

## ОСОБЕННОСТИ ДИАГНОСТИКИ ТЕХНИЧЕСКИХ СИСТЕМ С ИСПОЛЬЗОВАНИЕМ МУЛЬТИКЛАССОВОЙ КЛАССИФИКАЦИИ

В. Н. Клячкин<sup>1</sup>, И. Н. Карпунина<sup>2</sup>

<sup>1</sup> Ульяновский государственный технический университет, Ульяновск, Россия

<sup>2</sup> Ульяновский институт гражданской авиации, Ульяновск, Россия

<sup>1</sup> v\_kl@mail.ru, <sup>2</sup> karpunina531@yandex.ru

**Аннотация.** *Актуальность и цель.* Постоянно растущие требования по обеспечению безопасности и надежности технических систем приводят к необходимости более точной диагностики состояния объекта в условиях эксплуатации по результатам мониторинга показателей функционирования этого объекта. Иногда необходимо состояние объекта описать с помощью нескольких возможных вариантов. В этом случае проводится мультиклассовая классификация, при которой возможные состояния объекта подразделяются на несколько классов, например, по виду отказа. При этом могут эффективно использоваться методы машинного обучения. Особенности рассматриваемой задачи – ограниченный объем выборочных данных, а также несбалансированность обучающей выборки: информации о показателях функционирования при неработоспособных состояниях объекта, как правило, гораздо меньше, чем при работоспособных. Цель исследования – разработка технологии диагностики состояния технического объекта по заданным показателям его функционирования с учетом этих особенностей. *Материалы и методы.* Среди используемых методов машинного обучения для мультиклассовой классификации следует отметить как стандартные статистические, так и специальные: нейронные сети, композиционные модели, агрегированные классификаторы. В данной работе для мультиклассовой классификации применен метод Random Forest («случайный лес»), показавший высокое качество при решении различных задач машинного обучения. *Результаты и выводы.* Разработана технология мультиклассовой диагностики технических систем с применением «случайного леса» в системе Statistica. На примере диагностики компьютерной системы показано, что использование этого метода обеспечивает достаточно высокую точность классификации. В случае несбалансированности классов в качестве критерия классификации вместо доли ошибок используется  $F$ -мера. При необходимости количество показателей функционирования может быть сокращено с учетом их значимости.

**Ключевые слова:** техническая диагностика, показатели функционирования, машинное обучение, случайный лес, система Statistica

**Для цитирования:** Клячкин В. Н., Карпунина И. Н. Особенности диагностики технических систем с использованием мультиклассовой классификации // Надежность и качество сложных систем. 2022. № 2. С. 45–52. doi:10.21685/2307-4205-2022-2-5

## FEATURES OF DIAGNOSTICS OF TECHNICAL SYSTEMS USING MULTICLASS CLASSIFICATION

V.N. Klyachkin<sup>1</sup>, I.N. Karpunina<sup>2</sup>

<sup>1</sup> Ulyanovsk State Technical University, Ulyanovsk, Russia

<sup>2</sup> Ulyanovsk Institute of Civil Aviation, Ulyanovsk, Russia

<sup>1</sup> v\_kl@mail.ru, <sup>2</sup> karpunina531@yandex.ru

**Abstract.** *Background.* Constantly growing requirements for ensuring the safety and reliability of technical systems lead to the need for more accurate diagnostics of the state of the facility in operating conditions based on the results of monitoring the performance of this facility. Sometimes it is necessary to describe the state of an object using several possible options. In this case, a multiclass classification is carried out, in which the possible states of the object are divided into several classes, for example, by the type of failure. At the same time, machine learning methods can be effectively used. Features of the problem under consideration are a limited amount of sample data, as well as the imbalance of the training sample: information on the performance indicators of functioning in inoperable states of the object, as a rule, is much less than in the case of able-bodied ones. The purpose of the study is to develop a technology for diagnosing the state of a technical object according to the specified indicators of its functioning, taking into account these features. *Materials and methods.* Among the machine learning methods used for multiclass classification

are both standard statistical and special: neural networks, compositional models, aggregated classifiers. In this paper, the Random Forest method was used for multiclass classification, which showed high quality in solving various machine learning problems. *Results and conclusions.* A technology for multiclass diagnostics of technical systems using a random forest in the Statistica system has been developed. On the example of diagnostics of a computer system, it is shown that the use of this method provides a sufficiently high accuracy of classification. In case of class imbalance,  $F$  is used as a classification criterion instead of the error rate. – Measure. If necessary, the number of performance indicators can be reduced taking into account their importance.

