ISSN 0001-4370, Oceanology, 2017, Vol. 57, No. 2, pp. 265–269. © Pleiades Publishing, Inc., 2017.
Original Russian Text © M.A. Krinitskiy, 2017, published in Okeanologiya, 2017, Vol. 57, No. 2, pp. 297–302.

MARINE PHYSICS

Application of Machine Learning Methods to the Solar Disk State Detection by All-Sky Images over the Ocean

M. A. Krinitskiy

Institute of Oceanology, Russian Academy of Sciences, Moscow, Russia

e-mail: krinitsky@sail.msk.ru

Received September 22, 2016; in final form, October 6, 2016

Abstract—A new approach to automatic solar disk state detection by all-sky images using machine learning methods is developed and implemented. The efficiency of the most widely used machine learning algorithms is analyzed. The effect of reducing the dimensionality of the feature space on the classification accuracy is estimated. The multilayer artificial neural network model has shown the best accuracy in terms of the true score. The operation result demonstrates the effectiveness of machine learning methods applied to solar disk state detection by all-sky images.

DOI: 10.1134/S0001437017020126

INTRODUCTION

A number of algorithms [10, 12, 13] that yield biased estimates compared with observer evidence [1] are commonly used for automatic estimation of the total cloud cover (TCC) using digital all-sky images (Fig. 1). In [1], it was demonstrated that one of the important factors limiting the accuracy of TCC estimation is the solar disk state (SDS), shown in (Fig. 1). The SDS is also an important characteristic that affects the estimation of short-wavelength radiation fluxes incoming to the ocean surface and is a part of standard meteorological and actinometrical observations.

The SDS is estimated visually during marine and terrestrial observations [3]. According to the method [3], the SDS is divided into four classes: –1 (cloudy, the solar disk is not visible because of dense clouds), 0 (the Sun shines weakly; it can be seen, but objects do not cast shadows, the so-called zero degree Sun), 1 (the Sun shines through clouds, fog, smoke, and dust; objects cast shadows, the so-called first-degree Sun), 2 (there are no clouds, fog, smoke, or dust on the solar disk or in a 5° area from its center, the so-called squared Sun). Automatic and expert SDS detection from an all-sky image is hampered by the fact that there are no objects or their shadows in the frame. In order to estimate an operator's ability to detect the SDS from the photograph, tests were carried out on the sample balanced by SDS classes (1600 total images). As part of the test, human estimation from digital all-sky images was compared with observer evidence obtained in field conditions. It was taken into account that in certain cases, related SDS classes (0 and 1, 1 and 2) are not clearly distinguishable in field conditions. The described test showed that an expert can determine the SDS in a laboratory, i.e., from the information conveyed by the image without actual observation.

Assuming that the SDS determines the statistical characteristics of the image color fields and fields of the synthetic *GrIx* index (the so-called pixel grayness rate index) proposed in [1], we formed a set of numerical features on which the problem of automatic classification by modern machine learning methods is solved with an accuracy higher than the random selection accuracy. Some of the considered algorithms make it possible to achieve over 96% accuracy in terms of the true score. Taking into account the clustering result of the set of objects (digital all-sky images) presented in [1], a compactness hypothesis was formulated in order to correctly use machine learning methods: the objects belonging to the same SDS class form subsets compactly localized in the predictor space.

FORMULATION OF THE PROBLEM AND INITIAL DATA

The problem of SDS detection from an all-sky image is formulated as classification of images into four classes [3]. In this formulation, the model training set is prepared by experts during field observations and in a laboratory. Therefore, the problem is to select and optimize the parameters of the model from a family of machine learning algorithms that makes it possible to attribute a new independent object (a digital all-sky image) to one of the SDS classes.

