Temporal Maximum Margin Markov Network

Xiaoqian Jiang[1], Bing Dong[2], and Latanya Sweeney[1]

[1] Data Privacy Lab, School of Computer Science
[2] Center for Building Performance and Diagnostics, School of Architecture
Carnegie Mellon University

Abstract. Typical structured learning models consist of a regression component of the explanatory variables (observations) and another regression component that accounts for the neighboring states. Such models, including Conditional Random Fields (CRFs) and Maximum Margin Markov Network (M3N), are essentially Markov random fields with the pairwise spatial dependence. They are effective tools for modeling spatial correlated responses; however, ignoring the temporal correlation often limits their performance to model the more complex scenarios. In this paper, we introduce a novel Temporal Maximum Margin Markov Network (TM3N) model to learn the spatial-temporal correlated hidden states, simultaneously. For learning, we estimate the model's parameters by leveraging on loopy belief propagation (LBP); for predicting, we forecast hidden states use linear integer programming (LIP); for evaluation, we apply TM3N to the simulated datasets and the real world challenge for occupancy estimation. The results are compared with other state-of-the-art models and demonstrate superior performance.

1 Introduction

Traditional Markov random field models concentrate on either the spatial dependence or the temporal correlation. Lafferty et al. [6] developed a statistical framework, Conditional Random Fields (CRFs), which accounts for spatial dependence, in addition to the explanatory variables (observations). Later, Taskar [14] extended the Support Vector Machine (SVM) to the Maximum Margin Markov Network (M3N), which has the same modeling capacity of the CRFs but can be computed more efficiently. Similar models considering spatial dependence include, the Structured SVM [15] and the Maximum Margin Training [12]. All of these models aim to combine spatial dependence and the information from observations for a single end task, multivariate classification. They have been successfully applied to problems like optical character recognition [10], object detection [1] and scene understanding [4]. However, these models overlook the state correlations over time, hence, are insufficient to handle data with strong temporal pattern.

On the other hand, temporal correlated models has been developed over the decades, models including Kalman filter [5], HMM [17] have been carefully studied by the optimization and control community. Successful applications including time series forecasting [3], speech recognition [11] and behavior classification

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part I, LNAI 6321, pp. 587–600, 2010.

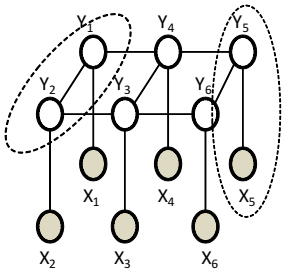

Fig. 1. Graphical model of CRFs and M3N. X_i and Y_i correspond to the local observations and their labels. Two dashed ovals encompass $[X_5, Y_5]$ and $[Y_1, Y_2]$, which correspond to a unary feature and a pairwise Markovian spatial feature, respectively.

[16]. These models are well known for their capability of capturing hidden temporal correlations; modeling the unknown state process from observations made in noisy environments. However, they ignore the structural correlations in the environment, which oftentimes hurt their performance.

Clearly, both temporal correlated models and spatial dependent models have limitations. [9] thus advocated a variational inference method for switching Linear Dynamical system (SLDS) that learns different dynamical processes at various time ticks; [7] extended this work to combine HMM and LDS with tractable computation. However, these methods treat temporal and spatial (structural) information once at a time; they fail to provide a comprehensive interface to model the temporal and spatial correlated real-world scenarios.

To close the gap, we propose a novel model that considers spatial correlations aggregated over time for tractable inference. The proposed model has advantages over models concentrating on either aspect, as the temporal and structural information are oftentimes complementary. We intend to provide a principled approach which accounts for spatial dependence and temporal correlations, simultaneously.

The remaining of the paper is organized as follows. In Section 2, we review the spatial-dependent structured learning framework. In Section 3, we suggest the spatial-temporal correlated framework extending the existing works. In Section 4, we instantiate the framework to propose a novel model: Temporal Maximum Margin Markov Network (TM3N). In Section 5, we introduce algorithms for estimating model parameters, leveraging on loopy belief propagation. In Section 6, we propose a linear integer programming interface for predicting hidden states. In Section 7, the TM3N model is applied to both the simulated datasets and a real world building occupancy estimation challenge. We compare the results with other state of the arts methods. Finally, in Section 8, we conclude the paper.

2 Overview

2.1 Notation

We summarize the basic notation of this paper in the following table.

Table 1. Summary of the notation

Variables	Summary
$X_{k,i,t}$	The k dimensional feature at site i in time tick t.
$Y_{i,t}$	The discrete valued state at site i in time tick t.
$\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})'$	The estimated states at time tick t.
$\hat{\mathbf{Y}}_t = (\hat{Y}_{1,t}, \ldots, \hat{Y}_{n,t})'$	The ground-truth states at time tick t.
$\theta^1_k, \theta^2_{ij,t}$ and $\theta^3_{it,t-1}$	The unary, spatial and temporal regression coefficients.
$\varphi(\cdot)$	The feature function.
$\ell_t(\mathbf{Y}_t)$	The loss at time t.
$h_\theta(\mathbf{X}_t)$	The state estimation function.

