Pattern Recognition 48 (2015) 1225–1234

Contents lists available at ScienceDirect

Pattern Recognition

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

Importance sampling based discriminative learning for large scale offline handwritten Chinese character recognition

Yanwei Wang [a,*], Qiang Fu [b], Xiaoqing Ding [a], Changsong Liu [a]

[a] State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, the Department of Electronic Engineering, Tsinghua University, China
[b] Microsoft Research Asia, China

ARTICLE INFO

Article history:
Received 13 November 2013
Received in revised form
11 September 2014
Accepted 18 September 2014
Available online 7 October 2014

Keywords:
Importance sampling
Discriminative learning
Sample importance weight
Handwritten Chinese character recognition

ABSTRACT

The development of a discriminative learning framework based on importance sampling for large-scale classification tasks is reported in this paper. The framework involves the assignment of samples with different weights according to the sample importance weight function derived from the Bayesian classification rule. Three methods are used to calculate the sample importance weights for learning the modified quadratic discriminant function (MQDF). (1) Rejection sampling method. The method selects important samples as a training subset and trains different levels of MQDFs by focusing on different types of samples. (2) Boosting algorithm. The algorithm modifies the sample importance weights iteratively according to the recognition performance. (3) Minimum classification error (MCE) rule. The parameter of the importance weight function is estimated using the MCE rule. In general, the cursive samples are usually misclassified or prone to be misclassified by the MQDF learned under the maximum likelihood estimation (MLE) rule. The proposed importance sampling framework thereby makes the MQDF classifier focus more on cursive samples than on normal samples. Such a strategy allows the MQDF to achieve higher accuracy while maintaining lower computational complexity. Comprehensive experiments on three Chinese handwritten character datasets demonstrated that the proposed framework exhibits promising character recognition accuracy.

1. Introduction

With the maturity and in-depth study of pattern recognition technologies, research on handwritten Chinese character recognition has attracted increasing attention and interest. There are two milestones in the history of Chinese character recognition. The first milestone is the demonstration of the extraction of a directional element feature (DEF) [1,2] from a binary image, and this achievement was subsequently extended to gray scale images [3,4] later. The directional feature greatly improves the feature representation ability and is currently commonly used for character recognition. The second milestone is the application of the MQDF classifier [5], which significantly improves the character recognition accuracy. Many research results on the recognition of constrained offline handwritten Chinese characters have been reported [6–10]. For example, the highest recognition accuracy reported is 98.56% [7] for the HCL2000 dataset [11]. However, the recognition accuracy of cursive characters is relatively low, being less than 95% [12]. Therefore, many challenges must be overcome in the development of a framework for the recognition of handwritten Chinese characters [13].

The recognition of cursive handwritten Chinese characters is a challenging large scale classification problem. Due to the wide diversity of cursive Chinese characters, the training samples are usually insufficient for training. In addition, the diverse writing styles of cursive characters reduce the inter-class variances and increase the intra-class variances, thereby increasing the ambiguity of character classification.

There are two main types of classification approaches for character recognition: the generative method and discriminative learning. The generative method, e.g., MQDF, mainly models each class as a specific probability distribution. The training process of the generative model is exactly a parameter estimation problem. Maximum likelihood estimation (MLE) and Bayesian estimation are two typical and effective methods used to estimate the parameters. MLE regards the estimated parameter as a certain variable with a fixed value, which enables the model to generate observations with the maximum likelihood probability. For Bayesian estimation, the parameters are considered random variables.

* Correspondence to: The Department of Electronic Engineering, Tsinghua University, qinghuayuan 1, Haidian district, Beijing 100084, China. Tel.: +86 135 0124 6279.
E-mail address: wangyw@ocrserv.ee.tsinghua.edu.cn (Y. Wang).

http://dx.doi.org/10.1016/j.patcog.2014.09.014

Due to its computational simplicity, MLE is widely used in large scale recognition problems. Duda [14] indicated that, in the absence of specific a priori knowledge, the estimation results of these two methods are similar. For cursive Chinese character recognition, the performance of the MQDF using the parameters estimated by MLE is far from satisfactory. The MQDF relies on the assumption that the features of each class satisfy a Gaussian distribution. However, such an assumption is poorly met by cursive characters. Furthermore, the MLE method does not consider discriminative information among different classes. Therefore, the MQDF based on the MLE method does not perform optimally. Discriminative learning is an effective means to improve the classification performance. However, its high computational complexity has limited its application in large scale classification problems. Therefore, discriminative learning for handwritten Chinese character recognition in which both the recognition accuracy and the computational complexity should be taken into consideration remains an open research problem.

To leverage the advantage of discriminative learning with low computational complexity, we propose an importance sampling based discriminative learning framework in which discriminative information is utilized in the MQDF training process. The basic concept is to assign higher weights to the misclassified (or prone to be misclassified) samples and lower weights to the correctly classified samples. Our preliminary idea and experimental results were partially reported in a previous report [15]. The theory and experimental results of our concept are extended and described in more detail in this paper.

