ECAI 2010
H. Coelho et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-606-5-809

Continuous Conditional Random Fields for Regression in Remote Sensing

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

Abstract. Conditional random fields (CRF) are widely used for predicting output variables that have some internal structure. Most of the CRF research has been done on structured classification where the outputs are discrete. In this study we propose a CRF probabilistic model for structured regression that uses multiple non-structured predictors as its features. We construct features as squared prediction errors and show that this results in a Gaussian predictor. Learning becomes a convex optimization problem leading to a global solution for a set of parameters. Inference can be conveniently conducted through matrix computation. Experimental results on the remote sensing problem of estimating Aerosol Optical Depth (AOD) provide strong evidence that the proposed CRF model successfully exploits the inherent spatio-temporal properties of AOD data. The experiments revealed that CRF are more accurate than the baseline neural network and domain-based predictors.

1 Introduction

A data mining approach for regression is based on learning relationships between attributes and the target variable. In the standard regression setting we are given a data set with N training examples, $D = (\mathbf{x}_i, y_i), i = 1 \dots N$, where $\mathbf{x}_i \in \mathbf{X} \subset R^M$ is an M dimensional vector of attributes and $y \in R$ is a real-valued target variable. The objective of regression is to learn a non-linear mapping f from training data D that predicts the output variable y as accurately as possible given an input vector $\mathbf{x}$.

Traditional supervised learning models, like neural networks (NN), are powerful tools for learning non-linear mappings. However, such models mainly focus on the prediction of a single output and could not exploit relationships that exist between multiple outputs. In structured learning, the model learns a mapping $f : \mathbf{X}^N \rightarrow R^N$ to simultaneously predict all outputs given all input vectors. For example, let us assume that the value of y_i is dependent on that of y_{i1} and y_{i+1}, as is the case in temporal data. Let us also assume that input $\mathbf{x}_i$ is noisy. A traditional model that uses only information contained in $\mathbf{x}_i$ to predict y_i might predict the value for y_i to be quite different from those of y_{i-1} and y_{i+1} because it treats them individually. A structured predictor uses dependencies among outputs to take into account that y_i is more likely to have value close to y_{i-1} and y_{i+1} thus improving final predictions. In structured learning we usually have some prior knowledge about relationships among the outputs y. Mostly, those relationships are application-specific where the dependencies are defined in advance, either by domain knowledge or by assumptions, and represented by statistical models.

[1] Center for Information Science and Technology, Temple University, Philadelphia, PA 19122, USA, email: {vladan, vucetic, zoran}@ist.temple.edu

Relationships among outputs can be represented by graphical models. In the case of spatial-temporal data, some popular models are the traditional Markov random fields [9] and the recently proposed Conditional Random Fields (CRF) [5]. Originally, CRF were designed for classification of sequential data [5]. Recently, it has found many applications in areas such as computer vision [4] and computational biology [6]. CRF for regression is a less explored topic. To address this gap, in this paper we build on the recently proposed Continuous CRF [7] and develop a solution that is applicable to regression on spatial-temporal data. We are particularly interested in applying the CRF for regression to a remote sensing problem of Aerosol Optical Depth (AOD) prediction.

Aerosols are minute particles suspended in the atmosphere originating from natural and man-made sources. AOD, as reported by surface and space-based passive remote sensing, is a measure of aerosol light extinction integrated vertically through the entire atmosphere. One of the biggest challenges of todays climate research is to characterize and quantify the effect of aerosols on Earths radiation budget.

The operational AOD prediction algorithms are deterministic typically manually tuned by domain scientists. In contrast to domain-driven methods, the AOD prediction is achievable by a completely data-driven approach. This statistical method consists of training a nonlinear regression model to predict AOD using the satellite observations as inputs. The targets are obtained from a network of unevenly distributed ground-based sites over the world (shown at Figure 1). This approach is possible when a data set is available that consists of satellite observations and collocated ground-truth measurements. A property of statistical prediction is that its high accuracy is guaranteed only for the conditions similar to those at the ground-based sites. However, some regions are underrepresented due to non-uniform distribution of ground-based sites over the world. The goal is to combine the two approaches so that in some regions where we know that deterministic algorithm performs better, we rely more on deterministic than on a statistical method, and vice versa.