2.2 Backgrounds

We first summarize the framework that encompasses spatial dependent models on a regular or irregular lattice [6, 13]. Define $s_1, \ldots, s_n$ to be the sites on a spatial lattice. For notational convenience, let $j \sim i$ denote $j \in N_i$, where N_i=$\{j$: s_j is a neighbor of $s_i\}$ defines the neighbors of the site s_i. Let $Y_1, \ldots, Y_n$ denote hidden states on the lattice, where $Y_i = Y(s_i) \in (1, \ldots, C)$, and C is the number of classes. The joint distribution of $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ can be formulated as:

$$p_\theta(\mathbf{X}, \mathbf{Y}) \propto \exp\left\{ \sum_{i=1}^{n} \sum_{k=1}^{p} \theta_k \varphi(X_{k,i}, Y_i) + \sum_{j \sim i} \theta_{ij} \varphi(Y_i, Y_j) \right\}, \tag{1}$$

$$p_\theta(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z_\theta(\mathbf{X})} \exp\left\{ \sum_{i=1}^{n} \sum_{k=1}^{p} \theta_k \varphi(X_{k,i}, Y_i) + \sum_{j \sim i} \theta_{ij} \varphi(Y_i, Y_j) \right\}, \tag{2}$$

where

$$Z_\theta(\mathbf{X}) = \sum_{Y} \exp\{ \sum_{i=1}^{n} \sum_{k=1}^{p} \theta_k \varphi(X_{k,i}, Y_i) + \sum_{j \sim i} \theta_{ij} \varphi(Y_i, Y_j) \} \tag{3}$$

is called the partition function; $X_{k,i} = X_k(s_i)$ denotes the k-th explanatory variable at site s_i; θ_k denotes the k-th regression coefficients correspond to the feature function $\varphi(X_{k,i}, Y_i)$, with $k = 1, \ldots, p$; θ_{ij} denotes the spatial-dependent regression coefficients for the i-th and j-th sites so that $\theta_{ij} = \theta_{ji}$ and $\theta_{ij} \geq 0$ if $j \sim i$.

Such model relates a discrete valued response variable to a hidden state by two regression components; and it is capable of estimating the probability of hidden

states at a given site; and predicting a certain outcome at an unsampled site. However, this formulation ignores the fact that observations are oftentimes made repeatedly over time and past states on the same spatial lattice may contribute to the states in a future time tick. That is, for a given location s_i at a given time tick t, the state is $Y(s_i,t) = Y_{i,t} \perp (Y_{i,t-1} \cup \{Y_{j,t}\}_{j\sim N_i})$, where $i = 1,\ldots,n$ and $t = 1,2,\ldots$. To close the gap, we will extend the model to include temporal correlations.

3 Spatial-Temporal Structured Model

We generalize the previous framework to include an additional temporal component. With the additional regression term, the new framework is capable of modeling: information carried by observations, spatial dependence at fixed time tick, and temporal correlations of the hidden states.

Consider a discrete valued spatial-temporal process $\{Y_{i,t} : i = 1,\ldots,n, t = 1,2,\ldots\}$, where $Y_{i,t} = Y(s_i,t) \in (1,\ldots,C)$ corresponds to the i-th site s_i at the time tick t; $i = 1,\ldots,n$ and $t = 1,2,\ldots$. For a given time tick t, let $\mathbf{Y}_t = (Y_{1,t},\ldots,Y_{n,t})'$ denote the discrete valued hidden states on a graph structure $\{(s_i),(s_i \times s_j)\}|_{i,j=1}^{n}$. We model $\{\mathbf{Y}_t : t = 1,2,\ldots\}$ by a n-dimensional Markov chain with the following transition probability:

$$p_\theta(\mathbf{Y}_t|\mathbf{Y}_{t-1}) = q(\mathbf{Y}_t|\mathbf{Y}_{t-1})/G_{\mathbf{X}}. \tag{4}$$

Here $G_{\mathbf{X}}$ is a normalization constant and,

$$q(\mathbf{Y}_t|\mathbf{Y}_{t-1}) = \exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{p}\theta_k^1\varphi^1(X_{k,i,t},Y_{i,t}) + \sum_{j\sim i}\theta_{ij,t}^2\varphi^2(Y_{i,t},Y_{j,t}) + \sum_{i=1}^{n}\theta_{it,t-1}^3\varphi^3(Y_{i,t},Y_{i,t-1})\right\}, \tag{5}$$

where $X_{k,i,t} = X_k(s_i,t)$ denotes the k-th explanatory variable at the site s_i and the time tick t; θ_k is the linear regression coefficients corresponding to explanatory feature $\varphi^1(X_{k,i,t},Y_{i,t}), k = 1,\ldots,p$; $\theta_{ij,t}^2$ represent the spatial regression coefficients. The difference between the Equation 5 and the Equation 2 is the additional parameters $\theta_{it,t-1}^3$ that represent the temporal coefficients. When $\theta_{it,t-1} = 0$, there is no correlation over time and the Markov network of $\{\mathbf{Y}_t\}$ reduces to a sequence of independent random vectors, each represents a set of spatially dependent observations at a given time tick. Clearly, the new framework incorporates the previous one described in Section 2 as $p_\theta(\mathbf{Y}_t|\mathbf{Y}_{t-1})$ reduces to $p_\theta(\mathbf{Y}_t)$.

