Computational and Structural Biotechnology Journal 16 (2018) 88–97

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Mini Review

# A Review of Matched-pairs Feature Selection Methods for Gene Expression Data Analysis

Sen Liang [a], Anjun Ma [b,c], Sen Yang [a], Yan Wang [a,*], Qin Ma [b,c,**]

[a] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China
[b] Bioinformatics and Mathematical Biosciences Lab, Department of Agronomy, Horticulture and Plant Science, Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007, USA
[c] BioSNTR, Brookings, SD, USA

ARTICLE INFO

*Article history:*
Received 18 September 2017
Received in revised form 14 February 2018
Accepted 19 February 2018
Available online 25 February 2018

*Keywords:*
Matched-pairs feature selection
Matched case-control design
Paired data
Gene expression

ABSTRACT

With the rapid accumulation of gene expression data from various technologies, e.g., microarray, RNA-sequencing (RNA-seq), and single-cell RNA-seq, it is necessary to carry out dimensional reduction and feature (signature genes) selection in support of making sense out of such high dimensional data. These computational methods significantly facilitate further data analysis and interpretation, such as gene function enrichment analysis, cancer biomarker detection, and drug targeting identification in precision medicine. Although numerous methods have been developed for feature selection in bioinformatics, it is still a challenge to choose the appropriate methods for a specific problem and seek for the most reasonable ranking features. Meanwhile, the paired gene expression data under matched case-control design (MCCD) is becoming increasingly popular, which has often been used in multi-omics integration studies and may increase feature selection efficiency by offsetting similar distributions of confounding features. The appropriate feature selection methods specifically designed for the paired data, which is named as matched-pairs feature selection (MPFS), however, have not been maturely developed in parallel. In this review, we compare the performance of 10 feature-selection methods (eight MPFS methods and two traditional unpaired methods) on two real datasets by applied three classification methods, and analyze the algorithm complexity of these methods through the running of their programs. This review aims to induce and comprehensively present the MPFS in such a way that readers can easily understand its characteristics and get a clue in selecting the appropriate methods for their analyses.


**Contents**



* Corresponding author.
** Correspondence to: Q. Ma, Bioinformatics and Mathematical Biosciences Lab, Department of Agronomy, Horticulture and Plant Science, Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007, USA.
*E-mail addresses:* wy6868@jlu.edu.cn (Y. Wang), qin.ma@sdstate.edu (Q. Ma).

https://doi.org/10.1016/j.csbj.2018.02.005

## 1. Introduction

During the last two decades, feature selection techniques have become an active and fruitful research field in machine learning [1–4], pattern recognition [5,6], and bioinformatics [7–9]. Feature selection, a.k.a. Variable selection or gene selection (in bioinformatics), is the process of selecting a subset of relevant features for model construction or interpretation of results. It improves model predictive accuracy and reduces model complexity by eliminating irrelevant and redundant features and provides a better understanding of the underlying processes [10]. Many novel methods have been proposed recently, such as the minimum-Redundancy-Maximum-Relevancy (mRMR) method proposed by Peng et al. which selects features using mutual information as a proxy for computing relevance and redundancy among features [11], and the Max-Relevance-Max-Distance (MRMD) method proposed by Zou et al. that selects features with strong correlation with labeled and lowest redundancy features subset [12]. With the rapid expansion of gene expression data, higher gene dimensionality has been generated in limited samples. The feature selection techniques are playing more and more pivotal roles in high-dimensional data analyses, especially in gene function enrichment analysis, cancer biomarker detection, and drug targeting identification in precision medicine. Recently, Zou et al. proposed a new method to predict TATA-binding proteins with feature selection and dimensionality reduction strategy [13]. Tang et al. proposed novel selection strategies to identify highly tissue-specific CpG sites and then constructed classifiers to predict primary sites of tumors [14].

However, it is still a challenge to choose the appropriate methods for specific problems and retrieve the most reasonable ranking features in gene expression data analysis. Nowadays, using the existing next-generation sequencing techniques, such as microarray and RNA-seq, developed for gene expression profiling, the paired gene expression data under matched case-control design (MCCD) is becoming increasingly popular. Such data has frequently been used in multi-omics studies and may increase the feature selection efficiency by offsetting similar distributions of confounding features [15]. Nevertheless, the appropriate feature selection methods specifically designed for paired data accounting on MCCD, which is so-called matched-pairs feature selection (MPFS), have not been maturely developed in parallel.

