Semi-Supervised Learning for Integration of Aerosol Predictions from Multiple Satellite Instruments

Nemanja Djuric, Lakesh Kansakar, Slobodan Vucetic
Department of Computer and Information Sciences, Temple University

Abstract

Aerosol Optical Depth (AOD) is estimated daily on a global scale by several Earth-observing satellite instruments. Each instrument has different coverage and sensitivity to atmospheric and surface conditions, and, as a result, the quality of AOD estimated by different instruments varies across the globe. We present a method for learning how to aggregate AOD estimations from multiple satellite instruments into a more accurate estimation. The proposed method is semi-supervised, as it is able to learn from a small number of labeled data where labels come from a few accurate and expensive ground-based instruments, and a large number of unlabeled data. The method uses a latent variable to partition the data, such that in each partition the expert AOD estimations are aggregated in a different, optimal way.
This work was supported by NSF grant IIS-1117433.

Aerosols – Important factor in Earth's climate

Aerosols are small particles suspended in atmosphere. Aerosol scattering of sunlight can reduce visibility and redden sunrises and sunsets, and they affect cooling of surface by absorbing and reflecting Solar radiation.

Figure 1. High concentration of aerosols in downtown Philadelphia, USA

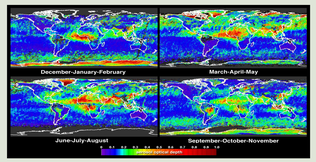

Figure 2. Distribution of aerosols varies significantly across different locations and time intervals

One of the biggest challenges in climate research is to characterize and quantify the distribution of aerosols. This can be done by measuring the Aerosol Optical Depth (**AOD**), which indicates the amount of depletion that light undergoes as it passes through the atmosphere.

There are two approaches to measure AOD:

- ground-based, using AERONET network of sensor instruments (Fig. 3);
- satellite-based, using sensor instruments aboard satellites such as Terra, Aqua, Aura, SeaWiFS, and others (Figure 4).

Sensor type	Ground-based	Satellite-based
Temporal coverage	HIGH	MODERATE
Spatial coverage	EXTREMELY LOW	HIGH
Accuracy	HIGH	LOW
Cost	HIGH	LOW

Table 1. Comparison of different AOD sensors

Satellite observations offer the potential of achieving continuous global coverage, necessary to understand effect of aerosols more completely.

Figure 3. Locations of AERONET sensors

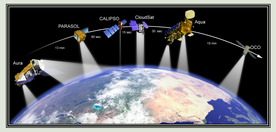

Figure 4. Satellites provide global coverage

Global estimation of Aerosol Optical Depth (AOD) – Problem overview

In order to provide global coverage of AOD, the question is how to combine AOD predictions from satellites considering they have different temporal and spatial coverage, as well as different measurement quality. Satellite measurements with collocated ground-based AERONET estimates are called labeled data (points *A* and *B* in Fig. 5), otherwise the data is unlabeled (points *C* and *D*).

In this work we tackle this task by finding an optimal linear combination of available satellite sensor AOD measurements.

Open issues inherent to remote sensing domain:

- Noisy experts' estimates may be correlated;
- Some expert predictions may be missing (due to lack of coverage: points *B* and *D* do not have MISR predictions available; or due to presence of clouds);
- Large parts of training data may be unlabeled (i.e., missing AERONET measurements);
- Experts should be combined differently for different subsets of the data (e.g., MISR does not maintain same quality of AOD estimates across the globe).

