

Evaluation of Decision Trees for Cloud Detection from AVHRR Data

Smadar Shiffman
Computational Sciences Division
NASA Ames Research Center / QSS
MS269/3
Moffett Field, CA 94035

Ramakrishna Nemani
Earth Sciences Division
NASA Ames Research Center
MS242-4
Moffett Field, CA 94035

Abstract—Automated cloud detection and tracking is an important step in assessing changes in radiation budgets associated with global climate change via remote sensing. Data products based on satellite imagery are available to the scientific community for studying trends in the Earth’s atmosphere. The data products include pixel-based cloud masks that assign cloud-cover classifications to pixels. Many cloud-mask algorithms have the form of decision trees. The decision trees employ sequential tests that scientists designed based on empirical astrophysics studies and simulations. Limitations of existing cloud masks restrict our ability to accurately track changes in cloud patterns over time. In a previous study we compared automatically learned decision trees to cloud masks included in Advanced Very High Resolution Radiometer (AVHRR) data products from the year 2000. In this paper we report the replication of the study for five-year data, and for a gold standard based on surface observations performed by scientists at weather stations in the British Islands. For our sample data, the accuracy of automatically learned decision trees was greater than the accuracy of the cloud masks $p < 0.001$.

I. INTRODUCTION

Understanding the role of clouds in the current climate is a prerequisite for predicting future climate change due to human activities [1]. Satellite-born instruments continually acquire data about the Earth’s oceans, land, and atmosphere. The data is processed to derive high-level observations, which are then distributed to the scientific community via online data products. The data products include cloud masks, which have dual functionality. The masks designate locations in which the observations may have limited quality due to cloud interference, and also provide estimated cloud amounts for each location. The cloud mask of interest in this study is included in products derived from data acquired by the Advanced Very High Resolution Radiometer (AVHRR) instrument on board the NOAA-14 weather satellite of the National Oceanic and Atmospheric Administration. The mask is called Clouds from AVHRR, phase 1 (CLAVR-1) [2].

The CLAVR-1 cloud mask is computed from measured reflectance and emission values using classification algorithms that scientists developed through experimentation with the data. To derive the algorithms, the scientists simulated clear-sky and cloud characteristics for a variety of surface and atmospheric conditions, and analyzed ambiguous manifestations of different physical phenomena, for example,

similar reflectance values for snow, ice and clouds. The algorithms employ sequential-threshold tests to arrive at decisions about the presence of clouds or about cloud composition [2-3]. The limitations of existing cloud masks [4] provided motivation for on-going research to develop improved cloud detection and characterization algorithms.

Cloud-detection methods must disambiguate clouds and other entities that have characteristics similar to clouds. Scientists have used a variety of machine-learning methods to learn models for remote sensing data, for example, neural networks [5], Bayesian classification [6], kernel methods [7-10], genetic algorithms [11], classification trees [12] and regression trees [13]. The results of these approaches range from promising preliminary results to validated algorithms that are deployed in high-level remote-sensing data products [14]. Of these machine-learning methods, the methods that resemble the sequential tests in cloud masks the most are classification trees. Because of this resemblance, we use classification trees in this study, and refer to them as automatically learned decision trees (ALDT).

In a previous study [15] we demonstrated the feasibility and potential of ALDT for improving the accuracy of cloud masks based on AVHRR data. In that study we compared cloud-detection results of the CLAVR-1 algorithm, which was devised by experts, to cloud-detection results of ALDT. We used ground observations collected by the National Aeronautics and Space Administration Clouds and the Earth’s Radiant Energy Systems S’COOL project as the gold standard. We found that for our sample data, the accuracy of ALDT was greater than the accuracy of the CLAVR-1 cloud masks, and that the difference in accuracy was statistically significant. The goal of this work was to corroborate the preliminary results in [15] by replicating the study and enhancing it in three ways: extending the time-period coverage of the sample data from one year to five years, using a refined ordinal scale for cloud quantity, and using a gold-standard generated by scientists.

II. BACKGROUND

A. AVHRR Data

The NOAA-14 AVHRR daily 8km global data product includes 12 scientific datasets (SDS), each of which

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE 25 JUL 2005		2. REPORT TYPE N/A		3. DATES COVERED -	
4. TITLE AND SUBTITLE Evaluation of Decision Trees for Cloud Detection from AVHRR Data				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Computational Sciences Division NASA Ames Research Center / QSS MS269/3 Moffett Field, CA 94035				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release, distribution unlimited					
13. SUPPLEMENTARY NOTES See also ADM001850, 2005 IEEE International Geoscience and Remote Sensing Symposium Proceedings (25th) (IGARSS 2005) Held in Seoul, Korea on 25-29 July 2005.					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 4	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

incorporates within a single plane a measured parameter, a flag, or a computed parameter. The SDS are: normalized difference vegetation index, CLAVR-1 cloud mask, quality control flag, scan angle, solar zenith angle, relative azimuth angle, surface reflectance in the visible wavelengths (channel 1), surface reflectance in the near-infrared wavelengths (channel 2), surface brightness temperature in the thermal infrared wavelengths (channels 3-5), and acquisition day and time [16].

