ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Hybrid HMM and HCRF model for sequence classification

Y. Soullard and T. Artières

University Pierre and Marie Curie - LIP6
4 place Jussieu 75005 Paris - France

Abstract. We propose a hybrid model combining a generative model and a discriminative model for signal labelling and classification tasks, aiming at taking the best from each world. The idea is to focus the learning of the discriminative model on most likely state sequences as output by the generative model. This allows taking advantage of the usual increased accuracy of generative models on small training datasets and of discriminative models on large training datasets. We instantiate this framework with Hidden Markov Models and Hidden Conditional Random Fields. We validate our model on financial time series and on handwriting data.

1 Introduction

Sequence and signal labelling is a fundamental task in many application domains (text, speech, handwriting). A particular case we focus on in this paper concerns sequence classification where one wants to assign a single label to an input sequence. We consider signal-like data where an input sequence is classically represented as a sequence of real-valued feature vectors.

There are two main approaches for tackling the sequence labelling and the sequence classification tasks. On the one hand generative approaches rely on the use of generative models, one for each class, such as Hidden Markov Models (HMMs) and their variants [12]. HMMs are the reference technology for dealing with speech and handwriting. One of their main strength is the existence of efficient algorithms for training and recognition. Their main limitation lies in their training criterion (Maximum Likelihood) that does not focus on discrimination. Note that a number of attempts have been proposed for training HMMs in a discriminative way, by using discriminative criterion such as Maximum Mutual Information (MMI) [13] and margin-based criterion [5].

On the other hand discriminative models such as Hidden Conditional Random Fields (HCRFs) have been proposed recently for signal labelling and classification tasks [6] [11]. These models are extensions of Conditional Random Fields (CRFs) (that were initially proposed by [7] for text data) for dealing with hidden states as traditionally done in HMMs. One limit of such models lies in their optimization. The non convexity of the training criterion, due to the introduction of hidden states, makes training very sensitive to initialization. One efficient way to initialize HCRFs seems to learn first a HMM, then to initialize the HCRF parameters so that it reproduces the same classification as the HMM (under certain conditions a HMM may be transformed in an "equivalent" HCRF while the reciprocal is wrong) [6].

Besides a number of researchers have studied the difference between generative approaches and discriminative ones. One key observation pointed out in [10] is that as the number of training examples increases one could expect that generative classifier may initially perform better while discriminative classifier, converging to an asymptotic lower error, would likely overtake the performance of the generative classifier. Based on such ideas, various combination schemes have been proposed to optimally combine generative and discriminative approaches [1], [3], [8], [9].

In this paper we build on these ideas and we propose a deep hybridation scheme of HMM and HCRF to combine their respective strengths. The main motivation is to help preventing overfitting of a discriminative HCRF by introducing constraints based on a learned HMM. The main expected advantage is to get a discriminative model that is more efficient than HMMs and HCRFs for small training datasets. As we will show our experiments seem to validate this expectation and furthermore show that the hybrid model may also outperform HMMs and HCRFs even for large training datasets.

2 Related Works

Recently, hybrid approaches that combine generative and discriminative models were successfully proposed for improving classical models [2]. One straight idea is to maximize a convex combination of the generative and of the discriminative log-likelihoods as in [3]:

$$\alpha \, log \, p(\mathbf{c}|\mathbf{x}, \Theta) + (1-\alpha) \, log \, p(\mathbf{x}, \mathbf{c}|\Theta) \tag{1}$$

where Θ is the parameter set and $0 \leqslant \alpha \leqslant 1$ allows tuning the combination from the pure generative case ($\alpha = 0$) to the pure discriminative case ($\alpha = 1$).

In addition, Minka [9] proposed to consider discriminative learning of generative models (modeling the joint distribution $p(\mathbf{x}, c|\theta)$) as the learning of a model belonging to a new family of discriminative models modeling the conditional and joint distribution as:

$$q(\mathbf{x}, c|\Theta, \Lambda) = p(c|\mathbf{x}, \Lambda)p(\mathbf{x}|\Theta) \tag{2}$$

and

$$q(\mathbf{x}, c, \Theta, \Lambda) = p(c|\mathbf{x}, \Lambda)p(\mathbf{x}|\Theta)p(\Theta, \Lambda) \tag{3}$$

where Θ and Λ are parameter sets that have the same type but may be different or even independent, and where $p(c|\mathbf{x}, \Lambda) = \frac{p(\mathbf{x},c|\Lambda)}{\sum_c p(\mathbf{x},c|\Lambda)}$ and $p(\mathbf{x}|\Theta)$ is given by $\sum_c p(\mathbf{x}, c|\Theta)$. Learning is performed by maximizing the joint likelihood of the data D, $q(D, \Theta, \Lambda)$. This may correspond to a variety of learning schemes according to the assumption on Θ and Λ. For instance, Lasserre and Bishop [1], [8], following Minka's work, explored discriminative training schemes that combine in a principled way generative and discriminative approaches by using specific priors linking Θ and Λ in Eq. (2).

