Pattern Recognition 43 (2010) 2466–2475

Contents lists available at ScienceDirect

Pattern Recognition

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

A regularization framework for multiclass classification: A deterministic annealing approach

Zhihua Zhang [a,*], Gang Wang [b], Dit-Yan Yeung [b], Guang Dai [b], Frederick Lochovsky [b]

[a] College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China
[b] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China

ARTICLE INFO

Article history:
Received 7 March 2009
Received in revised form
27 September 2009
Accepted 2 February 2010

Keywords:
Multiclass classification
Deterministic annealing
Maximum entropy
Fisher discriminant analysis
Logistic regression

ABSTRACT

In this paper, we propose a general regularization framework for multiclass classification based on discriminant functions. Since the objective function in the primal optimization problem of this framework is always not differentiable, the optimal solution cannot be obtained directly. With the aid of the deterministic annealing approach, a differentiable objective function is derived subject to a constraint on the randomness of the solution. The problem can be approximated by solving a sequence of differentiable optimization problems, and such approximation converges to the original problem asymptotically. Based on this approach, class-conditional posterior probabilities can be calculated directly without assuming the underlying probabilistic model. We also notice that there is a connection between our approach and some existing statistical models, such as Fisher discriminant analysis and logistic regression.

1. Introduction

Multiclass classification is pervasive in many machine learning applications. The problem can be formulated precisely as follows. Given a training set of n labeled examples, $\mathcal{T}=\{(\mathbf{x}_i, c_i)\}_{i=1}^n$, where each input vector $\mathbf{x}_i$ is drawn from some input space $\mathcal{X}\subset\mathbb{R}^d$ and the corresponding class label c_i is drawn from some index set $\mathcal{I}=\{1,\ldots,C\}$ with $C\geq 2$, we build a classifier which can be regarded as a mapping $f:\mathcal{X}\to\mathcal{I}$ that assigns a class label in $\mathcal{I}$ to each vector in $\mathcal{X}$. We say a training example $(\mathbf{x}_i, c_i)\in\mathcal{T}$ is correctly classified by the classifier if $f(\mathbf{x}_i)=c_i$.

Many multiclass classification methods have been proposed over the past few decades. Generally speaking, these methods can be grouped into two main categories:

- The first category is referred to as the *single-machine* approach [1] which addresses the problem by formulating it as a single optimization problem to solve. Logistic regression [2], Fisher discriminant analysis (FDA) [3] and relevance vector machine (RVM) [4] are some examples that belong to this category using the probabilistic approach. In recent years, due to the success of support vector machines (SVM) [5] for binary classification applications, many methods have been proposed to extend the original two-class SVM for multiclass classification, e.g., [6–10].
- The second category includes *one-vs-all* (OVA), *all-vs-all* (AVA) and *error-correcting* [11] approaches. The common characteristic of these approaches is that multiple binary classifiers are trained separately. When we classify a new example, its class label is determined by the outputs from these binary classifiers. The OVA approach requires training C real-valued binary classifiers, each of which is for distinguishing the examples in one class from those in all other classes. For AVA, $\binom{C}{2}$ binary classifiers are trained and each classifier is used for separating a pair of classes. The error-correcting approach uses ideas from the error-correcting coding theory. It chooses a collection of binary classifiers for training and uses a strategy to generate class labels by combining the results from binary classifiers. An obvious drawback of these approaches is the high computation cost required for training many binary classifiers. Different multiclass classification algorithms have been compared [12–14]. Based on many carefully controlled experiments, Ref. [1] defended the OVA approach by showing that OVA can be as accurate as any other state-of-the-art multiclass classification method provided that the underlying binary classifiers are well tuned.