B. The CLAVR-1 Cloud Mask

The CLAVR-1 algorithm includes four decision trees, one for each of daytime land scene, daytime ocean scene, nighttime land scene, and nighttime ocean scene. Each decision tree performs a series of threshold and uniformity tests on a 2x2 array of pixels, and classifies pixels as *clear*, *mixed*, or *cloudy*. The values used for each test are either retrieved channel values, or functions of retrieved values that incorporate acquisition parameters and estimates of emitted radiances [2]. Several tests were designed specifically to resolve ambiguities, for example, ambiguities due to reflectance greater than 44% in channel 1 or channel 2 for snow, ice, or sun glint. The thresholds used for the tests were derived empirically or via simulations of a variety of observation conditions as determined by cloud/surface/time combinations.

The sequential decision process in CLAVR-1 discriminates between clouds, first by their gross characteristics, and then by their subtle characteristics. The algorithm ensures that pixels that fail all the tests have a very small probability of having radiatively significant clouds. The sequential-test nature of CLAVR-1 makes it similar to ALDT, but unlike the latter, the CLAVR-1 algorithm is not based on an exhaustive analysis of the data space.

The CLAVR-1 algorithm has several limitations. First, the algorithm assumes that there is a representative sample of clear pixels in each image, however, this assumption does not hold for broadly overcast scenes. Second, the algorithm does not work well for polar-winter scenes or nighttime scenes, when only the thermal channels are available. Third, the ability of the algorithm to differentiate between clouds and other entities that appear as clouds in AVHRR images is limited.

C. CLAVR-1 Evaluation

Evaluation of cloud masks is difficult because there is no gold standard to which to compare the masks. Researchers estimate the quality of cloud masks by comparing them to masks produced by human analysts or by other algorithms. Stowe and colleagues [2] compared the results of CLAVR-1 to estimates of a human-expert analyst. Stowe's team found that the mismatch between CLAVR-1 and the expert estimates was at least 10%, and that the mismatch varied for different cloud amounts, geographical location, and season.

D. Decision Trees

Decision trees are classifiers that employ rules sequentially to determine the class to which an item belongs. Decision trees can be learned automatically from training data for which the classes are known using a computer program that generates trees via sequential binary partitioning of the training data [17]. The learning procedure searches in the space of all possible decision trees that fit the data for an optimal tree, where the optimization criterion is minimal prediction error.

III. METHODS

A. Data Preparation

We obtained ground observations of cloud characteristics from the British Atmospheric Data Centre (BADC) [18]. The BADC data included observation of cloud amounts in numerous weather stations within the British Islands. We selected all observations that were available for the year 1996-2000 from 1238 weather stations. Then, we retrieved 8km daily AVHRR data that matched the BADC data in acquisition date, time, longitude and latitude. We excluded from this dataset all records that exhibited one of the following criteria: a. the AVHRR data quality flag indicated out-of-range values or processing errors; b. the CLAVR-1 mask had a *no decision* value; c. there was no BADC *total-cloud-amount* observation, or the observation value indicated that the estimate was affected by obscuring fog or other meteorological phenomena. The average number of records per year was 18632. We used the BADC observations as the gold standard for labeling training and test data. We compared the labels of the test data to predictions made for the same data by CLAVR-1 and by the ALDT.

Although both CLAVR-1 and the BADC data utilized an ordinal scale for characterization of cloud amount, the scales were different and mapping one scale to the other could be done in more than one way. The CLAVR-1 mask had three possible values: *clear*, *mixed*, and *cloudy*. The BADC total cloud amount was specified in terms of okta, ranging from 0 – *clear sky* to 8 – *100% clouds*. We mapped the BADC grades onto the CLAVR-1 grades in the following way: *clear* – 0-1 okta, *mixed* – 2-5 okta, *cloudy* – 6-8 okta.

B. Experiments

We performed two experiments with the AVHRR data that we selected. The experiments differed in the set of variables that constituted the input to the decision-tree learning procedure. Experiment I included variables that represented sensor data: the radiances of channels 1 through 5, and the BADC label. Experiment II included the variables of Experiment I, as well as three additional function variables that are used within the CLAVR-1 daytime-land algorithm [2] (see [15] for a more detailed description of the functions we used).