**Keywords:** technical diagnostics, performance indicators, machine learning, random forest, Statistica system

**For citation:** Klyachkin V.N., Karpunina I.N. Features of diagnostics of technical systems using multiclass classification. *Nadezhnost' i kachestvo slozhnykh sistem = Reliability and quality of complex systems.* 2022;(2):45–52. (In Russ.). doi:10.21685/2307-4205-2022-2-5

### Постановка задачи

Постоянно растущие требования по обеспечению безопасности и надежности технических систем приводят к необходимости более точной диагностики состояния объекта в условиях эксплуатации по результатам мониторинга показателей функционирования этого объекта [1–4]. Современные компьютерные технологии, в частности, методы машинного обучения, дают возможность решить эту задачу [2, 5].

Диагностика может сводиться к распознаванию одного из двух состояний: к разделению состояний объекта на исправные или неисправные. В этом случае применяются методы бинарной классификации. Однако иногда необходимо более детальное исследование, когда состояние объекта необходимо описать с помощью нескольких возможных вариантов (а не только двух: исправен объект или неисправен). Например, такая задача решалась с применением нейронных сетей при классификации режимов работы авиационного газотурбинного двигателя [6, 7]. В этом случае проводится мультиклассовая классификация, при которой возможные состояния объекта подразделяются на несколько классов, например, по виду отказа. Как и при бинарной классификации здесь могут эффективно использоваться методы машинного обучения.

Выборка, полученная по результатам предварительных испытаний технического объекта (это могут быть как специально поставленный эксперимент, так и исследование в процессе эксплуатации), разбивается на две части: обучающую и тестовую. Обучающая часть предназначена для построения моделей, с помощью которых состояние объекта разделяется на заданное количество классов: требуется построить алгоритм, который для заданного набора показателей функционирования обеспечил бы достаточно точный результат о состоянии объекта. Тестовая часть выборки предназначена для оценки качества классификации.

Среди используемых методов следует отметить как стандартные статистические, так и специальные: нейронные сети, композиционные модели, агрегированные классификаторы [8–11]. Наиболее эффективно, как показывают исследования, глубокое обучение нейронных сетей, однако два обстоятельства часто препятствуют их практическому применению.

Первое – это необходимость достаточно большого объема выборочных данных, что для технических систем, как правило, нереально. Второе состоит в том, что для глубокого обучения необходима мощная вычислительная техника, которая далеко не всегда имеется в распоряжении исследователя в производственных условиях.

При бинарной классификации хорошие результаты показывают агрегированные методы, однако это не подтвердилось в условиях разделения состояния объекта на множество классов.

В рассматриваемой ситуации наиболее эффективно применение композиционных методов. В данной работе для мультиклассовой классификации применен метод Random Forest («случайный лес»), показавший высокое качество при решении различных задач машинного обучения [9].

Еще одна существенная особенность рассматриваемой задачи состоит в несбалансированности обучающей выборки: информации о показателях функционирования при неработоспособных состояниях объекта, как правило, гораздо меньше, чем при работоспособных. При сбалансированных данных критерием качества модели обычно служит функционал ошибок – среднее количество несопадений фактического и прогнозируемого состояний, или доля ошибок на тестовой выборке.

Для учета несбалансированности классов критерием качества классификации следует выбирать  $F$ -меру: гармоническое среднее между точностью и полнотой классификации; чем это значение ближе к единице, тем качество классификации лучше. Иногда, например, при совпадении значений этой меры для различных методов можно дополнительно использовать в качестве критерия площадь под кривой ошибок AUC [12–14].

При большом количестве показателей, характеризующих качество функционирования технической системы (иногда их может быть несколько десятков), целесообразно отобрать некоторое их подмножество. Для решения этой задачи могут быть построены регрессионные модели зависимости отклика (класса) от показателей функционирования, при этом значимые показатели отбираются по критерию Стьюдента. С другой стороны, значимость показателей может быть оценена и непосредственно средствами случайного леса.

Выборка представляется в виде матрицы  $X$  показателей функционирования системы, элементы которой  $x_{ij}$  – результат  $i$ -го наблюдения по  $j$ -му показателю;  $i = 1, \dots, l, j = 1, \dots, p$  ( $l$  – количество строк, или число наблюдений,  $p$  – количество столбцов, или число показателей), и вектора-столбца ответов  $Y$ , состоящего из номеров классов. Каждой строке  $x_i$  матрицы  $X$  соответствует определенное значение  $y_i$  вектора  $Y$ . Совокупность пар  $(x_i, y_i)$  образует выборку исходных данных – прецедентов.

Задача состоит в построении модели, которая по заданной строке показателей функционирования  $x_i$  предскажет номер класса  $y_i$ , определяющего состояние рассматриваемого объекта.

Цель исследования – разработка технологии диагностики состояния технического объекта по заданным показателям его функционирования.

