Weighted Iterated Conditional Modes for Random Fields: Application to Prostate Cancer Detection

James Monaco, Satish Viswanath, and Anant Madabhushi

Department of Biomedical Engineering, Rutgers University, USA. 
{jpmonoaco, anantm}@rci.rutgers.edu

Abstract. Random fields (RFs) provide a valuable means for modeling large numbers of dependent random variables. However, estimating RFs is nontrivial, requiring sophisticated techniques such as stochastic relaxation. Unfortunately, these techniques provide no means for adjusting classifier performance (e.g. sensitivity/specificity). Instead, they produce a single, hard classification at a static operating point, thus precluding, for example, the construction of receiver operator characteristic (ROC) curves. Addressing this deficiency, we introduce weighted maximum a posteriori (WMAP) estimation, a generalization of MAP estimation that allows certain classes to be weighted more heavily than others. We also introduce weighted iterated conditional modes (WICM), a novel adaptation of ICM capable of WMAP estimation on RFs. We demonstrate the use of WICM by integrating it into two separate Markov random field (MRF) based classification systems capable of detecting prostate cancer (CaP) in 1) whole-mount histological sections and 2) multi-protocol MRI. Specifically, we show how WICM can be used to arbitrarily adjust the CaP detection sensitivity of these systems, yielding ROC curves.

1 Introduction

Many estimation tasks require the ability to classify multiple objects simultaneously. For example, these objects could be calcifications in a mammogram or the pixels of a magnetic resonance image. Within a Bayesian framework each object is modeled as a random variable, and the collection of these random variables is called a random field (RF)\(^1\). If the random variables are assumed independent, we can estimate each in isolation. This estimation typically involves an exhaustive search. For example, obtaining the maximum a posteriori (MAP) estimate (for a single random variable) entails calculating the a posteriori probability for each possible class, and then choosing the class with the largest probability. However, if the random variables are not independent, the entire RF must be estimated collectively. Since the number of possible states of the random field

\(^1\) Though a Markov random field implies that the the inter-variable dependencies are restricted to local neighborhoods, a Markov random field and a random field are mathematically equivalent.
is prohibitively large, an exhaustive approach is untenable\(^2\). Consequently, more sophisticated schemes, such as iterated conditional modes [1] (ICM), simulated annealing [2], or maximum posterior marginals [3], become necessary. These schemes employ iterative techniques to converge to a single, hard labeling.

The capability of adjusting classifier performance (e.g., sensitivity/specificity) with respect to specific classes is essential for many applications. For example, the performance of commercial systems for detecting mammographic abnormalities is typically adjusted to the highest detection sensitivity that incurs no more than two false positive per image. In situations amenable to exhaustive searches, means for modifying performance, such as thresholding or weighting the \textit{a posteriori} probabilities, are well established [4]. Unfortunately, analogous methods compatible with the techniques required for classifying RFs have yet to be proposed. Consequently, most RF-based classification systems restrict their performance to a single, static operating point. To our knowledge, the only previous attempts [5,6] to adjust performance have resulted in \textit{ad hoc} schemes that leveraged the peculiarities of the ICM algorithm: Since ICM converges to a local maximum (mode) of the \textit{a posteriori} probability, varying the initial conditions (i.e., the initial state of the RF) can vary the classification results. However, there is no reason to believe that the modes of the \textit{a posteriori} probability are associated with meaningful classifications; and thus, such techniques are heuristic, lacking any mathematical justification.

In this paper we introduce the first mathematically sound means for adjusting the performance of an estimator compatible with RFs. Specifically, we introduce weighted MAP (WMAP) estimation, a generalization of MAP estimation that allows certain classes to be weighted more heavily than others. Furthermore, we introduce weighted ICM (WICM), a novel adaptation of ICM capable of WMAP estimation on RFs. We demonstrate their applicability in two separate classification systems based on Markov random fields (MRFs): 1) for detecting prostate cancer (CaP) in whole-mount histological sections (WMHISs) [7] and 2) for detecting CaP in multi-protocol (T2-weighted and dynamic-contrast enhanced) 3 Tesla in \textit{vivo} magnetic resonance imaging (MRI) [8]. Within this context, we illustrate how WICM can be used to vary classification performance, enabling the construction of receiver operator characteristic (ROC) curves.

