Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Recognition of the hardness of licorice seeds using a semi-supervised learning method and near-infrared spectral data

Liming Yang [a,*], Qun Sun [b,1]

[a] College of Science, China Agricultural University, Beijing, 100083, China
[b] College of Agriculture and Biotechnology, China Agricultural University, Beijing, 100193, China

ARTICLE INFO

Article history:
Received 12 November 2011
Received in revised form 26 February 2012
Accepted 15 March 2012
Available online 26 March 2012

Keyword:
Licorice seeds classification
Near-infrared spectroscopy
Support vector machine
Semi-supervised support vector machine
DC programming

ABSTRACT

The recognition of the hardness of licorice seeds is a challenging task. The purpose of this investigation is to identify the hardness of licorice seeds employing a semi-supervised learning method and near-infrared spectroscopy. An excellent semi-supervised learning model, the semi-supervised support vector machine (S^3VM), is built using the small labeled samples and the large unlabeled samples. Moreover, the proposed model is solved by employing an effective method, the robust DC (difference of convex functions) programming. The resulting algorithm only requires the solving of a few linear programs. Furthermore, this model is used for the direct classification of licorice samples. Comparing with the supervised support vector machine (SVM), experimental results on different spectral regions show that incorporating unlabeled samples in training improves the generalization when insufficient training information is available. Moreover, our method outperforms the existing S^3VM method by obtaining better performance in different spectral regions. These results show that it is possible to identify the hardness of licorice seeds using the proposed S^3VM and near-infrared spectroscopic data. We hope that the results obtained in this study will help further investigations of the hardness of crop seeds.

1. Introduction

Near-infrared (NIR) [1] spectroscopy has demonstrated great potential in the analysis of complex samples owing to its simplicity, rapidity and nondestructivity. Recently, in the field of agriculture, NIR spectroscopy has been applied to quantitative and qualitative analysis, such as the determining of seed moisture, seed vigor and seed purity [2–5]. NIR spectroscopy is based on the absorption of electromagnetic radiation in the region from 800 to 2500 nm (12,500 − 4000 cm^{-1}).

Licorice (*Glycyrrhiza uralensis Fisch*) is a traditional Chinese herbal medicine, and it has both hard seeds and soft seeds like many other legumes. The hard seeds are more suitable for storage and more resistant to adverse environmental conditions than the soft seeds. However, the soft seeds are morphologically very similar to the hard seeds. Thus, it is difficult to distinguish them from each other without damaging the seeds. Usually, the hard characteristics of licorice seeds are determined by soaking the seeds [1], but this method is time-consuming and sometimes destructive to the seeds. Therefore, developing a fast and nondestructive recognition technique for licorice seeds is an important and challenging subject.

Support vector machine (SVM) [6,7] is a promising supervised learning method in pattern recognition, and has been successfully applied to chemometrics [8,9]. It is firmly rooted in the minimization of structural risk which balances model complexity and empirical risk. SVM is thus superior to traditional learning methods which are usually based on the minimization of empirical risk. SVM performs the classification tasks only using the labeled samples. Thus when insufficient labeled samples are available, SVM is usually not satisfactory. A considerable drawback of SVM is that it requires a large number of labeled samples in order to construct accurate classifiers. However, in many real-world applications, the labeled samples may be very few or expensive to obtain, while unlabeled samples are easier to collect. In this setting, the supervised learning methods are difficult to use owing to the lack of labeled samples.

Using both labeled and unlabeled samples for the purpose of learning is called semi-supervised learning, where some knowledge of the unlabeled samples is taken into account to improve generalization during the training procedure. Semi-supervised SVM (S^3VM) [10–13] may seem to be the perfect semi-supervised learning approach since it combines the powerful regularization of SVM with a direct implementation of the cluster assumption [12]. S^3VM is trained using the small labeled sample while simultaneously assigning the large unlabeled samples to one of two classes so as to maximize the margin (distance) between the different classes. Such margin is a measure of the model complexity. The main drawback of S^3VM is that the objective function is nonconvex, and it is therefore difficult

[*] Corresponding author. Tel.: +86 010 62736511; fax: +86 010 62736777.
E-mail addresses: cauylm@126.com (L. Yang), sqcau@126.com (Q. Sun).
[1] Tel.: +86 010 62732775; fax: +86 010 62733404.

0169-7439/$ – see front matter. Crown Copyright © 2012 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2012.03.010

to find its exact solution. Some of the optimization algorithms have been applied to semi-supervised SVM such as the exact solution method [10] and approximation algorithms [11,13].

In this investigation, we present a semi-supervised learning model to classify licorice seeds using NIR spectroscopy data. This investigation is motivated by the following observations:

(1) NIR spectroscopy has already been used as a fast and nondestructive technique in chemometrics.
(2) S^3VM has become a powerful technology for classification problems. However, little attention has been applied to the recognition of the hardness of legume crops employing semi-supervised learning method.
(3) DC program algorithm (DCA) [14–17] is an efficient and robust algorithm for solving nonconvex problems, especially in the large-scale setting.

The main contributions of this work are as follows:

(1) We propose a framework for identifying the hardness of licorice seeds using semi-supervised SVM and NIR spectroscopy data.
(2) A new semi-supervised SVM formulation is presented and directly applied to identifying the hardness of licorice seeds employing NIR spectroscopy data.
(3) The proposed S^3VM model is solved by the DCA, and has low computational burden, only requiring to solve a few linear programs.

