知 能 と 複 雑 系 128−2
(2 0 0 2 . 5 . 2 3)

シーズ属性の逐次拡張に基づく属性選択

小森 麻央, 阿部 秀尚, 山口 高平

静岡大学 情報学部

本稿では，ラッパーメソッドの問題点である計算コストを改善する形で，シーズメソッドと呼ぶ新しい属性選択法を提案する．シーズメソッドはまず，所与の全属性集合でマイニングし，得られたルールセットにおいて出現率の高い属性の集合をシーズ属性集合として選択する．この初期属性集合を出発点とし，正解率が低下するまで属性集合を逐次的に拡張していく．今回はシーズメソッドを評価するために，ポピュラーな属性選択メソッドであるフィルターメソッドとラッパーメソッドとの比較実験を UCI ML リポジトリを用いて行った．今回のケーススタディでは，最高の正解率と 2 番目の計算コストの低さを達成できた．

Dynamic Incremental Extension of Seed Features in Data Pre-processing of KDD

Mao Komori, Hidenao Abe, and Takahira Yamaguchi

Faculty of Information, Shizuoka University

This paper presents a new feature selection method in data pre-processing of ML/DM (Machine Learning and Data Mining). Our new method starts with a set of seed features that comes up frequently in the results of ML/DM with all the features from a given data set . They can be extended with other features while the incremental extension process can keep higher accuracy of ML/DM than that of the immediate data mining. We have When the accuracy does not become higher, the above repetition stops and the feature subset at that time is taken as the input dataset of ML/DM schemes. compared our new feature selection method with the following popular feature selection methods: filter method and wrapper method. We have done a case study, using fourteen datasets from UCI ML Repository. The case study has shown us that our method gets to or a beyond the level of two popular feature selection methods.

1. はじめに

ML/DM(機械学習 & データマイニング) がより複雑なタスクを取り扱うにつれて，属性選択がますます重要になってきている．ML/DM の実タスクでは，データセットの属性数はかなり多いが，そのほとんどが無関係か冗長である．従って，所与の属性から有用な属性のみに絞り込む必要がある．これは属性選択 (feature selection) と呼ばれ，問題領域の意味を考えて実行すればよいが，専門家を必要とするためコストが大きくなり，その自動化が望まれている．

本稿では，ポピュラーな属性選択法であるフィルタメソッドとラッパーメソッド [1],[2] を紹介した後，ラッパーメソッドの問題点である計算コストを改善する形で，シーズメソッドと呼ぶ新しい属性選択法 [3] を提案する．さらに，本手法を評価するために UCI ML リポジトリ [4] の 14 種類のデータセットを用いて比較実験を行い，正解率，計算コスト，選択された属性数の観点から比較・評価する．

2. 属性選択メソッド

属性の数を N とすると属性パターン (部分集合) の数は 2^N となる．N が大きくなると組合せ数は莫大なものとなり，不要な属性をいかに効率よくふるい落せるかが問題となる．

優れた属性集合を選択する場合，ポピュラーな 2 つの手法が存在する．1 つはデータの一般的特徴に基づいて，スキームに対して独立した評価を行うもので，ML/DM を開始する前に最も見込みのある部分集合を求めるために属性集合をフィルタリングするため，フィルタメソッドと呼ばれる．もう一方は ML/DM スキームを用いて属性集合を評価するもので，ML/DM スキームが属性選択の手続きの中に含まれているため，ラッパーメソッドと呼ばれる．[5],[6]

2.1 フィルタメソッド [1],[2]

フィルタメソッドは，ML/DM スキームとは独立して属性選択を実行する．フィルタメソッドの代表的手法である Relief-F は，各属性に関連性の重みを割り当て，訓練集合から無作為にデータを抽出して，クラスが同じ，あるいは異なる近隣のレコード (ニヤヒットとニヤミス) を確認する．この時，ニヤヒットにおいて異なる値を取る属性は関連無しとして重みを減らし，ニヤミスで異なる値をとる属性は関連有りとして重みを増やす．この操作を繰り返したあと，正の重みを持つ属性のみを選択する．

