Reducing Need for Collocated Ground and Satellite based Observations in Statistical Aerosol Optical Depth Estimation

Debasish Das, Vladan Radosavljevic, Slobodan Vucetic, Zoran Obradovic
Center for Information Science and Technology, Temple University, USA
das@ist.temple.edu, vladan@ist.temple.edu vucetic@ist.temple.edu, zoran@ist.temple.edu

Abstract

One of the biggest challenges of current climate research is to characterize and quantify the effect of aerosols on the global and local weather. This requires an accurate prediction of Aerosol Optical Density (AOD) which is defined as the amount of loss a beam of light incurs when it passes through the atmosphere. In this paper a neural network-based data-driven prediction model is considered which uses collocated satellite (MODIS) observation and ground-based (AERONET) AOD retrievals as predictors and target respectively. This paper studies an active learning-based data collection method which will facilitate the learning of a sufficiently accurate AOD prediction model using a minimal set of labeled training data by querying the labels of only the most informative data points.

Key words: *AOD, Active Learning, Query by Committee*

1. INTRODUCTION

Aerosols are small solid or liquid particles suspended in air, emanating from natural or man-made sources. They proved to be very influential in shaping the short-term and long-term weather of the earth both globally and locally. They have both warming and cooling effects on earth's surface and atmosphere due to their absorption and reflection of solar radiation. An important metric of aerosol's concentration in the atmosphere is Aerosol Optical Density (AOD), which is a dimensionless quantity and is defined as the amount of depletion a beam of sunlight incurs due to absorption and scattering by aerosols when it travels through the atmosphere. Along with the atmospheric model, AOD can provide useful estimates of atmospheric effects on the transmission and reflection of solar radiation and therefore helps in building climate models.

There are two sources of observations which aid retrieval of AOD. AEROsol robotic NETwork (AERONET) is a global remote sensing network of about 540 ground-based radiometers which retrieves AOD several times an hour under clear-sky conditions. The AERONET retrievals are very accurate and therefore are considered as ground-truth value when satellite-based retrievals are validated. On the other hand The MODerate resolution Imaging Spectrometer (MODIS), aboard NASA's Terra and Aqua satellites, is one of the major instruments for satellite-based AOD retrieval. MODIS [1] observes reflected solar radiation over a large spectral range with a high spatial resolution and almost daily coverage of the entire Earth. MODIS has higher spatial coverage but moderate retrieval accuracy, whereas AERONET has very limited spatial coverage but provides highly accurate retrieval.

Most operational deterministic algorithms that retrieve AOD from satellite observation are constructed as inverse operators of high-dimensional non-linear functions derived from forward-simulation models according to the domain-knowledge of the aerosol physical properties. AOD is retrieved by matching observed reflectance with the simulated values stored in the Look-Up Tables (LUT). These algorithms are manually tuned by domain scientists periodically after validating the AOD values thus obtained against the AERONET retrievals. This makes these algorithm time consuming and at the same time fails to guarantee the most efficient use of highly accurate AERONET retrievals.

An efficient alternative is statistical or data-driven approach where a regression model is trained from the collocated satellite and ground-based observations using satellite-based observations as the predictors and ground-based observations as labels. This regression model can therefore be used to predict AOD from satellite observations where no ground-based retrievals are available. This method is computationally less expensive, flexible to different retrieval scenarios and more accurate than the deterministic algorithms if sufficient amount of training data is available. One problem with the data-driven approach is that it uses the ground-based data to build the prediction model which are very costly. One solution to this problem is to minimize the amount of training points needed to build an accurate model by selectively choosing training-data points having maximum information.

978-1-4244-2808-3/08/$25.00 ©2008 IEEE

2. NEURAL NETWORK MODEL

Neural networks are suitable for data-driven retrieval of atmospheric parameters because of their capability to approximate complex regression models. In a previous work [2] a neural network was trained to predict AOD over continental US was shown to significantly improve the accuracy of prediction over then operational C004 deterministic algorithm [3]. Because a universal predictor could not completely explain the complex spatial-temporal variation of aerosol density, a second algorithm [4] was proposed as an integration of global and local data-driven aerosol predictors. There, a global neural network was trained to predict AERONET AOD over U.S. in combination with region specific neural networks. The final AOD prediction was obtained as weighted average of global and local AOD predictors. The results showed that data-driven approach could be used as a complement to the traditional domain-based retrieval algorithms. However, in this paper, we have trained only a global neural network in order to prove the usefulness of active learning in reducing the need for collocated training data with an assumption that it will be equally applicable to a fusion model.