On the other hand, when $\theta_{it,t-1} \neq 0$, the framework considers state correlations over time; the magnitude of $\theta_{it,t-1}$ is related to the mean difference between

two consecutive time ticks of the same site. To simplify the representations, we abbreviate the model parameters by $\theta = \left(\{\theta^1\}, \{\theta^2\}, \{\theta^3\}\right)'$; model features by $\psi(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Y}_{t-1}) = \left(\{\varphi^1(X_{k,i,t}, Y_{i,t})\}, \{\varphi^2(Y_{i,t}, Y_{j,t})\}, \{\varphi^3(Y_{i,t}, Y_{i,t-1})\}\right)'$; and observations from T time points by $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$, where $\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})'$, $t = 1, \ldots, T$.

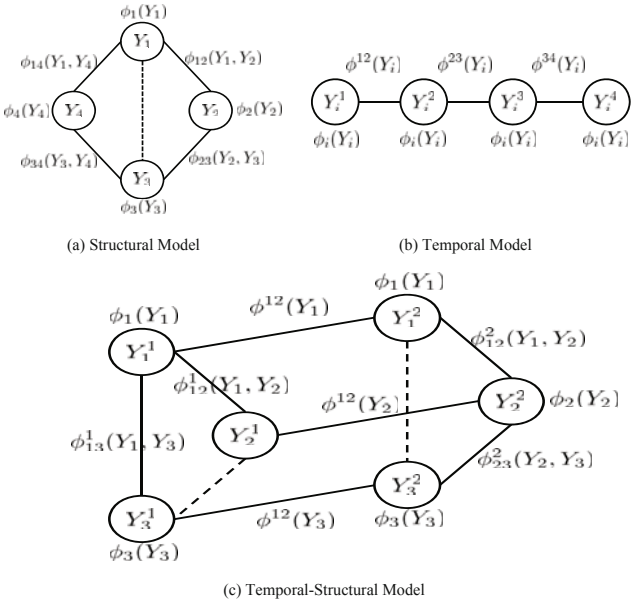

Fig. 2. (a) Typical spatial dependent model - first order Markov network: $\phi_i(Y_i) = \exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{p}\theta_k\varphi(X_{k,i}, Y_i)\right\}$ correspond to node potentials, $\phi_{i,i+1}(Y_i, Y_{i+1}) = \exp\left\{\sum_{j\sim i}\theta_{ij}\varphi(Y_i, Y_j)\right\}$ correspond to spatial edge potentials. (b) Typical temporal correlated model - first order Markov chain: $\phi_i(Y_i) = \exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{p}\theta_k\varphi(X_{k,i}, Y_i)\right\}$ correspond to node potentials, $\phi^{t-1,t}(Y_i) = \exp\left\{\sum_{i=1}^{n}\theta^3_{it,t-1}\varphi^3(Y_{i,t}, Y_{i,t-1})\right\}$ correspond to temporal edge potentials. (c) We propose a new framework that generalizes both spatial dependent models and temporal correlated models. For illustration purpose, we only show correlated states of two consequent time ticks but the framework indeed depicts a gigantic network over time. Thus, traditional approaches such as CRFs and M3N fail to solve it with tractable computation.

The equation 5 represents a general framework considering spatial-temporal correlations, which generalizes both temporal correlated models and spatial dependent models, as indicated by Figure 2. However, there are more states to be considered together in the new framework due to the spatial and temporal coupling. Thus, traditional solutions such as constructing a gigantic CRFs network would be computationally intractable.

4 Temporal Maximum Margin Markov Network

There are two typical tasks in a machine learning problem like Equation 5: learning and predicting. For learning, we want to estimate parameters θ so that

$$h_\theta(\mathbf{X}_t) = \text{argmax}_{\mathbf{Y}} \theta' \psi(\mathbf{X}_t, \mathbf{Y}, \hat{\mathbf{Y}}_{t-1}) \approx \hat{\mathbf{Y}}_t, \forall t, \tag{6}$$

where $\hat{\mathbf{Y}}_t$ is the ground-truth states. For predicting, we would like to infer the most likely states

$$\mathbf{Y}^*_{t+1} = \text{argmax}_{\mathbf{Y}} \theta' \psi(\mathbf{X}_{t+1}, \mathbf{Y}, \hat{\mathbf{Y}}_t), \tag{7}$$

given the parameter θ and the novel observation $\mathbf{X}_{t+1}$ and past states $\hat{\mathbf{Y}}_t$. We will now describe a convex instantiation of the spatial-temporal correlated framework to handle both tasks.

