メタ特徴の最適化処理による識別器構築アルゴリズム自動選択システム

南保 英孝（金沢大学 大学院自然科学研究科，nambo@ec.t.kanazawa-u.ac.jp）
大塚 敦史（金沢大学 大学院自然科学研究科，koko-peri@blitz.ec.t.kanazawa-u.ac.jp）
木村 春彦（金沢大学 大学院自然科学研究科，kimura@ec.t.kanazawa-u.ac.jp）
上田 芳弘（石川県工業試験場，ueda@irii.go.jp）

Automatic classifier algorithm selection using optimized meta-features

Hidetaka Nambo (Graduate School of Natural Science and Technology, Kanazawa University, Japan)
Atsushi Otsuka (Graduate School of Natural Science and Technology, Kanazawa University, Japan)
Haruhiko Kimura (Graduate School of Natural Science and Technology, Kanazawa University, Japan)
Yoshihiro Ueda (Industrial Research Institute of Ishikawa, Japan)

要約
ビッグデータ社会の到来とともに、現実世界の事例を基にモデルを作成し、作成したモデルに基づいて未知データの特性を予測するための識別器が広く応用されるようになってきた。しかし、識別器を構築するためのアルゴリズムは数多く存在するため、ビッグデータ解析に不慣れなユーザにとって、適切なアルゴリズムを選択することは非常に難しい。なぜならば、扱うデータの性質や目的によって、適したアルゴリズムが異なり、またデータの性質も多岐にわたるためである。そのため、精度の高い識別器を構築するために多くの時間と手間がかかってしまうと言う問題が生じる。そこで本論文では、この問題を解決するため、適切な識別器構築アルゴリズムを自動的に選択するシステムを提案する。提案システムでは、解析対象のデータの特徴を表すメタ特徴を導入し、各種データセットのメタ特徴と最適なアルゴリズムの組み合わせから学習を行うことによって、最適アルゴリズムを予測するモデルを構築する。

キーワード
データマイニング，機械学習，識別器構築アルゴリズム選択，メタ特徴，特徴選択

1. はじめに

近年、急速な情報化社会の発展に伴い、様々なセンサ技術からは多様なデータが取得できるようになり、ストレージには多くのデータを蓄積できるようになってきた。このようなビッグデータ社会の到来とともに今まで以上にデータマイニングに注目が集まるようになってきた。特にビッグデータをもとに学習し予測を行う機械学習がよく利用されている。機械学習において、正解がわかっている教師ありデータ（学習データ）をもとにモデルを作成し、作成したモデルを用いて未知データに関する予測を行うものを識別器と呼ぶ。識別器はビッグデータ社会の到来とともに、多くの場面で利用されるようになってきている。例えば、センサから取得したデータを基に機械の故障予測、独居老人の安否確認、医師の診断データを基に病気の予測など様々な場面で用いられているなど、識別器が用いられる場面や扱うデータは多種多様である。また、扱うデータは大量であり、予測精度を向上させる反面、学習に要する時間の増加といった問題も引き起こしている。さらに、現在データマイニング用の識別器は数多く存在しており、それぞれのデータに対して最適な識別器が異なるため、データマイニングの知識がなければ、どの識別器を使えばよいかわからないといった問題がある。このような問題に対して、データマイニングの知識があるならば、サポートベクタマシン（SVM）やニューラルネットワーク（NN）のような一般的に使われている識別器構築アルゴリズムを使うであろう。しかし、そのようにして構築された識別器は、様々なデータセットにおいて他の識別器より比較的識別精度は良い。しかし、すべての問題に万能なアルゴリズムは存在せず、ある問題にのみ特化したアルゴリズムが存在することがノーフリーランチ定理（David, 1996）により示されている。つまり、SVMやNNのような識別器構築アルゴリズムは一般的に精度が良い識別器を生成できると言われているが、必ずしも全てのデータに対して最適であるというわけではないということである。一方、数多くのアルゴリズムを全て試し、解析データに最適な識別器を構築する事は時間がかかるうえ、大変困難である。

