Neural Gaussian Conditional Random Fields

Vladan Radosavljevic[1]*, Slobodan Vucetic[2], and Zoran Obradovic[2]

[1] Yahoo Labs, Sunnyvale, CA, USA
vladan@yahoo-inc.com
[2] Temple University, Philadelphia, PA, USA
{vucetic,zoran.obradovic}@temple.edu

Abstract. We propose a Conditional Random Field (CRF) model for structured regression. By constraining the feature functions as quadratic functions of outputs, the model can be conveniently represented in a Gaussian canonical form. We improved the representational power of the resulting Gaussian CRF (GCRF) model by (1) introducing an adaptive feature function that can learn nonlinear relationships between inputs and outputs and (2) allowing the weights of feature functions to be dependent on inputs. Since both the adaptive feature functions and weights can be constructed using feedforward neural networks, we call the resulting model Neural GCRF. The appeal of Neural GCRF is in conceptual simplicity and computational efficiency of learning and inference through use of sparse matrix computations. Experimental evaluation on the remote sensing problem of aerosol estimation from satellite measurements and on the problem of document retrieval showed that Neural GCRF is more accurate than the benchmark predictors.

Keywords: Gaussian conditional random fields, neural networks, graphical models

1 Introduction

Learning from structured data is a frequently encountered problem in geoscience [1, 2], computer vision [3, 4], bioinformatics [5, 6], and other areas where examples exhibit sequential [7, 8], temporal [9, 10], spatial [11], spatio-temporal [12, 13], or some other dependencies. In such cases, the traditional unstructured supervised learning approaches could result in a weak model with low prediction accuracy [14]. Structured learning methods try to solve this problem by learning to simultaneously predict all outputs given all inputs. The structured approaches can exploit correlations among output variables, which often results in accuracy improvements over unstructured approaches that predict independently for each example. The benefits of structured learning grow with the strength of dependency between the examples and the data size.

In structured learning there is usually some prior knowledge about relationships among the outputs. Those relationships are application-specific and, very

* This study was conducted while the author was a postdoctoral associate at Temple University.

often, they can be modeled by graphical models. The advantage of the graphical models is that one can make use of sparseness in the interactions between outputs and develop efficient learning and inference algorithms. In learning from structured data, the Markov Random Fields [2] and the Conditional Random Fields (CRF) [7] are among the most popular models. Originally, CRFs were designed for classification of sequential data [7] and have found many applications in areas such as computer vision [3] and computational biology [6].

Using CRF for regression is a less explored topic. Continuous Conditional Random Fields (CCRF) [8] is a ranking model that takes into account relationships among ranks of objects in document retrieval. With minor modifications, it can be used for structured regression problems. The Conditional State Space Model (CSSM) [15], an extension of the CRF to a domain with continuous multivariate outputs, was proposed for regression of sequential data. CSSM is an undirected model that makes no independence assumptions between outputs, which results in a more flexible framework. In [4] a conditional distribution of pixels given a noisy input image is modeled using the weighted quadratic factors obtained by convolving the image with a set of filters. Feature functions in [4] are specifically designed for image de-noising problems and are not readily applicable in regression. The Gaussian CRF for structured regression problems with feature functions constrained to quadratic form was introduced in [1]. The Sparse GCRF [10] is a variant of the GCRF model that incorporates $l1$ regularization in optimization function, thus enforcing sparsity in GCRF parameters. GCRF has recently been successfully utilized in a variety of real world applications. In the computational advertising field, GCRF significantly improved accuracy of click through rate estimation by taking into account relationship among advertisements [11]. An extension of GCRF to the non-Gaussian case using the copula transform was used in forecasting wind power [16]. In combination with decision trees, GCRF was successfully applied to short-term energy load forecasting [17], while in combination with support vector machines it was applied on automatic recognition of emotions from audio and visual features [18]. A tractable fully connected GCRF, which captures both long-range and short-range dependencies, was developed in [19] and was successfully applied on image de-noising and geoscience problems.

To improve expressive power of GCRF, we propose a Neural GCRF (NGCRF) regression model where CCRF and GCRF can be considered as special cases. In addition to using the existing unstructured predictors, the proposed NGCRF allows training of additional unstructured predictors simultaneously with other NGCRF parameters. This idea is motivated by the Conditional Neural Fields (CNF) [20, 5] proposed for classification problems to facilitate modeling of complex relationships between inputs and outputs. Moreover, weights of NGCRF feature functions are themselves allowed to be nonlinear functions of inputs. In this way, NGCRF is able to capture non-homogeneous relationships among outputs and account for differing uncertainties in the unstructured predictors. We will show that learning and inference of NGCRF can be conducted efficiently through sparse matrix computations.

2 Gaussian conditional random fields

Let us denote as $\mathbf{x} = (x_1, \dots x_M)$ an M-dimensional vector of observations and as $\mathbf{y} = (y_1, \dots y_N)$ an N-dimensional vector of real-valued output variables. The objective is to learn a non-linear mapping $f : \mathcal{R}^M \to \mathcal{R}^N$ that predicts the vector of output variables $\mathbf{y}$ as accurately as possible given all inputs $\mathbf{x}$. A CRF models a conditional distribution $P(\mathbf{y}|\mathbf{x})$, according to the associated graphical structure

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{x})} e^{\phi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}, \mathbf{x})}, \tag{1}$$

with energy function

$$\phi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}, \mathbf{x}) = \sum_{i=1}^{N} A(\boldsymbol{\alpha}, y_i, \mathbf{x}) + \sum_{j \sim i} I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x}), \tag{2}$$

