# Constrained Classification on Structured Data

**Chi-Hoon Lee**[*] **Matthew Brown** **Russell Greiner**
{chihoon, mbrown, greiner}@cs.ualberta.ca
University of Alberta
Edmonton, Alberta, Canada

**Shoajun Wang**
shaojun.wang@wright.edu
Wright State University
Dayton, Ohio, U.S.A

**Albert Murtha**
albertmu@cancerboard.ab.ca
Cross Cancer Institute
Edmonton, Alberta, Canada

## Abstract

Most standard learning algorithms, such as Logistic Regression (LR) and the Support Vector Machine (SVM), are designed to deal with *i.i.d.* (independent and identically distributed) data. They therefore do not work effectively for tasks that involve non-*i.i.d.* data, such as "region segmentation". (Eg, the "tumor vs non-tumor" labels in a medical image are correlated, in that adjacent pixels typically have the same label.) This has motivated the work in random fields, which has produced classifiers for such non-*i.i.d.* data that are significantly better than standard *i.i.d.*-based classifiers. However, these random field methods are often too slow to be trained for the tasks they were designed to solve. This paper presents a novel variant, Pseudo Conditional Random Fields (PCRFs), that is also based on *i.i.d.* learners, to allow efficient training but also incorporates correlations, like random fields. We demonstrate that this system is as accurate as other random fields variants, but significantly faster to train.

## Introduction

Radiation therapy planners need to know the location of the tumor to determine what area to treat; this typically requires labelling each pixel within an magnetic resonance (MR) image as either tumor or non-tumor. We could view this segmentor as a standard discriminative classifier, and try to apply standard techniques — eg, logistic regression (LR) or support vector machine (SVM). Unfortunately, these systems work poorly for this task, as they assume that each pixel's label is independent of each other. This assumption is not true, as spatially adjacent pixels typically have the same label.

This has motivated many researchers to use systems that can model and use complex dependencies among data instances, such as Markov Random Fields (MRFs) and Conditional Random Fields (CRFs) (Kumar and Hebert 2003; Lee et al. 2007). This is not without costs, however. For instance, an MRF assumes that the observations are conditionally independent given their class labels; this clearly compromises their expressibility and hence their performance. CRFs allow more complex dependencies, but their underlying computations are intractable. This is partly because these systems *simultaneously* learn the parameters that optimize the label for each individual pixel *by itself*, and also the parameters for jointly labelling *pairs of adjacent pixels*. Although approximation techniques have been adopted to improve their computation efficiency, CRF variants, such as Discriminative Random Fields (DRFs) still require costly learning procedures (Kumar and Hebert 2003). Decoupled Conditional Random Fields (DCRFs) presents an efficient system by combining two classifiers that are each trained separately (Lee, Greiner, and Zaïane 2006). However, this often reduces the accuracy of these systems.

This paper presents the PCRF system that can efficiently use a simple discriminative classifier (LR) and still incorporate spatial compatibility, in a 2-D lattice. Our PCRF can be viewed as a regularized *i.i.d.* discriminative classifier, where the classification task is performed along with a regularized term that explicitly considers dependencies of labels. In particular, PCRF first learns a classifier under the *i.i.d.* assumption, and then relaxes this *i.i.d.* assumption during testing.

## Pseudo Conditional Random Fields – PCRFs

Let $\mathbf{X}$ be an observed input image, where the observation at pixel $i$ is represented by $\mathbf{x}_i$, $i \in S$, where $S$ is the set of observed image pixels. Let $\mathbf{Y}$ be joint set of labels over all pixels of an image. For simplicity we assume $y_i$ at pixel $i \in S$ is binary $y_i \in \{-1, 1\}$. For instance, $\mathbf{X}$ might be a magnetic resonance image of a brain and $\mathbf{Y}$ is a joint labelling over all pixels that indicates whether each pixel is normal or a tumor. We want to find the most-likely labelling $P_\theta(\,\mathbf{Y}\,|\,\mathbf{X}\,) = \prod_{i \in S} P_\theta(\,y_i\,|\,\mathbf{X}, \mathbf{Y} - y_i\,)$. Given feature vectors (observations) $\mathbf{x}_i$ for each pixel $i$ as well as the class label $y_j$ for each neighboring pixel $j \in N_i$, the PCRF formulation then defines

$$\begin{aligned} & P_\theta(\,y_i\,|\,\mathbf{x}_i, \mathbf{x}_{N_i}, y_{N_i}\,) \\ = \quad & \psi_\theta(\mathbf{x}_i, y_i) \times \textstyle\prod_{j \in N_i} \phi^o(\mathbf{x}_i, \mathbf{x}_j) \times \phi^c(y_i, y_j) \qquad (1) \end{aligned}$$

where the potential functions $\phi^o(\mathbf{x}_i, \mathbf{x}_j)$ quantifies the similarity of the feature vectors for pixels $i$ and $j$, and $\phi^c(y_i, y_j)$ models the interactions between the two class labels $y_i$ and $y_j$. The system designer can adjust $\phi^c(.)$ to alter the degree of continuity with respect to class labels expected by the

[*] Much of this work done while working at Yahoo!

model; *e.g.*, if $\phi^c$ gives high weight when neighboring pixels share the same class label, then PCRF will prefer having the same class labels among neighboring pixels. Alternatively, setting $\phi^o \equiv 1$ and $\phi^c \equiv 1$ would remove all spatial dependencies, leading to an *i.i.d.* classifier. Note we use a fixed pair of potential functions: we set $\phi^o(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$, as the similarity measure between neighboring pixels; note this measure is maximal value when the two vectors are colinear. We also set $\phi^c(y_i, y_j) = \alpha$ if $y_i \equiv y_j$, and $1 - \alpha$ otherwise where $\alpha$ weighs the importance that adjacent pixels share the same label.