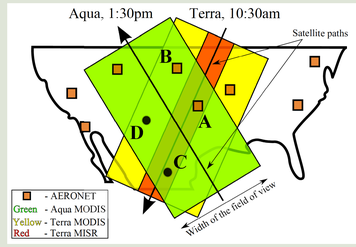

Figure 5. Daily coverage of AOD sensors

Semi-supervised aggregation of multiple satellite sensor predictions

Assumptions

We are given a training data set D with N data points sampled IID, each consisting of a ground truth and K experts' opinions,

$$D = \{\{\hat{y}_{ik}\}_{k=1,\dots,K}, y_i\}_{i=1,\dots,N}.$$

We assumed that the true labels y_i are sampled from a Gaussian distribution, and that the expert predictions, represented as a vector $\hat{\mathbf{y}}_i$, are sampled from a conditional multivariate Gaussian distribution,

$$y_i \sim N(\mu_y, \sigma_y^2) \quad \text{and} \quad \hat{\mathbf{y}}_i \mid y_i \sim N(y_i \mathbf{1}, \Sigma).$$

For example, in the aerosol domain that we study, the experts are satellite instruments and the predictions are their individual AOD estimates, while the ground truth is measurement given by AERONET. We assume there are N_u unlabeled and N_l labeled training data points, with $N = N_u + N_l$.

Then, the training task is to find the parameters $\Theta = \{\Sigma, \mu_y, \sigma_y^2\}$. Once the training is complete, aggregated prediction is found as mean of the posterior,

$$y_i \mid \hat{\mathbf{y}}_i \sim N(\overline{y}_i, (\mathbf{1}^T \Sigma^{-1} \mathbf{1})^{-1}),$$

where the mean can be found as follows,

$$\overline{y}_i = \frac{\hat{\mathbf{y}}_i'^T \Sigma'^{-1} \mathbf{1}}{\mathbf{1}^T \Sigma'^{-1} \mathbf{1}}, \text{ where } \hat{\mathbf{y}}_i' = [\hat{\mathbf{y}}_i^T, \mu_y]^T \text{ and } \Sigma' = \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \sigma_y^2 \end{bmatrix}.$$

Semi-supervised learning without missing experts

First, we derive equations when all experts are assumed available. The parameters can be learned by maximizing log-likelihood of the training data. We start by writing training data probability, $P(D \mid \Theta) = P(D_u \mid \Theta) \cdot P(D_l \mid \Theta)$.

The probability of unlabeled data set can be written as

$$P(D_u \mid \Theta) = \prod_{i=1}^{N_u} \int_y P(\hat{\mathbf{y}}_i \mid y, \Theta) P(y \mid \Theta) dy = \prod_{i=1}^{N_u} \left(\sqrt{\frac{|\Sigma'^{-1}|}{(2\pi)^{K-1} \mathbf{1}^T \Sigma'^{-1} \mathbf{1}}} \right) \exp\left(-\frac{1}{2} (\hat{\mathbf{y}}'_i - \overline{y}_i \mathbf{1})^T \Sigma'^{-1} (\hat{\mathbf{y}}'_i - \overline{y}_i \mathbf{1})\right),$$

and the probability of labeled data set as follows

$$P(D_l \mid \Theta) = \prod_{i=N_u+1}^{N} P(\hat{\mathbf{y}}_i \mid y_i, \Theta).$$

To simplify equations, in the remainder we assume $\sigma_y^2 \to \infty$, which amounts to an uninformative prior over target variable. Finding derivative of data log-likelihood with respect to Σ^{-1} and equating to 0, we obtain update expression,

$$\Sigma = \frac{1}{N} \left((\hat{\mathbf{Y}}_l - \mathbf{y}_l \mathbf{1}^T)^T (\hat{\mathbf{Y}}_l - \mathbf{y}_l \mathbf{1}^T) + \hat{\mathbf{Y}}_u^T \hat{\mathbf{Y}}_u + \frac{N_u \mathbf{1}\mathbf{1}^T}{\mathbf{1}^T \Sigma^{-1} \mathbf{1}} + \sum_{i=1}^{N_u} (\overline{y}_i^2 \mathbf{1}\mathbf{1}^T - \overline{y}_i (\mathbf{1}\hat{\mathbf{y}}_i^T + \hat{\mathbf{y}}_i \mathbf{1}^T)) \right).$$