3. ACTIVE LEARNING

The problem of learning a regression model can be defined as follows. Given a set of N training examples ($\boldsymbol{x}_i$,y_i, i=1,2…N) where $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, \ldots x_{iM}\}$ is a vector of M attributes and y_i is the corresponding target value, the task is to construct a function $F(\boldsymbol{x}_i)$ so that it can predict the true target y_i as accurately as possible. A standard measure of accuracy in regression is Mean Squared Error (MSE) defined as $mean((y - F(x))^2)$. For passive learning, N training examples are sampled from an unknown joint distribution $P(\boldsymbol{x},y)$ whereas in active learning, learner selects the input examples $\boldsymbol{x}_i$ and observes the corresponding targets y_i which are sampled from an unknown conditional distribution $P(y|\boldsymbol{x})$.

The method of pool-based active learning [5] is defined as a set of operators $\{C, Q, S, L, U\}$, where C is a learner which is trained on a set of labeled examples L, Q is query function which queries each example from the pool of unlabeled examples U to select the most informative sample(s), S is an oracle or supervisor which will provides the true label of the selected unlabeled examples before adding them to the pool of labeled examples. Initially L will be created taking random samples from U and labeling them using S. A good active learner should be able learn quickly even from a small set of examples.

The query function is central to any active learning algorithm and different algorithms mainly differ in defining this query function. In existing literature, two broad categories of designing the query function can be encountered. One of them is the statistical approach where the query function is designed to minimize the future error of the learner. But the these algorithms assume specific statistical model for the conditional probability distribution P(**x**|y) to calculate the future error of the learner. But in most practical cases this distribution model is unknown. The alternative approach for query design is based on some pragmatic methods where an indirectly calculated uncertainty score is attributed to each of the unlabeled examples and the example with maximum uncertainty is selected. One such approach is Query by Committee (QBC) proposed by *Seung et al* [6], where a committee of learners is trained and each prospective unlabeled example is assigned an uncertainty score based on the disagreement in the predicted label between the committee of learners. This approach is very popular due to its ease of implementation for a variety of learning problems. In this paper, we have adopted the QBC algorithm of active learning to reduce the need of collocated training points. But before further details of QBC algorithm implementation can be given, a brief description of the data set being used is given.

4. DATA-SET

A total of 5,074 collocated [7] data-points have been collected by fusing high-quality AERONET level 2.0 observations with observations from MODIS Terra and Aqua instruments. The collocated points were fused over 125 AERONET sites across the entire earth during the first eight months of 2005. Continental U.S. and Europe are most densely covered with AERONET sites whereas South America, Africa and continental Asia are covered sparsely. So, the global applicability of the learning algorithm will be somehow limited to only those regions where there is enough coverage.

Attributes extracted from the MODIS data are listed in table 1. Only these attributes are used as the inputs to the learning algorithm. The seven wavelengths taken from MODIS observations were between 440 nm-2100 nm, as these are sufficient for predicting AOD [3]. First three weeks of data from each month are kept as training data and the remaining one week of data are used for testing.

Attribute Index	Description
1-5	Solar Zenith, Solar Azimuth, Sensor Zenith, Sensor Azimuth and Scattering angles
6-19	Mean and Standard Deviations of reflectances in 50X50 blocks in 7 wavelengths respectively.
20	AERONET site elevation

Table 1: *Attributes collected from MODIS observations*

5. EXPERIMENTAL RESULTS

The committee of learners was created with ten neural networks, each having ten hidden nodes but different learning rates. Apart from these, two identical neural networks having 10 hidden nodes each were created to simulate passive and active learning. All of them were trained initially with 200 randomly chosen training points so that the initial accuracies of the networks were dependable enough to initialize the active learning.

