Vol.2014-AVM-87 No.14
2014/12/5

メタ特徴と特徴選択を用いた識別器自動選択システム

大塚敦史[†1]　　中村宗広[†2]　木村春彦[†1]

近年，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っている．しかし，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，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最適な識別器がわからないといった問題がある．そこで，本研究では与えられたデータセットに対して最適な識別器を推薦する識別器自動選択システムを提案する．

Automatic Selection of Classification Algorithms Using Meta-feature and Feature Selection

ATSUSHI OTSUKA[†1]　MUNEHIRO NAKAMURA[†2]
HARUHIKO KIMURA[†1]

With the arrival of big-data society, methods for classifying real-word problems have attracted much attention for researchers and developers in various fields. However, since a large variety of classification algorithms has been available, it is difficult for non-experts to find classification algorithms that achieve good results on a given data set. This paper presents a system of predicting the best classification algorithm for a given data set with respect to the accuracy.

1. はじめに

近年，急速な情報化社会の発展により大量で多様なデータが日々生成されている．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基に予測する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）を基にモデルを作成し未知のデータ（テストデータ）が入力されたときに作成したモデルを基に予測するものを識別器という．この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た．例えば，センサー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など様々な場面で用いられている．しかし，ビッグデータ社会の到来によるデータの増加は予測精度を向上させる反面，解析時間の増加といった問題も引き起こしている．さらには，データの多様化により，データに対して最適な識別器の多様化と言った問題も引き起こしている．現在データマイニング用の識別器は数多く存在している．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人ならば，サポートベクターマシンやニューラルネットワークのような識別器をまず使うだろう．しかし，そのような識別器は，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が，すべての問題に万能なアルゴリズムは存在しない，ある問題にのみ特価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[2]により示されている．つまり，サポートベクターマシンやニューラルネットワークのような識別器は一般的に精度が良いと言われているが必ずしも全てのデータに対して最も精度がよくなるというわけではない事が示されている．また，数多くの識別器を全て試し，解析データに最適な識別器を見つける事は時間がかかるうえ，大変困難である．

そこで本研究では，主にデータマイニング初学者を対象とした，数多くの識別器の中から解析データに有効な識別器を選択する「識別器自動選択システム」を提案する．

2. 提案システム

提案システムでは，メタ特徴が識別精度に関係があるのではないかという仮定のもとシステムの作成を行う．メタ特徴とは，データの特徴，つまりインスタンスの数や次元数などを特徴として扱うことであり，様々な文献で提案されたメタ特徴が数多く存在する．[4][9][10][11]

2.1 提案システム概要

本提案システムは以下の手順により動作する．

- システム利用者がデータを入力することにより，メタ特徴を抽出する．
- 抽出したメタ特徴を属性としたテストデータを作成する．

†1 金沢大学
Kanazawa University
†2 金沢工業大学
Kanazawa Institute of Technology

- 前処理として作成する学習データをもとに，テストデータを識別することにより得られる結果が入力されたデータに最適な識別器として選択される．

上記の流れを図 1 に示す．

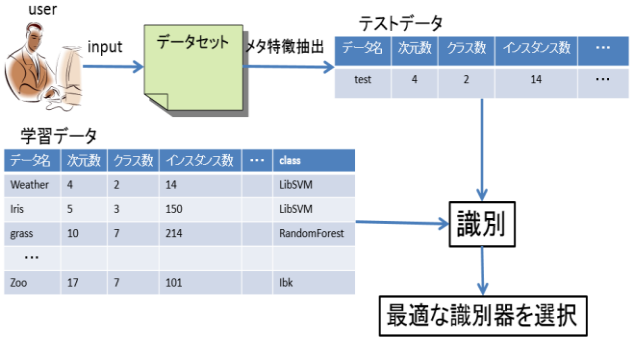

図 1　提案システム概要

Figure 1　Proposal system overview.

2.2 学習データ作成方法

提案システムに必要な学習データの作成を前処理として行う．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 Rapidmine というソフトを用いてメタ特徴の抽出を行う．同時に，複数の識別器を適用し，最適な識別器を求める．

上記の流れの概略図を以下の図 2 に示す．

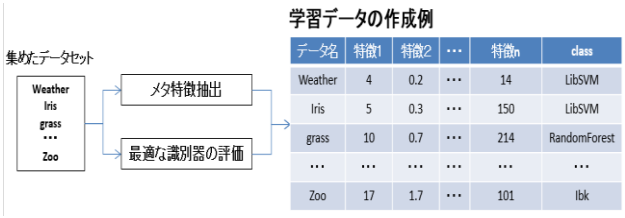

図 2　学習データ作成方法

Figure 2　How to create a leaning data.

2.3 評価方法

最適な識別器を求めるための評価方法は 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，再現率である Precision は式（2）適合率である Recall は式（3）に示す。tp は検索した結果の正解率，fp は検索結果の数，fn は正解の数を表している．

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad \cdots\cdots (1)$$

$$precision = \frac{tp}{tp + fp} \quad \cdots\cdots (2)$$

$$recall = \frac{tp}{tp + fn} \quad \cdots\cdots (3)$$

3. 実験

データマイニングツール Weka[7]に入っているベンチマークデータセットと UCIrvine Machine Learning Repository [1]のデータセットを合わせて 58 個のデータセットを取得した．システムは Eclipse をもとに構築する．Weka で使用可能な 30 種類の識別器を集めたデータセットに適用した．選択される識別器を決める際に使用した Weka 内の 30 種類の識別器を表 1 に示す．識別時のパラメーターはデフォルトで行った．

表 1　提案手法により実装された 30 種類の識別器

Table 1　30 classifiers in six categories implemented in the proposed system.