In this paper, the datasets obtained during field observations and by independent SDS estimation by three experts from the images were used to generate the training set. Original photographs were obtained

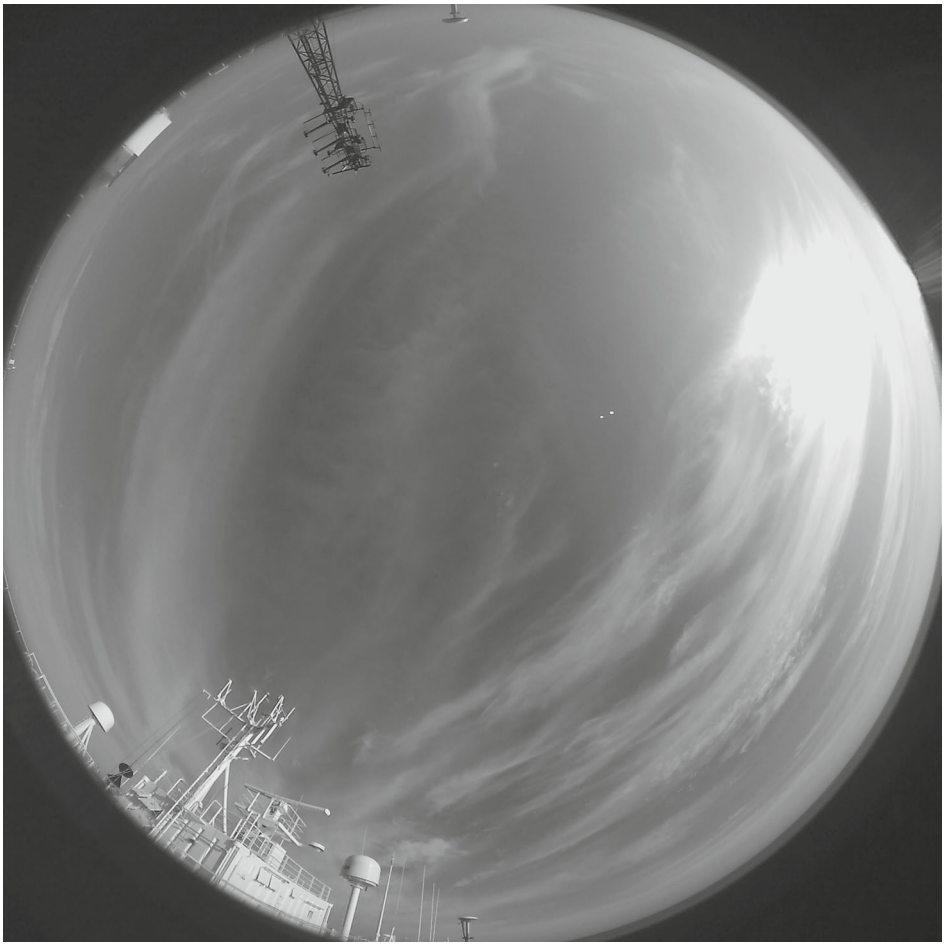

Fig. 1. Digital all-sky image.

by the TCC estimation unit [1] during cruise AI-49 on the R/V *Akademik Ioffe* and during the cruise 31 of the R/V *Akademik Nikolai Strakhov* (Fig. 2) with a recording period of 20 s. Related field observations were carried out with a period of 1 h during daylight hours. The readings from observers were assumed to be accurate SDS classes for images separated in time by less than 5 min from the observation time. At the preprocessing stage, emission objects were filtered for each of the numerical features. The total dataset size was 28724 objects including 15511 observation data and 13213 laboratory estimations. Figure 2 shows the resulting distribution of objects in the dataset by SDS classes. This distribution is uneven. Thus, the classes were balanced in order to generate training subsets. Two methods were used: duplication of samples of rare classes with noise addition to predictors and cutting of the volume of frequent classes. None of these methods showed a decisive advantage over the other in terms of the accuracy of the models.

The fields of the red (R), green (G), and blue (B) image channels in the RGB color model [9], the field of pixel brightness (Y), and the synthetic *GrIx* index calculated for each point of the image by the following formula were used to generate the numeric predictor space:

$$GrIx = 1 - \frac{StdDev(R,G,B)}{Y}, \quad (1)$$

where $StdDev(R, G, B)$ is the standard deviation of a series of values (R, G, B).

