Segmenting Brain Tumors Using Pseudo–Conditional Random Fields

Chi-Hoon Lee[1,4], Shaojun Wang[2], Albert Murtha[3], Matthew R.G. Brown[1], and Russell Greiner[1]

[1] Department of Computing Science, University of Alberta, Canada
{chihoon,mbrown,greiner}@cs.ualberta.ca
[2] Department of Computing Science, Wright State University, USA
shaojun.wang@wright.edu
[3] Cross Cancer Institute, University of Alberta, Canada
albertmu@cancerboard.ab.ca
[4] Yahoo! Inc, USA
chihoon@yahoo-inc.com

Abstract. Locating Brain tumor segmentation within MR (magnetic resonance) images is integral to the treatment of brain cancer. This segmentation task requires classifying each voxel as either tumor or non-tumor, based on a description of that voxel. Unfortunately, standard classifiers, such as Logistic Regression (LR) and Support Vector Machines (SVM), typically have limited accuracy as they treat voxels as *independent* and *identically distributed* (*iid*). Approaches based on random fields, which are able to incorporate spatial constraints, have recently been applied to brain tumor segmentation with notable performance improvement over iid classifiers. However, previous random field systems involved computationally intractable formulations, which are typically solved using some approximation. Here, we present *pseudo-conditional random fields* (PCRFs), which achieve accuracy similar to other random fields variants, but are significantly more efficient. We formulate a PCRF as a regularized discriminative classifier that relaxes the classification decision for each voxel by considering the labels and features of neighboring voxels.

1 Introduction

Segmenting brain tumors in magnetic resonance (MR) images involves classifying each voxel as tumor or non-tumor [1,2,3]. This task, a prerequisite for treating brain cancer using radiation therapy, is typically done by hand by expert medical doctors, who find this process laborious and time-consuming. Replacing this manual effort with a good automated classifier would save doctors time; the resulting labels may also be more accurate, or at least more consistent.

We treat this as a binary classification task, using a classifier to map each MR image voxel described as a vector of values $\mathbf{x} \in \Re^d$ to a bit $y \in \{+1, -1\}$, corresponding to either tumor or non-tumor. We first *learn* this classifier from a set of data instances $\{\langle \mathbf{x}_i, y_i \rangle\}_i$ [4]. Here, we focus on *probabilistic classifiers*

D. Metaxas et al. (Eds.): MICCAI 2008, Part I, LNCS 5241, pp. 359–366, 2008.

that actually return a class likelihood value $P(\, y = +1 \,|\, \mathbf{x}\,) \in [0, 1]$ for each voxel; our classifier can then return $+1$ (tumor) if $P(\, y = +1 \,|\, \mathbf{x}\,) \geq 0.5$. In general, given an entire $n \times m$ image, our classifier will seek the most likely labeling over $\{-1, +1\}^{n \times m}$: $\mathbf{Y}^{(*)} = \text{argmax}_{\mathbf{Y}}\, P(\, \mathbf{Y} \,|\, \mathbf{X}\,)$. (This use of probabilities distinguishes these approaches from many other segmentation approaches, such as those based on variational and level set techniques [5,6].)

Standard machine learners, such as Naïve Bayes, logistic regression (LR), and support vector machines (SVMs), produce effective classifiers in many domains [7,8]. However, these algorithms assume that the individual instances are iid. This is appropriate if the instances correspond to, say, a patients in a study, as finding that one patient responds well to some treatment does not mean that the next patient will also respond well. However, this assumption is problematic in our current situation, where each instance corresponds to a voxel: Here, finding that one voxel is labeled a tumor strongly suggests that its neighbors will have a similar label; similarly non-tumor voxels tends to be next to other non-tumor voxels. Algorithms that assume the data is iid typically perform poorly when the data is not, which is why these algorithms do relatively poorly at segmentation tasks.

This has motivated researchers to apply Markov Random Fields (MRFs; [9]) and Conditional Random Fields (CRFs; [10]) to various segmentation tasks. These techniques are able to represent complex dependencies among data instances, giving them higher accuracy on the segmentation task than iid classifiers [11,12]. However, these random field approaches are based on computationally intractable formulations. Although there are approximation techniques that can deal with these computational challenges, CRF variants such as Discriminative Random Fields (DRFs) and Support Vector Random Fields (SVRFs) still require computationally expensive learning procedures [11,13].

