Engineering Applications of Artificial Intelligence 37 (2015) 1–8

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Maximum mutual information regularized classification

Jim Jing-Yan Wang [a], Yi Wang [b], Shiguang Zhao [c], Xin Gao [a,*]

[a] Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
[b] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA
[c] Department of Neurosurgery, The First Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang 150001, PR China

ARTICLE INFO

Article history:
Received 13 March 2014
Received in revised form
13 August 2014
Accepted 14 August 2014
Available online 7 September 2014

Keywords:
Pattern classification
Maximum mutual information
Entropy
Gradient descend

ABSTRACT

In this paper, a novel pattern classification approach is proposed by regularizing the classifier learning to maximize mutual information between the classification response and the true class label. We argue that, with the learned classifier, the uncertainty of the true class label of a data sample should be reduced by knowing its classification response as much as possible. The reduced uncertainty is measured by the mutual information between the classification response and the true class label. To this end, when learning a linear classifier, we propose to maximize the mutual information between classification responses and true class labels of training samples, besides minimizing the classification error and reducing the classifier complexity. An objective function is constructed by modeling mutual information with entropy estimation, and it is optimized by a gradient descend method in an iterative algorithm. Experiments on two real world pattern classification problems show the significant improvements achieved by maximum mutual information regularization.

1. Introduction

The pattern classification problem is a problem of assigning a discrete class label to a given data sample represented by its feature vector (Cai et al., 2014; Ojala et al., 2002; Sun et al., 2014; Li et al., 2014; Li et al.). It has many applications in various fields, including bioinformatics (Alipanahi et al., 2009; Wang et al., 2013a; J.J. Wang et al., 2012; J. Wang et al., 2012; Liu et al., 2012), biometrics verification (Wang, 2009; Roy et al., 2011; Tafazzoli et al., 2010), computer networks (Yang et al.; Xu et al., 2014, 2013), and computer vision (Cai et al., 2013; Wang et al., 2014a, 2013b; Zhou et al., 2010). For example, in the face recognition problem, given a face image, the target of pattern classification is to assign it to a person who has been registered in a database (Jonathon Phillips et al., 2000; Zhao et al., 2003). This problem is usually composed of two different components — feature extraction (Sun et al., 2012; Zhou et al., 2013; Wang and Gao, 2013; Al-Shedivat et al., 2014; Wang et al., 2014b, 2013c, 2013d; J.-Y. Wang et al., 2012) and classification (Zhou et al.; Subbulakshmi and Afroze, 2013). Feature extraction refers to the procedure of extracting an effective and discriminant feature vector from a data sample, so that different samples of different classes could be separated easily. This procedure is usually highly domain-specific. For example, for the face recognition problem, the visual feature should be extracted using some image processing technologies, whereas for the problem of predicting zinc-binding sites from protein sequences,the biological features should be extracted using some biological knowledge (Chen et al., 2013). In terms of feature extraction of this paper, it is highly inspired by a hierarchical Bayesian inference algorithm proposed in Zhou et al. (2013). This new method created in Sun et al. (2012) has advanced the ground-truth feature extraction field and has provided a more optimal method for this procedure. On the other hand, different from feature extraction, classification is a much more general problem. We usually design a class label prediction function as a classifier for this purpose. To learn the parameter of a classifier function, we usually try to minimize the classification error of the training samples in a training set and simultaneously reduce the complexity of the classifier. For example, the most popular classifier is support vector machine (SVM), which minimizes the hinge losses to reduce the classification error, and at the same time minimizes the ℓ_2 norm of the classifier parameters to reduce the complexity. In this paper, we focus on the classification aspect.

Mutual information (E. Liu et al., 2014; Battiti, 1994) is defined as the information shared between two sets of variables. It has been used as a criterion of feature extraction for pattern classification problems (Sun and Xu, 2014). However, surprisingly, it has never been directly explored in the problem of classifier learning. Actually, mutual information has a strong relation to Kullback-Leibler divergence, and

* Corresponding author. Tel.: +966 12 8080323.
E-mail addresses: jimjywang@gmail.com (J.-Y. Wang), wayi@cse.ohio-state.edu (Y. Wang), guangsz@hotmail.com (S. Zhao), xin.gao@kaust.edu.sa (X. Gao).

http://dx.doi.org/10.1016/j.engappai.2014.08.009
0952-1976/

there are many works using KL-divergence for classifiers (Moreno et al., 2004; Liu and Shum, 2003). Moreno et al. (2004) proposed a novel kernel function for support vector classification based on Kullback-Leibler divergence, while Liu and Shum (2003) proposed to learn the most discriminating feature that maximizes the Kullback-Leibler divergence for the Adaboost classifier. However, both these methods do not use the KL-divergence based criterion to learn parameters of linear classifiers. To bridge this gap, in this paper, for the first time, we try to investigate using mutual information as a criterion of classifier learning. We propose to learn a classifier by maximizing the mutual information $I(f;y)$ between the classification response variable f and the true class label variable y. The classification response variable f is a function of classifier parameters and data samples. The insight is that mutual information is defined as the information shared between f and y. From the viewpoint of information theory, if the two variables are not mutually independent, and one variable is known, it usually reduces the uncertainty about the other one. Then mutual information is used to measure how much uncertainty is reduced in this case. To illuminate how the mutual information can be used to measure the classification accuracy, we consider the two extreme cases:

- On one hand, if the classification response variable f of a data sample is randomly given, and it is independent of its true class label y, then knowing f does not give any information about y and vice versa, and the mutual information between them could be zero, i.e., $I(f;y)=0$.
- On the other hand, if f is given so that y and f are identical, knowing f can help determine the value of y exactly as well as reduce all the uncertainty about y. This is the ideal case of classification, and knowing f can reduce all the uncertainty about y. In this case, the mutual information is defined as the uncertainty contained in f (or y) alone, which is measured by the entropy of f or y, denoted by $H(f)$ or $H(y)$ respectively, where $H(\cdot)$ is the entropy of a variable. Since f and y are identical, we can have $I(f;y)=H(f)=H(y)$.

