Talanta 132 (2015) 175–181

Contents lists available at ScienceDirect

Talanta

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

Dealing with heterogeneous classification problem in the framework of multi-instance learning

Zhaozhou Lin [a], Shuaiyun Jia [a], Gan Luo [a], Xingxing Dai [a], Bing Xu [a], Zhisheng Wu [a], Xinyuan Shi [a,b,*], Yanjiang Qiao [a,b,*]

[a] College of Chinese Medicine, Beijing University of Chinese Medicine, Beijing 100102, China
[b] Research Center of TCM-information Engineering, State Administration of Traditional Chinese Medicine of the People's Republic of China, Beijing 100102, China

ARTICLE INFO

Article history:
Received 21 April 2014
Received in revised form
28 August 2014
Accepted 3 September 2014
Available online 16 September 2014

Keywords:
Heterogeneous spectra
Multi-instance learning (MIL)
Error-Correcting Output Codes (ECOC)
Maximum count least square support vector machine (maxc-LS-SVM)
Geographical origins

ABSTRACT

To deal with heterogeneous classification problem efficiently, each heterogeneous object was represented by a set of measurements obtained on different part of it, and the heterogeneous classification problem was reformulated in the framework of multi-instance learning (MIL). Based on a variant of count-based MIL assumption, a maximum count least squares support vector machine (maxc-LS-SVM) learning algorithm was developed. The algorithm was tested on a set of toy datasets. It was found that maxc-LS-SVM inherits all the sound characters of both LS-SVM and MIL framework. A comparison study between the proposed approach and the other two MIL approaches (i.e., mi-SVM and MI-SVM) was performed on a real wolfberry fruit spectral dataset. The results demonstrate that by formulating the heterogeneous classification problem as a MIL one, the heterogeneous classification problem can be solved by the proposed maxc-LS-SVM algorithm effectively.

1. Introduction

The application of near infrared spectroscopy (NIR) to perform classification has been spread across the analysis of food, agricultural, petroleum, and pharmaceutical products [1–3]. However, spectra obtained on heterogeneous objects, such as corn and pharmaceutical tablet, are often of high variance. Applying common classification methods on these spectra will result in weak conclusion. This is one typical category of heterogeneous classification problem. However, the heterogeneous classification problem has not yet been solved thoroughly [4].

The key on solving heterogeneous classification problem is to represent each heterogeneous object efficiently. In the literatures, pre-treatment is the most successfully and widely used techniques for solving the heterogeneous classification problem. One builds certain measurement protocol to improve the representativeness of measurements. Spectra measured by using integrating sphere [5–10], rotating the sample during spectra collection, or grinding the samples [11–14] are of this type. Significantly better results were observed while the measurement was taken by a patented measurement method [15]. However, when the spectra are needed to be collected in-situ, none of the above methods are applicable. Hwang et al. [16] developed a fast and non-destructive analytical method to identify geographical origins of rice samples via transmission spectral collected through packed grains. But the variation of packing thickness will prevent the measurement from yielding reproducible spectra.

Instead of receiving a set of measurements, each for a sample, an object can be represented by a set of measurements (instances). Therefore, the heterogeneous classification problem can be solved by formulating it as a multi-instance learning one. Instead of receiving labeled instances as in traditional supervised learning, the MIL learner receives a set of labeled bags. The majority of MIL studies are concerned with binary classification problems [17]. Most of these studies assume that a bag is labeled positive if there is at least one instance in it being positive. Otherwise, a bag is labeled negative if all the instances in it are negative.

Based on the classical MIL assumption, numerous MIL methods have been proposed in the literatures, and most of them have been reviewed in earlier studies [17–19]. These algorithms use mainly the information of one instance from each positive bag. However, there is commonly more than one instance in any positive bags and much of the information contained in these instances is lost.

