Model and Feature Selection in Hidden Conditional Random Fields with Group Regularization

Rodrigo Cilla, Miguel A. Patricio, Antonio Berlanga, and José M. Molina

Computer Science Department. Universidad Carlos III de Madrid
Avda. de la Universidad Carlos III, 22
28270 Colmenarejo (Madrid). Spain
{rcilla,mpatrici}@inf.uc3m.es, {aberlan,molina}@ia.uc3m.es

Abstract. Sequence classification is an important problem in computer vision, speech analysis or computational biology. This paper presents a new training strategy for the Hidden Conditional Random Field sequence classifier incorporating model and feature selection. The standard Lasso regularization employed in the estimation of model parameters is replaced by overlapping group-L1 regularization. Depending on the configuration of the overlapping groups, model selection, feature selection,or both are performed. The sequence classifiers trained in this way have better predictive performance. The application of the proposed method in a human action recognition task confirms that fact.

1 Introduction

Sequence modelling methods are applied in multiple areas. They are employed by computational biologists to model proteins [1]. The natural language processing community uses them to solve chunking or part-of-speech tagging tasks [2]. They are also applied in action recognition from video [3].

Probabilistic graphical models [4] are employed in sequence modelling. The generative Hidden Markov Model has been employed in many works. Multiple variations have been proposed to capture the peculiarities of different sequence modelling scenarios. Efficient exact and approximate algorithms exist to perform the associated inference tasks. Recently, discriminative sequence models such the Hidden Conditional Random Field (HCRF)[5] have emerged as a new alternative. They provide compact parametrizations and have higher predictive power. However, they still have reduced applicability and have not displaced generative models.

This work wants to foster the spread of discriminative sequence classifiers incorporating model and feature selection to the training algorithm of the HCRF. The Occams Razor principle of machine learning stands that a model should not be more complex than strictly required. Model and feature selection are two ways of implementing it, obtaining a more compact result. Model selection in the

J.-S. Pan et al. (Eds.): HAIS 2013, LNAI 8073, pp. 140–149, 2013.

context of the HCRF refers to the determination of the optimal number of hidden state variables, while feature selection refers to the selection of informative features in the input sequences while discarding uninformative ones.

1.1 Contributions

The contributions of this paper might be summarized as follows:

- A new training procedure for the HCRF incorporating model and feature selection.
- Experimental evidence showing than the proposed training algorithm performs better than the standard HCRF in a standard action sequence classification task.

1.2 Paper Organization

Paper is organized as follows: section 2 introduces the standard HCRF model; the proposed training procedure is presented on section 3; experimental evidence of the higher performance of the proposed method in a human action classification task is reported on section 4; finally, 5 resumes the contributions of this work and presents new research directions.

2 Hidden Conditional Random Fields

The HCRF [5] is an undirected graphical from the exponential family. It might be understood as an extension of the Conditional Random Field with hidden variables to model correlations among different observations. Multiple structured prediction tasks might be represented with HCRFs. This work assumes, without loss of generality, a sequence classification task.

Formally, the HCRF defines the conditional probability distribution of a discrete random variable $y \in \{y_1, \ldots, y_N\}$ (a.k.a. sequence label) given a sequence of random variables $\mathbf{x} = x_1, \ldots, x_T$ (a.k.a. observations) employing a set of auxiliary discrete hidden variables $\mathbf{h} = h_1, \ldots, h_T$, $h_i \in \mathcal{H}$ not observed during training. These variables are introduced to model correlations among the observations in $\mathbf{x}$. In the case of sequence classification, these correlations correspond to the sequence dynamics. The conditional probability of the sequence label y and the hidden variable assignments $\mathbf{h}$ given the sequence of observations $\mathbf{x}$ is defined using the Hammersley-Clifford theorem of Markov Random Fields:

$$P\left(y, \mathbf{h} \mid \mathbf{x}, \theta\right) = \frac{e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}}{\sum_{y'} \sum_{h} e^{\Psi(y', \mathbf{h}, \mathbf{x}; \theta)}} \tag{1}$$

The conditional probability of the class label y given the observation sequence $\mathbf{x}$ is obtained marginalizing over all the possible value assignments to hidden parts $\mathbf{h}$:

$$P(y \mid \mathbf{x}, \theta) = \frac{\sum_{h} e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}}{\sum_{y'} \sum_{h} e^{\Psi(y', \mathbf{h}, \mathbf{x}; \theta)}} \tag{2}$$

