Contents lists available at ScienceDirect

Neurocomputing

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

Exploiting class label in generative score spaces

Bin Wang [a,c,*], Cungang Wang [b], Yuncai Liu [c]

[a] College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 200234, China
[b] School of Computer Science, Liaocheng University, Liaocheng 252400, China
[c] Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 14 January 2013
Received in revised form
31 March 2014
Accepted 7 May 2014
Communicated by H. Yu

Keywords:
Generative score spaces
Discriminative extension
Classification
Feature mapping

ABSTRACT

Generative score spaces have recently received increasing attention due to their state-of-the-art performance in a wide range of recognition tasks. These methods model the distribution of the training data using probabilistic generative models and derive the feature for each sample based on the generative models. The derived feature encodes the information of the sample, hidden variables and model parameters for classification, providing a staged way to integrate the abilities of generative models in inferring hidden information and discriminative models in classification. The underlying point is that the hidden information carried by hidden variables in generative models is informative and useful in classification. In this paper, we propose a general extension for the existing score space methods to exploit class label that encodes rich discriminative information, when deriving feature mappings. This is achieved by extending the regular generative models to class conditional models over both observed variable and class label, and deriving feature mapping over such extended models. The resulted methods take simple and intuitive forms which are weighted versions of existing methods, benefitting from the Bayesian inference of class label. The empirical evaluation over two typical generative models and 6 datasets shows its significant improvement over existing methods.

1. Introduction

Generative score space is a class of principled methods aiming to integrate the abilities of probabilistic generative models and discriminative models for classification. In these methods, generative models are used to drive feature mappings over observed variables, hidden variables and model parameters while discriminative models are used to perform classification in the derived feature spaces. The motivations of using generative models to derive feature mappings are threefold: (1) generative models are able to discover the information hidden in data by inferring hidden variables, which is additional and useful for classification; (2) generative models are good at dealing with structured data, e.g., variable length sequences or tree structure data; (3) the derived features are fixed dimension and can be straightforwardly delivered to discriminative classifiers for classification. Generative score space is an implementation of hybrid generative–discriminative classification methods [1–6] whose detailed reviews can be found in [5]. In this paper, we focus on generative score space due to its competitive performance in a number of challenging tasks [5], especially in the highly challenging task – image recognition [7].

Generative score space methods, first proposed in [1], can be categorized into two classes [6,5]: parameter based methods and random variable based methods. Let $P(\mathbf{x}|\theta)$ be the marginal distribution of an adopted generative model, where $\mathbf{x} \in \mathbb{R}^D$ is the observed variable and $\theta = \{\theta_1, \ldots, \theta_K\}$ is the set of K parameters. *Parameter based methods* are represented by Fisher Score (FS) [1]. It derives explicit feature mappings from a given generative model. The feature mappings measure how a sample affects the parameters θ, i.e., differential operation over the log likelihood $\log P(\mathbf{x}|\theta)$ with respect to parameters. This method is robust to the number of hidden variables [6], and shows state-of-the-art performance in image recognition [7]. *Variable based methods* include free energy score space (FESS) [8], posterior divergence (PD) [6] and augmented sufficient statistics (SS) [9–11]. FESS and PD derive feature mappings by mainly measuring how well a sample fits the distribution of the random variables, while SS is essentially the expectation over the sufficient statistics.

All the above methods derive feature mappings from the generative model trained using samples from all classes, without making use of the class label. An extension to utilize class label is tangent vector of posterior log-odds (TOP) [2] which trains a pair

[*] Corresponding author at: College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, 100 Guilin Road, Xuhui District, Shanghai 200234, China.
E-mail address: binley.wang@gmail.com (B. Wang).

http://dx.doi.org/10.1016/j.neucom.2014.05.004
0925-2312/© 2014 Elsevier B.V. All rights reserved.

Table 1
Notation lists.