The aerosol data are characterized by strong spatial and temporal dependencies that CRF is able to exploit by defining interactions among outputs using feature functions. The use of features to define the CRF models allows us also to include arbitrary properties of input-output pairs into the compatibility measure. In this study we propose CRF probabilistic model for structured regression that uses multiple non-structured predictors as its features. We construct features as squared prediction errors of deterministic and statistical models and show that this results in multivariate Gaussian conditional $P(\mathbf{y}|\mathbf{x})$ distribution. Learning becomes a convex optimization problem leading to a global solution for a set of parameters. Inference can be conveniently conducted through matrix computation. The performance of the proposed approach is compared to the baseline sta-

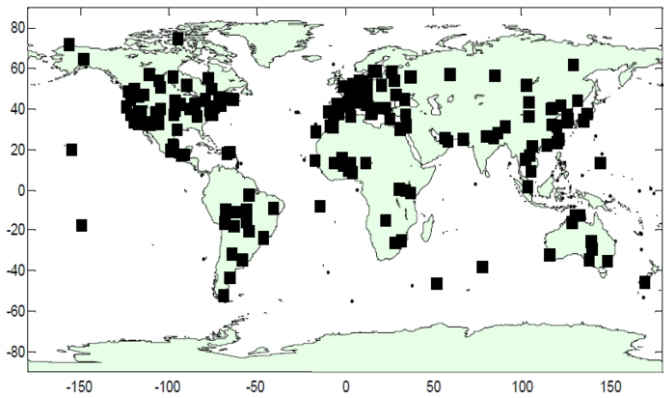

Figure 1. Global distribution of ground-based sites.

tistical and deterministic methods.

The rest of the paper is organized as follows. In Section 2, we introduce Continuous CRF. We then show how to apply Continuous CRF for AOD prediction in Section 3. We give experimental results in Section 4. Finally, we conclude the paper in Section 5.

2 CONTINUOUS CONDITIONAL RANDOM FIELDS

Conditional Random Fields (CRF) provide probabilistic framework for exploiting complex dependence structure among outputs by directly modeling the conditional distribution $P(\mathbf{y}|\mathbf{x})$. In regression problems, the output y_i is associated with input vectors $\mathbf{x} = (\mathbf{x}_1, \dots \mathbf{x}_N)$ by a real-valued function called association potential $A(\alpha, y_i, \mathbf{x})$, where α is K - dimensional set of parameters. The larger the value of A is the more y_i is related to $\mathbf{x}$. Usually, A is a combination of functions. We can use as many association functions as we find necessary to model input-output relations in data. In general, A takes as input all input data $\mathbf{x}$ to predict a single output yi meaning that it does not impose any independency relations among inputs $\mathbf{x}_i$.

To model interactions among outputs, a real valued function called interaction potential $I(\beta, y_i, y_j, \mathbf{x})$ is used, where β is an L dimensional set of parameters. Interaction potential represents the relationship between two outputs and in general can depend on an input $\mathbf{x}$. Different applications can have different interaction potentials. For example, in the AOD prediction problem, interaction potential can be modeled as a correlation between neighboring (in time and space) outputs. The larger the value of the interaction potential, the more related outputs are.

For the defined association and interaction potentials, CRF models a conditional distribution $P(\mathbf{y}|\mathbf{x})$, $\mathbf{y} = (y_1 \dots y_N)$, according to the associated graphical structure (an example of the structure is shown in Figure 2)

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x}, \alpha, \beta)} exp(\sum_{i=1}^{N} A(\alpha, y_i, \mathbf{x}) + \sum_{j \sim i} I(\beta, y_i, y_j, \mathbf{x})) \quad (1)$$

where $j \sim i$ denotes the connected outputs y_i and y_j (connected with solid line at Figure 2) and where $Z(\mathbf{x}, \alpha, \beta)$ is normalization