The remainder of the paper is organized as follows: Section 2 reviews the necessary nomenclature and introduces the WMAP formulation. In Section 3 we derive WICM. Section 4 provides an evaluation of WICM in the context of our CaP detection systems. Section 5 offers concluding remarks.

2 Weighted Maximum \textit{a Posteriori} Estimation for MRFs

2.1 Markov Random Field Nomenclature and Assumptions

Let the set \( S = \{1, 2, \ldots, N\} \) reference \( N \) sites to be classified. Each site \( s \in S \) has two associated random variables: \( X_s \in A = \{\omega_1, \omega_2, \ldots, \omega_L\} \) indicating its state.

\(^2\) If a random field contains \( N \) random variables, each of which can assume one of \( L \) classes, the total number of possible states is \( L^N \).
(class) and \( Y_s \in \mathbb{R}^D \) representing its \( D \)-dimensional feature vector. Particular instances of \( X_s \) and \( Y_s \) are denoted by the lowercase variables \( x_s \in \mathcal{X} \) and \( y_s \in \mathbb{R}^D \). Let \( X = (X_1, X_2, \ldots, X_N) \) and \( Y = (Y_1, Y_2, \ldots, Y_N) \) refer to all random variables \( X_s \) and \( Y_s \) in aggregate. The state spaces of \( X \) and \( Y \) are the Cartesian products \( \Omega = \mathcal{X}^N \) and \( \mathbb{R}^{D \times N} \). Instances of \( X \) and \( Y \) are denoted by the lowercase variables \( x = (x_1, x_2, \ldots, x_N) \in \Omega \) and \( y = (y_1, y_2, \ldots, y_N) \in \mathbb{R}^{D \times N} \). Let \( G = (S, E) \) establish an undirected graph structure on the sites, where \( S \) and \( E \) are the vertices (sites) and edges, respectively. A neighborhood \( \eta_s \) is the set containing all sites that share an edge with \( s \), i.e. \( \eta_s = \{ r : r \in S, r \neq s, \{r, s\} \in E \} \). If \( P \) is a probability measure defined over \( \Omega \) then the triplet \( (G, \Omega, P) \) is called a random field.

The conclusions in this paper are predicated on two assumptions. First, the states \( X \) are assumed to constitute a Markov random field (MRF). That is, \( X \) is a random field whose local conditional probability functions satisfy the Markov property: \( P(x_s | x_{s.a}) = P(x_s | x_{\eta_s}) \), where \( x_{s.a} = (x_1, \ldots, x_{s-1}, x_{s+1}, \ldots, x_N), x_{\eta_s} = (x_{\eta_s(1)}, \ldots, x_{\eta_s(|\eta_s|)}) \), and \( \eta_s(i) \in S \) is the \( i \)-th element of the set \( \eta_s \). Second, we assume that each feature vector \( Y_s \) is conditionally independent and identically distributed given its associated state \( X_s \): \( P(y | x) = \prod_{s \in S} P(y_s | x_s) = \prod_{s \in S} f(y_s | x_s) \), where the use of the single probability density function \( f \) indicates that \( P(y_s | x_s) \) is identically distributed across \( S \).

### 2.2 Weighted Maximum a Posteriori Estimation

Given an observation of the feature vectors \( Y \), we would like to estimate the states \( X \). Bayes rule advocates selecting the estimate \( \hat{x} \in \Omega \) that minimizes the conditional risk \([4]\)

\[
R(X | \hat{x}, y) = E \left[ C(X, \hat{x}) \mid y \right] = \sum_{x \in \Omega} C(x, \hat{x}) P(x \mid y),
\]

where \( E \) indicates expected value and \( C(x, \hat{x}) \) is the cost of selecting labels \( \hat{x} \) when the true labels are \( x \). For most classification tasks it is sufficient to assume that the cost of mislabeling any individual site is 1) independent of the remaining sites and 2) identical for every site. This implies the following: 

\[
C(x, \hat{x}) = \prod_{s \in S} C(x_s, \hat{x}_s),
\]

where \( C(x_s, \hat{x}_s) \) is the cost of selecting label \( \hat{x}_s \) when the true label is \( x_s \).