2.2 ラッパーメソッド [1],[2]

ラッパーメソッドでは，属性部分集合選択がブラックボックスとして ML/DM スキームを使用する．属性部分集合選択スキームは，ML/DM スキーム自身を評価関数の一部として使用し，よりよい属性部分集合の探索を導く．探索は，状態空間，初期状態，終了条件および探索アルゴリズムにより決定される．探索には属性の空集合で始まる前向き選択と，全属性集合から始まる後向き除去とがある，

2.3 シーズメソッド

ML/DM スキームを実行させながら属性を選択するので，ラッパーメソッドの方がフィルターメソッドよりも精度は高い

$Relief(S)$;
全ての重みを 0 に初期化
For j= 1 to No_of_Sample
ランダムにデータを 1 つ選択 ;
ニヤヒット (hit) とニヤミス ($miss$) を検索 ;
For 全ての属性 f_i
$W_i = W_i - (x_ij, hit_ji)^2 + \delta(x_ij, miss_ji)^2$
end_for
end_for
$W_i := W_i \setminus No\_of\_Sample$
For 全ての属性 f_i
if $W_i > 0$ $then$ $S_0 := S_0 \cup f_i$
end_for
S_0 を出力

図 1: Relief のアルゴリズム

が，計算コストはあまりにも大きくなり，非現実的になっている．そこで，我々はラッパーメソッドの問題点である計算コストを改善する形で，シーズメソッドと呼ぶ新しい属性選択法を提案する．

図 2 全体は 4 属性における探索空間であり，ラッパーメソッドの最大探索空間になるが，シーズメソッドは探索開始点としての初期属性集合を適切に見つけることで，この探索空間を縮小することが特色である．

以下，図 3 にシーズメソッドの詳細について説明する．まず，所与の全属性集合でマイニングし，得られたルールセットにおいて出現率の高い属性の集合をシーズ属性集合とする．属性数はデータセットにより異なるが，今回は全属性数の 2〜3 割を選択した．この初期属性集合を出発点とし，属性集合を拡張していく．すなわち，逐次的に非シーズ属性集合から属性を初期属性集合に付加し, 各属性数において最良の ML/DM 性能を有するデータセットを選択し，データセットの性能が劣化するまで属性の追加を続ける．正解率が向上しなくなった時点でこの繰り返しを終了し，最後の属性集合を ML/DM スキームの入力として与える．