First, we need to measure the error of the approximation $h(\cdot)$ using a loss function ℓ. Here we use a Hamming distance error measurement $\ell_t(\mathbf{Y}_t)$ to indicate the number of variables predicted incorrectlys, which essentially measure the loss on the label sequences,

$$\ell_t(\mathbf{Y}_t) = \sum_i \Delta(Y_{i,t}, \hat{Y}_{i,t}) \text{ and } \Delta(Y_{i,t}, \hat{Y}_{i,t}) = \begin{cases} 1 & Y_{i,t} \neq \hat{Y}_{i,t} \\ 0 & Y_{i,t} = \hat{Y}_{i,t} \end{cases}.$$

We adapt the hinge upper bound $\bar{\ell}(h_\theta(\mathbf{X}_t))$ on the loss function for structured classification inspired by max-margin criterion:

$$\bar{\ell}_t(h_\theta(\mathbf{X}_t)) = \max_{\mathbf{Y}_t} \Big[\theta' \psi(\mathbf{X}_t, \mathbf{Y}_t, \hat{\mathbf{Y}}_{t-1})) + \ell_t(\mathbf{Y}_t)\Big] - \theta' \psi(\mathbf{X}_t, \hat{\mathbf{Y}}_t, \hat{\mathbf{Y}}_{t-1})) \tag{8}$$

$$\geq \ell_t(h_\theta(\mathbf{X}_t)), \tag{9}$$

where $\bar{\ell}_t(h_\theta(\mathbf{X}_t)) = \bar{\ell}(h_\theta(\mathbf{X}_t), \hat{\mathbf{Y}}_t)$ and $\ell_t(h_\theta(\mathbf{X}_t)) = \ell(h_\theta(\mathbf{X}_t), \hat{\mathbf{Y}}_t)$. With this upper bound, the min-max formulation for structured classification problem is analogous to SVM,

$$\min_{\theta, \mathbf{Y}_t} \frac{\lambda}{2} ||\theta||^2 + \frac{1}{T} \sum_{t=1}^{T} \xi_t \tag{10}$$

$$s.t. \left\langle \theta, \Phi(\mathbf{X}_t, \mathbf{Y}_t, \hat{\mathbf{Y}}_{t-1}, \hat{\mathbf{Y}}_t) \right\rangle \geq \bar{\ell}(\mathbf{Y}_t, \hat{\mathbf{Y}}_t) - \xi_t, \forall t, \forall \mathbf{Y}_t, \tag{11}$$

where $\Phi(\mathbf{X}_t, \mathbf{Y}_t, \hat{\mathbf{Y}}_{t-1}, \hat{\mathbf{Y}}_t) = \psi(\mathbf{X}_t, \hat{\mathbf{Y}}_t, \hat{\mathbf{Y}}_{t-1}) - \psi(\mathbf{X}_t, \mathbf{Y}_t, \hat{\mathbf{Y}}_{t-1})$. This formulation incorporates the "maximum margin" criteria. We can interpret

$$\text{M} = \frac{1}{||\theta||} \left\langle \theta, \Phi(\mathbf{X}_t, \mathbf{Y}_t, \hat{\mathbf{Y}}_{t-1}, \hat{\mathbf{Y}}_t) \right\rangle \tag{12}$$

as the margin of the state configuration $\hat{\mathbf{Y}}_t$ over another state configuration $\mathbf{Y}_t$. Assuming ξ_i are all zeros (because λ is very small), the constraints enforce,

$$\theta' \Big(\psi(\mathbf{X}_t, \hat{\mathbf{Y}}_t, \hat{\mathbf{Y}}_{t-1}) - \psi(\mathbf{X}_t, \mathbf{Y}_t, \hat{\mathbf{Y}}_{t-1}) \Big) \geq \bar{\ell}(\mathbf{Y}_t, \hat{\mathbf{Y}}_t), \tag{13}$$

so minimizing $||\theta||^2$ essentially maximizes the smallest of such margins, scaled by the loss $\ell_i(\mathbf{Y}_t, \hat{\mathbf{Y}}_t)$. The above formulation is a standard QP and can be solved use optimization packages, but it is exponential in the size and computation is generally prohibitive. Another way to express this problem is the following representation,

$$\min_{\theta, \mathbf{Y}_t} \frac{\lambda}{2}||\theta||^2 + \frac{1}{T}\sum_{t=1}^{T}\xi_t \tag{14}$$

$$s.t. \theta'\psi(\mathbf{X}_t, \hat{\mathbf{Y}}_t, \hat{\mathbf{Y}}_{t-1}) + \xi_t \geq \max_{\mathbf{Y}_t}\left[\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \hat{\mathbf{Y}}_{t-1})) + \ell_t(\mathbf{Y}_t)\right], \forall t, \tag{15}$$

which is a convex quadratic program in θ, since

$$\max_{\mathbf{Y}_t}\left[\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \hat{\mathbf{Y}}_{t-1})) + \ell_t(\mathbf{Y}_t)\right], \tag{16}$$

is convex in θ. It might be easier to interpret Equation 14 in its alternative representation Equation 17 by eliminating the constraints,

$$\min_{\theta, \mathbf{Y}_t} \frac{\lambda}{2}||\theta||^2 + \frac{1}{T}\sum_{i=1}^{T}\left\{\max_{\mathbf{Y}_t}\left[\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \hat{\mathbf{Y}}_{t-1})) + \ell_t(\mathbf{Y}_t)\right] - \theta'\psi(\mathbf{X}_t, \hat{\mathbf{Y}}_t, \hat{\mathbf{Y}}_{t-1})\right\}, \tag{17}$$

careful readers might notice that $\theta'\psi(\mathbf{X}_t, \hat{\mathbf{Y}}_t, \hat{\mathbf{Y}}_{t-1})$ is invariant to $\mathbf{Y}_t$ and we can run the algorithm in two separate steps: first, fix θ and optimize $\max_{\mathbf{Y}_t}\left[\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \hat{\mathbf{Y}}_{t-1})) + \ell_t(\mathbf{Y}_t)\right]$; second, fix $\mathbf{Y}_t$ obtained in the first step to calculate θ that minimize Equation 17. The procedure is similar to the Expectation-Maximization algorithm and we are guaranteed not to increase the objective function at each step.

