Information Sciences 291 (2015) 1–18

Contents lists available at ScienceDirect

Information Sciences

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

Robust twin boosting for feature selection from high-dimensional omics data with label noise

Shan He [a], Huanhuan Chen [a], Zexuan Zhu [b,*], Douglas G. Ward [c], Helen J. Cooper [d], Mark R. Viant [d], John K. Heath [d], Xin Yao [a]

[a] CERCIA, School of Computer Science, University of Birmingham, UK
[b] College of Computer Science and Software Engineering, Shenzhen University, China
[c] School of Cancer Sciences, University of Birmingham, UK
[d] School of Biosciences, University of Birmingham, UK

ARTICLE INFO

Article history:
Received 18 June 2013
Received in revised form 2 August 2014
Accepted 22 August 2014
Available online 30 August 2014

Keywords:
Feature selection
Boosting
Ensemble learning

ABSTRACT

Omics data such as microarray transcriptomic and mass spectrometry proteomic data are typically characterized by high dimensionality and relatively small sample sizes. In order to discover biomarkers for diagnosis and prognosis from omics data, feature selection has become an indispensable step to find a parsimonious set of informative features. However, many previous studies report considerable label noise in omics data, which will lead to unreliable inferences to select uninformative features. Yet, to the best of our knowledge, very few feature selection methods are proposed to address this problem. This paper proposes a novel ensemble feature selection algorithm, robust twin boosting feature selection (RTBFS), which is robust to label noise in omics data. The algorithm has been validated on an omics feature selection test bed and seven real-world heterogeneous omics datasets, of which some are known to have label noise. Compared with several state-of-the-art ensemble feature selection methods, RTBFS can select more informative features despite label noise and obtain better classification results. RTBFS is a general feature selection method and can be applied to other data with label noise. MATLAB implementation of RTBFS and sample datasets are available at: http://www.cs.bham.ac.uk/~szh/TReBFSMatlab.zip.

1. Background

Omics technologies such as genomics, proteomics and metabolomics have become powerful tools for analyzing biological systems. In medical science, omics technologies have been applied to discover molecular signatures as biomarkers for disease diagnosis, prognosis and staging. Omics data are usually acquired on small number of patient samples, typically tens to a few hundred in each disease group, but each sample data often contains tens of thousands of variables or features.

Feature selection plays a crucial role in omics data analysis. Feature selection methods can select a subset, usually a parsimonious subset of important features for building fast and robust learning models with better generalization capability, therefore improve the learning accuracy and interpretability of the results [29,72,64]. Current approaches to feature selection include: filter methods [68], wrapper methods [15,56], embedded methods [29], and hybrid methods [74,73,72,12,25].

* Corresponding author.
E-mail addresses: s.he@cs.bham.ac.uk (S. He), h.chen@cs.bham.ac.uk (H. Chen), zhuzx@szu.edu.cn (Z. Zhu), d.g.ward@bham.ac.uk (D.G. Ward), h.j.cooper@bham.ac.uk (H.J. Cooper), m.viant@bham.ac.uk (M.R. Viant), j.k.heath@bham.ac.uk (J.K. Heath), x.yao@cs.bham.ac.uk (X. Yao).

http://dx.doi.org/10.1016/j.ins.2014.08.048
0020-0255/

Filter methods evaluate the goodness of features based on the intrinsic characteristic of the data and without any consideration of the learning models. They are computationally fast, but they could select features not suitable for the learning models. Contrarily, wrapper methods involve the learning models and evaluate features directly according to the learning performance. They tend to obtain better learning performance at the price of computation cost. Embedded methods, which could be seen as more efficient wrapper methods, perform feature selection while building the learning model, i.e., feature selection and learning model are optimized simultaneously. Hybrid methods try to take advantage of both filter and wrapper methods in some specific hybridization frameworks. The focus of this study is on the development of a generic embedded feature selection method by performing classification and feature selection simultaneously, for various omics data based on boosting algorithms.