既存研究（Nakamura et al., 2014）では、与えられたデータセットに対して54種類のメタ特徴から有効な26種類のメタ特徴を用いて、5つの識別器構築アルゴリズムから与えられたデータセットに最適なアルゴリズムを選択するシステムが提案されている。しかし、特徴選択手法を一度しか適用しておらず、特徴選択の際に有効な特徴を削除している可能性がある。そこで本研究では、特徴選択手法を複数回適用することによってより詳細な特徴選択を行い、既存研究より少ないメタ特徴数で解析データに最適な識別器構築アルゴリズムを選択する手法を提案する。

2. 既存研究

提案システムでは、メタ特徴が識別精度に関係があるという仮定のもと、識別器構築アルゴリズムの選択を行う。

メタ特徴とは、データセットの内容ではなくデータセット自体が持つ特徴、つまりインスタンス数や特徴（属性）数、クラス数などを指す。これまでも、様々な文献でメタ特徴が数

多く提案されている（Hilan and Alexandros, 2001; Yonghong et al., 2002; Pavel et al., 2003; Sarah et al., 2010）。

本論文では既存研究と同じメタ特徴を使用する。

2.1 提案システム概要

提案システムの処理手順は以下の通りである。

1. 前処理として学習用データセットよりメタ特徴を抽出し、さらに各データセットに最適な識別器構築アルゴリズムを求めておく。このデータを学習データとし、メタ特徴から最適な識別器構築アルゴリズムを求める識別器を生成する。
2. システム利用者が入力した未知のデータセットよりメタ特徴を抽出し、インスタンスを生成する。
3. 抽出されたインスタンスを1.の識別器に適用する。
4. 識別器によって最適な識別器構築アルゴリズムを選択し、出力する。

一連の手順を図1に示す.

2.2 学習データ作成方法

前処理として、提案システムに必要な学習データの作成を行う。学習データ作成方法の概略図を図2に示す。まず、教師ありデータセットを数多く用意する。それらのデータセットに対してメタ特徴の抽出を行う。同時に、各データセットに対して29種類の識別器構築アルゴリズムを適用し、それぞれに最適なアルゴリズムを求める。抽出したメタ特徴を属性とし、求めた最適アルゴリズムをclassとした1つのインスタンスを作成する。以上の作業をすべての教師ありデータセットに対して行うことで学習データの作成を行う。本研究で使用する識別器構築アルゴリズムはデータマイニングツールWeka（Mark et al., 2009）で使用できるものに限定した。用い

