Pattern Recognition 47 (2014) 1904–1916

Contents lists available at ScienceDirect

Pattern Recognition

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

Minimum-risk training for semi-Markov conditional random fields with application to handwritten Chinese/Japanese text recognition

Xiang-Dong Zhou [a,*,1], Yan-Ming Zhang [b], Feng Tian [c], Hong-An Wang [c], Cheng-Lin Liu [b]

[a] Intelligent Media Technique Research Center, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, PR China
[b] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguan East Road, Beijing 100190, PR China
[c] Beijing Key Lab of Human–Computer Interaction, Institute of Software, Chinese Academy of Sciences, Beijing 100190, PR China

A R T I C L E I N F O

Article history:
Received 6 July 2013
Received in revised form
24 October 2013
Accepted 2 December 2013
Available online 12 December 2013

Keywords:
Semi-Markov conditional random fields
Minimum-risk training
Character string recognition

A B S T R A C T

Semi-Markov conditional random fields (semi-CRFs) are usually trained with maximum a posteriori (MAP) criterion which adopts the 0/1 cost for measuring the loss of misclassification. In this paper, based on our previous work on handwritten Chinese/Japanese text recognition (HCTR) using semi-CRFs, we propose an alternative parameter learning method by minimizing the risk on the training set, which has unequal misclassification costs depending on the hypothesis and the ground-truth. Based on this framework, three non-uniform cost functions are compared with the conventional 0/1 cost, and training data selection is incorporated to reduce the computational complexity. In experiments of online handwriting recognition on databases CASIA-OLHWDB and TUAT Kondate, we compared the performances of the proposed method with several widely used learning criteria, including conditional log-likelihood (CLL), softmax-margin (SMM), minimum classification error (MCE), large-margin MCE (LM-MCE) and max-margin (MM). On the test set (online handwritten texts) of ICDAR 2011 Chinese handwriting recognition competition, the proposed method outperforms the best system in competition.

1. Introduction

Due to the large character set and the ambiguity of character segmentation, handwritten Chinese/Japanese text recognition (HCTR) is generally accomplished by an integrated segmentation and recognition approach based on character over-segmentation [1]. The input string (text line image for offline data or pen-tip trajectory for online data) is over-segmented into a sequence of components according to the overlapping between strokes (Fig. 1(a)), with each component (consisting a block of strokes) being a character or part of a character. Subject to constraints of character width, consecutive components are combined to generate candidate characters, which constitute the segmentation candidate lattice (Fig. 1(b) and (c)). On assigning each candidate character a number of candidate classes using a character classifier, we construct the segmentation-recognition candidate lattice (referred to as lattice for brevity). Each path in the lattice corresponds to a segmentation-recognition hypothesis, which is evaluated by a parameterized function combining the character recognition score, geometric and linguistic contexts, and the string recognition result is obtained by searching for the optimal path with maximum score.

The performance of integrated segmentation-recognition of character strings (handwritten texts) largely relies on the parameterized path evaluation function. Although many function forms have been proposed, which are usually the heuristic approximation of posterior probability for a segmentation-recognition path (see [1], for a review), only several papers address the problem of parameter learning [2–4]. In our previous work [5], a semi-Markov conditional random field (semi-CRF) [6] based approach has been proposed for HCTR. A semi-CRF outputs a segmentation S of the observation sequence X, together with the label sequence Y assigned to the segments (sub-sequences) of X. In other words, unlike the linear-chain CRF [7] which models $P(Y|X)$, the semi-CRF explicitly estimates $P(S,Y|X)$. For HCTR, if X is the component sequence after over-segmentation, the segments will be the candidate characters (cf. Fig. 1). Semi-CRFs have the advantages that they allow the use of segment features and between-segment dependencies. This attribute is important for HCTR, since the state-of-the-art Chinese character classifiers, such as the modified quadratic discriminant function (MQDF) [8], usually take the holistic character features as input. The semi-CRF model for HCTR is defined on the lattice to directly estimate the a posteriori probability of a segmentation-recognition hypothesis, in which the information of

* Corresponding author. Tel./fax: +86 2365935537.
E-mail addresses: zhouxiangdong@cigit.ac.cn, xiangdongzhou@foxmail.com (X.-D. Zhou), ymzhang@nlpr.ia.ac.cn (Y.-M. Zhang), tianfeng@iscas.ac.cn (F. Tian), hongan@iscas.ac.cn (H.-A. Wang), liucl@nlpr.ia.ac.cn (C.-L. Liu).

[1] This work was partially done when the first author was with the Beijing Key Lab of Human–Computer Interaction, Institute of Software, Chinese Academy of Sciences.

