Information Sciences 409–410 (2017) 68–86

Contents lists available at ScienceDirect

Information Sciences

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

Maximum relevance minimum common redundancy feature selection for nonlinear data

Jinxing Che[a,b,c], Youlong Yang[a,*], Li Li[c], Xuying Bai[a], Shenghu Zhang[d], Chengzhi Deng[b]

[a] School of Mathematics and Statistics, Xidian University, 266 Xinglong Section of Xifeng Road, Xi'an, Shaanxi, 710126, China
[b] Jiangxi Province Key Laboratory of Water Information Cooperative Sensing and Intelligent Processing, Nanchang Institute of Technology, China
[c] School of Science, Nanchang Institute of Technology, Nanchang, Jiangxi 330099, China
[d] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 20 January 2016
Revised 8 May 2017
Accepted 9 May 2017
Available online 10 May 2017

Keywords:
Feature selection
Mutual information
Normalization
Minimal common redundancy
Maximal relevance

ABSTRACT

In recent years, feature selection based on relevance redundancy trade-off criteria has become a very promising and popular approach in the field of machine learning. However, the existing algorithmic frameworks of mutual information feature selection have certain limitations for the common feature selection problems in practice. To overcome these limitations, the idea of a new framework is developed by introducing a novel maximum relevance and minimum common redundancy criterion and a minimax nonlinear optimization approach. In particular, a novel mutual information feature selection method based on the normalization of the maximum relevance and minimum common redundancy (N-MRMCR-MI) is presented, which produces a normalized value in the range [0, 1] and results in a regression problem. We perform extensive experimental comparisons over numerous state-of-art algorithms using different forecasts (Bayesian Additive Regression tree, treed Gaussian process, *k*-NN, and SVM) and different data sets (two simulated and five real datasets). The results show that the proposed algorithm outperforms the others in terms of feature selection and forecasting accuracy.

1. Introduction

Feature selection plays a critical role in regression systems, especially for nonparametric models [29,34]. Forecasting becomes unnecessarily complex or over-fitting if irrelevant or too many redundant attributes are included. Accordingly, to deal with large quantities of data, it is highly desirable to design a feature selection method which would include relevant features but exclude irrelevant or redundant ones [50]. To this end, we contribute this study to feature selection for nonlinear data, e.g., nonlinear classification or regression problems.

The feature selection problem for linear regression or simple parametric models has been studied extensively by different researchers and from different perspectives [9,14,39]. In this work, we consider feature selection for nonparametric relationships. Classical metaheuristic approaches consider feature selection as a discrete optimization problem with the so-

* Corresponding author.
E-mail addresses: jinxingche@163.com (J. Che), ylyang@mail.xidian.edu.cn, youlongyang2015@163.com (Y. Yang).

http://dx.doi.org/10.1016/j.ins.2017.05.013
0020-0255/

lution space spanning spanning all 2^m possible subsets from a candidate pool of m features, and require finding the best feature subset [5,19,47], where the objective functions of such discrete optimization problems are distinct in the context of individual evaluation criteria, such as the minimization of the sum-of-squares error in the regression. By retraining a neural network repeatedly, Setiono and Lui proposed a decision tree method to exclude irrelevant or redundant features one by one [40]. The method needs to retrain NN to explore almost every combination of feature subsets. Since 2^m is the total candidate number of feature subsets, the application of the method becomes extremely difficult to search the whole feature subset exhaustively for large values of m. This also brings such challenges as the management of the computational time complexity while extracting compact yet effective models [5]. A common way of overcoming these challenges is to use information measure based (e.g., feature subset selection) techniques.

In probability theory and information theory, the measure of dependence between two random variables is an important research topic [33], where the correlation coefficient and mutual information are the two major metrics. In fact, the mutual information (MI) between two variables X and Y measures how similar the joint distribution $p(X, Y)$ is to the product $p(X)p(Y)$ of the factored marginal distributions, which does provide a generalized measure of the variables' mutual dependence [1,26]. Specifically, such dependence, not limited to a linear dependence relationship such as the correlation coefficient, works for both linear and nonlinear cases. Accordingly, an MI-based algorithm is an attractive alternative to a correlation coefficient-based method.

