Neurocomputing 92 (2012) 36–43

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Biomarker discovery using 1-norm regularization for multiclass earthworm microarray gene expression data

Xiaofei Nan [a,*], Nan Wang [b], Ping Gong [c], Chaoyang Zhang [b], Yixin Chen [a], Dawn Wilkins [a]

[a] Department of Computer and Information Science, University of Mississippi, University, MS 38677, USA
[b] School of Computing, University of Southern Mississippi, Hattiesburg, MS 39406, USA
[c] SpecPro Inc., Environmental Services, Vicksburg, MS 39180, USA

ARTICLE INFO

Available online 22 February 2012

Keywords:
Biomarker discovery
1-norm regularization
Multiclass classification
Microarray gene expression data

ABSTRACT

Novel biomarkers can be discovered through mining high dimensional microarray datasets using machine learning techniques. Here we propose a novel recursive gene selection method which can handle the multiclass setting effectively and efficiently. The selection is performed iteratively. In each iteration, a linear multiclass classifier is trained using 1-norm regularization, which leads to sparse weight vectors, i.e., many feature weights are exactly zero. Those zero-weight features are eliminated in the next iteration. The empirical results demonstrate that the selected features (genes) have very competitive discriminative power. In addition, the selection process has fast rate of convergence.

1. Introduction

Discovery of novel biomarkers is one of the most important impetuses driving many biological studies including biomedical research. In this post-genomics era, many high throughput technologies such as microarrays have been applied to measure a biological system ranging from cell to tissue to whole animal. In the last decade, environmental scientists, particularly ecotoxicologists, have increasingly applied omics technologies in the hunt for biomarkers that display both high sensitivity and specificity. However, it is still a big challenge to sieve through high dimensional data sets and look for biomarker candidates that meet high standards and sustain experimental validation.

Previously, we developed an integrated statistical and machine learning (ISML) pipeline to analyze a multiclass earthworm gene expression microarray dataset [15]. As a continuation to this effort of biomarker discovery, here we developed a new feature selection method based on 1-norm regularization.

In machine learning, feature selection is a technique of seeking the most representative subset of features. It is the focus of research in applications where datasets have dramatic amounts of variables, e.g. text processing and gene expression data analysis. When applied to gene expression array analysis, the technique detects the influential genes by which biological researchers could discriminate normal instances from abnormal ones, and therefore, facilitates further biological research or judgments.

We only focus on supervised learning in this paper, which means a label is given for each instance. Unsupervised and semi-supervised learning could be found in other literatures [25,16,22]. Feature selection algorithms roughly fall into two categories, variable ranking (or feature ranking) and variable subset selection [9]. The latter essentially is divided into wrapper, filter, and embedded methods.

Variable ranking acts as a preprocessing step or auxiliary selection mechanism because of its simplicity and scalability. It ranks *individual features* by a metric, e.g. correlation, and eliminates features that do not exceed a given threshold. Variable ranking is computational efficient because it only computes feature scores. Nevertheless, this method only focuses on the predictive power of individual features. It is prone to the selection of redundant features.

Variable subset selection, on the other hand, attempts to select *subsets* of features that, jointly, produce good prediction performance. Filter methods consist of using a feature subset relevance criterion to yield a reduced subset of features which may be used for future prediction. Wrapper methods [13] search through feature subset space. Each subset is applied to a certain machine learning model and assessed by the learning performance. In these methods, learning models act as black boxes. Embedded approaches [14] implement feature selection in the process of learning. While wrapper methods search the space of all feature subsets, the searching step in embedded methods is guided by the learning algorithm. This guidance could be obtained from estimating changes in the objective function by adding or removing

[*] Corresponding author.
E-mail addresses: xiaofei.nan@gmail.com, xnan@olemiss.edu (X. Nan), nan.wang@usm.edu (N. Wang), Ping.Gong@usace.army.mil (P. Gong), Chaoyang.zhang@usm.edu (C. Zhang), ychen@cs.olemiss.edu (Y. Chen), dwilkins@cs.olemiss.edu (D. Wilkins).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.09.035

features. For example, Guyon et al. [10] proposed Support Vector Machine Recursive Feature Elimination (SVM-RFE) algorithm to recursively classify the instances by the SVM classifier and eliminate the feature(s) with the least weight(s). The number of features to be eliminated in each iteration is ad hoc. Moreover, there is no firm conclusion about when to terminate the recursive steps.

Most feature selection in the literature is designed for binary problems. When extended to real-life multiclass tasks, combining several binary classifiers are typically suggested, such as one-versus-all and one-versus-one [23]. For situations with k classes, one-versus-all constructs k binary classifiers, each of which is trained with all the instances in a certain class with positive labels and all other examples with negative labels. It is computationally expensive and has highly imbalanced data for each binary classifier. On the other hand, one-versus-one method constructs $k(k-1)/2$ binary classifiers for all pairs of classes. An instance is predicted for the class with the majority vote. Similar to the one-versus-all approach, the one-versus-one approach has heavy computational burden. Platt et al. [20] proposed a directed acyclic graph SVM (DAGSVM) algorithm whose training phase is the same as one-versus-one by solving $k(k-1)/2$ binary problems. However, DAGSVM uses a rooted acyclic graph to make a decision from $k(k-1)/2$ prediction results. Some researchers proposed methods solving multiclass tasks in one step: build a piecewise separation of the k classes in a single optimization. This idea is comparable to the one-versus-all approach. It constructs k classifiers, each of which separates a class from the remaining classes, but all classifiers are obtained by solving one optimization problem. Weston and Watkins [27] proposed a formulation of the SVM that enables a multiclass problem. But, solving multiclass problem in one step results in a much larger scale optimization problem. Crammer and Singer [5] decomposed the dual problem into multiple optimization problems of reduced size and solved them by a fixed-point algorithm. A comparison of different methods for multiclass SVM was done by Hsu and Lin [12].

