Chemometrics and Intelligent Laboratory Systems 142 (2015) 172–183

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Stress test procedure for feature selection algorithms ☆

A.M. Katrutsa [a,b,*], V.V. Strijov [a]

[a] Moscow Institute of Physics and Technology, Institutskiy lane 9, Dolgoprudny City 141700, Russian Federation
[b] Skolkovo Institute of Science and Technology, Novaya St., 100, Karakorum Building, 4th floor, Skolkovo 143025, Russian Federation

ARTICLE INFO

Article history:
Received 13 November 2014
Received in revised form 12 January 2015
Accepted 30 January 2015
Available online 7 February 2015

Keywords:
Regression analysis
Feature selection methods
Multicollinearity
Test data sets
The criterion of the selected feature redundancy

ABSTRACT

This study investigates the multicollinearity problem and the performance of feature selection methods in the case of data sets that have multicollinear features. We propose a stress test procedure for a set of feature selection methods. This procedure generates test data sets with various configurations of the target vector and features. This procedure provides more complex investigations of feature selection methods than procedures described in papers previously. A number of some multicollinear features are inserted in every configuration. A feature selection method results in a set of selected features for a given test data set. To compare given feature selection methods the procedure uses several quality measures. A criterion of the selected feature redundancy is proposed. This criterion estimates the number of multicollinear features among the selected ones. To detect multicollinearity it uses the eigensystem of the parameter covariance matrix. In computational experiments we consider the following illustrative methods: Lasso, ElasticNet, LARS, Ridge, Stepwise and Genetic algorithms and determine the best one, which solves the multicollinearity problem for every considered data set configuration.

1. Introduction

This study investigates the multicollinearity problem and proposes a test procedure for feature selection methods to evaluate their performance in solving this problem. *Multicollinearity* is a strong correlation between the features, which affect the target vector simultaneously. Due to multicollinearity the common methods of regression analysis like least squares build unstable models of excessive complexity. A model is called *stable* if a small change of the parameter vector leads to a small change of the target vector estimation. A model has *excessive complexity* if an adequate estimation of the target vector needs excessive number of features. The formal definitions of model stability and model complexity are given in Section 2.

Feature selection methods are intended to improve stability and reduce complexity of models by reducing dimensionality of the model parameter space and by removing irrelevant features [1,2]. This paper treats the multicollinear features as irrelevant ones. It analyses stability, complexity and redundancy of the models obtained by feature selection methods.

This study proposes and develops a new procedure to test feature selection methods. The main purpose of this procedure is to reveal the pros and cons of test feature selection methods in solving the multicollinearity problem. The test procedure uses artificial data sets generated by the developed test data generation routine. The input arguments of this routine are the number of features to generate and number of samples, and the output is the test dataset. A real data set obtained by spectroscopy [3] is included in the investigation. To construct test data set the data generation routine uses multicollinear features, features correlated to the target vector, orthogonal features and features orthogonal to the target vector. In this paper the computational experiment involves four data sets: 1) inadequate and correlated, 2) adequate and random, 3) adequate and redundant, and 4) adequate and correlated. Here the adequate data set means that there exists some linear combination of the features, which fits the target vector. The redundant data set means that there exist an excessive number of features, which fits the target vector. The correlated data set means that there exist some features correlated to each other. The data generation routine can construct another data set of any requested configuration. It has the parameters to control a data set configuration. The parameters are listed in Section 4. We consider these four data sets, because they have multicollinear features in an exhaustive set of positions with respect to the target vector and to each other. These configurations give a complex investigation of feature selection methods.

