1F5-07

構成型メタ学習と選択型メタ学習の比較評価

A Comparison of the Automatic Composition of Inductive Applications to Stacking Methods

阿部 秀尚 *1　　山口 高平 *2
Hidenao Abe　　Takahira Yamaguchi

*1静岡大学大学院理工学研究科
Graduate School of Science and Technology, Shizuoka University

*2静岡大学情報学部
Faculty of Information, Shizuoka University

In this paper, we present a comparison between selective meta-learning scheme and constructive meta-learning scheme with common data sets. Meta-learning scheme enhances the performance of classification tasks. Selective meta-learning scheme combines multiple base classifiers. It has done with bagging, boosting and stacking methods. On the other hand, constructive meta-learning scheme constructs the good classifiers to a given data set. It has done with our tool called CAMLET with a inductive method repository.

We have done the case study of a comparison of inductive applications composed by CAMLET to three stacking methods with StatLog common data sets. As a result of this case study, CAMLET shows us as good performance as the best stacking method, which uses the classification via linear regression for the meta-classifier.

1. はじめに

近年，分類学習の性能を高める手法としてメタ学習の研究が行われている．従来のメタ学習では，初めに所与の訓練データセットに対して分類器を学習し，次に学習された分類器の結果を統合してテストデータセットに対する予測を行わせている．学習された分類器の結果を統合する部分がメタレベルの学習にあたるため，これらの手法はメタ学習と呼ばれる．

従来のメタ学習には，図 1 に示すように，単一の学習アルゴリズムによる分類器を統合する手法と，複数の学習アルゴリズムによる分類器を統合する手法がある．

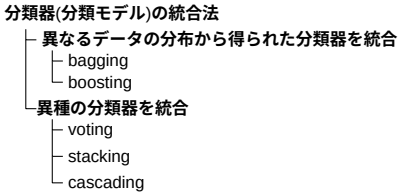

図 1: 選択型メタ学習アルゴリズム

前者は，分布の異なるデータセットに対し学習を行う bagging[Breiman 96]，誤分類のインスタンスの重みを増して学習を繰り返す boosting[Freund 96] がある．後者は，複数の学習アルゴリズムからの分類器による投票を行う voting，分類器の組合せを学習により制御する stacking[Wolpert 92] や cascading[Gama 00] がある．

従来のメタ学習では，学習アルゴリズムを分解せずに利用するため，我々は，このメタ学習を選択型メタ学習と呼ぶ．

選択型メタ学習には，所与のデータセットに対して基本となる学習アルゴリズムが抱える性能面での悪影響を乗り越えられないという問題がある．

この問題に対し，我々は，既存の学習アルゴリズムを機能単位であるメソッドに分割し，体系化したメソッドリポジトリを用いて，所与のデータセットに対する要求を満足する学習アルゴリズムを再構成できる枠組を提案してきた [阿部 02]．この構成型メタ学習では，メソッドリポジトリを用いることによって，既存の学習アルゴリズムを含めたより多くの学習アルゴリズムを構成可能である．

本稿では，選択型メタ学習において，広く利用されている stacking との比較を通して，提案手法である構成型メタ学習の有意性を評価する．

連絡先: 阿部　秀尚，静岡大学大学大学院理工学研究科，〒 432-8011 浜松市城北 3-5-1 静岡大学情報学部 山口研究室，Tel:053-478-1473，Fax:053-473-6421，hidenao@ks.cs.inf.shizuoka.ac.jp

2. 構成型メタ学習:CAMLET

構成型メタ学習を実現するツールである CAMLET では，まず人手により従来の学習アルゴリズムを分析し，機能単位であるメソッドを同定する．次に，これらのメソッドを体系化し，データセットにあわせて学習アルゴリズムを再構成することを可能にする．メソッドを体系化したものを，メソッドリポジトリと呼ぶ．CAMLET では，メソッドリポジトリを用いて，実際に学習アルゴリズムを所与のデータセットに対して実行し，ユーザから入力された要求を満たす帰納アプリケーションを構築する．

2.1 帰納メソッドリポジトリ

帰納メソッドリポジトリを構築するため，我々は 8 種類の帰納学習アルゴリズムの分析を行った [阿部 02]．分析対象となったのは，ヴァージョン空間法・AQ15・Classifier Systems・ID3・C4.5・ニューラルネットワーク・Bagged/Boosted C4.5 である．これらのアルゴリズムから似たような機能部分 (メソッド) を 6 種類同定し，アルゴリズムを構成する 8 種類の制御構造を整理したものが，図 2 で示す帰納メソッドと制御構造ある．

