*ОКЕАНОЛОГИЯ, 2017, том 57, № 2, с. 297–302*

## ФИЗИКА МОРЯ

*УДК 551.501.776,551.508.761*

# МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧЕ ОПРЕДЕЛЕНИЯ СОСТОЯНИЯ ДИСКА СОЛНЦА ПО ШИРОКОУГОЛЬНЫМ СНИМКАМ ВИДИМОЙ ПОЛУСФЕРЫ НЕБА НАД ОКЕАНОМ

**© 2017 г. М. А. Криницкий**

*Институт океанологии им. П.П. Ширшова РАН, Москва, Россия*

*e-mail: krinitsky@sail.msk.ru*

Поступила в редакцию 22.09.2016 г.
После доработки 06.10.2016 г.

Разработан и реализован новый подход к автоматическому определению состояния диска Солнца по широкоугольному снимку видимой полусферы неба с использованием методов машинного обучения. Проанализирована эффективность наиболее широко используемых алгоритмов машинного обучения, а также оценено влияние понижения размерности пространства признаков на точность классификации. Модель многослойной искусственной нейронной сети показала наилучшее качество по показателю доли верных ответов. Результат работы демонстрирует эффективность использования методов машинного обучения в приложении к задаче определения состояния диска Солнца по широкоугольным снимкам неба.

**DOI:** 10.7868/S0030157417020125

## ВВЕДЕНИЕ

При автоматической оценке общего балла облачности (далее ОБО) по цифровым широкоугольным снимкам видимой полусферы неба (рис. 1) традиционно используется ряд алгоритмов [10, 12, 13], которые дают смещенную оценку по сравнению с показаниями наблюдателя [1]. В [1] продемонстрировано, что одним из важных факторов, ограничивающих точность оценки ОБО, является состояние диска Солнца (далее СДС), представленного на фотографии (рис. 1). СДС также является важной характеристикой, влияющей на оценку приходящих коротковолновых радиационных потоков на поверхность океана и входящей в состав стандартных метеорологических и актинометрических наблюдений.

При проведении морских и наземных наблюдений СДС оценивается визуально [3]. Согласно методике [3], СДС делится на четыре класса: –1 (“пасмурно”, солнечного диска не видно сквозь плотные облака); 0 (Солнце слабо просвечивает, видно место его нахождения, но тени от предметов не наблюдаются; т.н. “Солнце в нулевой степени”); 1 (Солнце просвечивает сквозь облака, туман, дым, пыль; наблюдаются тени от предметов; т.н. “Солнце в первой степени”); 2 (на солнечном диске и в зоне 5° от его центра не заметно следов облаков, тумана, дыма, пыли; т.н. “Солнце в квадрате”). Автоматическое и экспертное определения СДС по широкоугольному снимку неба затруднены отсутствием предметов и их теней в кадре. Для оценки способности оператора определять СДС по снимку было проведено тестирование на выборке, сбалансированной по классам СДС, общим объемом 1600 снимков. В рамках этого тестирования оценка человека по широкоугольной цифровой фотографии сравнивалась с показаниями наблюдателя в натурных условиях. При этом учитывалось, что смежные классы СДС (“0” и “1”, “1” и “2”) в некоторых ситуациях недостаточно четко различимы в натурных условиях. Описанное испытание показало, что специалист способен определить СДС в лабораторных условиях, т.е согласно общим представлениям о передаче визуальной сцены фотоснимком без наблюдения реальной ситуации.