In this paper, we propose a novel regularization framework for multiclass classification based on discriminant functions. Our method belongs to the first category based on the single-machine approach. It is a general formulation in the sense that it can work with different loss functions and discriminant functions. The empirical risk function we try to minimize is not differentiable,

* Corresponding author. Tel.: +86 571 8795 3053; fax: +86 571 8795 1250.
E-mail addresses: zhzhang@zju.edu.cn (Z. Zhang), daiguang116@gmail.com (G. Dai).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.02.003

thus, the optimization for the solution is not straightforward. We apply deterministic annealing to solve this optimization problem.

The deterministic annealing approach has demonstrated substantial performance improvement in many clustering, classification and optimization problems [15–19]. It is an attractive approach with two important advantages: (a) It can minimize a cost function even when its gradient vanishes almost everywhere. (b) It can avoid many poor local minima in the cost function. The deterministic annealing approach for optimization has been strongly inspired by analogies to statistical physics [20]. It regards the optimization problem in question as a thermal system, which has a temperature parameter T to control the level of randomness of the system and the objective function corresponds to the free energy of the system. The minimum of the free energy determines the state of the system at thermal equilibrium. To achieve the equilibrium state, one tracks the minimum of the free energy while gradually lowering the temperature. At the limit of zero temperature, the minimum free energy is reached. In other words, deterministic annealing performs an annealing process as it maintains the objective function at its minimum while gradually lowering the temperature. With careful annealing, this process can avoid many shallow local minima of the objective function and can finally lead to a non-random solution.

In machine learning, most classification algorithms may be considered as performing either soft or hard classification [21]. Soft classification, such as logistic regression, has the form of a posterior probability in its formulation. On the contrary, hard classification, such as least square regression or SVM, does not use a probabilistic formulation. Its output does not indicate the posterior probability that a point belongs to a certain class. A number of methods have been proposed for mapping predictions from hard classification to posterior probabilities [22,23]. However, such posterior probability estimates via postprocessing are unreliable in many cases [4]. In this paper, we seek to bridge the gap between these two kinds of classification directly from the basis. We first formulate the multiclass classification problem in a hard notion where each point belongs to a category with probability of either 0 or 1. Through the deterministic annealing method, the posterior probabilities are introduced during the optimization procedure without assuming an underlying probabilistic model. We will discuss two implementations of this general approach. It is interesting to note that there is a connection between our algorithms and some statistical models such as Fisher discriminant analysis and logistic regression. It should be pointed out here that for the sake of simplicity, this paper only discusses the linear discriminant functions. Similar to [24,25], the proposed implementations can be easily extended to the corresponding nonlinear cases by employing the so-called kernel tricks. In addition, we illustrate in the experiments that our algorithm outperforms the one-vs-all SVM algorithm for most of the data sets, especially when the number of classes is large.

The remainder of this paper is organized as follows. In Section 2, we formulate the problem by devising a regularization framework for multiclass classification and then apply deterministic annealing to solve the optimization problem in Section 3. In Section 4, we consider two choices of the discriminant functions leading to two different realizations of the proposed method. Section 5 reports the experimental setup and results. The last section presents some concluding remarks.

2. Problem formulation

A classifier is formulated in terms of a set of C discriminant functions $\mathbf{g}(\mathbf{x})=\{g(\mathbf{x};\boldsymbol{\theta}_j)|g:\mathcal{X}\to\mathbb{R},\ j=1,\ldots,C\}$, where the parameter $\boldsymbol{\theta}_j$ is a column vector. $\boldsymbol{\theta}=(\boldsymbol{\theta}_1^T,\ldots,\boldsymbol{\theta}_C^T)^T$ is the vector integrating the parameter of each discriminant function. An input vector $\mathbf{x}$ is assigned to a class c if and only if [1]

$$g(\mathbf{x};\boldsymbol{\theta}_c)>g(\mathbf{x};\boldsymbol{\theta}_j)\quad\text{for all } j\neq c. \tag{1}$$