$A(\boldsymbol{\alpha}, y_i, \mathbf{x})$ - association potential with parameters $\boldsymbol{\alpha}$,
$I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x})$ - interaction potential with parameters $\boldsymbol{\beta}$,
$i \sim j$ - y_i and y_j are connected by an edge in the graph structure,

and the normalization function $Z(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{x})$ defined as

$$Z(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{x}) = \int_{\mathbf{y}} e^{\phi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y}, \mathbf{x})} d\mathbf{y}. \tag{3}$$

The output y_i is associated with vector of observations $\mathbf{x} = (x_1, \dots x_M)$ by a real-valued function called the association potential $A(\boldsymbol{\alpha}, y_i, \mathbf{x})$, where $\boldsymbol{\alpha}$ is a K-dimensional set of parameters. In general, A takes as input any appropriate combination of attributes from vector of observations $\mathbf{x}$. To model interactions among outputs, a real valued function called the interaction potential $I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x})$ is used, where $\boldsymbol{\beta}$ is an L dimensional set of parameters. Interaction potential represents the relationship between two outputs and in general can depend on inputs $\mathbf{x}$. Different applications can impose different interaction potentials. The larger the value of the interaction potential, the more related the two outputs are.

In CRF applications, A and I could be conveniently defined as linear combinations of a set of fixed features in terms of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, as in [7]

$$\begin{aligned} A(\boldsymbol{\alpha}, y_i, \mathbf{x}) &= \sum_{k=1}^{K} \alpha_k f_k(y_i, \mathbf{x}), \\ I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x}) &= \sum_{l=1}^{L} \beta_l g_l(y_i, y_j, \mathbf{x}). \end{aligned} \tag{4}$$

The use of feature functions is convenient because it allows us to model arbitrary relationships between inputs and outputs. In this way, any potentially relevant feature function could be included to the model and the learning algorithm can automatically determine their relevance.

Models with real valued outputs pose quite different challenges with respect to feature function complexity than in the discrete-valued case. Discrete valued models are always feasible, because Z is finite and defined as a sum over finitely many possible values of y. On the contrary, to have a feasible model with real valued outputs, Z must be integrable. Proving that Z is integrable in general might be difficult due to the complexity of association and interaction potentials.

2.1 Feature functions

Construction of appropriate feature functions in CRF is a manual process that depends on prior beliefs of a practitioner about what features could be useful. The choice of features is often constrained to simple constructs to reduce the complexity of learning and inference from CRF.

If A and I are defined as quadratic functions of $\mathbf{y}$, $P(\mathbf{y}|\mathbf{x})$ becomes a multivariate Gaussian distribution such that learning and inference can be accomplished in a computationally efficient manner.

In the following, we describe the feature functions that led to Gaussian CRF. Let us assume we are given K unbiased unstructured predictors, $R_k(\mathbf{x})$, $k = 1, \dots K$, that predict single output y_i taking into account $\mathbf{x}$ (in a special case, only corresponding x_i can be used as $\mathbf{x}$). To model the dependency between the prediction and output, we use quadratic feature functions

$$f_k(y_i, \mathbf{x}) = -(y_i - R_k(\mathbf{x}))^2, k = 1, \dots K. \tag{5}$$

These feature functions follow the basic principle for association potentials in that their values are large when predictions and outputs are similar. To model the correlation among outputs, we use the quadratic feature function

$$\begin{aligned} g_l(y_i, y_j, \mathbf{x}) &= -e_l(i, j, \mathbf{x})(y_i - y_j)^2, \\ e_l(i, j, \mathbf{x}) &= \begin{cases} w_l(i, j, \mathbf{x}),\ (i, j) \in G_l \\ \hfill 0,\ (i, j) \notin G_l, \end{cases} \end{aligned} \tag{6}$$

which imposes that outputs y_i and y_j have similar values if they are connected by an edge in the graph G_l. $w_l(i, j, \mathbf{x})$ represents the weight of an edge (i, j) in graph G_l. It should be noted that using multiple graphs G_l can facilitate modeling of different aspects of correlation between outputs (for example, spatial and temporal).

2.2 Multivariate Gaussian model

Conditional distribution $P(\mathbf{y}|\mathbf{x})$ for the CRF model in Eq. (1), which uses quadratic feature functions defined in the previous section, can be represented as a multivariate Gaussian distribution. The resulting energy function of the GCRF model can be written as

$$\phi = -\sum_{i=1}^{N}\sum_{k=1}^{K}\alpha_k(y_i - R_k(\mathbf{x}))^2 - \sum_{i,j}\sum_{l=1}^{L}\beta_l e_l(i,j,\mathbf{x})(y_i - y_j)^2. \tag{7}$$

The energy function is a quadratic function in terms of $\mathbf{y}$. Therefore, $P(\mathbf{y}|\mathbf{x})$ can be transformed to a Gaussian form by representing ϕ as

$$\phi = -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}). \tag{8}$$

To transform $P(\mathbf{y}|\mathbf{x})$ to Gaussian form we determine $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ by matching Eq. (7) and (8)

$$\boldsymbol{\Sigma}^{-1}_{i,j} = 2\begin{cases} \sum_{k=1}^{K}\alpha_k + \sum_{n=1,n\neq j}^{N}\sum_l \beta_l e_l(i,n,\mathbf{x}),\ i=j \\ \qquad\qquad\qquad\qquad -\sum_l \beta_l e_l(i,j,\mathbf{x}),\ i\neq j, \end{cases} \tag{9}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{b}, \tag{10}$$

where $\mathbf{b}$ is a vector with elements

$$b_i = 2\sum_{k=1}^{K}\alpha_k R_k(\mathbf{x}). \tag{11}$$

If we calculate Z using the transformed exponent, we obtain

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}}e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}. \tag{12}$$

Therefore, the resulting conditional distribution is Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. We observe that $\boldsymbol{\Sigma}$ is a function of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and interaction potential graphs G_l, while $\boldsymbol{\mu}$ is also a function of inputs $\mathbf{x}$. The resulting CRF is the Gaussian CRF (GCRF). In order for the model to be feasible, the conditional distribution has to be well defined. This means that we have to ensure that the precision matrix $\boldsymbol{\Sigma}^{-1}$ is positive semi-definite [1], which we will address in the following sections.