Notation	Description
$\mathbf{x} \in \mathbb{R}^D$	Observed input data
$y \in \{1, \ldots, C\}$	Output label indexed by c
$S = \{\mathbf{x}^t, y^t\}_{t=1}^N$	Training set of N input–output pairs indexed by t
H	Hidden variable set
$P(\mathbf{x}, H\|\theta)$	Joint distribution of a generative model parameterized by θ
$Q(H)$	Approximate posterior for the real posterior $P(H\|\mathbf{x}, \theta)$
$-\mathcal{F}(Q, \theta)$	Variational lower bound for the log likelihood log $P(\mathbf{x}\|\theta)$
$\Phi(\mathbf{x}^t)$	Score function or feature mapping for the sample $\mathbf{x}^t$
$\mathbf{w} \in \mathbb{R}^{D'}, b \in \mathbb{R}$	Weight and bias for the linear classifier

of models $P(\mathbf{x}|\theta_{+1})$ and $P(\mathbf{x}|\theta_{-1})$ using the positive samples and the negative samples respectively. Then TOP derives the feature mappings from the maximum a posteriori (MAP) discriminant function which is defined on the pair of models. Another extension is proposed in [3] which trains C models using C classes of samples respectively (i.e., train the model $P(\mathbf{x}|\theta_y)$ using the samples from the class y, where $y \in \{1, \ldots, C\}$). The method then derives C sets of feature mappings from the C trained models respectively, and concatenates them as the final feature mapping. The *assumptions* underlying these two extensions are that all the class-conditional models $P(\mathbf{x}|\theta_y)$ represent samples equally well, and are equally important when deriving feature mappings. However, these assumptions are not always true. For instance, if a sample $\mathbf{x}^t$ takes a higher probability on the model θ_c than on the model $\theta_{c'}$ (i.e., $P(\mathbf{x}^t|\theta_c) > P(\mathbf{x}^t|\theta_{c'})$), the model θ_c should represent the sample $\mathbf{x}^t$ better than the model $\theta_{c'}$. Moreover, a model can be more important than others when it represents more training samples.

In this paper, we propose a discriminative extension for generative score space methods [1,6,5,9–11]. The proposed extension can fully exploit the class label which is informative in classification when deriving feature mapping, but release from the *assumptions* made in [2,3]. The proposed extension models the joint distribution of the observed data $\mathbf{x}$ and its label y, and applies two representative score space methods [1,5] to the joint model respectively. The resulted feature mappings are very simple, i.e., the concatenation of the weighted feature mappings respectively derived from the class-conditional models $\{\theta_1, \ldots, \theta_C\}$. The difference between the proposed extension and the extension in [3] is the weights. The weights measure how important a model is in deriving feature mappings and therefore our extension avoids making the assumptions in [2,3]. Moreover, we prove that, for multi-class classification, the error rate of a zero-one loss linear classifier working with feature mappings derived by the proposed extension is at least as lower as that of maximum a posterior (MAP) classifier. It is worth noting that previous score space methods [1,6,5,9–11] as well as their extensions [2,3] have no such a guarantee.

The remainder of this paper is organized as follows. In Section 2, we review related score space methods and two discriminative extensions. Section 3 proposes our discriminative extension and justifies its error rate. We evaluate the proposed method and related methods in Section 4, and draw a conclusion in Section 5. For readability, we summarize the involved notations in Table 1.

2. Generative score spaces revisit

In this section, we will revisit the generative score space methods. First, we introduce the variational lower bound of the log likelihood function of generative models, on which score space methods work. Then, we review two representative score space methods [1,5] and two discriminative extensions [3,2].

2.1. Variational lower bound of log likelihood function

Let $\mathbf{x} \in \mathbb{R}^D$ be the observed variable and $\mathcal{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$ be a set of N training samples of the variable $\mathbf{x}$. We consider a general case that the probabilistic distribution over $\mathbf{x}$ is modeled by a hierarchical probabilistic generative model with a set of hidden variables H introduced, and is parameterized by a vector of parameters θ. Let $P(\mathbf{x}, H|\theta)$ be the joint distribution and $P(\mathbf{x}|\theta)$ be the marginal distribution. For most generative models, the marginal distribution $P(\mathbf{x}|\theta)$ is unavailable since the integration $\int P(\mathbf{x}, H|\theta)\, dH$ is intractable [12]. A number of approximation methods [13] are developed to attack this problem. The common idea of these methods is to construct an approximate posterior distribution $Q^t(H)$ to estimate the real posterior distribution $P(H|\mathbf{x}^t, \theta)$. Then we have [12,13]