**Learning** One key factor that constrains the form of typical CRF variant models is to compute exact expectations, as required to learn parameters (Kumar and Hebert 2003).

Our PCRF defines $\psi_\theta(\mathbf{x}_i, y)$ as $\sigma(\theta^T \mathbf{x}_i)$, where $\sigma(t) = \frac{1}{1+exp(-t)}$ corresponds to a standard local discriminative classifier, Logistic Regression. This explicitly quantifies the posterior probability of being in class $y$ given observation $\mathbf{x}_i$. Our PCRF learner is simple, and more efficient than CRF variants, since we only need to fit the parameter $\theta$ for a local potential function $\psi(.)$. (Recall we hand-defined the $\phi^c$ and $\phi^o$ functions.)

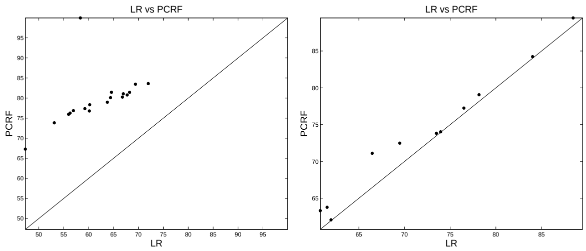
**Inference** Our PCRF system incorporates the spatial correlations in the *inference* step. In general, our objective in the inference process seeks $\mathbf{Y}^*$ as:

$$\begin{aligned} \mathbf{Y}^* \;=\; & \arg\max_{\mathbf{Y}} \Big( \sum_{i \in S} \log \psi_\theta(\mathbf{x}_i, y_i) \; + \\ & \sum_{i \in S} \sum_{j \in N_i} \log \phi^c(\mathbf{x}_i, \mathbf{x}_j) + \log \phi^o(y_i, y_j) \Big) \qquad (2) \end{aligned}$$

The above equation requires searching an exponential search space (i.e. $2^{|S|}$ for binary case) to find an optimal $\mathbf{Y}^*$. To efficiently solve (approximate) our objective, we formulate Eq. 2 using graph cuts algorithm that solves image pixel classification tasks by minimizing an energy function when spatial correlations among pixels are "independent" of the observations; this involves using linear programming to find the max-flow/min-cut on a graph in which nodes correspond to pixels and edges correspond to connections between neighboring pixels (Boykov, Veksler, and Zabih 1999). We reformulate this graph cuts approach to apply to our PCRF framework (Eq. 2), in which neighbor relationships are dependent on both the labels and the observations (feature vectors). Refer to (WEB )

## Experiments

We examine the performance of our PCRF on a binary classification task for both synthetic image sets and real world problem (MR image segmentation) and compare it with the baseline classifier – Logistic Regression (LR) – to highlight the importance in modeling the spatial constraint. To quantify the performance of each model, we used the Jaccard score $J = \frac{TP}{(TP+FP+FN)}$ , where TP denotes true positives, FP false positives, and FN false negatives. We generated 18 synthetic binary images, each with its own shape. The intensities of pixels in each image were then independently corrupted by Gaussian noise $\mathcal{N}(0, 1)$. For our real world problem, datasets are MRI scans from 11 patients with brain tumors; for each patient, we annotated each pixel with values based on three different MR imaging modalities: *T1*, *T2*, and *T1 with gadolinium contrast* ("T1c"). Figure. 1 clearly presents robustness of PCRF from synthetic images (a) and brain tumor segmentation (b). Each point above the diagonal line indicates that PCRF produced a higher Jaccard score (percentage) for a given test image. We also compare the PCRF's performance with DRF's for brain tumor segmentation. On average, PCRF produced as accurate Jaccard scores (percentage) as DRFs: 73.69 versus 73.03, respectively. However, the PCRF's learning time over 11 patients (38 seconds) is significantly more efficient than DRFs' (1697 seconds average). Our PCRF was over 40 times faster than the DRF ($p < 10^{-37}$, paired-samples $t$-tests). Refer to (WEB ) for more details about experiments.



(a) Synthetic data sets (b) MR image analysis

Figure 1: Jaccard Scores (percentage) for PCRF vs. LR

## Conclusion

As standard *i.i.d.* classifiers do not model *interdependencies* of labels, they typically do very poorly for such tasks spatially constrained classification. In this paper, we proposed an efficient model – Pseudo Conditional Random Fields (PCRF) – that takes advantage of a typical discriminative classifier (LR) when training, then relaxes the classifier's *i.i.d.* assumption during the inference. We demonstrate the effectiveness and the efficiency of PCRFs from both synthetic and real world data sets — showing that its performance is comparable to state-of-the-art random field systems, but its training time is significantly more efficient.

## References

Boykov, Y.; Veksler, O.; and Zabih, R. 1999. Fast approximate energy minimization via graph cuts. In *ICCV*.

Kumar, S., and Hebert, M. 2003. Discriminative fields for modeling spatial dependencies in natural images. In *NIPS*.

Lee, C.-H.; Wang, S.; Jiao, F.; Schuurmans, D.; and Greiner, R. 2007. Learning to model spatial dependency: Semi-supervised discriminative random fields. In *NIPS*.

Lee, C.-H.; Greiner, R.; and Zaïane, O. R. 2006. Efficient spatial classification using decoupled conditional random fields. In *PKDD*.

WEB. http://www.cs.ualberta.ca/∼chihoon/aaai2008/.