В предположении, что СДС обуславливает статистические характеристики цветовых полей изображения, а также поля синтетического индекса *GrIx* (т.н. “индекс степени серости пикселя”), предложенного в [1], нами был сформирован набор численных признаков, на которых задача автоматической классификации с применением современных методов машинного обучения имеет решение с точностью, превышающей точность случайного выбора. При этом некоторые из рассмотренных алгоритмов позволяют достигать качества по показателю доли верных ответов более 96%. Принимая во внимание результат кластеризации множества объектов (цифровых снимков видимой полусферы неба), приведенный в [1], для корректности использования методов ма-

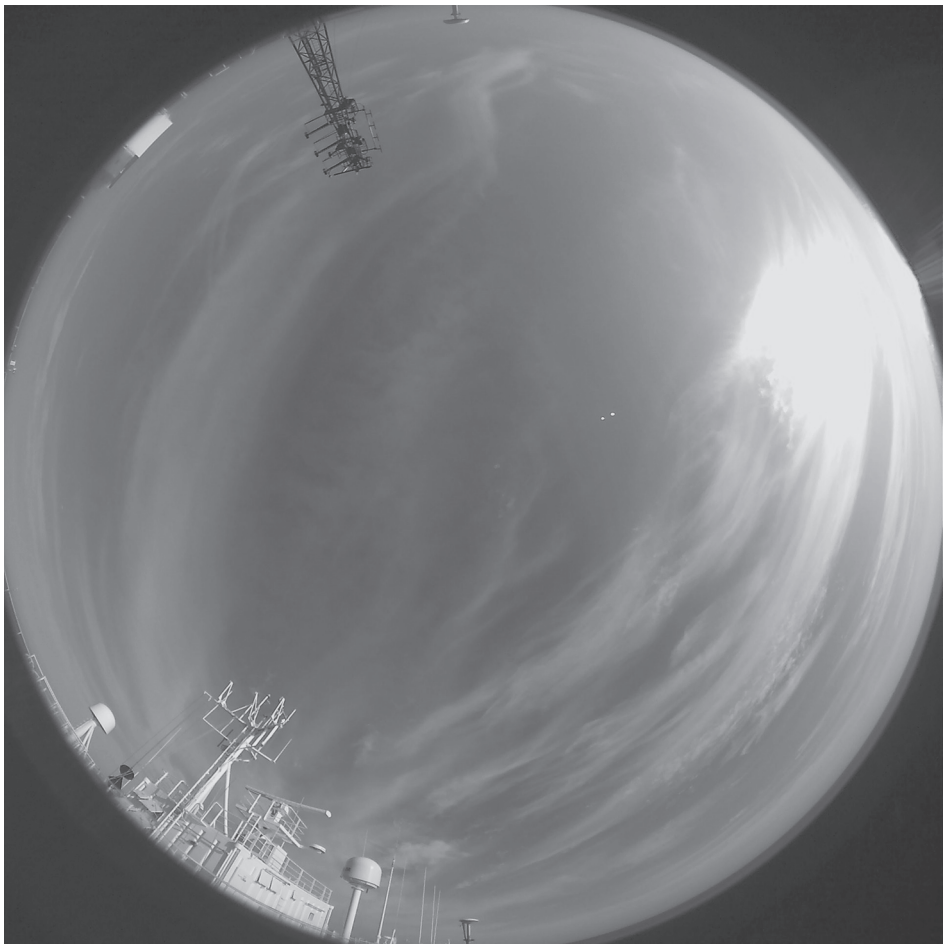

**Рис. 1.** Цифровой широкоугольный оптический снимок видимой полусферы неба.

шинного обучения нами была выдвинута гипотеза компактности: объекты, относящиеся к одному классу СДС, образуют подмножества, компактно локализованные в пространстве предикторов.

## ПОСТАНОВКА ЗАДАЧИ И ИСХОДНЫЕ ДАННЫЕ

Задача определения СДС на широкоугольном снимке видимой полусферы неба сформулирована как классификация фотографий по четырем классам [3]. Тренировочная выборка для обучения моделей в этой постановке предварительно формируется экспертами совокупно в процессе натурных наблюдений и в лабораторных условиях. Задача, таким образом, состоит в подборе и оптимизации параметров модели из семейства алгоритмов машинного обучения, позволяющей отнести новый независимый объект (цифровой широкоугольный снимок неба) к одному из классов СДС.