5 Learning

Recall the objective in Equation 17 is a convex function, an intuitive way to estimate its parameters θ is to use a gradient descent approach. In this case, the gradients only depends on the most violated state configuration,

$$\mathbf{Y}_t^* = \text{argmax}_{\mathbf{Y_t}}\left(\theta'\psi(\mathbf{X}_t, \mathbf{Y}_t, \hat{\mathbf{Y}}_{t-1})) + \ell_t(\mathbf{Y}_t)\right), \tag{18}$$

which can be computed as:

$$g(\theta) = \lambda\theta + \frac{1}{T}\sum_{i=1}^{T}\left(\psi(\mathbf{X}_t, \mathbf{Y}_t^*, \hat{\mathbf{Y}}_{t-1}) - \psi(\mathbf{X}_t, \hat{\mathbf{Y}}_t, \hat{\mathbf{Y}}_{t-1})\right), \tag{19}$$

the following algorithm thus summarizes the procedure of gradient optimization,

Algorithm 1. Subgraident Optimization

Input: training data $\mathcal{D} = \{(\mathbf{X}_t, \mathbf{Y}_t)\}|_{t=1}^{T}$, regularization parameter λ, step size σ, tolerance ϵ, number of iterations T
Output: parameter vector θ

```
Initialize θ ← 0, t ← 1
repeat
    for t = 1 to T do
        Set violation function
        H(Y_t) = θ'ψ(X_t, Y_t, Ŷ_{t−1})) + ℓ_t(Y_t) − θ'ψ(X_t, Ŷ_t, Ŷ_{t−1})
        Find most violated label for (X_t, Y_t) : Y*_t = arg max_{Y_t} H(Y_t)
    end for
    Compute g(θ), update θ^(t) ← θ^(t−1) − σg(θ).
    Update t ← t + 1
until t ≥ T or MSE(||θ^(t)|| − ||θ^(t−1)||) ≤ ε
```

A critical part of Algorithm 1 is to compute the most violated constraint at each time step efficiently. The exact inference of this step is usually intractable as irregular lattices often involve loops that cannot be handled by deterministic algorithms in polynomial time.To this end, we leverage on a well established approximation algorithm, loopy belief propagation (LBP) [8] to solve this. To use LBP, we define the following potentials:

– **Unary potentials** represent the impact of local observation in $\mathbf{X}_t$ to the states $\mathbf{Y}_t$, this potential function at each site s_i takes the form,

$$\exp\left(\sum_{k=1}^{p} \theta_k^1 \varphi^1(X_{k,i,t}, Y_{i,t}) + \ell_t(Y_{i,t})\right), \forall i, \tag{20}$$

– **Environmental potentials** represent the influence between states and over time, these potential functions take the form,

$$\exp\left(\theta_{ij,t}^2 \varphi^2(Y_{i,t}, Y_{j,t})\right), \forall i, j \sim i, \text{ Structural Potential}, \tag{21}$$

$$\exp\left(\theta_{it,t-1}^3 \varphi^3(Y_{i,t}, Y_{i,t-1})\right), \forall i, \quad \text{Temporal Potential}. \tag{22}$$

6 Predicting

Now we will introduce our linear integer programming interface for predicting. The goal is to predict a hidden state as the most likely configuration:

$$\mathbf{Y}_{T+1}^* = \text{argmax}_{\mathbf{Y_T}} \left(\theta' \psi(\mathbf{X}_{T+1}, \mathbf{Y}_T, \hat{\mathbf{Y}}_T)\right). \tag{23}$$

Denote $Z^t = (\{z_i^t\}|_{i=1}^n, \{z_{ij}^t\}|_{i=1}^{n,j\sim i}, \{z_i^{t,t-1}\}|_{i=1}^n)$ as indicator variables at time t so that: $z_i(m) = 1$ indicates i-th site takes state m, $z_{ij}(m,n)$ indicates i and j-th sites take states m and n, and $z_i^{t,t-1}(m,n) = 1$ indicates i-th site take

states m and n at time t and $t-1$, respectively. If we factorize Equation 23, the following linear integer programming interface defines an exact mapping,

$$\max_{Z^t} \sum_{i,m} z_i^t(m) \left[\theta^1_{(\cdot)} \varphi^1(X_{(\cdot),i,t}, m)\right] + \sum_{i,j,m,n} z_{ij}^t(m,n) \left[\theta^2_{ij,t} \varphi^2(m,n)\right] + \sum_i \left[\theta^3_{it,t-1} \varphi^3(m, \hat{Y}_{i,t-1})\right], \tag{24}$$