In the above fields, the following statistics were calculated: the minimum and maximum values, arithmetic mean, empirical central distribution moments (variance, skewness, and kurtosis), the standard deviation, the 5th–95th percentiles in increments of 5 (5, 10 ... 95), the 99the percentile, and the mean square over the field. In addition, the height and azimuth of the Sun were used. A Euclidean metric of distances between events in the feature space was selected. Thus, a 142-dimensional space was generated by the above-listed real variables.

The most popular algorithms were used to compare the classification accuracy:

—Linear discriminant analysis (LDA) method [7].

—Random forests (RF) method [4, 6].

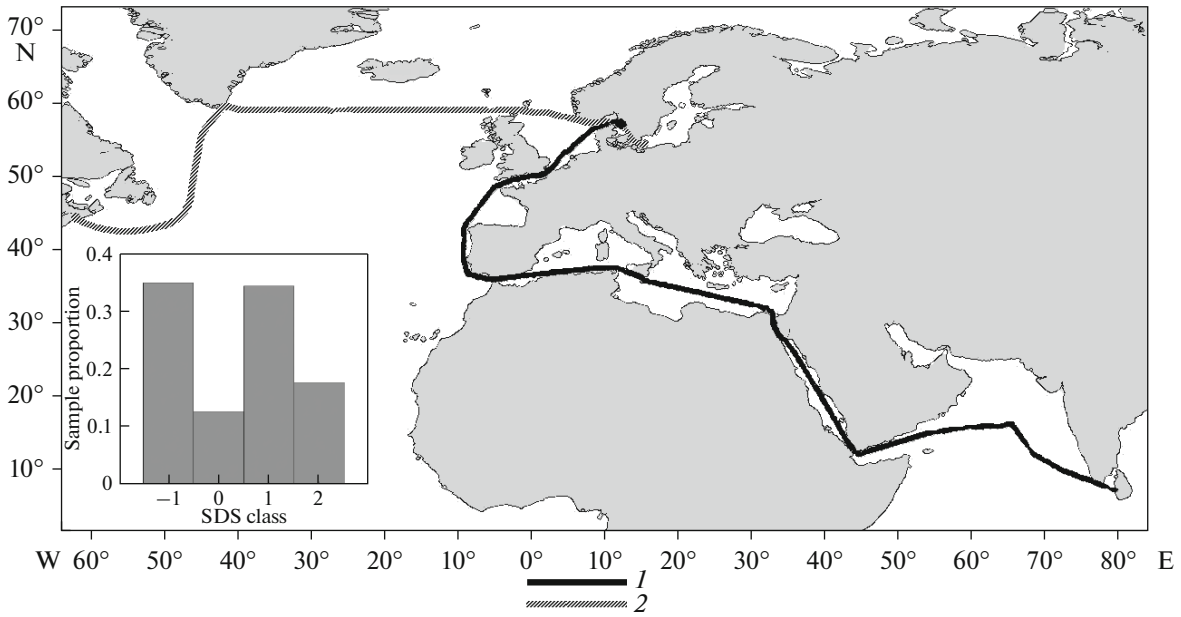

Fig. 2. (1) Route of cruise 31 of R/V *Akademik Nikolai Strakhov* from December 15, 2015, to January 21, 2016; (*2*) route of cruise AI-49 on R/V *Akademik Ioffe* from June 12, 2015, to July 2, 2015, and (inset) distribution of objects of training set by SDS classes.

—Gradient boosting trees (GBT) method [5, 8].

—Deep artificial neural network (DANN) method [2, 5].

Each of these families of machine learning models has its own approach to feature significance estimation. For the families of the LDA, RF, and GBT algorithms, feature ranking by significance estimation was used. The effect of dimensionality N of the space of the most significant predictors on the classification accuracy was investigated. In the case of the DANN model, the optimal brain damage (OBD) approach [11] for the layer of network input parameters was used to estimate feature significance.

