Postprint: Feature Correlation-Based Partial Least Squares Feature Selection Method

Authors: Zeng Qingxia, Du Jianqiang, Zhu Zhipeng, Nie Bin, Yu Riyue, Yu Fang

Date: 2018-05-02T00:00:00+00:00

Abstract

To address the issues of traditional partial least squares methods that only consider the importance of individual features and suffer from redundancy and multicollinearity among features, statistical correlation among features is introduced into conventional partial least squares analysis to construct a feature-correlation-based partial least squares model. First, feature correlation is utilized to evaluate and pre-select feature subsets, which are then trained in the partial least squares model to assess the acceptability of the feature subset. Combined with a forward greedy search strategy, candidate features are evaluated sequentially, and the candidate feature that minimizes the objective function is selected and added to the set of selected features. The method was applied to analyze the cough-suppressing and asthma-relieving effects of the monarch drug in Maxing Shigan Decoction and UCI datasets, respectively. Experimental results demonstrate that this feature selection method can effectively identify superior feature subsets.

Full Text

Preamble

Journal: Computer Application Research (ChinaXiv Cooperative Journal)
Article: PLS Feature Selection Method Based on Feature Correlation
Authors: Zeng Qingxia[1], Du Jianqiang[1], Zhu Zhipeng[1], Nie Bin[1], Yu Riyue[2], Yu Fang[1]
Affiliations: [1]School of Computer Science; [2]School of Pharmacy, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, China

Abstract: The traditional partial least squares method only considers the importance of single features and suffers from redundancy and multicollinearity among features. To address these issues, this paper introduces statistical

correlation between features into traditional partial least squares analysis and constructs a PLS feature selection model based on feature correlation. First, feature relevance is used to evaluate and pre-select feature groups, which are then trained in the partial least squares model to assess their suitability. Combined with a forward greedy search strategy, candidate features are evaluated sequentially, and the candidate that minimizes the objective function is added to the selected features. Experiments were conducted using cough-relieving and asthma-relieving data from the monarch drug of Maxing Shigan Decoction and UCI datasets. The results demonstrate that this feature selection method can effectively identify optimal feature groups.

Keywords: Traditional Chinese Medicine information; partial least squares; feature correlation; feature selection

0 Introduction

With scientific advancement, the objects processed in data mining have become increasingly complex, and their dimensionality has grown dramatically. High dimensionality often leads to the "curse of dimensionality," where computational complexity increases significantly while classifier performance deteriorates sharply as dimensions increase. Consequently, dimensionality reduction is essential and can be achieved through two approaches: feature selection and feature extraction. Feature selection refers to the process of selecting an optimal feature subset from the original feature space that maximizes performance for a given task. It is a critical preprocessing step in pattern recognition, machine learning, and related fields. The primary goal is to select an optimal feature subset that removes irrelevant or redundant features without significantly reducing classification accuracy, thereby enhancing the discriminative power of the remaining features.

Evaluation criteria represent a key component of feature selection algorithms, encompassing distance measures, information measures, dependency measures, and consistency measures. Based on these criteria, feature selection methods can be categorized into three types: filter, wrapper, and embedded. Filter methods require a scoring function to evaluate feature relevance and a threshold to select the highest-scoring subset. While fast to train, they often exhibit significant bias relative to the performance of subsequent learning algorithms. Wrapper methods use the training accuracy of subsequent learning algorithms to evaluate feature subsets, offering lower bias but requiring substantial computational resources, making them unsuitable for large datasets. Embedded methods were developed primarily to address the high reconstruction costs of wrapper methods when processing different datasets. By integrating feature selection with the learning process of classification models, embedded methods achieve efficient spatiotemporal performance and good classification accuracy.

Partial Least Squares (PLS) is a multivariate regression modeling method pro-

posed for situations with high correlations among independent variables, effectively addressing multicollinearity issues. Leveraging this advantage, Li Jiangeng et al. proposed a feature selection method based on stepwise extraction of PLS principal components, repeatedly utilizing PLS to extract principal components and select genes with larger weights. Li Sheng et al. introduced an improved quantum genetic PLS feature selection method that assigns initial values to the population and designs a novel fitness function combined with PLS for feature selection. Nguyen et al. employed PLS as a dimensionality reduction method and used Logistic Discrimination (LD) and Quadratic Discrimination Analysis (QDA) algorithms to build classifiers for data classification.