The potential function $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$ measures the compatibility of the input $\mathbf{x}$ with the assignments to the hidden variables $\mathbf{h}$ and the class label y. There are multiple possibilities about the form of this function. Here it is defined as:

$$\Psi(y, \mathbf{h}, \mathbf{x}; \theta) = \sum_{t=1}^{T} \phi(x_t)\, \alpha(h_t) + \sum_{t=1}^{T} \beta(h_t, y) + \sum_{t=1}^{T} \gamma(h_t, h_{t+1}, y) \qquad (3)$$

where $\phi(x_t) \in \mathcal{R}^d$ is the feature vector associated with the observation x_t and $\theta = [\alpha \quad \beta \quad \gamma]$ is the vector of model parameters, indexed according to the values given to the hidden variables $\mathbf{h}$ and label y. The first term, parametrized by $\alpha(h_t) \in \mathcal{R}^d$ measures the compatibility of the observation at instant x_t with the assignment to the hidden variable h_t. The second term measures the compatibility of the values given to the hidden parts h_t with the class label y and is parametrized by $\beta(y, h_i) \in \mathcal{R}$. Finally, the third term, parametrized by $\gamma(y, h_t, h_{t+1}) \in \mathcal{R}$ models sequence dynamics, measuring the compatibility of adjacent hidden variable assignments h_t and h_{t+1} with the class y.

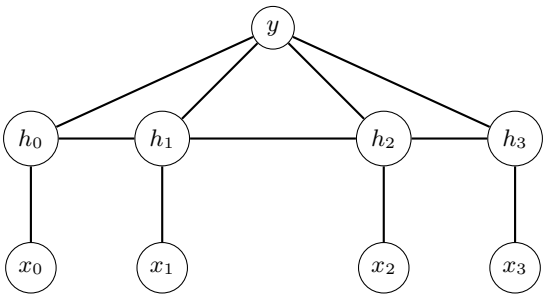


Fig. 1. Graphical model representing the structure of the HCRF induced by the function Ψ

The function Ψ induces the structure of the undirected graphical model defined by the HCRF. The structure of this graph can be observed on figure 1. Exact inference of the conditional probability distribution defined in equation 2 is possible, as the dependencies among the values given to the hidden variables $\mathbf{h}$ form a chain. Efficient inference is achieved employing belief propagation [4].

2.1 Parameter Estimation

Optimal model parameters θ^* are estimated from a set of K training samples $\left(\mathbf{x}^i, y^i\right), 1 \leq i \leq K$, minimizing the L_2 regularized negative conditional log-likelihood function:

$$\theta^* = \arg\min_{\theta} L(\theta) = \arg\min_{\theta} -\sum_{i=1}^{K} \mathcal{L}\left(\mathbf{x}^i, y^i; \theta\right) + \lambda R(\theta). \qquad (4)$$

The first term measures how model parameters are adjusted to predict each one of the K training samples, while the second term acts as a regularization prior over model parameters. The standard regularization employed in the HCRF is the Ridge regularizer, defined as $R(\theta) = ||\theta||_2^2$, imposing a zero-mean gaussian prior on the values of θ to prevent overfitting. The parameter λ defines a trade-off between regularization and adjustment. A value of $\lambda = \frac{1}{2\sigma^2}$ is equivalent to a gaussian with variance σ^2. The conditional log-likelihood function $\mathcal{L}(\mathbf{x}, y; \theta)$ is defined as:

$$\mathcal{L}(\mathbf{x}, y; \theta) = \log P(y \mid \mathbf{x}, \theta) = \log \left(\frac{\sum_h e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}}{\sum_{y'} \sum_h e^{\Psi(y', \mathbf{h}, \mathbf{x}; \theta)}} \right) \tag{5}$$

Due to the presence of the hidden variables $\mathbf{h}$, the objective function in equation 4 is non-convex [6]. However, a local optimum θ^* for the model parameter values might be obtained employing standard convex optimization techniques, as the function in 4 has an smooth gradient.

Different search strategies might be employed to find the optimal parameter values. Among them, the LBFGS quasi-newton method is the most popular [7], updating the descent direction with an approximation of the Hessian based on previous gradient estimations. Others have proposed to employ an online stochastic gradient descent algorithm [7], achieving a fast convergence rate but at the cost of obtaining a worst quality solution. In any case, the non-convexity of the objective function to optimize makes necessary to run the search multiple times from different starting points.

2.2 Limitations

The standard method to estimate HCRF optimal parameters leaves some open issues that are going to be discussed in order to motivate the proposal in subsequent section. These are:

- **How many hidden state variables employ?** $|\mathcal{H}|$ i.e., the number of different values that the hidden state variables in $\mathbf{h}$ can take, should be specified *a priori*. If it is too small, the model is not enough expressive to capture the required correlations. However, if it is too big, noisy correlations are modelled and the result has a low predictive performance. Thus, it is necessary to adjust it to the right number. In practice, this is done employing cross-validation, evaluating the predictive performance for different choices and selecting the best. The non-concavity of the loss function in equation 4 complicates this process, as many trials should be made per choice to obtain a fair estimation of the optimality of each value. Thus, an efficient procedure is needed.
- **What happens if there are irrelevant features in the input sequences?** The L2 norm in equation 4 gives a non-zero weight to the parameters $\alpha(h_t)$ corresponding to irrelevant features. Thus, the result model does not have an optimal performance, as noise is incorporated to the inference

process. Thus, it is necessary to incorporate a method to select appropriate features from the input while discarding the irrelevant.

Other problem in the estimation of optimal HCRF parameters is how to adjust the trade-off between parameter fitting and regularization, i.e., what value give to λ in equation 4. This problem is shared by every regularized log-linear model. In practice, λ is adjusted employing cross-validation, needing to try different values until the one with the best performance is obtained. This adds another cross-validation dimension, as it should be already employed in the selection of the right number of hidden state values. The problem of estimating the right value for λ is out of the scope of this paper.

3 Model and Feature Selection in Hidden Conditional Random Fields

This section presents an overlapping group-L1 regularization strategy to estimate optimal parameters for the HCRF sequence classifier. As described in previous section, the components of the HCRF parameter vector θ are divided into three groups $\alpha(h_t)$, $\beta(h_t, y)$ and $\gamma(h_t, h_{t+1}, y)$, respectively indexed by the values of h_t, h_t and y and h_t, h_{t+1} and y. To obtain a model selection effect it is necessary to obtain zero values for all the parameters related to each unnecessary h. In a similar way, to perform feature selection it is necessary to obtain a zero value for all the parameters related to irrelevant input features.

Model and feature selection in log-linear models has been reported replacing L2 regularization of the objective function by L1 regularization[8]. However, L1 regularization is not enough to obtain model and feature selection in the HCRF as it only gives zero values to single variables and not to groups of them.

One way of obtaining zeros in groups of variables is employing overlapping group L1 regularization [9,10]. Be $\mathcal{G}$ the power set of the parameter vector θ, and $G \subseteq \mathcal{G}$ an arbitrary subset of the power set. The overlapping group-L1 regularized training of the HCRF is given by the solution to the optimization problem:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\theta) + \sum_{g \in G} \lambda_g \left\| \theta_g \right\|^2 \tag{6}$$

The overlapping group-L1 norm sums the L2 norm of the different groups defined in G. At the optimal, some of the groups will have a zero norm, as all the components from those groups will have become zero. Depending on the way the set G is defined, model selection, feature selection, both or even other advanced effects might be achieved:

- If $G \equiv G_{fs} = \cup_{d=1}^{D} \{\alpha(\cdot)_d\}$ feature selection is performed, as the L2 norm of the input features is penalized. A zero weight is expected for all the parameters corresponding to an input feature. Note that beta and gamma parameters are also regularized in order to prevent a big value on them, causing overfitting.

- If $G \equiv G_{ms} = \cup_{h=1}^{|\mathcal{H}|} \{\alpha(h) \cup \beta(h,\cdot) \cup \gamma(h,\cdot,\cdot) \cup \gamma(\cdot,h,\cdot)\}$ model selection is performed, as the L2 norm of the parameters corresponding to a hidden variable is minimized. A zero weight is expected to the parameters corresponding to non necessary hidden parts.
- If $G \equiv G_{fs} \cup G_{ms}$ model and feature selection are performed at the same time.

3.1 Optimization Algorithms

The convex optimization methods employed to estimate the optimal parameters of the standard HCRF are no longer valid. The new regularization term makes the objective function to optimize non-smooth. In particular, the gradient has a singularity at the points where a group gets a zero L2 norm. It is necessary to transform the problem into a smooth one before applying a gradient based method.

The unconstrained optimization problem in equation 6 might be reformulated into an equivalent constrained optimization problem as suggested by [11]:

$$\begin{aligned} \theta^* &= \min_{\theta} \mathcal{L}(\theta) + \sum_{g \in G} \lambda_g h_g \\ s.t. & \\ \forall g \quad & \|\theta_g\|_2 \leq h_g \end{aligned} \tag{7}$$

The overlapping group-L1 regularization term is replaced by a set of constraints, one for each group of variables in G. Each one of the constraints in the new optimization problem defines a norm cone of radius h_g, ensuring that the L2 norm of each group is smaller than h_g. A norm cone is a convex set, and the intersection of a set of convex sets is also a convex set [6]. Thus, the feasible region defined by the restrictions is convex. The norms of the different groups are added to the objective function. At the optimum the constraints are fulfilled with equality (it is trivial to probe that if they are not then it is not the optimal).

The objective function of the optimization problem in equation 7 is smooth, as the cause for the singularities has been removed. The estimation of the optimal parameters is made employing a gradient descent method, projecting the obtained values into the feasible set defined by the restrictions.

Dykstra's algorithm [12] solves the problem of projecting a point $w_0 \in \mathcal{R}^k$ into the intersection of a set of convex sets $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_q$, alternately projecting the point into each set and removing the residual from the previous step.

To obtain the optimal parameter values different search methods have been proposed in [11]. Here the Projected Quasi-Newton (PQN) optimization method is employed. It builds a second-order approximation of the objective function around the current point to find the minimizing direction. The method avoids evaluating the objective function int the neighbourhood, assuming that computing the projections is cheaper than evaluating the objective function. Readers are referred to the original publication for further details on the method.

4 Experimental Evaluation

This section provides experimental evidence about the improvements that overlapping group-regularized training of HCRF models produces in their predictive power.

4.1 Experimental Setup

The system presented in figure 2 has been built to test the proposed method in a human action sequence classification task. The distance transform [13] is computed for each one of the human silhouettes extracted from the frames in the input sequence. A 3072 dimensional descriptor is obtained for each frame. The resulting sequence is introduced to the trained HCRF model to predict action class.

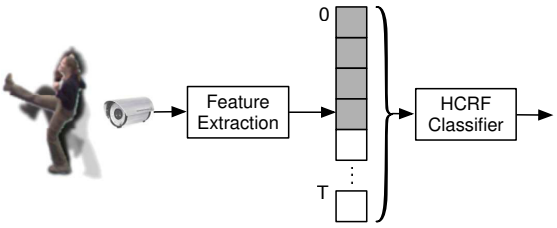


Fig. 2. Action Recognition Pipeline employed for evaluation

The models to be tested in order to evaluate the proposal are.

1. HCRF: The standard HCRF model as shown on section 2, employing L2 regularization. Optimal model parameters are obtained with LBFGS optimization.
2. MFS-HCRF: The Hidden Conditional Random Field trained with L1 group regularization to perform feature and model selection, as shown in section 3.

The predictive performance of these algorithms is going to be measured employing Weizmann dataset[1]. It contains 10 different actions performed by 9 actors once, to give a total of 90 clips. Note that perfect classifications has been already reported for the dataset in [14]. However, the purpose of the experiments to be presented is to compare the performance of the presented algorithms in the task and not to try to provide a better way of performing Human Action Recognition.

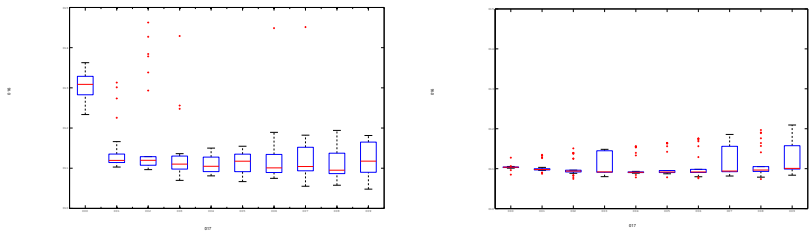
The models are trained employing $|\mathcal{H}| = 20$ hidden parts, twice the number of action classes in Weizmann dataset. Iterative algorithms are applied until convergence. The non-convexity of the objective functions to be optimized forces to employ of a monte-carlo approach to evaluate each configuration to obtain fair results. Thus, each configuration in tested 30 times averaging the obtained results.

[1] `http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`

4.2 Experiment I: Finding a Good Regularization Trade-Off

The first experiment to be conducted is to find for the different models a good value for the regularization parameter λ, providing a good equilibria between adaptation to the training data and regularization. The optimum is defined as the value minimizing the median negative log-likelihood obtained in the prediction of a test set. To this end sequences from Weizmann dataset are split in different subsets according to the actor. Sequences from actor 1 are employed as test set, while sequence from actors 2-9 are employed to train models.

Boxplots on figures 3(a) and 3(b) respectively show negative conditional log-likelihood values obtained for different values of λ for HCRF and MFS-HCRF. The negative log-likelihood values obtained for MFS-HCRF are smaller than the obtained for HCRF. Thus, the MFS-HCRF has a better predictive performance than the HCRF. Boxplots also show that the variance in negative log-likelihood values for the MFS-HCRF are slower than for the HCRF. This fact might be motivated by a softer objective function landscape, where local minima from the loss term of the objective function gets more penalized by the group regularization term.



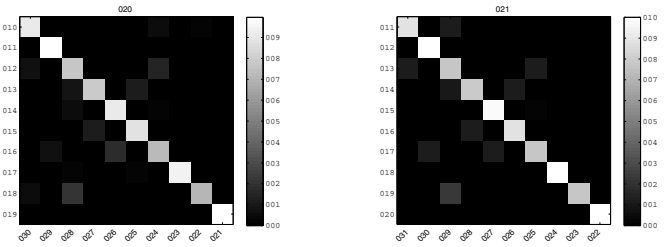
(a) HCRF (b) MFS-HCRF

Fig. 3. Negative log-likelihood values achieved for different values of λ

4.3 Experiment II: Action Recognition Results

Previous experiment has shown that MFS-HCRF has higher predictive performance than HCRF for action sequences from a single actor. Now model performance is going to be measured in the prediction of the complete Weizmann dataset, measuring just predictive accuracy. This is done employing Leave One Actor Out Cross-Validation. Dataset is split again in different subsets according to the actor performing the sequence. The sequences from one actor are employed to measure the performance of models trained with the remaining actors. The process is repeated until every actor has been employed in the evaluation, joining the obtained results. The parameter λ is adjusted for the minimum value found in previous experiment.

Figures 4(a) and 4(b) present the confusion matrices respectively obtained for HCRF and MFS-HCRF. MFS-HCRF has a performance about a 2% higher than



(a) HCRF (b) MFS-HCRF

Fig. 4. Confusion matrices obtained for the different models in the prediction of Weizmann dataset

HCRF. Thus, the overlapping group-L1 regularized training of HCRF produces models with a higher predictive performance for the prediction of the action classes in Weizmann Dataset than those trained with standard L2 regularization.

5 Conclusions

This paper has presented a new training algorithm for the HCRF based on overlapping group-L1 regularization. Models trained with the proposed algorithm are more compact than the obtained by the standard algorithm, as model and feature selection is performed during training. Experiments have shown that the proposed algorithm recovers models with a higher predictive performance than the standard in an action recognition task.

Future works will validate the proposed method in other sequence classification tasks beyond human action recognition. The proposed algorithm might be adapted to provide model and feature selection in the estimation of optimal parameter values of other discriminative graphical models with hidden variables.

Acknowledgement. This work was supported in part by Projects MINECO TEC2012-37832-C02-01, CICYT TEC2011-28626-C02-02, CAM CONTEXTS (S2009/TIC-1485)

References

1. Krogh, A., Brown, M., Mian, I.S., Sjolander, K., Haussler, D.: Hidden markov models in computational biology: Applications to protein modeling. Journal of Molecular Biology 235, 1501–1531 (1994)
2. Collins, M.: Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 1–8. Association for Computational Linguistics (2002)

3. Yamato, J., Ohya, J., Ishii, K.: Recognizing human action in time-sequential images using hidden markov model. In: Proceedings of the 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1992, pp. 379–385 (1992)
4. Bishop, C., et al.: Pattern recognition and machine learning, vol. 4. Springer, New York (2006)
5. Quattoni, A., Wang, S., Morency, L.P., Collins, M., Darrell, T.: Hidden conditional random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 1848–1853 (2007)
6. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press (2004)
7. Zhu, C., Byrd, R., Lu, P., Nocedal, J.: Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software (TOMS) 23, 550–560 (1997)
8. Ng, A.: Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In: Proceedings of the Twenty-First International Conference on Machine Learning, p. 78. ACM (2004)
9. Huang, J., Zhang, T.: The benefit of group sparsity. The Annals of Statistics 38, 1978–2004 (2010)
10. Szabó, Z., Póczos, B., Lorincz, A.: Online group-structured dictionary learning. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2865–2872. IEEE (2011)
11. Schmidt, M.: Graphical model structure learning with l1-regularization. PhD thesis, University of British Columbia (2010)
12. Bauschke, H., Lewis, A.: Dykstras algorithm with bregman projections: A convergence proof. Optimization 48, 409–427 (2000)
13. Wang, L., Suter, D.: Visual learning and recognition of sequential data manifolds with applications to human movement analysis. Computer Vision and Image Understanding 110, 153–172 (2008)
14. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 2247–2253 (2007)