3 Neural Gaussian CRF

In this section we propose a new Neural Gaussian CRF model, which enhances GCRF and increases its representational power.

3.1 Neural GCRF Model

First, motivated by the recently proposed Conditional Neural Fields [20, 5], we introduce the adaptive feature function defined as

$$f_a(y_i, \mathbf{x}) = -(y_i - R_a(\mathbf{w}, \mathbf{x}))^2, \tag{13}$$

where $R_a(\mathbf{w}, \mathbf{x})$ is a function of weights $\mathbf{w}$ that can be trained simultaneously with other GCRF parameters. In this way, $R_a(\mathbf{w}, \mathbf{x})$ can be trained directly with the goal of maximizing the log-likelihood such that it complements the existing predictors R_k. In this paper, we will assume that predictor $R_a(\mathbf{w}, \mathbf{x})$ is a feedforward neural network.

Second, as defined in Eq. (4), Gaussian CRF assigns weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to the feature functions. Considering that feature functions for the association potential are defined as squared errors of unstructured predictors, the role of weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is to measure their prediction uncertainty. Since it is likely that the quality of different predictors changes with $\mathbf{x}$, we enhance GCRF such that parameters α_k and β_l are replaced with the uncertainty functions $\alpha_k(\boldsymbol{\theta_k}, \mathbf{x})$ and $\beta_l(\boldsymbol{\psi_l}, \mathbf{x})$, where $\boldsymbol{\theta_k}$ and $\boldsymbol{\psi_l}$ are the parameters. We allow using feedforward neural networks for the uncertainty functions. By using the adaptive feature and uncertainty functions, we have

$$\begin{aligned} A(\boldsymbol{\theta}, y_i, \mathbf{x}) &= -\sum_{k=1}^{K} \alpha_k(\boldsymbol{\theta}_k, \mathbf{x})(y_i - R_k(\mathbf{x}))^2 - \alpha_a(\boldsymbol{\theta}_a, \mathbf{x})(y_i - R_a(\mathbf{w}, \mathbf{x}))^2, \\ I(\boldsymbol{\psi}, y_i, y_j, \mathbf{x}) &= -\sum_{l=1}^{L} \beta_l(\boldsymbol{\psi_l}, \mathbf{x})(y_i - y_j)^2. \end{aligned} \tag{14}$$

In this way, $\alpha_k(\boldsymbol{\theta_k}, \mathbf{x})$ models the varying degree of importance of predictor R_k over different conditions. Similarly, $\beta_l(\boldsymbol{\psi_l}, \mathbf{x})$ models varying importance of correlation between outputs. As a result, $\boldsymbol{\Sigma}$ from Eq. (9) becomes dependent on inputs, thus allowing for error heteroscedasticity. Conditional distribution of the enhanced GCRF is Gaussian as in Eq. (12). Since both adaptive feature and uncertainty functions are assumed to be feedforward neural networks, we call the resulting model the Neural GCRF (NGCRF).

Let us analyze the feasibility condition for the NGCRF model. In order for the model to be feasible, the precision matrix $\boldsymbol{\Sigma}^{-1}$ has to be positive semi-definite. A common approach used in practice [21] is to enforce sufficient condition given by Gershgorin's circle theorem [22], which says that a symmetric matrix is positive definite if all diagonal elements are non-negative and if the matrix is diagonally dominant.

Definition 1. *A square matrix $\boldsymbol{\Sigma}^{-1}$ is diagonally dominant if the absolute value of each diagonal element is greater than the sum of absolute values of the non-diagonal elements in corresponding row $|\boldsymbol{\Sigma}^{-1}_{i,i}| > \sum_{j \neq i} |\boldsymbol{\Sigma}^{-1}_{i,j}|, \forall i$.*

Theorem 1. *If the values of functions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in Eq* (14) *are always greater than 0, then the precision matrix $\boldsymbol{\Sigma}^{-1}$ that corresponds to NGCRF model defined by association and interaction potentials in Eq.* (14) *is diagonally dominant and hence positive definite.*

Proof. For each i, the absolute value of a diagonal element $\boldsymbol{\Sigma}_{i,i}^{-1}$ of precision matrix $\boldsymbol{\Sigma}^{-1}$ can be represented as

$$\begin{aligned} |\boldsymbol{\Sigma}_{i,i}^{-1}| =& |\sum_{k=1}^{K} \alpha_k(\boldsymbol{\theta}_k, \mathbf{x}) + \sum_{j \neq i} \sum_{l=1}^{L} \beta_l(\boldsymbol{\psi_l}, \mathbf{x})| \\ =& \sum_{k=1}^{K} \alpha_k(\boldsymbol{\theta}_k, \mathbf{x}) + \sum_{j \neq i} \sum_{l=1}^{L} \beta_l(\boldsymbol{\psi_l}, \mathbf{x}), \end{aligned} \tag{15}$$

where we use the fact that values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are always greater than 0. Similarly, the absolute value of each off-diagonal element $\boldsymbol{\Sigma}_{i,j}^{-1}$ equals

$$|\boldsymbol{\Sigma}_{i,j}^{-1}| = |\sum_{l=1}^{L} \beta_l(\boldsymbol{\psi_l}, \mathbf{x})| = \sum_{l=1}^{L} \beta_l(\boldsymbol{\psi_l}, \mathbf{x}). \tag{16}$$