In this paper, we present a novel supervised learning system, PCRF, that can efficiently produce high-quality segmenters, incorporating spatial constraints among MR image voxels. PCRF can be viewed as a regularized iid discriminative classifier that is first trained assuming the data is iid; this makes the training computationally efficient. It then relaxes the iid assumption during inference, by including a regularizing term that uses the class labels and feature vectors of neighboring voxels of a given voxel. We demonstrate that PCRF is robust and efficient by illustrating its performance at segmenting MR images of the brains of tumor patients. We show that PCRF is significantly more accurate than the corresponding base iid classifiers, and is significantly more efficient that other random field methods during training, while producing similar accuracy.

Section 2 reviews related work, including random field models. Section 3 introducing our framework and novel PCRF system. Section 4 presents experiments that empirically demonstrate the efficiency and effectiveness of our model.

2 Background

We view brain tumor segmentation on a 2D MR image as classifying each image voxel as either tumor or non-tumor. The challenge is finding the most likely

configuration of (tumor vs. non-tumor) labels $\mathbf{Y} = (y_1, y_2, \ldots, y_r) \in \{-1, +1\}^r$ for the voxels of a 2D MR image $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r)$, where each set ranges over the set of indices S of all voxels in the $r = m \times n$ image, each $y_i \in \{-1, +1\}$ is the label for voxel i, and $\mathbf{x}_i$ is the feature vector for voxel i.

A pair-wise MRF is formulated as

$$P(\mathbf{Y} \mid \mathbf{X}) \propto P(\mathbf{Y}, \mathbf{X}) = \frac{1}{Z(\mathbf{X})} \exp\left(\sum_{i \in S} D(\mathbf{x}_i, y_i) + \sum_{i \in S} \sum_{j \in N_i} V(y_i, y_j)\right) \quad (1)$$

where $D(\mathbf{x}_i, y_i)$ corresponds to the local log likelihood $\log(P(\mathbf{x}_i \mid y_i))$ of $\mathbf{x}_i$ given a class label y_i; $V(y_i, y_j)$ is a potential function that explicitly encodes the dependencies between labels at i and its neighbor j, based on N_i, which is the set of voxels neighboring $\mathbf{x}_i$. $Z(\mathbf{X})$ is a normalizing factor to make the formulation a probability distribution. We can read-off the MRF assumptions from Equation 1: that the voxels are conditionally independent given their class labels, and that spatial correlations are modelled based only on the labels of neighboring voxels (y_i and y_j) but not on the observations ($\mathbf{x}_i$ and $\mathbf{x}_j$). These factors limit the advantages of using MRFs to model spatial dependencies in MR images [11,12,13].

CRFs attempt to overcome these disadvantages by relaxing the conditional independence assumption and incorporating observations into the formulation of spatial dependencies.

$$P(\mathbf{Y} \mid \mathbf{X}) = \frac{1}{Z(\mathbf{X})} \exp\left(\sum_{i \in S} A(y_i, \mathbf{X}) + \sum_{i \in S} \sum_{j \in N_i} I(y_i, y_j, \mathbf{X})\right) \quad (2)$$

where $A(y_i, \mathbf{X})$ corresponds to the *conditional* probability distribution (while the MRF's $D(\mathbf{x}_i, y_i)$ corresponds to the log conditional probability), and the $I(y_i, y_j, \mathbf{X})$ term incorporates observations of data instances (unlike MRF's $V(y_i, y_j)$ which does not). The Discriminative Random Field (DRF) is a variant of the CRF that performs robustly in 2D image region classification problems [13]. The Support Vector Random Field (SVRF) is a modification of DRFs that address high dimensional feature vectors and imbalanced datasets effectively [11].