Для формирования обучающей выборки в работе использовались наборы данных, полученные двумя разными способами: в ходе сопутствующих съемке натурных наблюдений, а также независимой оценкой СДС тремя экспертами по снимкам. Исходные фотографии были получены установкой оценки ОБО [1] в экспедиции АИ-49 НИС “Академик Иоффе” и в экспедиции № 31 НИС “Академик Николай Страхов” (рис. 2) с периодом съемки 20 с. Сопутствующие натурные наблюдения проводились с периодом 1 ч в светлое время суток. Полученные показания наблюдателей считались верными классами СДС для снимков, отстоящих по времени менее чем на 5 мин от момента наблюдения. На этапе предварительной обработки применялась процедура фильтрации объектов-выбросов по каждому из числовых признаков. Общий объем выборки составил 28 724 объекта, включая 15 511 данных наблюдений и 13 213 оценок в лабораторных условиях. Полученное распределение объектов набора данных по классам СДС представлено на рис. 2. Это распределение неравномерное, поэтому для составления тренировочных подмножеств применялась процедура балансировки классов. Были

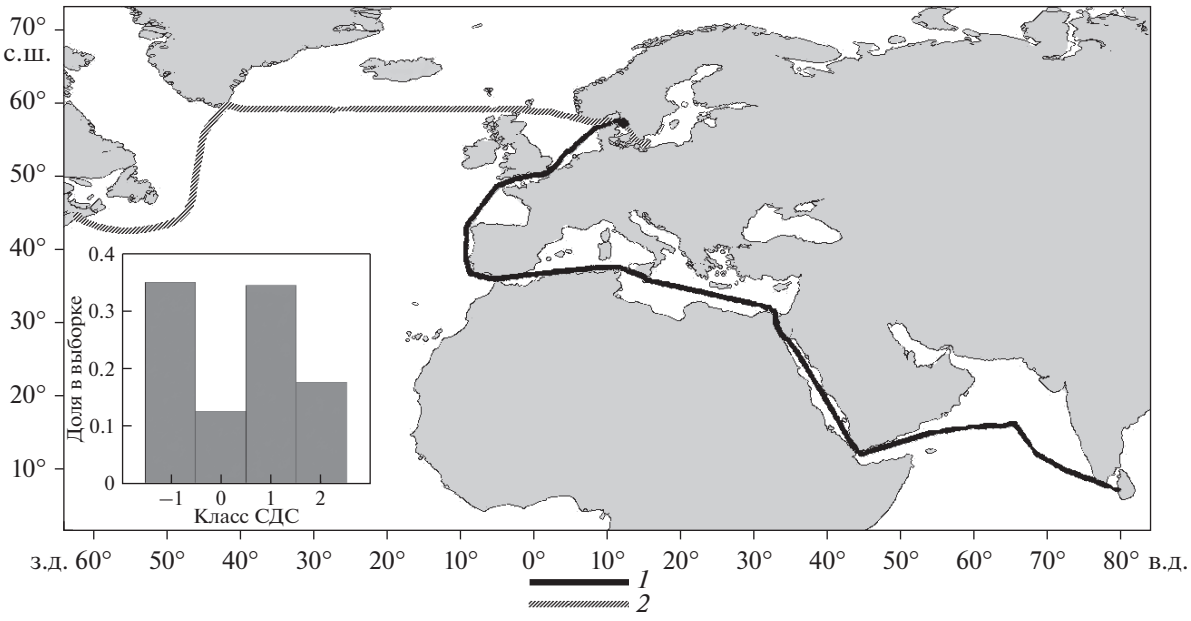


**Рис. 2.** *1* – Маршрут экспедиции № 31 НИС “Академик Николай Страхов” с 15.12.2015 г. по 21.01.2016 г.; *2* – маршрут экспедиции АИ-49 НИС “Академик Иоффе” с 12.06.2015 г. по 02.07.2015 г.; на врезке – распределение объектов обучающей выборки по классам СДС.