カテゴリ	識別器名
Function	MultilayerPerceptron
	LibSVM
	SimpleLogistic
	SMO
Lazy	IB1
	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

表 1 に示した識別器の中から選ばれた回数が多かった上位 5 つの識別器を求めたところ MultilayerPerceptron，RandomForest[6]，LMT[8]，LADTree[3]，FT[5]となった．

今回はこの 5 つの識別器を選択される識別器とする．これら 5 つの識別器を正解とし，58 個のデータセットから 54 種類のメタ特徴を特徴とした学習データを作成した．

3.1 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し，有効なメタ特徴のみを選択する．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存在しているが，今回はラッパーアプローチを用いた．ラッパーアプローチとは，サンプリングした学習データに対して，識別器を繰り返し適用し制度が最もよくなる部分集合を得る手法である．ラッパーアプローチとして WrapperSubsetEval と検索方法として遺伝的アルゴリズムの Genetic Search を用いて特徴選択手法を適用する．それぞれのパラメーターを表 2 と表 3 に示す．

表 2　Wrapper Subset Eval のパラメーター
Table 2　Parameter of Wrapper Subset Eval.

Classifier	IB1
Fold	5
Seed	1
Threshold	0.01

表 3　Genetic Search のパラメーター
Table 3　Parameter of Genetic Search.

crossobcerProb	0.6
maxGenerations	20
mutationProb	0.033
populationSize	20
reportFrequency	20
Seed	1

上記の特徴選択を用いることにより，IB1 によって識別する際の 54 種類のメタ特徴の信頼性を求める事ができる．求めた信頼性を基に特徴選択を行う．

3.2 実験内容

実験は以下の手順で行う．

- 特徴選択
- 閾値以下の特徴を削除
- 識別・精度検証

上記の 3 つの手順を閾値以下の特徴がなくなるまで行うことにより最も精度の良くなる特徴を残す．今回は 10%，20%，30%，40%，50%を閾値として特徴選択，削除を行った．精度の検証方法としては，学習データの 1 つをテストデータ，残りを学習データとして精度を求める Leave-one-out cross-validation を用いる．

3.3 実験結果

実験結果を表 4 に示す．実験結果より，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．つまり，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 IB1 の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．閾値 30%以下の特徴を削除した場合の残された特徴を表 5 に示す．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．

表 4　実験結果
Table 4　Result of the evaluation experiment.

閾値[%]	属性数	精度[%]	F 値[%]
0	54	44.8	43.4
10	47	46.55	44.8
20	19	62.07	61.5
30	5	65.52	65.8
40	14	62.07	61.7
50	7	62.07	61.7

表 5　閾値 30%の時の特徴選択結果
Table 5　Result of attribute selection when threshold 30%.

信頼性[%]	特徴名
100	knn
100	max_entropy
100	numerical
100	classes
98	mean_level

4. まとめ

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，解析したいデータに対して最適な識別器を見つけることは大変困難で時間がかかるといった問題がある．そこで本研究では，解析データを入力することにより最適な識別器を選択する識別器自動選択システムの提案を行った．今回は現実に存在する 58 個のデータセットのメタ特徴をもとに学習データを作成し，作成した学習データを基に識別を行うことにより，解析データに有効な識別器を選択するシステムの精度の検証を行った．属性選択手法を用いて 5 種類のメタ特徴を用いた学習データを使い IB1 によって識別を行うことにより 65.5%の精度で予測可能という実験結果が得られた．今後の展望としては，本来識別器を使用す

る際識別器のパラメーターの最適化により大きく精度が変化するが，現在パラメーターをデフォルトで行っているため，今後はパラメーターの最適化を考慮した識別器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識別器 5 種類を対象とした選択システムの提案を行ったが，今後は選択識別器の増加を行う．

参考文献

1) Blake C. L., & Merz, C. J. (1998), UCI Repository of Machine Learning Databeses, Retrieved from http://archive.ics.uci.edu/ml/
2) David, H. W. (1996). The lack of a priori distinctions between learning algorithms. Neural Computing, 8(7), 1341-1390.
3) Geoffrey, H., Bernhard, P., Richard, K., Eibe, F., & Mark, H. (2001). Multiclass alternating decision trees. Proceedings from ECML'02: The 13th European Conference on Machine Learning (pp. 161-172). London, UK.
4) Hilan, B., & Alexandros, K. (2001). Estimating the predictive accuracy of a classifier. Lecture Notes in Computer Science, 2167, 25-36
5) Joao G. (2004) , Functional Trees, Machine Learning, 55(3), 219-250
6) Leo B. (2001) , Random Forests, Machine Learning, 45(1), 5-32.
7) Mark, H., Eibe, F., Geoffrey, H., Bernhard, P., Peter, R., & Ian, H. W. (2009). The WEKA data miming software: An update. SIGKDD Explorations, 11(1), 10-18.
8) Niels L., & Mark H. (2005) , Logistic Model Trees, Machine Learning, 95(1-2), 161-205.
9) Pavel, B. B., Carlos, S., & Joaquim, P. C. (2003). Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. Machine Learning, 50(3), 251-277.
10) Sarah, D. A., Faisal, S., Matthias, R., & Markus, G. (2010). Landmarking for meta-learning using RapidMiner. Proceedings from RapidMiner Community Meeting and Conference (RCOMM-10). Dortmund, Germany.
11) Yonghong, P., Peter, A. F., Carlos, S., & Pavel, B. (2002). Improved dataset characterisation for meta-learning. Lecture in Notes in Computer Science, 2534, 193-208.