The rest of this paper is organized as follows: The related works are briefly reviewed in Section 2. The MQDF is introduced in Section 3, and the importance sampling based discriminative learning framework is described in Section 4. Rejection sampling, boosting and MCE, which are used for estimating the sample importance weights, are introduced in Sections 5, 6 and 7, respectively. The experimental results and a summary of the proposed algorithms are presented in Section 8. Finally, the conclusion is presented in Section 9.

2. Related works

Discriminative learning for large scale classification is an important and challenging research topic in pattern recognition. Many works have contributed to improving the recognition accuracy of handwritten Chinese characters. Although MQDF remains the most widely used classifier for Chinese character recognition, it suffers from the non-Gaussian characteristics of the data distribution. To better represent the data distribution, the Gaussian mixture model (GMM) is adopted [16]. Theoretically, when the order of sub-Gaussian mixtures approximates infinity and there are a sufficient number of training samples, GMM could simulate any probability distribution. However, usually only limited training samples can be collected. In addition, the number of parameters grows drastically with an increase in the order of the mixtures. Therefore, the number of training samples will always seem insufficient for training higher order Gaussian mixtures, which indicates that only few Gaussian mixtures can be successfully trained.

Different from generative methods, discriminative learning methods do not depend on the specific probability assumption. Discriminative learning methods directly build classification boundaries in the feature space. The representative methods include AdaBoost [17], support vector machines (SVM) [18], learning vector quantization (LVQ) [19], neural network (NN) [20] etc. These methods can explore complex interfaces in a feature space and have been successfully applied in small scale classification problems [20,21]. The discriminative learning method was first proposed for two-class classification problem and then extended to multi-class problems [22,23]. SVM must solve the constrained quadratic programming problem and thus suffers from high computational complexity. The fast version of SVM has greatly reduced its computational complexity [24]. Researchers have successfully explored the fast versions of SVM in large scale problems [25]. As reported in the literature, the accuracy of SVM is slightly lower than that of MQDF for Chinese character recognition. However, the training computational complexity remains large, e.g., SVM required over six hundred hours for training on 2,144,489 samples of 3755 classes [26].

The concept of discriminative learning provides a new way to improve the performance of a generative model. Because the generative model is determined by probability distribution parameters, refining the parameters by integrating discriminative information is an effective means to improve the classifier performance. These methods are categorized into two types. One type optimizes the parameters according to an objective function with respect to the classification performance, e.g., minimum classification error (MCE) [27], maximum mutual information (MMI) [28], etc. Some methods use parameters estimated under the MLE as the initial parameters and continue to optimize them under the objective function [29–32]. These methods have a sound theoretical framework and can ensure convergence. However, these methods also suffer from high computational complexity. In reality, usually only part of the parameters are optimized in discriminative learning [7]. The other type of method uses heuristic rules to adjust parameters based on the samples, e.g., mirror image learning (MIL) [33]. MIL attempts to use samples located near the classification boundary to modify the classifier parameters. There is no rigorous theoretical support for convergence in MIL, however, it works effectively for many practical problems.

In Chinese character recognition, the MQDF under MLE is strong enough to recognize most samples successfully. As a result, MLE was chosen as the baseline MQDF in this paper. Many experiments indicate that the error mainly originates from cursive character samples. Further improving the accuracy is still important not only for research but also for real applications. The obtaining of cursive samples can be viewed as rare events if sample acquisition is simulated as a sampling process. In computer vision, importance sampling is used to improve the probability of the rare events and provides an important way to reduce the estimated variance [34]. This approach provides the basic idea for the proposed importance sampling based discriminative learning for large scale classification.

3. MQDF

In this paper, MQDF based on MLE was used as the baseline classifier. MQDF is derived from the quadratic discriminant function (QDF), which is a Bayesian classifier that assumes that the samples of each class satisfy a Gaussian distribution. QDF can be expressed as the following equation:

$$d_i(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \log |\boldsymbol{\Sigma}_i| \tag{1}$$

The mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ are usually estimated by MLE. By applying eigenvalue decomposition on $\boldsymbol{\Sigma}_i$, the QDF classifier can be rewritten as [5]

$$d_i(\boldsymbol{x}) = \sum_{j=1}^{n} \frac{1}{\lambda_j} [\boldsymbol{\varphi}_{ij}^T(\boldsymbol{x}-\boldsymbol{\mu}_i)]^2 + \ln \prod_{j=1}^{n} \lambda_{ij}, \tag{2}$$

where λ_{ij} is the jth eigenvalue of the covariance matrix of the ith class and $\boldsymbol{\varphi}_{ij}$ is the corresponding eigenvector. Kimura [5] illustrated that the performance of a QDF classifier is mainly affected