The MIL assumption is first used to solve the problem of *musk* drug activity prediction problem [20]. From then on, many problems have been formulated as MIL problem, such as image categorization, object detection, and human active recognition

* Corresponding authors at: College of Chinese Medicine, Beijing University of Chinese Medicine, Beijing 100102, China. Tel.: +86 10 84738621; fax: +86 84738661.
E-mail addresses: xyshi@126.com (X. Shi), yjqiao@263.net (Y. Qiao).

http://dx.doi.org/10.1016/j.talanta.2014.09.007
0039-9140/

[18,21,22]. Although all early work in MIL assumes a specific concept class known to be appropriate for a drug activity prediction like domain, the classical MIL assumption is not guaranteed to hold in other domains. Recently, a significant amount of researches in MIL are concerned with cases where the classical view of MIL problem is relaxed, and alternative assumptions are considered instead [17]. Although not all these works clearly state what particular assumption is used and how it relates to other assumptions, the use of alternative assumptions has been clarified by reviewing the studies in this area.

In this work, a variant of count-based binary MIL assumption was adopted. It assumes that a bag is labeled positive if the product of the positive posteriors of instances, weighted by the bag prior, is larger than that of the negative ones. Otherwise, a bag is labeled negative. Base on this assumption, the maxc-LS-SVM algorithm was proposed to deal with the heterogeneous classification problem. For multi-class cases, the original classification problem was decomposed in the framework of Error-Correcting Output Codes (ECOC) by one-versus-one design [23]. The maxc-LS-SVM algorithm was modified correspondingly and its performance was compared with mi-SVM and MI-SVM [24] algorithm. The results of applying these approaches on toy datasets and a real herbal dataset clearly showed the advantage of maxc-LS-SVM algorithm.

2. Theory and algorithm

2.1. Least squares-support vector machine (LS-SVM)

This study will briefly focus on the basic concepts of LS-SVM because the theory of LS-SVM has been described extensively in the literatures [25,26]. LS-SVM encompasses similar advantages as SVM, but additionally, it requires solving only a set of linear equations, which is much easier and computationally very simple. The objective function defined in LS-SVM is [27]

$$\min_{\omega,b,e} J_P(\omega,e)=\tfrac{1}{2}\omega^T\omega+\gamma\tfrac{1}{2}\sum_{i=1}^{N}e_i^2$$
$$\text{subject to(s.t.)}: y_i[\omega^T\varphi(x_i)+b]=1-e_i,\quad i=1,\dots,N \tag{1}$$

where w denotes the normal vector to the classification hyperplane, γ is the hyperparameter tuning the amount of regularization versus the sum squared error, e is the error variable, and $\varphi(x)$is the nonlinear map from original space to the high (and possibly infinite) dimensional space.

To solve the optimization problem efficiently, a Lagrange function is constructed and translated into its dual form

$$L(\omega,b,e;\alpha)=J_P(\omega,e)-\sum_{i=1}^{N}\alpha_i\{y_i[\omega^T\varphi(x_i)+b]-1+e_i\} \tag{2}$$

where α_i values are the Lagrange multipliers.

The conditions for optimality are

$$\begin{cases}\frac{\partial L}{\partial\omega}=0\to\omega=\sum_{i=1}^{N}\alpha_i y_i\varphi(x_i)\\ \frac{\partial L}{\partial b}=0\to\sum_{i=1}^{N}\alpha_i y_i\\ \frac{\partial L}{\partial e_i}=0\to\alpha_i=\gamma e_i, & i=1,\dots,N\\ \frac{\partial L}{\partial\alpha_i}=0\to y_i[\omega^T\varphi(x_i)+b]-1+e_i=0, & i=1,\dots,N\end{cases} \tag{3}$$