For the feature selection problem, we aim to classify all candidate features into three subsets [6]: (1) the relevant feature subset which is required for any modeling tool; (2) the indifferent feature subset which consists of useless features that have bad effects on data analysis and increase the complexity of modeling; and (3) the redundant feature subset which includes features that are useful, but depend on relevant features, so that if we make some mistake in measuring a relevant feature, the predictor may work badly, but if the predictor is derived considering highly correlated redundant features, the mistakes can be avoided once a measurement error occurs in one of the relevant features. Based on the above analysis, it is desired to include certain redundant features to improve the robustness of the predictor. Accordingly, it is expected that any feature selection problem should consider [7]: (1) including relevant features; (2) excluding indifferent features; and (3) using redundant features.

To reduce the number of combinations, Battiti introduced a mutual information feature selector (MIFS) which makes use of mutual information between inputs and outputs [2], and demonstrated the effectiveness of mutual information in feature selection. Since then, many feature selection approaches based on the Max-Relevance and Min-Redundancy criterion have been proposed to improve the performance of feature selection, such as the NMIFS [20], MIFS-U [30] and mRMR [38], but these methods also have some limitations. For example, in most cases, only a part of the redundancy term is contained in the relevance term. Moreover, the NMIFS may produce a value outside the range [0, 1].

For this, conditional mutual information, which integrates both terms of relevance and redundancy, was introduced for feature selection, e.g., the Joint Mutual Information (JMI) [51], Double Input Symmetrical Relevance (DISR) [35], Conditional Mutual Information Maximization [22,42], JMIM and NJMIM [3]. However, two difficulties arise. First, accurate calculation of conditional mutual information is hard, due to the amount of calculations and the limited number of observations available for the evaluation of the 3-dimensional probability density function. Second, to extend the calculation of $I(X_i, X_j, Y)$ to $I(X_i, S, Y)$, these methods use either cumulative sum approximation or 'maximum of the minimum' approximation, which may exclude relevant features or include indifferent features.

Recently, some improved feature selection methods based on the max-relevance and min-redundancy (mRMR) criterion were proposed to further use information about both relevance and redundancy. Wang et al. built a multi-objective optimization problem by considering two objectives, namely, the maximization of relevance and minimization of redundancy, and performed an evolutionary algorithm for the feature selection process [46]. Using the information about already selected features, Chernbumroong et al. showed how a feature complements already selected features, and then established feature selection according to maximum relevancy and maximum complementary [10].

Inspired by these works, this paper proposes a new filter framework which introduces a novel criterion in terms of maximum relevance and minimum common redundancy (MRMCR), and results in a 'maximum of the minimum' nonlinear approach. It can properly select relevant features and control the use of redundant features, while discarding indifferent features at the same time. More specifically, to make the relevance and redundancy term comparable, common redundancy is first calculated to evaluate the common information of the candidate feature, already selected features and response feature. Then, a novel feature selection method based on the normalization of maximum relevance and minimum common redundancy (N-MRMCR-MI) is presented for the studied nonlinear optimization problem, which produces a normalized value in the range [0, 1], and further extends the NMIFS approach to the regression problem.

The key contributions of this paper are as follows: (i) the goal of any feature selection problem is introduced; (ii) the existing mutual information (MI) feature selection methods are classified into two frameworks, and their limitations are analyzed; and (iii) a new framework to attain the above goal is proposed, where a novel Max-Relevance and Min-Common-Redundancy criterion is developed to make relevance and redundancy comparable.

This paper is organized as follows. In Section 2, the basic principles of the information theory and forecast methods are briefly presented, including concepts such as mutual information, the Bayesian Additive Regression tree, and treed Gaussian process. In Section 3, a review of related works is provided, and then the limitations of previous approaches are discussed in detail. In Section 4, an improved feature selection criterion for regression and classification is proposed. In Section 5, the