2D-03

変数を細分類化する特徴選択方法

鄭　弯弯

東京理科大学工学部情報工学科

1. はじめに

高次元空間でデータを扱うことは，過学習の回避，学習モデルの解釈性と計算コストの抑制などの理由から望ましくない．次元削減は，高次元空間から低次元空間へデータを変換しながら，元データの特性を保持する手法であり，画像処理，音声処理，バイオインフォマティクス，テキストマイニングと金融・経済などさまざまな研究分野に頻繁に適用され，重要な研究トピックとして挙げられている．次元削減は，特徴選択（Feature Selection）と特徴抽出（Feature Extraction）に分けられる．特徴選択はデータセットから有用な特徴量を選択する．元の変数が保持されるため，解釈しやすいが，情報損失が必ず生じるため，汎化能力に欠如している．特徴抽出はデータセットを要約するような新しい特徴量を作り出す．有用な潜在変数を作成できるが解釈しむずかしい．特徴選択と特徴抽出の重要性と応用性により，数多くの手法の提案と改善が重ねて研究されてきた．本研究では，特徴選択を扱う．

特徴選択において，変数とクラスの相関，変数と変数の相関，変数と変数の交互作用の三つの要素から変数の重要度を判定できる．しかし，現時点の特徴選択方法はどちら一方しか測っていない問題が存在している．例えば，フィルタ法は各変数が独立であることを仮定しているため，単独で無効であるが他の変数と併用して有効な変数は選ばれない；また，高相関を持つ変数が選択されるため，変数セットは冗長になる．また，変数は(1)独立で有効，(2)他の変数とセットとして有効，(3)1と2の性質両方がある，(4)冗長である，(5)無効である五つの種類に分けられるが，既存方法は主に無効な変数を取り除くことを目的としている．本研究は，相関と相互作用の機能を考慮し，変数を細分類化する特徴選択方法を構築した．

A feature selection method with subdivisions of variables
†Wanwan Zheng, Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science

2. 提案モデル

本研究はベースの特徴選択方法として Boruta（Kira & Rendell, 1992a, 1992b）と Relief（Kursa & Rudnicki, 2010）を用いる．Boruta はシャウド特徴量（shadow features）を作って，ランダムフォレストで学習を繰り返して元変数ごとにシャドウ特徴量より重要度が上回る回数が低いものを除外する．変数はセットで学習を行っているため，変数間の相関関係を考慮している．また，シャドウ特徴量の作成は，低次元データにも適応でき，特にサンプル数が多くて変数数が少ないデータに効果が顕著であると報告されている．Relief の変数重要度は最近傍サンプルペア間の特徴量値の差に基づいている算出されている．同じクラスの隣接するサンプルペアで特徴量値の差が大きいほど，変数重要度が小さくなる．一方，クラスが異なるサンプルペアで特徴量値の差が大きいほど，変数重要度が大きくなる．回帰問題において，二つのサンプルが同じクラスに属するかどうかの知識を必要とする代わりに，二つのサンプルの予測値が同じクラスである確率を使う．この確率は，二つのサンプルの予測値間の相対距離で評価する．Relief は変数間の交互作用に敏感であると指摘されている．

本研究では，各変数に対して，Boruta で相関，Relief で交互作用の性能を変数重要度とランクで評価し，二つの合計を各変数の総合的な性能にする．さらに，図 1 に示すモデル SoV（subdivisions of variables）を使用し，変数を独立で有効，他の変数とセットとして有効，冗長，無効 4 つの種類に分ける．SoV は変数とクラスの相関を起点として，変数と変数の相関関係の議論を経て，その結果に基づいて変数を分類し，最適な変数セットを決めるような流れになる．

