2F3-04

データセットの特徴分析に基づく因子分析と属性選択の統合手法の提案と評価

The Integrating Factor Analysis and Attribute Selection Based on Feature Analysis of Data Sets

久米 俊二[*1] 渡邉 悠司[*1] 阿部 秀尚[*2] 山口 高平[*3]
Shunji Kume Yuji Watanabe Hidenao Abe Takahira Yamaguchi

[*1]静岡大学大学院情報学研究科
Graduate School of Informatics, Shizuoka University

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

[*3]慶應義塾大学理工学部
Faculty of Science and Technology, Keio University.

At the stage of data pre-processing in knowledge discovery, attribute selection is so important with many attributes from given data sets. This paper specifies how is attributes selection going; finding out starting point, search method, evaluating attribute sets, and stopping conditions. After we build up method repositories based on the specification, we invent constructive meta-level attribute selection to compose proper attribute selection algorithms from method repositories. We are still going on the evaluation of our constructive meta-level attribute selection.

1. はじめに

データマイニングにおいて，所与の属性群には不要な属性が含まれていることが多いため，属性選択はデータ前処理の重要なタスクになっている．我々は先行研究として，探索開始点を適切に設定することによりラッパーメソッド [Kohavi97] の計算コストの問題を改善したシーズメソッド [小森 02] を開発してきた．シーズメソッドは，ラッパーメソッドより小さな計算コストでラッパーメソッドと同等の分類精度を導出する属性群を選択することができるが，適切な探索開始点の設定に失敗することがあり，この場合に導出される属性群の持つ分類精度は非常に劣ったものであった．そこで，シーズメソッドの探索開始点設定問題において因子分析を利用し，不適切な属性を事前に除去する手法を考案・適用した [渡辺 03]．この時，探索法を固定した場合では有用性が確認されたが，探索法を変更すると有意な差が見られなくなることが判明した．すなわち，よりよい属性選択法を模索するためには，探索開始点だけを考慮すればいいということはなく，属性選択アルゴリズム全体を総合的に考慮しなければならないことがわかった．

以上の背景より，本稿では，探索開始点を選定するだけでなく，属性選択法をメソッド (意味を保つ範囲で最小のアルゴリズム構成要素) レベルで分解し，データセット毎に適した属性選択法をメソッド単位から合成していく手法を提案・評価する．

2. 探索開始点の選定による効果

属性選択法の性質を決定するメソッドは次の 4 種類，**探索開始点，探索法，属性評価法，探索終了条件**であると同定できる．**探索開始点**は，その名の通り探索を開始する点を示すものである．探索法によっては，探索開始点の選定によって探索空間が大幅に縮小される．**探索法**は，探索の進行の仕方を規定するもので，最優良探索などがある．**属性評価法**は，属性探索空間に存在する各属性群に評価を与える方法のことである．評価は探索の進行や探索の終了に関わる．**探索終了条件**は，探索を打ち切る条件を規定するものである．シーケンシャルな探索法の探索終了条件として「評価が向上しなくなったら終了」など，ランダムな探索法の探索終了条件として「一定数の繰り返し処理をして終了」などがある．

図 1 は，4 属性における属性探索空間と，各メソッドの役割を表す．黒丸は選択属性，白丸は非選択属性を表す．

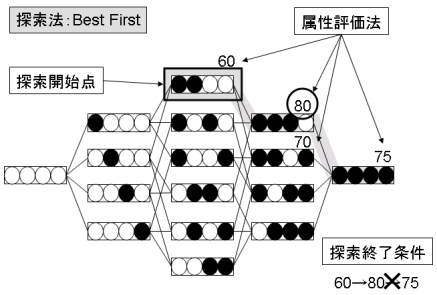

図 1: 4 属性の属性探索空間と各メソッドの役割

2.1 シーズメソッド

シーズメソッドは，前向き探索を行う属性選択法において**探索開始点**を適切に選定することにより，不要な属性が選択される可能性を減らし，探索空間を制限して，精度の向上を図りつつ計算コストも抑えることができる手法である．

シーズメソッドは，**探索開始点**を選定するために以下の処理を行う．

1. データセットに対して因子分析を適用し，固有値が 1 以上の共通因子を抽出
2. 各因子に対して因子負荷量最小の属性を除去
3. RELIEF-F[Kononenko 94] によりクラス関連度の高い属性を抽出

連絡先: 久米 俊二, 静岡大学大学院情報学研究科, 〒 432-8011
静岡県浜松市城北 3-5-1,TEL:053-478-1510,
e-mail : cs9030@cs.inf.shizuoka.ac.jp