Missing experts and incorporation of prior knowledge

Assume that the i^{th} data point has q missing experts, and we reorganize vector $\hat{\mathbf{y}}_i$ and precision matrix (using the permutation function Π_i) so the first a elements are from available experts, and the last q elements are missing,

$$\hat{\mathbf{y}}_i = [\hat{\mathbf{y}}_{ai}^T, \hat{\mathbf{y}}_{qi}^T]^T \text{ and } \Pi_i(\Sigma^{-1}) = \begin{bmatrix} \mathbf{U} & \mathbf{V} \\ \mathbf{V}^T & \mathbf{Q} \end{bmatrix}.$$

Given the learned parameters and available experts' predictions, it follows,

$$y_i \mid \hat{\mathbf{y}}_{ai} \sim N(\overline{y}_i, (\mathbf{1}^T \mathbf{U}'_i \mathbf{1})^{-1}), \text{ where } \overline{y}_i = \frac{\hat{\mathbf{y}}_{ai}^T \mathbf{U}'_i \mathbf{1}}{\mathbf{1}^T \mathbf{U}'_i \mathbf{1}} \text{ and } \mathbf{U}' = \mathbf{U} - \mathbf{V}\mathbf{Q}^{-1}\mathbf{V}^T.$$

Before maximizing log-likelihood, we rewrite probability of unlabeled data as

$$P(\hat{\mathbf{y}}_{ai} \mid \Theta) = \iint_{y, \hat{\mathbf{y}}_{qi}} P([\hat{\mathbf{y}}_{ai}^T, \hat{\mathbf{y}}_{qi}^T]^T \mid y_i, \Theta) P(y \mid \Theta) dy \, d\hat{\mathbf{y}}_{qi} = \left(\sqrt{\frac{|\Sigma^{-1}| |\mathbf{Q}_i|^{-1}}{(2\pi)^{K+q-1} \mathbf{1}^T \mathbf{U}'_i \mathbf{1}}} \right) \exp\left(-\frac{1}{2} (\hat{\mathbf{y}}_{ai} - \overline{y}_i \mathbf{1})^T \mathbf{U}'_i (\hat{\mathbf{y}}_{ai} - \overline{y}_i \mathbf{1})\right),$$

and the probability of labeled data as $\hat{\mathbf{y}}_{ai} \mid y_i \sim N(y_i \mathbf{1}, \mathbf{U}'^{-1}_i)$.

If we have prior knowledge about the relationship between experts, we can impose a prior over the precision matrix in a form of Wishart distribution.

Data partitioning using latent variables and the EM algorithm

We consider partitioning the data points into R groups, called *regimes*, where each regime is governed by a different multivariate Gaussian. We assume that we have available feature vector $\mathbf{x}_i$ for the i^{th} data point that can be used to assign it to an appropriate regime. Then,

$$\overline{y}_i = E[y_i \mid \hat{\mathbf{y}}_{ai}, \mathbf{x}_i, \Theta] = \sum_{r=1}^{R} \pi_{ir}(\mathbf{x}_i) \frac{\hat{\mathbf{y}}_{ai}^T \mathbf{U}'_{ir} \mathbf{1}}{\mathbf{1}^T \mathbf{U}'_{ir} \mathbf{1}}.$$

Probability of observing the i^{th} labeled and unlabeled data point equals

$$P(\hat{\mathbf{y}}_{ai} \mid y_i, \mathbf{x}_i, \Theta) = \sum_{r=1}^{R} \pi_{ir}(\mathbf{x}_i) P_r(\hat{\mathbf{y}}_{ai} \mid y_i) \quad \text{and} \quad P(\hat{\mathbf{y}}_{ai} \mid \mathbf{x}_i, \Theta) = \sum_{r=1}^{R} \pi_{ir}(\mathbf{x}_i) P_r(\hat{\mathbf{y}}_{ai}).$$