We randomly selected approximately 10% of the data points to form a dataset that would be used exclusively as a test set for validation. Then, for each of Experiment I and Experiment II, we used the remaining data to conduct 100 bootstrapped [19] training trials to learn and evaluate multiple decision trees. For each trial, we randomly partitioned the data into a training set and a test set with a size ratio of 9:1. We learned a decision tree from the training set with the *treefit* procedure, which is an implementation of classification and regression trees [17] available within the MATLAB® statistics toolbox. We then classified the data in the corresponding test set as *clear*, *mixed*, or *cloudy* using the decision tree. We compared the classification results to the corresponding BADC observations.

To measure accuracy for each experiment we computed two mismatch rates. First, we computed the rate of mismatch between classification results of the ALDT and the BADC observations. Second, we computed the rate of mismatch between the CLAVR-1 cloud masks and the BADC observations. We ran two-sided paired *t*-tests to determine if there were significant differences between rates of classification mismatch, for CLAVR-1 and for each of the decision trees, and for each pair of decision trees. Finally, we used the ALDT to classify the test set we had initially set aside, and we compared the rate of classification mismatch to that of CLAVR-1.

IV. RESULTS

Table I lists the mean and standard deviation classification mismatch for each experiment. Note that the training sets used in the bootstrapping trials were not independent, and the test sets were not independent as well. However, the validation test set that was set aside initially was independent of all other sets. Columns 3, 4 in the table show mean and standard deviation of the rates of mismatch between CLAVR-1 and the gold standard, and between ALDT and the gold standard. The statistics were calculated for the 100 training trials for the five years. Columns 5, 6 show similar statistics calculated for the validation test set for the five years. In each of the two experiments, the difference in classification-mismatch rates between CLAVR-1 and ALDT was significant $p < 0.001$.

Across experiments, the difference in classification-mismatch rates between ALDT was not statistically significant.

V. DISCUSSION

The two experiments that we performed showed that ALDT classified 8km daily AVHRR data for the years 1996-2000 more accurately than CLAVR-1. The two types of decision trees—trees based on only sensor data, and trees based on sensor data and functions of the sensor data—had similar accuracy. Thus the sensor data alone were sufficient to obtain an improvement over CLAVR-1. The sample data we used was limited in two ways. First, the sample was influenced by the availability of the BADC gold standard. Second, the sample was restricted to a geographical area that had a high prevalence of clouds. In effect, the ALDT were trained to predict BADC observations from AVHRR data. Thus, our ability to conclude the true presence or absence of clouds based on the results of ALDT depended on the accuracy of the BADC observations.

The mismatch rates with respect to the gold standard that we obtained in this study were higher for both CLAVR-1 and ALDT compared to the initial feasibility study [15]. The higher rates could be explained in part by errors due to scale differences. Although scale differences occurred in [15] as well, the final scale in [15] had two values only: *clear* and *cloudy*, and it was CLAVR-1 that was down-scaled, from three to two levels. In this study, the CLAVR-1 scale was coarser than the BADC scale, and to match the CLAVR-1 scale we down-scaled BADC data from nine to three levels. Here, the mapping between scales involved loss of information in the gold-standard.

We chose to use AVHRR data and the CLAVR-1 mask to demonstrate the feasibility and contribution of ALDT because of their relative simplicity. The promising results that we obtained indicate that it would be a worthy effort to replicate the study with data acquired by more advanced instruments such as the Moderate-Resolution Imaging Spectroradiometer (MODIS) and with the corresponding newer versions of the CLAVR mask. Two possible extensions of this work are to use ALDT to classify clouds into multiple cloud types, and to

TABLE I
CLASSIFICATION MISMATCH RATES WITH RESPECT TO THE GOLD STANDARD

Experiment	Method	Mean - training	Standard deviation - training	Mean - validation	Standard deviation - validation
I	CLAVR-1	0.523	0.05	0.494	0.02
I	Decision trees based on only channels	0.287	0.014	0.278	0.02
II	CLAVR-1	0.522	0.048	0.497	0.022
II	Decision trees based on channels, acquisition parameters, and functions	0.286	0.015	0.277	0.013

use regression trees to predict the amount of cloud cover and visual opacities.

VI. CONCLUSION

Our work demonstrated that a sequential testing approach similar to that used by experts, combined with a comprehensive analysis of training data via an automated procedure for learning decision trees, contributed to the development of an improved cloud mask.

ACKNOWLEDGEMENTS

The gold-standard data for this study were obtained from the British Atmospheric Data Centre. We thank Ashok Srivastava, Mark Schwabacher, Nikunj Oza, Dawn McIntosh, and Marianne Mosher for useful discussions and comments on earlier version of this manuscript. This work was funded by the Intelligent Systems / Intelligent Data Understanding program for Virtual Sensors.