Then, for each i we have

$$|\boldsymbol{\Sigma}_{i,i}^{-1}| - \sum_{j \neq i} |\boldsymbol{\Sigma}_{i,j}^{-1}| = \sum_{k=1}^{K} \alpha_k(\boldsymbol{\theta}_k, \mathbf{x}) > 0. \tag{17}$$

which proves the theorem. □

Therefore, one way to ensure that the NGCRF model is feasible is to impose the constraints $\boldsymbol{\alpha} > 0$ and $\boldsymbol{\beta} > 0$, which is analytically tractable [8, 1], but is known to be conservative [21]. To analyze the effect of constraining $\boldsymbol{\alpha} > 0$, we will assume that the interaction potential is not used (output variables are assumed to be conditionally independent). The prediction for each y_i becomes a weighted average of the unstructured predictors, where weights are positive values with their sum equal to 1. This constrains the range of outputs to $y_i \in [min(R_k(\mathbf{x})), max(R_k(\mathbf{x}))]$, which has negligible effect on NGCRF since we assumed that unstructured predictors are unbiased. In [21] it was empirically verified that constraint $\boldsymbol{\beta} > 0$ reduces parameter search space more and more with decreasing sparsity and increasing number of parameters in beta functions. This leads to limited improvements when using NGCRF with constraint $\boldsymbol{\beta} > 0$ on more dense graphs.

3.2 Learning and Inference of NGCRF

Learning The learning task is to choose values of parameters $\boldsymbol{\theta}$, $\boldsymbol{\psi}$ and $\mathbf{w}$ to maximize the conditional log-likelihood on the set of training examples $\mathcal{D} = \{(\mathbf{x_t}, \mathbf{y_t}), t = 1 \ldots T\}$

$$\begin{gathered} (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}, \hat{\mathbf{w}}) = \underset{\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{w}}{\operatorname{argmax}}(\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{w})) \\ \text{where } \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{w}) = \sum_{t=1}^{T} \log P(\mathbf{y_t} | \mathbf{x_t}). \end{gathered} \tag{18}$$

By setting $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to be greater than 0, learning becomes a constrained optimization problem. To convert it to unconstrained optimization, we adopt a technique used in [8, 1] that applies the exponential transformation of the functions to guarantee that their values are positive. We apply an exponential transformation on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

$$\begin{aligned} \alpha_k =& e^{u_k(\boldsymbol{\theta_k}, \mathbf{x})}, \text{ for } k = 1, \dots K, \\ \alpha_a =& e^{u_a(\boldsymbol{\theta_a}, \mathbf{x})}, \\ \beta_l =& e^{v_l(\boldsymbol{\psi_l}, \mathbf{x})}, \text{ for } l = 1, \dots L. \end{aligned} \tag{19}$$

where u_k and v_l are differentiable functions with respect to parameters $\boldsymbol{\theta_k}$ and $\boldsymbol{\psi_l}$.

All the parameters are learned by a gradient-based optimization. To apply the gradient-based method for learning, we need to find the gradient of the conditional log-likelihood. The derivatives of $\mathcal{L}$ with respect to θ, ψ, and $\mathbf{w}$ are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta_k} =& \frac{\partial \mathcal{L}}{\partial \alpha_k} \frac{\partial \alpha_k}{\partial u_k} \frac{\partial u_k}{\partial \theta_k}, \\ \frac{\partial \mathcal{L}}{\partial \psi_l} =& \frac{\partial \mathcal{L}}{\partial \beta_l} \frac{\partial \beta_l}{\partial v_l} \frac{\partial v_l}{\partial \psi_l}, \\ \frac{\partial \mathcal{L}}{\partial \mathbf{w}} =& \frac{\partial \mathcal{L}}{\partial R_a} \frac{\partial R_a}{\partial \mathbf{w}}. \end{aligned} \tag{20}$$

The gradient of $\mathcal{L}$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ has three components. The first components are $\partial \mathcal{L} / \partial \alpha_k$ and $\partial \mathcal{L} / \partial \beta_l$. The expression for $\partial \mathcal{L} / \partial \alpha_k$ is

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \alpha_k} =& -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \alpha_k} (\mathbf{y} - \boldsymbol{\mu}) + (\frac{\partial \mathbf{b}^T}{\partial \alpha_k} - \boldsymbol{\mu}^T \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \alpha_k})(\mathbf{y} - \boldsymbol{\mu}) \\ &+ \frac{1}{2} Tr(\boldsymbol{\Sigma} \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \alpha_k}). \end{aligned} \tag{21}$$

To calculate $\partial \mathcal{L} / \partial \beta_l$, we use $\partial \mathbf{b} / \partial \beta_l = 0$ and obtain

$$\frac{\partial \mathcal{L}}{\partial \beta_l} = -\frac{1}{2} (\mathbf{y} + \boldsymbol{\mu})^T \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \beta_l} (\mathbf{y} - \boldsymbol{\mu}) + \frac{1}{2} Tr(\boldsymbol{\Sigma} \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \beta_l}). \tag{22}$$

