Activity Recognition from Physiological Data using Conditional Random Fields

Hai Leong Chieu[1], **Wee Sun Lee**[2], **Leslie Pack Kaelbling**[3]
[1]Singapore-MIT Alliance, National University of Singapore
[2]Department of Computer Science, National University of Singapore
[3]CSAIL, Massachusetts Institute of Technology

***Abstract*— We describe the application of conditional random fields (CRF) to physiological data modeling for the application of activity recognition. We use the data provided by the Physiological Data Modeling Contest (PDMC), a Workshop at ICML 2004. Data used in PDMC are sequential in nature: they consist of physiological sessions, and each session consists of minute-by-minute sensor readings. We show that linear chain CRF can effectively make use of the sequential information in the data, and, with Expectation Maximization, can be trained on partially unlabeled sessions to improve performance. We also formulate a mixture CRF to make use of the identities of the human subjects to further improve performance. We propose that mixture CRF can be used for transfer learning, where models can be trained on data from different domains. During testing, if the domain of the test data is known, it can be used to instantiate the mixture node, and when it is unknown (or when it is a completely new domain), the marginal probabilities of the labels over all training domains can still be used effectively for prediction.**

Index Terms: Machine Learning, Graphical Models, Applications

I. Introduction

This paper describes the application of conditional random fields (CRF) [1] to the task of activity recognition from physiological data. We apply CRF to the two activity recognition tasks proposed at the Physiological Data Modeling Contest (PDMC), a workshop at the Twenty-First International Conference on Machine Learning (ICML-2004). The physiological data provided at PDMC were sequential in nature: they consists of sessions of physiological signals, and each session consists of minute-by-minute sensor readings. Three tasks were defined at PDMC, a gender prediction task and two activity recognition tasks. In this paper, we only work on the activity recognition tasks. We show that the linear chain CRF (L-CRF) outperforms all participants at the PDMC, and we formulate Generalized Expectation Maximization [2] updates for CRF to make use of partially labeled sequences.

The data provided at PDMC consists of physiological sessions. Each session is provided with a user identity number, and two characteristics of the users. Each minute of the session consists of nine types of sensor readings. The semantics behind the characteristics and the sensors, as shown in Table I, were provided only after the contest. The training data is also provided with a gender for each session, and an activity code for each minute of a session. However, as observed in [3] and [4], it is in general desirable to normalize sensor readings for each user. To take user information into account, we formulate a mixture CRF (M-CRF), which allows inference either with or without user information. When the user identity is known and has been seen in training, we can leverage on this information by instantiating the mixture node with the correct user identities. On the other hand, if we do not know the user identity, or if we are faced with a new user, mixture CRF can also allow inference to be done by taking the marginal probability of the labels by summing the joint probabilities for all users seen in training. We show that with this mode of inference, M-CRF outperforms L-CRF.

TABLE I
Semantics of the characteristics of the human subjects and the sensor readings

Name	Semantics
characteristic1	age
characteristic2	handedness
sensor1	gsr low average
sensor2	heat flux high average
sensor3	near body temp average
sensor4	pedometer
sensor5	skin temp average
sensor6	longitudinal accelerometer SAD
sensor7	longitudinal accelerometer average
sensor8	transverse accelerometer SAD
sensor9	transverse accelerometer average

II. Conditional Random Fields

Conditional random fields [1] are discriminative, undirected graphical models. They have been shown to perform well in a variety of tasks including part-of-speech tagging[1], shallow-parsing [5] and object recognition in machine vision [6]. In this paper, we used the linear chain CRF for activity recognition, and propose a mixture CRF for transfer learning when the human subject has already been seen in training.

We denote $\mathbf{X}$ as a random variable over data sequences to be labeled, and $\mathbf{Y}$ a random variable over corresponding sequences. $\mathbf{X}$ corresponds to the observed sensor readings of entire sequences, and $\mathbf{Y}$ corresponds to entire sequences of labels to each node in the sequence (see Figure 1). Each component Y_i of $\mathbf{Y}$ range over an alphabet Y. In the application of activity recognition from physiological signals, each sequence (or linear chain) is a session of physiological data which consists of readings taken at each minute of the session. In this setting, each component X_i of $\mathbf{X}$ is a vector of sensor

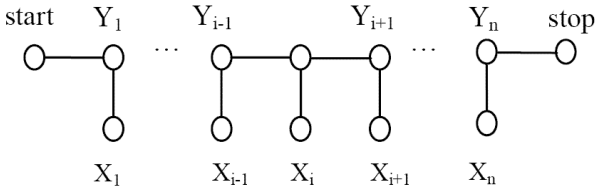


Fig. 1. Linear chain conditional random fields

readings taken at each minute, and each component Y_i of $\mathbf{Y}$ ranges over the activities to be recognized.