Control and selection of the hyperparameters of models by the cross-validation method in stratified form with division into ten blocks were used for reliable estimation of accuracy and control of errors during training and to avoid overfit.

MODEL ACCURACY ASSESSMENT

The accuracy of models was assessed using a hold-out subset in terms of the true score:

$$Acc = \frac{Tc}{Tc + Fc}, \qquad (2)$$

where Tc is the number of classifier responses that coincide with expert responses, and Fc is the number of wrong classifier responses. The hold-out subset amounted to 25% of the total set of events. Figure 3 shows the dependence of the accuracy Acc of the LDA, GBM, and RF models on the dimensionality N of the space of the most significant numerical features.

For the LDA family models, the accuracy did not exceed $Acc = 87\%$ irrespective of the composition of predictors. The best accuracy obtained using the RF family models with a hold-out subset was $Acc = 93.6\%$. For the GBT family models, in some cases an accuracy of $Acc = 94\%$ was achieved. For these algorithms, the increase in accuracy Acc with increasing dimensionality N of the space of real predictors is significantly slowed, since $N = 20$ (Fig. 3). Thus, for the

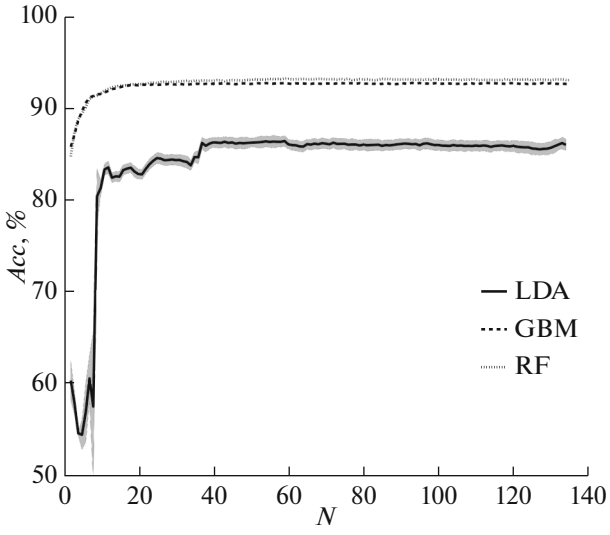

Fig. 3. Classification accuracy in terms of Acc true score as a function of dimensionality N of space of numerical features for models of LDA, GBM, and RF families. In accuracy graph of LDA family models, gray fill indicates confidence intervals at each value N with 95% confidence level.

