S^3Bagging による高速な分類器生成

寺邊 正大 *, 鷲尾 隆 **, 元田 浩 **

三菱総合研究所 *, 大阪大学 産業科学研究所 **

データマイニングの過程では，意思決定者がデータから必要とする知識を得ることができるまでに，分類器の生成を繰り返さなければならないことが多い．よって，データマイニングツールには大量のデータから正確な知識を抽出するだけでなく，意思決定者の要求に応えるべく高速に抽出することも必要である．高速に分類器を抽出する方法としてはサブサンプリングを行うことにより訓練データ数を減らすことが考えられるが，一般的には訓練データ数が減少すると分類器の分類精度が劣化する．本論文では，サブサンプリングとコミッティ学習手法の１種である Bagging を組み合せた新しい分類器生成手法である S^3Bagging (**S**mall **S**ub**S**ampled **Bagging**) を提案する．S^3Bagging では，複数の分類器をサブサンプリングされたデータを用いて並列に生成し，生成された分類器を用いてコミッティ分類器を構成することにより，サブサンプリングによる最終的な分類精度の劣化を防ぐことができる．また，S^3Bagging の性能について実験を通じて確認した結果を合わせて示す．

Fast Classifier Generation by S^3Bagging

Masahiro Terabe*, Takashi Washio** and Hiroshi Motoda**

Mitsubishi Research Institute, Inc.*, I.S.I.R, Osaka University**

In the data mining process, it is often necessary to induce classifiers iteratively until the human analysts complete extracting valuable knowledge from data. Therefore, the data mining tools need extract accurate knowledge from a large amount of data fast enough in responce to the human demand. One of the approaches to answer this request is to reduce the training data size by subsampling. In many cases, the accuracy of the induced classifiers becomes worse when the training data is subsampled. We propose S^3Bagging (**S**mall **S**ub**S**ampled **Bagging**) that adopts both subsampling and a method of committee learning, i.e., Bagging. S^3Bagging can induce classifiers efficiently by reducing the training data size by subsampling and parallel processing. Additionally, the accuracy of the classifiers is maintained by aggregating the result of each classifier through the Bagging process. The performance of S^3Bagging is investigated by carefully designed experiments.

1　はじめに

分析者がデータマイニングを行う過程では，自らが必要とする知識の抽出が完了するまでに，それまでに得られた知見にもとづいて抽出方法に工夫を加え，さらに知識抽出を繰り返さなければならない場合が多い．このため，データマイニングツールには，高い精度の知識を抽出するだけでなく，高速に知識を抽出し分析者に提示することにより新たな知識抽出への示唆を与えることが期待される．

データマイニングツールにより抽出される知識の代表的なものである決定木などの分類器を高速に抽出するための方法については，これまでに多くの研究が行われているが，これらのアプローチは以下のように類型化できる [3].

- 大量のデータを高速に扱えるように分類器を導出するアルゴリズムを改良する方法.
- サンプリングなどによりデータ分類器を導出するアルゴリズムに投入する方法.

前者については，これまで処理の高速化に有効なヒューリスティクスの導入やアルゴリズム内部の並列化などの研究が行われてきた．一方，後者のサンプリング，あるいは事例選択の研究については，分類器の精度向上に焦点が当てられており，サンプリング処理自体に多くの時間を要するものが多く，高速化の観点からの議論が少ない．

一方，機械学習の分野では，分類器の分類精度を高めるための方法として，コミッティ学習 (committee learning) に関する研究が盛んである [8]．コミッティ学習の枠組みでは，元学習データに対してサンプリングを行うことにより元学習データと事例の構成を異ならせた学習データを複数準備し，それぞれから生成した分類器（メンバー分類器）の集合を最終的な分類器（コミッティ分類器）とする．コミッティ分類器による分類結果は，メンバー分類器による多数決をとることにより決定される．

コミッティ学習の主なものに，Boosting[6] と Bagging[1] がある．Boosting は，Bagging に比べて分類精度を向上させることが多いが，メンバー分類器の生成を系列的に行わなければならないため，コミッティ分類器の構成に多くの処理時間を要する．また，Bagginig に比べると分類性能が学習データに含まれるノイズの影響を受けやすいことら知られている．これに対して，Bagging はデータを問わず安定して分類精度を向上させることができる．さらにメンバー分類器の生成を並行して行えるために，並列処理が可能な環境下では１つの分類器を生成するのとほぼ同じ時間でコミッティ分類器を構成できるという特長を持つ．