REFERENCES

- [1] Wielicki, B.A., Harrison, E.F., Cess, R.D., King, M.D., Randall, D.A. Mission to planet earth: role of clouds and radiation in climate, *Bulletin of the American Meteorological Society*, Vol. 76, No. 11, pp. 2125-2154, 1995.
- [2] Stowe, L.L., Davis, P.A. Davis, and McClain, E.P. Scientific basis and initial evaluation of the CLAVR-1 global clear/cloud classification algorithm for the advanced very high resolution radiometer. *Journal of Atmospheric and Oceanic Technology*, Vol. 16(6), pp. 656-681, 1999.
- [3] Trepte, Q.Z., Minnis, P., and Arduini, R.F. Daytime and nighttime polar cloud and snow identification using MODIS. In Huang, H-L., Lu, D., Sasano, Y. *Proc. SPIE 3rd International Asia-Pacific Environmental Remote Sensing Symposium 2002: Remote Sensing of the Atmosphere, Ocean, Environment, and Space*, Hangzhou, China, October 23-27, pp. 449-459, 2003.
- [4] Stroeve, J. Assessment of greenland albedo variability from the avhrr polar pathfinder data set, *Journal of Geophysical Research*, Vol. 33, pp. 989-1034, 2002.
- [5] Promcharoen, S., Rangsaner, Y., Suwit Ongsomwang, S., Jarupatt, J. Supervised classification of multispectral satellite images using fuzzy logic and neural network. In *Proceeding of the 20th Asian Conference on Remote Sensing*, November 22-25, Hong Kong, China, 1999.
- [6] Murtagh, F., Barreto, D. Marcello, J. Decision boundaries using Bayes factors: the case of cloud masks. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 14, pp. 2952-2958, 2003.
- [7] Lee, Y., Wahba, G., Ackerman, S.A. Cloud classification of satellite radiance data by multicategory support vector machines. *Journal of Atmospheric and Oceanic Technology*, Vol. 21, No. 2, pp. 159-169, 2004.
- [8] Srivastava, A.N. Mixture density Mercer kernels: a method to learn kernels directly from data. In M.W. Berry, U. Dayal, C. Kamath, and D. Skillicorn. (Eds.) *Proceedings of the Fourth SIAM International Conference on Data Mining*. Lake Buena Vista, Florida, April 22-24, 2004.
- [9] Srivastava, A.N., Oza, N.C., and Stroeve, J. Virtual sensors: using data mining techniques to efficiently estimate remote sensing spectra. *IEEE Transactions on Geosciences and Remote Sensing, Advances Special Issue*, 2005. Accepted.
- [10] Garay, M.J., Mazzoni, D., Davies, R., and Diner, D.J., The application of support vector machines to analysis of global satellite data sets from MISR. In *Proceedings of the 4th AMS Conference on Artificial-Intelligence Applications to Environmental Science*, San Diego, Jan 9-13, 2005.
- [11] Brumby S.P., Hirsch K.L., Davis A.B., Harvey N.R., and Rohde C.A. Genetic refinement of cloud-masking algorithms for the multi-spectral thermal imager (MTI). In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, Sydney, Australia, July 9-13, pp. 1152-1154, 2001.
- [12] Hansen, M., Dubayah, R., and DeFries, R. Classification trees: an alternative to traditional land cover classifiers. *International Journal of Remote Sensing*, Vol. 17, pp. 1075-1081, 1996.
- [13] Schwabacher, M., Langley, P. Discovering communicable scientific knowledge from spatio-temporal data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. pp. 489-496. Williamstown, MA: Morgan Kaufmann, 2001.
- [14] Zhan, X., Sohlberg, R.A., Townshend, J.R.G, DiMiceli, C., Carroll M.L., Eastman, J.C., Hansen, M.C. , DeFries, R.S. Detection of land cover changes using MODIS 250m data. *Remote Sensing of Environment*, Vol. 83, No. 1-2, pp. 336-350, 2002.
- [15] Shiffman, S. Cloud detection from satellite imagery: a comparison of expert-generated and automatically generated decision trees. In *Proceedings of the Eighth Workshop on Mining Scientific and Engineering Datasets*, held in conjunction with the 2005 SIAM International Conference on Data Mining, April 21-23, Newport Beach, CA, 2005.
- [16] Agbu, P.A. and James, M.E. *The NOAA/NASA Pathfinder AVHRR Land Data Set User's Manual*. Goddard Distributed Active Archive Center, NASA, Goddard Space Flight Center, Greenbelt, MD, 1994.
- [17] Brieman, L., Friedman, J., Stone, C.J., and Olshen, R.A. Classification and regression trees. Chapman & Hall, CRC Press, Boca Raton, FL, 1984.
- [18] British Atmospheric Data Center web site <http://badc.nerc.ac.uk/home/>, last accessed April 28, 2005.
- [19] Mooney, C.Z., and Duval, R.D. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Newbury Park, CA: Sage Publications, 1993.