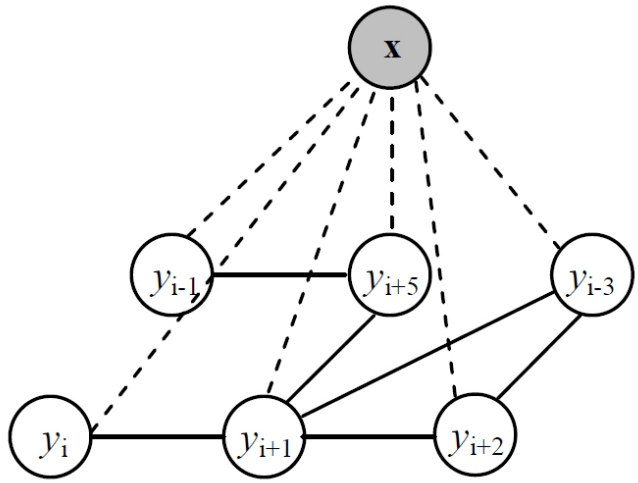

Figure 2. Continuous CRF graphical structure. x-inputs (observations); y-outputs; dashed lines-associations between inputs and outputs; solid lines-interactions between outputs.

function defined as

$$Z(\mathbf{x}, \alpha, \beta) = \int_y exp(\sum_{i=1}^{N} A(\alpha, y_i, \mathbf{x}) + \sum_{j \sim i} I(\beta, y_i, y_j, \mathbf{x})) \quad (2)$$

The learning task is to choose values of parameters α and β to maximize the conditional log-likelihood of the set of training examples

$$\begin{aligned} L(\alpha, \beta) &= \log P(\mathbf{y}|\mathbf{x}) \\ (\hat{\alpha}, \hat{\beta}) &= \underset{\alpha, \beta}{\operatorname{argmax}}(L(\alpha, \beta)) \end{aligned} \quad (3)$$

This can be achieved by applying standard optimization algorithms such as gradient descent. To avoid overfitting, we regularize $L(\alpha, \beta)$ by adding $\alpha^2/2$ and $\beta^2/2$ terms to formula (3) that prevents the parameters from becoming too large.

The inference task is to find the outputs $\mathbf{y}$ for a given set of observations $\mathbf{x}$ and estimated parameters α and β such that the conditional probability $P(\mathbf{y}|\mathbf{x})$ is maximized,

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x}) \quad (4)$$

The following dLearning and inference in models with real valued targets pose quite different challenges than in the discrete-valued case. The most important difference is that the normalizing function Z is an integral instead of the sum. Discrete valued models are always feasible as Z is a finite number 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 directly that Z is integrable might be difficult due to the complexity of association and interaction potentials.

In CRF applications, A and I could be defined as linear combinations of a set of fixed features in terms of α and β [5]

$$\begin{aligned} A(\alpha, y_i, \mathbf{x}) &= \textstyle\sum_{k=1}^{K} \alpha_k f_k(y_i, \mathbf{x}) \\ I(\beta, y_i, y_j, \mathbf{x}) &= \textstyle\sum_{l=1}^{L} \beta_l g_l(y_i, y_j, \mathbf{x}) \end{aligned} \quad (5)$$

The use of features to define the model is convenient because it allows us to include arbitrary properties of input-output pairs into the compatibility measure. This way, any potentially relevant feature

could be included to the model because parameter estimation automatically determines their actual relevance by feature weighting.

In general, to evaluate $P(\mathbf{y}|\mathbf{x})$ needed during training and inference, one would need to use time consuming sampling methods such as Markov Chain Monte Carlo-based algorithms. However, if A and I are defined as quadratic function in terms of $\mathbf{y}$, then the sum $A+I$ can be transformed to $(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)+const$. This expression corresponds to multivariate Gaussian distribution with mean μ and covariance $\mathbf{\Sigma}$. If this is the case, learning and inference are convenient. Learning becomes a convex optimization problem leading to a global solution for α and β. Inference can be conveniently conducted through matrix computation. For inference, given new observation $\mathbf{x}$, the output $\mathbf{y}$ is calculated as the conditional expectation $E(\mathbf{y}|\mathbf{x})$. By exploiting the sparsity inherent to spatio-temporal data, the inference can be performed in time linear with the number of spatio-temporal observations.

However, we need to make sure that $P(\mathbf{y}|\mathbf{x})$ is feasible conditional distribution. The condition for $P(\mathbf{y}|\mathbf{x})$ to be multivariate Gaussian is that normalization function Z is finite. Z defined in (2) is finite if covariance matrix $\mathbf{\Sigma}$ is positive semi-definite. Hence, when learning the parameters we have to imply constraint that covariance matrix $\mathbf{\Sigma}$ is positive semi-definite.