本稿では，大量データから高速に分類器を生成する手法としてサンプリングとコミッティ学習の 1 つである Bagging を併用する手法である S^3Bagging(**S**mall **S**ub**S**ampled **Bagging**) を提案する．S^3Bagging は，サンプリングを行うことにより分類器の生成に要する処理時間の短縮し，さらに Bagging によりコミッティ分類器を構成することにより，サンプリングによる分類精度の劣化を防ぐことを狙った手法である．

2 S^3 Bagging

高速に精度の高い分類器を生成する手法として，Bagging とサブサンプリングを併用する S^3Bagging を提案する．

まず，元データ D の r%の事例数分をサブサンプリングした学習データ D_t を T 個生成する．ここで，D_t を生成するためのサブサンプリングは，独立に復元抽出法を用いて行う．また，$r < 100$ であるので，サンプリングされた学習データの事例数は元学習データの事例数 $|D|$ よりも小さい．次に各学習データ D_t から分類器 C_t を生成する．この生成された T 個のメンバー分類器 C_t により構成される C が，S^3Bagging により生成されるコミッティ分類器である．

ここで，各サブサンプリングと分類器の生成過程は，独立に行われるので並列処理が可能であることに注目されたい．T 個のサブサンプリングと分類器の生成が並列に行われる場合には，各サブサンプリングとメンバー分類器生成の処理に要する時間を $proc\_time_t(t = 1, \ldots, T)$ とすると，コミッティ分類器の構成までに要する時間 $proc\_time_c$ は，$proc\_time_c = \max_t\ proc\_time_t$ である．よって，全体の処理時間はサブサンプリングを行い分類器を 1 つ生成するのに要する時間とほぼ同等のものである．また，通常の Bagging と異なり，サブサンプリングを行っているため，分類器の生成に要する時間も全学習データを用いる場合に比べて短縮することができる．

このように S^3Bagging は，サブサンプリングによる分類器の生成に要する時間の短縮と Bagging による分類精度の向上，および並列処理への適性という特長を活かしたアルゴリズムとなっている．

S^3Bagging のアルゴリズムを特徴づける主なパラメータは，サブサンプリング率 r である．この値を大きくすると，学習データに用いる事例数が増えるために分類器の生成に要する時間が増加するが，サブサンプリングによる分類精度の劣化は防ぐことができる．一方，サブサンプリング率を小さくすると，逆に学習データの事例数が減ることにより，一般に分類器の生成に要する時間は短縮されるが分類精度は劣化すると考えられる．次章以後では，このサブサンプリング率の設定とコミッティ分類器の生成に要する時間，および分類精度との関係を確認するために実験を行った結果を示す．

3 実験

3.1 サブサンプリング率と分類精度

S^3Bagging におけるサブサンプリングを小さく設定することによる分類精度の劣化と，コミッティ分類器を構成することによる分類精度向上の効果について確認することを目的として実験を行った．学習器のアルゴリズムには，決定木アルゴリズムの代表的なものである Quinlan による C4.5 Release 8[4] を用いている．C4.5 の実行時のオプションは，導出される決定木に影響を与えるものについては全てデフォルトのものを用いている．また，以下の分類精度の評価は全て枝刈り後の決定木を用いた．なお，S^3Bagging および Bagging におけるコミッティ分類器を構成するメンバー分類器の個数は，一般的なものとして $T = 10$ とした [5]．

表 1: 実験データ

データ名	データ数	属性		クラス
		数値	離散	
breast-w	699	9	–	2
credit-a	690	6	9	2
credit-g	1,000	7	13	2
hepatitis	155	6	13	2
pima	768	8	–	2
sick	3,772	7	22	2
tic-tac-toe	958	–	9	2
vote	435	10	16	2