использованы два метода: дублирование сэмплов редких классов с зашумлением предикторов и обрезание объема частых классов. Ни один из этих способов не показал решающего преимущества перед другим по точности получаемых моделей.

Для формирования пространства числовых предикторов были использованы поля красного ($R$), зеленого ($G$) и синего ($B$) каналов изображения в цветовой модели RGB [9], поле яркости пикселей ($Y$), а также синтетический индекс $GrIx$, вычисляемый для каждой точки изображения по следующей формуле:

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

где $StdDev(R, G, B)$ – стандартное отклонение ряда из значений ($R$, $G$, $B$).

На вышеперечисленных полях вычислялись следующие статистики: минимальное и максимальное значения, арифметическое среднее, эмпирические центральные моменты распределения (дисперсия, коэффициент асимметрии, коэффициент эксцесса), среднеквадратическое отклонение, набор перцентилей от 5 до 95 с шагом 5 (5, 10 ... 95), перцентиль 99, среднее квадратическое по полю. Кроме того, использовались высота и азимут Солнца. Метрика расстояний между событиями в пространстве признаков была выбрана евклидова. Таким образом, вышеперечисленными вещественными переменными было сформировано 142-мерное пространство.

Для сравнения точности классификации выбирались наиболее популярные алгоритмы:

метод линейного дискриминантного анализа (далее LDA) [7];

метод произвольного ансамбля решающих деревьев (Random Forest, далее RF) [4, 6];

метод градиентного бустинга над решающими деревьями (gradient boosting trees, далее GBT) [5, 8];

метод глубокой искусственной нейронной сети (deep artificial neural network, далее DANN) [2, 5].

Для каждого из вышеперечисленных семейств моделей машинного обучения существует свой подход к оценке значимости признаков. Для семейств алгоритмов LDA, RF и GBT применялось ранжирование признаков по оценке значимости. Было проведено исследование влияния размерности $N$ пространства наиболее значимых предикторов на точность классификации. В случае модели типа DANN для оценки значимости признаков применялся подход “optimal brain damage” (далее OBD) [11] для слоя входных параметров сети.

Для достоверной оценки точности и контроля ошибок в процессе тренировки, а также для исключения переобучения использовался подход контроля и подбора гиперпараметров моделей методом кросс-валидации в стратифицированном варианте с разбиением на 10 блоков.

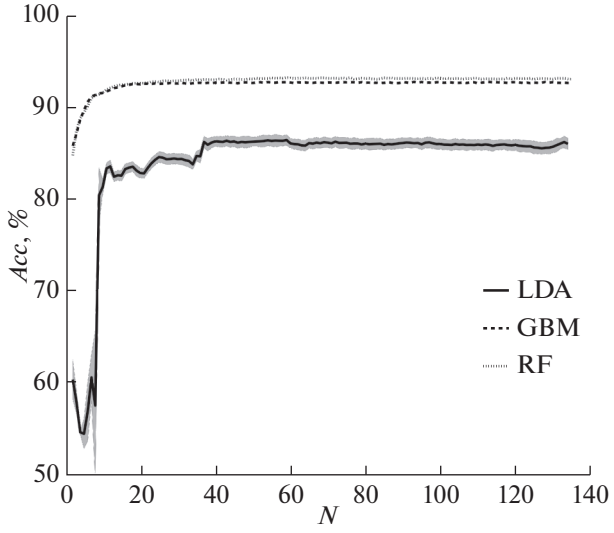


**Рис. 3.** Точность классификации по показателю *Acc* доли верных ответов в зависимости от размерности *N* пространства числовых признаков для моделей из семейств LDA, GBM и RF. Для графика точности моделей семейства LDA серой заливкой обозначены доверительные интервалы на каждом значении *N* с уровнем доверия 95%.