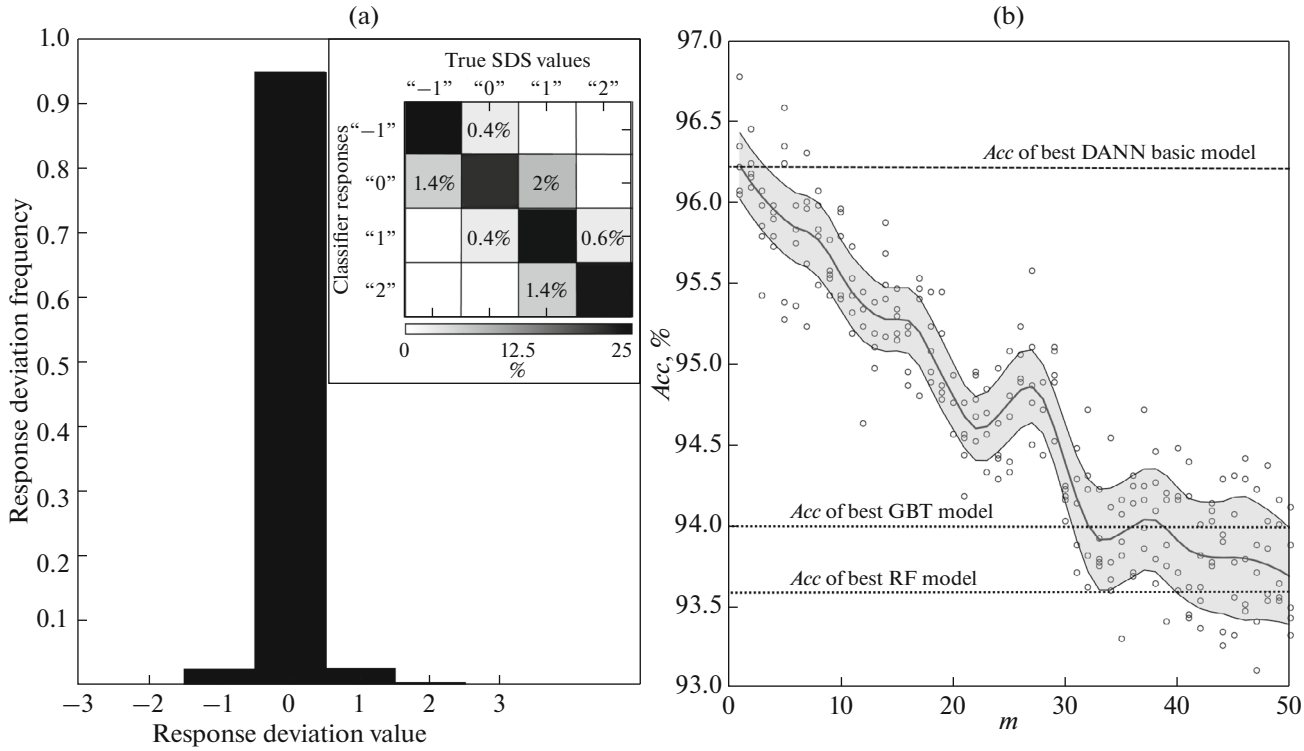

Fig. 4. (a) Distribution of response deviations of best classifier of DANN family compared with observer responses. Confusion matrix of best classifier of DANN family can be found in inset. (b) Accuracy of DANN family models in terms of true score *Acc* within OBD approach after exclusion of *m* minimum significant predictors. Gray fill indicates confidence intervals *Acc* on each value *m* with 95% confidence level.

LDA, RF, and GBT models, using more than 20 of the most significant features in the SDS detection problem makes no practical sense.

DANN FAMILY MODELS

An alternative SDS detection method is based on fully connected deep artificial neural networks [2]. An accuracy of *Acc* = 96.42% was achieved with DANN family algorithms for the full set of features. Figure 4a show, respectively, the distribution of response deviations and the confusion matrix of the best obtained classifiers compared with observer readings.

At the intermediate training stage, the OBD approach [11] is commonly used to estimate the feature significance in the case of DANN family models. This approach does not make it possible to plot the resulting accuracy *Acc* versus *N* in the case of an accumulative increase in the dimensionality of the predictor space. However, the OBD method makes it possible to estimate the significance of each feature separately. The following predictors showed the greatest significance (with a loss of more than 5% of *Acc* in the OBD approach) in descending order: the asymmetry coefficient of the fields and *G* and *Y*, the dispersion of *R* values, and 70th, 15th, 20th, 80th, 75th, 25th, 5th, 35th, and 30th *GrIx* percentiles. Thus, the *GrIx* field is an important variable in SDS classification from all-sky images over the ocean.

The OBD approach implies the possibility of additional training of the DANN model after the exclusion of *m* predictors with the lowest significance. Figure 4 shows the dependence of accuracy *Acc* on *m* excluded features after such a procedure. This graph shows that after additional training on some training subsamples, in the case of reduction of space dimensionality *N,* it is possible to obtain a classification accuracy comparable to the quality with the full set of predictors for *N* = 142. In some cases, the model accuracy on the hold-out subset can exceed the basic one after additional training. However, the general tendency shows that features with a minimum significance estimate should be excluded only within 10–15 predictors without significant loss of accuracy (over 1%).

CONCLUSIONS

The results show that machine learning methods can be effectively used for SDS detection from all-sky images over the ocean. High accuracy is achieved with a space of numerical features generated by the statistics of image color fields and the synthetic *GrIx* index [1], as well as additional predictors calculated based on the coordinates and time of the photo shooting.