Naturally, we hope that the classification response f can predict the true class label y as accurately as possible, and knowing f can reduce the uncertainty about y as much as possible. Thus, we propose to maximize the mutual information between f and y with regard to the parameters of a classifier. To this end, we proposed a mutual information regularization term for the learning of classifier parameters. An objective function is constructed by combining the mutual information regularization term, a classification error term and a classifier complexity term. The classifier parameter is learned by optimizing the objective function with a gradient descend method in an iterative algorithm.

The rest parts of this paper are organized as follows: in Section 2, we introduce the proposed classifier learning method. The experiment results are presented in Section 3. In Section 4 the paper is concluded.

2. Proposed method

In this section, we introduce the proposed classifier learning algorithm to maximize the mutual information between the classification response and the true class label.

2.1. Problem formulation

We suppose that we have a training set denoted as $X=\{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i\in\mathbb{R}^d$ is the d-dimensional feature vector for the i-th training sample, and n is the number of training samples. The class label set for the training samples is denoted as $Y=\{y_i\}_{i=1}^n$, where $y_i\in\{+1,-1\}$ is the class label of the i-th sample. To learn a classifier to predict the class label of a given sample with its feature vector $\mathbf{x}$, we design a linear function as a classifier,

$$g(\mathbf{x};\mathbf{w})=\mathrm{sign}(f)=\mathrm{sign}(\mathbf{w}^\top\mathbf{x}), \tag{1}$$

where $\mathbf{w}$ is the classifier parameter vector, $f=\mathbf{w}^\top\mathbf{x}$ is the classification response of $\mathbf{x}$ given the classifier parameter $\mathbf{w}$, and $\mathrm{sign}(\cdot)$ is the signum function which transfers the classification response to the final binary classification result. We also denote the classification response set of the training samples as $F=\{f_i\}_{i=1}^n$ where $f_i=\mathbf{w}^\top\mathbf{x}_i\in\mathbb{R}$ is the classification response of the i-th training sample. To learn the optimal classification parameter $\mathbf{w}$ for the classification problem, we consider the following three problems:

2.1.1. Classification loss minimization

To learn the optimal classification parameter $\mathbf{w}$, we hope the classification response f of a data sample $\mathbf{x}$ obtained with the learned $\mathbf{w}$ can predict its true class label y as accurately as possible. To measure the prediction error, we use a loss function to compare a classification response against its corresponding true class label. Given the classifier parameter $\mathbf{w}$, the loss function of the i-th training sample $\mathbf{x}_i$ with its classification response $f_i=\mathbf{w}^\top\mathbf{x}_i$ and true class label y_i is denoted as $L(f_i,y_i;\mathbf{w})$. There are a few different loss functions which could be considered.

Hinge Loss is used by the SVM classifier (Wu and Liu, 2007; Yildiz and Alpaydin, 2013; Bach et al., 2013), and it is defined as

$$L(f_i,y_i;\mathbf{w})=\max(0,1-y_if_i)=\tau_i\times(1-y_i\mathbf{w}^\top\mathbf{x}_i), \tag{2}$$

where τ_i is defined as

$$\tau_i=\begin{cases}1 & \text{if } y_i\mathbf{w}^\top\mathbf{x}_i\le 1\\ 0 & \text{otherwise.}\end{cases} \tag{3}$$

Squared Loss is usually used by regression problems (X. Wang et al., 2013; Luo, 2012; Luo et al., 2012), and it is defined as

$$L(f_i,y_i;\mathbf{w})=(1-y_if_i)^2=(1-y_i\mathbf{w}^\top\mathbf{x}_i)^2. \tag{4}$$

Logistic Loss is defined as follows, and it is also popular in regression problems (Park et al., 2008),

$$L(f_i,y_i;\mathbf{w})=\log[1+\exp(-y_if_i)]=\log[1+\exp(-y_i\mathbf{w}^\top\mathbf{x}_i)]. \tag{5}$$

Exponential loss is another popular loss function which could be used by both classification and regression problems (X. Wang et al., 2013), which is defined as

$$L(f_i,y_i;\mathbf{w})=\exp(-y_if_i)=\exp(-y_i\mathbf{w}^\top\mathbf{x}_i). \tag{6}$$

Obviously, to learn an optimal classifier, the average loss of all the training samples should be minimized with regard to $\mathbf{w}$. Thus the following optimization problem is obtained by applying a loss function to all training samples,

$$\min_{\mathbf{w}}\frac{1}{n}\sum_{i=1}^n L(f_i,y_i;\mathbf{w}). \tag{7}$$

2.1.2. Classifier complexity reduction

To reduce the complexity of the classifier to prevent the overfitting problem, we also regularize the classifier parameter by a ℓ_2