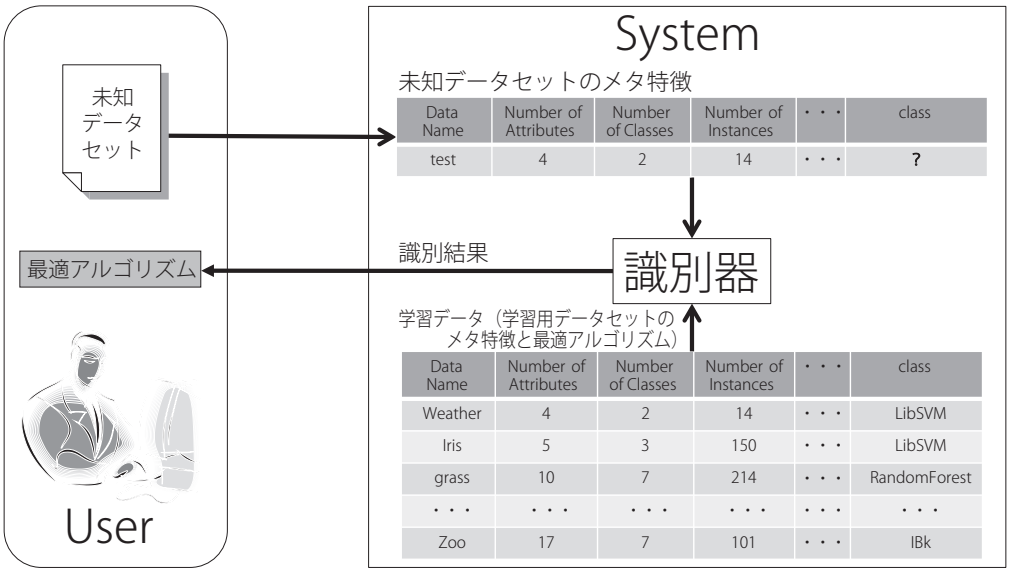

図1：提案システム処理手順概要

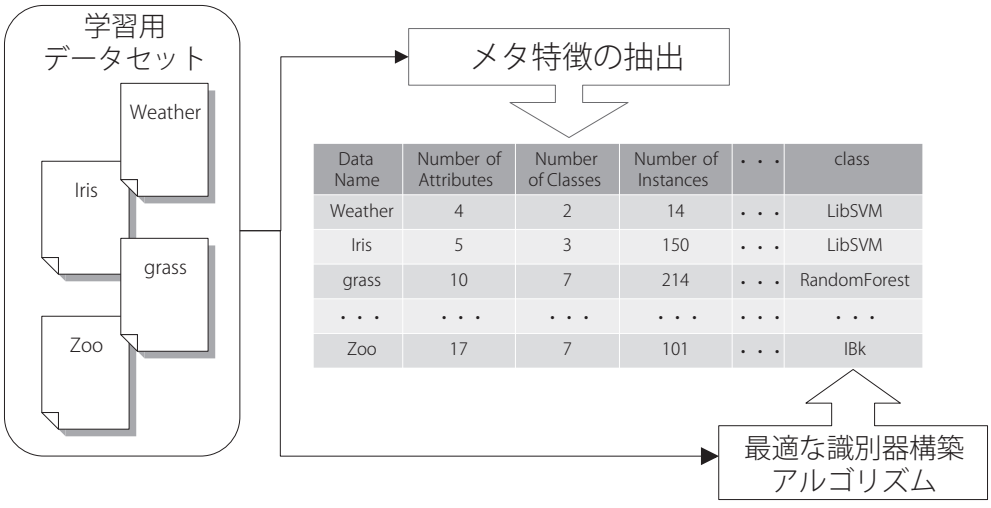

図2：学習データの生成方法

表1：用いた識別器構築アルゴリズム

カテゴリ	識別器名
Function	MultilayerPerceptron
	LibSVM
	SimpleLogistic
	SMO
Lazy	IBk
	Kstar
	LWQ
Rules	ConjunctiveRule
	DecisionTable
	JRip
	NNge
	OneR
	PART
	ZeroR
Bayes	Bayes Net
	NaiveBayes
	NaiveBayesUpdateable
Misc	HyperPipes
	VFI
Trees	DecisionStump
	Ft
	J48
	J48graft
	LDATree
	LMT
	NBTree
	RandomForest
	RandomTree
	REPTree

た29種類のアルゴリズムを表1に示す。

2.3 特徴選択手法

抽出したメタ特徴の中から有効な特徴のみを選択するためにWekaで使用可能な特徴選択手法を適用する。特徴選択手法としては大きく分けてフィルターアプローチとラッパーアプローチの二種類が存在しているが、今回はより正確な特徴選択が可能であるラッパーアプローチを用いた。ラッパーアプローチのおおまかな適用の流れを以下に示す。

1. 学習データの部分特徴集合をサンプリングし、選択された特徴を用いて識別器を構築し、分類精度を求める。
2. すべての部分集合に対して、1.を繰り返す
3. 分類精度（あるいはその期待値）が最大となるときの特徴の集合を選択する。

2.4 評価方法

学習データの評価にはIBk（David et al., 1991）という識別器を用いて交差検定であるLeave-one-out cross-validationを行う。Leave-one-out cross-validationとは学習データの1つのインスタンスをテストデータ、残りのインスタンスを学習データとし、学習データの精度を求める検証方法である。最適な識別器を求めるための交差検定の評価方法はF値を用いる。F値とは識別器の評価方法によく用いられており、再現率と適合率の調和平均により求められ、以下の式（1）で示される。また、*Pr*は適合率であり式（2）で、*Re*は再現率であり式(3)で与えられる。

$$\text{F値} = \frac{2 \cdot Pr \cdot Re}{Pr + Re} \tag{1}$$

$$Pr = \frac{tp}{tp + fp} \tag{2}$$

$$Re = \frac{tp}{tp + fn} \tag{3}$$

tp : 正解クラスに正しく識別されたインスタンス数
fp : 正解クラスに誤って識別されたインスタンス数
fn : 正解クラスに識別されなかったインスタンスのうち正解クラスのインスタンス数