Therefore, this paper proposes a PLS feature selection method based on feature correlation. The method utilizes feature relevance to evaluate and pre-select feature subsets, which are then trained in a PLS model to assess their suitability. Combined with a forward greedy search strategy, candidate features are evaluated sequentially, and the feature that minimizes the objective function is added to the selected set. This approach not only offers fast training speed and local optimality but also compensates for the drawbacks of wrapper methods, such as high computational cost and unsuitability for large datasets, thereby identifying superior feature subsets.

---

1 Correlation-Based Feature Selection

In 1999, Hall proposed the Correlation-based Feature Selection (CFS) method, a typical filter-based approach that heuristically evaluates the contribution of individual features to each class to obtain the final feature subset. CFS estimates feature subsets and ranks them rather than individual features. Its core employs a heuristic approach to evaluate the worth of feature subsets by calculating correlations between features and between features and class labels. The goal is to select features that are mutually uncorrelated yet highly correlated with the class label. The CFS heuristic equation is:

$$Merit_S = \frac{k \cdot \overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

where $Merit_S$ represents the 'merit' (class discriminative ability) of feature subset S containing k features, $\overline{r_{cf}}$ denotes the average correlation coefficient between class c and features f ($f \in S$), and $\overline{r_{ff}}$ is the average inter-feature correlation coefficient among features in S, with all variables requiring standardization. The numerator represents the class prediction capability of subset S, while the denominator indicates redundancy within S. Thus, a larger numerator signifies stronger class prediction ability, and a smaller denominator indicates less redundancy.

However, in CFS, features must be discrete random variables, and correlations are computed using conditional entropy and mutual information. This makes it difficult to handle continuous random variables. For continuous data, Pearson correlation coefficients can be used to calculate correlations between features and between features and class labels, where larger absolute values indicate stronger correlations and values approaching zero indicate weaker correlations.

1.2 Searching Feature Subset Space

CFS first computes feature-class and feature-feature correlation matrices from the training set, then employs a forward selection search strategy (FS) to explore the feature subset space. Alternative search methods include Best First Search (BFS) and Backward Elimination (BE). Forward selection starts with an empty set and greedily adds one feature at a time until no suitable feature can be added. Backward elimination begins with all features and greedily removes one feature at a time until the merit estimate no longer decreases. Best-first search is similar to the other two methods but can start from either an empty or full set. Starting from an empty set, it begins with no features selected and generates all possible single features, computes their merit values (represented by M), and selects the feature with the largest M value to enter the subset S. It then selects the second feature with the largest M value to enter S. If the M value of these two features is smaller than the original M value, the second feature is removed, and the process continues recursively to find the feature combination that maximizes merit.

---

2 PLS Feature Selection Based on Feature Correlation

2.1 PLS Regression Modeling Concept

Partial Least Squares Regression (PLS) is a novel multivariate statistical analysis method. Unlike traditional least squares regression, PLS regression investigates multivariate regression modeling of multiple dependent variables on multiple independent variables, particularly when variables exhibit multicollinearity or when sample size is smaller than the number of variables. Given independent variable set X and dependent variable set Y, to best summarize the original data information, the first component t_1 is extracted from X to maximize its variance, and the first component u_1 is extracted from Y to maximize its variance while maximizing the correlation between t_1 and u_1. Multivariate linear regression is then performed on t_1 and u_1 to obtain residual vectors, and the same method is applied iteratively. Cross-validation determines the number of principal components to extract in PLS regression, after which iteration stops and the PLS regression model is established.

2.2 Forward Selection Search Strategy Algorithm Based on PLSCF

Feature selection aims to choose a set of features that form a subset highly correlated with the class but with minimal inter-feature correlation. Thus, larger $Merit_S$ values indicate greater contribution of subset S to classification, representing a good feature subset. However, the CFS merit criterion only considers feature-class and feature-feature correlations without accounting for multicollinearity or incorporating model training effects. This paper proposes using the CFS metric to measure class discriminative ability of corresponding feature subsets and the Sum of Squares for Error (SSE) of the PLS regression model as the evaluation criterion for feature subset selection, termed the PLSCF evaluation criterion. The search strategy employs forward selection.

