神嶌 敏弘, 濱崎 雅弘, 赤穂 昭太郎, "飼いならしを用いた協調タグ付けのタグ予測", 2008年度統計関連学会連合大会講演報告集, p.88 (2008)

飼いならしを用いた協調タグ付けのタグ予測

神嶌 敏弘，濱崎 雅弘，赤穂 昭太郎
産業技術総合研究所

文献 [2] にて，我々は新たな学習問題である**飼いならし (Taming)** と，そのための **BaggTaming アルゴリズム**を提案した．今回は，このアルゴリズムの改良について報告する．

飼いならし学習では，**飼育データ (tame data)** と**野生データ (wild data)** の 2 種類の訓練事例集合が混在している．飼育データでは，これから学習したい目標概念と無矛盾なラベルが注意深く選ばれて与えられている．もう一方の，野生データのラベルは，厳密には管理されておらず，目標概念に合致しているものも，そうでないものもあり，完全には信頼はできない．ここで，管理コストが大きいため飼育データを大量に準備するのは困難だが，野生データは大量に獲得できると仮定する．この大量の野生データを用いて，飼育データのみの場合よりも，より高精度の予測を行うことが飼いならし学習の目標である．

野生データの一例として，http://del.icio.us/ に代表される協調タグ付け (collaborative tagging) によって得られるデータがある．協調タグ付けでは，利用者は自身が好きな Web ページを登録し，そのページを表現するタグと呼ぶキーワードを付加できる．さらに，これらのタグを他の利用者と共有することで，登録ページの検索に利用できる．このタグだが，各利用者が個人的な規準に従って自由にタグを付加できる．そのため，多様な規準に基づくタグが使われる．この多様性のため，ある利用者がラベル付けしたタグは，他の利用者にとって適切とは限らない．そこで，ある特定の利用者が自身の一貫した基準で付けたタグを飼育データ，それ以外の利用者が付けたタグを野生データとして扱う．そして，この野生データを併用することで，飼育データの利用者の基準に基づいたタグ付けを，より高精度で予測する問題を扱う．

この飼いならし問題のために BaggTaming と呼ぶ手法を提案した．これは Bagging [1] と同様の方法だが，野生データからブートストラップサンプリングした訓練事例から弱分類器を学習する点と，飼育データに対する予測精度によって弱分類器をフィルタリングする点が異なる．この BaggTaming に次の 2 点の改良を行った．

(1) 予定個数の，受理可能な弱分類器が得られるまで何度も弱分類器の学習を反復していたが，これを，一定個数の弱分類器を学習し，そこから受理可能なものだけを使った．受理された弱分類器が全くなかった場合には，飼育データから学習した分類器をデフォルトとして利用する．

(2) 以前は，飼育・野生を合わせたデータで訓練した分類器の飼育データに対する予測精度を，弱学習器のそれが上回るかどうかというヒューリスティックな受理基準であった．これを，飼育データのみで訓練した分類器より有意に予測精度が悪いとはいえない場合（実験では危険率 5%で）に弱分類器として採用するようにした．

20 種のタグそれぞれについて，そのタグを付加すべきかどうかを識別する二値識別問題（実験の詳細は文献 [2]）に適用した結果を示す．なお旧手法の弱学習器数は 30 個，新手法では最大 100 個（実際の数は平均約 45 個）とした．

データ数	ALL	1/2	1/4	1/8	1/16
旧手法	5/2	8/3	8/2	10/2	11/1
新手法	2/0	6/1	8/1	10/0	9/0

20 種のタグのうち，危険率 1%で正解率の差を検定し，BaggTaming が，飼育データに対する Bagging より良かった場合の数を「/」の左に，悪かった数を右に，いわば勝敗表として示した．表の右にゆくほど飼育データ数が減るが，そうした場合に BaggTaming がより有効であることが分かる．今回の改良で，実行時間は平均で 30%ほどに短縮され，飼いならしによって予測精度が悪くなってしまう状況をほぼ抑制できた．

[1] L. Breiman. Bagging predictors. *Machine Learning*, Vol. 24, pp. 123–140, 1996.

[2] 神嶌敏弘, 濱崎雅弘, 赤穂昭太郎. 飼いならし — 飼育・野生混在データからの学習. 人工知能学会全国大会（第 22 回）論文集, 2D1-3, 2008.