When feature selection is applied to omics data, one problem is the fact that there is considerable noise in the omics data. Firstly, there is data noise arising from biological variation and analytical variance. More crucially, many studies reported that label noise [31] is also common in omics data sets [45,70,7]. Label noise might be introduced by false diagnosis, which usually occurs near the decision boundary of the feature space. Such noise is called mislabeling. It is also very likely during decision making, experts may introduce mislabeling because they are distracted. Another reason is that in some heterogeneous disease such as cancer, subgroups may behave differently – a subgroup might only be one or a few individuals in these small studies and would appear to be outliers, which would introduce significant label noise. Label noise will lead to unreliable inferences from the omics data and consequently result in the selection of unreliable features. Such a problem presents significant challenges to feature selection algorithms to extract reliable combination of biomarkers from omics data with high discriminant power.

The label noise problem has been recognized by machine learning research community. In the past few years, several studies on label noise problem have been proposed [22,45,70,39,7,71,27,55,21,67,37]. However, most of the studies focus on designing classifiers that are robust to label noise. The only exception is [21] which proposes a novel feature selection method to select reliable features from data with label noise. The method can robustly evaluate the mutual information [63] between features and labels using a probabilistic label noise model and a nearest neighbors-based entropy estimator, then a backward greedy search procedure is applied to search for relevant sets of features based on the mutual information. The method is essentially a filter based method, therefore it ignores the interaction with the classifier, a general drawback of filter based methods as mentioned above, which may lead to worse classification performance compared with wrapper and embedded feature selection methods. Apart from this drawback, methods based on mutual information such as in [21] "does not always guarantee to decrease the misclassification probability" [20], which may also lead to worse classification performance.

To address the challenge of selecting reliable features from data with label noise, a novel embedded feature selection algorithm called robust twin boosting feature selection (RTBFS) is proposed in this study. The RTBFS algorithm is based on the twin boosting framework, which is a novel ensemble feature selection algorithmic framework proposed in [8]. When applied to several omics datasets with known label noise, the standard twin boosting algorithm in [8] could not select features of satisfactory performance (discussed in Section 2). Therefore, in this study, the twin boosting framework is improved for better robustness to label noise by incorporating a novel robust loss function, i.e., robust eta-loss [36] and a robust weak learner, i.e., robust componentwise linear least squares.

In the experimental study, the performance of RTBFS is firstly thoroughly investigated using a feature selection test bed based on a real-world microarray dataset of which the optimal features are known. In order to evaluate the robustness of RTBFS against label noise, RTBFS is tested on the contaminated test bed to which different percentage of label noise is deliberately introduced. And then the RTBFS algorithm is applied to seven publicly available real-world omics datasets including three microarray transcriptomic data, four Mass Spectrometry Time-Of-Flight (MS-TOF) proteomic data and one Ion Molecule Reaction MS (IMR-MS) metabolomic data. Among these datasets, the three microarray datasets have been reported to have label noise [45,70,39,7]. It is also worth mentioning that, while the label noise problem has received attention in genomic research community, it was until recently researchers in other omics, e.g., proteomics research communities started to discuss this issue using simulated data [50].

For comparison purpose, several twin boosting variants and several state-of-the-art feature selection algorithms, e.g., a fast correlation based feature selection algorithm [69] and an ensemble embedded feature selection algorithm [1] based on the popular support vector machine recursive feature elimination (SVM-RFE) [30] are implemented. Apart from the algorithms mentioned above, the results of RTBFS are also compared with the results from several novel feature selection algorithms published in literature recently. The results show that RTBFS can robustly extract parsimonious informative features from noisy omics data that generate better classification models, which is highly desirable for clinical omics data analysis where accuracy and interpretability are important.

2. Boosting based feature selection and label noise problem

Boosting is a type of ensemble machine learning method, which constructs a set of classifiers to classify new data points in some way (typically by weighted or unweighted voting of their predictions) [16]. Since the publication of Adaboost [23], boosting algorithms have attracted much attention in the machine learning and statistics communities due to their simplicity and competitive prediction accuracy. Boosting is also a good feature selection scheme especially for high-dimensional