1F5-04

因子分析と属性選択の統合に基づくデータ前処理機構

Data pre-processing Based on the Integration of Factor Analysis and Feature Selection

渡邉悠司[*1] 小森麻央[*1] 阿部秀尚[*1] 山口高平[*2]
YUJI Watanabe MAO Komori HIDENAO Abe TAKAHIRA Yamaguchi

[*1]静岡大学大学院情報学研究科 Graduated School of Informatics,Shizuoka University
[*2]静岡大学情報学部 Faculty of Informatics,Shizuoka University

Here is discussed how to integrate factor analysis and feature selection in data pre-processing of data maining. Factor analysis removes the fatal features that never make highest the accuracy of data mining applications. Seed method, which has been developed in our previous work, leaves important features and deletes trivial features. Then we obtain the data set with selected features to get into a decision tree learner.

We have done a case study, using ten data sets from UCI ML Repository and StatLog, comparing our method with filter and wrapper methods,from the points of accuracy,computational costs and the number of features that make learned rules. This case study has shown us that our method gets beyond the performance of two popular feature selection methods.

1. はじめに

データマイニングや機械学習において，データの前処理のコストは非常に高いとされている．本稿では，その前処理における属性選択に焦点をあて，その自動化に目標をおく．まず先行研究のシーズメソッドの問題点を述べ，その問題を解決するために因子分析とシーズメソッドの統合について考察する．またベンチマークデータセットを用いて，本手法とフィルタメソッド，ラッパーメソッドとの性能を比較，評価する．

2. シーズメソッド

本節では，我々の先行研究のシーズメソッドについて紹介する．まず，その処理手順を紹介し，次に最良属性集合とシーズメソッドの選択した属性との比較，そしてその問題点について示す．

2.1 処理手順

以下にシーズメソッドの概要を示す．

- Step 1: Relief アルゴリズムによりクラス分類に強く関連する属性集合を抽出する．
- Step 2: 決定木学習を 1 回実行し，生成された決定木での属性の出現頻度をもとに大域分類に役立つ属性を抽出し，初期属性とする．
- Step 3: 分類精度を評価基準として初期属性集合を拡張し，最良属性集合を探索する．

2.2 最良属性集合との比較

本節では，シーズメソッドの選択した属性と最良属性集合との比較実験について示す．

使用したデータセットは，UCI ML リポジトリの 7 種類のデータセット (breast,crx,glass,hayes-roth,labor-neg,pima,wine) と，ポルト大学 LIACC による Stat Log プロジェクトの 3 種類のデータセット (australian,diabetes,heart) である．そして実験結果を表 1 に示す．

連絡先: 渡邉悠司，静岡大学，432-8011 静岡県浜松市城北 3-5-1，yuji@ks.cs.inf.shizuoka.ac.jp

表 1: 最良属性集合とシーズメソッドが選択した属性集合の比較

No	データ名	最良属性集合		シーズメソッド	
		精度	属性数	精度	属性数
1	breast	96.14%	3	96.14%	5
2	crx	87.68%	6	87.68%	8
3	glass	76.64%	5	73.39%	7
4	hays-roth	92.86%	3	92.86%	3
5	labor-neg	82.35%	2	82.35%	2
6	pima	75.65%	3	75.65%	7
7	wine	97.75%	4	97.50%	6
8	australian	87.83%	6	87.83%	6
9	diabetes	75.65%	3	75.65%	6
10	heart	84.07%	5	82.60%	6
Ave.		85.66%	4	85.17%	5.6

2.3 問題点

表 1 に示すように，No3,No7,No10 の 3 つのデータセットにおいて，シーズメソッドが最高精度に達していない．これは学習の性能を劣化させる不適切な属性が，初期属性に含まれていたことを示している．また No1,No2,No6,No9 の 4 つのデータセットは，最高精度には達しているものの，属性数が最良属性集合より多いため，冗長な属性が初期属性に含まれていることを示している．初期属性の決定の方法に問題があることがこの結果よりわかった．

3. シーズメソッドと因子分析の統合

シーズメソッドでは，初期属性集合に冗長または不適切な属性が選択される場合があるために，分類精度が劣化する場合や可読性の高い（サイズが小さい）決定木が作成されない場合がある．

本節では適切な初期属性集合選択を目標とし，Relief アルゴリズムで残存してしまう無関連属性の除去のため因子分析を適用し，2 節で使用した共通データセットに対して実験・評価を行った．