The most prevalent cost function for MRFs (though this cost is rarely expressed explicitly) is \( C_{MAP}(x, \hat{x}) = 1 - \prod_{s \in S} \delta(x_s - \hat{x}_s) \), where \( \delta \) is the Kronecker delta. That is, mislabeling any of the sites results in an identical cost of 1. To allow certain decisions to be weighted more heavily than others \( C_{MAP} \) can be generalized as follows: 

\[
C_{WMAP}(x, \hat{x}) = \left[ \prod_{s \in S} \alpha(x_s) \right] \left[ 1 - \prod_{s \in S} \delta(x_s - \hat{x}_s) \right].
\]

This case mislabeling a system of sites whose true labels are \( x \) has an associated cost of \( \prod_{s \in S} \alpha(x_s) \).

The connection between cost functions and estimation becomes clear when \( C_{WMAP} \) is inserted into (1)

\[
R_{WMAP}(X | \hat{x}, y) = \sum_{x \in \Omega} \left\{ P(x \mid y) \left[ \prod_{s \in S} \alpha(x_s) \right] \left[ 1 - \prod_{s \in S} \delta(x_s - \hat{x}_s) \right] \right\}
\]
\[ = \sum_{x \in \Omega} \left[ P(x|y) \prod_{s \in S} \alpha(x_s) \right] - P(\tilde{x}|y) \prod_{s \in S} \alpha(\tilde{x}_s). \quad (2) \]

Since the first term in (2) is not a function of \( \tilde{x} \), minimizing (2) over \( \tilde{x} \) is equivalent to maximizing the second term. For convenience we will refer to this term as the conditional benefit \( B(x|y) \), i.e.

\[ B(x|y) = P(x|y) \prod_{s \in S} \alpha(x_s). \quad (3) \]

(Since it will no longer be necessary to differentiate the true labels from their estimates, the notation \( \tilde{x} \) is henceforth dropped.) Note that if \( \alpha(\cdot) \equiv 1 \) then the conditional benefit reduces to the \textit{a posteriori} probability, and the minimization of risk becomes maximum \textit{a posteriori} (MAP) estimation. In this context maximizing the weighted \textit{a posteriori} probability in (3) over \( x \in \Omega \) can be seen as weighted MAP (WMAP) estimation, where \( \alpha(x_s) \) are the weights.

### 3 Weighted Iterated Conditional Modes

In [1] Besag introduced iterated conditional modes (ICM) as a means for maximizing \( P(x|y) \). By adapting ICM we can introduce a method for maximizing the weighted \textit{a posteriori} probability in (3). We begin the derivation of weighted ICM (WICM) by reformulating \( B(x|y) \) as follows:

\[ B(x|y) = P(x|y) \prod_{s \in S} \alpha(x_s) = \alpha(x_s) P(x_s|x_{-s},y) P(x_{-s}|y) \prod_{r \in S, r \neq s} \alpha(x_r) \]

\[ \propto \alpha(x_s) P(x_s|x_{\eta_s}, y_s) P(x_{-s}|y) \prod_{r \in S, r \neq s} \alpha(x_r) \quad (4) \]

The final step follows from the two assumptions of Section 2.1. Increasing the first two terms of (4) necessarily increases \( B(x|y) \). This suggests a global optimization strategy that sequentially visits each site \( s \in S \) and determines the label \( x_s \in \Lambda \) that maximizes \( \alpha(x_s) P(x_s|x_{\eta_s}, y_s) \). The maximization of this quantity is straightforward since \( P(x_s|x_{\eta_s}, y_s) \propto P(y_s|x_s) P(x_s|x_{\eta_s}) \). Note that WICM converges to a local maximum of \( B(x|y) \). The WICM algorithm is as follows:

