Knowledge-Based Systems 55 (2014) 15–28

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

PLS-based recursive feature elimination for high-dimensional small sample

Wenjie You [a,b], Zijiang Yang [b], Guoli Ji [a,c,*]

[a] Department of Automation, Xiamen University, 361005 Xiamen, China
[b] School of Information Technology, York University, Toronto M3J 1P3, Canada
[c] Innovation Center for Cell Biology Research, Xiamen University, 361102 Xiamen, China

ARTICLE INFO

Article history:
Received 13 January 2013
Received in revised form 29 September 2013
Accepted 1 October 2013
Available online 14 October 2013

Keywords:
High-dimensional small samples (HDSS)
Partial least squares (PLS)
Recursive feature elimination (RFE)
Feature subset consistency
Feature subset compactness

ABSTRACT

This paper focused on feature selection for high-dimensional small samples (HDSS). We first presented a general analytical framework for feature selection on a HDSS including selection strategy (single-feature ranking and multi-feature ranking) and evaluation criteria (feature subset consistency and compactness). Then we proposed partial least squares (PLS) based feature selection methods for HDSS and two theorems. The proposed methodologies include a PLS model for classification, parameter selection, PLSRanking, and PLS-based recursive feature elimination. Furthermore, we compared our proposed methods with several existing feature selection methods such as Support Vector Machine (SVM) based feature selection, SVM-based recursive feature elimination (SVMRFE), Random Forest (RF) based feature selection, RF-based recursive feature elimination (RFRFE), ReliefF algorithm and ReliefF-based recursive feature elimination (ReliefFRFE). Using twelve high-dimensional datasets from different areas of research, we evaluated the results in terms of accuracy (sensitivity and specificity), running time, and the feature subset consistency and compactness. The analysis demonstrated that the proposed approach from our research performed very well when handling both two-category and multi-category problems.

1. Introduction

Classification is one of the main tasks of machine learning. The existence of a large number of redundant features weakens the generalization ability of classifiers [1]. Thus feature selection is an important part of machine learning. It aims at selecting a subset of key features from a problem and determining the source of the specific issues arising from the study. In recent years, a large number of high-dimensional and ultrahigh dimensional datasets [2–4] have appeared in applications such as microarray, image recognition, text categorization, and visual perception. More detailed information of objective phenomenon is provided with the increase of data dimensions. However, relative to thousands of features (even hundreds of thousands), the number of samples is usually small in many cases. The challenge is to create an effective mathematical model to deal with such a small sample that contains a large number of features.

Significant increase in the number of dimensions and the emergence of redundant features has made subsequent data analysis extremely difficult. Dimension growth has certainly led to a rapid increase in computation complexity. And more importantly, a small sample size of high dimensional data implies that statistical asymptotic properties are no longer guaranteed. The robustness of traditional methods is not reliable anymore. With an increase in the proportion of redundant features, multicollinearity occurs, which means that slight changes in input values will cause significant changes in learning outcomes. As a result, the existence of a large number of redundant features leads to degradation of a learner's generalization ability [3,4]. All of these are remarkable challenges to pattern recognition and knowledge discovery on high-dimensional small samples. Currently there are no particular data mining methods which are generally applicable to data with various characteristics. Many data mining algorithms lack efficiency or even fail in such cases [1]. The common method used to address the problem of high-dimensional small samples is to compress the dimension of their features.

One research direction in the field of pattern recognition and machine learning is to build effective feature selection methods for high-dimensional small samples. The objective of a feature selection process is to define metrics to retain the most effective features from the original features. This contains two aspects: (1) How to select the most effective feature subset from the original features? (2) How to choose appropriate evaluation criteria to determine the effectiveness of feature selection?

Effectiveness of feature selection has been usually measured by discrimination, generalization and reducing the degree of the

* Corresponding author.
E-mail addresses: zyang@yorku.ca, zyang@mathstat.yorku.ca (Z. Yang), glji@xmu.edu.cn (G. Ji).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.10.004

feature space dimension, or having a tradeoff among these scales. In fact, feature selection can be viewed as an optimization problem. The key is to establish evaluation criteria used to identify the feature subset that can help classification and has less redundancy. Different evaluation functions may give different results. According to the relationship between the evaluation function and the classifier, feature selection can be divided into three types: filter methods, wrappers methods and embedded approaches. In a filter method the evaluation function is independent of the classifier while wrapper methods usually use classification error probability as an evaluation function. Although using a filter method may be faster, the classification results from wrappers are usually better due to the fact that they are tuned to the specific interaction between a classification algorithm and its training data [5]. For embedded approaches, feature selection process is an integral part of a machine learning algorithm. Embedded approaches are usually more computationally tractable than wrapper methods. The computational complexity and extension to multiclass problems are major issues of embedded approaches when the number of features becomes excessively large [6]. Good feature selection should meet the following criteria [7]: (1) Take full account of the interaction between features; (2) Is based on the feature subset rather than individual features associated with classification; (3) Detect features with a relatively small main effect, but with a strong interaction effect [8]; (4) The feature selection algorithm should be reasonable and efficient. By definition, the selected feature subsets should contain a smaller number of features when compared to the set of all features.

