Expert Systems With Applications 76 (2017) 1–11

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria

Alexandr Katrutsa[a,b,*], Vadim Strijov[a]

[a] Moscow Institute of Physics and Technology, Institutskiy lane, 9, 141700, Dolgoprudny city, Russia
[b] Skolkovo Institute of Science and Technology, Nobel St., 3, 143025, Skolkovo, Russia

ARTICLE INFO

Article history:
Received 14 August 2016
Revised 25 January 2017
Accepted 26 January 2017
Available online 26 January 2017

Keywords:
Data fitting
Feature selection
Multicollinearity
Quadratic programming
Evaluation criteria
Test data sets

ABSTRACT

This paper provides a new approach to feature selection based on the concept of feature filters, so that feature selection is independent of the prediction model. Data fitting is stated as a single-objective optimization problem, where the objective function indicates the error of approximating the target vector as some function of given features. Linear dependence between features induces the multicollinearity problem and leads to instability of the model and redundancy of the feature set. This paper introduces a feature selection method based on quadratic programming. This approach takes into account the mutual dependence of the features and the target vector, and selects features according to relevance and similarity measures defined according to the specific problem. The main idea is to minimize mutual dependence and maximize approximation quality by varying a binary vector that indicates the presence of features. The selected model is less redundant and more stable. To evaluate the quality of the proposed feature selection method and compare it with others, we use several criteria to measure instability and redundancy. In our experiments, we compare the proposed approach with several other feature selection methods, and show that the quadratic programming approach gives superior results according to the criteria considered for the test and real data sets.

1. Introduction

This paper presents a new approach to avoiding multicollinearity in feature selection. *Multicollinearity* is a strong correlation between features that affect the target vector simultaneously. In the presence of multicollinearity, common methods of regression analysis, such as least squares, build unstable models of excessive complexity. The formal definitions of model stability, complexity and redundancy are given in Section 5.

Most existing feature selection methods that solve the multicollinearity problem are based on heuristics (Leardi, 2001; Oluleye, Armstrong, Leng, & Diepeveen, 2014), greedy searches (Guyon, 2003; Ladha & Deepa, 2011) or regularization techniques (El-Dereny & Rashwan, 2011; Zou & Hastie, 2005). These approaches do not take into account the data set configuration and do not guarantee the optimality of the specially designed feature subset (Katrutsa & Strijov, 2015). In contrast, we propose a *quadratic programming approach* (Rodriguez-Lujan, Huerta, Elkan, & Cruz, 2010) to solving the multicollinearity problem that avoids disadvantages mentioned above. This approach is based on two ideas: representing feature presence as a binary vector, and defining the feature subset quality criterion in quadratic form. The first term of the quadratic form is the pairwise feature similarity, and the linear term is the relevance of features to the target vector. Therefore, we can state the feature selection problem with a quadratic objective function and a Boolean vector domain.

Measures of feature similarity and relevance are problem-dependent and need to be defined according to the application before performing feature selection. These measures should take into account the data set configuration to remove redundant, noisy and multicollinear features, selecting those that are significant for target vector approximation. We consider the correlation coefficient (Hall, 1999) and mutual information (Estévez, Tesmer, Perez, & Zurada, 2009) between features as measures of feature similarity and between features and the target vector as a measure of feature relevance. These measures guarantee a positive semidefinite quadratic form.

To solve the *convex optimization problem* we need to relax the binary domain to a continuous domain. This relaxation allows the *convex optimization problem* to be efficiently solved by state-of-the-art solvers such as CVX, a package for specifying and solving convex programs (Grant & Boyd, 2008; 2014). To translate the contin-

[*] Corresponding author.
E-mail addresses: aleksandr.katrutsa@phystech.edu (A. Katrutsa), strijov@ccas.ru (V. Strijov).

http://dx.doi.org/10.1016/j.eswa.2017.01.048
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

uous solution to a binary solution, we set a *significance threshold* that defines a number of features to be selected. If the feature similarity function does not give a positive semidefinite matrix, then the optimization problem is not convex and convex relaxation is required. In this case, we propose using a semidefinite programming relaxation (Naghibi, Hoffmann, & Pfister, 2015). Such feature similarity functions are out of the scope of this paper. In addition, the proposed approach gives a simple visualization of the feature weights in the target vector approximation. This visualization helps to tune the threshold.