It is not easy to maximize log-likelihood due to sum in the above equations. To facilitate optimization, we introduce R latent binary variables z_{ir} indicating whether the i^{th} data point was generated by the r^{th} regime, and write

$$P(\hat{\mathbf{y}}_{ai}, \mathbf{z}_i \mid y_i, \mathbf{x}_i, \Theta) = \sum_{r=1}^{R} (\pi_{ir}(\mathbf{x}_i) P_r(\hat{\mathbf{y}}_{ai} \mid y_i))^{z_{ir}}, \quad \text{where} \quad \pi_{ir} = \frac{\exp(-(\mathbf{x}_i - \mathbf{q}_r)^T \Lambda_r (\mathbf{x}_i - \mathbf{q}_r))}{\sum_{m=1}^{R} \exp(-(\mathbf{x}_i - \mathbf{q}_m)^T \Lambda_m (\mathbf{x}_i - \mathbf{q}_m))},$$

where we introduced per-regime parameters, prototype $\mathbf{q}_r$ and scaling matrix Λ_r. We can maximize log-likelihood using the EM algorithm. In the E-step, we fix model parameters and compute h_{ir}, the expectation of latent variable z_{ir}. In the M-step, we fix h_{ir} and update the model parameters.

Experiments on synthetic data

We generated data by sampling ground-truth y_i from zero-mean Gaussian with unit-variance, then sampling K expert predictions from the multivariate Gaussian $N(y_i \mathbf{1}, \Sigma)$. We removed each expert with probability 0.5 to simulate missing experts.

We first set $R = 1$, $K = 5$, and $\Sigma = \text{diag}([0.1, 0.2, 0.3, 0.4, 0.5])$. For $R = 2$ we set $\Sigma_1 = \text{diag}([0.1, 0.2, 0.3, 0.4, 0.5])$, $\Sigma_2 = \text{diag}([0.5, 0.4, 0.3, 0.2, 0.1])$ and $\mathbf{q}_1 = [1, 1]$, $\mathbf{q}_2 = [-1, -1]$. The results for different training data sizes and fractions of labeled data points for $R = 1$ and $R = 2$ after 100 repetitions are shown in Figures 6a and 6b, respectively. The performance of unsupervised method is already better than simple averaging. Moreover, as we increase the number of labeled points, the semi-supervised method further improves the accuracy, approaching the lower bound on RMSE achieved by the optimal combination of experts.

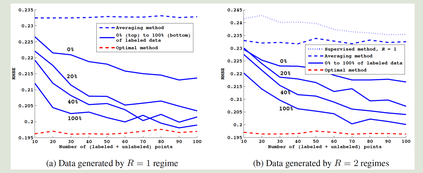

Figure 6. Results on the synthetic data

Experiments on aerosol data

We considered AERONET data from 33 sites within the USA, and collocated data from 5 satellite instruments (Terra MODIS, MISR, Aqua MODIS, OMI, SeaWiFS) spanning years 2006 to 2010. This resulted in N = 6,913 data points, with 58% of missing expert predictions. We used this data set for two sets of experiments: (**1**) evaluating usefulness of partitioning; and (**2**) evaluating usefulness of unlabeled data. We used longitude and latitude of the AERONET site as features $\mathbf{x}_i$ for the i^{th} data point, and performed leave-one-site-out cross-validation (see Table 2).

➔ **For (1)**, from each site we randomly sampled 100 points, and assumed that 50 are labeled and 50 unlabeled. For baseline we computed average of available experts. By increasing number of clusters from 1 to 2, there was a drop in RMSE of nearly 5%. In Figure 7 we see that clustering roughly corresponds to partitioning proposed by domain scientists.

➔ **For (2)**, we randomly selected 2, 4, and 6 sites and took 100 points from each as labeled data. Then, we selected 100 points from remaining sites and treated them as unlabeled. We trained one model which used only labeled data, and one using both labeled and unlabeled. We see unlabeled data were helpful and led to significant reductions in RMSE.

Method	# clusters	RMSE
Averaging	--	0.0818
All sites, semi-super.	1	0.0677
All sites, semi-super	2	0.0648
2 sites, supervised	2	0.0795
2 sites, semi-super.	2	0.0752
4 sites, supervised	2	0.0728
4 sites, semi-super.	2	0.0704
6 sites, supervised	2	0.0694
6 sites, semi-super.	2	0.0688

Table 2. Performance of aggregation methods

Figure 7. Found clustering of AERONET sites for $R = 2$