On Conditional Random Fields:
Applications, Feature Selection, Parameter Estimation and Hierarchical Modelling

ABSTRACT

There has been a growing interest in stochastic modelling and learning with complex data, whose elements are structured and interdependent. One of the most successful methods to model data dependencies is *graphical models*, which is a combination of graph theory and probability theory. This thesis focuses on a special type of graphical models known as Conditional Random Fields (CRFs) [3], in which the output state spaces, when conditioned on some observational input data, are represented by *undirected* graphical models. The contributions of thesis involve both (a) broadening the current applicability of CRFs in the real world and (b) deepening the understanding of theoretical aspects of CRFs.

On the application side, we empirically investigate the applications of CRFs in two real world settings. The first application is on a novel domain of Vietnamese *accent restoration*, in which we need to restore accents of an accent-less Vietnamese sentence. Experiments on half a million sentences of news articles show that the CRF-based approach is highly accurate. In the second application, we develop a new CRF-based *movie recommendation* system called *Preference Network* (PN). The PN jointly integrates various sources of domain knowledge into a large and densely connected Markov network. We obtained competitive results against well-established methods in the recommendation field.

On the theory side, the thesis addresses three important theoretical issues of CRFs: *feature selection*, *parameter estimation* and *modelling recursive sequential data*. These issues are all addressed under a general setting of *partial supervision* in that training labels are not fully available.

For feature selection, we introduce a novel learning algorithm called *AdaBoost.CRF* that incrementally selects features out of a large feature pool as learning proceeds. AdaBoost.CRF is an extension of the standard boosting methodology to structured and partially observed data. We demonstrate that the AdaBoost.CRF is able to eliminate irrelevant features and as a result, returns a very compact feature set without significant loss of accuracy.

Parameter estimation of CRFs is generally intractable in arbitrary network structures. This thesis contributes to this area by proposing a learning method called *AdaBoost.MRF* (which stands for AdaBoosted Markov Random Forests). As learning proceeds AdaBoost.MRF incrementally builds a tree ensemble (a forest) that covers the original network by selecting the best spanning tree at a time. As a result, we can approximately learn many rich classes of CRFs in linear time.

The third theoretical work is on modelling *recursive, sequential* data in that each level of resolution is a Markov sequence, where each state in the sequence is also a Markov sequence at the finer grain. One of the key contributions of this thesis is *Hierarchical Conditional Random Fields* (HCRF), which is an extension to the currently popular sequential CRF and the recent semi-Markov CRF [4]. Unlike previous CRF work, the HCRF does not assume any fixed graphical structures. Rather, it treats structure as an uncertain aspect and it can estimate the structure automatically from the data. The HCRF is motivated by Hierarchical Hidden Markov Model (HHMM) [2]. Importantly, the thesis shows that the HHMM is a special case of HCRF with slight modification, and the semi-Markov CRF is essentially a flat version of the HCRF.

Central to our contribution in HCRF is a polynomial-time algorithm based on the Asymmetric Inside Outside (AIO) family developed in [1] for learning and inference. Another important contribution is to extend the AIO family to address learning with missing data and inference under partially observed labels. We also derive methods to deal with practical concerns associated with the AIO family, including numerical overflow and cubic-time complexity. Finally, we demonstrate good performance of HCRF against rivals on two applications: indoor video surveillance and noun-phrase chunking.

References

[1] H. H. Bui, D. Q. Phung, and S. Venkatesh. Hierarchical hidden Markov models with general state hierarchy. In D. L. McGuinness and G. Ferguson, editors, *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI)*, pages 324–329, San Jose, CA, Jul 2004.

[2] Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998.

[3] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine learning (ICML)*, pages 282–289, 2001.

[4] Sunita Sarawagi and William W. Cohen. Semi-Markov conditional random fields for information extraction. In Bottou L Saul LK, Weiss Y, editor, *Advances in Neural Information Processing Systems 17*, pages 1185–1192. MIT Press, Cambridge, Massachusetts, 2004.