$$s.t.\ z_i^t(m) \in \{0,1\},\ z_{ij}^t(m,n) \in \{0,1\}, \tag{25}$$

$$\sum_m z_i^t(m) = 1, \tag{26}$$

$$\sum_n z_{ij}^t(m,n) = z_i^t(m), \tag{27}$$

$$\sum_m z_{ij}^t(m,n) = z_j^t(n). \tag{28}$$

The constraint Equation 26 enforces only one state is allocated for each site s_i; the constraint equation 27 enforces the structural consistency. Note we assign $z_i^{t,t-1}(m, \hat{Y}_{i,t-1}) = 1, \forall i$ so that $Y_{i,t}$ is influenced by its previous state $\hat{Y}_{i,t-1}$of the same site s_i. The above linear integer programming is an intractable combinatorial problem but we can obtain an approximated solution by relaxing the binary constraint in Equation 25 to be $z_i^t(m) \geq 0,\ z_{ij}^t(m,n) \geq 0$. A threshold χ, usually equals to 0.5, is used to discretize the final outputs Z^t for predicting the states.

7 Experiments

7.1 Simulation Results

We use the following temporal-spatial correlated Linear Dynamic System (LDS) to generate the simulation. This system specifies the hidden state Y_t^i, which depends temporally on the previous state Y_{t-1}^i and correlates spatially with the states of the neighboring sites Y_t^j, $j \in N_{-i}$.

$$Y_t^i = \alpha Y_{t-1}^i + (1-\alpha) \sum_{j \in N_{-i}} \beta^j Y_t^j + e_1,\ , e_1 \sim N(0, \sigma^2_{e_1}), \tag{29}$$

$$X_t^i = A Y_t^i + e_2,\ , e_2 \sim N(0, \sigma^2_{e_2}), \tag{30}$$

where N_{-i} corresponds to the neighboring sites of i; A is a projection vector that maps hidden states to the observations; X_t^i corresponds to the observations at site i, time tick t; e_1 and e_2 are the environmental Gaussian noises;α represents the temporal/spatial trade-off parameter. If α is set to be zero, the simulation considers no time dependence. Otherwise, if α is set to be one, the simulation ignores spatial correlations.

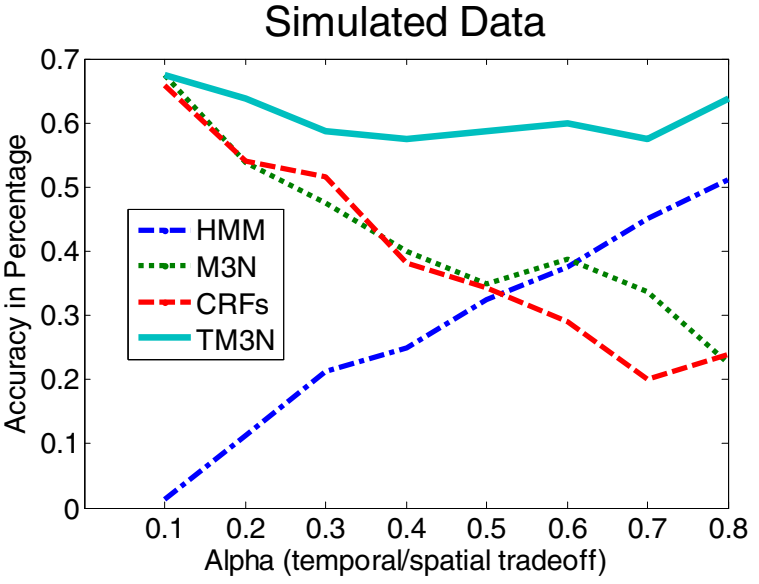

Fig. 3. Model comparison on simulated temporal-spatial correlated data. The X axis corresponds to the Alpha (temporal/spatial trade-off parameter) value and Y axis represents the accuracy in percentage. HMM's performance increases as the temporal influence becomes larger while CRFs/M3N's accuracy decreases at the meantime. TM3N outperforms all these three models and demonstrates its efficacy.

We initialize $Y_t^i \sim \text{Uniform}(0,1)$, $\beta^j \sim \text{Uniform}(0,1)$; set total sites number N equals to four; specify the error term $e_1 \sim N(0, 0.05)$ and $e_l \sim N(0,1)$; and let the projection matrix $A = [10; 20]$. To simulate the hidden states, we use an approach similar to Gibbs sampling that iteratively samples Y_t^i until the system converges. These simulated states are rounded to be real valued states and the simulated observations are calculated use Equation 30.

In the experiment, we vary the temporal/spatial trade-off parameter α from 0.1 to 0.8 at an interval of 0.1 to evaluate the performances of four different models: HMM, M3N, CRF and TM3N. For every α value, we run the simulation for 50 times and calculate the averaged accuracy. The results are demonstrated in Figure 3, where the blue curve corresponds to the accuracy of TM3N model at various α values. Obviously, TM3N shows superior performance comparing to HMM, CRF and M3N.