We perform experiments on special test data sets generated according to the procedure proposed in Katrutsa and Strijov (2015). These data sets demonstrate different cases of multicollinearity between features and correlation between features and the target vector. Experiments show that the proposed approach outperforms the other feature selection methods considered on every type of test data set. Quadratic programming feature selection also gives better quality results on the test and real data sets according to various simultaneous evaluation criteria in contrast to other feature selection methods.

The main contributions of this paper are:

- addressing the multicollinearity problem with a quadratic programming approach and investigating its properties;
- evaluating the performance of the quadratic programming feature selection method on test data sets according to various criteria;
- comparing the proposed feature selection method with other methods on test and real data sets, and showing that the proposed method gives better feature subsets than the other methods. The feature subset quality is measured by external criteria.

1.1. Related works. A comprehensive survey of feature selection algorithms can be found in Li et al. (2016), which gives a systematic analysis of filter, wrapper, and embedded methods.[1] Various strategies have been proposed for detecting multicollinearity and solving the multicollinearity problem (Askin, 1982; Belsley, Kuh, & Welsch, 2005; Leamer, 1973). One way to solve the multicollinearity problem is to use feature selection methods (Belsley et al., 2005; Liu & Motoda, 2012). These are based on scoring functions that estimate the quality of a feature subset, or on a heuristic sequential search procedure.

This paper considers feature selection methods based on scoring functions, such as least angle regression (LARS) (Efron, Hastie, Johnstone & Tibshirani, 2004), Lasso (Tibshirani, 1994), Ridge (El-Dereny & Rashwan, 2011), and the Elastic Net (Zou & Hastie, 2005), and based on sequential search, such as Stepwise (Harrell, 2001) and the genetic algorithm (Ghamisi & Benediktsson, 2015). The Lasso scoring function is the weighted sum of the ℓ_2 norm of the residuals and the ℓ_1 norm of the parameter vector. This scoring function gives a good approximation to the target vector and penalizes large elements in the parameter vector. Moreover, the ℓ_1 norm of the parameter vector induces sparsity in the obtained parameter vector and therefore performs feature selection. The Ridge scoring function is the same as in Lasso, but uses the ℓ_2 norm instead of the ℓ_1 norm. This approach makes the solution more stable, but does not give a sparse parameter vector and selects features less aggressively than Lasso. The Elastic Net (Zou & Hastie, 2005) uses a linear combination of the ℓ_1 and ℓ_2 norms of the parameter vector as a penalty for the residual norm. This penalty allows us to combine the advantages of both Lasso and Ridge. Two common problems for these feature selection methods are tuning the weights corresponding to the penalty terms and taking into account the structure of a data set. A study of feature selection methods that use sequential search can be found in Aha and Bankert (1996). The genetic algorithm (Ghamisi & Benediktsson, 2015) uses a random search that maximizes the objective function and adds or removes some features on each iteration, while Stepwise starts from an empty feature set and sequentially adds a single feature on each interation according to the importance determined by an F-test.