The variance of the ten NNs at 200 randomly chosen unlabeled data-points was measured to compare the level of prediction disagreement. The points with relatively high variance are those at which more information is needed for correct modeling. It follows therefore that resampling of a labeled point corresponding to maximum variance is required. In each iteration, 20 such points were added to the pool of labeled points and the active learner was trained on them. On the other hand, for passive learner, data points are chosen purely at random. This process is repeated until a pre-defined number (1000 in this case) of labeled points are added to the pool or a pre-defined level of R^2-accuracy (0.85 in this case) is attained by any of the learners. Table 2 gives the pseudo-code for QBC algorithm implemented in this study. The reason for applying the selective sampling only on 200 randomly selected unlabeled data is to remove any possible correlation that might exist between the unlabeled points if all of them are considered. Correlated points do not add new information to the learner.

```
1.  Create 1 NN to work as actual learner and 10 NN to serve
    as a committee.
2.  Select 200 points randomly from U and add them to L after
    labeling them.(see section 3 for definitions of U and L)
     While |L|<1000
3.      Train the committee and actual learner on L.
4.      Select 200 points randomly from U.
5.      Apply the committee of NN to predict the labels of
        selected points and measure the variance of predictions
        of 10 NN for each of those.
6.      Get the labels of 20 points with maximum variance in
        predictions and add them to L.
    end
```

Table 2: *Implementation of QBC algorithm for active learning*

R^2-accuracy of both learners is measured in each iteration on the test data and reported along with the number of labeled training points used in that iteration. The entire experiment was run 60 times and the average accuracies in each iteration were recorded. The plots of accuracy in different iterations against the number of training points for both passive and active learning are shown in Figure 1. These curves are called learning curves. It can be seen from the Figure 1 that active learning performs consistently better than passive learning in terms of learning rate. For an example, with active learning, around 480 labeled data points are required to attain an accuracy of 0.75 whereas passive learning needs almost 600 data points to reach this level, i.e. active learner can save almost 20% of supervision cost required by its passive counterpart.

A second experiment was designed to study the performance of active learning under an initially biased set of training data. Passive learner was initially trained with 200 randomly chosen data points as before. But for active learning, all 200 initial data-points were intentionally chosen from winter in order to enforce a poor initial performance. A learner which is trained only on data from winter is expected to predict poorly on a random test data where other seasons are dominant. This is due to large distributional difference of AOD vs. observed radiances in presence of white background in winter vs. darker background situations of other seasons. With this initial set-up active learner was allowed to select data-points from the entire pool of unlabeled data using QBC algorithm described earlier and passive learner selected data randomly. Apart from recording the accuracies in each iteration for both learners, the seasonal distribution of actively selected data-points was also recorded. Figure 2 shows the comparison of learning curves of two types of learning and Figure 3 shows the seasonal distribution of actively selected data-points for first 10 iterations.

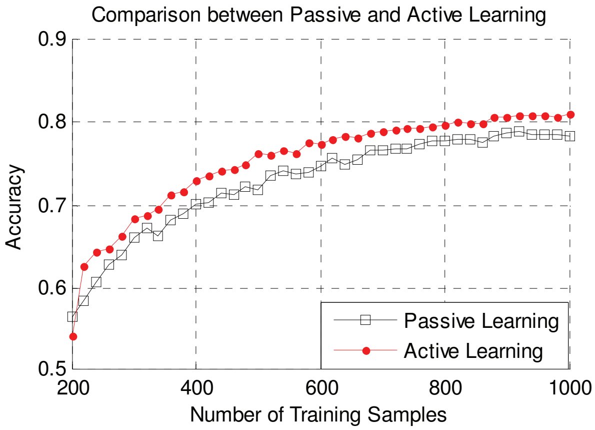

Fig 1: *Comparison of learning curves for passive and active learner of AOD predictors*

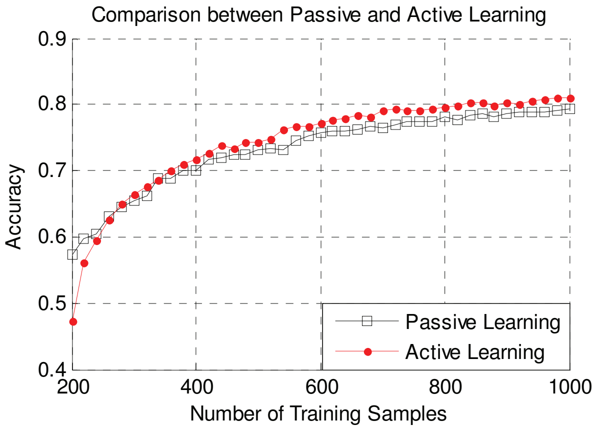

Fig 2: *Comparison of learning curves for passive learner with random initial dataset and active learner with a biased initial data set*