Unfortunately, DRFs and SVRFs are computationally expensive, especially during learning, as their computations are exponential in the number of data points. This is basically due to their need to compute the partition function, corresponding to the $Z(\mathbf{X})$ in Equation 2. (Note that Gaussian assumption of MRF makes $Z(\mathbf{X})$ in Equation 1 simpler.) This has led to many approximation methods, such as pseudo-likelihood, contrastive divergence, and pseudo-marginal approximation [10,13,14,12]. Unfortunately these approximations reduce the accuracy of the learned segementor. This motivated *Decoupled Conditional Random Fields* (DCRFs [15]), which speed up the CRF-based computation by approximating a CRF as the combination of two classifiers that are each trained separately. As the DCRF framework searches for the parameter values that optimize each model separately, the combined parameter values are not necessarily globally optimal.

3 Pseudo Conditional Random Fields – PCRFs

The PCRF framework attempts to obtain the advantage of both the MRF and CRF approaches by relaxing the iid assumption of a simple discriminative classifier by adding a regularization term. We want to find the most-likely labelling $P_\theta(\mathbf{Y} \mid \mathbf{X}) = \prod_{i \in S} P_\theta(y_i \mid \mathbf{X}, \mathbf{Y} - y_i)$. Given feature vectors (observations) $\mathbf{x}_i$ for each voxel i as well as the class labels y_{N_i} over neighboring voxels $j \in N_i$, the PCRF formulation defines

$$P_\theta(y_i \mid \mathbf{x}_i, \mathbf{x}_{N_i}, y_{N_i}) \quad = \quad \psi_\theta(\mathbf{x}_i, y_i) \times \prod_{j \in N_i} \phi^o(\mathbf{x}_i, \mathbf{x}_j) \times \phi^c(y_i, y_j), \qquad (3)$$

where the potential functions $\phi^o(\mathbf{x}_i, \mathbf{x}_j)$ quantifies the similarity of the feature vectors for voxels i and j, and $\phi^c(y_i, y_j)$ models the interactions between the two class labels y_i and y_j. We can adjust $\phi^c(.)$ to alter the degree of continuity with respect to class labels expected by the model; *e.g.*, if we set ϕ^c to give high weight when neighboring voxels share the same class label, then the resulting PCRF will prefer having the same class labels among neighboring voxels. Alternatively, setting $\phi^o \equiv 1$ and $\phi^c \equiv 1$ would remove all spatial dependencies, leading to an iid classifier. Note we use a fixed pair of potential functions: here we set $\phi^o(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$, as the similarity measure between neighboring voxels; note this measure is maximum value when the two vectors are co-linear. We also set $\phi^c(y_i, y_j) = \alpha$ if $y_i \equiv y_j$, and $1 - \alpha$ otherwise, where α weighs the continuity of identical class labels. Here we used $\alpha = 0.6$.

For now, we define $\psi_\theta(\mathbf{x}_i, y_i) = \sigma(\theta^T \mathbf{x}_i) = \frac{1}{1+exp(-\theta^T \mathbf{x}_i)}$ as a simple logistic regression classifier. We chose a discriminative approach rather than a generative one because the former empirically shows better performance than the latter [8].

Learning. Learning the PCRF parameters is more efficient than for other CRF variants as a PCRF needs to fit only the parameter vector θ for a local potential function $\psi_\theta(.)$, which does not involve any spatial interactions. Here, we use the standard way to maximize the conditional log-likelihood $\theta^{(*)} = \arg\max_\theta \sum_{i \in S} \Big[y_i \log \sigma(\theta^T \mathbf{x}_i) + (1 - y_i) \log(1 - \sigma(\theta^T \mathbf{x}_i)) \Big]$.

Inference. The PCRF inference process incorporates regularization based on neighbor relationships. In general, the objective of inference is to maximize the log likelihood:

$$\begin{aligned} \mathbf{Y}^* &= \arg\max_{\mathbf{Y}} \log P(\mathbf{Y} \mid \mathbf{X}) \\ &= \arg\max_{\mathbf{Y}} \sum_{i \in S} \log \psi_\theta(\mathbf{x}_i, y_i) \; + \; \sum_{i \in S} \sum_{j \in N_i} \log \phi^c(\mathbf{x}_i, \mathbf{x}_j) + \log \phi^o(y_i, y_j) \quad (4) \end{aligned}$$

The graph cuts algorithm 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 whose nodes correspond to voxels and edges correspond to connections between neighboring voxels [16]. We reformulate this graph cuts approach

to apply to our PCRF framework (Equation 4), where neighbor relationships are dependent on both the labels and the observations (feature vectors).