3. 実験

本提案システムでは、識別器構築アルゴリズムを提案するための基盤となる学習データの作成方法が重要となる。よって、本実験では、精度向上につながる学習データの作成を目標とする。使用するデータマイニングツールWekaはバージョン3.6.10を使用した。

3.1 学習データ

本実験で使用する学習データは既存研究と同様のものである。インスタンスは、UCIrvine Machine Learning Repository（Bache and Lichman, 2013）より取得した58種類のデータセット、属性は58種類のデータセットからRapidMinerというソフトを利用し取得した54種類のメタ特徴を用いた。クラスは、集めた学習データに対して29種類のアルゴリズムで識別器を構築し精度を測定した結果、最適であった回数が多い上位5つの識別器構築アルゴリズムとなったMultilayerPerceptron、RandomForest（Leo, 2001）、LMT (Niels et al., 2005)、LADTree（Geoffrey et al., 2002）、FT（Joao, 2004）を用いた。

3.2 特徴選択手法適用方法

作成した学習データに対して既存研究を基に特徴選択手法を複数回適用し、有効な特徴のみを抽出する。特徴選択手法の属性検証方法としてラッパー法アプローチのWrapperSubsetEval (Ron and George, 1997)を用いた。その際、パラメータはclassifier = IBk、fold = 5、threshold = 0.01とし、検索方法は遺伝的アルゴリズムのGenetic Searchを用いた。Generic SearchのパラメータはcrossoverProb = 0.6、maxGenerations = 20、mutationProb = 0.033、populationSize = 20、reportFrequency = 20とした。

Wekaによって特徴選択手法を適用することにより、クロ

```
=== Attribute selection 150 fold cross-validation
(stratified), seed: 1 ===

number of folds (%)  attribute
          18( 12 %)  1 sepallength
          13(  9 %)  2 sepalwidth
         133( 89 %)  3 petallength
         150(100 %)  4 petalwidth
```

図3：特徴選択出力結果の例

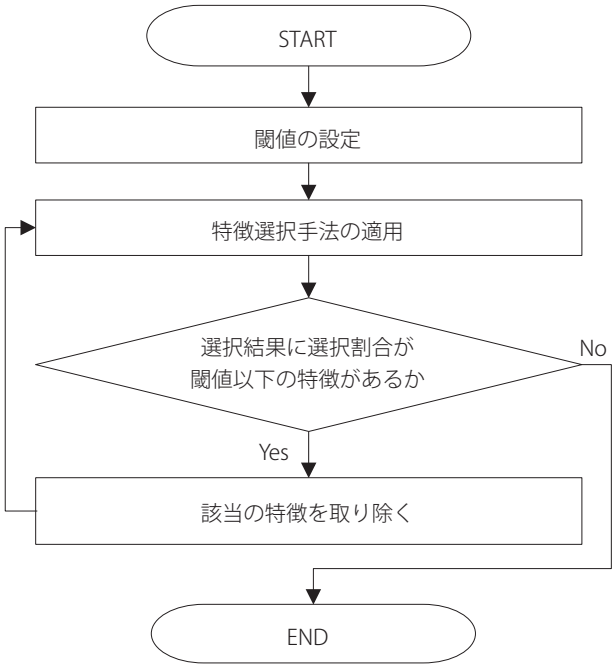

図4：特徴選択手法の処理の流れ

スバリデーションを実行した際に、各特徴が選択された回数が出力される。Wekaによる特徴選択手法の出力例を図3に示す。本実験では特徴が選択された割合を基準とし、割合が事前に設定した閾値以下の特徴を削除していくことにより特徴選択を行った。

特徴選択手法の処理の流れを図4に示す。

3.3 実験内容

特徴選択を行い、選択された割合が事前に設定した閾値以下の特徴がなくなるまで特徴選択を繰り返す。閾値は、10 %、20 %、30 %、40 %、50 %を用いた。その後、学習データの選択された特徴を用いてIBkにより識別器を構築し、Leave-one-out cross-validationを行い精度の検証を行う。