3 Proposed Method

We propose a new model that we formalize as an HCRF and which makes use of an existing HMM for guiding its learning. As we will show this model outperforms HCRF and HMM on our datasets.

3.1 HCRF

Hidden CRF (HCRF) have been proposed as an extension of CRFs for dealing with more complex and structured data [11]. In CRF-based systems there is usually one state per class (e.g. a POS tag) while there are few states corresponding to a given class in HRCF, alike in HMMs. We first recall basics of HCRFs.

Let $\mathbf{x} = (x_1, ..., x_T)$ be an observed sequence of length T and $\mathbf{s} = (s_1, ..., s_T)$ be a state sequence. Let c be the class of $\mathbf{x}$ and $S(c)$ the set of all possible hidden state sequences corresponding to the model of class c. Let note $\Phi(\mathbf{x}, c, \mathbf{s})$ the feature vector corresponding to a state sequence $\mathbf{s}$, which we assume to decompose as $\Phi(\mathbf{x}, c, \mathbf{s}) = \sum_t \phi(\mathbf{x}, c, s_t, t)$ with $\phi(\mathbf{x}, c, s_t, t)$ being a feature vector at time t in state s_t of model of class c. Then, the class conditional probability of a HCRF model is given by:

$$q(c|\mathbf{x}, \Lambda) = \frac{1}{Z(\mathbf{x}, \Lambda)} \sum_{\mathbf{s} \in S(c)} exp^{\lambda_{\mathbf{s}} \cdot \Phi(\mathbf{x}, c, \mathbf{s})} = \frac{1}{Z(\mathbf{x}, \Lambda)} \sum_{\mathbf{s} \in S(c)} exp^{\sum_t \lambda_{s_t} \cdot \phi(\mathbf{x}, c, s_t, t)} \quad (4)$$

where Λ is a parameter set, λ_s is the subset of Λ corresponding to a particular state s, and $Z(\mathbf{x}, \Lambda)$ is a normalization term. When given an input sequence $\mathbf{x}$, its class is determined according to $argmax_c q(c|\mathbf{x}, \Lambda)$.

3.2 Hybrid Modelling

Our aim is to make the learning of the discriminative model focus on segmentation which are likely according to a learned HMM. We propose to define a probabilistic conditional model based on a HMM (with parameters Θ) and on a HCRF (with parameters Λ) with identical structures (number of states and topology) as follows:

$$q(c|\mathbf{x}, \Lambda, \Theta) = \sum_{\mathbf{s} \in S(c)} p(\mathbf{s}|\mathbf{x}, c, \Theta) p(c, \mathbf{s}|\mathbf{x}, \Lambda) \quad (5)$$

In this formulation, assuming the generative model is already learned, one may see that the learning of the discriminative model will focus on likely state sequences. It is expected that such a soft constraint on the learning of the HCRF will help preventing overfitting. This formulation exhibits some similarity with Eq. (2). In case one uses a HCRF as discriminative model, one gets a hybrid model of the form:

$$q(c|\mathbf{x}, \Lambda, \Theta) = \frac{1}{Z(\mathbf{x}, \Lambda, \Theta)} \sum_{\mathbf{s} \in S(c)} p(\mathbf{s}|\mathbf{x}, c, \Theta) exp^{\lambda_{\mathbf{s}} \cdot \Phi(\mathbf{x}, c, \mathbf{s}, \Theta)} \tag{6}$$

where $Z(\mathbf{x}, \Lambda, \Theta)$ is a normalization term. Of course, according to usual HMM assumptions, $p(\mathbf{s}|\mathbf{x}, c, \Theta)$ may be written as:

$$p(\mathbf{s}|\mathbf{x}, c, \Theta) = \frac{p(\mathbf{x}, \mathbf{s}|c, \Theta)}{p(\mathbf{x}|c, \Theta)} = \frac{1}{p(\mathbf{x}|c, \Theta)} \prod_{t=1}^{T} p(x_t, s_t|s_{t-1}, \Theta) \tag{7}$$

Hence $p(\mathbf{s}|\mathbf{x}, c, \Theta)$ may be decomposed as:

$$p(\mathbf{s}|\mathbf{x}, c, \Theta) = exp^{log(p(\mathbf{s}|\mathbf{x}, c, \Theta))} = \frac{1}{p(\mathbf{x}|c, \Theta)} exp^{\sum_{t=1}^{T} log(p(x_t, s_t|s_{t-1}, \Theta))} \tag{8}$$

Then one may represent the distribution in Eq. (6) very similarly to a standard HCRF distribution by introducing an extended feature vector $\tilde{\phi}(\mathbf{x}, c, s_t, t, \Theta) = [\phi(\mathbf{x}, c, s_t, t), log(p(x_t, s_t|s_{t-1}, c, \Theta))]$ and by introducing an extended parameter vector with an additional and constant parameter $\tilde{\lambda}_{s_t} = [\lambda_{s_t} 1]$. Then Eq. (6) rewrites:

$$q(c|\mathbf{x}, \Lambda, \Theta) = \frac{1}{Z(\mathbf{x}, \Lambda, \Theta)} \frac{1}{p(\mathbf{x}|c, \Theta)} \sum_{\mathbf{s} \in S(c)} exp^{\sum_{t=1}^{T} \tilde{\lambda}_{s_t} \cdot \tilde{\phi}(\mathbf{x}_t, c, s_t, t, \Theta)} \tag{9}$$

This model is very close to a HCRF model and may be trained using similar optimization algorithms (SGD, LBFGS etc) for maximizing the conditional likelihood of training data (note that we used a standard L2 regularization term).

4 Experiments

We experimented our hybrid model on financial time series and handwriting data. The financial time series dataset is a dataset of what are called chart patterns. A chart pattern is a particular shape of a stock exchange series of interest for financial operators. We used two databases, the first one (*CP4*) includes 448 series corresponding to the 4 most popular patterns *Head and Shoulders*, *Double Top*, *Reverse Head and Shoulders* and *Reverse Double Top* (see Figure 1). The second dataset *CP8* includes 892 patterns from 8 classes, the four previous ones and four additional chart patterns :*Triple Top* (and the reverse pattern), *Ascending Triangle* and *Descending Triangle*. We report cross-validation results (4 folds for CP8 and 7 folds for CP4). HMM and HCRF model of a particular class has a number of states corresponding to the number of ideal segments of the pattern (e.g. 6 states for *Head and Shoulders* and 4 states for *Double Top*).

The handwriting database is a subset of the benchmark IAM database [4], it includes images of handwritten English characters. These images are transformed into series of feature vectors by using a sliding window moving from the

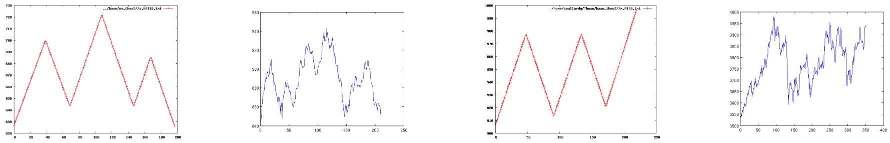

Fig. 1: From left to right: ideal shape of a Head and Shoulder pattern (HS), observed HS, ideal shape of an Ascending Triangle pattern (AT), observed AT.

left to the right of the image (and using preprocessing as in [4]). The dataset is divided into a training set (200 samples per class), a validation set (50 samples) and a test set (50 samples). We studied three settings corresponding to three training dataset sizes: a small setting with 25 training samples per class (HLS), a medium with 50 training samples (HLM) and a larger one with 100 training samples per class(HLL). For each training set size, we report averaged results gained by using subsets of the training dataset (from 8 experiments for $HLS$to 2 for HLL). HMM and HCRF model of every class has 8 states.

In all experiments learning parameters are selected based on best results on the validation set and performance is measured on the test set. We compared our hybrid model to HMMs with one gaussian distribution per state with either a diagonal or a full covariance matrix, and to HCRFs using a HMM based initialization as in [6]. Note that HCRF using random initialization perform poorly, with for instance a performance of 86.4% on CP4 and of 68.4% on CP8.

Table 1 reports comparative results gained with all the models on the four datasets. These results call for some comments. First, except for HLL case, best results are achieved by hybrid systems with strong improvements over HMMs and over HCRFs. Second, focusing on Diagonal Covariance Gaussian HMMs (HMMD) based systems, one sees that initializing from a learned HMMD allows improving upon HMMD performance. Third, looking at diagonal and full cases, one sees that HCRF using a HMM based initialization improve over HMMs but hybrid models reach best performances most of the times.