In general, a CRF is defined as follows [1]:

Definition: Let $G = (\mathbf{V}, \mathbf{E})$ be a graph such that $\mathbf{Y} = (Y_v)_{v \in V}$, so that $\mathbf{Y}$ is indexed by the vertices of G. Then $(\mathbf{X}, \mathbf{Y})$ is a conditional random field in case, when conditioned on $\mathbf{X}$, the random variables Y_v obey the Markov property with respect to the graph: $p(Y_v|\mathbf{X}, \mathbf{Y}, w \sim v) = p(Y_v|\mathbf{X}, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G.

By the fundamental theorem of random fields [7], the general form of the joint distribution of the labeled sequence $\mathbf{Y}$ given $\mathbf{X}$ has the form:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp(\sum_{c \in C} \Phi(\mathbf{y_c}, \mathbf{x_c})),$$

where $C = \{\{\mathbf{y_c}, \mathbf{x_c}\}\}$ is the set of cliques in the graph G, and $Z(\mathbf{x})$ is a normalization factor.

In the case of a linear chain, each edge of the form $(Y_i, \mathbf{X_i})$ or (Y_i, Y_{i+1}) forms a clique, and the joint distribution can be expressed as

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \quad (\sum_{e \in E, k} \lambda_k f_k(e, \mathbf{y}|_\mathbf{e}, \mathbf{x}) + \sum_{v \in V, k} \mu_k g_k(v, \mathbf{y}|_\mathbf{v}, \mathbf{x})),$$

where $\mathbf{x}$ is a data sequence, $\mathbf{y}$ a label sequence and $y|_s$ is the set of components of $\mathbf{y}$ associated with the vertices in the subgraph S. The functions f_k and g_k are features: f_k are features on the edges of the form (Y_i, Y_{i+1}) and g_k features on the edges of the form $(\mathbf{X_i}, Y_i)$ in the linear chain. To simplify notation, from here onwards, we will not distinguish between f_k and g_k in our formulation, and simply write

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp(\mathbf{\Lambda} \cdot \mathbf{F}(\mathbf{y}, \mathbf{x})),$$

where $F(\mathbf{y}, \mathbf{x})$ is the global feature vector for the input sequence $\mathbf{x}$ and label sequence $\mathbf{y}$, comprising of the f_k's and the g_k's.

The parameter estimation problem is to determine, from the training data $D = \{(\mathbf{x^{(j)}}, \mathbf{y^{(j)}})\}_{j=1..N}$, the parameters in Λ. We determine Λ by maximizing the log-likelihood of the training data:

$$L_\Lambda = \sum_j [\mathbf{\Lambda}.\mathbf{F}(\mathbf{y^{(j)}}, \mathbf{x^{(j)}}) - \log Z_\Lambda(\mathbf{x^{(j)}})].$$

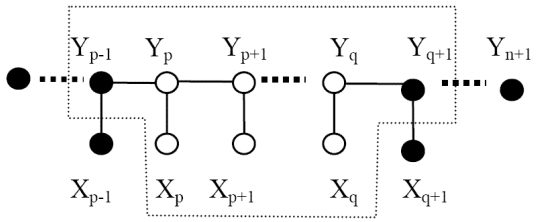


Fig. 2. Linear chain for partially labeled sequences. The black nodes are labeled and the white nodes are unlabeled

It is often useful to define a Gaussian prior over the parameters to avoid overfitting (a process that is sometimes called regularization), which changes the above objective function into

$$L_\Lambda = \sum_j [\mathbf{\Lambda}.\mathbf{F}(\mathbf{y^{(j)}}, \mathbf{x^{(j)}}) - \log Z_\Lambda(\mathbf{x^{(j)}})] - \frac{\|\mathbf{\Lambda}\|^2}{2\sigma^2}.$$

We use a gradient based algorithm for maximizing the log likelihood, which requires the calculation of the gradient of the regularized log-likelihood [5]:

$$\nabla L_\Lambda = \sum_j [\mathbf{F}(\mathbf{y^{(j)}}, \mathbf{x^{(j)}}) - E_{p_\Lambda(\mathbf{Y}|\mathbf{x^{(j)}})} F(\mathbf{Y}, \mathbf{x^{(j)}})] - \frac{\lambda}{\sigma^2}.$$

The above gradient term requires the calculation for each sequence $\mathbf{X}$ of the expected feature values over all possible Y over the entire sequence $\mathbf{X}$. For the linear chain, this can be done efficiently by the forward backward algorithm. The gradient based approach we used is the limited memory variable metric (lmvm) algorithm provided in the Toolkit for Advanced Optimization [8].