From Eq. (19), the second components are $\partial \alpha_k / \partial u_k = \alpha_k$ and $\partial \beta_l / \partial v_l = \beta_l$. The third components depend on the chosen functions u_k and v_l. The gradient of $\mathcal{L}$ with respect to $\mathbf{w}$ depends on the functional form of R_a. Since $\boldsymbol{\Sigma}^{-1}$ does not depend on R_a, $\partial \mathcal{L} / \partial R_a$ becomes

$$\frac{\partial \mathcal{L}}{\partial R_a} = 2 \boldsymbol{\alpha_a}^T (\mathbf{y} - \boldsymbol{\mu}). \tag{23}$$

We observe that an update for the adaptive model R_a is proportional to the difference between true output and the mean of the NGCRF model. This means that R_a will be updated only if NGCRF is not able to predict the output correctly

Algorithm 1 Learning of NGCRF Parameters

```
Input: $\mathbf{x}$, $R_k(\mathbf{x})$, $\mathbf{y}$.
1. Initialize $\boldsymbol{\theta_k}$, $\boldsymbol{\psi_l}$.
2. Estimate $\boldsymbol{\theta_k}$, $\boldsymbol{\psi_l}$ by applying gradient based approach and Eq. (21) and (22),
without taking into account $R_a$.
3. Initialize $\boldsymbol{\theta_a}$.
4. Learn predictor $R_a$ using Eq. (23).
repeat
    Apply gradient based optimization to estimate all parameters.
until Convergence
```

and R_a will be updated more aggressively when the error is larger. This justifies our hypothesis that R_a will work as a complement of the existing non-structured models.

To ensure convergence, the iterative procedure presented in Algorithm 1 [23, 20] is used for learning model parameters according to update formulas derived earlier in this section. To avoid overfitting, which is a common problem for maximum likelihood optimization, we added regularization terms for $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, β, $\boldsymbol{\psi}$ to the log-likelihood. In this way, we penalize large outputs of $\boldsymbol{\alpha}$ and β as well as large weights $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$.

Inference The inference task is to find the outputs $\mathbf{y}$ for a given set of observations $\mathbf{x}$ and estimated parameters $\hat{\boldsymbol{\alpha}}$ and $\hat{\beta}$ such that the conditional probability $P(\mathbf{y}|\mathbf{x})$ is maximized. The NGCRF model is Gaussian and, therefore, the maximum a posteriori estimate of $\mathbf{y}$ is obtained as the expected value $\boldsymbol{\mu}$ of the NGCRF distribution

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x}) = \boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{b}, \tag{24}$$

while $\mathbf{\Sigma}$ is a measure of uncertainty of the point estimate.

3.3 Complexity

If the size of the training set is N and the learning takes I iterations, the straightforward matrix computation results in $\mathcal{O}(IN^3)$ time to train the model. The main cost of computation is matrix inversion, since during the gradient-based optimization we need to find $\boldsymbol{\Sigma}$ as an inverse of $\boldsymbol{\Sigma}^{-1}$. However, this is the worst case performance. Since matrix $\boldsymbol{\Sigma}^{-1}$ is typically very sparse (it depends on the imposed neighborhood structure), the training time can be decreased to $\mathcal{O}(IN^2)$ by using sparse matrix apparatus or even to $\mathcal{O}(IN)$ if we do not consider interaction potential [21]. During inference, we need to compute $\boldsymbol{\mu}$, which takes $\mathcal{O}(N)$ time. As we eventually need to calculate the trace of the matrix, only the elements that correspond to the main diagonal should be stored. Therefore, memory requirements depend only on the imposed neighborhood structure.

4 Experiments

To demonstrate the strength of the NGCRF model, we applied it on two real-world structured regression applications. The experimental results indicate that NGCRF improves prediction accuracy by efficiently utilizing information from structured data.

4.1 The NGCRF Model for Document Retrieval

In this application the objective is to retrieving the most relevant documents with respect to the given query. In order to make a comparison to the GCRF method, we replicated the experimental setup from [8]. We obtained query-document data from OHSUMED dataset from LETOR [24], which is a standard data source used in document retrieval research (the same dataset was used in [8]). The OHSUMED dataset contains search queries, where each query is associated with a number of relevant documents. There are 106 queries, 348,566 documents and a total of 16,140 query-"relevant document" pairs. From the NGCRF perspective, each query-"set of relevant documents" represents an example $(\mathbf{x}, \mathbf{y})$. Each component of $\mathbf{y}$ represents a relevance of the corresponding document to a query, while $\mathbf{x}$ contains extracted features. Features $\mathbf{x}$ were used to construct $K = 25$ unstructured predictors $R_k(\mathbf{x})$ that predict document relevance for a given query. The outputs of unstructured predictors are available in OHSUMED (more details are in [24]). OHSUMED considers three levels of relevance - highly, partially and not relevant (each component in $\mathbf{y}$ can take values 2, 1, or 0 respectively). In addition, OHSUMED contains information about similarity between documents i and j, $w(i, j, \mathbf{x})$, which was determined based on similarity of their contents. Having this setup, the goal is to estimate relevance of each document in the database for a given query.

Benchmark methods As benchmark methods we use the following (all parameters were set using a small validation set)

Unstructured retrieval by neural network (NN) We trained NN with five hidden units to predict relevance of documents for a given query. The inputs to NN were outputs of unstructured predictors.

