The Infinite Factorial Hidden Markov Model
CS886 - Fall 2010
Paper Critique 2
Karim Ali

1 Overview

The authors start their paper by discussing the evolution of the Hidden Markov Model (HMM) to the Factorial Hidden Markov Model (FHMM). The main reason behind this extension is the HMM's limited representational power of latent variables in modeling discrete time series data. To overcome this, the FHMM factors each hidden state into multiple hidden (i.e. latent) state variables. In other words, FHMM formally defines M latent variables as M latent Markov chains that evolve according to Markov dynamics at each timestep t. Although this strengthens the representational power of the FHMM for latent variables, it introduces a new free parameter M. What the authors are after is **learning this parameter from data rather than specifying it beforehand** (Section 1, page 2, end of first paragraph).

The main contribution of the paper is presenting a non-parametric FHMM, Infinite FHMM (iFHMM), that is based on a new stochastic process for latent feature representation of time series called the Markov Indian Buffet Process (mIBP). The mIBP can be dervied in three steps (Section 2):

1. Describe a distribution over binary matrices with a finite number of columns.
2. Integrtate out the parameters of the model by careful selection of the hyperparameters.
3. Take the limit as the number of features (i.e. columns) goes to ∞.

The mIBP provides a matrix S which is interpreted as an arbitrarily large set of parallel Markov chains. S is then used as the building block of the iFHMM (Section 3).

2 Significance and Originality

This work is very recent (2009) so it does not have that many citations yet. However, one can see the significance of the work in the field of modeling discrete time series data (e.g. blind source separation) once one realizes that infinite features (or previously unknown number of features) can be modelled using the iFHMM.

The novelity of the methods presented by the authors stems from relieving the constraint of knowing M (number of features) beforehand in the FHMM and dealing with inifite (or uknown) number of features in the iFHMM. This has various applications in vision, audio processing (e.g. uknown number of speakers), and natural language processing.

3 Model Evaluation

Although the authors presented a very interesting idea (iFHMM), they did not give enough explanation or clarification for some key points in their processes. My only explanation for that is space limits. Here are some of thos key points that I think required more details:

- Section 2.1, page 3: how they got to equation 5 from the previous explanation.

- Section 2.2, page 3: the derivation for the size of *[S]*.
- Section 2.3, page 4: I did not get how equation 9 shows that the suggested model is exchangeable in the columns and Markov exchangeable in the rows. Going back to reference [6], I am even more confused. If the columns are exchangeable, then it should not matter where the position of a feature in the matrix is. However, this is not true as the probability that a customer orders a dish m (corresponds to a feature in the matrix) depends on whether the cutsomer in front of him ordered that dish or not.
- Section 3.1, page 5: the authors did not mention what other alternatives other than Laplace(0,1) could be used to IID sample the entries in matrix X. Using Laplace(0,1) was justified because it fits their data better. What about other types of data?
- Section 3.1, page 6: the authors did not provide further details about how their likelihood calculations satisfy the two technical conditions for proper iFHMM likelihoods.
- Section 3.2, page 6: no details about how to extend the representation of S, X, W were given (first step in the iterative sampling).

4 Experiment Evaluation

The experimental evaluation done by the authors is very poorly explained and not rigorous enough. My remarks on their experimental evaluation can be summarized as follows:

- Section 3.2, end of page 6: the authors threw some claims without backing them up by empirical evidence, or references to such evidence. For example, poor performance of naive Gibbs sampler; dynamic programming does not perform better than blocked Gibbs sampler; the bulk of computation is used for estimating X and W.
- Much more information was needed about the dataset used for the experiments. For example, which speakers were used as the dataset (there're 34 speakers in the original dataset); the randomness of the artificial pauses; how long the sentence were; was ther any background noise in the speech; how the input from different mics was merged together to form the speech of one speaker.

5 Related Work

7/14 references are self references.

6 Readability

7 Suggestions