7.2 Real World Applications

Recently, buildings began to have sensor networks installed for energy and comfort management. The control strategy for lighting, heating ventilation and air-conditioning (HVAC) can be updated adaptively as needed [2]. For the cost-efficient operation, understanding occupancy behavior in buildings is becoming

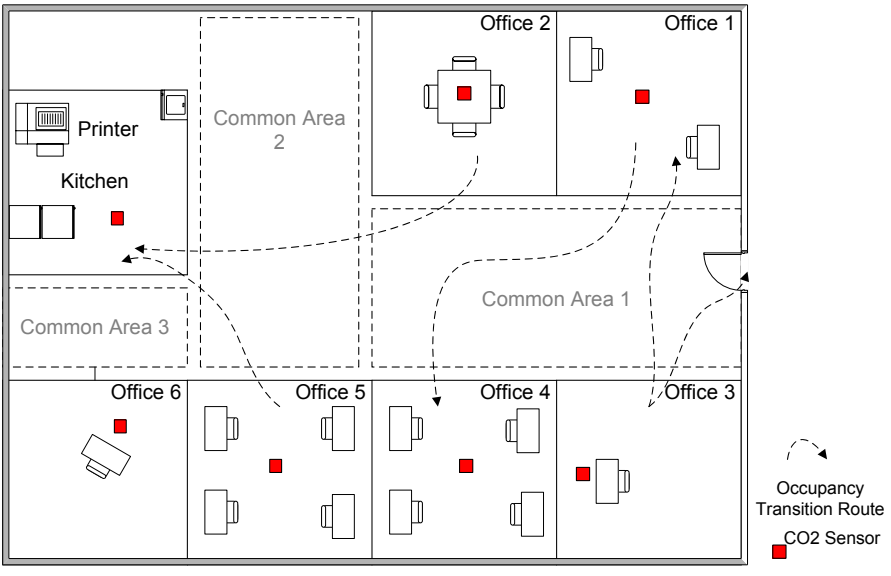

Fig. 4. Geometric flat view of the office area testbed

crucial problems to success. One specific question is to estimate office occupant number over the time, which naturally fits our proposed model.

7.2.1 Data Collection

The sensor network is setup in an open plan office space with six rooms and one kitchen/printer room. It provides offices for two faculty members and ten Ph.D. students. Since it is an open plan office, the faculties and students have discussions frequently. The entire indoor environment can be considered very dynamic. Occupants have different activities such as reading, talking on the phone, drop-by and discussion.An occupant may leave his own area and go to other areas, such as printer room, kitchen, and restroom. The physical sensor network includes a wired CO_2 network and a data server. One CO_2 sensor is installed in the center of each office at the nose level (1.1m) above the ground. To establish ground truth about occupancy information, we use a network of commercial cameras. Figure 4 shows the geometric view of the test-bed. Note the CO_2 sensors are preferred over the vision sensors because of the privacy reasons, e.g., we cannot easily distinguish the occupants by the CO_2 measurements.

Data collection for this paper was for one continuous period, with a sampling rate of every one minutes, capturing CO_2 measurements and the number of occupants in four offices. The time period is three weeks from March 17th, 2008 to April 4th, 2008 excluding weekends. Occupancy data was recorded from 8:00am to 8:00pm from the four offices (2, 3, 4 and 5). Office 2 and 5 have four Ph.D. students; office 4 has two graduate students ; and office office 3 have 1 faculty. We synchronize the measurements from all sensors; and aggregate measurements

Table 2. Comparison results of the average accuracy in the building occupancy estimation task

Algorithm	Accuracy
HMM	36.5%
CRFs	49.81%
M3N	49.05%
TM3N	**69.76**%

for every 10 minutes to predict the averaged occupant numbers in a ten minute window.

7.2.1 Results and Discussions

We split the data into training and testing: the first week from 3/17 to 3/21 is used as training data; the second and third week from 3/24 to 4/4 are used as testing. Along with our proposed model, we implement and test CRFs, M3N and HMM models.

Table 2 summarizes the comparison results of the four different methods. If we consider only temporal correlations and assume structural independence, HMM model gives an average accuracy of 36.5% for four offices. Clearly, first order Markov model is not suffice to capture the dynamics in the environment. On the other hand, structural models such as CRFs and M3N shows similar results, although slightly better than HMM, are still unsatisfactory. A significant improvement in average accuracy is observed when we combine both temporal and structural influence into a unified model, TM3N. This improvement owns a big part of its success to the joint modeling of both temporal and structural information, which oftentimes complement to each other.

To make the figures uncluttered, we only show the prediction results of proposed TM3N model on 3/25/2008. Figure 5 illustrates the results in four offices from 8am to 8pm. The solid blue curve corresponds to the prediction results of TM3N approach and the dashed magenta curve corresponds to the ground-truth value. The X axis show the time tick for every 50 minutes from 8am to 8pm. The Y axis shows the number of occupants. The estimation accuracies are 0.76, 0.86, 0.74 and 0.67 for office 2, 3, 4 and 5, respectively. As indicated by these figures, the TM3N occupant estimation results are close to ground-truth, which shows our method captures the occupancy dynamics quickly.

For the faculty office 3, the occupancy number is usually between zero and one during the day. For student's offices, there are much faster changes of occupant numbers, for example, the occupant number in office 2 and office 5 changes hourly. These changes are due to the "drop-by" activities of visitors.

In some of those cases, the estimation does not capture it well as the length of our estimation window is ten minutes while the visitors stay for a few minutes. However, such abrupt changes usually will not affect the operation of building energy management systems because these systems such as HVAC cannot response in high frequencies. Hence, in the practical application, this abrupt change will be ignored.