**Weighted Iterated Conditional Modes**

**Input:** Initial labeling \( x^0 \), weights \( \alpha(\omega) \)

**Output:** Final labeling \( x^k \) after iteration \( k \)

1. \( k = 0 \)
2. do
3. \( k = k + 1 \)
4. \( x^k = x^{k-1} \)
5. for \( \forall s \in S \) do
6. \( x_s^k = \arg \max_{\omega \in \Lambda} \left[ \alpha(\omega) P(\omega|x_{\eta_s}^k, y_s) \right] \)
7. end for
8. while $x^k \neq x^{k-1}$

It is illuminating to consider the case of binary classes, i.e. $A = \{\omega_1, \omega_2\}$. In this situation step 6 in the above algorithm simplifies to the following:

$$x_s^k = \begin{cases} 
\omega_1 & \text{if} \ P(\omega_1 | x_s^k, y_s) > \frac{\alpha(\omega_1)}{\alpha(\omega_1) + \alpha(\omega_2)} = T_{wicm} \\
\omega_2 & \text{otherwise.}
\end{cases}$$

Consequently, increasing $T_{wicm} \in [0, 1]$ results in a greater preference for $\omega_2$, while decreasing $T_{wicm}$ increases the tendency toward $\omega_1$. Note that when $T_{wicm} = 1/2$, WICM is equivalent to ICM.

4 Experimental Results and Discussion

In this section we evaluate WICM by incorporating it into two separate MRF-based classification systems for detecting prostate cancer (CaP). The goal of both systems is to classify their respective sites $S$ (specifically glands or pixels) into one of two classes: $X_s \in A = \{\omega_1, \omega_2\}$, where $\omega_1$ and $\omega_2$ indicate malignancy and benignity. The basic procedure for both systems is similar: 1) Using the distribution $f(y_s | x_s)$, a Bayesian classifier assigns each site $s$ a probability of malignancy $P(x_s | y_s)$ based solely on its feature vector $y_s$. 2) If this probability exceeds the threshold $T_f$, the state $x_s$ of site $s$ is labeled malignant; otherwise it is labeled benign. 3) Using these labels as the initial conditions $x_0$, WICM produces a final labeling $x^k$. The performance (e.g. sensitivity and specificity) of this procedure is a function of the two thresholds $T_f$ and $T_{wicm}$. $T_f$ is fixed at an empirically chosen value. (We have observed that system performance is remarkably consistent over a wide range of $T_f$.) We vary $T_{wicm}$ to adjust classifier performance. In the following subsections we use this approach to construct ROC curves for our two classification systems. These ROC curves and the corresponding areas under them reflect the ability of each system to detect CaP.

It is worth mentioning that classification performance could conceivably be adjusted by fixing $T_{wicm}$ and varying $T_f$. That is, since modifying $T_f$ alters the initial conditions $x^0$, it can cause WICM to converge to a different local maximum of (3). However, there is no reason to assume that the individual modes (local maxima) of the weighted a posteriori distribution correspond to meaningful classifications; and consequently, this method has no obvious justification.

4.1 Detecting Cancerous Glands on WMHS

Methodology: The goal of this classification system is to detect malignant glands in whole-mount histological sections (WMHSs) of the prostate. Figure 1(a) illustrates a prostate WMHS. The black circle delimits the approximate spatial extent of CaP as delineated by a pathologist. The numerous white regions
are the glands, which our system automatically identifies and segments. Figure 1(b) illustrates the segmented gland boundaries in blue. Figure 1(c) provides a magnified view of the white box in Figure 1(b). Let the set $S = \{1, 2, \ldots, N\}$ reference the $N$ segmented glands in a WMHS. The random variable $Y_s \in \mathbb{R}$ indicates the area of gland $s$. Two glands are neighbors if the distance between their centroids is less than 0.7 mm. The MRF is implemented using pairwise probabilistic Markov models [7]. The distribution $p(y_s|x_s)$ is described using a parametric model—specifically, a mixture of Gamma functions. All results were produced using leave-one-out cross-validation over 20 WMHSs from 19 patients.