Structured retrieval by baseline GCRF We trained GCRF to predict relevance of documents. As unstructured predictors we used R_k, which are readily available in OHSUMED. GCRF also utilized relationship among documents by incorporating weights $w(i, j, \mathbf{x})$ from OHSUMED into the interaction potential.

Structured retrieval by GCRF+NN We trained a GCRF model using unstructured predictors R_k from OHSUMED and pre-trained NN. We call this model GCRF+NN.

RankSVM State-of-the-art retrieval method [25], which predictions are available as a part of OHSUMED.

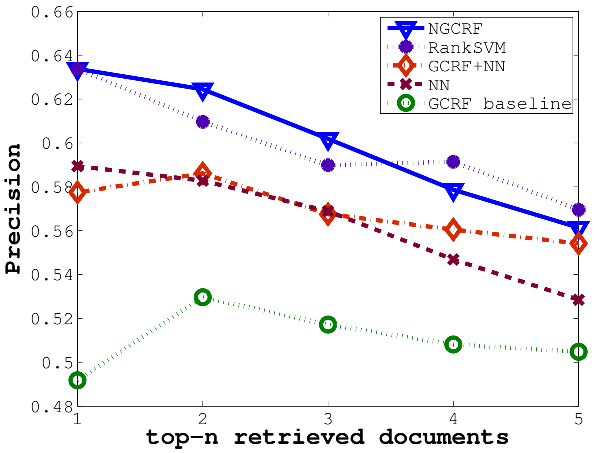

Fig. 1. Comparison of retrieval performance in terms of precision when top-n documents are retrieved.

The NGCRF model We trained the NGCRF model where unstructured predictors were R_k, $\boldsymbol{\alpha}$ was a function of unstructured predictors, $\boldsymbol{\beta}$ a function of similarity between documents, and adaptive NN was a function of R_k.

Evaluation In our experiments, for each method we averaged results over 5 fold cross validation data sets provided in OHSUMED. As an evaluation measure, we used *precision*@n, which represents a percentage of relevant publications in top-n publications retrieved ($n = 1 \ldots 5$ in our experiments). To fetch top-n relevant publications we retrieved those publications which corresponded to the n largest predictions. In Figure 1 we see that NN and GCRF+NN outperform baseline GCRF, which can be explained by the ability of NN to capture nonlinearity in feature space. Furthermore, if we allow NN to be adaptive, we see that NGCRF outperforms all other alternatives. We see that NGCRF is comparable to state-of-the-art retrieval method RankSVM, which is specifically designed for ranking problems (while NGCRF has general applicability) and which also used R_k and $w(i, j, \mathbf{x})$ as its inputs.

4.2 The NGCRF Model for AOD Prediction

We evaluated the proposed Neural GCRF model on a high impact regression problem from remote sensing, namely, prediction of aerosol optical depth (AOD) in the atmosphere from satellite measurements. AOD is a measure of aerosol light extinction integrated vertically through the atmosphere. AOD prediction is important because one of the main challenges of today's climate research is to characterize and quantify the effect of aerosols on Earth's radiation budget [26].

We considered data from MODIS, an instrument aboard NASA's Terra satellites [27]. We used ground-based data obtained from the AERONET [28], which is a global remote sensing network of radiometers that measure AOD several times per hour at specific geographic locations. The data can be obtained from the official MODIS website of NASA [29].

We extracted satellite-based attributes that are used as inputs to domain-based deterministic prediction algorithms [27]. In addition, we extracted information about the location of each data point (longitude and latitude) and a quality of observation (QA) assigned to each point provided by domain scientist. Data quality index was provided at four levels from the lowest quality QA=0 to the highest quality QA=3. We collected 28,374 data points distributed over the entire globe at 217 AERONET sites during the years 2005 and 2006.

Benchmark methods Here we list benchmark methods that we compared NGCRF to.

Deterministic prediction algorithm C005 The primary benchmark for comparison with our CRF predictors was the most recent version of the MODIS operational algorithm, called C005 [27]. This is a deterministic algorithm that retrieves AOD from MODIS observations relying on the domain knowledge. It is based on the inversion of physical forward models developed by the domain scientists.

Statistical prediction by a neural network As a baseline statistical algorithm we used a neural network model trained to predict AERONET AOD from all MODIS attributes excluding location and quality flag. It has been shown previously that neural networks achieve comparable accuracy to C005 on the AOD prediction problem [30]. The neural network has a hidden layer with 10 nodes and an output layer with one node. In nested 5-cross-validation experiments we trained 5 neural networks. When tested on 2006 data, we used a single network trained on the entire training set.

Structured prediction by GCRF The aerosol data are characterized by strong spatial and temporal dependencies that a CRF is able to exploit by defining interactions among outputs using feature functions. Given a data set that consists of satellite observations and ground-based AOD measurements, a statistical prediction model (R_a) can be trained to use satellite observations as attributes and predict the labels which are ground-based AODs. The deterministic AOD prediction models (DP) are based on solid physical principles and tuned by domain scientists. To model the association potential, i.e the dependency between the predictions and output AOD, we introduce the following two feature functions,

$$\begin{aligned} f_1(y_i, \mathbf{x}_i) &= -\left(y_i - DP(\mathbf{x}_i)\right)^2, \\ f_a(y_i, \mathbf{x}_i) &= -\left(y_i - R_a(\mathbf{x}_i)\right)^2. \end{aligned} \tag{25}$$

To model the interaction potential we introduce feature function

$$g_1(y_i, y_j, \mathbf{x}) = -(y_i - y_j)^2. \tag{26}$$

Table 1. RMSE and FRAC of C005, NN, GCRF and NGCRF on data with four quality flags.