III. Generalized Expectation Maximization

In this section, we formulate Expectation Maximization (E.M.) [2] updates for CRFs in partially labeled graphs, where some (but not all) of the variables are hidden during training.

Partially labeled $\mathbf{Y}$ can be used under E.M. settings for the L-CRF. Under E.M. settings, we maximize the expected log likelihood LL of the incomplete data given the labeled data at each iteration:

$$\begin{aligned} LL &= \sum_z P(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathbf{\Lambda^t}) \log P(\mathbf{z}, \mathbf{y}|\mathbf{x}, \mathbf{\Lambda}) \\ &= \sum_z P(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathbf{\Lambda^t}) \log \frac{1}{Z(x)} exp(\sum_{c \in C} \Phi(\mathbf{y_c}, \mathbf{x_c})) \\ &= \sum_z P(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathbf{\Lambda^t})(\sum_{c \in C} \Phi(\mathbf{y_c}, \mathbf{x_c})) \\ &\quad - \sum_z P(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathbf{\Lambda^t}) log(Z(\mathbf{x})) \\ &= \sum_z P(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathbf{\Lambda^t})(\sum_{c \in C} \Phi(\mathbf{y_c}, \mathbf{x_c})) - logZ(\mathbf{x}) \end{aligned}$$

The gradient for the expected log likelihood of the incomplete data is

$$\frac{\delta LL}{\delta \lambda_i} = \sum_z P(\mathbf{z}|\mathbf{x},\mathbf{y},\mathbf{\Lambda^t}) f_i - \sum_{y,z} P(\mathbf{y},\mathbf{z}|\mathbf{x},\mathbf{\Lambda}) f_i$$
$$= E_{p^{(t-1)}(z|x,y)}[f_i] - E_{p^t(y,z|x)}[f_i]$$

where **z** is the unlabeled sub sequence, **x** the observations, **y** the labeled sub sequence, **t** the parameters at the last iteration (iteration t), and $\mathbf{\Lambda}$ the parameters to be optimized. The E-step in E.M. requires calculation of expected feature values for unlabeled nodes given the rest of the graph. In the M-step, a gradient based approach is used to maximize the expected log likelihood of the incomplete data with the above gradient.

If all variables are hidden, since components of the gradients (of the form $E_{\tilde{p}}[f_k] - E_p[f_k]$) will be zero, gradient based optimization techniques will ignore the unlabeled data. However, if some of the variables are instantiated, $E_{\tilde{p}}[f_k]$ for unlabeled nodes will be the expected feature value in the partially instantiated graph, and this will be different from $E_p[f_k]$, the expected feature value in the totally un-instantiated graph. We show in Figure 2 a partially labeled linear chain, where black nodes represent labeled instances and white nodes represent unlabeled instances. Note that the two nodes Y_{p-1} and Y_{q+1} d-separate the unlabeled chain from the rest of the chain (i.e. the unlabeled chain is independent of the labeled chain given Y_{p-1} and Y_{q+1}). The probability of the unlabeled chain $P(\mathbf{z}|\mathbf{x},\mathbf{y},\mathbf{t})$ can hence be calculated by the same forward backward algorithm within the sub-chain starting at node Y_{p-1} and ending at node Y_{q+1} (these two nodes are labeled). The transition matrices $\mathbf{M_i}$ in the subchain are the same as those in the original chain. Initialization of the forward backward vectors are $f_0^m(y|x) = \delta(y, y_{p-1})$ and $b^m_{q-p+2}(y|x) = \delta(y, y_{q+1})$. During training, we need to calculate expected feature values for the unlabeled nodes given current parameters $\mathbf{\Lambda}$ and the labeled portion of each chain, and this can be done from the above forward and backward vectors.

As we are using an iterative method (lmvm) for optimizing log likelihood, using E.M. requires parameters to converge at each E.M. iteration. Each lmvm iteration takes a long time due to the data size. As a result, we use generalized E.M (G.E.M.) [2], and run only a few iterations of the lmvm algorithm during each E.M. iteration.