This algorithm combines the PLSCF feature evaluation criterion with forward selection search strategy. It first adds the single feature with the strongest class discriminative ability, then iteratively adds features that, when combined with already selected features, provide the strongest class discriminative ability. A floating component then determines whether to retain the added feature based on whether the SSE of the PLS model corresponding to the updated feature subset decreases. If SSE decreases after training on the current feature subset, the added feature is retained; otherwise, it is removed. This process repeats until all features have been tested. The features remaining in the subset constitute the selected optimal feature subset. The pseudocode description is presented in Algorithm 1.

Algorithm 1: PLSCF-Based Forward Selection Hybrid Feature Selection Algorithm

Input: Current training set and test set
Output: Feature subset C

Step 1: Preprocess the data
Step 2: Feature evaluation
Let F be the set of all features, and C be the selected feature subset, initially empty ($C = \emptyset$). While $S \neq \emptyset$, compute the discriminative ability of each feature on the training set and select the most important feature.

Step 3: Evaluate candidate features using forward selection search strategy
Train PLS using features in C to obtain a PLS prediction model. Record the SSE for training set (SSETrain) and test set (SSETest). If SSETrain preSSETrain, then remove the feature; otherwise, retain it.

Step 4: Repeat until termination conditions are met.

3 Experimental Results and Analysis

3.1 Experimental Data Description

The experimental data primarily consists of cough-relieving data (MXZK) and asthma-relieving data (MXPC) for Maxing Shigan Decoction from the Key Laboratory of Jiangxi University of Traditional Chinese Medicine, along with UCI datasets including Air Quality, CASP, Slump, Housing, and CCPP_Folds5x2_pp. The basic information for these datasets is shown in Table 1 .

3.2 Experimental Results and Analysis

To validate the feasibility and effectiveness of the proposed PLSCF feature selection method, experiments were conducted on seven datasets using Support Vector Machine (SVM), Correlation-based Feature Selection (CFS), and the proposed Partial Least Squares Feature Selection based on Feature Correlation (PLSCF), all employing forward selection search strategy. Data was randomly partitioned in a 7:3 ratio, with 70% used for training and 30% for testing. To obtain statistically meaningful results, model parameters were adjusted to achieve optimal performance, and the three algorithms were compared under the same training conditions.

The evaluation metrics included Sum of Squares for Error of training set (SSE-Train) and Sum of Squares for Error of test set (SSETest). The experimental results are presented in Table 2 .

According to Table 2, across the seven datasets, the SSETrain and SSETest values obtained by SVM and CFS algorithms with forward selection search strategy are similar, indicating comparable feature selection performance for these data types. For example, on the CCPP dataset, the training and test SSE values for both algorithms are 100.3872 and 112.4920, and 4.2302 and 6.5398, respectively. In contrast, the proposed PLSCF method with forward selection search strategy achieves significantly lower SSE values for both training and test sets. For instance, on the AQ dataset, the three algorithms yield test SSE values of 4.6118, 3.7188, and 0.2328, and training SSE values of 0.0385, 0.0894, and 0.0106, respectively. Similar improvements are observed on CASP, Housing, and CCPP datasets. However, on MXZK, MXPC, and Slump datasets, the differences are less pronounced. Notably, on the Slump dataset, CFS achieves a lower training SSE than PLSCF (0.3049 vs. 0.3091) but a higher test SSE (0.0312 vs. 0.03041). This variation arises because different datasets produce different experimental effects, and the selected feature subsets are not guaranteed to be globally optimal but rather near-optimal. SVM and CFS generally have shorter runtime than PLSCF because the PLSCF feature selection algorithm requires PLS during each feature evaluation, increasing computational time.

For more intuitive visualization, Figures 1 [Figure 1: see original paper] and 2 [Figure 2: see original paper] illustrate the fluctuations in SSETrain and

SSETest. Since different datasets have different magnitudes, the SSE values were centralized to [0,1] for comparison. The centralization formula is applied to both training and test SSE values to facilitate comparison across datasets at the same scale.

As shown in Figures 1 and 2, on the Slump dataset, PLSCF' s training SSE is higher than CFS but lower than SVM, indicating suboptimal performance compared to CFS. Its test SSE is much higher than SVM, suggesting slightly worse performance than SVM. However, except for the Slump dataset where improvement is less significant, PLSCF demonstrates clear performance enhancements across all other metrics, outperforming both SVM and CFS. This is because different datasets yield different experimental results, and feature selection subsets exhibit some randomness, preventing guaranteed global optimality but achieving near-optimal solutions. The results indicate that due to variations in experimental data, algorithm effectiveness differs accordingly. Nonetheless, the training and test SSE values for PLSCF show clear downward trends in other experimental datasets.

In summary, across the seven experimental datasets, the PLSCF method significantly outperforms SVM and CFS, though the improvement is less pronounced for certain datasets. This occurs because while the feature correlation-based PLS evaluation criterion selects representative feature subsets, different experimental data produce varying effects, suggesting that the selected subsets may not be globally optimal but are near-optimal solutions.

---

4 Conclusion

This paper addresses the limitations of traditional partial least squares methods that only consider single feature importance and suffer from redundancy and multicollinearity among features. By introducing statistical correlation between features into traditional PLS analysis, we propose a feature selection method based on feature-correlation partial least squares. This approach fully leverages the discriminative ability of feature subsets and combines it with PLS regression, which can model effectively with small samples while maximizing relationships between independent and dependent variables. Experimental comparisons on Traditional Chinese Medicine data and UCI datasets demonstrate that the proposed PLSCF method selects more representative feature subsets compared to SVM and CFS algorithms.

References

[1] Lin Yaojin, Li Jinjin, et al. Feature selection via neighborhood multi-granulation fusion [J]. Knowledge-Based Systems, 2014, 67(3): 162-168.
[2] Zhang Ce, Arun K, Christopherré, Materialization optimizations for feature selection workloads [J]. ACM Trans on Database Systems, 2016, 41(1): 2.

[3] Cao Jin, Zhang Li, Li Fanzhang. A feature selection algorithm based on support vector data description [J]. Journal of Intelligent Systems, 2015(2): 215-220.
[4] Zhu Z, Ong Y S, Dash M. Wrapper-filter feature selection algorithm using a memetic framework [J]. IEEE Trans on Systems Man & Cybernetics Part B: Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, 2007, 37(1): 70-76.
[5] Zhang X. A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine [J]. Geographical Research, 2008, 37(3): 71471J-71471J-9.
[6] Wang Xiang, Hu Xuegang. Survey on feature selection for high-dimensional small-sample classification problems [J]. Computer Applications, 2017, 37(9): 2433-2438.
[7] Shang Zhigang, Dong Yonghui, Li Mengmeng, et al. Robust feature selection and classification algorithm based on partial least squares regression [J]. Computer Applications Research, 2017, 37(3): 871-875.
[8] Li Jiangeng, Geng Tao, Ruan Xiaogang. Feature selection method based on stepwise extraction of partial least squares principal components [J]. Journal of Biology, 2010, 27(4): 85-87.
[9] Li Sheng, Zhang Peilin, Li Bing, et al. Application of improved quantum genetic partial least squares feature selection method [J]. Computer Engineering and Applications, 2017, 53(3): 242-252.
[10] Nguyen D V, Rocke D M. On partial least squares dimension reduction for microarray-based classification: a simulation study [J]. Computational Statistics & Data Analysis, 2004, 46(2): 407-425.
[11] Hall M A. Correlation-based feature selection for discrete and numeric class machine learning [C]// Proc of the 17th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2000.
[12] Saad Z S, Glen D R, Gang C, et al. A new method for improving functional-to-structural MRI alignment using local Pearson correlation [J]. Neuroimage, 2009, 44(3): 839-848.
[13] Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics [J]. Chemometrics & Intelligent Laboratory Systems, 2001, 58(2): 109-130.
[14] UCI machine learning repository [EB/OL]. [2016-07-18]. http://archive.ics.uci.edu/ml/.

Note: Figure translations are in progress. See original paper for figures.