Pattern Recognition 60 (2016) 908–920

Contents lists available at ScienceDirect

Pattern Recognition

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

# Orthogonal optimal reverse prediction for semi-supervised learning

Hongbin Yu [a], Hongtao Lu [a,b,*]

[a] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, PR China
[b] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai 200240, PR China

ARTICLE INFO

*Article history:*
Received 12 December 2015
Received in revised form
1 July 2016
Accepted 3 July 2016
Available online 5 July 2016

*Keywords:*
Semi-supervised learning
Optimal reverse prediction
Orthogonal constraint
Orthogonal optimal reverse prediction

ABSTRACT

Optimal reverse prediction (ORP) has recently been proposed as a semi-supervised framework to unify supervised and unsupervised training methods such as supervised least square, principal component analysis (PCA), *k*-means clustering and normalized graph-cut. ORP has an ability to deal with classification tasks in which the labeled data are insufficient. But, the performance of ORP and its kernelized version is still not satisfactory for classification applications. To further improve performance of ORP, motivated by recently proposed orthogonal *k*-means clustering, in this paper we propose an orthogonal optimal reverse prediction (OORP), together with its kernelized and Laplacian regularized extensions. With only limited additional computations, our algorithms can greatly enhance the classification performance, compared to the original ORP.

Extensive experiments on synthetic and benchmark data collections consistently prove the effectiveness and efficiency of our OORP in comparison with several competing approaches.



## 1. Introduction and brief review of SSL

Semi-supervised learning (SSL) is a popular learning paradigm when labeled data are scarce. In many machine learning applications, obtaining data samples is much easier than labeling them. For example, in the speech recognition research, labeling speech signal is more complicated than signal collecting. Also, for web information processing, there are billions of web pages that are available online. However, it is costly to reliably label them.

When labeled data are insufficient, it is possible to exploit unlabeled data to facilitate the classification performance [1,2]. Semi-supervised learning algorithms, which incorporate labeled and unlabeled data simultaneously into models, have recently been extensively studied and widely used in classification, dimensionality reduction [3,4] and clustering.

Probably the earliest idea of adopting the unlabeled data to improve the classification performance is the self-learning algorithm [1], which is a wrapper-algorithm and has appeared in many literatures for some times. The procedure of this algorithm can be generally divided into three steps: First, it repeatedly trains a classifier model based on labeled data. Second, the algorithm can predict a part of unlabeled data according to the current classifier model. Third, the algorithm adds the predicted unlabeled data into the labeled data set for the next training phase. At that time, the idea of self-learning was widely used in semi-supervised learning algorithm. However, the performance of the algorithm strongly depends upon the classifier's prediction accuracy.

The transductive support vector machine (TSVM) [5,6] is the mostly used semi-supervised learning algorithm, which seeks the largest separation between labeled and unlabeled data through regularization. It has been shown to give good performance on the text classification job [7]. In our five experiments, TSVM gives a good classification performance, only inferior to our KOORP method in four experiments, however, TSVM is a time consuming method which limits its applications in reality. SVMs combining other algorithms such as locality preserving projection [8], called semi-supervised induction, have been proposed for images retrieval tasks. Experimental results have shown its efficiency and effectiveness in [9].

LapSVM and LapRLS [10] are two important regularizer based semi-supervised learning methods. They exploit the geometry structure of data to enhance learning by imposing a data-dependent regularization to the learning model. LapSVM and LapRSL are graph based learning methods, where graph is constructed based on the labeled and unlabeled data and the instances which connected by heavy edge tend to have the same labels. Graph-based algorithms have shown their superiority in many classification tasks [10–12]. However, this approach usually involves the extra graph computation operations so that resulting those algorithms are time consuming when dealing with large data set. Many algorithms, such as anchor graph algorithm [13,14] and Nystrom

* Corresponding author at: Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, PR China.
*E-mail addresses:* alexander.yuu@hotmail.com (H. Yu), lu-ht@cs.sjtu.edu.cn (H. Lu).

http://dx.doi.org/10.1016/j.patcog.2016.07.005
0031-3203/

approximation [15,16], have been proposed to reduce their computation complexities. Besides the graph based algorithms, kernel based algorithms have also been proposed in [17] as well as other regularized semi-supervised learning algorithms such as those in [18–20].

Recently, Xu et al. have proposed a new SSL model called optimal reverse prediction (ORP) [21,22], which adopt the least square regression-like approach to minimize the least square loss by optimizing over both the mapping matrix and guessed labels of the test data. It has been shown that supervised least squares, unsupervised principal components analysis, $k$-means clustering and normalized graph-cut can all be expressed as special cases of the ORP method. However, their paper only analyzed the relationship between ORP and the traditional methods and has not presented efficient algorithms to solve it and the performance of the method is also unsatisfactory.

In [23,24], the concept of orthogonal $k$-means clustering has been proposed. In this work, pair-wise orthogonal constraint has been resorted to the cluster center matrix, which not only can improve the information retrieval accuracy but also make the $k$-means clustering algorithm more scalable. As we know that discrete optimization is the most feasible approach to solve $k$-means clustering problem, however, as $k$-means problem is not submodular, discrete optimization cannot obtain a global optimal solution. The same problem exists in the ORP problem. In [21], the algorithm of ORP first relaxes the label matrix to be continuous matrix and then thresholds it to get the discrete label matrix. Some algorithms imposed the sparse constraints on the objective to implement the clustering task and got excellent performance [25,26]. In the orthogonal $k$-means the cluster matrix is constrained to be columns orthogonal and the matrix can be represented as a product of a rotation matrix and a diagonal scaling matrix, which makes the optimization procedure much easier. More details are referred to [24,23].

