Yale Song | yalesong@mit.edu

Multi-View Hidden Conditional Random Fields

Yale Song, Louis-Philippe Morency, and Randall Davis

Massachusetts Institute of Technology

CSAIL

Abstract

Human Behavior Understanding

Speech
Facial expression
Gesture
…

Stock Market Analysis

Inflation
Company earnings
World news
Interest rates
…

Multi-view dynamic learning deals with sequential data, where the views are generated by multiple interacting time-series processes that encode different sources of information. We present a new family of models, called Multi-View HCRF, which generalize traditional HCRF to explicitly learn the interaction between multiple views. Knowledge about the underlying structure of the data is formulated as a chain-structured latent model, learning the interaction between views using disjoint sets of hidden variables. This significantly reduces the model complexity as compared to traditional HCRF (early fusion), from $O(D^C)$ to $O(DC)$, with C views and D hidden variables per view. Experimental results show that our model can capture the hidden interaction between views with less training data, using fewer model parameters, making the sequence modeling task more accurate and efficient.

Multi-View Sequences

- May have distinctive dynamics, e.g., different distributions, noise, variance, and/or frame rates.
- May interact with each other in both time and space, either synchronously or asynchronously.
- May not be conditionally independent given the class label .

Early Fusion Latent Models

Early fusion with a latent model, e.g., HCRF, needs a set of latent variables that is the product set of the latent variables from each original view. This increase in complexity is exponential. Early fusion thus requires much more data to estimate the underlying distributions correctly, which makes this solution impractical for many real world applications.

Multi-View HCRF

The Model

MV-HCRF is a conditional probability distribution that factorizes according to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_W, \mathcal{E}_B)$, with a set of vertices $\mathcal{V}$, *within-view* edges $\mathcal{E}_W$, and *between-view* edges $\mathcal{E}_B$.

$$p(y \mid \hat{\mathbf{x}}; \Lambda) = \sum_{\hat{\mathbf{h}}} p(y, \hat{\mathbf{h}} \mid \hat{\mathbf{x}}; \Lambda) = \frac{1}{Z} e^{\Phi(y, \hat{\mathbf{h}}, \hat{\mathbf{x}}; \Lambda)}$$

$$e^{\Phi(\cdot)} = \sum_{k,(s,c)\in\mathcal{V}} \lambda_k f_k\left(y, h_s^{(c)}, \mathbf{x}^{(c)}\right) + \sum_{k,(s,t,c,d)\in\mathcal{E}} \omega_k g_k\left(y, h_s^{(c)}, h_t^{(d)}, \hat{\mathbf{x}}\right)$$

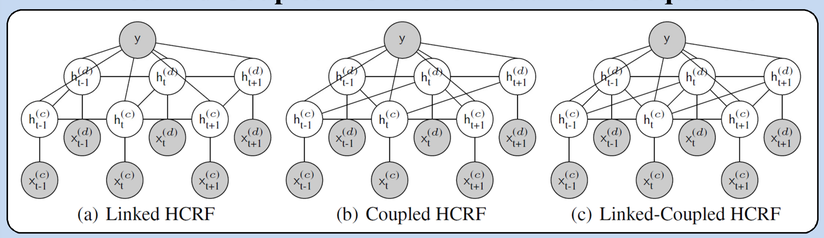
Feature Functions

- $f_k\left(y, h_s^{(c)}\right)$: *label* feature function. $\sum_C |\mathcal{Y}| \times \left|\mathcal{H}^{(c)}\right|$
- $f_k\left(h_s^{(c)}, \mathbf{x}^{(c)}\right)$: *observation* feature function. $\sum_C \left|\mathcal{H}^{(c)}\right| \times \left|\phi\left(\mathbf{x}^{(c)}\right)\right|$
- $g_k(\cdot)$: depends on the definitions of $\mathcal{E}_W$ and $\mathcal{E}_B$