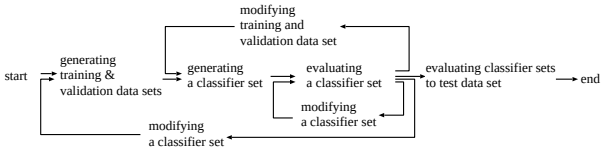

図 2: 帰納メソッドと制御構造

図 2 での 6 種類の帰納メソッドに対して，実装レベルと対応するよう各メソッドを体系化したものが，図 3 に示すメソッ

ドリポジトリである．階層を生成するため，各ノードには「入力」「出力」「参照」のデータ構造と前後のメソッドを記述する．

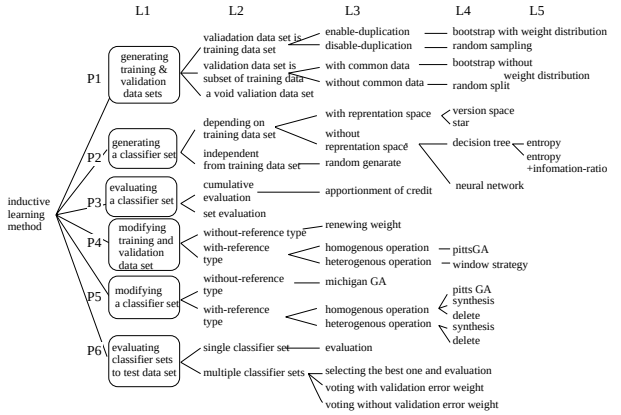

図 3: メソッドリポジトリ

図 4 は，上記の帰納メソッドで扱われるデータ構造の階層を示している．データ構造の階層は，メソッドリポジトリの階層化に利用される．

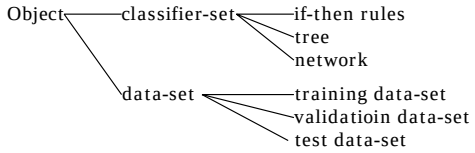

図 4: データ構造階層

2.2 CAMLET の基本動作

CAMLET の基本動作を図 5 に示す．CAMLET の入力は，訓練データセットとテストデータセット*1，要求 (現在はテストデータに対する正解率) であり，出力は要求を満たす帰納アプリケーションの仕様と所与のデータセットに対する実行結果である．

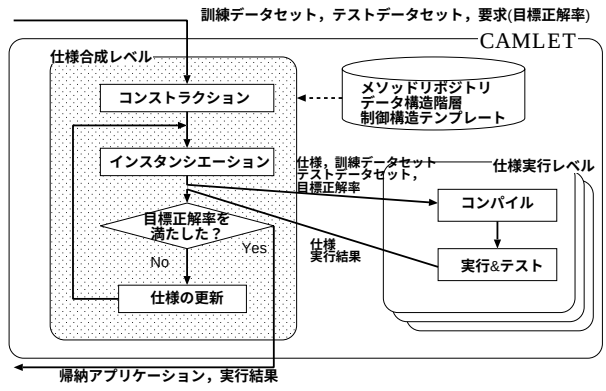

図 5: CAMLET の基本動作

*1 n 回交差検定が必要な場合，訓練データセットとテストデータセットの組が n 個必要

コンストラクションでは，帰納アプリケーションの基本となる仕様が決定される．最初に図 2 の 8 種類の制御構造から 1 つを選択する．次に，各メソッドについての具体的なメソッドをメソッドリポジトリの各ノードに与えられた宣言的仕様に基づいて選択する．**インスタンシエーション**では，帰納アプリケーションの仕様を実体化する．そのため，メソッド間で出力と入力のデータ構造が違う場合，変換可能なデータ構造については変換メソッドを挿入するなど仕様の調整を行う．**コンパイル**では，実体化された仕様から，実行コードを生成する．**実行&テスト**では，所与のデータ集合に対して帰納アプリケーションを実行し，評価に必要な正解率などを得る．実行された帰納アプリケーションが要求を満たさない場合，**仕様の更新**によって，新たな仕様が生成され，**インスタンシエーション**へと渡される．**仕様の更新**において，評価が改善するように仕様の更新を行うことは，すなわち仕様の洗練化となる．