![Fig. 1. (a) H&E stained prostate histology section; black ink mark provided by pathologist roughly indicates CaP extent. (b) Gland segmentation boundaries. (c) Magnified view of white box in (b). (d) Centroids of cancerous glands after initial classification with $T_f = 0.3$. (e)-(g) Final labels after WICM for $T_{wicm} \in \{0.9, 0.5, 0.3\}$. (h) ROC curve indicates CaP detection performance over all 20 WMHSs.](image)

**Results:** The light green dots in Figure 1(d) indicate the centroids of those glands initially classified as malignant (i.e. $x^c$) with $T_f = 0.3$. Figures 1(e)-1(g) illustrate the final labels (i.e. $x^b$)—dark green dots indicate the centroids of the malignant glands—for $T_{wicm} \in \{0.9, 0.5, 0.3\}$. Notice that as $T_{wicm}$ decreases, indicating a increase in the weighting of the cancer class as compared to the
benign class, the sensitivity increases. The black line in Figure 1(h) indicates system performance\(^3\) over all 20 WMHSs as \(T_{wicm}\) varies from 0 to 1.

4.2 CaP Detection in Multi-protocol MRI

**Methodology:** In this section we consider a classification system for detecting CaP in multi-protocol in vivo MRI. This system combines functional and structural information from dynamic-contrast enhanced (DCE) and T2-weighted (T2-w) 3 Tesla MRI from six patients (18 2D slices). Figures 2(a) and 2(b) illustrate a T2-w MR image and a DCE MR image sampled at the first time point. The green overlays indicate the cancerous extent as specified by a radiologist. Let the set \(S = \{1, 2, \ldots, N\}\) reference the \(N\) pixels in a T2-w MR image that reside within the prostate. The random vector \(Y_s \in \mathbb{R}\) represents the 14 features associated with pixel \(s\). These features are comprised of the T2-w image intensity, six textural features [8] extracted from the T2-w image, and the DCE image intensity sampled at seven time points. The neighborhood \(\eta_s\) of a pixel \(s\) is the typical 8-connected region. The MRF is implemented using pairwise probabilistic Markov models. A random forest (i.e. bagging multiple decision trees classifiers) was used to determine the distribution \(f(y_s|x_s)\). All results were produced using leave-one-out cross-validation.

**Results:** The intensities in Figure 2(c) indicate the probability of malignancy at each pixel. Those pixels labeled as malignant during the initial classification with \(T_f = 0.1\) are indicated by the red overlay in Figure 2(d). Figures 2(e)-(g) illustrate the malignant labels after WICM for \(T_{wicm} \in \{0.8, 0.275, 0.008\}\). The black ROC curve in Figure 2(h) indicates system performance over all 18 studies as \(T_{wicm}\) varies from 0 to 1.

5 Concluding Remarks

The inter-variable dependencies within random fields (RFs) necessitate the use of sophisticated estimation strategies. Unfortunately, these strategies provide no clear means for varying classification performance (e.g. sensitivity/specificity). Instead, they produce a single, hard labeling at a static operating point. Addressing this deficiency, we introduced weighted maximum a posteriori estimation (WMAP), a generalized form of MAP estimation that allows certain classes to be weighted more heavily than others. We also introduced weighted iterated conditional modes (WICM), an RF-compatible classification strategy capable of WMAP estimation. To illustrate the value of WICM we applied it to separate MRF-based classification systems for detecting prostate cancer (CaP) in whole-mount histological sections and multi-protocol MRI. We demonstrated how WICM could arbitrarily vary the CaP sensitivity of these systems by appropriately altering their associated cost functions.

\(^3\) The true positive rate is the ratio of malignant sites correctly classified to the total number of malignant sites. The false positive rate is the ratio of benign sites incorrectly classified to the total number of benign sites.
Though the concept of weighting probabilities as a means for adjusting classifier performance is well established, its extension to RFs is completely novel. This extension, in retrospect, may seem obvious; yet, previous authors [5, 6], requiring such capabilities, have instead resorted to using ad hoc methods.