3 THE CCRF MODEL FOR AOD PREDICTION

In the following we describe in detail the proposed CRF for regression in remote sensing, using the AOD prediction as the motivating example. Given a data set that consists of satellite observations and ground based AOD measurements, a statistical prediction model (SP) 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 target AOD, we introduce two feature functions,

$$\begin{array}{l} f_1(y_i,\mathbf{x}_i) = -(y_i - SP(\mathbf{x}_i))^2 \\ f_2(y_i,\mathbf{x}_i) = -(y_i - DP(\mathbf{x}_i))^2 \end{array} \quad (6)$$

where for a given observation $\mathbf{x}_i$, $SP(\mathbf{x}_i)$ and $DP(\mathbf{x}_i)$ are outputs of statistical and deterministic models, respectively. These feature functions follow the basic principle for association potentials (their values are larger for more accurate predictions). Learned parameters α of the linear combination of these features,

$$A(\alpha, y_i, \mathbf{x}_i) = -\alpha_1(y_i - SP(\mathbf{x}_i))^2 - \alpha_2(y_i - DP(\mathbf{x}_i))^2 \quad (7)$$

provide some insight on how much to trust the SP and DP prediction algorithms. For example, large α_1 places large penalty on mistakes of SP model and is an indicator of large quality of this predictor.

To improve expressiveness of the CRF model we introduce various indicator functions. Here are some examples of possible indicator functions

$$\begin{array}{l} I_1(\mathbf{x}_i) = \left\{ \begin{array}{ll} 1, & \text{if } \mathbf{x}_i \text{ belongs to North America} \\ 0, & \text{otherwise} \end{array} \right. \\ I_2(\mathbf{x}_i) = \left\{ \begin{array}{ll} 1, & \text{if } \mathbf{x}_i \text{ is a data point of high quality} \\ 0, & \text{otherwise} \end{array} \right. \end{array} \quad (8)$$

Association potential now becomes

$$A(\alpha, y_i, \mathbf{x}_i) = \sum_{k=1}^{K/2} (\alpha_{2k-1} I_k(\mathbf{x}_i)(y_i - SP(\mathbf{x}_i))^2 + \alpha_{2k} I_k(\mathbf{x}_i)(y_i - DP(\mathbf{x}_i))^2) \quad (9)$$

By introducing indicator functions we essentially partition the whole data set into smaller subsets. Learned α represents our belief in SP and DP in different subsets, corresponding to different prediction conditions.

To model the interaction potential we introduce feature function

$$g_1(y_i, y_j, \mathbf{x}) = -w_{ij}(y_i - y_j)^2 \quad (10)$$

In AOD prediction problem data are irregularly sampled in both space and time. Weight w_{ij} is positive number representing a measure of spatio-temporal proximity between data points i and j (closer points are given larger weight). The corresponding interaction potential is

$$I(\beta, y_i, y_j, \mathbf{x}) = -\beta w_{ij}(y_i - y_j)^2 \quad (11)$$

Here, the learned parameter β represents the level of spatio-temporal correlation of neighboring outputs (large β indicates that spatio-temporal correlation is large).

Finally, the resulting CRF model is

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z} exp(-\sum_{i=1}^{N}\sum_{k=1}^{K/2}(\alpha_{2k-1} I_k(\mathbf{x}_i)(y_i - SP(\mathbf{x}_i))^2 - \alpha_{2k} I_k(\mathbf{x}_i)(y_i - DP(\mathbf{x}_i))^2) - \beta w_{ij}(y_i - y_j)^2) \quad (12)$$

In the following we show that (12) can be represented as multivariate Gaussian distribution. In (12), the exponent E is quadratic function in terms of y, and therefore $P(\mathbf{y}|\mathbf{x})$ can be transformed to a Gaussian form by representing E as

$$E = \frac{1}{2}(\mathbf{y}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\mu) = \mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} + \mathbf{y}^T\mathbf{\Sigma}^{-1}\mu + const \quad (13)$$