A. Mixture of Conditional Random Fields

In this section, we introduce the mixture CRF. In many applications in machine learning, it is often necessary to apply models trained in one domain to test data from a different domain. However, machine learning algorithms often assume that the distribution of the test data is the same as that of the training data. We propose a mixture node for CRFs, that allows training on a few different domains. During testing, if the domain is known, the mixture node can be instantiated with the correct domain. If the domain is unknown, the model can still be used by calculating the marginal probability over all domains. In the context of physiological data modeling, we use the user identity as the mixture node. Without transfer learning, one can either (i) make use of user information by training separate models for each user or (ii) ignore user information and train one model for all users. PDMC participants were advised to ignore user information and all used approach (ii). Moreover, there were human subjects in the test data that were not seen in the training data. We show that we can use M-CRF to leverage on user identities when they have been seen in training, and for new users, M-CRF still performs well by taking marginal probabilities over the users.

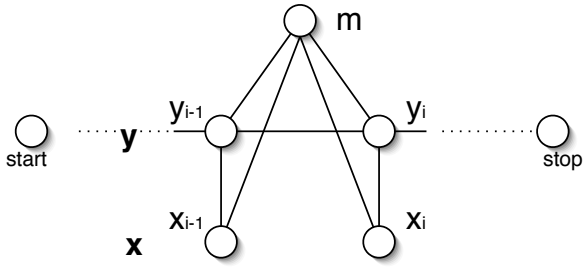


Fig. 3. Mixture conditional random fields

The structure of a M-CRF is as shown in Figure 3. The maximal clique size in the M-CRF is 3: inference can be done efficiently using belief propagation on junction trees [9]. Here, we formulate inference algorithms for the M-CRF using the forward backward procedure. In this formulation, we use an incomplete parameterization of the M-CRF, allowing only features conditioned on the pairs $(M, \mathbf{X_i})$, (M, Y_i), $(\mathbf{X_i}, Y_i)$ and (Y_i, Y_{i+1}). These features include the usual features of the L-CRF, $f_{y,y'}(Y_i, Y_{i+1})$ and $h_{y,xk}(Y_i, \mathbf{X_i})$, where $f_{y,y'}$ is the indicator function for state transitions from y to y', and $h_{y,xk}$ is the value of the k^{th} element of $\mathbf{X_i}$ if Y_i equals y, and zero otherwise. Besides these features, the M-CRF also uses features $p_{m,y}(M, Y_i)$, which are indicator functions of the mixture node-state pair (m,y), and $q_{m,xk}(M, \mathbf{X_i})$, which equals to the k^{th} element of $\mathbf{X_i}$ when M equals m and zero otherwise. With these features, the sequence **y** with different mixture nodes will share the parameters for the features $f_{y,y'}$ and $h_{y,xk}$, but will have different parameters for the features $p_{m,y}$ and $q_{m,xk}$. If the mixture node represents the domain of the sequence, then model information is shared across domains, while each individual domain could still have a model that can account for features specific to itself.

The conditional probability of the mixture node M and the **Y** chain given the observations **X** is as follows:

$$P(m,\mathbf{y}|\mathbf{x}) = \frac{1}{\sum_M Z(x,m)} \exp\Big(\sum_i \alpha_{y^i,y^{i+1}} + \beta_{m,y^i} + \sum_k \gamma^{(k)}_{y^i,k} x_i^{(k)} + \delta^{(k)}_{m,k} x_i^{(k)}\Big)$$

$$Z(\mathbf{x},m) = \sum_{\mathbf{y}} \exp\Big(\sum_i \alpha_{y^i,y^{i+1}} + \beta_{m,y^i} + \sum_k \gamma^{(k)}_{xk,y^i} x^{(k)} + \delta^{(k)}_{xk,m} x^{(k)}\Big),$$

where α, β, δ, and γ are parameters for f, p, h, and q respectively, and $x_i(k)$ is the k^{th} element of the vector $\mathbf{x_i}$. (In the above expressions, feature values have been evaluated to either $x_i(k)$ for $h_{y,xk}$ and $q_{m,xk}$, or 1 for $f_{y,y'}$ and $p_{g,y}$). By writing the numerator of $P(m, \mathbf{y}|\mathbf{x})$ as $exp(\mathbf{\Lambda}.\mathbf{F}(\mathbf{y}, \mathbf{x}))$, where $\mathbf{\Lambda}$ is the vector of the parameters and $\mathbf{F}$ is the global feature vector over the entire sequence, we maximize the log-likelihood L_Λ of the data D regularized with a spherical Gaussian prior, by a gradient based approach, with