Here we test the following methods: LARS [4], Lasso [5], ElasticNet [6], Ridge [7], Stepwise [8] and Genetic algorithms [9]. In Section 6 we compare them according to various quality measures. In addition, we propose the criterion of the selected feature redundancy described in Section 5. It ranks feature selection methods according to the redundancy of selected models. The proposed criterion estimates the number of multicollinear features among the selected ones for some given limit value of the error function. The feature selection methods are ranked according to increasing number of multicollinear features in the set of

☆ This publication is based on the work funded by Skolkovo Institute of Science and Technology (SkolTech) No. 081-R within the framework of the SkolTech/MITInitiative.
* Corresponding author at: Moscow Institute of Physics and Technology, Institutskiy lane 9, Dolgoprudny City 141700, Russian Federation.
E-mail address: aleksandr.katrutsa@phystech.edu (A.M. Katrutsa).

http://dx.doi.org/10.1016/j.chemolab.2015.01.018
0169-7439/

selected ones. The best method selects a feature set with the minimum number of multicollinear features.

1.1. Related works

The multicollinearity problem, multicollinearity detection methods and approaches in solving this problem are discussed in [10–12]. One way to solve the multicollinearity problem is to use feature selection methods [12]. They are also used in the following machine learning problems: dimensionality reduction [13,14], simplification usage of the standard machine learning algorithms [15], removing irrelevant features [16] and increasing the generalisation ability of applying algorithm [17]. The papers [18,2,1] review existing feature selection methods, and classify them according to error functions and optimum feature subset search strategies.

The papers devoted to test feature selection methods study a specific type of data sets [19] or specific algorithms with different upgrades [20,21]. The most related paper [22] presents a comparison of some feature selection methods on the multicollinearity problem. However, the proposed test procedure provides a more complex investigation of the considered feature selection methods. It evaluates the quality measures while the parameter of multicollinearity and data set parameters are changing continuously. In spite of the paper [22], this study investigates the model stability and complexity.

2. Feature selection problem statement

Let $\mathfrak{D} = \{(\mathbf{X}, \mathbf{y})\}$ be the given data set, where the design matrix

$$\mathbf{X} = \left[\mathcal{X}_1, \ldots, \mathcal{X}_j, \ldots, \mathcal{X}_n\right], \mathbf{X} \in \mathbb{R}^{m \times n} \text{ and } j \in \mathcal{J} = \{1, \ldots, n\}.$$

The vector $\mathcal{X}_j$ is called the j-th feature and the vector $\mathbf{y} = [y_1, \ldots, y_m]^\top \in \mathbb{Y} \subset \mathbb{R}^m$ is called the target vector. Assume that the target vector $\mathbf{y}$ and design matrix $\mathbf{X}$ are related through the following equation:

$$\mathbf{y} = \mathbf{f}(\mathbf{w}, \mathbf{X}) + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{f}$ maps the cartesian product of the feasible parameter space and the space of the $m \times n$ matrices to the target vector domain, and $\boldsymbol{\varepsilon}$ is the residual vector. The data fit problem is to estimate the parameter vector $\mathbf{w}^*$,

$$\mathbf{w}^* = \underset{\mathbf{w} \in \mathbb{R}^n}{\arg\min}\, S(\mathbf{w} | \mathfrak{D}_{\mathcal{L}}, \mathcal{A}, \mathbf{f}), \tag{2}$$

where S is the error function. The set $\mathfrak{D}_{\mathcal{L}} \subset \mathfrak{D}$ is a training set and the set $\mathcal{A} \subseteq \mathcal{J}$ is the *active index set* used in computing the error function S. In the stress test procedure we use the quadratic error function

$$S = \|\mathbf{y} - \mathbf{f}(\mathbf{w}, \mathbf{X})\|_2^2 \tag{3}$$

and the linear regression function $\mathbf{f}(\mathbf{w}, \mathbf{X}) = \mathbf{X}\mathbf{w}$. The introduced stress test procedure could be applied to the generalised linear model selection algorithms, where the model is $\mathbf{f} = \boldsymbol{\mu}^{-1}(\mathbf{X}\mathbf{w})$ and $\boldsymbol{\mu}$ is a link function.