Here the discriminant functions are general and we will discuss some possible implementations in Section 4. The classification rule in (1) is an example of hard classification [26], which assigns an input vector to a class with a probability of either 0 or 1. If we denote the posterior probability of class c given $\mathbf{x}_i$ by $p_{ic}=p(c|\mathbf{x}_i)$, the hard classification simply implies that

$$p_{ic}=\begin{cases}1 & \text{if } c=\arg\max_j g(\mathbf{x}_i;\boldsymbol{\theta}_j),\\ 0 & \text{otherwise.}\end{cases} \tag{2}$$

For convenience, the class label c_i of $\mathbf{x}_i$ is represented as a C-dimensional binary vector $\mathbf{y}_i=(y_{i1},y_{i2},\ldots,y_{iC})'$, with all values equal to 0 except the c_ith position which is equal to 1. We shall use c_i and $\mathbf{y}_i$ interchangeably in this paper to denote the class label of $\mathbf{x}_i$. Let $\mathbf{g}_i=(g(\mathbf{x}_i;\boldsymbol{\theta}_1),g(\mathbf{x}_i;\boldsymbol{\theta}_2),\ldots,g(\mathbf{x}_i;\boldsymbol{\theta}_C))'$ and $\|\mathbf{g}_i\|_\infty=\max_{1\le j\le C}\{g(\mathbf{x}_i;\boldsymbol{\theta}_j)\}$. For a hypothesis indexed by the parameter $\boldsymbol{\theta}$, if a vector $\mathbf{x}_i$ is correctly classified according to (1), then we have $\|\mathbf{g}_i\|_\infty=\mathbf{y}_i'\mathbf{g}_i$. Otherwise we have $\|\mathbf{g}_i\|_\infty>\mathbf{y}_i'\mathbf{g}_i$. This observation motivates us to define the empirical loss as

$$L(\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^n;\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^n l(\|\mathbf{g}_i\|_\infty-\mathbf{y}_i'\mathbf{g}_i), \tag{3}$$

where $l(\cdot)$ is the loss function specified by the user. Here the loss function is pointwise as we assume that the data points are independent and identically distributed (i.i.d.). One possibility is to define the loss function in terms of the following misclassification error:

$$L_1(\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^n|\|\mathbf{g}_i\|_\infty-\mathbf{y}_i'\mathbf{g}_i|=\frac{1}{n}\sum_{i=1}^n(\|\mathbf{g}_i\|_\infty-\mathbf{y}_i'\mathbf{g}_i). \tag{4}$$

If a training example $(\mathbf{x}_i,c_i)$ is correctly classified by the classifier, it will attain its minimum value of zero. Otherwise, we pay a penalty which is equal to the difference between the largest $g(\boldsymbol{\theta}_j)$ $(\forall j\neq c_i)$ and $g(\boldsymbol{\theta}_{c_i})$.

To solve the optimization problem for multiclass classification under a regularization framework, we minimize the following regularized risk function:

$$R(\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^n l(\|\mathbf{g}_i\|_\infty-\mathbf{y}_i'\mathbf{g}_i)+\lambda\sum_{j=1}^C\|\boldsymbol{\theta}_j\|_p, \tag{5}$$

where $\|\cdot\|_p$ is the L_p-norm with p typically being 1 or 2. Since the value of $\|\mathbf{g}_i\|_\infty$ in (5) depends on the output of each discriminant function, $R(\boldsymbol{\theta})$ is not differentiable with respect to $\boldsymbol{\theta}$ and hence cannot be optimized directly. In this paper, we employ the deterministic annealing approach to solve the optimization problem.

3. Deterministic annealing approach

In spirit, deterministic annealing is similar to penalty methods in the optimization literature [27]. For a general optimization problem

$$\min_{\mathbf{u}} f(\mathbf{u}), \tag{6}$$

we first introduce a family of probability distributions $\mathcal{P}$ from $[0,1]^m$, where m is problem-dependent, in order that

$$f(\mathbf{u})=f(\mathbf{u},\mathbf{p}^*)=\max_{\mathbf{p}\in\mathcal{P}} f(\mathbf{u},\mathbf{p}). \tag{7}$$

[1] We assume for simplicity that the discriminant functions give different values.