Signal Processing 114 (2015) 171–182

Contents lists available at ScienceDirect

Signal Processing

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

High-dimensional semi-supervised learning via a fusion-refinement procedure

Zhikun Lei [a], Renfu Li [a,*], Xuelei Sherry Ni [b], Xiaoming Huo [c]

[a] Department of Aerospace Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[b] Department of Statistics and Analytical Sciences, Kennesaw State University, Kennesaw, GA 30144, USA
[c] School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA

ARTICLE INFO

Article history:
Received 24 September 2014
Received in revised form
2 February 2015
Accepted 2 March 2015
Available online 11 March 2015

Keywords:
Semi-supervised learning (SSL)
Sufficient dimension reduction (SDR)
Fusion-refinement (FR) procedure

ABSTRACT

This paper develops a sufficient dimension reduction (SDR) approach for the high-dimensional semi-supervised learning (SSL) problem. In the proposed technique, we first modify the fusion-refinement (FR) procedure, which was proposed in [1], to extract the essential features for a lower-dimensional representation. We then apply an SSL algorithm (e.g., the low density separation (LDS)) in the lower-dimensional feature space to tackle the SSL problem. Numerical experiments are conducted on some widely-used data sets. We carry out a comparison between the proposed procedure and some recently proposed semi-supervised learning approaches (including greedy gradient Max-Cut (GGMC), semi-supervised extreme learning machines (SS-ELM)) and dimension reduction procedures (such as the semi-supervised local Fisher discriminant analysis (SELF), the trace ratio based flexible semi-supervised discriminant analysis (TR-FSDA), and trace ratio relevance feedback (TRRF)). In extensive numerical simulations, the new technique outperforms its competitors in many cases, demonstrating its effectiveness.

1. Introduction

In various fields, people often deal with data that have a limited number of labeled data points together with a large amount of unlabeled ones. The modeling problem in such scenario, where both labeled and unlabeled data points are present, is called a semi-supervised learning (SSL) problem [2,3]. Such modeling problem can be dated back to 90s [4,5]. SSL has drawn extensive attention in many areas, such as face recognition [6], text analysis [7], signal processing [8], and remote sensing [4]. Studies have shown that adequate use of unlabeled data can help to develop much better classifiers [5,9,10].

Meanwhile, due to rapid advances in data collection and storage technology, the data sets we have nowadays are more and more complicated, and often in a high-dimensional format. Because of the massive size of data, solving an SSL problem in a high-dimensional space is much more difficult than doing so in a relatively low-dimensional space. In the literature, various dimension reduction approaches have been proposed [11–14,6,15–18] to help overcome the curse of dimensionality. These methods can be categorized into two groups: (1) the methods using pairwise constraints (including [11–14,17]) and (2) the methods that are based on class labels (including [6,15,16,18]).

The main idea of this paper is to precede an SSL solver with a dimension reduction method. So, we now briefly survey some representative dimension reduction approaches on

* Corresponding author. Tel./fax: +86 27 87559384.
E-mail addresses: zhikunlei@hust.edu.cn (Z. Lei), renfu.li@hust.edu.cn (R. Li), xni2@kennesaw.edu (X. Sherry Ni), xiaoming@isye.gatech.edu (X. Huo).

http://dx.doi.org/10.1016/j.sigpro.2015.03.003
0165-1684/© 2015 Elsevier B.V. All rights reserved.

the high-dimensional SSL problem. Some of them will be used as our benchmarks.

- The following two methods adopt pairwise constraints. Paper [11] constructs an adjacency matrix using pairwise constrains known as must-link constraints and cannot-link constraints. With the help of this matrix, the SSDR-CMU method is then proposed to preserve the structure of the unlabeled data as well as the pairwise constraints of the labeled cases in the projected low-dimensional space. The trace ratio relevance feedback (TRRF) method is proposed in [17], which also utilizes pairwise constraints, however it constructs some scatter matrices. A linear regression model subsequently is adopted to explore the local structure of the unlabeled data. Then a trace ratio objective function is reformulated, whose solution defines the projection matrix.
- The following methods deal with class labels. Paper [16] proposes a semi-supervised local Fisher discriminant analysis (SELF), which reports improved performance in relative to locality preserving projection (LPP) [19] and local Fisher discriminant analysis (LFDA) [20]. Semi-supervised discriminant analysis (SDA) in [6] leverages the Laplacian matrix [21] to avoid the overfitting phenomenon in Fisher discriminant analysis (FDA) [22] and to exploit the unlabeled information at the same time. In [18], the Trace Ratio (TR) criterion is introduced for dimension reduction. A flexible regularizer is added into the objective function to cope with nonlinear data representations.