To transform $P(\mathbf{y}|\mathbf{x})$ to Gaussian form we need determine $\mathbf{\Sigma}$ and μ by matching (12) and (13). We will first represent quadratic terms of y in association and interaction potentials as -$\mathbf{y}^T\mathbf{Q}_1\mathbf{y}$ and -$\mathbf{y}^T\mathbf{Q}_2\mathbf{y}$ respectively. Then we will combine them to get $\mathbf{\Sigma}^{-1} = 2(\mathbf{Q}_1 + \mathbf{Q}_2)$. The quadratic term of $\mathbf{y}$ in association potential can be represented as -$\mathbf{y}^T\mathbf{Q}_1\mathbf{y}$, where $\mathbf{Q}_1$ is diagonal matrix with elements

$$\mathbf{Q}_{1ij} = \left\{ \begin{array}{ll} \sum_{k=1}^{K/2}(\alpha_{2k-1} I_k(\mathbf{x}_i) + \alpha_{2k} I_k(\mathbf{x}_i)), & i = j \\ 0, & i \neq j \end{array} \right. \quad (14)$$

The quadratic term of y in interaction potential is -$\mathbf{y}^T\mathbf{Q}_2\mathbf{y}$, where $\mathbf{Q}_2$ is symmetric matrix with elements

$$\mathbf{Q}_{2ij} = \left\{ \begin{array}{ll} \sum_j \beta w_{ij}, & i = j \\ -\beta w_{ij}, & i \neq j \end{array} \right. \quad (15)$$

To get μ we match terms linear in E to linear terms in exponent of (12). If we represent the linear terms of (12) as $\mathbf{y}^T\mathbf{b}$, then we get $\mu = \mathbf{\Sigma}\mathbf{b}$, where $\mathbf{b}$ is vector with elements

$$b_i = 2\sum_{k=1}^{K/2}(\alpha_{2k-1} I_k(\mathbf{x}_i) SP(\mathbf{x}_i) + \alpha_{2k} I_k(\mathbf{x}_i) DP(\mathbf{x}_i)) \quad (16)$$

If we calculate Z using transformed exponent, we will get

$$Z(\alpha, \beta, \mathbf{x}) = (2\pi)^{N/2}|\mathbf{\Sigma}|^{1/2} exp(const) \quad (17)$$

Exponent of $const$ term from Z and $P(\mathbf{y}|\mathbf{x})$ cancels out, so that we finally get

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\mathbf{\Sigma}|^{1/2}} exp(-\frac{1}{2}(\mathbf{y}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\mu)) \quad (18)$$

where $\mathbf{\Sigma}$ and μ are defined previously.

Let us analyze feasibility condition for this model. In order for the model to be feasible, covariance matrix $\mathbf{\Sigma}$ has to be positive semi-definite. We can analyze the equivalent that $\mathbf{\Sigma}^{-1}$ is positive semi-definite. $\mathbf{\Sigma}^{-1}$ is defined as a double sum of $\mathbf{Q}_1$ and $\mathbf{Q}_2$. $\mathbf{Q}_2$ is a symmetric matrix with a property that the absolute value of a diagonal element is equal to the sum of absolute values of non-diagonal elements from the same row

$$|\mathbf{Q}_{2ii}| = \sum_{j \neq i} |\mathbf{Q}_{2ij}| \tag{19}$$

By Gershgorin's circle theorem [1], a symmetric matrix is positive semi-definite if all diagonal elements are non-negative and if matrix is diagonally dominant. A matrix is diagonally dominant if for every row of the matrix, the value of the diagonal element in that row is larger than the sum of the absolute values of non-diagonal elements in that row. If we sum elements from diagonal matrix $\mathbf{Q}_1$ to matrix $\mathbf{Q}_2$, and if all α's and β are positive then the values on diagonal of $\mathbf{\Sigma}^{-1}$ will be non-negative and the matrix will be diagonally dominant. Therefore, to ensure that our model is feasible, we have to impose constraint that all parameters have to be greater than 0.

In this setting, learning is a constrained optimization problem because we need to guarantee that all $\alpha_k > 0$ and $\beta > 0$. Gradient ascent cannot be directly applied to a constrained optimization problem. Here we adopt a technique similar to that in [7] and then employ gradient ascent. Specifically, we maximize log-likelihood with respect to $\log \alpha_k$ and $\log \beta$ instead of α_k and β. As a result, the new optimization problem becomes unconstrained. Derivatives of log-likelihood function and updates of α's and β in gradient ascent can be computed as

$$\begin{aligned} \frac{\partial L}{\partial \log \alpha_k} &= \alpha_k \frac{\partial L}{\partial \alpha_k}, \log \alpha_k^{new} = \log \alpha_k^{old} + \eta \frac{\partial L}{\partial \log \alpha_k} \\ \frac{\partial L}{\partial \log \beta} &= \beta \frac{\partial L}{\partial \beta}, \log \beta^{new} = \log \beta^{old} + \eta \frac{\partial L}{\partial \log \beta} \end{aligned} \tag{20}$$

where η is is the learning rate.