	C005	NN	GCRF+NN	NGCRF
RMSE	0.123	0.112 ± 0.002	0.105 ± 0.0006	0.102 ± 0.0008
FRAC	0.65	0.68 ± 0.03	0.71 ± 0.005	0.74 ± 0.007

This interaction potential will reflect correlation between spatio-temporal data examples i and j (closer examples are given larger weight). The learned parameter β represents the level of spatio-temporal correlation of neighboring outputs (large β indicates that spatio-temporal correlation is large). We partitioned data into four subsets corresponding to quality flags QA=0, 1, 2, and 3. We determined eight a parameters corresponding to C005 and NN predictions over these subsets. To model interaction potential we defined spatial-temporal neighbors as a pair of observations where temporal distance $temporalDist(i, j)$ is less than 7 days and spatial distance $spatialDist(i, j)$ is less than 50km. This choice is based on previous studies of aerosol dynamics by geoscientists. We multiply feature g with weights $w(i, j, \mathbf{x})$, that are products of Gaussians

$$w(i,j,\mathbf{x}) = \begin{cases} e^{-\frac{spatialDist(i,j)^2}{2\sigma_s^2} - \frac{temporalDist(i,j)^2}{2\sigma_t^2}} & , i \sim j \\ & 0, otherwise \end{cases} \tag{27}$$

where $\sigma_s = 50$ and $\sigma_t = 10$ were determined using a small validation set.

The NGCRF model Here we use similar attributes as in the previous section but in the spirit of the proposed NGCRF model. Instead of defining manual partitions of the dataset, we use all observations as inputs to the $\boldsymbol{\alpha}$ functions. We define $\boldsymbol{\alpha}$ as an exponential function of linear combinations of observations. To incorporate potential bias, one observation is a vector with all ones.

$$\alpha_k(\boldsymbol{\theta}, \mathbf{x}^{(i)}) = e^{\sum \theta_t x_t^{(i)}}, \tag{28}$$

where $\mathbf{x}_1^{(i)}$ is a vector with all ones, $\mathbf{x}_{2,3,4,5}^{(i)}$ are quality flags. As an adaptive model R_a we used NN defined in previous sections. Its weight α_a follows the definition in Eq. (28).

To model spatio-temporal correlation, we use spatial and temporal distance between i and j as two observations for the β function. Similar to Eq. (28) we define β as

$$\beta(\boldsymbol{\psi}, \mathbf{x}^{(i,j)}) = e^{\sum \psi_l x_l^{(i,j)}}, \tag{29}$$

where $\mathbf{x}_1^{(i,j)}$ is a vector with all ones, $\mathbf{x}_2^{(i,j)}$ represents spatial distance between i and j and $\mathbf{x}_3^{(i,j)}$ represents their temporal proximity.

Evaluation To evaluate proposed methods, we trained the models on 2005 data and used 2006 data for testing. There are many possible measures that could be used to assess AOD prediction accuracy. Given vector $\mathbf{t} = (t_1, t_2, \dots t_N)^T$ of N outcome values and vector $\mathbf{y} = (y_1, y_2, \dots y_N)^T$ of the corresponding predictions, we measure the root mean squared error (RMSE). We also report accuracy on the domain specific measure called *the fraction of successful predictions* (FRAC) that penalizes errors on small AOD more than errors on large AOD [27]

$$FRAC = \frac{I}{N} \times 100\%, \qquad (30)$$

where I is the number of predictions that satisfy $|y_i - t_i| \leq 0.05 + 0.15t_i$.

RMSE error of the four models is presented in Table 1, where smaller numbers mean more accurate predictions. FRAC accuracy of these four models is also shown in Table 1, where larger numbers correspond to better predictions. We can see that in our experiments NN was more accurate than the operational C005 algorithm. GCRF showed an improvement in accuracy over both NN and C005 by taking advantage of a combination of models and spatio-temporal correlation in data. NGCRF achieves even better accuracy by utilizing nonlinear weights, an adaptive statistical model, and learning instead of assuming the level of correlation between points. Although NGCRF is a non-convex approach, it has only slightly larger variance in predictions than GCRF+NN.

The obtained results provide strong evidence that adaptive structured learning approaches can be successfully applied to AOD prediction, where even a small improvement of prediction accuracy results in huge uncertainty reduction in many geophysical studies that rely on AOD predictions [26].

5 Conclusion

Structured learning, as a fairly new research area in machine learning, has great success in classification, but its application on regression problems has not been explored sufficiently. In this article we proposed a method to adaptively combine the outputs of powerful non-structured regression models such as neural networks and a variety of correlated knowledge sources into a single prediction model by utilizing possible correlation among outcome variables. It is worth pointing to differences between our NGCRF model and the GCRF model proposed in [4]. The GCRF in [4] models a conditional distribution of pixels given a noisy input image using the weighted quadratic factors obtained by convolving the image with a set of filters. GCRF is designed for image de-noising problems, while NGCRF can be applied to general regression problems. By taking a closer look at GCRF we find that features in Eq. (5) and (6) are represented in GCRF, while GCRF does not model the adaptive component of NGCRF in Eq. (13). The proposed NGCRF is also readily applicable to other regression applications, where there is a need for knowledge integration and exploration of structure in outputs.

6 Acknowledgment

This work is supported in part by DARPA Grant FA9550-12-1-0406 negotiated by AFOSR, and NSF grant IIS-1117433.