Model	CP4	CP8	HLS	HLM	HLL
HMMD	87.5%	70.4%	38.3%	40.3%	44.3%
HMMF	91.3%	74.3%	41.9%	46.9%	**50.6%**
HCRF (HMMD init)	90.2%	76.7%	39.7%	43.5%	46.7%
HYBRID HCRF-HMMD	90.2%	77.5%	39%	41.6%	45.1%
HYBRID HCRF-HMMF	**92%**	**79.4%**	**42.5%**	**47.2%**	49.3%

Table 1: Accuracy on Chart Pattern and on Handwriting datasets using diagonal covariance gaussian HMM (HMMD), full covariance gaussian HMMs (HMMF), HCRF initialized based on a learned HMMD as in [6], and hybrid models exploiting a HMMD and a HMMF.

5 Conclusion

We proposed a new hybrid framework that combines two well-known modelling, Hidden Markov Models and Hidden Conditional Random Fields. Our main idea is to introduce a HMM-based weighting in the conditional probability of the HCRF which constrains the discriminative learning, yielding improved accuracy.

References

[1] C. M. Bishop, J. Lasserre, Generative or discriminative? getting the best of both worlds. In J. M. Bernardo and al., editor, *proceedings of the* 8^{th} *World Meeting on Bayesian Statistics* (ISBA 2006), Oxford University Press, vol. 8, pages 3-24, Spain, 2007.

[2] G. Bouchard, Bias-Variance Tradeoff in Hybrid Generative-Discriminative Models, *proceedings of the* 6^{th} *International Conference on Machine Learning and Applications* (ICMLA 2007), IEEE Computer Society pub., pages 124-129, December 13-15, Cincinnati (USA), 2007.

[3] G. Bouchard and W. Triggs, The trade-off between generative and discriminative classifiers. In J. Antoch, editor, *proceedings of the* 16^{th} *Symposium of the International Association for Statistical Computing* (CompStat 2004), Physica-Verlag pub., pages 721-728, August 23-27, Czech, 2004.

[4] H. Bunke and U. Marti, A full english sentence database for off-line handwriting recognition, *proceedings of the* 5^{th} *International Conference on Document Analysis and Recognition* (ICDAR 1999), pages 705-708, September 20-22, India, 1999.

[5] T-M-T. Do and T. Artières(2009), Large margin training for hidden Markov models with partially observed states, *proceedings of the* 26^{th} *International Conference on Machine Learning* (ICML 2009), pages 265-272, Montreal (Canada), 2009.

[6] A. Gunawardana and M. Mahajan and A. Acero and J. C. Platt, Hidden Conditional Random Fields for Phone Classification, *proceedings of Interspeech* (Interspeech 2005), pages 1117-1120, 2005.

[7] J. Lafferty and A. McCallum and F. Pereira, Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In C. E. Brodley and A. P. Danyluk, editor, *proceedings of the* 18^{th} *International Conference on Machine Learning* (ICML 2001), M. Kaufmann pub., pages 282-289, 2001.

[8] J. Lasserre, C. M. Bishop, T. P. Minka, Principled Hybrids of Generative and Discriminative Models. *proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR 2006), IEEE Computer Society pub., vol. 1, pages 87-94, June 17-22, New York City (USA), 2006.

[9] T. Minka, Discriminative Models, not discriminative training. Technical Report, Microsoft Research, Cambridge, TR-2005-144, England, October 2005.

[10] Andrew Y. Ng and Michael I. Jordan, On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In T. G. Dietterich and S. Becker and Z. Ghahramani, editor, *Advances in Neural Information Processing Systems 14* (NIPS 2001), MIT Press pub., pages 841-848, December 3-8, Vancouver (Canada), 2001.

[11] A. Quattoni, S. Wang, L. P. Morency, M. Collins, T. Darrell, Mit Csail, Hidden-state Conditional Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (IEEE TPAMI 2007), IEEE Computer Society pub., vol. 29, issue 10, pages 1848-1852, October 2007.

[12] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, A. Waibel and K-F. Lee, editor, *proceedings of the IEEE* (IEEE 1989), Morgan Kaufmann pub., vol. 77, No. 2, pages 257-286, February 1989.

[13] P. Woodland P. and D. Povey. Large scale discriminative training of hidden markov models for speech recognition. *Computer Speech and Language*, 2002, Elsevier, Vol. 16, issue 1, pages 25-47, January 2002.