$$L_\Lambda = \sum_{j \in D} \left(\mathbf{\Lambda}.\mathbf{F}(\mathbf{y_j}, \mathbf{x_j}) - \log \sum_M Z(\mathbf{x_j}, M) \right) - \frac{\|\Lambda\|^2}{2\sigma^2;}$$

$$\frac{\partial L_\Lambda}{\partial f} = E_{\tilde{p}}[f] - E_p[f] - \frac{\lambda}{\sigma^2},$$

where $E_{\tilde{p}}[f]$ is the empirical average of the feature f, and $E_p[f]$ is the expected feature value given the current model p. Expressions for $E_p[f]$ are as follows:

$$E_p[f_{y1,y2}] = \sum_{D,i} \sum_m P(m, y_i = y_1, y_{i+1} = y_2|\mathbf{x}) \ ;$$

$$E_p[h_{xk,y}] = \sum_{D,i} \sum_m P(m, y_i = y|\mathbf{x}) x_i^{(k)};$$

$$E_p[p_{m,y}] = \sum_{D,i} P(m, y_i = y|\mathbf{x}) \ ;$$

$$E_p[q_{xk,m}] = \sum_{D,i} \left[\sum_{y1,y2} P(m, y1, y2|\mathbf{x}) \right] x_i^{(k)}.$$

With these expressions, the structure in Figure 3 can be decomposed into $|M|$ separate linear chains, one for each value of M. These chains share the same parameters α's and γ's, but have different β's and δ's (indexed by M). We define the transition matrix $M_i^m(y, y'|\mathbf{x})$, from which the normalization factors $Z(\mathbf{x}, m)$ and the forward backward vectors can be calculated:

$$\begin{aligned}
M_i^m(y, y'|\mathbf{x}) &= \exp(\alpha_{y,y'} + \beta_{m,y'} + \\
&\quad \sum_k \gamma_{xk,y'}^{(k)} x^{(k)} + \delta_{xk,m}^{(k)} x^{(k)}); \\
Z(\mathbf{x}, m) &= \left(\prod_{i=1}^{n} M_i^m(\mathbf{x}) \right)_{start,stop} ; \\
f_i^m(\mathbf{x})^T &= f_{i-1}^m(\mathbf{x})^T M_i^m(\mathbf{x}); \\
b_i^m(\mathbf{x}) &= M_i^m(\mathbf{x}) b_{i+1}^m(\mathbf{x}); \\
f_0^m(y|\mathbf{x}) &= \delta(y, start); \\
b_{n+1}^m(y|\mathbf{x}) &= \delta(y, stop).
\end{aligned}$$

The probabilities $P(m|\mathbf{x}), P(m, y|\mathbf{x})$ and $P(m, y_1, y_2|\mathbf{x})$ can then be calculated as follows:

$$P(m, y|\mathbf{x}) = \frac{f_i^m(y) b_i^m(y)}{\sum_M Z(\mathbf{x}, M);}$$

$$P(m|\mathbf{x}) = \frac{Z(\mathbf{x}, m)}{\sum_M Z(\mathbf{x}, M);}$$

$$P(m, y1, y2|\mathbf{x}) = \frac{f_i^m(y1) M_i^m(y1, y2) \beta_i^m(y2)}{\sum_M Z(\mathbf{x}, M)}.$$

TABLE II
COMPARISON OF TEST SCORES WITH THE TOP THREE SYSTEMS AT PDMC

Number of	Training	Test
Total Minutes	580,264	720,792
Total Sessions	1,410	1,713
Minutes of TV	4,413	5,813
Minutes of TV	98,172	103,666
Minutes of TV	66	72
Minutes of TV	235	244

Note that this model is discriminative for the pair (M, **Y**), and no longer discriminative for either M or **Y** alone. Evaluation of gradient and log likelihood for the M-CRF can be performed in $O(|Y|^2.L.|M|)$ where L is the length of the chain, as compared to $O(|Y|^2.L)$ for the L-CRF.

In the PDMC task, the mixture node correspond to user identities for each physiological session, and the sequence of label $\mathbf{y}$ still corresponds to the activities at each minute. For activity prediction, inference can be done in two ways: (a) if user identity is known and has been seen in training, the mixture node can be instantiated with the user identity and we can take the label y* with the highest joint probability $y^* = \arg\max_y P(m = user, y|\mathbf{x})$. (b) if the user identity is unknown, or when if its a new user, then we take the most likely label y* given the entire sequence of observations: $y^* = \arg\max_y P(y|\mathbf{x}) = \arg\max_y \sum_m P(m, y|\mathbf{x})$ by marginalizing over all users seen in training.