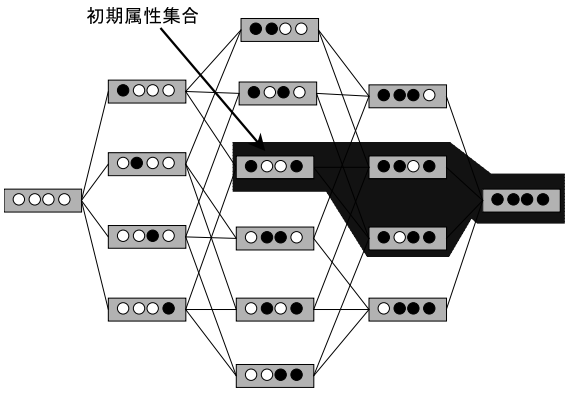

図 2: ラッパーメソッドとシーズメソッドの探索空間

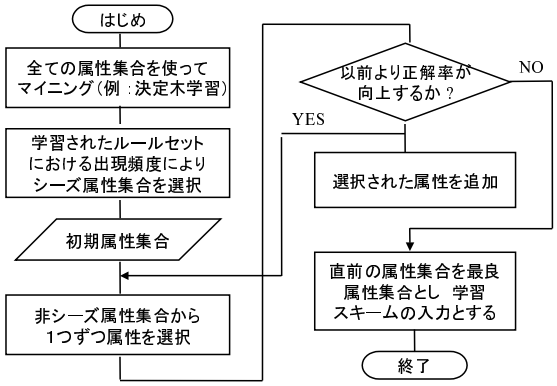

図 3: シーズメソッドの概要

表 1: データセットの概要

No	データ名	属性数	クラス数	訓練データ	テストデータ
1	breast	10	2	699	CV
2	crx	15	2	690	CV
3	Hayes-Roth	5	3	132	28
4	labor-neg	16	2	40	17
5	pima	8	2	768	CV
6	sick	29	2	2800	972
7	audiology	69	24	200	26
8	chess	36	2	3198	CV
9	glass	10	7	214	CV
10	lung-cancer	56	3	33	CV
11	wine	13	3	178	CV
12	monk 1	6	2	124	432
13	monk 2	6	2	124	432
14	monk 3	6	2	124	432

3. ケーススタディ

本節では，シーズメソッドとフィルタメソッドとラッパーメソッドを比較して，その有用性について検討する．比較実験を行うために，UCI ML リポジトリの 14 種類のデータセットを利用し，分類器としては C4.5 を用いた．表 1 はデータセットの概要であり，属性数，クラス数，訓練集合，テスト集合のサイズを示している (CV は 10 フォールドクロスバリデーションを意味している)．

今回は各属性選択メソッドの評価基準として，正解率，計算時間，選択された属性数を用いた．実験にはワイカト大学で開発された DM ツール，Weka(Waikato Environment for Knowledge Analysis)[6] を使用した．フィルタメソッドについては前述した Relief-F スキームを使用している．また，ラッパーメソッドは，初期状態を空集合とする前向き探索を採用し，探索法については最優良探索を使用した．評価のために使用した 10 フォールドクロスバリデーションは，属性部分集合選択スキームとは独立した外部のループとしている．

表 2: 各属性選択メソッドの正解率

No	データセット名	属性選択なし	フィルターメソッド	ラッパーメソッド	シーズメソッド
1	breast	95.14%	95.14%	+ 95.71% ○	+ 95.56%
2	crx	85.94%	− 85.65%	+ 86.96% ○	− 85.80%
3	Hayes-Roth	92.86% ○	92.86% ○	− 71.43%	92.86% ○
4	labor-neg	82.35% ○	82.35% ○	− 76.47%	82.35% ○
5	pima	74.08%	74.08%	+ 75.0 % ○	+ 75.0 % ○
6	sick	98.77%	98.77%	98.77%	+ 99.18% ○
7	audiology	84.62%	− 81.62%	+ 92.31% ○	+ 92.31% ○
8	chess	99.53% ○	− 99.47%	− 97.87%	99.53% ○
9	glass	65.89%	65.89%	+ 71.96%	+ 72.90% ○
10	lung-cancer	81.25%	81.25%	+ 84.38% ○	+ 84.38% ○
11	wine	94.94%	94.94%	+ 97.50% ○	+ 97.20%
12	monk1	75.69%	+ 88.89% ○	+ 88.89% ○	+ 88.89% ○
13	monk2	65.05%	− 62.50%	+ 67.13% ○	+ 67.13% ○
14	monk3	97.22% ○	97.22% ○	97.22% ○	97.22% ○
	Ave	85.24%	85.76%	85.83%	87.88%

表 3: 各属性選択メソッドの計算時間 (秒)

No	データ名	フィルタメソッド	ラッパーメソッド	シーズメソッド
1	breast	37	26	44
2	crx	43	782	100
3	Hayes-Roth	1	1	4
4	labor-neg	1	16	64
5	pima	29	304	20
6	sick	126	1708	352
7	audiology	2	183	752
8	chess	1753	9060	668
9	glass	4	411	28
10	lung-cancer	1	44	216
11	wine	3	176	68
12	monk1	1	3	20
13	monk2	1	2	12
14	monk3	1	3	16
	Ave	143.07	992.14	168.86