A multiclass optimization cost function typically comprises two parts, empirical error and model complexity. The model complexity is usually approximated by a regularizer, e.g. 2-norm or 1-norm [3]. The use of 1-norm was advocated in many applications, such as multi-instance learning [4], ranking [19] and boosting [6], because of its sparsity-favoring property. Several literatures discussed the multiclass problem based on 1-norm regularization and various loss functions for the empirical error. For example, Friedman et al. [8] introduced 1-norm into multinomial logistic regression which is capable of handling multiclass classification problems. Bi et al. [2] chose ϵ-insensitive loss function. Liu and Shen [17] defined a specific loss function ψ-loss that replaces the convex SVM loss function by a nonconvex function. Other works mainly used hinge loss with different variations [24,26]. In this paper, the hinge loss function we apply is similar to that in [27], but has not be used in any 1-norm multiclass work.

Feature selection under the framework of 1-norm multiclass regularization is achieved by discarding the least significant features, i.e., features with zero weights. The sparsity of the weights is determined by a regularization parameter that controls the trade off between empirical error and model complexity. However, the selection of a proper regularization parameter is a challenging problem. We only know the trend of tuning the parameter to make the number of selected features smaller or larger, but it is difficult to associate a parameter value with a particular feature subset and at the same time achieve a high learning performance, unless the entire regularization path is computed. As 1-norm is non-differentiable (so is hinge loss), calculating the accurate regularization path is difficult (some other loss functions, such as logistic loss, have defined gradients). Even though the regularization path is piecewise linear, path-following methods are slow for large-scale problems. Instead of computing an approximate regularization path, we introduce an iterative 1-norm multiclass feature selection method that selects a small number of features with high performance.

In this paper, we propose a multiclass 1-norm regularization feature selection method, L1MR (Linear 1-norm Multiclass Regularization), and its simple variation SL1MR, that solve a single linear program. An iterative feature elimination framework is proposed to obtain a minimum feature subset. The sparsity favoring property of 1-norm regularization enables fast convergence of the iterative feature elimination process. In our empirical studies, the algorithm typically converges in no more than ten iterations. The reminder of the paper is organized as follows. Section 2 proposes the 1-norm multiclass regularization. Section 3 describes the iterative feature elimination process. Section 4 demonstrates the experimental results. Conclusions are presented in Section 5 along with a discussion of future work.

2. Learning a multiclass linear classifier via 1-norm regularization

Consider a set of l instances (X,Y) from an unknown fixed distribution, where $X \in \mathbb{R}^n$ is the earthworm microarray gene expression data, and output Y is the class label. In a k-category classification task, y is coded as $\{1,\ldots,k\}$. For the earthworm data studied in this article, $k=3$ (control, TNT, RDX), $n=869$ and $l=248$.

Given k linear decision functions $f_1,\ldots,f_k$ where f_c corresponds to class c, each decision function is defined as $f_c(x)=w_c^T x+b_c$, $c=1,\ldots,k$, where, the parameters $w_c=[w_{c,1},\ldots,w_{c,n}]^T \in \mathbb{R}^n$ and $b_c \in \mathbb{R}$. We consider a winner-takes-all classification rule specified as

$$\Phi(x)=\arg\max_c f_c(x),$$

which assigns input x to class $\Phi(x)$ with the highest decision value. If the instances are separable, there exist decision functions that satisfy

$$f_{y_i}(x_i) \ge f_c(x_i), c \ne y_i.$$

The above inequality is equivalent to

$$(w_{y_i}-w_c)x_i+(b_{y_i}-b_c) \ge 1, c \ne y_i.$$

To handle the non-separable cases, we introduce slack variables into the model, i.e.,

$$(w_{y_i}-w_c)x_i+(b_{y_i}-b_c) \ge 1-\xi_{y_i,c}, c \ne y_i, \tag{1}$$

where $\xi_{y_i,c} \ge 0$ is a slack variable.

The learning of the classifier can be formulated as an optimization problem. Our goal is to seek f that minimizes the sum of empirical error and model complexity. Empirical error can be computed as the proportion of non-separable instances, i.e., errors on training data, or approximated using various loss functions, such as hinge loss, logistic loss. These loss functions were originally defined for binary classes. Extending to multiclass case, there are different variations. For example, [26] utilized hinge loss function $\sum_{c \ne y_i}[f_c(x_i)+1]_+$, where $(\cdot)_+ \equiv \max(\cdot,0)$. In this paper, we apply the hinge loss function:

$$\sum_{c \ne y_i}[1-(f_{y_i}(x_i)-f_c(x_i))]_+,$$

which was introduced by Weston and Watkins [27]. This hinge loss function has not been used in the multiclass setting with 1-norm penalty before. And it is equivalent to the slack variable defined in (1).

Model complexity is approximated using 1-norm. Zhu et al. [29] argued that the 1-norm regularization yields sparse results,