IV. EXPERIMENTAL RESULTS

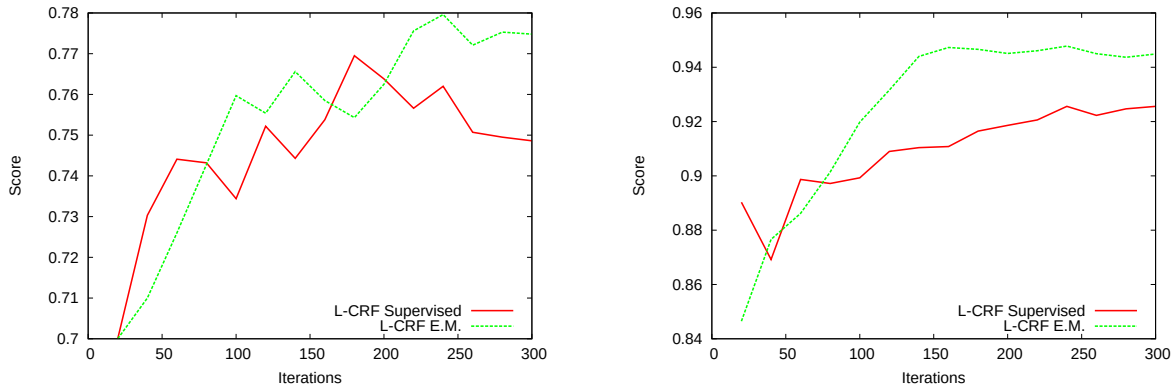
The scoring metric used at PDMC for the activity recognition tasks is as follows:

$$score = 0.3 \frac{TP}{TP + FN} + 0.7 \frac{TN}{TN + FP},$$

where TP=true positives, FP=false positives, TN=true negatives, and FN=false negatives. With this metric, the baseline of guessing all negatives will achieve a score of 0.7. The two target activities for prediction is Watching TV and Sleeping. The number of positive training instances for each of the two tasks is shown in Table II. While there are lots of positive training examples for Sleeping, there are fewer positive training examples for Watching TV. At PDMC, Sleeping has been shown to be the easier task and almost all participants performed better than the baseline of 0.7. For Watching TV, however, a number of participants did worse than this baseline.

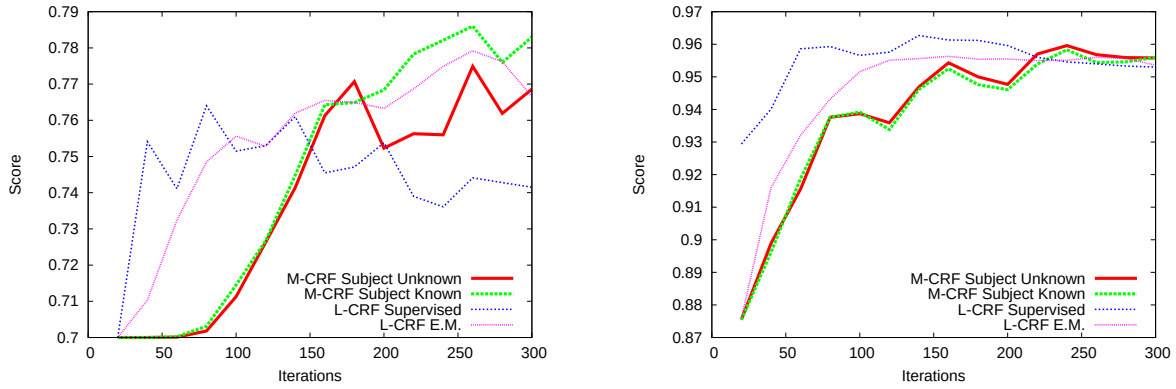
A. Linear CRF

Instead of using the feature values as they are, we find it beneficial to cluster each sensor values into 3 Gaussians using E.M. Each sensor value is then converted into a vector of 3 values, which are the probabilities that it belong to each of the 3 Gaussians. We run the L-CRF on the PDMC data under two settings: (i) we use all features shown in Table I and (ii) we exclude the two characteristics and use only the nine sensor values as features. In (i), we clustered age and the 9 sensor values into 3 clusters each. For handedness which is boolean, we keep it as one boolean. In (ii), each of the 9 sensors are