3.1 因子分析の統合

決定木を構成するノード（属性）は大きく 2 つに分けることができる．木の上位にある属性は多くのルールに共通して現れる属性，すなわち大域的に分類することに貢献する属性 (共通属性) であり，木の下位にある属性は少数のルールに現れる，特殊な分類に貢献する属性 (固有属性) である．

シーズメソッドは，ML/DM スキームを実行する前に，Relief アルゴリズムにより属性を絞り込み，その属性集合に対し，決定木による学習を 1 回実行し，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初期属性とした．

しかし，Relief アルゴリズムは局所的にクラス分類に貢献する属性を判断しているため，データ全体という視点からでは不適切である属性を選択してしまうことがあった．その解決法として，クラスに依存せずに不適切，冗長な属性を見出すために，因子分析の共通因子を利用する．共通因子に影響を与えない属性は不適切，冗長な属性であると仮定し，削除することで，適切な初期属性集合を抽出することを試みた．

3.2 処理手順

因子分析と統合したシーズメソッドの処理手順を図 1 に示す．

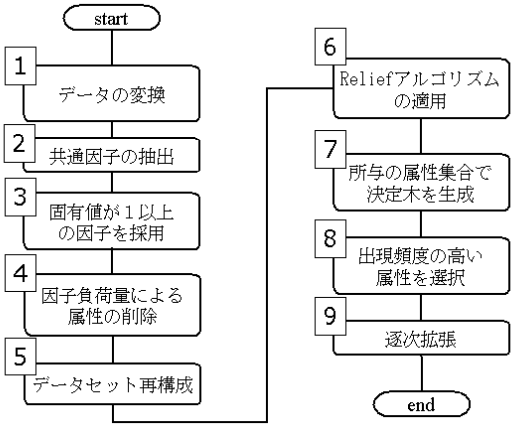


図 1: 因子分析とシーズメソッドの統合処理手順

Step1 ： データの変換 準備されている元データを因子分析で扱えるように，全ての属性を数値属性に変換する．本来なら，専門家の意見などに従って変換するのが理想的だが，今回はデータセットに付属している資料等を参考にした．

Step 2 ：共通因子の抽出 数値属性に変換されたデータを用いて，因子分析を行う．ここでの目的は，共通因子に影響しない属性を削除することで，不適切または冗長な属性を減少させることある．

Step 3 ： 因子の採用 因子数の決定には固有値を用いる．今回の実験では共通因子の妥当性を考慮し，固有値が 1 以上の因子を採用した．

Step 4 ： 因子負荷量による属性削除 採用された共通因子の因子負荷量が最小の属性となる削除する．また, ここで削除された属性は初期属性を拡張する段階でも使用しない．

Step 5 ： データセット再構成 ステップ 4 で削除されなかった属性のデータセットを再構成する．数値属性への変換は因子分析のためのものであるため，再構成されたデータセットは元データのままの形の属性である．

Step 6 ：Relief アルゴリズムの適用 ステップ 2 で再構成された属性集合に対して，それぞれ Relief アルゴリズムを用いてクラスへの関連度による重み付けを行う．今回は Relief アルゴリズムで選択された属性のうち，上位 2〜3 割程度の属性を順に抽出することにした．

Step 7 ：決定木の作成 ステップ 6 で抽出された属性集合で ML/DM スキームを 1 回実行し，決定木を生成する．

Step 8 ： 出現頻度による属性選択・初期属性集合決定 大域的に分類することに貢献する属性が初期属性集合として選択されるため，Step7 で生成された決定木において出現頻度の高い属性を初期属性集合とする．

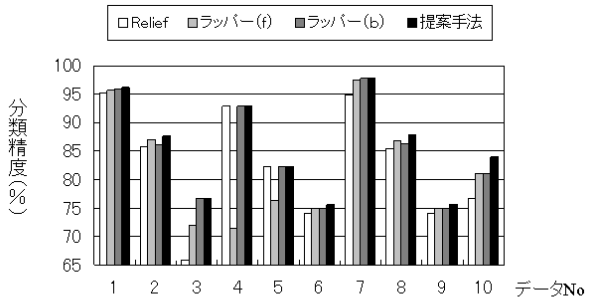
Step 9 ：初期属性集合の拡張 初期属性集合を探索の開始点として，前向き探索による属性集合の拡張を行う．このステップでは，特殊なクラス分類に貢献する固有属性を逐次的に追加することで，分類精度の向上を目的としている．精度の向上が終わるまで属性集合の拡張を続け，精度が向上しなくなった時点で処理を終了し，最後の属性集合を ML/DM スキームの入力として与える．