実験で生成した各分類器は，(1) S^3 Bagging (コミッティ): S^3Bagging により構成されたコミッティ分類器，(2) S^3 Bagging (メンバー) : S^3Bagging においてコミッティを構成する各メンバー分類器，(3) Bagging (コミッティ): 通常の Bagging により構成されるコミッティ分類器，(4) Bagging (メンバー) : 通常の Bagging においてコミッティを構成する各メンバー分類器，(5) All : 全学習データから導出された分類器の 5 つである．なお，メンバー分類器に関する実験結果は，コミッティ分類器を構成する各メンバー分類器の平均をとったものである．

UCI Machine Learning Repository のデータの中から，Quinlan[5] らが実験に用いているデータを参考に選択した．各データの特徴を表 1 にまとめる．実験は，10 分割の交差検定により行った．S^3Bagging と Bagging の実験結果については，各アルゴリズム中のサンプリングによる影響を考慮するために，交差検定に使う元データから，それぞれ 10 回サンプリングをして分類器を生成し，平均をとっている．

さらに，コミッティ分類器内のメンバー分類器による分類結果間の多様性を見るための指標として，エントロピー $E = -\sum_{i=1}^{|C|} p(c_i) \log_2 p(c_i)$ を計算した．ここで，$p(c_i)$ はコミッティ内である事例についてクラス $C = c_i$ と分類したメンバー分類器の割合である．今回用いた実験データでは全てクラス数が 2 であるため $0 \leq E \leq 1$ であり，メンバーが全て同じ分類結果を出したときに 0 となり，各クラスと分類したメンバー分類器が同数のときに 1 となる．

実験結果のうち，コミッティ分類器の誤分類率 (B) とメンバー分類器の誤分類率 (A) の比率 (B/A) を図 1 に示す．比率は全て 100%より小さくなっており，複数のメンバー分類器によりコミッティ分類器を構成することで分類精度を改善することができることが確認できた．また tic-tac-toe データを除いては，サブサンプリング率が 5%以後については，サブサンプリング率が小さいほどコミッティ学習による分類精度の改善が大きい．このようにコミッティ学習は，通常の Bagging だけでなくサブサンプリングを行う S^3Bagging においても分類精度の改善に効果があり，かつ

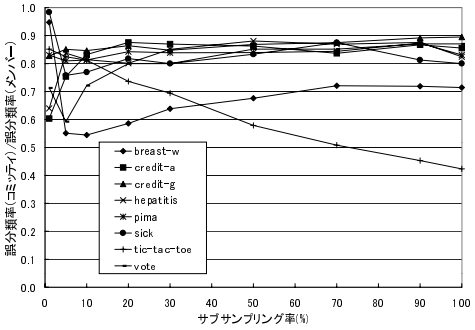

図 1: コミッティ分類器の誤分類率 (B) とメンバー分類器の誤分類率 (A) の比率 (B/A)

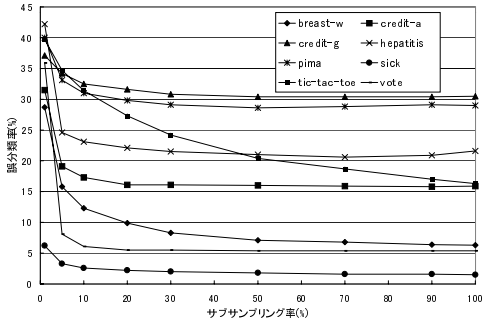

図 3: メンバー分類器の分類精度（誤分類率 (%)）

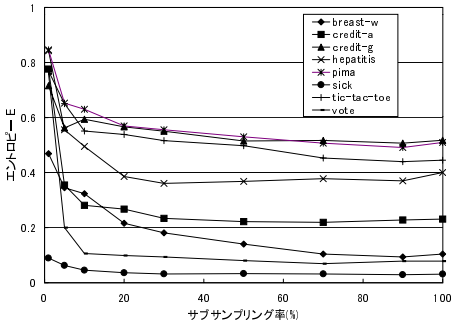

図 2: コミッティ分類器を構成するメンバー分類器による分類結果の多様性（エントロピー）

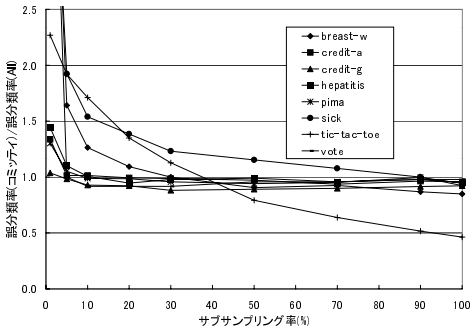

図 4: コミッティ分類器の誤分類率 (B) と全データから導出された分類器の誤分類率 (C) の比率 (B/C)