Parameter Estimation and Inference

- Given a training dataset $\mathcal{D} = \{y_i, \hat{\mathbf{x}}_i\}$, we find $\Lambda^* = \{\lambda^*, \omega^*\}$ by optimizing $L(\Lambda) = \sum_{i=1}^{N} \log p(y_i \mid \hat{\mathbf{x}}_i; \Lambda)$ with L-BFGS.
- Adding hidden variables makes optimization non-convex. We find Λ^* by initializing from multiple random starting points and searching for the best local maximum.
- We use loopy BP with a random message update schedule.
- Classification rule: $y^* = \arg\max_{y\in\mathcal{Y}} p(y \mid \hat{\mathbf{x}}; \Lambda^*)$.

Linked and Coupled HCRFs for Dual-View Sequences



(a) Linked HCRF (b) Coupled HCRF (c) Linked-Coupled HCRF

$$g_k\left(y, h_s^{(c)}, h_t^{(d)}\right) = \mathbf{1} \Leftrightarrow \begin{cases} (s+1 = t \wedge c = d) \vee (s = t \wedge c \neq d) & \text{(LHCRF)} \\ (s+1 = t) & \text{(CHCRF)} \\ (s+1 = t) \vee (s = t \wedge c \neq d) & \text{(LCHCRF)} \end{cases}$$

Future Work

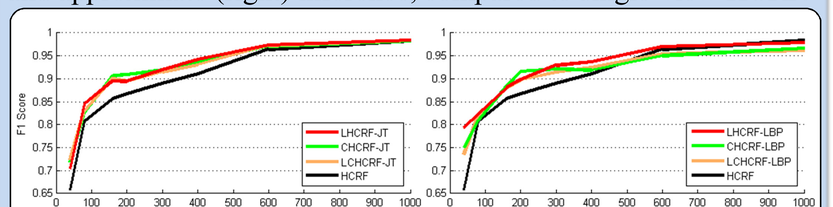
We plan to work with unsegmented continuous sequences using CRF extensions, e.g., LDCRF. We also plan to explore structural learning in multi-view models, i.e., learning interaction patterns.

Experiments

Synthetic Data

Data: Two first-order Markov chains were coupled as LCHCRF; data was generated using Gibb's sampler. To simulate strong interaction between views, we set the weights on $\mathcal{E}_B$ as [-10,10], while on $\mathcal{E}_W$ they were [-1,1]. We set $|\mathcal{Y}| = 2$, $\left|\mathcal{H}^{(1)}\right| = \left|\mathcal{H}^{(2)}\right| = 8$.

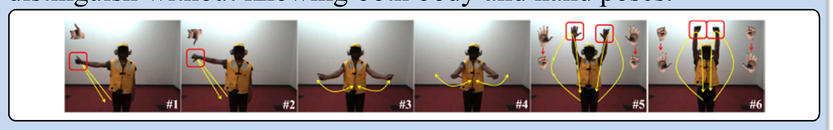
Result: The graphs below show F1 score obtained using exact (left) and approximate (right) inference, compared to single-view HCRF.



MV-HCRF outperformed HCRF, even with fewer training data. For the HCRF, $|\mathcal{H}|^*$ was 64, which was $\left|\mathcal{H}^{(1)}\right| \times \left|\mathcal{H}^{(2)}\right|$. This resulted in a sizable difference in the model complexity, with 8,960 parameters to estimate using HCRF. Comparatively, for LHCRF it was 496, for CHCRF 624, and for LCHCRF 752.

Body-and-Hand Gesture Data (NATOPS dataset[1])

Data: We used three pairs of NATOPS gestures that are difficult to distinguish without knowing both body and hand poses.



[1] Y. Song, D. Demirdjian, and R. Davis. Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database. In *FG*, pp. 500-506. 2011.

Result: MV-HCRF outperformed HCRF and HMM. Our results show that by automatically learning the hidden interaction between multiple views, MV-HCRF can efficiently differentiate all body-and-hand gestures with fewer parameters than single-view models.