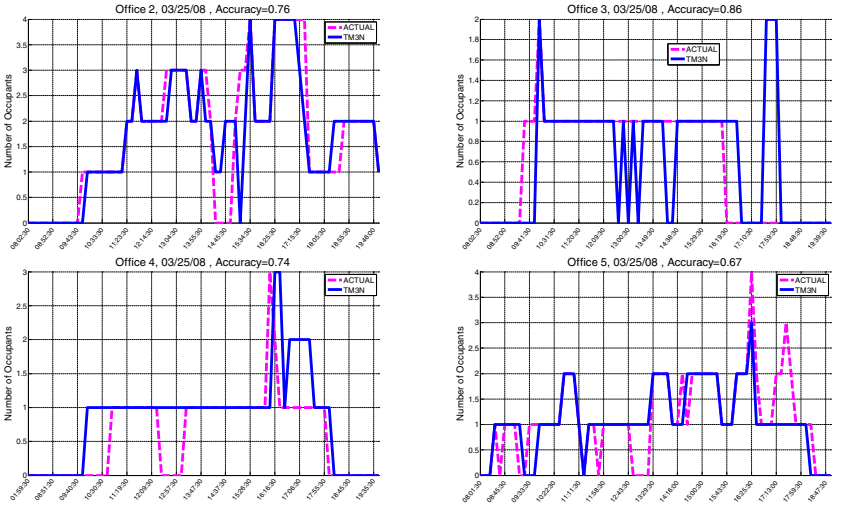

Fig. 5. Occupancy Estimation of 4 offices on March 25. The dashed magenta curve corresponds to the ground-truth; the solid blue curve corresponds to the estimation results of TM3N. The states are discrete valued variables, e.g. 0, 1, 2, 3 and 4. We connect these discrete states for visualization purpose; thus, a sharp jump mismatch does not mean a large deviation from the ground-truth.

8 Conclusion

This paper presents a maximum margin structured learning model, TM3N to model temporal-spatial correlated environments. The main goal of this work is to synthesize information from different perspectives to model real world systems more faithfully. We demonstrate how the proposed framework incorporates, generalizes, and extends existing approaches presented in the literature. The experiments show superior performance of proposed model against other state of the arts approaches in the building occupancy estimation task.

References

[1] Desai, C., Ramanan, D., Fowlkes., C.: Discriminative models for multi-class object layout. In: ICCV (2009)

[2] Dong, B., Andrews, B., Lam, K., Hoynck, M., Chiou, Y., Zhang, R., Benitez, D.: An Information Technology Enabled Sustainability Test-Bed (ITEST) For Occupancy Detection Through An Environmental Sensing Network. Energy and Buildings. Energy and Buildings 42(7) (2010)

[3] Fildes, R.: Forecasting structural time series models and the kalman filter: Andrew harvey, p. 554. Cambridge University Press, Cambridge (1989); International Journal of Forecasting 8(4), 635–635 (December 1992), http://ideas.repec.org/a/eee/intfor/v8y1992i4p635-635.html, ISBN: 0-521-32196-4

[4] Gould, S., Fulton, R., Koller, D.: Decomposing a scene into geometric and semantically consistent regions. In: ICCV (2009)
[5] Kalman, R.E.: A new approach to linear filtering and prediction problems. Trans. Am. Soc. Mech. Eng., Series D. Journal of Basic Engineering 82 (1960)
[6] Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, pp. 282–289 (2001)
[7] Li, R., Tian, T.P., Sclaroff, S.: Simultaneous Learning of Nonlinear Manifold and Dynamical Models for High-dimensional Time Series. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8 (October 2007), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4409044
[8] Murphy, K.P., Weiss, Y., Jordan, M.I.: Loopy belief propagation for approximate inference: An empirical study. In: Proceedings of Uncertainty in AI, pp. 467–475 (1999)
[9] Oh, S.M., Ranganathan, A., Rehg, J.M., Dellaert, F.: A Variational inference method for Switching Linear Dynamic Systems Need for Variational methods Variational method. Computing (2005)
[10] Qian, X., Jiang, X., Zhang, Q., Huang, X., Wu, L.: Sparse higher order conditional random fields for improved sequence labeling. In: ICML, p. 107 (2009)
[11] Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 257–286 (1989)
[12] Sarawagi, S., Gupta, R.: Accurate max-margin training for structured output spaces. In: ICML 2008: Proceedings of the 25th International Conference on Machine Learning, pp. 888–895. ACM, New York (2008)
[13] Taskar, B., Guestrin, C., Koller, D.: Max-margin markov networks. Advances in Neural Information Processing Systems 16, 25–32 (2003)
[14] Taskar, B.: Learning structured prediction models: A large margin approach. Stanford University, Ph.D. thesis (2004)
[15] Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res. 6, 1453–1484 (2005)
[16] Duong, T.V., Phung, D.Q., Bui, H.H., Venkatesh, S.: Behavior recognition with generic exponential family duration modeling in the hidden semi-markov model. In: Proceedings of the 18th International Conference on Pattern Recognition, vol. 3 (2006)
[17] Welch, L.R.: Hidden markov models and the baum-welch algorithm. IEEE Information Theory Society Newsletter 53(4) (December 2003), http://www.itsoc.org/publications/nltr/it_dec_03final.pdf