表 4: 各属性選択メソッドにより選択された属性数

No	データ名	全属性数	属性選択なし	フィルタメソッド	ラッパーメソッド	シーズメソッド
1	breast	10	7	10	4	5
2	crx	15	9	14	8	8
3	Hayes-Roth	4	3	3	1	3
4	labor-neg	16	2	13	1	2
5	pima	8	6	8	6	7
6	sick	29	10	25	11	24
7	audiology	69	14	42	9	9
8	chess	36	22	28	12	24
9	glass	10	9	9	5	7
10	lung-cancer	56	2	31	2	2
11	wine	13	3	13	4	6
12	monk1	6	5	3	3	4
13	monk2	6	6	3	0	3
14	monk3	6	2	3	2	2
	Ave	20.3	7.1	14.6	4.9	7.6

3.1 正解率

表 2 は属性選択を行なわない場合と各属性選択メソッドを用いた場合の正解率をそれぞれ示している．正解率の左側には，属性選択なしの場合よりも正解率が向上したものには + を，正解率が低下したものには − を付けた．+ は 1 点，− は −1 点として各メソッドの総合得点を計算すると，シーズメソッドが最高得点であった．また，表 2 の右側には，最高正解率を有するものに ○ を付けた．ここでもシーズメソッドが最高得点であった．さらに各属性選択メソッドの正解率の平均でもシーズメソッドが 1 位であった．従って，正解率の観点からでは，シーズメソッドは最もうまく機能していると言える．

3.2 計算時間

表 3 は各属性選択メソッドによって費された計算時間を示している．平均計算時間を比較すると，フィルタメソッドのコストが最小であり，シーズメソッドは 3 つの中間に位置している．ラッパーメソッドは属性部分集合を選択する際に分類器を何回も実行するため，データサイズの大きい No.6 “sick” と No.9 “chess” には多くの計算時間を費している．一方，シーズメソッドの計算時間は探索空間の大きさに依存するため，全属性数が多い No.7 “audiology” と No.10 “lung-cancer” は他の 2 つのメソッドと比べ，多くの計算時間を費している．

3.3 選択された属性数

表 4 はデータの全属性数，属性選択を行わない場合に使用された属性数，各属性選択メソッドで選択された属性数を示している．選択された属性数を平均すると，ラッパーメソッドの属性数が最小であり，次にシーズメソッド，フィルタメソッドの順であった．ML/DM 結果を理解しやすくするためには，属性数が少ないことは効果的であるため，理解容易性の観点からするとラッパーメソッドが最もよいと言える．ただし，この結果になった理由は今後さらに実験を進め，詳しく検討していく予定である．

4. おわりに

本稿では，ラッパーメソッドと同等程度の精度で，計算コストも抑制できるシーズメソッドを提案し，3 つのメソッドの比較実験を通して，最高の正解率と 2 番目の計算コストの低さを達成できた．今後は，他のデータに対しても適用し，理論的な分析を行う予定である．さらにフィルタメソッドで考察されている指標と学習器の実行結果の相関などを分析することを通して，両メソッドを融合した属性選択法を考察していきたい．

参考文献

[1] R. Kohavi, G.H. John : Wrappers for feature subset selection, Artificial Intelligence 97 (1997) 273-324.

[2] A.L. Blum, P. Langley : Selection of relevant features and examples in machine learning, Artificial Intelligence 97 (1997) 245-271.

[3] 知識発見システムにおける属性処理と属性値処理に関する一考察，小森 麻央，阿部 秀尚，畑澤 寛光，橘 恵昭，山口 高平，人工知能学会全国大会 (第 15 回)

[4] C.J. Merz and P.M. Murphy : UCI repository of machine learning databases (1996)

[5] 機械学習とデータマイニング， 元田 浩，鷲尾 隆， 人工知能学会誌， Vol.12, No7, pp11-18,1997

[6] Ian H. Witten，Eibe Frank : Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations - , Morgan Kaufmann Publishers (2000)

[7] Mao Komori, Hidenao Abe,Takahira Yamaguthi : A New Feature Selection Method Based on Dynamic Incremental Extension of Seed Features, JCKBSE2002 (in submission)