4. 抽出された属性を使って決定木学習 (C4.5)[Quinlan 93] を実行，木の上位に現れる属性を**探索開始点**として選定

2.2 共通データセットによる有用性の評価

シーズメソッドの効果を確かめるため，ケーススタディを行った．データセットは UCI ML リポジトリの 7 データセットを使用し，評価は決定木学習（C4.5）の 10foldCV による分類精度とした．このときシーズメソッドは,「**探索開始点**:Seed Set，**探索法**:Best First Forward Search，**属性評価法**:WrapperEvaluation，**探索終了条件**:Not Improved」で定義される属性選択法である．比較対象として，シーズメソッドの**探索開始点**部分に空集合 (Null Set) を代入した属性選択法 (空集合属性選択法) の分類精度と，全探索から得られた最良分類精度を用意した．結果は表 1 に示す．◯印は最良分類精度と同じ数値であることを示す．表 1 から，適切な探索開始点を選定するシーズメソッドの方が，選定を行わない属性選択法より優れた分類精度を導出することがわかる．

表 1: 探索開始点選定の効果

	空集合属性選択法 (%)	シーズメソッド (%)	全探索による最良精度 (%)
breast	◯ 94.7067	◯ 94.7067	◯ 94.7067
glass	68.2243	◯ 77.5701	◯ 77.5701
labor	◯ 85	◯ 85	◯ 85
pima	73.8281	◯ 75.9115	◯ 75.9115
wine	96.0674	96.0674	◯ 97.191
australian	85.5072	84.6377	◯ 87.2464
heart	◯ 85.1852	◯ 85.1852	◯ 85.1852

2.3 課題

これまで前向き探索を使用する属性選択法において，探索開始点を適切に選定した場合の効果を検証し，その有用性を確認した．しかし，前向き探索ではない属性選択法，例えば遺伝的アルゴリズム (以下，GA) を利用した探索法 (Genetic Search) を含む属性選択法は，探索開始点の選定を行わなくともシーズメソッドと同等以上の分類精度を導出する．Genetic Search を含む属性選択法の一例として,「**探索開始点**:Random set，**探索法**:Genetic Search，**属性評価法**:Wrapper Evaluation，**探索終了条件**:Number of Generations」で定義される属性選択法 (GA 利用属性選択法) を用意し，実験・比較した結果を表 2 に示す．

表 2: 探索法の違いによる分類精度の比較

	GA 利用属性選択法 (%)	シーズメソッド (%)	全探索による最良精度 (%)
breast	◯ 94.7067	◯ 94.7067	◯ 94.7067
glass	◯ 77.5701	◯ 77.5701	◯ 77.5701
labor	◯ 85	◯ 85	◯ 85
pima	◯ 75.9115	◯ 75.9115	◯ 75.9115
wine	◯ 97.191	96.0674	◯ 97.191
australian	86.8116	84.6377	◯ 87.2464
heart	84.8148	◯ 85.1852	◯ 85.1852

GA 利用属性選択法は探索開始点をランダムに選定しているにも関わらず，適切に探索開始点を選定しているシーズメソッドの分類精度より優れた分類精度を導出している．

このことから，一部の属性選択法の探索開始点に対する改善は，属性選択アルゴリズム全体からみれば小さな改善に過ぎず，真に優れた属性選択法を開発するためには総合的な改善が必要であると考えられる．

3. 構成型メタレベル属性選択法

前節では，属性選択法を構成する各メソッドに対する個別の改善は一定の効果を挙げることを示した．しかし，改善したメソッド以外の他のメソッドの影響によってその改善の効果が吸収されてしまうことも示された．以上の検証から，属性選択法が持つ全てのメソッドを考慮した総合的な改善が重要であることを確認できた．

そのための手法として，メソッド全体を記述した属性選択法単位でデータセット毎に適したものをメタ学習で選び出す**選択型メタレベル属性選択法**がまず考えられる．しかし，メソッドリポジトリを拡張していくにつれ，選択肢となる属性選択法の数が膨大なものとなり，選択型では対応しきれなくなることが想定される．

その点を踏まえた上で，メソッド全体を考慮した総合的な属性選択法として，我々は**構成型メタレベル属性選択法**を提案する．構成型メタレベル属性選択法とは，属性選択法をメソッドレベルで分解し，それらの組み合わせとしてデータセット毎に適切な属性選択法を合成していく手法のことである．

構成型メタレベル属性選択法の実装方法・実験結果等を以下に示していく．

3.1 メソッドの分類・整理

本稿の実験で構成型メタレベル属性選択法 (以下，提案手法) が使用した各メソッドを分類・整理したメソッドリポジトリを図 2 に示す．これらのメソッドは Weka(Waikato Environment for Knowledge Analysis) [Witten 00] の実装と対応する．メソッドの組み合わせによっては実行不可能なもの，無意味なものが存在するので，それらの可能性を除去すると，このメソッドリポジトリからは **48 種類**の属性選択法を合成することができる．