From the conditions for optimality, it can be concluded that no α_i values will be exactly equal to zero, meaning that the advantages of automatic sparseness are lost. However, the model can be trained much more efficiently after constructing the Lagrangian by solving the linear Karush–Kuhn–Tucker (KKT) system, since it yields a linear system instead of a quadratic programming problem

$$\begin{bmatrix}0 & 1_N^T\\ 1_N & \boldsymbol{\Omega}+\gamma^{-1}I_N\end{bmatrix}\begin{bmatrix}b\\ a\end{bmatrix}=\begin{bmatrix}0\\ \boldsymbol{y}\end{bmatrix} \tag{4}$$

where $\boldsymbol{y}$ is a vector containing the reference values, $\boldsymbol{1}_N$ is a $[N\times 1]$ vector of ones and $\mathbf{I}$ is an $[N\times N]$ identify matrix. $\boldsymbol{\Omega}$ is the kernal matrix defined by $\boldsymbol{\Omega}_{ij}=\varphi(x_i)^T\varphi(x_j)=\mathbf{K}(x_i,x_j)$.

The classifier in the dual space takes the form

$$y(x)=sign\left[\sum_{i=1}^{N}\alpha_i y_i K(x,x_i)+b\right] \tag{5}$$

2.2. Maximum margin formulation of MIL

In traditional supervised learning framework, an object is represented by one single instance, i.e. measurement vector, and associated with a class label. The goal is to induce a classifier to label instances. While, MIL groups instances into bags and each bag is attached with a class label. More formally, an object is represented by a bag $\boldsymbol{B}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_p\}$, which contains a set of D-dimensional instances. Each bag is associated with a label $\boldsymbol{Y}$.

Both mi-SVM and MI-SVM approaches [24] are modified and extended from Support Vector Machines (SVMs). The mi-SVM approach explicitly treats the instance labels as unobserved integer variables subjected to constrain defined by the (positive) bag labels. A generalized soft-margin SVM is defined as follows in primal form

$$\min_{\{y_i\}}\min_{w,b,\xi}\tfrac{1}{2}\|w\|^2+C\sum_i\xi_i$$
$$\text{s.t. }\forall i:\ y_i(\langle w,x_i\rangle+b)\ge 1-\xi_i,\xi_i\ge 0,y_i\in\{-1,1\} \tag{6}$$

where ξ is a non-negative slack variable, C is the hyperparameter tuning the amount of the degree of misclassification versus the sum squared error, y_i is the label of instance.

The mi-SVM formulation leads to a mixed integer programming problem. One has to maximize a soft-margin criterion jointly over possible label assignments and hyperplane. While, the MI-SVM approach extends the notion of margin to bag and maximizing the bag margin directly. The bag margin with respect to a hyperplane is defined by

$$\gamma_I\equiv Y_I\max_{i\in I}(\langle w,x_i\rangle+b) \tag{7}$$

Incorporating the above bag margin, an MIL version of the soft-margin classifier is defined by

$$\min_{w,b,\xi}\tfrac{1}{2}\|w\|^2+C\sum_I\xi_I$$
$$\text{s.t. }\forall I:\ Y_I\max_{i\in I}(\langle w,x_i\rangle+b)\ge 1-\xi_I,\xi_I\ge 0 \tag{8}$$

Unfolding the max operation by introducing inequality constraint per instance for the negative bags, or by introducing a selector variable $s(I)\in I$ which denotes the positive instance selected for the positive bags, one can obtain the following equivalent formulation

$$\min_s\min_{w,b,\xi}\tfrac{1}{2}\|w\|^2+C\sum_I\xi_I$$
$$\text{s.t. }\begin{array}{l}\forall I:\ Y_I=-1\wedge-\langle w,x_i\rangle-b\ge 1-\xi_I,\forall i\in I,\\ \text{or}Y_I=1\wedge-\langle w,x_{s(I)}\rangle-b\ge 1-\xi_I,\xi_I\ge 0.\end{array} \tag{9}$$

MI-SVM can also be cast as mixed-integer program. One has to find both the optimal selectors and hyperplane.

A heuristic optimization scheme has been proposed to solve the mixed-integer programs by alternating the following two steps (i) for a given integer variables, solve the associated Quadratic Programming (QP) and find the optimal discriminant function,