サブサンプリング率が比較的小さい段階で効果が大きい．

コミッティ分類器を構成するメンバー分類器による分類結果の多様性を測るための指標としてエントロピーを計算した結果を図 2 に示す．コミッティ分類器が分類精度をメンバー分類器に比して分類精度を改善するためには，(a) コミッティ分類器を構成するメンバー分類器が多様であること，および (b) メンバー分類器の分類精度が一定の水準以上であることが必要である．(a) の条件については，サブサンプリング率が 20%以上の場合には，credit-g，pima，tic-tac-toe がエントロピーが大きい[1]．一方，(b) の条件に関する実験結果として，メンバー分類器の分類精度を図 3 に示す．tic-tac-toe についてはサブサンプリング率が 20%のときには誤分類率が 27.3%であり，サブサンプリング率を大きくすることによりさらに分類精度が高まっているのに対して，credit-g，pima ではそれぞれメンバー分類器の分類精度が 30%前後と悪く，サブサンプリング率を大きくしても改善されていない．このメンバー分類器の分類精度の違いが，tic-tac-toe ではコミッティ学習による分類精度の改善が大きく，サブサンプリング率を高めることにより，さらに改善が大きくなるのに対して，credit-g，pima では同じく分類結果の多様性が大きいにも関わらず分類精度が大きく改善されないというコミッティ学習による分類精度の改善効果の違いに表れていると考えられる．また，breast-w ではサブサンプリング率が大きくなるにつれてコミッティ分類器構成による分類精度の改善が小さくなっている．これは，メンバー分類器の分類精度は良くなっているが，図 2 のようにメンバー分類器間の多様性が小さくなっていることによるものと考えられる．

次に，S^3Bagging により生成されたコミッティ分類器の誤分類率 (B) と全データを使って生成した分類器 (All) の誤分類率 (C) の比率 (B/C) を図 4 に示す．全てのデータにおいて，サブサンプリング率を高く設定すれば S^3Bagging を行うことにより分類精度を改善することができることがわかる．一方，各サブサンプリング率におけるメンバー分類器の分類精度を比較するとわかるように，サブサンプリングを行うことにより生成されるメンバー分類器の分類精度は悪くなる．特にサブサンプリング率が 1%のときのように極端に学習事例が少なくなる場合には，分類精度が大幅に悪化する．これは，学習データの事例数が正しく学習を行うのに必要な数に対して不足しているためである．

3.2 サブサンプリング率と処理時間

次に，S^3Bagging により期待される分類器生成までの処理時間の短縮効果について確認するために，大規模デー

[1] ここでは，エントロピーが 0.4 以上を基準にしている

タを使った分析を行った．処理時間 (CPU Time) とは，サンプリングに要する時間と分類器の学習に要する時間の和である．また，S^3Bagging や Bagging などのコミッティ学習では，各メンバー分類器の生成は並列に行えるものとして，コミッティを構成する各メンバー分類器の生成に要した時間のうち最も長いものをコミッティ分類器の生成に要した時間としている．なお，実験に用いた計算機は，OS が Linux OS，CPU が PentiumIII 700MHz，メインメモリが 256M のものである．

実験に用いるデータとして UCI KDD Repository のデータから census-income データを選択した．事例数が 30 万弱，属性数が 40 と比較的大規模なデータである．

実験の結果，処理時間はサブサンプリング率にほぼ比例することが分った [7]．よって S^3Bagging では，サブサンプリング率を低くすることにより処理時間が大幅に短縮することができる．例えば，サブサンプリング率が 5%のときで S^3Bagging に要する時間は 8.3 秒程度であり，全学習データを用いて単独分類器を生成した場合 (All) の 3%弱である．