相関を測るときには，スピアマンの順位相関係数を使用する．相関が高いか低いかを判断する閾値をパラメータとして設定する．変数重要度を表すスコアの極端に小さい値は，

次の式で算出した閾値を用いて判断する．

Threshold = 第1四分位数 − r×IQR
IQR = 第3四分位数 − 第1四分位数

3. 数値実験

3.1 シミュレーションデータ

サンプル i のボーナス（bonus）に対して，勤務年数（experience），業績評価（performance），契約数（sales），遅刻欠勤の日数（days_late）四つの共変量の乱数データを生成する．勤務年数と遅刻欠勤の日数は整数乱数（それぞれの範囲は[0, 50]，[0, 30]），業績評価と契約数は一様分布に従う乱数（それぞれの範囲は[0, 10]，[0, 100]）とする．結果変数ボーナスY_iを次式で定義する．

$$Y_i = 2 \times experience_i + 7 \times performance_i + 8 \times sales_i + -6.5 \times day\_late_i + \epsilon_i, \epsilon_i \sim N(0,1)$$

また，ノイズ変数n_{i1}，n_{i2}，n_{i3}三つ，冗長変数r_{i1}，r_{i2}，r_{i3}三つを追加する．r_{i1}，r_{i2}，r_{i3}は契約数との相関はそれぞれ0.85，0.9，0.95とする．サンプル数 n=2000 とする．

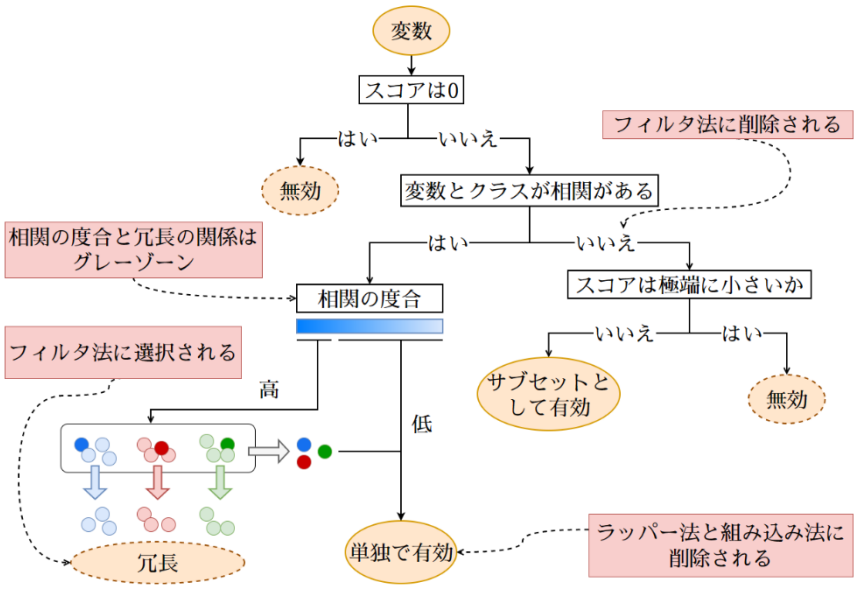

図1　SoVの全体構造

3.2 シミュレーション結果

分類器は，サポートベクター回帰（support vector regression, SVR）を使用する．SoVによる特徴選択の結果は表1に示す．勤務年数，業績評価，契約数，遅刻欠勤の日数の四つの共変量はすべて正しく単独で有効である特徴量として選択された．ノイズ変数n_1にはノイズ，冗長変数r_1，r_2，r_3には冗長変数に分類できた．n_2とn_3は他の変数と交互作用がある変数に分類されたが，変数重要度は選択された特徴量の中に最も小さかった．同じデータに対してBorutaとLassoを使って特徴選択の結果と比較した．

表1　シミュレーションによる特徴選択の結果

	変数	重要度	変数	重要度	変数	重要度
単独で有効	sales	1.68	days_late	1.37		
	experience	0.82	performance	0.59		
他の変数とセットとして有効	n_3	0.26	n_2	0.14		
ノイズ	n_1	0				
冗長	r_3	1.32	r_2	1.08	r_1	0.93

4. まとめ

本研究は，変数を細分類化する特徴選択方法を提案した。人工データを用いたシミュレーションの結果，正しく各種類の変数を細分類できた．しかし，ノイズは有効な変数として選択されたため，変数重要度の極端小さい値の判断に使用する閾値の設定の課題が残されている．

参考文献

[1] Kursa, M. B. and Rudnicki, W. R., Feature selection with the Boruta package, Journal of Statistical Software, 36(11), 2010.