2. Feature selection problem statement

Let $\mathbf{X} = [\boldsymbol{\chi}_1, \dots, \boldsymbol{\chi}_n] \in \mathbb{R}^{m\times n}$ be a design matrix, where $\boldsymbol{\chi}_j \in \mathbb{R}^m$ is the jth feature. Denote by $\mathcal{J} = \{1, \dots, n\}$ the feature index set, and let $\mathcal{A} \subseteq \mathcal{J}$ be a feature index subset. Let $\mathbf{y} \in \mathbb{R}^m$ be the target vector. The data fitting problem is to find a parameter vector $\mathbf{w}^* \in \mathbb{R}^n$ such that

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^n} S(\mathbf{w}, \mathcal{A}|\mathbf{X}, \mathbf{y}, \mathbf{f}), \quad (1)$$

where S is the error function, which validates the quality of the parameter vector $\mathbf{w}$ and the corresponding feature index subset $\mathcal{A}$ given a design matrix $\mathbf{X}$, a target vector $\mathbf{y}$ and a function $\mathbf{f}$. The function $\mathbf{f}$ approximates the target vector $\mathbf{y}$.

This study explores the linear function

$$\mathbf{f}(\mathbf{X}, \mathcal{A}, \mathbf{w}) = \mathbf{X}_{\mathcal{A}}\mathbf{w},$$

where $\mathbf{X}_{\mathcal{A}}$ is the reduced design matrix consisting of features with indices in $\mathcal{A}$, and the quadratic error function

$$S(\mathbf{w}, \mathcal{A}|\mathbf{X}, \mathbf{y}, \mathbf{f}) = \|\mathbf{f}(\mathbf{X}, \mathcal{A}, \mathbf{w}) - \mathbf{y}\|_2^2. \quad (2)$$

The features are assumed to be noisy, irrelevant or multicollinear, which leads to additional error in estimating the optimum vector $\mathbf{w}^*$ and increases the instability of this vector. Feature selection methods can be used to remove certain features from the design matrix $\mathbf{X}$. The feature selection procedure reduces the dimensionality of problem (1) and improves the stability of the optimum vector $\mathbf{w}^*$. The feature selection problem is

$$\mathcal{A}^* = \arg\min_{\mathcal{A}\subseteq\mathcal{J}} Q(\mathcal{A}|\mathbf{X}, \mathbf{y}), \quad (3)$$

where $Q : \mathcal{A} \to \mathbb{R}$ is a quality criterion that determines the quality of a selected feature index subset $\mathcal{A} \subseteq \mathcal{J}$. Problem (3) does not necessarily require estimation of the optimum parameter vector $\mathbf{w}^*$. It uses the relationships between the features $\boldsymbol{\chi}_j, j \in \mathcal{J}$ and the target vector $\mathbf{y}$.

Let $\mathbf{a} \in \mathbb{B}^n = \{0, 1\}^n$ be an indicator vector such that $a_j = 1$ if and only if $j \in \mathcal{A}$. Then problem (3) can be rewritten as

$$\mathbf{a}^* = \arg\min_{\mathbf{a}\in\mathbb{B}^n} Q(\mathbf{a}|\mathbf{X}, \mathbf{y}), \quad (4)$$

where $Q : \mathbb{B}^n \to \mathbb{R}$ is another form of the criterion Q with domain $\mathbb{B}^n$. The vector $\mathbf{a}^*$ and the index set $\mathcal{A}^*$ are related by

$$a_j^* = 1 \Leftrightarrow j \in \mathcal{A}^*, \ \ j \in \mathcal{J}. \quad (5)$$

2.1. Multicollinearity problem

In this subsection, we give a formal definition and some special cases of the multicollinearity problem. Assume that the features $\boldsymbol{\chi}_j$ and the target vector $\mathbf{y}$ are normalized:

$$\|\mathbf{y}\|_2 = 1 \text{ and } \|\boldsymbol{\chi}_j\|_2 = 1, \ \ j \in \mathcal{J}. \quad (6)$$

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

Definition 2.1. The features with indices in the set $\mathcal{A}$ are *multicollinear* if there exist an index j, coefficients λ_k, an index $k \in \mathcal{A} \setminus j$

[1] Implementations of several feature selection algorithms are available from a library developed by Arizona State University (http://featureselection.asu.edu).