Inspired by the success of orthogonal $k$-means, we propose an orthogonal optimal reverse prediction (OORP) algorithm, then extend it to the kernelized OORP (KOORP) and Laplacian OORP (LOORP). Our algorithms can give a better classification accuracy than the original ORP and kernelized ORP (KORP) algorithms at the cost of limited additional computations in OORP and KOORP.

To illustrate the superiority of our algorithms, we designed the classification experiments on a synthetic data and four public data sets to compare our algorithms' performance with other semi-supervised learning algorithms including self learning+SVM (SLSVM) [27], ORP, KORP, TSVM, LapRLS, LapSVM and supervised learning algorithm SVM. Experimental results show that our proposed OORP and its extensions KOORP and LOORP outperform the other semi-supervised learning algorithms, SLSVM, LapRLS, LapSVM and TSVM, respectively. Compared with the supervised SVM, our KOORP performs better, and it achieves the best performance among all methods in four experiments of five. So, our OORP and its extensions are useful for semi-supervised learning algorithms.

Besides, we compare our algorithms with the new proposed $\mathfrak{U}$-SemiAdaboosts.MH algorithms [28] and some other universum algorithms on the text clustering task. The experimental results show that our proposed algorithms conquer the $\mathfrak{U}$-SemiAdaboosts.MH on three experiments of four, and our algorithms have a more stable performance than $\mathfrak{U}$-SemiAdaboosts.MH and other universum algorithms.

The contributions of this paper can be summarized as follows:

- We impose the column orthogonal constraints on the mapping operator of ORP and propose the OORP algorithm which gives better classification performance than that of the original ORP algorithm.
- We extend our OORP to kernelized version and propose the KOORP algorithm. An efficient optimization algorithm is designed to solve the KOORP.
- We also extend the OORP to the Laplacian regularized OORP (LOORP) algorithm.

The remainder of the paper is constructed as follows: we briefly review the optimal reverse prediction and orthogonal $k$-means [24,23] in Section 2. We propose our orthogonal optimal reverse prediction and its two extensions, kernelized and Laplacian regularized optimal reverse prediction models in Section 3. Algorithm interpretation and performance evaluation through experimental results are given in Section 4, and in Section 5 we compare the computation complexity of our algorithms with the baseline algorithms and the new proposed $\mathfrak{U}$-SemiAdaboosts.MH [28] as well as other universum algorithms. Finally, we conclude our work in Section 6.

## 2. Related work

In this section we will review the optimal reverse prediction proposed by Xu et al. in [21] and the orthogonal $k$-means algorithm given in [24,23], which are closely related to our work.

### 2.1. Optimal reverse prediction (ORP)

Assuming that we are given the input data $X = (X^{(L)}, X^{(U)})$ in a $d \times (N^{(L)} + N^{(U)})$ matrix, which contains the labeled data set $X^{(L)} \in \mathbb{R}^{d \times N^{(L)}}$ with corresponding label $Y^{(L)} \in \mathbb{R}^{k \times N^{(L)}}$, and the unlabeled data set $X^{(U)} \in \mathbb{R}^{d \times N^{(U)}}$, where $N^{(L)}$ denotes the number of labeled instances, $N^{(U)}$ denotes the number of unlabeled instances, $d$ is the dimensionality of instances and $k$ is the number of classes in the classification problem. For the regression problems $k$ is the targets' dimensionality. Different from previous clustering model using the categorical value to present the clustering assignment, label variable $Y^{(L)}$ adopts the 1-of-$k$ encoding scheme, each column in $Y^{(L)}$ indicates the class label of the corresponding data points in $X^{(L)}$. In another words, $Y^{(L)} \in \{0, 1\}^{k \times N^{(L)}}$ and $Y^{\top}\mathbf{1} = \mathbf{I}$, here $\mathbf{1}$ is a column vector with entries are all 1 and $\mathbf{I}$ is the identity matrix of appropriate dimension. We adopt $Y^{(U)}$ to denote the predicted label matrix corresponding to the unlabeled data, $Y^{(U)}$ use as same encoding scheme as $Y^{(L)}$. Adopting the 1-of-$k$ encoding scheme for $Y$ as in [21] has two advantages. First, it is convenient for us to unify the classification and regression problem with one same equation. Second, it is feasible for us to introduce the reverse prediction and unify supervised and unsupervised learning algorithms in a same framework. In ORP learning problem, $X^{(L)}$, $Y^{(L)}$, $X^{(U)}$ are known and $Y^{(U)}$ is initially unknown.

Conventionally, training a supervised learning model often involves finding parameters $W$ for $f_W : X^{(L)} \to Y^{(L)}$ that minimizes some loss function with respect to the targets. Here call $W$ the forward coefficient. Take the least square model as an example, training supervised model is equal to minimizing the following objective function with respect to forward coefficient $W$:

$$\min_{W} \operatorname{tr}\left(\left(WX^{(L)} - Y^{(L)}\right)^{\top}\left(WX^{(L)} - Y^{(L)}\right)\right) \tag{1}$$

Here, $\operatorname{tr}(\cdot)$ denotes the trace of a matrix. We call (1) least square forward prediction, which means we predict label $Y^{(L)}$ based on the input data $X^{(L)}$. Conversely, the following equation is called the least square reverse prediction, which predicts the input data $X^{(L)}$ from the target $Y^{(L)}$: