# AUC-maximized Deep Convolutional Neural Fields for Sequence Labeling

**Sheng Wang , Siqi Sun, Jinbo Xu**
Toyota Technological Institute at Chicago
wangsheng, siqi.sun, j3xu@ttic.edu

## Abstract

Learning from complex data with imbalanced label distribution is a challenging problem, especially when the data/label form structure, such as linear-chain or tree-like. The widely-used training methods, such as maximum-likelihood and maximum labelwise accuracy, do not work well on imbalanced structured data. To model the complex relationship between the data and the structured label, we presents Deep Convolutional Neural Fields (DeepCNF), which is an integration of Deep Convolutional Neural Networks (DCNN) and Conditional Random Field (CRF). To handle the imbalanced structured data, we train DeepCNF by directly maximizing the empirical Area Under the ROC Curve (AUC), which is an unbiased measurement for imbalanced data. To fulfill this, we formulate AUC in a pairwise ranking framework and approximate it by a polynomial function and then apply a gradient-based procedure to optimize this approximation. We then test our AUC-maximized DeepCNF on three very different protein sequence labeling tasks the results confirm that maximum-AUC greatly outperforms the other two training methods.

## 1 Introduction

Deep Convolutional Neural Networks (DCNN), originated by Yann LeCun at 1998 [15] for document recognition, is being widely used in a plethora of machine learning (ML) tasks ranging from speech recognition [10], to computer vision [13], and to computational biology [5]. DCNN is good at capturing medium- and/or long-range structured information in a hierarchical manner. To handle structured data, [2] has integrated DCNN with fully connected Conditional Random Fields (CRF) for semantic image segmentation. Here we present Deep Convolutional Neural Fields (DeepCNF), which is an integration of DCNN and linear-chain CRF, to address the task of sequence labeling and apply it to three important biology problems: solvent accessibility prediction (ACC), disorder prediction (DISO), and 8-state secondary structure prediction (SS8) [17; 12].

A protein sequence can be viewed as a string of amino acids and we want to predict a label for each residue. In this paper we consider three types of labels: solvent accessibility, disorder state and 8-state secondary structure. These three structure properties are very important to the understanding of protein structure and function. The solvent accessibility is important for protein folding [6], the order/disorder state plays an important role in many biological processes [19], and protein secondary structure(SS) relates to local backbone conformation of a protein sequence [20]. The label distribution in these problems varies from almost uniform to highly imbalanced. For example, only $\sim$6% of residues are shown to be disordered [8]. Some SS labels, such as 3-10 helix, beta-bridge, and pi-helix are extremely rare [21]. The widely-used training methods, such as maximum-likelihood [14] and maximum labelwise accuracy [7], perform well on data with balanced labels but not on highly-imbalanced data [4].

This paper presents a new maximum-AUC method to train DeepCNF for imbalanced sequence data. Specifically, we train DeepCNF by maximizing Area Under the ROC Curve (AUC), which is a good measure for class-imbalanced data [3]. Taking disorder prediction as an example, random guess can obtain $\sim$94% per-residue accuracy, but its AUC is only $\sim$0.5. AUC is insensitive to changes in class distribution because the ROC curve specifies the relationship between false positive (FP) rate and true positive (TP) rate, which are independent of class distribution [3]. However, it is very challenging to directly optimize AUC. A few algorithms have been developed to maximize AUC on unstructured data [11; 9; 18], but to the best of our knowledge, there is no such an algorithm for imbalanced structured data (e.g., sequence data addressed here). To train DeepCNF by maximum-AUC, we formulate the AUC function in a ranking framework, approximate it by a polynomial Chebyshev function [1] and then use L-BFGS [16] to optimize it.

Our experimental results show that when the label distribution is almost uniform, there is no big difference between the three training methods. Otherwise, maximum-AUC results in better AUC and Mcc than the other two methods. Tested on several publicly available benchmarks, our AUC DeepCNF model obtains the best performance on all the three protein sequence labeling tasks. In particular, at a similar specificity level, our method obtains better precision and sensitivity for those labels with a much smaller occurring frequency.

## References


[1] Toon Calders and Szymon Jaroszewicz. Efficient auc optimization for classification. In *Knowledge Discovery in Databases: PKDD 2007*, pages 42–53. Springer, 2007.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[3] Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, 16(16):313–320, 2004.

[4] Gaël De Lannoy, Damien François, Jean Delbeke, and Michel Verleysen. Weighted conditional random fields for supervised interpatient heartbeat classification. *Biomedical Engineering, IEEE Transactions on*, 59(1):241–247, 2012.

[5] Pietro Di Lena, Ken Nagata, and Pierre Baldi. Deep architectures for protein contact map prediction. *Bioinformatics*, 28(19):2449–2457, 2012.

[6] Ken A Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990.

[7] Samuel S Gross, Olga Russakovsky, Chuong B Do, and Serafim Batzoglou. Training conditional random fields for maximum labelwise accuracy. In *Advances in Neural Information Processing Systems*, pages 529–536, 2006.

[8] Bo He, Kejun Wang, Yunlong Liu, Bin Xue, Vladimir N Uversky, and A Keith Dunker. Predicting intrinsic disorder in proteins: an overview. *Cell research*, 19(8):929–949, 2009.

[9] Alan Herschtal and Bhavani Raskutti. Optimising area under the roc curve using gradient descent. In *Proceedings of the twenty-first international conference on Machine learning*, page 49. ACM, 2004.

[10] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[11] Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pages 377–384. ACM, 2005.

[12] David T Jones and Domenico Cozzetto. Disopred3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, 31(6):857–863, 2015.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[15] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[16] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[17] Christophe N Magnan and Pierre Baldi. Sspro/accpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, 30(18):2592–2597, 2014.

[18] Harikrishna Narasimhan and Shivani Agarwal. A structural {SVM} based approach for optimizing partial auc. In *Proceedings of the 30th International Conference on Machine Learning*, pages 516–524, 2013.

[19] Christopher J Oldfield and A Keith Dunker. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annual review of biochemistry*, 83:553–584, 2014.

[20] Linus Pauling, Robert B Corey, and Herman R Branson. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4):205–211, 1951.

[21] Zhiyong Wang, Feng Zhao, Jian Peng, and Jinbo Xu. Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics*, 11(19):3786–3792, 2011.