4 考察

3.1 節の実験結果から，10%程度の低いサブサンプリング率でサンプリングを行った場合でも，S^3Bagging が採用しているコミッティ学習を行うことによって，全学習データより分類器を生成した場合からの分類精度の劣化を防ぐことができることが確認された．また，3.2 節の実験結果からは，S^3Bagging が採用しているサブサンプリングを行うことにより，分類器の生成に必要な処理時間を大幅に削減できることを確認した．

データマイニングの過程では，一度の知識抽出で必要な知識を取り出せることは少なく，抽出された知識を分析者が吟味しながら知識抽出を重ねなければいけない場合が一般的である．このように，意思決定者とのインタラクションの中で知識抽出を行うことを考えると，短い時間で知識を提示することは，意思決定者にその後の知識抽出のきっかけを与えるという点で非常に重要であると考えられる．よって，S^3Bagging のように抽出される分類器の質の劣化を抑えながら，大量データから高速に分類器を生成するような枠組みは，データマイニングツールの実用性の観点から有用である．この分類器の生成に必要な処理時間のコストと，分類精度により代表される分類器の質との間のバランスという観点から 3.1 節および 3.2 節の実験結果を見ると，データにより異なるが，サブサンプリング率で 10%–30%程度，あるいは学習データ数で数百から数千ぐらいに設定するのが適切であると考えられる．ただし，これは今回の実験で用いた C4.5 を学習器として適用した場合の結果にもとづいたものであり，ニューラルネットワークなど他のアルゴリズムを学習器に適用した場合については，実験を通じた検討が必要である．

5 おわりに

本論文では，大規模なデータから高速に分類器を生成する手法として，サブサンプリングと Bagging を併用する S^3Bagginig を提案し，テストデータを用いた実験を通じて性能を確認した．S^3Bagging では，サブサンプリングにより学習データの事例数を減らし，かつコミッティを構成する分類器の生成を並列に行うために処理時間は，サブサンプリング率にほぼ比例する程度である．大規模データを用いた実験からはサブサンプリング率を 20%程度に設定した場合には，全学習データを用いる場合に比べて計算時間を 20%程度に抑えることが可能であることを確認した．また，S^3Bagging は，Bagging の方法によりコミッティ学習を行うことにより，サブサンプリングによる分類精度の低下を抑えることができる．実験結果から，サブサンプリング率が 20%程度にした場合に誤分類率の劣化は多くの場合 20%程度に抑えることができることを確認した．このように，提案手法は並列処理環境下において，大量データから高速に，かつ分類精度の高い分類器を生成するのに有効である．

今後の検討課題として，以下のようなものを考えている．

- 属性間の相関を考慮したサンプリング率決定のための指標の検討：サブサンプリング率を決定する際の参考指標として提案している記述長 [7] は，属性間の相関を考慮していないものである．注目する属性数を絞り込むことにより処理時間を小さくしながら，属性間の相関も考慮することができる指標について検討する必要がある．
- S^3Bagging により抽出された情報の効果的な提示方法の検討：S^3Bagging では抽出されるコミッティ分類器以外にも，メンバー分類器ごとの分類結果，およびその分布などデータマイニングを進める上で有用な情報が含まれている．これら情報の意思決定者への効果的な提示方法について検討していきたい．

参考文献

[1] Breiman, L.: Bagging Predictors, *Machine Learning*, **24** (2), pp.123–140 (1996).

[2] Catlett, J.: Megainduction: a Test Flight, In *Proceedings of the Eighth International Workshop on Machine Learning*, pp.596–599. (1991)

[3] Provost, F. and Kolluri, V.: A Survey of Methods for Scaling Up Inductive Algorithms, *Knowledge Discovery and Data Mining*, **3** (2) pp.131–169 (1999).

[4] Quinlan, R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann (1993).

[5] Quinlan, R.: Bagging, boosting , and C4.5., In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pp.725–730 (1996).

[6] Shapire, R.E.: A brief introduction to boosting, *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pp.1401–1406 (1999).

[7] 寺邊，鷲尾，元田：S^3Bagging による高速な分類器生成，情報処理学会論文誌 数理モデル化と応用,（投稿中）

[8] Zheng, Z.: Generating Classifier Committees by Stochastically Selecting both Attributes and Training Examples, In *Proceedings of PRICAI-98*, pp.12–33, (1998).