There are many popular MPFS methods and strategies for bioinformatics research. Several studies have been managed to account for paired data in their algorithms, which can be categorized into three groups. First, the test statistic uses original and modified paired *t*-test to rank relevant features by evaluating significant levels which is often followed by a classification approach to improve model predictive accuracy. Such kind of methods is comparatively time-consuming and may return a preliminary feature selection results. Second, the conditional logistic regression (CLR) [16] is a modeling approach widely be used in MCCD studies to identify features significantly associated with case-control status. CLR has considerations of the interaction between features and make a better selection results when potential correlations exist. Third, the boosting strategy addresses classification problems with matched case-control responses. In machine learning, boosting is usually combined with many weak classifiers to build a powerful committee. Since Friedman et al. [17] described boosting as a method for the additive model using an exponential loss criterion, researchers employed boosting to identify significant features with paired data within a classifier task [18]. The boosting strategy is more powerful and time-consuming, which always need to be wrapped with a classifier, e.g., support vector machines (SVM) [19].

This review provides a survey of existing MPFS methods and applications for paired gene expression data under MCCD. Two real gene expression datasets from The Cancer Gene Atlas database (TCGA) [20] and Gene Expression Omnibus database (GEO) [21] were selected to evaluate the performance of MPFS methods and traditional unpaired feature selection methods. The rest of the paper is organized as follows: Section 2 introduces the feature selection techniques in general and presents overall classification strategies according to different data properties. In Section 3, the MPFS problem is defined and then the existing MPFS methods are summarized according to the above three feature selection groups. In Section 4, we compare the performance of ten methods, including eight MPFS methods and two traditional unpaired methods on the two real datasets and three classification methods, i.e., SVM, Gaussian Naïve Bayesian (GNB) [22], and Logistic Regression [23]. The running times of these methods are also recorded simultaneously as another vital criterion to help readers select the appropriate method for different environments. We further discuss several challenges for the development of the MPFS techniques and their further applications in many other bioinformatics research fields in Section 5. Finally, the conclusions are clearly drawn in the last section.

## 2. Feature Selection Techniques

The most acceptable benefit of feature selection is to help improving accuracy and reducing model complexity, as it can remove redundant and irrelevant features to reduce the input dimensionality and help biologists identify the underlying mechanism that connects gene expression with diseases or interested phenotype.

Feature selection techniques have been successfully applied in many real-world applications, such as large-scale biological data analysis [24–26], text classification [27], information retrieval [28], near-infrared spectroscopy [29], mass spectroscopy data analysis [30], drug design [31,32], and especially the quantitative structure-activity relationship (QSAR) modeling [33,34]. In cancer research community, feature selection has also been widely applied in different omics data analyses: mRNA data [9,35], miRNA data [36,37], whole exome sequencing data [38], DNA-methylation data [39,40], and proteomics data [41,42]. Recently, some researchers have applied feature selection techniques on integrative analysis of multi-omics data. Chen et al. reviewed multivariate dimension reduction approaches which can be applied to the integrative exploratory analysis of multi-omics data [43]; Mallik et al. developed a new feature selection framework for identifying statistically significant epigenetic biomarkers using maximal-relevance and minimal-redundancy criterion based on multi-omics dataset [44]; and Liu et al. [45] developed two methods based on the proportional hazards regression [46], named SKI-Cox and wLASSO-Cox approaches, to perform feature selection on different multi-omics datasets.

### 2.1. Unpaired Feature Selection Methods

It is not trivial to choose the appropriate feature selection method for a given scenario, hence, several classification strategies of unpaired feature selection techniques have been approached. The most widely-used classification strategy classified the methods into the filter, wrapper and embedded, based on the integrated classifiers [7,10,47]. The filter approach separates feature selection from classifier construction and assesses the relevance of features only relying on the intrinsic properties of data [48,49], which have frequently been used in