4 Brain Tumor Segmentation

We applied our PCRF model to the challenging real world problem of segmenting brain tumors in MR images. Since a PCRF can be viewed as a regularized discriminative iid classifier, we first show the differences between PCRF and its degenerate iid classifier – LR.

To quantify the performance of each model, we used the percentage Jaccard score $J = 100 \times \frac{TP}{(TP+FP+FN)}$, where TP denotes the number of true positives, FP false positives, and FN false negatives, taken over the entire image. We used this score for brain tumor segmentation task as this data is very imbalanced in that only a small percentage of voxels are in the "tumor" class; hence scores like "accuracy" would be high as the "true negative" class is typically huge.

We applied several different models – LR, PCRF, SVRF – to the task of classifying MR image slices, where each slice is defined with 258 by 258 pixels, each of which is described using 33 features [17]. We considered data from 11 patients with brain tumors; for each patient, we annotated each voxel with values based on three different MR imaging modalities: *T1*, *T2*, and *T1 with gadolinium contrast*("T1c"). We focus on 2D images; this is sufficient to illustrate the challenges as the neighborhood structure here involves cycles, which makes both inference and learning procedures computationally challenging[1]. Testing and training were done in a patient-specific manner: for each patient, each algorithm was trained on a subset of the patient's data, then tested on another (disjoint) subset. This is similar to the approach taken in many other studies of automatic brain tumor segmentation such as [18,19,20,21].

Our systems attempted to segment the "enhancing" tumor area — the region that appears bright on T1c images. Note that it is not sufficient to simply threshold T1c images by "brightness" because other tissues can have the same range of intensities. In the case of glioblastomas with necrotic cores, which appear dark on T1 images, we defined the enhancing rim of the tumor as well as the dark necrotic core as the target tumor region.

Fig. 1 shows one example of segmentation results. One test and its correct label ("ground truth") slice are shown in first two columns respectively. The result from LR, shown in third from Fig. 1, indicates that LR correctly classifies the tumor region but that it also misclassifies several small non-tumor regions as "tumor". PCRF's result, which appears on the far right, is more accurate. (See [22] for the complete set of larger images.)

Fig. 2(a) presents the Jaccard percentage scores from the 11 studies, where points above the diagonal line denote instances in which the PCRF performed better than its degenerative model, LR. Overall, the PCRF's accuracy was statistically significantly higher than LR's at $p < 0.005$ on a paired sample t-test.

[1] We are beginning to explore extending this approach to 3D, which involves simply redefining the neighborhood structure.

(a) Testing Slice (b) Ground Truth (c) LR(J=66.45) (d) PCRF(J=71.11)

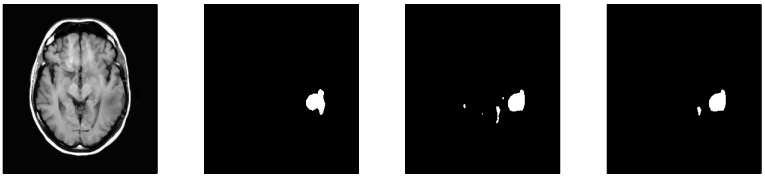
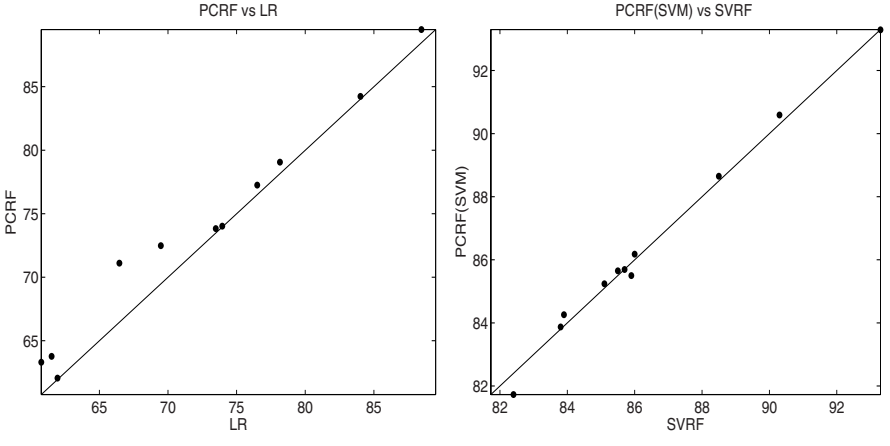