In inference, since the model is Gaussian, the prediction will be expected value, which is equal to the mean μ of the distribution,

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x}) = \mathbf{\Sigma b} \tag{21}$$

4 EXPERIMENTS

4.1 Data sources and collocation

In this study we consider data from MODerate resolution Imaging Spectrometer (MODIS), an instrument aboard NASA's Terra and Aqua satellites [8]. Instruments mounted on Terra observe the Earth during morning whereas those mounted on Aqua observe the Earth during afternoon. In this study, we use data only from the Terra satellite. Ground-based data are obtained from the AErosol RObotic NETwork (AERONET) [2] which is a global remote sensing network of radiometers that measure AOD several times per hour from specific geographic locations.

MODIS has high spatial resolution (pixel is as small as 250x250 m^2) and achieves global coverage daily. On the other hand, AERONET sites, situated at fixed geographical locations, acquire data at intervals of 15 min on average. This gives rise to the need for both spatial and temporal data fusion. The fusion method involves aggregating MODIS pixels into blocks of size 50x50 km^2 and spatially collocating them with an AERONET site. The MODIS observations are said to be temporally collocated with the corresponding AERONET AOD predictions if there is a valid AERONET AOD prediction within 30 minutes of the satellite overpass. The data collocated in this way can be obtained from the official MODIS website of NASA [3].

There are several levels of AERONET AOD measurements [2]. To avoid potential problems with outliers in ground truth data, AERONET Level 2.0 observations were considered since they were cloud screened and manually verified.

For our study we collected MODIS Terra observations collocated with AERONET Level 2.0 points. We extracted satellite-based attributes that are used as inputs to knowledge-based prediction algorithms. The radiances at four wavelengths were taken from the MODIS range 440nm 2100nm, as these are sufficient to describe aerosol properties [8]. An average and standard deviation of radiances of pixels in 50x50 km^2 blocks were then estimated. Along with radiances we also extracted ancillary attributes. Information about geometry is characterized by solar and sensor angles. As surface elevation affects estimated AOD, it was also included in the set of attributes and has been extracted from AERONET data. 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. There are four levels of qualities from lowest quality QA=0 to highest quality QA=3.

By convention, AOD is reported at the 550nm wavelength. Since AERONET sites do not provide AOD value at that particular wavelength, we performed a standard linear interpolation in the log scale of AERONET AOD measurements at 440nm and 670nm to estimate AOD at 550nm [8]. We collected 28374 data points distributed over entire globe at 217 AERONET sites (Figure 1) during years 2005 and 2006.

4.2 Evaluation

To assess the efficiency of the proposed methods, we performed training on 2005 data and used 2006 data for testing. Because we jointly train NN on 2005 data and then its predictions as inputs to CRF, we applied a nested cross-validation. First, we split AERONET locations into 5 subsets and created five data sets $D_i, i = 1 \ldots 5$, each with data points from one of the AERONET subsets in year 2005. We reserved one of D_i datasets for testing and merged data from the remaining 4 datasets $D_j, j \neq i$, for training. The trained NN predictor was tested on D_i. The procedure was repeated five times, for values $j = 1 \ldots 5$. Finally, we get five NN models and NN predictions for all points in training set.

There are many possible measures that could be used to assess AOD prediction accuracy. Given vector $t = [t_1, t_2, \ldots t_N]$ of N target values and vector $y = [y_1, y_2, \ldots y_N]$ of the corresponding predictions, the

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - t_i)^2} \tag{22}$$

standard root mean squared error (RMSE) is defined as We also report accuracy on domain specific measure fraction of successful predictions (FRAC) that penalizes errors on small AOD more than errors on large AOD. The AOD prediction can be considered successful if the absolute error is

Table 1. RMSE and FRAC of C005, NN, and NN+CCRF using features defined over five regions, without ($\beta = 0$) and with spatio-temporal correlation ($\beta \neq 0$)