3.4 実験結果

実験結果を表2に示す。実験結果より、特徴選択手法を複数回繰り返し、閾値30 %以下の特徴を削除することで最も識別率がよくなるという結果が得られた。閾値30 %以下の特徴を削除した際に残った特徴を以下に示す。また、閾値30 %以下の時の実験結果を表3に示す。実験結果より、特徴選択手法を複数回適用することによって試行回数5回目は例外とし、

表2：特徴選択結果

閾値(%)	メタ特徴数	精度(%)	F値(%)
0	54	44.8	43.4
10	47	46.6	44.8
20	19	62.1	61.5
30	5	65.5	65.8
40	14	62.1	61.7
50	7	62.1	61.7

表3：閾値30 %時の特徴選択試行回数と精度

試行回数	メタ特徴数	精度(%)	F値(%)
1	40	53.5	52.3
2	34	55.2	54.1
3	28	55.2	54.4
4	25	55.2	54.6
5	21	53.5	53.6
6	15	60.3	60.7
7	6	63.8	64.0
8	5	65.5	65.8

徐々に精度が向上する事がわかる。

- knn：k-nearest neighborによる識別精度
- max_entoropy：情報量の最大値
- numerical：数値特徴の数
- classes：クラスの数
- mean_level：決定木を作成した際の深さの平均

既存研究の学習データとの比較を表4に示す。表より使用するメタ特徴の数は減少し、精度、F値ともに改善していることから、本研究で提案した特徴選択手法が有効であることがわかる。また、閾値30 %以下の特徴を削除した時のそれぞれのクラスのクロス表を表5に示す。括弧内の数字は既存

表4：既存研究との結果の比較

	メタ特徴数	精度(%)	F値(%)
既存研究	26	60.3	60.6
提案手法	5	65.5	65.8

表5：閾値30 %時の分類結果

クラス	分類結果				
	LAD	FT	LMT	MP	RF
LAD	6(4)	0(0)	2(2)	1(1)	0(2)
FT	1(2)	7(7)	2(1)	0(0)	0(0)
LMT	4(1)	2(4)	9(9)	1(4)	1(2)
MP	1(2)	0(0)	1(0)	9(10)	1(0)
RF	1(3)	1(0)	0(1)	1(1)	7(5)

研究のクロス表の結果となっている。クロス表を確認すると、比較的LADTreeに識別されている割合が高いが、過学習の可能性は低いと考える。既存研究のクロス表と比べると、全体的に正しく識別されている割合が増えていることがわかる。

4. 考察

実験結果より、特徴選択により得られる信頼性が30％より大きい特徴に関してはIBkによる識別の際、精度に何らかの影響が与えられていることがわかる。また、信頼度30％以下の特徴に関しては比較的精度にはかかわらない特徴であることがわかる。5つの特徴のうちknn、max_entropy、numericalに関しては、他の閾値で特徴選択を行った場合でも残される確率が高かったことから、IBkの識別に有効な特徴ということがわかる。交差検定により間違えたインスタンスを確認したところ、予測したクラスと正解クラスの精度が同じ事例がいくつかあった（表6）。

表6：最適な識別器が複数存在する例

データセット名	LAD	FT	LMT	MP	RF
labor.arff	84.2	89.5	89.5	86.0	87.7
hypothyroid.arff	99.5	99.3	99.5	94.2	99.1
mushroom.arff	99.9	100.0	99.9	100.0	100.0

注：MP/MultiLayerPerceptron, LAD/LADTree, RF/RandomForest

このようなデータセットのインスタンスを正解としたところ、正解数は42/58となり、正解率は72.4%となった。
次にWebから取得した58種類のデータセットすべてに対して一般的に使用されている識別器構築アルゴリズムを用いて識別器を構築したと考え、各データセットに対する識別器の正解率の平均を求めた。つまり、よく用いられているアルゴリズムを盲目的に使用する場合と、提案システムを用いて識別器構築アルゴリズムを選択する場合の比較を行った。結果を表7に示す。

以上の結果より、一般的に使用されている識別器構築アルゴリズムを用いるよりも、提案システムにより選択されたアルゴリズムを用いた方が、適切なアルゴリズムを用いることが出来ることが分かる。このことから、提案システムの有効性が確認できる。

