ISS-P-118

特徴ランキングに基づく 動的な探索域による特徴選択

能登　優太† 　森　康久仁† 　松葉　育雄†
† 千葉大学融合科学研究科

1. はじめに

近年，情報技術の発展に伴い高次元のデータを扱う機会が増え，パターン認識の分野においてはデータを低次元化する特徴選択が重要な処理となっている．本稿では探索域を変えながら特徴部分集合を求める，新たな特徴選択法を提案する．

2. 特徴選択

特徴選択は元のデータに含まれる特徴の内，不適切または冗長な特徴と取り除き，できるだけ識別精度を落とさずにデータを低次元化することである．特徴選択法の分類の中にはラッパー法とフィルター法が存在する．ラッパー法は選択する特徴部分集合を変えながら，識別器を構成，評価をおこない最適な特徴部分集合を探索する方法である．一方，フィルター法は特徴をなんらかの基準をもとに順位をつけて各特徴を選択していくことで特徴部分集合を決定する方法である．

一般にラッパー法は実際の識別器により評価を行うため識別精度は高くなるが，識別器を大量に作るため計算コストが高い．フィルター法はラッパー法に比べ，計算コストは低いが識別精度が低いとされている．近年では，ラッパー法とフィルター法を組み合せて使う手法が多く提案されている[1, 2].

3. 提案手法

本提案ではd次元のデータからd'次元（$d > d'$）を選択する問題を考える．はじめに，元データの特徴を用いてフィルター法により特徴ランキングを求める．次に特徴をランキング順にg個単位の特徴グループに分割する．この特徴グループを利用しラッパー法による探索をおこなう．最初は最上位の特徴グループに属するg個の特徴から現在よりも高評価の特徴部分集合を探索する．そのような特徴部分集合が存在した場合は解を更新し，次回の探索域を最上位の特徴グループにする．現在よりも高評価の特徴部分集合が存在しなかった場合は，次回の探索域を一つ下位の特徴グループにする．これを繰り返し，最下位の特徴グループでも現在より高評価の特徴部分集合が見つからなければ停止する(図 1).

この方法は，各特徴が評価される機会の多さをフィルター法によるランキングで重みづけすることができる．これによりフィルター法による閾値を用いた特徴選択では捨てられるランキング下位の特徴が選択される可能性を残しつつ，計算コストを抑えることができる．

4. 実験及び結果

提案手法の有効性を確認するために，高次元データを持つ文書分類問題である real-sim（クラス数: 2，特徴数: 20958，データ数: 72309）を用いて実験をおこなった．本

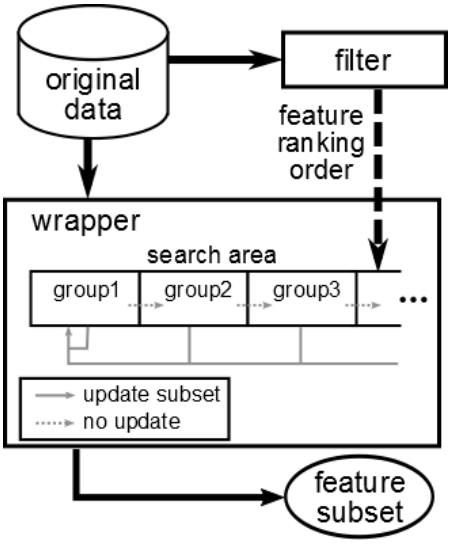

図 1. 提案手法の概要

実験ではデータは訓練用データ 52309 個，テストデータ 20000 個に分割した．このデータから 10 個の特徴を選択する($d' = 10$)こととする．特徴ランキングを求めるフィルター法は RFE-SVM を用い，ラッパー法には plus-1-minus-1 探索と SVM による識別率による評価値を用いた．特徴グループには 100 個の特徴を入れる($g = 100$)．その結果，毎回全特徴を探索する方法に比べ，高速に解の探索をおこなうことができた(表1).

表 1. 実験結果

方法	時間（秒）	識別率（%）
提案手法	25410	81.99
全域探索	107462	81.99

5. まとめ

特徴の探索域を変えながら探索をおこなうことで効率的に解を求めることができた．今後は，特徴グループの特徴数gや用いる特徴ランキングによる結果の比較をしたい．

参考文献

[1] A.E. Akadi, A. Amine, A.E. Ouardighi, and D. Aboutajdine, “A two-stage gene selection scheme utilizing MRMR filter and GA wrapper” Knowledge and Information Systems, Vol.26, Issue 3, pp.487-500, March 2011.

[2] P. Bermejo, L.D.L. Ossa, J.A. Gámez, J.M. Puerta, “Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking” Knowledge-Based Systems, Vol.25, Issue 1, pp.35-44, February 2012.