	RMSE				FRAC			
Region	C005	NN	CRF,$\beta = 0$	CRF,$\beta \neq 0$	C005	NN	CRF,$\beta = 0$	CRF,$\beta \neq 0$
Whole Globe	0.123	0.112±0.001	0.1067±0.0005	0.1056±0.0005	0.65	0.667±0.002	0.704±0.003	0.708±0.004
N. America	0.098	0.085±0.001	0.083±0.001	0.081±0.001	0.64	0.667±0.008	0.71±0.01	0.71±0.01
S. America	0.140	0.110±0.005	0.104±0.003	0.098±0.002	0.55	0.56±0.01	0.58±0.02	0.60±0.02
Europe	0.080	0.080±0.001	0.0736±0.0005	0.0728±0.0005	0.76	0.762±0.005	0.807±0.006	0.812±0.006
Africa	0.172	0.154±0.001	0.152±0.001	0.149±0.001	0.53	0.560±0.006	0.568±0.007	0.577±0.006
Asia&Aus.	0.161	0.156±0.001	0.145±0.001	0.148±0.001	0.64	0.66±0.01	0.71±0.01	0.70±0.01

Table 2. RMSE and FRAC of C005, NN, and NN+CCRF using features defined over four subsets of data of different quality (QA=0 lowest, QA=3 highest), without ($\beta = 0$) and with spatio-temporal correlation ($\beta \neq 0$)

	RMSE				FRAC			
Data Quality	C005	NN	CRF,$\beta = 0$	CRF,$\beta \neq 0$	C005	NN	CRF,$\beta = 0$	CRF,$\beta \neq 0$
Entire Set	0.123	0.112±0.001	0.1065±0.0005	0.105±0.0005	0.65	0.667±0.002	0.705±0.004	0.709±0.004
QA=0	0.151	0.128±0.002	0.123±0.001	0.121±0.001	0.59	0.60±0.01	0.64±0.01	0.64±0.01
QA=1	0.130	0.108±0.001	0.109±0.001	0.107±0.001	0.58	0.623±0.006	0.65±0.01	0.652±0.005
QA=2	0.118	0.110±0.002	0.104±0.001	0.101±0.001	0.64	0.65±0.01	0.686±0.007	0.689±0.007
QA=3	0.105	0.104±0.002	0.097±0.001	0.096±0.001	0.70	0.714±0.007	0.755±0.003	0.761±0.002

$$|y_i - t_i| \leq 0.05 + 0.15 t_i \tag{23}$$

We may now define the FRAC as

$$FRAC = \frac{I}{N} \times 100\% \tag{24}$$

where I is the number of predictions that satisfy relation (23).

4.3 Benchmark Methods

4.3.1 Deterministic prediction algorithm C005

The primary benchmark for comparison with our predictors was the most recent version of the MODIS deterministic algorithm called C005. The deterministic algorithms that retrieve AOD from MODIS observations rely on the domain knowledge of aerosol properties and are based on lookup tables representing the most common atmospheric conditions.

4.3.2 Statistical prediction by neural network

As a baseline statistical algorithm we used a neural network trained to predict AERONET AOD from all MODIS attributes except location information and quality flag. 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 whole training set.

4.4 The CRF model

4.4.1 Integration of models

We first consider the case when interaction potential does not exist ($\beta = 0$). NN and C005 predictions are inputs to CRF. We partitioned the world into five regions: North America, South America, Europe, Africa, and Asia and Australia. Asia and Australia were treated together due to the small number of data points in each of them. Then, we defined five indicator functions. Each function indicates belonging to one of five regions. We determined ten α parameters corresponding to C005 and NN predictions over these regions. Results are presented in Table 1. Over all regions CRF achieved better accuracy than either NN or C005 alone. Values of obtained α parameters suggest that we should trust NN more in the North America (ratio of α's is NN:C005=24:13 approximately) while in Africa we should trust C005 a little bit more (ratio of α's is NN:C005=8:9 approximately). Also, CRF improves domain-based accuracy measure FRAC (Table 1).

Second, we check how much we should rely on NN and C005 over observations with different qualities. We partitioned data into four subsets having quality flags QA=0, 1, 2, and 3. We introduced four indicator functions to indicate belonging to each of subsets. We determined eight α parameters corresponding to C005 and NN predictions over these subsets. Results are presented in Table 2. For all data qualities CRF achieved better accuracy than either NN or C005 alone. As expected, error of the deterministic predictor C005 decreases as data quality increases. Values of obtained α parameters also suggest that we should trust NN more for low data quality QA=0 (ratio of α's is NN:C005=21:10 approximately) while for high data quality we should equally trust to C005 and NN (ratio of α's is NN:C005=16:16 approximately). FRAC is also improved by CRF, Table 2.