Definition 2.1. Let $\mathcal{A}^*$ denote *the optimum index set*, the solution of the problem

$$\mathcal{A}^* = \underset{\mathcal{A} \subseteq \mathcal{J}}{\arg\min}\, S_{\mathcal{M}}(\mathcal{A} | \mathbf{w}^*, \mathfrak{D}_{\mathcal{C}}, \mathbf{f}), \tag{4}$$

where $\mathfrak{D}_{\mathcal{C}} \subset \mathfrak{D}$ is the test set, $\mathbf{w}^*$ is the solution for problem (2) and $S_{\mathcal{M}}$ is an error function corresponding to a feature selection method $\mathcal{M}$ (5).

The feature selection problem (4) is to find the optimum index set $\mathcal{A}^*$. It must exclude indices of noisy and multicollinear features. It is expected that if one uses features indexed by the set $\mathcal{A}^*$ then it brings more stable solution for problem (2), in comparison with the case of $\mathcal{A} \equiv \mathcal{J}$.

In the computational experiment we consider the feature selection methods from the set $\mathfrak{M} = \{\text{Lasso, LARS, Stepwise, ElasticNet, Ridge}\}$.

Definition 2.2. A feature selection method $\mathcal{M} \in \mathfrak{M}$ is a map from the complete index set $\mathcal{J}$ to active index set $\mathcal{A} \subseteq \mathcal{J}$:

$$\mathcal{M} : \mathcal{J} \to \mathcal{A}. \tag{5}$$

According to this definition we consider the terms feature selection problem and the model selection problem to be synonyms.

Definition 2.3. Let a model be a pair $(\mathbf{f}, \mathcal{A})$, where $\mathcal{A} \subseteq \mathcal{J}$ is an index set. The model selection problem is to find the optimum pair $(\mathbf{f}^*, \mathcal{A}^*)$ which minimizes the error function S (3).

Definition 2.4. Let *the model complexity* C be the cardinality of the active index set $\mathcal{A}$, number of the selected features:

$$C = |\mathcal{A}|.$$

Definition 2.5. Define *the model stability* R as logarithm of the condition number κ of the matrix $\mathbf{X}^\top\mathbf{X}$:

$$R = \ln\kappa = ln\frac{\lambda_{\max}}{\lambda_{\min}},$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and the minimum non-zero eigenvalue of the matrix $\mathbf{X}^\top\mathbf{X}$. The features with indices from the corresponding active set $\mathcal{A}$ are used in computing the condition number κ.

3. Multicollinearity analysis in feature selection

In this section we give definitions of multicollinear features, correlated features and features correlated with the target vector. In the following subsections we list and study the multicollinearity criteria.

Assume that the features $\mathcal{X}_j$ and the target vector $\mathbf{y}$ are normalized:

$$\|\mathbf{y}\|_2 = 1 \text{ and } \left\|\mathcal{X}_j\right\|_2 = 1, \quad j \in \mathcal{J}. \tag{6}$$

Consider active index subset $\mathcal{A} \subseteq \mathcal{J}$.

Definition 3.1. The features with indices from the set $\mathcal{A}$ are called *multicollinear* if there exist the index j, the coefficients a_k, the index $k \in \mathcal{A} \setminus j$ and sufficiently small positive number $\delta > 0$ such that

$$\left\|\mathcal{X}_j - \sum_{k \in \mathcal{A} \setminus j} a_k \mathcal{X}_k\right\|_2^2 < \delta. \tag{7}$$

The smaller δ is, the higher *degree of multicollinearity*.

Definition 3.2. Let the features indexed i, j be *correlated* if there exists sufficiently small positive number $\delta_{ij} > 0$ such that:

$$\left\|\mathcal{X}_i - \mathcal{X}_j\right\|_2^2 < \delta_{ij}. \tag{8}$$

From this definition it follows that $\delta_{ij} = \delta_{ji}$. In the special case $a_k = 0$ $k \neq j$ and $a_k = 1$ $k = j$ the inequalities (8) and (7) are identical.