3. 選択型メタ学習:Stacking

本稿では，選択型メタ学習アルゴリズムのうち，stacking についてデータマイニングツール WEKA[Witten 00] での実装 Stacking を利用する．この Stacking の学習過程を図 6 に，予測過程を図 7 に示す．学習過程の入力は，ユーザから与えられた訓練データセットであり，出力は基本レベル分類器とメタレベル分類器である．予測過程では，学習過程からの基本レベル分類器，メタレベル分類器を用いて，ユーザから入力されたテストデータセットを評価する．

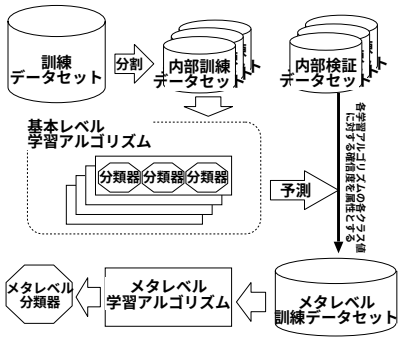

図 6: Stacking の学習過程

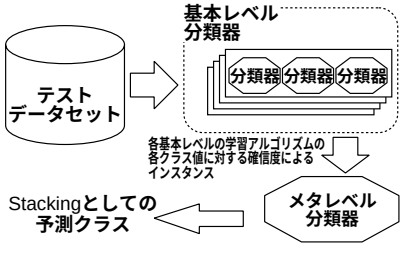

図 7: Stacking の予測過程

Stacking では，所与の訓練データセットを指定した数の交差検定用の訓練データセットと検証データセットに分割し，分割された訓練データセットを用いて基本レベルのアルゴリズムで分類器を学習する．次に，訓練データセットと組になった検証データセットを用いて，メタレベルの訓練データセットを作

成する．メタレベル訓練データセットの属性は，基本レベルの各学習アルゴリズムによる各クラス値への予測確信度である．メタレベルの訓練データセットは入力された検証データセットのインスタンス x に対して，次のように作成される．基本レベル分類器を生成する学習アルゴリズム L の数を m，クラス数 C を n，内部での交差検定の回数を F としたとき，各メタ属性の属性値 $Attibute$ は以下の式で与えられる．ここで，関数 $Prediction$ は，入力 x に対して，基本レベル分類器 L_{ik} がクラス値 C_j を予測する確信度を返す関数である．

$$Attibute(L_{i=1}^{m}, C_{j=1}^{n}, x) = \sum_{k=1}^{F} Prediction(L_{ik}, C_j, x)/F$$

以上のように生成されたメタレベルの訓練データセットに対して，なんらかの学習アルゴリズムを適用して，メタレベル分類器を学習する．メタレベル分類器は，メタ属性を用いて，基本レベルのクラスを予測する分類モデルとなる．

予測過程では，テストデータセットの各インスタンスに基本レベルの分類器を適用し，上記のメタ属性をもつメタレベルのインスタンスに変換する．このメタレベルのインスタンスに対し，メタレベルの分類器を適用してクラスを予測させ，成否を判断する．

4. StatLog データセットによる比較

本実験では，StatLog プロジェクト [Michie 94] で用いられ，公開されている 10 データセット[*2]を利用する．データセットの内容は表 1 に示す．各データセットの検定法は，“dna”,“letter-liacc”,“satimage”,“shuttle” については，既に割り当てられた訓練・テストデータセットを用いる．それ以外のデータセットについては，データセットを分割し，10 回交差検定を行う[*3]．

表 1: 実験に使用する StatLog データセット (数値：数値属性数)

Dataset	クラス数	属性数 (数値)	インスタンス数
australian	2	14(8)	690
diabetes	2	8(8)	768
dna	3	180(0)	2,000/1,186
german	2	20(3)	1,000
heart	2	13(7)	270
letter-liacc	26	16(16)	16,000/4,000
satimage	6	36(36)	4,435/2,000
segment	7	18(18)	2,317
shuttle	7	9(9)	43,500/14,500
vehicle	4	18(18)	846

Stacking の基本レベル学習アルゴリズムとして，J4.8[*4]・IBk[*5]・Naive Bayes・Part ルール学習器 [Frank 98] ・Bagged J4.8 ・Boosted J4.8 の 6 種類の学習アルゴリズムを用いる．ただし，Bagging と Boosting については生成する決定木を 5 と 10 にしたものを含めるため，計 8 種類が基本レベル学習器となる．メタレベル学習アルゴリズムについては，J4.8・Naive Bayes・線形回帰モデルを利用した分類 (以降，Linear Regression と呼ぶ) の 3 つを用いる．