4.4.2 Integration of spatio-temporally correlated models

Here we consider the case when interaction potential does exist ($\beta \neq 0$). NN and C005 predictions are inputs to CRF. To model interaction potential we need to define weights w_{ij} in (11). After analysis of spatial and temporal AOD autocorrelation (results not shown) we decided to define spatial-temporal neighbors as a pair of observations where temporal distance $temporalDist(i, j)$ is less than 60 days and spatial distance $spatialDist(i, j)$ is less than 100km. Therefore, we used weighted distance in defining w_{ij}, weights are multiplica-

tion of Gaussians

$$w_{ij} = \begin{cases} exp(-\frac{spatialDist(i,j)^2}{2\sigma_s^2} - \frac{temporalDist(i,j)^2}{2\sigma_t^2}, & i \sim j \\ 0, & otherwise \end{cases} \quad (25)$$

where $\sigma_s = 50$ and $\sigma_t = 10$ were determined empirically.

Taking into account spatio-temporal correlation and comparing to the CRF model with ($\beta = 0$) when the world was partitioned into five regions, we get better results globally and over all regions separately except Africa where two models were equally good and Asia&Australia where the latter model was better (Table 1). This result suggests that level of spatio-temporal correlation is different in different regions, and each region should have its own β. β was estimated to 0.049, which does not indicate significant correlation, but it is still enough to improve single-output based predictors.

Including spatial-temporal correlation in the model when data were partitioned based on quality also improves final prediction (Table 2), β was estimated to 0.06.

5 CONCLUSION

Structured learning is a new research area in machine learning that had great success in classification, but its application on regression problems has not been explored sufficiently. We proposed a method to combine the outputs of a powerful non-linear regression tool such as NN by incorporating a variety of correlated knowledge sources into single prediction model.

We reported the results on remote sensing application of predicting AOD from satellite-based observations. Presented results provide strong evidence that structured learning approaches can be successfully applied to not only the AOD prediction problem but also other remote sensing regression problems. Furthermore, the presented model can be applied to any regression application where there is a need for knowledge integration and exploration of structure in outputs.

REFERENCES

[1] S. Gerschgorin, 'Uber die abgrenzung der eigenwerte einer matrix', *Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk*, **7**, 749–754, (1931).
[2] Eck T. F. Slutsker I. Tanre T. Buis J. P. Setzer-A. Vermote E. Reagan J. A. Kaufman Y. J. Nakajima T. Lavenu F. Jankowiak I. Holben, B. N. and A. Smirnov, 'Aeronet: A federated instrument network and data archive for aerosol characterization', *Remote Sensing of Environment*, **66**, 1–16, (1998).
[3] http://modis.gsfc.nasa.gov. Official modis website.
[4] S. Kumar and M. Hebert, 'Discriminative random fields: A discriminative framework for contextual interaction in classification', *In Proceedings of ICCV*, 1150–1159, (2003).
[5] McCallum A. Lafferty, J. and F. Pereira, 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data', *In Proceedings of ICML*, 282–289, (2001).
[6] Carbonell J. Klein-Seetharaman J. Liu, Y. and V. Gopalakrishnan, 'Comparison of probabilistic combination methods for protein secondary structure prediction', *Bioinformatics*, **20**, 3099–3107, (2004).
[7] Liu T. Zhang-X. Wang D. Qin, T. and H. Li, 'Global ranking using continuous conditional random fields', *In Proceedings of NIPS*, 1281–1288, (2008).
[8] Kaufman Y.J. Tanre-D. Mattoo S. Chu D.A. Martins J.V. Li R-R. Ichoku C. Levy R.C. Kleidman R.G. Eck T.F. Vermote E. Remer, L.A. and B.N. Holben, 'The modis aerosol algorithm, products and validation', *Journal of the Atmospheric Sciences*, **62**, 947–973, (2005).
[9] Taxt-T. Solberg, A. H. S. and A. K. Jain, 'A markov random field model for classification of multisource satellite imagery', *IEEE Trans. Geosci. Remote Sensing*, **34**, 110–113, (1996).