Fig. 1. Classification results. The PCRF shows almost 4% improvement of Jaccard score over LR.



(a) PCRF vs. LR (b) PCRF(SVM) vs. SVRF

Fig. 2. Jaccard Scores (percentage)

We also compared our PCRF system with the state-of-the-art CRF variant, the Support Vector Random Field (SVRF [11]), whose potential functions are based on Support Vector Machines (SVMs). Here, we implemented PCRF(SVM), which differed from the PCRF system only by using an SVM to compute the $\psi(\mathbf{x}, y)$ (from Equation 4) which models the relationship between a voxel's feature vector and its label. An SVM produces the distance between a hyperplane and a data instance as its decision value $f_{SVM}(\mathbf{x}_i) \in (-\infty, +\infty)$. To normalize this unbounded range, we fit this value to a sigmoid function: $g_{\beta_0,\beta_1}(\mathbf{x}) = P(\, y = +1 \,|\, \mathbf{x}\,) = \frac{1}{1+\exp(\beta_0+\beta_1(\mathbf{x}))}$, estimating the parameters β_0 and β_1 from the training data $\{(f_{SVM}(\mathbf{x}_i),\, y_i)\}_i$ [11]. Figure 2(b) compares the percentage Jaccard scores of PCRF(SVM)[2] vs SVRF. It is clear that PCRF(SVM) is comparable with SVRF.

We next considered the timing. As our PCRF did not need to learn parameters for its spatial correlation model, we anticipated it would be significantly faster

[2] PCRF(SVM) outperformed the SVM, which is a robust i.i.d. classifier ($p < 0.001$); see [22] for details.

during the learning stage. The learning times (average across 11 patients, in seconds) confirm this:

	DRF	SVRF	DCRF	PCRF
Tumor segmentation	1697	1276	63	**38**

Our PCRF was over 40 times faster than the DRF and over 30 times faster than the SVRF ($p < 10^{-37}$ and $p < 10^{-29}$, paired-samples t-tests for DRFs and SVRFs, respectively). Even DCRF, known as the fastest CRF variant, is significantly slower than our PCRF ($p < 10^{-26}$).

5 Conclusion

We found that the PCRF(SVM) system, which uses a linear SVM to map from voxel to label, worked effectively. We might be able to obtain further performance improvements by using a non-linear kernel function. In addition, we might be able to produce a more robust model by incorporating a prior $P(\theta)$ over θ to further reduce the possibility of overfitting. We are extending this work to develop effective systems to overcome the limitations of patient-specific training, by taking advantage of semi-supervised learning principles.

Contributions. This paper has presented the Pseudo Conditional Random Field (PCRF) model, a CRF-inspired formulation that incorporates a specified potential function to model the relationships between neighboring voxels. Our PCRF is fast to train as it does not need to fit parameters that model the neighbor relationships. It can be viewed as a regularized iid classifier, whose classification decisions for each pixel involve the labels and features of neighboring voxels. Thus, during inference, PCRF avoids the iid assumption, which is inappropriate for image segmentation tasks. We demonstrate that PCRF is effective by showing it can effectively segment brain tumors from MR images, achieving state-of-the-art segmentation results, but at a small fraction of the training time.

Acknowledgments. R. Greiner is supported by NSERC and the Alberta Ingenuity Centre for Machine Learning (AICML). C-H Lee is supported by the AICML. M. Brown is supported by Alberta Cancer Board. Our thanks to Dale Schuurmans for helpful discussions on problem formulation and to BTAP members for help in data processing.

