Polynomial Conditional Random Fields for signal processing

Poster: Trinh-Minh-Tri Do and **Thierry Artières**[1]

Abstract. We describe Polynomial Conditional Random Fields for signal processing tasks. It is a hybrid model that combines the ability of Polynomial Hidden Markov models for modeling complex dynamic signals and the discriminant power of Conditional Random Fields. We detail the learning of these models and report experimental results on handwriting recognition.

1 INTRODUCTION

Classification, segmentation and labelling of sequential data are major problems in many application fields such as bioinformatics, signal processing (e.g. handwriting), information extraction and so on. Generally speaking these tasks consist in transforming an observed input signal or sequence into a sequence of labels. For decades, Hidden Markov Models (HMMs) have been the most popular approach although they rely on strong independence assumptions and despite they are learned using a non discriminant criterion, namely Maximum Likelihood Estimation. This comes from the fact that HMMs are generative models and that they define a joint probability distribution on the sequence of observations *X* and the associated label sequence *Y*.

A number of models were proposed for relaxing classical independence assumptions; most rely on segmental HMMs where trajectory models (e.g. polynomial) are used in each state to implement probability density functions at the segment level [1]. Besides conditional models were recently proposed for sequence processing tasks, Conditional Random Field (CRF) is the most popular one [2]. CRFs aim at modelling directly the conditional distribution *P(Y/X)*, they are attractive since first they are learned with a dicriminant criterion where each label sequence is viewed as a possible class for an input sequence and second, being conditional, these models do not explicitly model input signals distribution so that they do not require any simplifying assumptions over *X*. However, CRF rely on the definition by hand of set of (eventually many) features, computed from observations *X* and label sequence *Y*. In a way underlying assumptions over *X* are implicit and hidden in the manual design of features. Up to now CRFs have been used in specific text processing tasks such as information extraction, FAQ segmentation, POS-tagging for which efficient features may be easily designed. The use of conditional models for continuously varying signals is less straightforward.

This paper aims at developing a model which combines the accurate modelling of polynomial HMMs (PHMMs) with the efficiency of CRF discriminant learning. We first present Polynomial CRF (PCRF) and then we report comparative experimental results on on-line digit recognition.

[1] Pierre and Marie Curie University, France, email: Trinh-Minh-Tri.Do@lip6.fr, Thierry.Artieres@lip6.fr

2 POLYNOMIAL CRF

Let $X = (x_1, x_2, ..., x_T)$ with $x_t \in \Re^d$ be an input signal. Assuming that successive observations are generated according to a polynomial function of time, this sequence of observations may be modelled with a polynomial regression function:

$$x_t = f(t) = \sum_{k=0}^{P} a_k f_k(t) + \varepsilon(t) \quad (1)$$

where P is the degree of f, a_k are coefficients of polynomial functions f_k and $\varepsilon(t)$ is an i.i.d. residual. $\{f_k\}_k$ consists of a fixed set of polynomials (e.g. x^k). Assuming $\varepsilon(t)$ is a white Gaussian noise, the likelihood of a sequence may be computed as:

$$P(X) \propto e^{-\frac{1}{2}\sum_{t=1}^{T}\left\| x_t - \sum_{k=0}^{P} a_k f_k(t) \right\|^2} \quad (2)$$

Such polynomial functions have been embedded in HMM to implement more accurate probability density functions (*pdf*) than traditional Gaussian distributions. These Polynomial HMMs (PHMMs) have been shown to be more accurate in speech modelling and recognition tasks, they are trained using a Maximum Likelihood training algorithm which is a generalization of the standard HMM training algorithm [1].

Besides in *CRFs* the conditional probability of the output *Y* (i.e. the label sequence) given the input *X* (i.e. the sequence of observations) is defined with:

$$P(Y/X) = \frac{e^{W^t F(X,Y)}}{Z(X)} \quad (3)$$

where $F(X,Y)$ is a feature vector, W is the weight vector, $W^t F(X,Y)$ is their dot product, and $Z(X)$ is a normalization factor. Training consists in maximizing the conditional log likelihood over the training set w.r.t. *W*:

$$L = \sum_i \log P(Y_i / X_i) \quad (4)$$

Our model is an extension of CRFs for signal classification using segmental (polynomial) features, i.e. features that are defined on segments of successive observations. In this model, we use an ideal polynomial curve describing the sequence of successive observations in each state. Let *S* be a segmentation of *X* in the model, i.e. a state sequence. It may be decomposed into a sequence

of segments where a segment stands for a sequence of successive observations that are assigned to the same state. Let note s_j the j^{th} segment and let β_j stands for the starting time of s_j. Then $s_j = (x_{\beta_j}, ... x_{\beta_{j+1}-1})$. The probability $P(Y/X)$ may then be defined as:

$$P(Y/X) = \frac{1}{Z(X)} \sum_{S} e^{score(S,Y,X)} \quad (5)$$

where $score(S,Y,X)$ stands for the score of the label sequence Y along a particular segmentation S for input X. It may be further decomposed into:

$$score(S,Y,X) = \sum_{j=1}^{|S|} score(s_j, q_{j,Y}, X) \quad (6)$$

Where $q_{j,Y}$ is the state corresponding to the j^{th} segment in segmentation Y and $score(s_j, q, X)$ stands for the score associated to j^{th} segment and state q. Following works on polynomial regression, this score is linked to how well the segment fits with the trajectory model associated to the state. Following works in [1] and using Eq. (1) and Eq. (2), it is computed as:

$$score(s_j, q, X) = \sum_{t=\beta_j}^{\beta_{j+1}-1} \left\| x_t - \sum_{k=0}^{P} a_k^q f_k(t-\beta_j) \right\|^2 \quad (7)$$