2. Theory and method

2.1. Semi-supervised support vector machine (S^3VM)

S^3VM is an extension of the supervised SVM with an additional regularization term for unlabeled samples. Specifically, assume that the sample set consist of m labeled samples and p unlabeled samples. The labeled samples are represented by the matrix A of size $m \times n$, and the unlabeled samples are represented by the matrix B of size $p \times n$. The labels for the labeled samples are given by a diagonal matrix D of m-th order with values of ± 1. Each row of matrix A (resp. B) denotes a labeled sample (resp. unlabeled sample). For each unlabeled sample x_j, the variables r_j and $s_j (j = 1,2,\ldots p)$ represent the two possible misclassification errors. The final class of the unlabeled sample is the one that results in the smallest misclassification error. Finding a linear hyperplane $w^T x = b$ far away from both the labeled and unlabeled samples can be formulated to minimize the objective function

$$||w||_1 + \nu e^T \xi + \mu e^T min\{r, s\} \tag{1}$$

subject to the constraints:

$$\begin{cases} D(Aw - eb) + \xi \geq e, \xi \geq 0 \\ Bw - eb + r \geq e, r \geq 0 \\ -Bw + eb + s \geq e, s \geq 0 \end{cases} \tag{2}$$

where an arbitrary dimension vector of ones is denoted by e. The component by component minimum of two vectors r and s is denoted by $min\{r, s\}$, with component j being: $min\{r_j, s_j\} (j = 1,\ldots p)$. Two parameters $\nu > 0$ and $\mu > 0$ balance the model complexity and misclassification error. In addition, they control over the influence of labeled and unlabeled samples. Variable ξ represents the classification error for the labeled samples. The first two terms of the objective function, together with the first constraint correspond to a supervised 1-norm SVM, which attempts to classify the labeled samples. The last term in the objective function, together with the remaining constraints assign each unlabeled sample x_j to the positive class or negative class, whichever generates a lower misclassification error: $min\{r_j, s_j\}$. Bennett and Demiriz formulated this problem as a mixed integer program (MIP) in literature [10], where a globally optimal solution of this problem was found. However, this method (called MIP-S^3VM) do not work on large unlabeled sample sets.

1-norm of w ($||w||_1 = \sum |w_i|$) instead of 2-norm is used as a regularization term in the objective function, which corresponds to maximizing the classification margin using the infinity norm of w, namely $\frac{1}{||w||_1}$. One major benefit of $||w||_1$ over $||w||_2$ in the objective function is variable reduction since minimizing $||w||_1$ leads to most elements of vector w are zero. When the ith component of w is zero, the ith component of the observation vector x is irrelevant in deciding the class of x using linear decision function $f(x) = w^T x - b$. Thus variable selection [18,19] and classification can be conducted jointly through the S^3VM formulation.

2.2. DC programming

We outline the main algorithmic results for DC programming [14–17]. The key to DC programs is to decompose an objective function into the difference of two convex functions, from which a sequence of approximations of the objective function yields a sequence of solutions converging to a stationary point, possibly an optimal solution. Generally speaking, a so-called DC program (P_{dc}) is to minimize a DC function:

$$f(x) = g(x) - h(x), x \in R^n (P_{dc}) \tag{3}$$

with $g(x)$ and $h(x)$ being convex functions. A DC program is called a polyhedral DC program when either $g(x)$ or $h(x)$ is a polyhedral convex function (i.e., the pointwise supremum of a finite collection of affine functions).

The DCA is an iterative algorithm based on local optimality conditions and duality [14–17]. The idea of DCA is simple (to simplify, we omit here the dual part): at each iteration, one replaces in the primal DC problem (P_{dc}) the second component h by its affine minorization: $h(x^k) + (x - x^k)^T y^k$, to generate the convex program:

$$minimize : \left\{ g(x) - h\left(x^k\right) - \left(x - x^k\right)^T y^k, x \in R^n, y^k \in \partial h\left(x^k\right) \right\} \tag{4}$$

Where ∂h is the subdifferential of convex function h. In practice, a simplified form of the DCA is used. Two sequences $\{x^k\}$ and $\{y^k\}$ satisfying $y^k \in \partial h(x^k)$ are constructed, and x^{k+1} is a solution to the convex program (4). The simplified DCA scheme is described as follows.

```
Initialization: Choose an initial point $x^0 \in R^n$ and let $k = 0$
Repeat
    Calculate $y^k \in \partial h(x^k)$
    Solve convex program (4) to obtain $x^{k+1}$
    Let k:=k+1
Until some stopping criterion is satisfied.
```

DCA is a descent method without line search, and it converges linearly for general DC programs. In particular, for polyhedral DC programs, the sequence $\{x^k\}$ contains finitely many elements, and in a finite number of iterations the algorithm converges to a stationary point satisfying the necessary optimality condition [14,16].

DCA has been successfully applied to many optimizations. For example, only using the information from the labeled samples, a variable selection model for the supervised SVM was proposed based on DCA in literature [17]. This supervised learning machine is trained without considering unlabeled samples.

2.3. DC formulation for solving S^3VM

Let $t = r - s$; thus $min\{r, s\} = \frac{1}{2}(r + s - |t|)$. Furthermore, we introduce the variable z such that $|w| \leq z$, and then S^3VM (1)–(2) is equivalent to minimizing the objective function

$$e^T z + \nu e^T \xi + \frac{1}{2} \mu e^T (r + s - |t|) \tag{5}$$