0031-3203/$ - see front matter
http://dx.doi.org/10.1016/j.patcog.2013.12.002

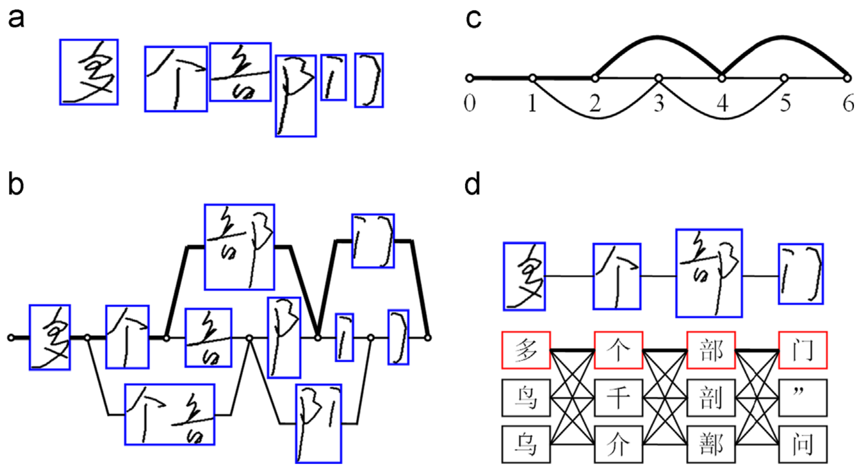

Fig. 1. Generation of segmentation-recognition candidate lattice. (a) Component sequence, (b) candidate characters, (c) segmentation candidate lattice, where each node denotes a candidate segmentation point, each edge corresponds to a candidate character, and the bold lines indicate the desired segmentation and (d) candidate classes of the desired segmentation path.

character recognition, geometric and linguistic contexts are defined as feature functions [5]. This model provides principled tools for both parameter learning and decoding under the maximum a posteriori (MAP) criterion and enables the fusing of high-order features (long range context, such as the trigram language model).

According to the Bayesian decision theory [9], the optimal decision should be made by minimizing the overall risk associated with a cost function measuring the loss of misclassification. When the 0/1 cost is adopted in the Bayes decision rule, we get the popular MAP criterion. The 0/1 cost simply assigns no loss to a correct prediction and a unit loss to an error. Consequently, for HCTR, it aims to minimize the string error rate rather than the character error rate. For example, a hypothesized path (cf. Fig. 1) containing one or more character-level errors, or a totally different path, as compared to the correct, will incur the same amount of loss. However, the performance of HCTR is usually measured in terms of character errors (insertions, deletions and substitutions) [2,5,10], instead of the errors of whole sentence or text line. So, for HCTR, the use of 0/1 cost will lead to a mismatch between classifier training and performance evaluation.

For structured prediction on sequence labeling problems, various discriminative learning techniques have been proposed for training hidden Markov models (HMMs) and CRFs (see Section 2 for a review). Generalization ability is one of the key issues in discriminative training, since the learned models will be finally tested on the unseen data. According to statistical learning theory [11], the test-set error rate is bounded by the sum of the empirical error rate on the training set and a generalization term associated with the margin. Traditional discriminative learning methods, such as maximum mutual information (MMI, an instance of MAP criterion) [12], minimum classification error (MCE) [13] and minimum phone/word error (MPE/MWE) [14], focus more on reducing the empirical error rate rather than decreasing the generalization term [15]. Many attempts have been made to incorporate the principle of large margin into the training of HMMs or CRFs to further improve the generalization abilities [16–18]. However, the performances of these training criteria have not been comprehensively evaluated on HCTR tasks.

In this paper, based on our previous work on HCTR [5], we propose a lattice-based minimum-risk (MR) estimation framework for parameter learning of semi-CRFs. By incorporating the non-uniform (non-0/1) misclassification cost, this criterion is more directly related to the character error rate in contrast to the MAP rule which aims at minimizing the string error rate. With this method, the cost functions initially used for training HMMs in speech recognition can be conveniently applied to HCTR. We also investigate edge selection in MR training, attempting to improve the generalization ability and reduce the computational complexity. Further, we compare the proposed method with several prevalently used learning criteria, including conditional log-likelihood (CLL) [19], MCE [13], max-margin (MM) [17,18], and margin-based extensions of CLL and MCE. We believe that it is the first work that evaluates these training techniques on HCTR tasks. In experiments on three online handwriting databases, the proposed MR training method has yielded superior string recognition performances compared to the state-of-the-art methods.

This work is an extension of a conference paper [20]. The extension includes the comparison with other training criteria, the details of derivations, the effects of edge selection, extensive experimental results and discussions. The remainder of this paper is organized as follows: Section 2 reviews the related work. Section 3 gives a brief introduction to the semi-CRF model defined on the candidate lattice. Section 4 details the minimum-risk training framework for semi-CRFs. Section 5 describes the learning criteria for comparison. Section 6 presents our experimental results and Section 7 draws concluding remarks.

2. Related work

In speech recognition, it has been shown that discriminative learning of HMMs is able to produce consistent improvements in performance compared to the conventional maximum likelihood training criterion, which aims at modeling the data distribution instead of directly separating class categories [21]. In contrast, discriminative learning typically bypass the stage of building the joint-probability model while directly managing to minimize the classification errors. A central issue in the development of discriminative learning methods is the construction of objective function (learning criterion). Popular discriminative learning techniques for HMMs are MMI [12], MCE [13] and MPE/MWE [14]. MMI estimation tries to maximize the a posteriori probability of the training utterances, whereas in MCE training, an approximation to the sentence error rate on the training data is minimized. In contrast to MMI and MCE, which are typically designed to optimize the string-level errors, MPE/MWE aims at performance optimization at the substring pattern level, such as phones and words. Traditional discriminative training aims to find classification boundaries that minimize empirical error rates on training sets, which may not be well generalized to test sets [16]. Many attempts have been made to incorporate the principle of large-