これに対し，CAMLET に与える要求は，最も平均正解率が優れていた Linear Regression を用いた Stacking の各データセットに対する正解率である．CAMLET の帰納アプリケーション探索終了条件は，要求を満たすか，100 個の帰納アプリケーションの実行である．100 個の帰納アプリケーションを実行しても要求を満たせない場合は，それまでで最高正解率のものを採用する．探索手法としては，[阿部 02] で用いられた GA に基づく探索を採用した．

テストデータセットに対する正解率の結果を表 2 に示す．

表 2: メタ学習アルゴリズムによる正解率 (%) の比較 (NB:Naive Bayes, LR:Linear Regression)

Dataset	J4.8	NB	LR	CAMLET
australian	85.1	85.9	87.1	87.1
diabetes	73.3	74.9	76.2	76.2
dna	93.9	95.3	94.9	95.2
german	72.6	73.1	74.4	74.4
heart	80.7	82.2	83.0	83.3
letter-liacc	95.0	93.6	96.6	94.6
satimage	89.3	88.3	91.1	90.6
segment	97.8	97.6	98.1	97.7
shuttle	99.96	99.94	99.99	99.99
vehicle	72.0	75.2	74.7	75.2
Average	86.0	86.6	87.6	87.4

表 2 が示すように，CAMLET によって合成された帰納アプリケーションは，10 データセット中 7 データセットで Linear Regression を用いた Stacking の正解率を満たすことができた．要求した正解率に満たなかった 3 データについても，全体が有意差になるほどの差ではない．このことは，CAMLET が基本学習アルゴリズムに対して，Stacking と同等の分類性能向上能力を有することを示している．また，CAMLET により合成された帰納アプリケーションの仕様には，Bagging や Boosting に由来する制御構造やメソッドが，全ての帰納アプリケーションに見られた．これは，CAMLET が正解率のみを基準としているため，より多数の分類器を生成してテストデータセットを評価する帰納アプリケーションへと探索が偏ったためだと考えられる．

5. 関連研究

本稿の実験で用いた WEKA は，MLC++[Kohavi 96] のようにデータマイニングアプリケーションを多数集めたツールである．WEKA では，各アルゴリズムとアルゴリズム内のパラメータに対して説明が付与され，これらは GUI を通して提供される．しかし，これらのツールでは，ユーザが入力したデータセットに対し最適なアルゴリズムの選択を支援する機能は提供されない．

アルゴリズム選択の支援を目指す研究としては，メタルールによる最適なマイニングアルゴリズムの選択 [Brazdil 94][Gama 95] や，マイニングアルゴリズムの順位付けをする手法 [Brazdil 00] がある．なかでも，Metal プロジェクト [METAL] により開発された手法は，入力データセットの特徴などから，過去の実績による評価値を算出し，マイニング

∗2 http://www.liacc.up.pt/ML/statlog/

∗3 データセットの分割には，WEKA の weka.filters.SplitDatasetFilter を用いた．パラメータであるランダムの種はデータセット毎に同一とした．

∗4 C4.5 の Java 実装版

∗5 k-Nearest Neighbours

アルゴリズムの順位付けを行う方法である．さらに，データの前処理まで含めたデータマイニングプロセスを対象としたアプローチとして，IDEA[Bernstein 01] があげられる．IDEA は，前処理段階でフィルターの組合せを行えるものの，マイニングアルゴリズム自体を分解し，データセットにあわせて再合成することはできない．

順位付けとは異なるマイニングプロセスの選定支援方法論として，Engels らによる user-guidance model[Engels 96] の一連の研究がある．このモデルでは，システムが自動的にマイニングプロセスを選定する方法論が示されており，今後，構成型メタ学習にも採り入れて行く必要があると考える．

stacking に関しては，基本レベル学習アルゴリズムとしてどのようなアルゴリズムを用意すれば良いのか，メタ属性はどのように用意すれば良いのか，メタレベル学習アルゴリズムは何を適用すれば良いのか，の 3 点が主な研究課題となっている．[Todorovski 00] では，基本レベル学習アルゴリズム毎の最大クラス確信度・クラス予測のエントロピー・訓練データでの該当クラス比をメタ属性として，メタレベルの学習によって基本レベル学習アルゴリズムの動的選択を行う分類器を得る Meta Decision Trees という手法が提案されている．