## ОЦЕНКА КАЧЕСТВА МОДЕЛЕЙ

Оценка качества моделей производилась на отложенной выборке по показателю доли верных ответов:

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

где *Tc* – количество ответов классификатора, совпадающих с экспертным, *Fc* – количество ошибочных ответов классификатора. Объем отложенной выборки составлял 25% общей совокупности событий. Зависимость качества *Acc* моделей LDA, GBM и RF от размерности *N* пространства наиболее значимых числовых признаков приведена на рис. 3.

Для моделей из семейства LDA вне зависимости от состава предикторов точность не превысила *Acc* = 87%. Наилучшее качество на отложенной выборке, полученное с использованием моделей семейства RF, составило *Acc* = 93.6%. С использованием моделей семейства GBT в некоторых случаях была достигнута точность *Acc* = 94%. Для этих алгоритмов рост качества по показателю *Acc* с ростом *N* размерности пространства вещественных предикторов существенно замедляется, начиная с *N* = 20 (рис. 3). Таким образом, для моделей LDA, RF и GBT использование больше 20 наиболее значимых признаков в задаче определения СДС не имеет практического смысла.

## МОДЕЛИ СЕМЕЙСТВА DANN

Альтернативный метод определения СДС базируется на полносвязных глубоких искусственных нейронных сетях [2]. С применением алгоритмов семейства DANN на полном составе признаков была достигнута точность *Acc* = 96.42%. Распределение отклонений ответов и матрица ошибок лучшего из полученных классификаторов в сравнении с показаниями наблюдателя представлены на рис. 4а.

Для оценки значимости признаков в случае моделей семейства DANN обычно применяется подход OBD [11] на промежуточном этапе обучения, который не позволяет построить зависимость результирующей точности *Acc* от *N* при накопительном наращивании размерности пространства предикторов. Однако метод OBD дает возможность оценить значимость каждого признака в отдельности. Предикторы, продемонстрировавшие наибольшую значимость (с потерей более 5% по *Acc* в подходе OBD) в порядке убывания значимости: коэффициент асимметрии по полям *G* и *Y*, дисперсия значений *R*, а также перцентили 70, 15, 20, 80, 75, 25, 5, 35, 30 *GrIx*. Таким образом, поле *GrIx* является важной переменной в задаче классификации СДС по широкоугольным снимкам видимой полусферы неба над океаном.

Подход OBD подразумевает возможность дообучения модели DANN после исключения *m* предикторов с наименьшей значимостью. На рис. 4б представлена зависимость точности *Acc* после такой процедуры от *m* исключенных признаков. Этот график показывает, что, дообучаясь на некоторых тренировочных подвыборках, при сокращении размерности пространства *N* возможно получить качество классификации, сравнимое с качеством на полном наборе предикторов при *N* = 142. В некоторых случаях точность модели на отложенной выборке после дообучения может превысить базовую. Однако общая тенденция демонстрирует, что исключать признаки с минимальной оценкой значимости без существенной потери качества (более 1%) следует лишь в пределах 10–15 предикторов.

## ВЫВОДЫ

Полученные результаты показывают, что методы машинного обучения могут эффективно использоваться в задаче определения СДС по широкоугольным снимкам видимой полусферы неба над океаном. Высокая точность достигается на пространстве числовых признаков, сформированных на основании статистик цветовых полей изображения и синтетического индекса *GrIx* [1], а также дополнительных предикторов, вычисляемых на основании координат и времени съемки.

