Neurocomputing 211 (2016) 191–201

Contents lists available at ScienceDirect

Neurocomputing

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

Novel Grouping Method-based support vector machine plus for structured data

Qiuling Hou, Ling Zhen, Naiyang Deng, Ling Jing *

College of Science, China Agricultural University, Beijing 100083, China

ARTICLE INFO

Article history:
Received 21 September 2015
Received in revised form
18 March 2016
Accepted 26 March 2016
Available online 8 June 2016

Keywords:
Support vector machine
Support vector machine plus
Hidden information
Group information
Clustering technique

ABSTRACT

In the traditional Support Vector Machine Plus (SVM+), the grouping method has great randomness and only takes into account part of the structural information of the dataset. In order to overcome these shortcomings, in this paper, we propose a novel framework termed as FCSVM+ to improve the performance of SVM+ by combining clustering technique and feature selection. The new framework strategy is expected to not only take fully into account the structural information of the training data, but also partition the training data into more meaningful groups. To prove the advantage of the framework, in particular, we adopt two simplest feature selection methods, i.e. F-score and Laplacian score methods, to select the features, then apply a recently proposed clustering technique to get a better partition of training data by the selected features, in which the number of clusters could be found automatically. Three major contributions of this paper can be concluded as: (1) improving the performance of the existing SVM+ classifier; (2) extracting the potential structural information of the training data by using more feature attributes instead of one; (3) replacing the truncation method in SVM+ with a clustering technique. The comprehensive experimental results on the UCI benchmark datasets illustrate the validity and advantage of our approach.

1. Introduction

In the data-rich world, there often exists additional hidden information about training samples, which is only available at the training stage and never available for the test dataset. For example, in medical diagnosis, suppose our goal is to find a rule to predict the outcome of a treatment in a year under the condition of being given the current symptoms, and the development of symptoms in three months, in six months, and in nine months can be considered as the additional hidden information. Group information can be considered as a special case of hidden information. In this case, each training data has an additional group label, and this information can be used to help improving the generalization of the algorithms. For example, the heart disease predictive model is estimated by using a training dataset of male and female patients, and the gender can be considered as the additional group information. Another example, for landmine detection, the data from different landmine fields have different characteristics, so data from the same landmine field can be regarded as a group.

To introduce additional hidden information, Vapnik proposed an advanced learning paradigm termed as Learning Using Hidden Information (LUHI) [1]. Later, [2] proposed SVM+ to implement algorithms which addressed the LUHI paradigm. Refs. [3,4] discussed the details of the new paradigm and corresponding algorithms and demonstrated the superiority of the new learning paradigm over the classical one. Ref. [5] gave out a growing body of work on theoretical analysis of LUHI. Ref. [6] showed two fast algorithms for solving the optimization problem of SVM+. In [7,8], the learners studied the relationship between the SVM+ approach and the multi-task learning scenario. The SVM+ idea was applied to metric learning in [9] and boosting algorithms in [10]. Ref. [11] used privileged information to data clustering. In [12], an SVM-minus method was proposed to compute similarity scores in video face recognition. Ref. [13] showed that any SVM+ solution was also a weighted SVM (WSVM) solution with appropriately chosen weights, at the same time it also gave out the necessary and sufficient condition for equivalence between SVM+ and WSVM. Ref. [14] identified some issues which might affect SVM+'s applicability in practice.

Learning Using Structured Data (LUSD), as a special case of LUHI, was first investigated in [15]. Ref. [15] presented another SVM-based optimization formulation also termed as SVM+ for LUSD, and this formulation took into account the known structure of the training data. To be more precise, the training data could be partitioned into several related groups according to one feature attribute, then SVM+ methodology provided a way to incorporate

* Corresponding author.
E-mail address: jingling@cau.edu.cn (L. Jing).

http://dx.doi.org/10.1016/j.neucom.2016.03.086
0925-2312/

this additional knowledge into a formal optimization formulation. Ref. [16] applied this SVM+ to biomedical engineering, and analyzed the characteristics of group variables. Ref. [17] applied group information to one-class SVM, and proposed a new one-class SVM based on the hidden information.

Though LUSD paradigm has made a great success, grouping method in SVM+ under LUSD is seldom discussed by now. The group information is usually from the priori knowledge which is actually difficult to obtain. In order to solve this problem, the traditional SVM+ partitions the training data into several groups by truncating only one feature attribute. But this grouping method has great randomness and only takes into account part of the structural information of the dataset. In fact, the partition of the training data will directly affect classification results. In order to overcome these shortcomings of SVM+, in this paper, we propose a novel framework termed as FCSVM+ to improve the performance of SVM+ by combining clustering technique and feature selection. The new framework strategy is expected to not only take fully into account the structural information of the training data, but also partition the training data into more meaningful groups to obtain a better classifier. Of course, under this framework, there are many different formulations if different clustering techniques and feature selection methods are adopted. While, it is worth noting that the adopted clustering techniques should be able to automatically find the appropriate number of clusters. In this paper, to prove the advantage of our proposed framework, in particular, we adopt a recently proposed clustering technique [20] and two simplest feature selection methods to validate the performance of the framework. Concretely, first, we adopt F-score or Laplacian score feature selection method to select the features, then apply the recently proposed clustering technique to get a better partition of training data by the selected features, in which the number of clusters could be found automatically, finally, we obtain the classifier by SVM+. We call these two combined formulations as FFCSVM+ and LFCSVM+ corresponding to F-score and Laplacian score feature selection methods, respectively.