表7：一般的な識別器構築アルゴリズムを常用した場合との精度比較

分類器名	平均精度(%)
LibSVM	71.3
SMO	81.9
MultilayerPerceptron	82.5
NaiveBayes	79.6
IBk	80.9
J48	80.9
RandomForest	82.0
Selected classifier	85.4

5. まとめ

現在、ストレージやセンサ技術の発展とともに多様なデータが日々生成されている。これらのデータセットから有用な知見の応用のために機械学習に注目が集まっている。しかし、機械学習で使用される識別器は現在数多く存在しており、データセットによって最適な識別器構築アルゴリズムが異なるため最適なアルゴリズムを見つける事は現状では大変困難である。そこで、本研究ではシステム利用者が解析したいデータセットを入力することによって、データセットに対して5種類の識別器構築アルゴリズムの中から最適な亜リゴリズムを選択する識別器構築アルゴリズム自動選択システムの提案を行った。本実験では58種類のデータセットをもとに、5種類のメタ特徴を用いた学習データに対してIBk（k＝1）によって識別することにより正解率65.5 %、F値65.8 %の精度で識別できることを確認した。また、予測した識別器が正解識別器とは違うものの、同一精度だった場合を正解と判断した場合、正解率は72.4 %となった。

今後の展望としては、本来識別器構築アルゴリズムを用いる際には、アルゴリズムのパラメータにより大きく精度が変化するが、現在はパラメータはデフォルト値を用いている。今後はパラメータの最適化を考慮したアルゴリズムの選択システムを構築したい。そして、今回は出現回数の多かった5種類のアルゴリズムを対象とした選択システムを提案したが、今後は対象とする識別器構築アルゴリズムを追加する。また、現在はF値が高いアルゴリズムを有効なものとして定義しているが、正解率、再現率、適合率、平均絶対誤差や処理時間なども考慮し、総合的に有効なアルゴリズムを提案するシステムの作成を行う。

引用文献

Bache, K. and Lichman, M. (2013). UCI machine learning repository. Retrieved from http://archive.ics.uci.edu/ml/.

David, W. A., Dennis, K., and Marc, K, A. (1991). Instance-based learning algorithms. *Machine Learning*, Vol. 6, 37-66.

David, H. W. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computing*, Vol. 8, No. 7, 1341-1390.

Geoffrey, H., Bernhard, P., Richard, K., Eibe, F., and Mark, H. (2002). Multiclass alternating decision trees. *Proceedings of European Conference on Machine Learning*, 161-172.

Hilan, B. and Alexandros, K. (2001). Estimating the predictive accuracy of a classifier. *Lecture Notes in Computer Science*, Vol. 2167, 25-36.

Joao, G. (2004). Functional trees. *Machine Learning*, Vol. 55, Issue 3, 219-250.

Leo, B. (2001). Random forests. *Machine Learning*, Vol. 45, Issue 1, 5-32.

Mark, H., Eibe, F., Geoffrey, H., Bernhard, P., Peter, R., and Ian, H. W. (2009). The WEKA data miming software: An update. *SIGKDD Explorations*, Vol. 11, No. 1, 10-18.

Nakamura, M., Otsuka, A., and Kimura, H. (2014). Automatic selection of classification algorithms for non-experts using

meta-features. *China-USA Business Review*, Vol. 13, No. 3, 199-205.

Niels, L., Mark, H., and Eibe, F. (2005). Logistic model trees. *Machine Learning*, Vol. 59, Issue 1-2, 161-205.

Pavel, B. B., Carlos, S., and Joaquim, P. C. (2003). Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. *Machine Learning*, Vol. 50, No. 3, 251-277.

Ron, K. and George H. J. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, Vol. 97, 273-324.

Sarah, D. A., Faisal, S., Matthias, R., and Markus, G. (2010). Landmarking for meta-learning using RapidMiner. *Proceedings from RapidMiner Community Meeting and Conference*.

Yonghong, P., Peter, A. F., Carlos, S., and Pavel, B. (2002). Improved dataset characterisation for meta-learning. *Lecture in Notes in Computer Science*, Vol. 2534, 193-208.

（受稿：2016年7月4日　受理：2016年7月21日）