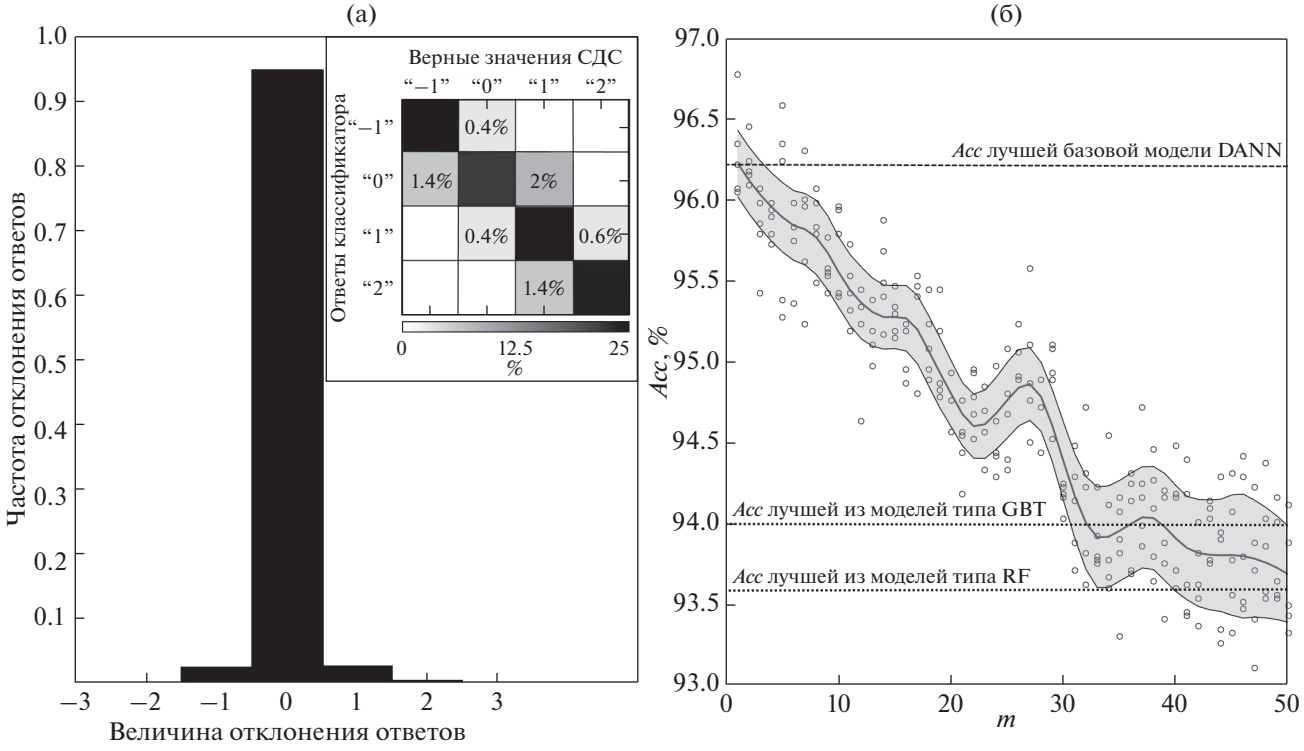


**Рис. 4.** (а) – Распределение отклонений ответов лучшего классификатора семейства DANN в сравнении с показаниями наблюдателя; на врезке – матрица ошибок лучшего классификатора семейства DANN; (б) – качество моделей семейства DANN по показателю *Acc* доли верных ответов в подходе OBD при исключении количества *m* предикторов минимальной значимости. Серой заливкой обозначены доверительные интервалы *Acc* на каждом значении *m* с уровнем доверия 95%.

Следует отметить, что ограниченная точность в определении СДС наблюдателем может вносить неопределенность в предложенный метод. Кроме того, вероятно снижение точности при определении в условиях съемки, редко представленных в тренировочной выборке (например, сильная аэрозольная или пылевая загрязненность атмосферы).

В рассматриваемой задаче лучшее качество по показателю доли верных ответов дают модели семейства DANN [2], где точность классификации снимков достигает $Acc = 96.42\%$. При этом поле индекса *GrIx* [1] оказывается одной из наиболее значимых переменных. Общая тенденция потери качества по мере исключения предикторов в моделях семейства DANN позволяет говорить о бесперспективности сокращения размерности пространства признаков.