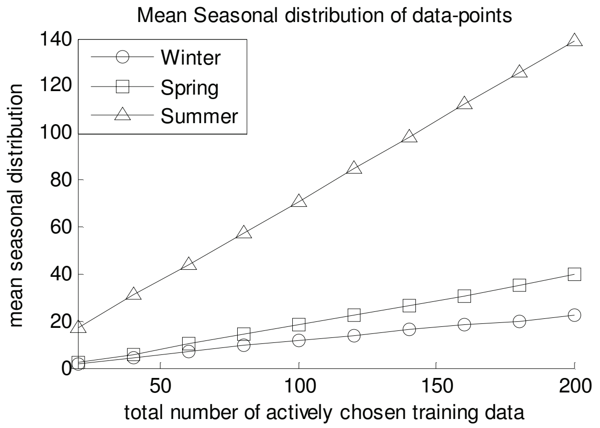

Fig 3: *Distribution of actively chosen data in different seasons with all initial training data from winter.*

The results show that even if a poor performance is enforced for an active learner by an initially biased training set, it manages to recover and even supersede the performance of a passive learner by picking up highly informative data-points. For example, in the previous experiment, all initial data points were chosen from winter. Therefore in subsequent iterations, active learner has chosen most of the points from summer and spring and a few points from winter because data points from summer and spring adds more variation and thereby more information to the training set.

6. CONCLUSION AND FUTURE WORKS:

In this paper, a simple QBC-based active learning scheme is proposed to reduce the need of costly collocated training data required to train a neural network-based AOD prediction model. Experimental evidence shows that the above scheme can significantly reduce the supervision cost and improve the accuracy of the model at a faster rate than random sampling. It also shows that this scheme can be used to identify potentially informative data points in cases where initial training set is biased and full of data points which are under-represented in the entire unlabeled pool of potential training points.

As an extension of this work we are currently using active learning in order to predict the location of next AERONET site that can exploit maximum usable information for the AOD predictors. This is an important decision to the Earth-scientists as these instruments are costly and have limited spatial coverage.

One problem with the QBC algorithm is that it requires iterative retraining of several neural networks which is a computationally costly process. Efforts are being made to reduce this computational cost by forming the committee with learners such as kernel regressor which have a shorter training time. Another alternative method under consideration is to use single-learner query functions where uncertainty of prospective training samples can be calculated directly from the current parameters of the neural network in each iteration. One such method is proposed by *Aires et al* [8] that calculates the Bayesian estimate of uncertainty of neural network outputs and which can be used as the informativeness score of each prospective unlabeled example.

Acknowledgement

This work is funded in part by NSF grant IIS-0612149.

REFERENCES:

[1] R. C. Levy, L. A. Remer, et al., "Evaluation of the MODIS aerosol retrievals over ocean and land during CLAMS", *J. Atmos. Sci.,* 62(4): 974-992, 2005.

[2] B. Han, Z. Obradovic, Z. Li and S. Vucetic, "Data mining support for improvement of MODIS aerosol retrievals," *Proc. IEEE Int'l Geoscience and Remote Sensing Symp.*, Denver, CO, Aug. 2006.

[3] L. A. Remer, D. Tanré and Y. J. Kaufman, "Algorithm for remote sensing of tropospheric aerosol from MODIS for collection 005,
"modis.gsfc.nasa.gov/data/atbd/atbd_mod02.pdf, 2006.

[4] B. Han, S. Vucetic, A. Braverman, Z. Obradovic, "Construction of a geospatial predictor by fusion of global and local models," Proc. *2005 8th International Conference on Information Fusion,* vol. 1, July 2005, pp. 25-28.

[5] M. Li, 2006. Confidence-based active learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 8 (Aug. 2006), 1251-1261.

[6] H.S. Seung, M. Opper, H. Somepolinsky, "Query by committee," *Proc. 5th Annual. Workshop on Computation Learning Theory,* pp. 287-294, 1992.

[7] C. Ichoku, D. A. Chu, et al., "A spatio-temporal approach for global validation and analysis of MODIS aerosol products," *Geophys. Res. Lett*. 29(12): no.616, 2002.

[8] F. Aires, C. Prigent and W. B. Rossow 2004. Neural network uncertainty assessment using Bayesian statistics: a remote sensing application. *Neural Comput.* 16, 11 (Nov. 2004), 2415-2458