**Acknowledgment**

This work was made possible via grants from the Wallace H. Coulter Foundation, New Jersey Commission on Cancer Research, National Cancer Institute (R01CA136535-01, ARRA-NCI-3 R21 CA127186-02S1, R21CA127186-01, R03CA128081-01), the Society for Imaging Informatics in Medicine (SIIM), the Cancer Institute of New Jersey, and the Life Science Commercialization Award from Rutgers University. Additionally, we would like to thank Michael D. Feldman and John E. Tomaszewski from the Dept. of Surgical Pathology at the University of Pennsylvania for the whole mount histology sections, and Dr. Nicholas B. Bloch and Dr. Neil Rofsky from the Beth Israel Deaconess Medical Center for the prostate MRI data.
References

Commentary on *Weighted Iterated Conditional Modes for Random Fields: Application to Prostate Cancer Detection*

Paulo R. S. Mendonça

GE Global Research, One Research Circle, Niskayuna, NY 12309, USA
mendonca@crd.ge.com

**Abstract.** The paper “Weighted Iterated Conditional Modes for Random Fields: Application to Prostate Cancer Detection,” in *Prob. Models for Med. Image Analysis*, 2009, by J. Monaco et al., presents an extension of maximum a posteriori (MAP) estimation in a Markov random field (MRF) by introducing weights in the cost function optimized by MAP, and applies this idea to the generation of receiver operating characteristic curves for classification. Perhaps unknowingly to the authors, this conceptually simple generalization of MAP estimation of MRFs has an interesting connection with other areas of decision theory. A discussion of this connection provides theoretical insight into the authors’s contribution and allows for practical elements of the proposed algorithm to be implemented on firmer footing.

**Commentary**

In their excellent paper, Monaco et al. [1] demonstrate how a simple extension to maximum a posteriori (MAP) estimation in a Markov random field (MRF) allows for a natural mechanism for producing receiver operating characteristic (ROC) curves on the MRF’s performance for a given classification task. This contrasts with the static operating point typically found in the literature.

This extension is related to a powerful connection, unexplored in the paper, between Bayesian methods and techniques from decision theory based on utility functions. In particular, the introduction of the weights that generalize equation (1) into equation (2) makes clear that $C_{MAP}$ (as well as $C_{WMAP}$) can be interpreted as a log-additive utility function [2]. Moreover, for a log-additive utility function $U$ given by $U = \alpha c_1 + \beta c_{II}$, where $\alpha$ and $\beta$ are the probabilities of a type I error (false rejection of a true null hypothesis) and a type II error (false rejection of a false null hypothesis), respectively, and $c_1$ and $c_{II}$ are application-dependent costs associated with each type of error, the minimum of $U$ is achieved at a specific point along the ROC curve of the associated Bayesian classifier [3,4]. Therefore, the proposed method works by varying the probabilities of the loss incurred by type I and type II errors. Although this connection is well known, the typical approach for sweeping ROC curves for a Bayesian classifier is to use different thresholds on *Bayes factors* [5], and in doing so the connection between these different approaches is lost.

A practical consequence of this observation is the provision of a theoretical justification for varying thresholds $T_f$ and $T_{wicm}$ simultaneously. One could already ask why
the decision thresholds would be different depending on whether neighborhood information is \( P(x_i|\mathbf{x}_{\text{nb}}, y_i) \) or is not \( P(x_i|y_i) \) used. In view of the discussion above this question could be reformulated as: why would \( U = U_f \) obtained from \( P(x_i|y_i) \) be kept fixed while \( U = U_{\text{wicm}} \) obtained from \( P(x_i|\mathbf{x}_{\text{nb}}, y_i) \) is allowed to vary? From a theoretical standpoint it seems natural to assume \( c_I \) and \( c_{II} \) the same in both cases and produce the ROC curve by jointly varying \( T_f \) and \( T_{\text{wicm}} \). Given the authors's important observation that "system performance is [...] consistent over a wide range of \( T_f \)" (Section 4), this change will have limited effect on performance on the data in which the proposed method was tested. Nevertheless, this may not be the case for different data, and it certainly adds to the elegance of the approach.

References