(a) Watching TV (b) Sleeping

Fig. 4. Performance on the activity recognition tasks using the two characteristics and the nine sensors as features



(a) Watching TV (b) Sleeping

Fig. 5. Performance on the activity recognition tasks using only the nine sensors as features

TABLE III
COMPARISON OF TEST SCORES WITH THE TOP THREE SYSTEMS AT PDMC

System	TV	Sleep	Average
L-CRF-EM(ii)	0.7665	0.9536	0.8601
L-CRF-EM(i)	0.7748	0.9449	0.8502
L-CRF(ii)	0.7415	0.9530	0.8473
L-CRF(i)	0.7486	0.9256	0.8371
Informedia-3	0.7375	0.9125	0.8250
NLM-3	0.7208	0.8938	0.8073
SmartSignal	0.7498	0.8684	0.8091

clustered into 3 clusters. For all our experiments (for both L-CRF and M-CRF), we use a Gaussian prior with a variance of 10, as recommended in [10].

The reason why we chose to omit the two characteristics for (ii) was because it seems that age and handedness would not help in predicting the two target activities at PDMC: Watching TV or Sleeping. However, as the PDMC participants do not know the semantics of the characteristics and sensors, a few of them used all features (e.g. [11]), while a few others did feature selection that excluded these features (e.g. [3]). We show that L-CRF under settings (ii) did better than (i), but both outperform all participants at PDMC. We plotted the score against the number of iterations of lmvm performed by the TAO toolkit in Figure 4 and 5. From the graphs, we see that using unlabeled data with E.M. is generally beneficial. For comparison with the performance of PDMC participants, we tabulated the perform of the L-CRFs at 300 iterations of lmvm under conditions (i) and (ii) in Table III. Among the top systems for activity recognition at PDMC, Informedia-3 [11] used support vector machines (SVM) with an rbf kernel for

minute-by-minute classification, NLM-3 [12] used atemporal Bayesian networks, and Smartsignal [3] used feature selection and a similarity based approach to predict windows of the activities. Both Informedia-3 and NLM-3 ignored sequence information. Informedia tried using SVM-based Markov models, but failed to achieve good performance, citing skewed data distribution as the reason. NLM-1 and NLM-2 [12] used sequence information with dynamic Bayesian networks, but performance was worse than NLM-3 which used an atemporal Bayesian network approach. We show that CRF could effectively make use of sequence information: all our CRF systems outperform all entries at PDMC on both activity recognition tasks.

Without using E.M., the unlabeled instances have to be discarded and for those in the middle of a session, removing them requires cutting such sequences into two separate sequences. Unlabeled instances makes up the bulk (70%) of the training data. Sequences that are entirely unlabeled are removed since they do not influence learning with E.M. in CRF. Among partially labeled sequences, unlabeled instances still make up the majority. For CRFLinear-EM, instead of using all such sequences, we remove unlabeled instances at the beginning or end of a session, and use only those in between labeled ones. This reduces unlabeled instances to be about 32% of the total data used in runs with E.M.

B. Mixture CRF

In this section, we investigate the effectiveness of incorporating user information into the model by using M-CRF. The training data at PDMC consists of physiological sessions of 18 different subjects, and the test data consists of data from 30 different subjects, 17 of which have been seen in training.

As L-CRF has shown that the two characteristics does not help in classifying the two activities, we run M-CRF only in settings (ii), where only the nine sensor values are used as features. Inference with the M-CRF are done in two ways, (a) taking the label with the highest marginal $p(y|\mathbf{x})$ by summing out the mixture node and (b) in cases where the user is known, make use of the user identities and use the label with maximal joint probability $p(y, m = userid|\mathbf{x})$.

We plotted the performance of the M-CRF in Figure 5. From the graphs, it can be seen that M-CRF outperforms L-CRF for the TV task even when the user is assumed to be unknown during testing (settings (a)). When the user is known, M-CRF does even better on the TV task by leveraging on the user information. On the Sleep task, however, performance remains more or less the same. It seems that for the Sleep task, the signals themselves provides sufficient evidence and user identity does not help to improve prediction.