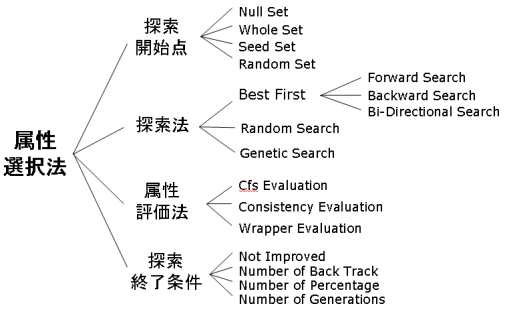

図 2: メソッドリポジトリ

3.2 遺伝的アルゴリズムの適用

メソッドを合成・評価するプロセスには GA を採用した．

4 種類のメソッドをそれぞれ 1 遺伝子としてコード化し，個体はそれらのコードの組み合わせから成るストリングとして表現する．例えば属性評価法においては，Cfs Evaluation，Consistency Evaluation，Wrapper Evaluation という 3 つの

メソッドに対してそれぞれ A，B，C のアルファベットコードを割り当てる．すなわち個体の表現は「ABCD」のような 4 桁のアルファベットコードになる．このとき 1 個体は，合成された 1 つの属性選択法を表現している．

本稿の実験で適用した GA の各パラメータを以下に示す．

- Population：5
- Max Generation：10
- Crossover Probability：50%
- Mutation Probability：1%

個体の評価は，その個体（属性選択法）が導出する属性群の分類精度（C4.5）を用いた．また，最も評価の高い個体を交叉や突然変異の対象にせず，必ず次代に残す“ エリート保存戦略 ”を採用した．

GA によって，データセットに適した遺伝子 (メソッド) が生き残り，適した 4 遺伝子 (メソッド) 同士の組み合わせを得られることが期待できる．

3.3 共通データセットによる有用性の評価

提案手法の有用性を確かめるためのケーススタディを行った．データセットは UCI ML リポジトリの 7 データセットを使用し，評価は決定木学習（C4.5）の 10foldCV による分類精度とした．提案手法は，データセット毎に属性選択法を合成し，評価した．比較対象として，単一の属性選択法として最良の分類精度を導出した属性選択法「**探索開始点**:Random set，**探索法**:Genetic Search，**属性評価法**:Wrapper Evaluation，**探索終了条件**:Number of Generations」(最良単一属性選択法) を用意した．結果を表 3 に示す．

表 3: 構成型メタレベル属性選択法の分類精度

	提案手法 (%)	最良単一属性選択法 (%)	全探索による最良精度 (%)
breadt	◯ 94.7067	◯ 94.7067	◯ 94.7067
glass	◯ 77.5701	◯ 77.5701	◯ 77.5701
labor	◯ 85	◯ 85	◯ 85
pima	◯ 75.9115	◯ 75.9115	◯ 75.9115
wine	96.6292	◯ 97.191	◯ 97.191
australian	◯ 87.2464	86.8116	◯ 87.2464
heart	◯ 85.1852	84.8148	◯ 85.1852

表 3 から，提案手法と最良単一属性選択法の分類精度を比較してもそれほど大きな差はないことがわかる．また wine において，提案手法は適切な属性選択法の合成に失敗していることから，“データセット毎に適切な属性選択法を合成する”という目標は未達成と言える．

次に，提案手法が適切な属性選択法を合成するまでに要した探索コストを表 4 に示す．

breast や labor のような正解が数多い，合成が簡単なデータセットにおいて，GA が始まる前の 0 世代の段階で適切な属性選択法を合成していることがわかる．しかしこれは確率と偶然によって得られただけに過ぎず，提案手法中の GA の効果とは言えない．

逆に，wine や australian のような正解が数少ない，合成が難しいデータセットにおいては，探索コストが高コスト (全探索ですら 48 探索空間で済む)，あるいは最後まで適切な属性選択法を合成できていないことがわかる．これは GA による合成がそれほど効率的・効果的に行われなかったことを示している．

表 4: 探索コスト

	提案手法が要した探索空間 (世代数)	適切属性選択法数/属性選択法全体数
breast	5(0)	32/48
glass	35(6)	5/48
labor	5(0)	34/48
pima	45(8)	10/48
wine	55(10)	2/48
australian	45(8)	3/48
heart	15(2)	8/48

4. おわりに

本稿では，属性選択の仕様に基づいて属性選択メソッドリポジトリを整備した後，リポジトリを利用した構成型メタレベル属性選択法を提案した．リポジトリの整備は始めたばかりなので，合成されるアルゴリズム本数は少なく，そのため，優秀な一つの属性選択法と比較して，有意な差を示す程の効果は確認されていない．リポジトリの規模拡大，合成プロセスとしてのＧＡの最適化，より多くのデータセットによる実験評価などが今後の課題である．

謝辞

本研究は，文部科学省科学研究費補助金特定領域研究 (13131205)「メタ学習機構に基づくアクティブマイニング」の助成によるものである．

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).

[Quinlan 93] Quinlan, J,R.: “C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers (1993).

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

[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” , 第 16 回人工知能学会全国大会, 1A4-02, (2002).

[渡辺 03] 渡邉悠司，小森麻央，阿部英尚，山口高平:” 因子分析と属性選択の統合に基づくデータ前処理機構” , 第 17 回人工知能学会全国大会, 1F5-04, (2003).