References

1. Corso, J.J., Sharon, E., Yuille, A.L.: Multilevel segmentation and integrated bayesian model classification with an application to brain tumor segmentation. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 790–798. Springer, Heidelberg (2006)

2. Gering, D.T.: Diagonalized nearest neighbor pattern matching for brain tumor segmentation. In: Ellis, R.E., Peters, T.M. (eds.) MICCAI 2003. LNCS, vol. 2879, pp. 670–677. Springer, Heidelberg (2003)
3. Corso, J.J., Sharon, E., Dube, S., El-Saden, S., Sinha, U., Yuille, A.: Efficient Multilevel Brain Tumor Segmentation with Integrated Bayesian Model Classification. IEEE Transactions on Medical Imaging 27(5), 629–640 (2008)
4. Mitchell, T.: Machine Learning. McGraw-Hill, New York (1997)
5. Liu, J., Udupa, J.K., Odhner, D., Hackney, D., Moonis, G.: A system for brain tumor volume estimation via mr imaging and fuzzy connectedness. Computational Medical Imaging and Graphics 29(1), 21–34 (2005)
6. Cobzas, D., Birkbeck, N., Schmidt, M., Jagersand, M., Murtha, A.: A 3D variational brain tumor segmentation using a high dimensional feature set. In: MMBIA (2007)
7. Joachims, T.: Making large-scale svm learning practical. In: Scholkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press, Cambridge (1999)
8. Ng, A., Jordan, M.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: NIPS, vol. 14 (2002)
9. Li, S.Z.: Markov Random Field Modeling in Image Analysis. Springer, Tokyo (2001)
10. Lafferty, J., Pereira, F., McCallum, A.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML (2001)
11. Lee, C.H., Greiner, R., Schmidt, M.: Support vector random fields for spatial classification. In: Jorge, A.M., Torgo, L., Brazdil, P.B., Camacho, R., Gama, J. (eds.) PKDD 2005. LNCS (LNAI), vol. 3721, pp. 121–132. Springer, Heidelberg (2005)
12. Lee, C.H., Wang, S., Jiao, F., Schuurmans, D., Greiner, R.: Learning to model spatial dependency: Semi-supervised discriminative random fields. In: NIPS, vol. 19 (2007)
13. Kumar, S., Hebert, M.: Discriminative fields for modeling spatial dependencies in natural images. In: NIPS (2003)
14. Kumar, S., August, J., Hebert, M.: Exploiting inference for approximate parameter learning in discriminative fields: An empirical study. In: Rangarajan, A., Vemuri, B.C., Yuille, A.L. (eds.) EMMCVPR 2005. LNCS, vol. 3757, pp. 153–168. Springer, Heidelberg (2005)
15. Lee, C.H., Greiner, R., Zaiane, O.R.: Efficient spatial classification using decoupled conditional random fields. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 272–283. Springer, Heidelberg (2006)
16. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. In: ICCV, pp. 377–384 (1999)
17. Schmidt, M.: Automatic brain tumor segmentation. Master's thesis, University of Alberta (2005)
18. Garcia, C., Moreno, J.: Kernel based method for segmentation and modeling of magnetic resonance images. In: Lemaître, C., Reyes, C.A., González, J.A. (eds.) IBERAMIA 2004. LNCS (LNAI), vol. 3315, pp. 636–645. Springer, Heidelberg (2004)
19. Zhang, J., Ma, K., Er, M., Chong, V.: Tumor segmentation from magnetic resonance imaging by learning via one-class support vector machine. In: Int. Workshop on Advanced Image Technology, pp. 207–211 (2004)
20. Chen, T., Metaxas, D.N.: Gibbs prior models, marching cubes, and deformable models: A hybrid framework for 3d medical image segmentation. In: Ellis, R.E., Peters, T.M. (eds.) MICCAI 2003. LNCS, vol. 2879, pp. 703–710. Springer, Heidelberg (2003)
21. Kaus, M., Warfield, S., Nabavi, A., Black, P., Jolesz, F., Kikinis, R.: Automated segmentation of MR images of brain tumors. Radiology 218, 586–591 (2001)
22. `http://www.cs.ualberta.ca/~btap/research/pcrf/` (2008)