It should be noted that the limited accuracy in the SDS detection by the observer can introduce uncertainty in the proposed method. In addition, the accuracy is likely to decrease under shooting conditions that are poorly represented in the training set (e.g., a strong aerosol or dust pollution of the atmosphere).

In the considered problem, the best accuracy in terms of the true score are given by the DANN family models [2], where the image classification accuracy reaches $Acc = 96.42\%$. Simultaneously, the field of the *GrIx* index [1] is one of the most significant variables. The general tendency of accuracy deterioration as predictors are excluded in DANN family models suggests the futility of reducing the dimensionality of the feature space.

High classification accuracy using machine learning algorithms confirms our hypothesis of compactness for objects (digital all-sky images) on the generated space of real predictors, which makes it possible to assume the effectiveness of the proposed methods for similar problems, such as estimation of the total cloud cover, classification of observed cloud cover types, etc.

ACKNOWLEDGMENTS

Calculations and analysis of results were supported by the Russian Science Foundation, project no. 14-50-00095. Preparation and processing of experimental data were supported by the Ministry of Education of the Russian Federation, project no. 14.607.21.0023 (RFMEF160714X0023).

REFERENCES

1. M. A. Krinitskiy and A. V. Sinitsyn, "Adaptive algorithm for cloud cover estimation from all-sky images over the sea," Oceanology (Engl. Transl.) **56**, 315–319 (2016).
2. M. Minsky and S. Papert, *Perceptrons: An introduction to Computational Geometry*, Chap. 13: *Perceptrons and Pattern Recognition* (MIT Press, Cambridge, MA, 1969).
3. *RD 52.04.562-96: Recommendations for Hydrometeorological Stations and Posts*, No. 5, Part 1: *Actinometric Observations at the Stations*, Chap. 7.2: *Meteorological Parameters and Optic Characteristics of Atmosphere Determined by Actinometric Observations* (Rosgidromet, Moscow, 1997), pp. 15–17.
4. S. P. Chistyakov, "Random forests: a review," Tr. Karel. Nauch. Tsentr, Ross. Akad. Nauk, No. 1, 117–136 (2013).
5. L. Breiman, "Bagging predictors," Mach. Learn. **24** (2), 123–140 (1996). doi 10.1023/A:1018054314350
6. L. Breiman, "Random forests," Mach. Learn. **45** (1), 5–32 (2001). doi 10.1023/A:1010933404324
7. R. A. Fisher, "The use of multiple measurements in taxonomic problems," Ann. Eugen. **7** (2), 179–188 (1936).
8. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer-Verlag, New York, 2008), pp. 353–360.
9. G. H. Joblove and D. Greenberg, "Color spaces for computer graphics," *Proceedings of the 5th Annual Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH'78)* (Association for Computing Machinery, New York, 1978), pp. 20–25. doi 10.1145/800248.80736210.1145/800248.807362
10. J. Kalisch and A. Macke, "Estimation of the total cloud cover with high temporal resolution and parameterization of short-term fluctuations of sea surface insolation," Meteorol. Z. **17** (5), 603–611 (2008).
11. Y. Le Cun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in Neural Information Processing Systems 2 (NIPS*89)*, Ed. by D. Touretzky (Morgan Kaufman, Denver, CO, 1990), Vol. 2, pp. 598–605.
12. C. N. Long and J. J. Deluisi, "Development of an automated hemispheric sky imager for cloud fraction retrievals," *10th Symp. on Meteorological Observations and Instrumentation* (Phoenix, AZ, 1998), pp. 171–174.
13. M. Yamashita, M. Yoshimura, and T. Nakashizuka, "Cloud cover estimation using multi-temporal hemisphere imageries," *XXth ISPRS Congress Technical Commission VII, July 12–23, 2004, Istanbul, Turkey* (Istanbul, 2004), Vol. 35, Part B7, pp. 826–829.

Translated by O. Pismenov