Applied Soft Computing 58 (2017) 245–255

Contents lists available at ScienceDirect

Applied Soft Computing

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

# Semi-supervised classification by discriminative regularization

Jun Wang, Guangjun Yao, Guoxian Yu*

*College of Computer and Information Science, Southwest University, Chongqing 400715, China*

ARTICLE INFO

*Article history:*
Received 6 September 2016
Received in revised form 26 March 2017
Accepted 19 April 2017
Available online 27 April 2017

*Keywords:*
Semi-supervised classification
Manifold assumption
Discriminative graph
Neighborhood graph
Discriminative regularization

ABSTRACT

One basic assumption in graph-based semi-supervised classification is manifold assumption, which assumes nearby samples should have similar outputs (or labels). However, manifold assumption may not always hold for samples lying nearby but across the boundary of different classes. As a consequence, samples close to the boundary are quite likely to be misclassified. In this paper, we introduce an approach called semi-supervised classification by discriminative regularization (SSCDR for short) to address this problem. SSCDR first constructs a $k$ nearest neighborhood graph to capture the local manifold structure of samples, and a discriminative graph to encode the discriminative information derived from constrained clustering on labeled and unlabeled samples. Next, it separately treats the discriminative graph and the neighborhood graph in a discriminative regularization framework for semi-supervised classification, and forces nearby samples across the boundary to have different labels. Experimental results on various datasets collected from UCI, LibSVM and facial image datasets demonstrate that SSCDR achieves better performance than other related methods, and it is also robust to the input values of parameter $k$.

## 1. Introduction

In many practical tasks of pattern recognition and data mining, the acquisition of sufficient labeled data is quite expensive and time consuming, while large volume of unlabeled data is easy to obtain. Applying supervised classifiers (i.e., $k$ nearest neighbor ($k$NN) [1], support vector machine (SVM) [2]) on scarce labeled samples often results in poor generalization capability. On the other hand, the valuable information embedded in the labeled samples is wasted when applying unsupervised learning techniques (i.e., $k$-means [3], spectral clustering [4]). Consequently, semi-supervised learning, which can take advantage of scarce labeled samples and abundant unlabeled samples to boost the learning performance, has been attracting increasing interests in various domains [5–8].

In this paper, we focus on semi-supervised classification (SSC), and more specifically on graph-based SSC (GSSC). GSSC takes the labeled and unlabeled samples as nodes in a graph, and weights of edges as the similarity between samples. GSSC often uses a graph based regularization term to exploit labeled and unlabeled samples [5]. Zhu et al. [9] proposed the Gaussian random fields and harmonic function to classify unlabeled samples by label propagation on a $k$NN graph. Zhou et al. [10] introduced a local and global consistent method on a $k$NN graph to classify unlabeled samples in this graph. In essence, most graph-based semi-supervised classification methods are transductive, and can not directly classify new samples outside of the graph [5]. Belkin et al. [11] proved that a neighborhood graph can be used to discretely approximate the local manifold structure of samples, and introduced a general semi-supervised learning framework called manifold regularization. Different from these two representative transductive methods [9,10], this framework can directly apply to new samples outside of the graph.

The adopted graph is one of the key factors that determine the performance of GSSC [5]. Some researchers move towards graph optimization based semi-supervised classification [12–15]. Liu and Chang [12] tried to optimize a symmetry-favored $k$NN graph and showed the optimized graph can boost the performance of GSSC. Wang et al. [13] minimized the locally linear reconstruction error regularized by $\ell_2$ norm to construct an $l_2$ graph, and introduced a linear neighborhood propagation classifier on the graph. To improve the performance of GSSC in high-dimensional data, Yan and Wang [14] firstly sought the sparse representation coefficient vectors by optimizing an $\ell_1$ norm regularized sparse representation problem [16], and then constructed an $\ell_1$ graph based on these coefficients. Similarly, Fan et al. [15] introduced a sparsity regularized least square classification method. Zhao et al. [17] integrated sparse representation with low rank representation [18] for semi-supervised classification. Zhao et al. [19] recently proposed a method called compact graph based semi-supervised learning, which infers labels of unlabeled samples by using a compact graph.

* Corresponding author.
*E-mail address:* gxyu@swu.edu.cn (G. Yu).

http://dx.doi.org/10.1016/j.asoc.2017.04.041
1568-4946/

Particularly, this graph is constructed by utilizing neighborhood samples of a sample and reciprocal neighborhood samples of its neighbors. Most GSSC methods perform well on low-dimensional samples but not so on high-dimensional samples [20]. The reason is that the graph is affected by redundant and noisy features of high-dimensional samples. To further improve the performance of GSSC on high-dimensional samples, Yu et al. [20] proposed an algorithm called semi-supervised ensemble classification in subspaces. This algorithm constructs neighborhood graphs in randomly partitioned subspaces instead of original space and trains a semi-supervised linear classifier (SSLC) on a *k*NN graph in each subspace, and then it combines these classifiers into an ensemble classifier via majority-voting rule. As well as that, some researchers try to construct multiple graphs on high-dimensional data, and then integrate these graphs into a composite graph for GSSC [21], or combine these graphs by classifier ensemble [22,23].