This paper aims to provide an analytical framework as well as related theory for high-dimensional small samples. A partial least squares (PLS) model suitable for feature selection is proposed. First we show how to select parameters for the PLS. Next we present the PLS-based feature selection method (PLSRanking). Motivated by the SVMRFE analysis from Guyon et al. [9], we introduce a recursive feature elimination strategy on PLSRanking and propose PLS-based recursive feature elimination (PLSRFE). In order to evaluate the effectiveness of feature selection, we also define two metrics: feature subset consistency and feature subset compactness. The proposed method is compared with several state-of-the-art methods on multiple high-dimensional datasets. Compared to SVM-based feature selection, RF-based feature selection and the Relief (ReliefF) algorithm, our proposed PLS-based feature selection algorithm has several advantages: (1) It is computationally efficient, especially for high-dimensional dataset [8]; (2) It can be applied to both two-category and multi-category problems without limitation; (3) It can improve the consistency of the selected feature subset; (4) It makes the selected feature subset more compact.

The rest of the paper is organized as follows. A literature review is provided in Section 2 which discusses feature selection on high-dimensional small samples. Section 3 introduces PLS basics. Section 4 presents an analytical framework and theories to analyze high-dimensional small samples. PLSRanking and PLSRFE algorithms are also proposed in this section. The proposed model is tested with twelve high-dimensional datasets covering both two-category and multi-category problems. The test results and the comparison with the other existing methods are provided in Section 5. Conclusions and future work are discussed in Section 6.

2. Literature review

There is a vast amount of literature available on dimension reduction techniques for high-dimensional problems. Here we only focus on the feature selection methods relevant to our analysis.

The key issue of pattern recognition is how to select an optimal feature subset from the original dataset. Kohavi has proved that the optimal feature subset selection is a NP hard problem [10]. For high-dimensional small samples, feature ranking is an important method used to select an optimal feature subset. It is the trade-off between computational efficiency and classification performance. Instead of checking all possible feature subsets, Feature Ranking only ranks all the features based on defined heuristic rules. When reasonable evaluation criteria is applied, it can produce results that are as good as outputs from global optimal search strategies (branch and bound [11]), or random search strategies (genetic algorithm [1,12,13]). Relative to these stratifies; a key advantage of using feature ranking is its computational efficiency, which is particularly important for high-dimensional or ultrahigh-dimensional data.

Most of the traditional feature selection methods define some metrics to evaluate each individual feature, such as signal-to-noise ratio (SNR) [14] and information gain (InfoGain) [15]. They also utilize many statistical hypothesis testing techniques such as parametric *t*-test [16,17], *F*-test [18] and non-parametric Wilcoxon test (Mann–Whitney U test) [19,20] and their corresponding *p*-values [21]. The shortcoming of single feature ranking (univariate method) is that it ignores the correlation and non-linear relationship among the features. An ideal approach needs to consider the joint distribution among features. It must take all the features into account in order to detect those features with smaller impact on their own but strong interaction effects when combined with other features [8]. As a result, classification results based on single feature ranking are not satisfactory.

Multi-feature ranking methods take the correlation between features into account to a certain extent. SVM-based SVMRFE algorithm [9] and the Relief algorithm [22] based on iteratively adjusted margin are the best methods published so far for multi-feature ranking. However, the application of these two methods is limited as they can only handle two-category classification problems. The improved version of Relief algorithm, ReliefF [23], addressed this issue and can be used for multi-category classification and regression. Random Forest (RF), which is an ensemble learning method based on the decision tree classifier, is another algorithm that can be used to resolve multi-category problems [24]. RF can be used for variable importance analysis during data classification process. By analyzing feature importance, feature selection is implemented in Random Forests [25]. All these algorithms consider the correlation between features and the resulting feature subset is usually more compact.

For high-dimensional small samples, margin-based SVM algorithm [9,26] and their extensions, such as the two-stage SVM-RFE algorithm [27] and SVM-RFE with MRMR [28] are often used. Since SVM is originally designed for two-category classification, SVM-based multi-category classification is currently an important part of studies [26]. There are usually two solution strategies. One is using a decomposition algorithm such as one-versus-one (OVO), one-versus-all (OVA), directed acyclic graph (DAG), Support Vector Machine (DAG–SVM) and other methods [29–31]. The other solution is an overall approach which directly takes all categories into account in an optimization formula [32]. However, the overall approach is slow and classification accuracy is not satisfactory. As a result, OVO and OVA are the most often used methods in practice. OVO and OVA based multiclass SVM-RFE [33] and extensions of SVM-RFE are used in multiclass gene selection [31]. Relief and ReliefF [23] can also be used for this type analysis. But one key issue of Relief and ReliefF is that these algorithms cannot remove redundant features. All features with high correlation to a category are given a higher weight regardless whether they are redundant to the rest of the features. Global objective function is not optimized. The Iterative Relief (I-Relief) method [34] designs a margin-based global objective function which optimizes the feature weights similar to EM algorithm. The weakness of this method is that it is not robust for different datasets [35].