V. Related Work

Physiological signals provide an interesting platform for machine learning algorithms as they are context dependent, noisy, and sequential in nature. As physiological sensing equipment becomes wearable, it is sometimes less invasive than alternative surveillance equipments such as video. Previous work on modeling physiological signals have mainly focused on emotion recognition. [13] detect emotions such as anger, hate, and love by using physiological signals gathered from four sensors. They described methods to extract, select and transform features, and used a generative MAP approach by fitting Gaussians to the transformed data. [4] used the Bodymedia armband (the same armband used for gathering PDMC data) to collect physiological signals for emotion classification, and showed that Discriminant Function Analysis performed better than a k-Nearest Neighbor approach on their dataset. In their experiments, they normalized features with corresponding data collected during relaxation periods for the same user. We show that by using a mixture CRF, performance is indeed improved when user information is known.

Conditional random fields were defined as discriminative learning algorithm for undirected graphical models [1]. Most work using CRF use the linear chain CRF, for which there are efficient inference algorithms. The linear chain CRF has previously been used for part-of-speech tagging [1], shallow parsing [5] and named entity recognition [14]. Besides linear chain CRF, [15] have also cast the information extraction problem as a graph partitioning problem for CRF, but this generalization means the efficient dynamic programming that works for the linear chain CRF are no longer applicable and approximations have to be made for calculations to remain tractable. [16] has used CRF for transfer learning with factorial CRFs: during training, the models for the subtasks were trained independently, and during testing, the learned weights are combined into a single grid-shaped factorial CRF. In our formulation with mixture CRF, both training and testing were performed jointly on all training data. [6] proposed learning CRFs with hidden variables for object recognition, where the hidden variables correspond to parts of objects. For their application, the object class corresponds to the mixture node, and the variables $\mathbf{y}$ are the hidden variables.

In this paper, we use the linear chain CRF for activity recognition, and defined a mixture CRF to leverage of information of the user identity to further improve performance. We formulated exact and efficient algorithms for training and inference in this CRF. We believe mixture CRFs, in the same way as sentence mixture models, can be used in applications such as language modeling or named entity recognition, where it is often useful to model the topic (e.g. finance, sports) or the zone (e.g. headline) of a sentence. The mixture node can be used for this purpose.

VI. Conclusion

In this paper, we used the linear chain CRF for activity recognition from physiological signals, and defined a mixture CRF for transfer learning between different users' physiological data models. We believe that mixture CRF can be used in applications where mixture Markov models have been used, such as in language modeling. Empirical performance on the PDMC dataset shows that both linear chain CRF and mixture CRF outperforms top participants at PDMC for the activity

recognition tasks. We show that mixture CRF can be used for transfer learning, where the mixture node defines the domain of the data, which can be either used to improve performance during testing, or ignored if the domain is unknown.

REFERENCES

[1] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 282–289.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[3] J. E. Mott and R. M. Pipke, "Physiological data analysis," in *Proceedings of the ICML Physiological Data Modeling Contest Workshop*, 2004.

[4] F. Nasoz, C. Lisetti, K. Alvarez, and N. Finelstein, "Emotional recognition from physiological signals for user modeling of affect," in *Proceedings of the 3rd Workshop on Affective and Attitude User Modeling*, 2004.

[5] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the Twentieth International Conference on Machine Learning*, 2003, pp. 282–289.

[6] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *Proceedings of the Eighteenth Annual Conference on Neural Information Processing Systems*, 2004.

[7] J. M. Hammersley and P. Clifford, "Markov fields on finite graphs and lattices," 1971.

[8] S. J. Benson, L. C. McInnes, J. Mor, and J. Sarich, "Tao user manual (revision 1.7)," Mathematics and Computer Science Division, Argonne National Laboratory, Tech. Rep., 2004.

[9] C. Huang and A. Darwiche, "Inference in belief networks: A procedural guide," *International Journal of Approximate Reasoning*, vol. 15, 1996.

[10] C. Sutton, K. Rohanimanesh, and A. McCallum, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," in *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.

[11] W. H. Lin and A. Hauptmann, "Informedia at pdmc," in *Proceedings of the ICML Physiological Data Modeling Contest Workshop*, 2004.

[12] M. Kayaalp, "Bayesian methods for diagnosing physiological conditions of human subjects from multivariate times series sensor data," in *Proceedings of the ICML Physiological Data Modeling Contest Workshop*, 2004.

[13] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, 2001.

[14] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields," in *Proceedings of Conference on Computational Natural Language Learning*, 2003.

[15] B. Wellner, A. McCallum, F. Peng, and M. Hay, "An integrated, conditional model of information extraction and coreference with application to citation matching," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2004.

[16] C. Sutton and A. McCallum, "Composition of conditional random fields for transfer learning," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2005.