References

1. Radosavljevic, V., Vucetic, S., Obradovic, Z.: Continuous conditional random fields for regression in remote sensing. In: European Conference on Artificial Intelligence (ECAI). (2010) 809–814
2. Solberg, A.H.S., Taxt, T., Jain, A.K.: A markov random field model for classification of multisource satellite imagery. IEEE Transactions on Geoscience and Remote Sensing **34**(1) (1996) 100–113
3. Kumar, S., Hebert, M.: Discriminative random fields: A discriminative framework for contextual interaction in classification. Proceedings Ninth IEEE International Conference on Computer Vision (2003) 1150–1157 vol.2
4. Tappen, M.F., Liu, C., Adelson, E.H., Freeman, W.T.: Learning gaussian conditional random fields for low-level vision. IEEE Conference on Computer Vision and Pattern Recognition (2007)
5. Peng, J., Bo, L., Xu, J.: Conditional neural fields. In: Advances in Neural Information Processing Systems 22. (2009) 1419–1427
6. Liu, Y., Carbonell, J., Klein-Seetharaman, J., Gopalakrishnan, V.: Comparison of probabilistic combination methods for protein secondary structure prediction. Bioinformatics **20**(17) (2004) 3099–3107
7. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings International Conference on Machine Learning. (2001)
8. Qin, T., Liu, T., Zhang, X., Wang, D., Li, H.: Global ranking using continuous conditional random fields. Neural Information Processing Systems (2008)
9. Grbovic, M., Vucetic, S.: Tracking concept change with incremental boosting by minimization of the evolving exponential loss. In: ECML/PKDD (1)'11. (2011) 516–532
10. Wytock, M., Kolter, Z.: Sparse gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In Dasgupta, S., Mcallester, D., eds.: Proceedings of the 30th International Conference on Machine Learning (ICML-13). Volume 28., JMLR Workshop and Conference Proceedings (May 2013) 1265–1273
11. Xiong, C., Wang, T., Ding, W., Shen, Y., Liu, T.Y.: Relational click prediction for sponsored search. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. WSDM '12, New York, NY, USA, ACM (2012) 493–502
12. Grbovic, M., Li, W., Xu, P., Usadi, A.K., Song, L., Vucetic, S.: Decentralized fault detection and diagnosis via sparse {PCA} based decomposition and maximum entropy decision fusion. Journal of Process Control **22**(4) (2012) 738 – 750
13. Djuric, N., Radosavljevic, V., Coric, V., Vucetic, S.: Travel speed forecasting by means of continuous conditional random fields. Transportation Research Record (2263) (2011) 131–139
14. Neville, J., Gallagher, B., Eliassi-Rad, T., Wang, T.: Correcting evaluation bias of relational classifiers with network crossvalidation. Knowledge and Information Systems **30** (2012) 31–55

15. Kim, M., Pavlovic, V.: Discriminative learning for dynamic state prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(10) (2009) 1847–1861
16. Wytock, M., Kolter, J.: Large-scale probabilistic forecasting in energy systems using sparse gaussian conditional random fields. In: Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on. (Dec 2013) 1019–1024
17. Guo, H.: Modeling short-term energy load with continuous conditional random fields. In Blockeel, H., Kersting, K., Nijssen, S., Zelezn, F., eds.: ECML/PKDD (1). Volume 8188 of Lecture Notes in Computer Science., Springer (2013) 433–448
18. Baltrušaitis, T., Banda, N., Robinson, P.: Dimensional affect recognition using continuous conditional random fields. In: IEEE Conference on Automatic Face and Gesture Recognition. (2013)
19. Ristovski, K., Radosavljevic, V., Vucetic, S., Obradovic, Z.: Continuous conditional random fields for efficient regression in large fully connected graphs. In desJardins, M., Littman, M.L., eds.: AAAI, AAAI Press (2013)
20. Do, T.M.T., Artieres, T.: Neural conditional random fields. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Volume 9., JMLR (5 2010)
21. Rue, H., Held, L.: Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis (2005)
22. Gerschgorin, S.: Uber die abgrenzung der eigenwerte einer matrix. Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk **7** (1931) 749–754
23. Nix, D.A., Weigend, A.S.: Learning local error bars for nonlinear regression. In Tesauro, G., Touretzky, D.S., Leen, T.K., eds.: Advances in Neural Information Processing Systems 7. Cambridge, MA: MIT Press (1995) 489–495
24. Liu, T.Y., Xu, J., Qin, T., Xiong, W., Li, H.: Letor: Benchmark dataset for research on learning to rank for information retrieval. In: SIGIR '07: Proceedings of the Learning to Rank workshop in the 30th annual international ACM SIGIR conference on Research and development in information retrieval. (2007)
25. Joachims, T.: Optimizing search engines using clickthrough data. In: KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, ACM (2002)
26. Kaufman, Y.J., Tanre, D., Boucher, O.: A satellite view of aerosols in the climate system. Nature **419**(6903) (2002) 215–223
27. Remer, L.A., Kaufman, Y.: The modis aerosol algorithm, products and validation. Journal of the Atmospheric Sciences **62** (2005) 947–973
28. Holben, B.N., Eck, T.F.: Aeronet: A federated instrument network and data archive for aerosol characterization. Remote Sensing of Environment **66** (1998) 1–16
29. http://modis.gsfc.nasa.gov: Official modis website
30. Radosavljevic, V., Vucetic, S., Obradovic, Z.: A data-mining technique for aerosol retrieval across multiple accuracy measures. Geoscience and Remote Sensing Letters, IEEE **7**(2) (April 2010) 411–415