Where a_k^q stand for the coefficients of the polynomial function in state q. To perform sequence classification we build a CRF whose architecture is illustrated in Fig. 1. There is one left-right chain model for each class. At test time, the most probable sequence of states (i.e. left-right chain) indicates the recognized class [3].

Training aims at estimating model's parameters, i.e. a_k^q. As for standard CRFs, this is done through maximizing the criterion in Eq. (4). There are two main problems for training polynomial CRF. First, unlike what is usually done with standard CRFs, the training database is only partially labelled. This means that the segmentation S corresponding to a training sequence is not known, only its class is known. This problem may be solved by introducing, in the training algorithm, hidden variables that correspond to segmentations S of the training sequences [3,4]. One can then use an EM-like learning algorithm consists then in iterating two steps. In the E-step, we one estimates the distribution over hidden variables $P(S/X,Y)$. In the M-step one updates the model parameters to reach a better likelihood [3]. Second, unlike standard CRFs, training involves a non linear optimization problem since the usual dot product in Eq. (3) is replaced here by a non linear term (Eq. (7)). Hence, in the M-step algorithm, parameters a_k^q are updated using a standard gradient descent method till convergence.

3 EXPERIMENTS

We carried out experiments on on-line digit recognition. On-line handwriting signals are captured with an electronic pen or a digitalized tablet; there are temporal sequences of pen coordinates. As is usually done in handwriting processing, raw signals are first normalized and transformed into a sequence of direction feature (angle of the trajectory with the horizontal axis). There are about 50 samples for every digit '0' to '9'. We report in Table 1 comparative results of a generative system (PHMMs) and our discriminative system (PCRFs) that have been obtained using cross validation. Performances are reported for various settings concerning the number of states in the models (1 to 4) and the degree of polynomials (0 to 3). Almost whatever the number of states and the degree of polynomials PCRFs significantly outperform PHMMs. Although increasing the number of states in PHMMs improves accuracy it does not easily reach performance of PCRFs. In order to investigate deeper the behaviour of the models we generated from the learned models "mean samples" using polynomial trajectories learned in each state. Figure 2 compares these samples for the generative PHMM system as well as for the discriminant PCRF. One may see that although PHMM perform much lower, they allow modelling accurately the input signals. On the contrary PCRF, being discriminant, have focused on what distinguishes input signals of the ten digits so that mean samples do not represent well the original digits.

Table 1. Accuracy of PHMM and PCRF for digit recognition as a function of the number of states per digit model and of the polynomial degree.

	#states	P=0	P=1	P=2	P=3
PHMMs	1	67.4	88.2	93.0	93.6
	2	92.2	97.2	98.0	98.4
	3	96.4	98.6	98.6	98.6
	4	96.6	98.6	98.4	98.6
PCRFs	1	65.4	89.8	96.8	97
	2	98.0	99.0	99.0	98.8
	3	99.2	99.8	99.4	99.2
	4	98.6	98.8	98.8	98.2

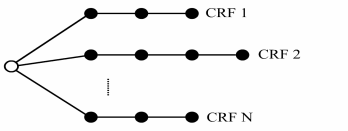

Figure 1. Mixture of chained CRF (i.e. with a chain structure) for sequence classification. Each chain corresponds to a class.

Figure 2. Mean samples of digits learned by a generative system (PHMM) on the left and a discriminant system (PCRF) on the right. Both systems are based on two-states models.

4 CONCLUSION

We presented Polynomial Conditional Random Fields for signal processing and discussed their learning. These models are hybrid and allow discriminant learning of trajectory models for signal classification and segmentation. PCRFs are shown to outperform non discriminant comparable polynomial generative models.

5 REFERENCES

[1] Deng L., Aksmanovic M., Sun X. and Wu C.F.J, *Speech Recognition Using Hidden Markov Models with Polynomial Regression Functions as Nonstationary States*, IEEE Transactions on Speech and Audio Processing, 2:4, 507-520 (1994).

[2] Lafferty J., McCallum A. and Pereira F., *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, International Conf. on Machine Learning, 282–289. Morgan Kaufmann, San Francisco, CA (2001).

[3] Do T-M-T., Artières T., *Conditional Random Field for tracking user behavior based on his eye's movements*, NIPS'05 Workshop on Machine Learning for Implicit Feedback and User Modeling (2005).

[4] Sarawagi S. and Cohen W., *Semi-Markov Conditional Random Fields for Information Extraction*, Advances in Neural Information Processing Systems NIPS (2004).