All the aforementioned techniques can be formulated as label propagation on a graph or multiple graphs under manifold assumption, which assumes that nearby samples on the data manifold are likely to have the same label [5,10]. However, manifold assumption may violate the fact that two nearby samples may be across the boundary of different classes [24]. As a result, methods solely based on the manifold assumption may misclassify the boundary samples of different classes.

Some researchers have been aware of the boundary samples and incorporated dissimilarity or discrimination into GSSC. Goldberg et al. [25] defined a mixed graph by considering both the similarity and dissimilarity between samples, and then adopted manifold regularization on the mixed graph. Wang and Zhang [26] utilized a unsupervised discriminative kernel based on discriminant analysis to derive two specific algorithms, semi-supervised discriminative regularization and semi-parametric discriminative semi-supervised classification. Xue et al. [27] focused on the underlying discriminative and geometrical information of labeled samples rather than only on the smoothness between them, and proposed a supervised method called discriminatively regularized least-squares classification. Wu et al. [24] proposed a method called semi-supervised discriminative regularization (SSDR). SSDR first constructs intra-class and inter-class graphs based on labeled samples, and then optimizes the corresponding intra-class compactness and inter-class separation, along with the smoothness on a *k*NN graph. Zheng and Ip [28] incorporated similarity and dissimilarity between samples into semi-supervised classification, and developed a graph-based label propagation framework.

All the aforementioned methods only utilize the discrimination derived from labeled samples, while the labeled samples usually are rather limited. To further improve the performance of GSSC, Wang et al. [29] constructed a discriminative graph by using unsupervised clustering over all samples and the prior label information of labeled samples, and then exploited this graph to develop a discrimination-aware Laplacian regularized least square classifier (DA-LapRLSC). DA-LapRLSC can only utilize the discriminative information to some extent, since it ignores labeled samples in the process of clustering, and requires to set an important but vulnerable threshold.

In this paper, we attempt to improve the performance of GSSC based on manifold assumption, and propose a method called semi-supervised classification by discriminative regularization (SSCDR in short). SSCDR firstly constructs a *k*NN graph to capture the local manifold structure of samples, and a discriminative graph to capture the discriminative information of samples by constrained *k*-means [30] on all training samples, including labeled and unlabeled ones. Next, it incorporates these two graphs into a discriminative regularization framework for semi-supervised classification. The empirical study on various datasets collected from University of California at Irvine (UCI) machine learning repository, Library for Support Vector Machines (LibSVM), and publicly available facial images shows that SSCDR achieves significantly better performance than DA-LapRLSC and other related methods. In addition, the comparative study shows that the integration of discriminative information with manifold assumption does not necessarily always enhance the prediction accuracy.

Several features distinguish SSCDR from GSSC methods based on graph optimization. They are listed as follows:

(i) SSCDR constructs a discriminative graph to encode the discriminative information derived from constrained clustering, it can avoid misclassifying the boundary samples of different classes to some extent and improve the performance of GSSC.
(ii) Unlike SSDR that utilizes intra-class and inter-class graphs based on only labeled samples to capture discriminative information between samples, SSCDR utilizes a discriminative graph constructed by constrained *k*-means [30] on all training samples (including labeled and unlabeled ones) to employ global discriminative information.
(iii) Different from DA-LapRLSC and SSDR that utilizes graph Laplacian matrix [31] defined by a mixed graph, SSCDR separately treats the *k*NN graph and discriminative graph in a discriminative framework. As well as that, SSCDR is a linear classifier that can directly apply to new samples without storing all the training samples, whereas DA-LapRLSC is a nonlinear one, which has to store all the training samples and to compute the similarity between new samples and these training samples for prediction.

The structure of this paper is organized as follows. Section 2 provides more detailed review of SSDR and DA-LapRLSC, and motivates our SSCDR. Section 3 firstly gives a toy example to illustrate the importance of discriminative regularization and then elaborates on the proposed SSCDR. Section 4 provides the experimental results and analysis, followed with conclusions and future work in Section 5.

## 2. Related work

One basic assumption in GSSC is manifold assumption, which states that nearby samples are likely to have the same label. Based on this assumption, the learned classifier should be smooth on the intrinsic data manifold [26]. However, a classifier may be not always consistently smooth, especially for samples lying nearby but across the decision boundary of different classes [27]. Given that, the smoothness constraint may take the risk of ignoring the discrimination among boundary samples, and mislead the classification of nearby samples lying across the decision boundary of different classes. However, since traditional regularization methods do not encode underlying discriminative information (within/between class or cluster) in the design of a classifier, they neglect to utilize the useful discriminative information for classification [27]. Some researchers recognized that the underlying discriminative information of samples can be used to mitigate the aforementioned issue of manifold assumption to some extent [24,27–29]. To be self-inclusive, here we re-introduce two representative solutions SSDR [24] and DA-LapRLSC [29], which will be used for comparative study in the experiments.

### 2.1. Semi-supervised discriminative regularization (SSDR) [24]

Suppose among $N$ samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, the first $l$ samples are labeled and the remaining $u = N - l$ samples are unlabeled. $\mathbf{y} = [y_1, y_2, \ldots, y_l]^T$ is the label vector of $l$ labeled samples, and $y_i = c$ $(1 \le c \le C)$ means the $i$th sample ($\mathbf{x}_i$) belonging to the $c$th class, and