4. 共通データセットによる有用性の評価

本節では，因子分析とシーズメソッドの統合手法の有用性を確かめるため，2 節で使用したデータセットを用いて，分類精度，処理時間，選択された属性数の 3 点について，フィルタメソッド（Relief アルゴリズム），ラッパーメソッド（空集合から逐次的に属性を追加して適切な属性集合を見出す前向き探索．図 2〜図 4 では f と示す.），ラッパーメソッド（全集合から逐次的に属性を削除して適切な属性集合を見出す後ろ向き探索．図 2〜図 4 では b と示す）との比較・評価を行った．評価結果を図 2,3,4 に示す．

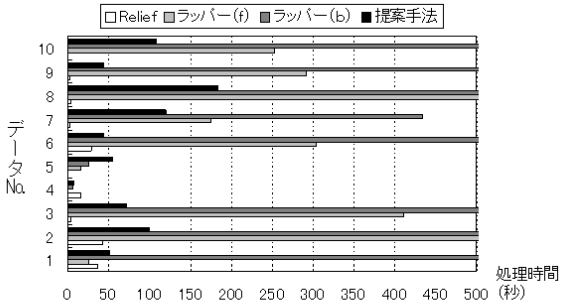
分類精度では全てのデータセットに対して本手法が最高の性能を示し，また最良属性集合と全て同じ精度を示した．処理時間については Relief アルゴリズムには劣るものの，ラッパーメソッドと比較して現実的な処理時間を得ることができた．選択された属性数についても，Relief アルゴリズム，ラッパーメソッドを上回る性能を示した．

以上のことから，本提案手法は従来の属性選択法に比較して，有効に働くことが示されたといえる．



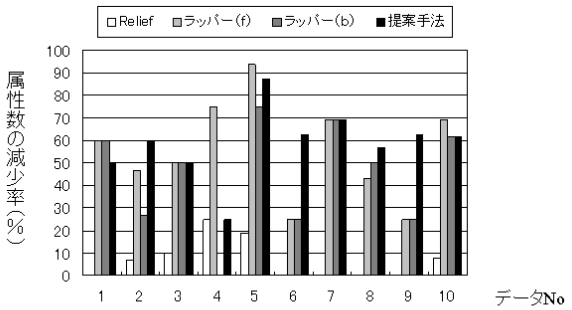
Relief	ラッパー(f)	ラッパー(b)	提案手法
82. 72%	81. 79%	84. 19%	85. 66%

図 2: 分類精度



Relief	ラッパー(f)	ラッパー(b)	提案手法
14. 1	288	2403. 2	73. 6

図 3: 処理時間



Relief	ラッパー(f)	ラッパー(b)	提案手法
6. 9%	55. 68%	44. 24%	58. 53%

図 4: 選択された属性数

5. おわりに

シーズメソッドと因子分析を統合することによって，データ全体から見て不適切，または冗長である属性を排除することができ，初期属性集合として決定木の上位に位置する属性を選択することに成功した．

今回は因子分析を冗長，不適切な属性の削除のため使用したが，共通因子を新規属性とする新しい属性構築の検討，そして因子分析ではなく他の統計手法とシーズメソッドの統合による属性選択についての検討していきたいと思う．

参考文献

[Kohavi97] R. Kohavi, G.H. John : “Wrappers for feature subset selection”, *Artificial Intelligence 97* , pp.273-324 (1997).

[Kononenko 94] “Estimating attributes: analysis and extensions of Relief”, *Proceedings European Conference on Machine Learning*, (1994).

[松尾 02] 松尾 太加志，中村 知靖:誰も教えてくれなかった因子分析，北大路書房 (2002).

[元田 97] 元田 浩，鷲尾 隆 : 機械学習とデータマイニング，人工知能学会誌， Vol.12, No7, pp11-18,(1997).

[小森 02] 小森麻央，阿部秀尚，山口高平：シーズ属性の拡張に基づく属性選択法の提案と評価人工知能学会全国大会（第 16 回）