が CV によるエラー率となる。サンプルのどの要素も１度はテスト用サンプルとして使用される。

本研究では実際にバッギングの有効性を確かめるために次の４つの実験を行った。

１．有効性が実証されているC4.5にバッギングが有効であるかどうかを確認する。
２．有効でないとされているNNCにバッギングが有効でないことを確認する。
３．有効性のわかっていないNBに対する有効性を調べる。
４．バッギングが有効な特定のサンプル、予想機を用いて、エラー率を求める実験を行う。このときブートストラップの回数を変える。これによりブートストラップの回数と予想の正確さの関係を調べる。

本研究の実験を行うに先立ち３つのシステムを作成した。これらはそれぞれ中心に予想機としてNNC, C4.5, NBを持つ。これらシステムはCVの部位を持つ。これらのシステムはバッギングの部位を持つ。そしてこのシステムは１つの入力サンプルに対して１つのCVエラー率を計算する。これらシステムはUnix環境上にC言語で開発された。実験に使用するサンプルはよく知られたものを使用した。それらのサンプルはUCIの機械学習用データベースから入手した。

実験の結果は次のようである。バッギング手法が C4.5 には有効であることが確認された。NNCには有効でないことが確認された。NBについては有効でないことが実

証された。そしてブートストラップ回数とエラー率の関係の実験はBreimanの「２５回以上のブートストラップは無駄だ」とする主張を支持した。