6. おわりに

本稿では，構成型メタ学習を行う CAMLET と stacking の比較を行った．StatLog 共通データセットを用いた正解率比較では，構成型メタ学習が従来のメタ学習である Stacking(WEKA での実装) と比較して同等の性能であることを示した．

今後，選択型メタ学習への CAMLET の適用として，基本レベル学習アルゴリズムとメタレベル学習アルゴリズムを CAMLET により構築することが考えられる．これは，stacking の課題のうち，基本レベル学習アルゴリズムの選定とメタレベル学習アルゴリズムの選定を CAMLET により支援する，ということである．本来は，メタ属性についても支援すべきであるが，現在の CAMLET にはデータの前処理に関するメソッドのリポジトリが整備されていない．これを実現するため，データの前処理まで含めたメソッドリポジトリの拡充を行って行きたい．

さらに，現在ではデータマイニングを行う際の評価基準として，正解率以外の基準が重視されることも多い．構成型メタ学習は，正解率以外にも計算コストや興味深さの指標など，様々な評価基準に，柔軟に対応できる枠組であると考えられるため，多くの評価基準を構成型メタ学習で扱えるよう，リポジトリを整備していく予定である．

参考文献

[阿部 02] 阿部 秀尚，山口 高平：“メソッドリポジトリに基づく帰納アプリケーションの並列合成とその洗練化”，人工知能学会誌 Vol.17, No.5, pp.647–657 (2002).

[Breiman 96] Breiman, L.: “Bagging Predictors”, *Machine Learning*, 24(2), pp.123–140 (1996).

[Bernstein 01] Bernstein, A., and Provost, F.: “An Intelligent Assistant for Knowledge Discovery Process”, IJCAI 2001 Workshop on Wrappers for Performance Enhancement in KDD, (2001).

[Brazdil 94] Brazdil, P., Gama, J., and Henery, B.: “Characterizin the Applicability of Classification Algorithms Using Meta-Level Learning”,in *Proc. of the European Conferece on Machine Learning (ECML-94)*, pp. 83–102 (1994).

[Brazdil 00] Brazdil, P., and Soares, C.: “A Comparison of Ranking Methods for Classification Algorithm Selection”, in *Proc. 11th European Conference on Machine Learning (ECML-2000)*, pp. 63–74 (2000).

[Engels 96] Engels, R. : “Planning in Knowledge Discovery in Databases; Performing Task-Oriented User-Guidance”, Insitute for AIFB (1996).

[Frank 98] Frank, E., and Witten, I. H.: “Generating Accurate Rule Sets without Grobal Optimization”, in Proc. of Fifteenth International Conference on Machine Learning, pp. 144–151 (1998).

[Freund 96] Freund, Y., and Schapire, R. E.: “Experiments with a new boosting algorithm”,in *Proc. of Thirteenth International Conference on Machine Learning*, pp. 148–156 (1996).

[Gama 95] Gama, J., Brazdil, P.: “Characterization of Classification Algorithms”, 7th Portuguese Conference on Artificial Intelligence, EPIA ’95, pp. 189–200 (1995).

[Gama 00] Gama, J. and Brazdil, P.: “Cascade Generalization”, Machine Learning, 41(3),Kluwer Academic Publishers, Boston, pp.315–343, (2000).

[Kohavi 96] Kohavi, R., and Sommerfield, D.: “Data Mining using MLC++ — A Machine Learning Library in C++”, in *Proc. 8th International Conference on Tools with Artificial Intelligence*, pp. 234–245 (1996).

[METAL] http://www.metal-kdd.org/

[Michie 94] Michie,D., Spergelhalter,D., Taylor,C. (eds.): “Machine Learning, Neural and Statistical Classification ”, Ellis Horwood, (1994).

[Todorovski 00] Todorovski, L., and Dzerovski, J. : “Comnibing Multiple Models with Meta Decision Trees”, in *Proceedings of Principles of Data Mining and Knowledge Discovery*, pp. 54–64 (2000).

[Witten 00] Witten, I., and Frank, E. :“Data Mining: Practical machine learning tools and techniques with Java implementations”, Morgan Kaufmann Publishers (2000).

[Wolpert 92] Wolpert, D. :“Stacked Generalization”, Neural Network 5(2), pp.241-260 (1992).