Pattern Recognition 42 (2009) 93 – 104

Contents lists available at ScienceDirect

Pattern Recognition

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

Discriminatively regularized least-squares classification

Hui Xue[a], Songcan Chen[a,*], Qiang Yang[b]

[a]Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, PR China
[b]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Article history:
Received 27 March 2008
Received in revised form 15 July 2008
Accepted 16 July 2008

Keywords:
Classifier design
Discriminative information
Manifold learning
Pattern recognition

ABSTRACT

Abstract

Over the past decades, regularization theory is widely applied in various areas of machine learning to derive a large family of novel algorithms. Traditionally, regularization focuses on smoothing only, and does not fully utilize the underlying *discriminative* knowledge which is vital for classification. In this paper, we propose a novel regularization algorithm in the least-squares sense, called discriminatively regularized least-squares classification (DRLSC) method, which is specifically designed for classification. Inspired by several new geometrically motivated methods, DRLSC directly embeds the discriminative information as well as the local geometry of the samples into the regularization term so that it can explore as much underlying knowledge inside the samples as possible and aim to maximize the margins between the samples of different classes in each local area. Furthermore, by embedding equality type constraints in the formulation, the solutions of DRLSC can follow from solving a set of linear equations and the framework naturally contains multi-class problems. Experiments on both toy and real world problems demonstrate that DRLSC is often superior in classification performance to the classical regularization algorithms, including regularization networks, support vector machines and some of the recent studied manifold regularization techniques.

1. Introduction

Regularization methods for machine learning have made great progress recently. Such methods have been extended to several subareas of machine learning, including regression, clustering and classification [1–9].

A related area under extensive development is the manifold learning area, where methods have been developed to take advantage of the locality information while performing dimensionality reduction. In this area, Belkin et al. [5,10] further introduced the underlying sample distribution information of the data with manifold structures into the traditional regularization, resulting in manifold regularization (MR), which aims to retain the manifold structure of the samples in each given class. In the framework of MR, two regularization terms are introduced: one controls the complexity of the classifier, and the other controls the complexity measured by the manifold geometry of the sample distribution [5].

However, when focusing on classification problems, we notice that each of the above methods alone suffers from some deficiencies. First, although the traditional regularization methods have been widely applied to the classifier design, it is essentially derived from multivariate functional fitting or regression problems instead of classification problems [2,11–13]. It constructs the regularization term by focusing more on the smoothness of the function. However, in classification, similar inputs near the discriminant boundaries are more likely to belong to different classes, implying that just a smoothness constraint may not be sufficient for discrimination among classes. In particular, a classifier may not be always smooth everywhere, especially when we are near the boundaries between classes. Furthermore, the primary goal of classification is to separate the samples of different classes in the output space as far as possible. Hence, the underlying *discriminative* information is crucial for classification. However, since the regularization terms of the traditional regularization methods do not inject more underlying class information in a classifier's design, they may not incorporate all the useful discriminative information for classification.

Second, although MR performs well in semi-supervised learning such as sensor networks [14], for supervised learning, MR suggests constructing a graph or Laplacian matrix for each class, which results

* Corresponding author. Tel.: +86 25 84896481x12106; fax: +86 25 84498069.
E-mail addresses: xuehui@nuaa.edu.cn (H. Xue), s.chen@nuaa.edu.cn (S. Chen), qyang@cse.ust.hk (Q. Yang).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.07.010

in an equal number of the regularization terms that is the same as the number of classes. As a result, dependency on the number of given classes makes MR difficult to scale well. The algorithm may perform badly in cases of small number of classes (e.g., three or so classes), whereas the computational complexity in the training phase of MR will increase sharply, because making an optimal tuning for the many regularization parameters is impractical.

In this paper, we propose a novel method for classification that, by the well-known "No Free Lunch" Theorem [15], integrates as much underlying knowledge inside the samples as possible, including the discriminative and geometrical information, into a unified regularization framework. We call our method DRLSC, which stands for discriminatively regularized least-squares classification. By making the best of the underlying discriminative information rather than only emphasizing the smoothness of the classifier in the traditional regularization methods, DRLSC introduces a new discriminative regularization term in the framework. Furthermore, inspired by the new supervised dimensionality reduction methods, DRLSC also uses two graphs to characterize the intra-class compactness and inter-class separability, respectively, and thus can further maximize the margins between the samples of the different classes in each local area. DRLSC integrates the underlying *discriminative* and *geometrical* information into a single regularization term. A major advantage is that it can scale well with the number of the classes. In addition, by introducing the equality constraints in the formulation, the solutions of DRLSC can be found by solving a set of linear equations, which makes the algorithm simpler and more stable. Experiments are conducted to demonstrate the superiority of our DRLSC algorithm compared well with the state-of-the-art regularization methods such as regularization networks (RN), generalized radial basis function networks (GRBFN), support vector machines (SVM), least squares support vector machines (LS-SVM) and manifold regularization (MR).