Although there have been many methods proposed for the high-dimensional SSL problem, our literature search came back with no paper that incorporates the *sufficient* dimension reduction (SDR). Recall that SDR is to represent data in a lower-dimensional space with 'minimal loss of information'. In this paper, the meaning of 'loss of information' will be interpreted as searching for the *central spaces*, which will be defined later.

Facing the labeled and unlabeled data in SSL, the drawback of representative unsupervised dimension reduction methods, like PCA method, is that they cannot make use of labeled data. And LFDA — supervised dimension reduction method — exploits the labeled ones and omits the abundant unlabeled ones. Although labeled and unlabeled data are utilized in SELF, a semi-supervised dimension reduction, they are not used efficiently, which means not considering minimal loss of information. SDR methods consider minimal loss of information via the central spaces, which makes it more efficient for semi-supervised dimension reduction problems.

We will explore the usage of SDR in the framework of SSL. Our literature search indicates that this is the first time that the concept of SDR is employed in SSL. Conceptually, we will first employ Fusion Refinement (FR) [1]—a sufficient dimension reduction technique—to compute for a lower-dimensional representation of both labeled and unlabeled data points. This representation is then used to facilitate the subsequent classification by incorporating some proper SSL algorithms, such as the low density separation (LDS) [23] and the manifold regularization (LapRLS/LapSVM) [24]. To study the performance of the proposed FR based method, numerical simulations are conducted to compare them with LDS, LapRLS/LapSVM, greedy gradient Max-Cut (GGMC) [25], and semi-supervised extreme learning machines (SS-ELM) [26]. We choose these methods because they are the current state-of-the-art methods. Experiments have also been carried out to evaluate the difference between the FR related methods and three recent semi-supervised dimension reduction methods: semi-supervised local Fisher discriminant analysis (SELF) [16], TR based flexible SDA (TR-FSDA) [18], and TRRF [17]. The comparison with other dimension reduction competitors, principal component analysis (PCA) and LFDA [20], is presented as well. We will show that FR can lead to better performance in SSL in many cases.

We summarize the main contributions of this paper as follows.

- We find that the FR method can reveal a compact representation in the framework of SSL; this leads to a fast implementation, due to the compact representation.
- We propose a combined semi-supervised classification method, FR+LDS, which can capture the lower-dimensional data structure embedded in the original high-dimensional classification problem. This can lead to better numerical performances in some cases.
- We bring in sufficient dimension reduction concept in SSL. Moreover in our experiments, the modified FR method demonstrates its effectiveness comparing with other semi-supervised dimension reduction methods.

The remainder of the paper is organized as follows. Section 2 is the literature review of semi-supervised learning and sufficient dimension reduction. Section 3 outlines the idea of applying an FR-based procedure on SSL problems. A complete algorithm is developed and presented in this section as well. Section 4 describes the two sets of systematic experiments. The first set of experiments is a comparison between the SSL-only methods and the SDR-enhanced methods, i.e., an FR procedure followed by a particular SSL algorithm. The second set of experiments is to compare FR with some existing semi-supervised/supervised/unsupervised dimension reduction methods: e.g., SELF [16], TR-FSDA [18], TRRF [17], LFDA [20], and PCA. Conclusions are drawn in Section 5.

2. Background

This section gives an overview of both semi-supervised learning and sufficient dimension reduction.

2.1. A very brief survey on SSL

We denote the labeled data as $\{(x_i, y_i)\}_{i=1}^{l}$, and the unlabeled data as $\{x_i\}_{i=l+1}^{l+u}$. In the SSL framework, the objective is to find a function $f: \mathcal{X} \mapsto \mathcal{Y}$, $f \in \mathcal{F}$, so that $y_i \approx f(x_i)$, where $\mathcal{F}$ is a prescribed functional space, $\mathcal{X}$ is the input space, and $\mathcal{Y}$ is the label space. An input instance is denoted by $x \in \mathcal{X}$, and $y \in \mathcal{Y}$ denotes the corresponding label.

One typically assumes a probability distribution $\mathcal{P}$ on $\mathcal{X} \times \mathcal{Y}$, from which the labeled data are generated. Correspondingly, an unlabeled datum $(x \in \mathcal{X})$ is drawn from the