Высокое качество классификации при использовании алгоритмов машинного обучения подтверждает выдвинутую нами гипотезу компактности для объектов (цифровых широкоугольных снимков видимой полусферы неба) на сформированном пространстве вещественных предикторов, что позволяет выдвинуть предположение об эффективности предложенных методов в аналогичных задачах, таких как оценка общего балла облачности, классификация наблюдаемых типов облачности и др.

Проведение вычислительной части исследований и анализ результатов выполнены за счет РНФ (проект № 14-50-00095). Подготовка и обработка экспериментальных данных осуществлена в рамках проекта Минобрнауки РФ 14.607.21.0023, идентификационный номер RFMEF160714X0023.

## СПИСОК ЛИТЕРАТУРЫ

1. *Криницкий М.А., Синицын А.В.* Адаптивный алгоритм оценки общего балла облачности над морем по широкоугольным снимкам неба // Океанология. 2016. Т. 56. № 3. С. 341–345.
2. *Минский М., Пейперт С.* Персептроны. Гл. 13. Персептроны и распознавание образов. М.: Мир, 1971. С. 226–231.
3. РД 52.04.562-96. Наставление гидрометеорологическим станциям и постам. Выпуск 5, часть 1. Актинометрические наблюдения на станциях. Гл. 7.2. Метеорологические параметры и оптические характеристики атмосферы, определяемые при выполнении актинометрических наблюдений. М.: Росгидромет, 1997. С. 15–17.
4. *Чистяков С.П.* Случайные леса: обзор // Тр. КарНЦ РАН. 2013. № 1. С. 117–136.

5. *Breiman L.* Bagging predictors // Mach. Learn. 1996. V. 24. № 2. P. 123–140. doi 10.1023/A:1018054314350
6. *Breiman L.* Random Forests // Mach. Learn. 2001. V. 45. № 1. P. 5–32. doi 10.1023/A:1010933404324
7. *Fisher R.A.* The use of multiple measurements in taxonomic problems // Ann. Eugen. 1936. V. 7. № 2. P. 179–188.
8. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Springer. 2008. P. 353–360.
9. *Joblove G.H., Greenberg D.* Color spaces for computer graphics // Proceedings of the 5th annual conference on Computer graphics and interactive techniques (SIGGRAPH '78). ACM, New York, NY, USA. 1978. P. 20–25. doi 10.1145/800248.807362
10. *Kalisch J., Macke A.* Estimation of the total cloud cover with high temporal resolution and parameterization of short-term fluctuations of sea surface insolation // Meteorologische Zeitschrift. 2008. V. 17. № 5. P. 603–611.
11. *Le Cun Y., Denker J.S., Solla S.A.* Optimal Brain Damage // Touretzky, David (Eds), Advances in Neural Information Processing Systems 2 (NIPS*89). Morgan Kaufman, Denver, CO. 1990. V. 2. P. 598–605.
12. *Long C.N., Deluisi J.J.* Development of an automated hemispheric sky imager for cloud fraction retrievals // 10th Symp. on Meteorological Observations and Instrumentation, January 11 to 16. 1998, Phoenix, AZ, USA. P. 171–174.
13. *Yamashita M., Yoshimura M., Nakashizuka T.* Cloud cover estimation using multi-temporal hemisphere imageries. // International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences. 2004. V. 35. Part B7. P. 826–829.

## Machine Learning Methods Applied to Solar Disk State Detection using All-Sky Images over the Ocean

**M. A. Krinitskiy**

New approach has been developed and implemented for automatic determination of solar disk state from all-sky optical images with the use of machine learning techniques. Efficiency of the most widely used machine learning algorithms has been analysed. Influence of dimensionallity reduction of the feature space on classification accuracy has been estimated. Multilayer artificial neural network model demonstrated the best accuracy. The obtained results demonstrate applicability of the machine learning approach to determination of the solar disk state using all-sky optical images.