The rest of the paper is organized as follows. Section 2 introduces the related works in regularization. Our contributions are simply described in Section 3. Section 4 presents the proposed DRLSC. The analytic solution to DRLSC is derived in Section 5. In Section 6, the experiment analysis is given. Some conclusions are drawn in Section 7.

2. Related works

Ill-posed problems widely exist in science and engineering regions, which denotes that given the available input samples, the solution to the problem is nonunique or unstable [2,16]. Early in the 1960s, Tikhonov had proposed a classical method named *regularization* to solve these problems [17,18]. By incorporating the right amount of prior information into the formulation, the regularization techniques have been shown to be powerful in making the solution stable [2,16]. In the past few decades, the regularization theory was introduced to the machine learning community on the premise that the learning can be viewed as a multivariate functional fitting problem [2,11–13]. Consequently, in the classical Tikhonov regularization, the most common form of prior information involves the assumption that the input–output mapping function, i.e., the solution to the fitting problem, is *smooth* [16,19]

$$\min_{f\in F}\left\{\frac{1}{2}\sum_{i=1}^{N}V(y_i,f(\mathbf{x}_i))+\frac{\lambda}{2}\|Df\|^2\right\} \tag{1}$$

where $V(y_i,f(\boldsymbol{x}_i))$ is the loss function, which indicates the penalty we pay when we see $\boldsymbol{x}_i$, predict $f(\boldsymbol{x}_i)$, and the true value is y_i [7]. In the regularization term, D is a linear differential operator that is applied to the function f, in which the prior information about the form of the solution is embedded [16]. D is also referred to as a stabilizer because the smoothness prior involved in it makes the solution stable [2,16]. Moreover, the regularization parameter λ controls the trade-off between fitting the training samples and the roughness of the solution [2,7].

Tikhonov [17,18] presented that when the loss function is designated to be the simple square-loss function

$$V(y,f(\mathbf{x}))=(y-f(\mathbf{x}))^2 \tag{2}$$

the solution $f_\lambda(\mathbf{x})$ to the Tikhonov regularization problem can be represented as a linear combination of the Green's function [16]

$$f_\lambda(\mathbf{x})=\sum_{i=1}^{N}w_iG(\mathbf{x},\mathbf{x}_i) \tag{3}$$

Poggio and Girosi [11,12] showed that a regularization algorithm for learning is equivalent to a multilayer neural network with the Green's function as the activation function, resulting in the RN. Haykin [16] indicated that if we select a multivariate Gaussian function as the Green's function, the solution by RN will be an optimal interpolant in the sense that it minimizes the Tikhonov regularization formula. GRBFN is an approximation of the RN, for its number of the hidden units is typically less than that of the RN's, which is equivalent to the number of the training samples.

In the classical regularization theory, a recent trend in studying the smoothness of the function is to put the function into the reproducing kernel Hilbert space (RKHS) [6,20], which has been well developed in several areas [2]. In the RKHS, the Tikhonov minimization problem can be rewritten as [21]:

$$\min_{f\in H}\left\{\frac{1}{2}\sum_{i=1}^{N}V(y_i,f(\mathbf{x}_i))+\frac{\lambda}{2}\|f\|_K^2\right\} \tag{4}$$

Following the so-called Representer Theorem [20,22,23], under very general conditions on the loss function V, the minimizer of Eq. (4) will have the form

$$f(\mathbf{x})=\sum_{i=1}^{N}c_iK(\mathbf{x},\mathbf{x}_i) \tag{5}$$

Corresponding to different selections of V, the classical Tikhonov regularization method can be used to derive a large family of the state-of-the-art algorithms in machine learning. When selecting V as the square-loss function, we obtain regularized least-squares classification (RLSC) [21]. Similarly, we can obtain SVM [1,24] by choosing V to be the hinge-loss function defined as

$$V(y,f(\mathbf{x}))=\begin{cases}0 & \text{if } yf(\mathbf{x})\geqslant 1\\ 1-yf(\mathbf{x}) & \text{otherwise}\end{cases} \tag{6}$$

Specifically, if we introduce error terms into the hinge-loss function and consider the equality constraints instead of inequalities in SVM, we obtain the LS-SVM with the formulation in the least-square sense [25]. Though introducing dissimilar loss functions, these regularization algorithms have many inherently similar properties. Evgeniou et al. [26] described a unified framework for RN and SVM. Rifkin [21] indicated that RLSC empirically performs as good as SVM.

Although traditional regularization has been widely applied to the classifier design, it focuses more on the smoothness of the classification function owing to the essential derivation from ill-posed multivariate functional fitting problems as we mentioned above, to enforce the constraint that similar inputs correspond to similar outputs. This constraint is natural for regression problems. But, it also seems to be too general for classification. Since the regularization terms of the traditional regularization methods do not inject more underlying class information, they may not incorporate all the useful discriminative information for classification.