$$\log\ P(\mathbf{x}^t|\theta) = -\mathrm{KL}(Q^t(H) \,\|\, P(\mathbf{x}^t, H|\theta)) + \mathrm{KL}(Q^t(H) \,\|\, P(H|\mathbf{x}^t, \theta)) \tag{1}$$

where KL denotes the Kullback–Leibler divergence. The log likelihood is decomposed into two terms. The second term measures the residual error of using $Q^t(H)$ to approximate $P(H|\mathbf{x}^t, \theta)$ and takes zero when $Q^t(H)$ is expressive enough (for instance, $Q^t(H)$ is given by exact inference). In this case, the first term is the exact log likelihood. As did in [5,6], we focus on the variational inference [12] which resorts to a lower bound of the log likelihood

$$\log\ P(\mathbf{x}^t|\theta) \geq -\mathrm{KL}(Q^t(H) \,\|\, P(\mathbf{x}^t, H|\theta)) = -\mathcal{F}(Q^t, \theta) \tag{2}$$

where $-\mathcal{F}(Q^t, \theta)$ is the variational lower bound; $\mathcal{F}(Q^t, \theta)$ is the free energy function. A choice for $Q^t(H)$ is that it takes the same form with $P(H)$ but with different parameters [12]. Note that the above formulation involves two approximations: (1) using the approximate posterior distribution $Q^t(H)$ to approach the real posterior distribution $P(H|\mathbf{x}^t, \theta)$; (2) using the lower bound $-\mathcal{F}(Q^t, \theta)$ to approach the real log likelihood log $P(\mathbf{x}^t|\theta)$. However, using the above two approximations will not lose generality, because, when $Q^t(H)$ is given by exact inference methods, the approximate posterior $Q^t(H)$ exactly equals to the real posterior $P(H|\mathbf{x}^t, \theta)$, and the lower bound $-\mathcal{F}(Q^t, \theta)$ exactly equals to the real log likelihood log $P(\mathbf{x}^t|\theta)$.

2.2. Generative score space methods

The formulations of generative score space methods are based on the log likelihood function or its variational lower bound. Although their motivations of the score space methods [1,5,6,10] are different, their formulations can be written as an unified expression, i.e., expressing the variational lower bound as the linear combination of the score function of either score space. In this section, we will review two representative score space methods, Fisher score (FS) [1], free energy score space (FESS) [5], and their discriminative extensions.

2.2.1. Fisher score (FS)

Let $\theta = \{\theta_1, \ldots, \theta_K\}$ be the set of K parameters of an adopted generative model. For an input sample $\mathbf{x}^t$, the i-th element $\Phi_i(\mathbf{x}^t)$ of the score function of Fisher score (FS) [1] is defined as the differential of the log likelihood log $P(\mathbf{x}^t|\theta)$ with respect to the i-th model parameter θ_i

$$\Phi_i(\mathbf{x}^t) = \nabla_{\theta_i} \log\ P(\mathbf{x}^t|\theta) \quad \text{or} \ \Phi_i(\mathbf{x}^t) = \nabla_{\theta_i}(-\mathcal{F}(Q^t, \theta)) \tag{3}$$

The right expression is also referred to as gradient FESS (gFESS) in [5]. For brevity, we refer to both expressions as FS, and use the right expression in the following part. The complete score function of FS can be written as

$$\Phi_{\mathrm{FS}}(\mathbf{x}^t) = (\Phi_1(\mathbf{x}^t), \ldots, \Phi_K(\mathbf{x}^t))^T$$

Given the above derivation of FS, the lower bound $-\mathcal{F}(Q^t, \theta)$ can be expressed as the linear combination of the elements of the FS

Please cite this article as: B. Wang, et al., Exploiting class label in generative score spaces, Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2014.05.004