Actually, F-score feature selected method, as a simple and common way, has been combined with SVM before, [18] investigated the performance of combining SVM and various feature selection strategies, such as F-score+SVM, F-score+RF+SVM and RF+RM-SVM. There is some difference between our proposed FFCSVM+ and F-score+SVM. In F-Score+SVM method, SVM is performed on the selected features with high F-scores. While in FFCSVM+, the selected features with high F-scores are only used to group the training data, and SVM+ is applied still on the original data with all features.

In all, three major contributions of this paper can be concluded as: (1) improving the performance of the existing SVM+ classifier by a novel framework; (2) extracting the potential structure information of the training data by using more feature attributes instead of one; (3) replacing the truncation method in SVM+ with a clustering technique which is able to automatically find the appropriate number of clusters. All the comprehensive experimental results on the UCI benchmark datasets show the robustness and effectiveness of our proposed approach.

The rest of the paper is organized as follows. In Section 2, a brief overview of the useful background is given. In Section 3, we will introduce our proposed method in detail. Section 4 includes the experimental results, and Section 5 concludes the paper.

2. Background

2.1. Support vector machine (SVM) [19]

SVM [19] is a useful and promising technique for pattern classification and regression. It emerges from the research in statistical learning theory on how to regulate the trade-off between structural complexity and empirical risk. And its main attempt is to reduce generalization error by maximizing the margin between two disjoint half planes.

Given a training dataset $T = \{(x_1, y_1), \ldots, (x_l, y_l)\}$, where $x_i \in R^n$, $y_i \in \{+1, -1\}$, $i = 1, 2, \ldots, l$. The goal of SVM is to find an optimal separating hyperplane to correctly separate the positive points and the negative points

$$(w \cdot \phi(x)) + b = 0 \tag{1}$$

where $\phi(\cdot)$ is a map from the space R^n to the Hilbert space, w is a weight vector corresponding to the Hilbert space, and $b \in R$.

The optimization problem corresponding to SVM is as follows:

$$\begin{aligned} &\min_{w,b,\xi} \ \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{l}\xi_i \\ &\text{s. t.} \quad y_i((w \cdot \phi(x_i)) + b) \ge 1 - \xi_i \\ &\xi_i \ge 0, \quad i = 1, \ldots, l, \end{aligned} \tag{2}$$

where $C > 0$ is a penalty parameter, and $\xi_i, i = 1, \ldots, l$, are the slack variables. $\frac{1}{2}\|w\|^2$ is the regularization term which is equivalent to the maximization of the margin between two classes. Generally, the solution of (2) is obtained by means of solving its dual problem.

The dual problem of (2) can be expressed as

$$\begin{aligned} &\min_{\alpha} \ \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{j=1}^{l}\alpha_j \\ &\text{s. t.} \sum_{i=1}^{l} y_i\alpha_i = 0, \\ &0 \le \alpha_i \le C, \quad i = 1, \ldots, l. \end{aligned} \tag{3}$$

where α_i, $i = 1, \ldots, l$, are the Lagrangian multipliers, and $K(x, x') = (\phi(x), \phi(x'))$ is the kernel.

Suppose the solution of (3) is $\alpha^* = (\alpha_1^*, \ldots, \alpha_l^*)^T$, then

$$w^* = \sum_{i=1}^{l}\alpha_i^* y_i \phi(x_i), \quad b^* = y_j - \sum_{i=1}^{l}\alpha_i^* y_i K(x_i, x_j) \tag{4}$$

where α_j^* is a component of α^*, and $\alpha_j^* \in (0, C)$.

A new sample is classified as +1 or −1 according to the final decision function as follows

$$f(x) = \text{sgn}\left(\sum_{i=1}^{l}\alpha_i^* y_i K(x_i, x) + b^*\right) \tag{5}$$

2.2. Support vector machine plus (SVM+) [15]

Given a training dataset $T = \{(x_1, y_1), \ldots, (x_l, y_l)\}$, where $x_i \in R^n$, $y_i \in \{+1, -1\}$, $i = 1, 2, \ldots, l$. Suppose the training dataset is a union of t related groups: $T = \{\{X_r, Y_r\}, r = 1, \ldots, t\}$, where $\{X_r, Y_r\} = \{(x_{r_1}, y_{r_1}), \ldots, (x_{r_{n_r}}, y_{r_{n_r}})\}$, $T_r = \{r_1, \ldots, r_{n_r}\}$. To control the capacity of the slack variables, the slack variable corresponding to x_i belonging to the r-th group can be defined as a realization of a correcting function

$$\xi_i^r = \xi_r(x_i), \quad i \in T_r, \ r = 1, \ldots, t. \tag{6}$$

That is, the slack variables from different groups should correspond to different constraints.

To define the correcting function $\xi_i^r = \xi_r(x_i)$ for the group r, Vapnik proposed to map the input vectors x_i, $i \in T_r$, into two different spaces: the decision space Z and the correcting space Z_r. In the decision space, the decision function is defined as the standard SVM, while in the correcting space, the correcting functions are