### Метод Random Forest для мультиклассовой классификации

Random Forest, или случайный лес, – это алгоритм машинного обучения, предложенный Л. Брейманом [9], он использует ансамбль (комитет) решающих деревьев. Алгоритм сочетает в себе бэггинг (случайный выбор с возвращением) и метод случайных подпространств. Он состоит из множества независимых деревьев решений, при этом используется случайная выборка наблюдений из обучающего набора и случайный набор показателей при принятии решений о разбиении узлов. Случайный лес применяется для решения задач классификации, регрессии и кластеризации.

Классификация объектов проводится путем голосования: каждое дерево комитета относит классифицируемый объект к одному из классов, а побеждает класс, за который проголосовало наибольшее число деревьев. Оптимальное число деревьев подбирается таким образом, чтобы минимизировать ошибку классификатора на тестовой выборке.

Метод имеет высокую точность предсказания, нечувствителен к монотонным преобразованиям значений показателей, редко переобучается: добавление деревьев почти всегда только улучшает композицию, но после достижения определенного количества деревьев кривая обучения выходит на асимптоту. К недостаткам относят то, что в отличие от одного дерева, результаты случайного леса сложнее интерпретировать; кроме того, требуется много памяти для хранения модели вследствие большого размера получающихся моделей.

Точность прогнозирования случайного леса зависит от ряда факторов – количества показателей в случайном наборе, объема подвыборки (случайной выборки из обучающего набора), количества деревьев, максимальной глубины деревьев (максимального количества уровней), максимального количества узлов в деревьях, минимального числа объектов в листьях, минимального количества объектов в дочернем узле.

Модуль случайного леса включен в библиотеку scikit-learn в языке Python, поэтому задача легко программируется. Однако часто удобнее воспользоваться реализацией этого метода, включенной в систему Statistica [15], который позволяет варьировать различные факторы, влияющие на качество обучения, и может быть использован для решения поставленной задачи мультиклассовой классификации для диагностики состояния технического объекта по результатам мониторинга показателей его функционирования.

### Численное исследование

Для наблюдения за процессом функционирования компьютера проводился тест его характеристик с помощью встроенных функций программы AIDA64 Extreme вместе с параллельной нагрузкой другими программами в течение определенного периода времени. Исследовалось влияние различных показателей работы компьютера (загрузки и температуры процессора и ядер, динамической памяти, напряжения, мощности и других – всего 9 показателей  $X_1 \dots X_9$ ) на его состояние: класс 1 – компьютер исправен и 4 вида неисправностей: класс 2 – имело место зависание, класс 3 – искажение изображения, 4 – заторможенность курсора, 5 – заедание звука.

В табл. 1 показана часть исходных данных, из которых случайным образом формировалась обучающая и тестовая выборки. Первые девять столбцов – это матрица  $X$  показателей функционирования системы, десятый – вектор-столбец  $Y$  – номера классов.

Таблица 1

Часть исходной выборки

X1	X2	X3	X4	X5	X6	X7	X8	X9	Y
8	23	15	0	113	50	45	0,817	4,78	1
25	17	17	10	97	50	52	1,013	2,69	1
15	38	15	202	97	50	44	0,693	5,88	1
25	44	7	10	101	46	43	0,693	2,27	1
22	29	18	202	101	46	51	1,015	1,79	1
25	44	25	202	102	46	44	0,908	2,08	1
25	100	100	1252	184	61	60	1,013	5,58	2
25	100	100	1252	184	60	62	1,009	10,92	2
25	100	100	1252	184	63	65	1,009	10,83	2
25	100	48	1752	171	58	61	1,013	8,17	3
25	100	100	1252	128	66	70	1,009	10,43	3
8	100	100	1252	83	49	68	1,009	6,69	5
25	100	100	1252	85	64	64	1,009	10,81	5
25	100	100	1252	85	64	62	1,013	10,71	5
8	12	12	1752	76	65	51	0,688	9,75	1
8	11	18	1752	76	58	57	1,011	6,49	1
22	25	29	1752	82	48	53	1,011	1,51	1

Всего получено 345 наблюдений, из которых нарушения имели место в 126 случаях (37 %), распределение нарушений по классам представлено в табл. 2. Например, состояние  $y = 3$  (искажение изображения) имело место в 32 наблюдениях.

Таблица 2

Распределение классов

Класс	Количество
1	219
2	36
3	32
4	29
5	29

Таким образом, задача исследования – используя матрицу  $X$  показателей функционирования компьютерной системы размерностью 345 строк и 9 столбцов и вектор-столбец ответов о состоянии системы  $Y$ , разработать модель классификатора, которая обеспечивала бы по введенным данным функционирования компьютера диагностику одного из пяти возможных состояний.

Для решения задачи использовался модуль «случайный лес» локализованной версии 13.3 пакета Statistica. Настройки показаны на рис. 1: приняты доля тестовой выборки – 0,3 (30 % от всего набора данных), доля подвыборки – 0,5; число случайных показателей (предикторов) – 4 и т.д.

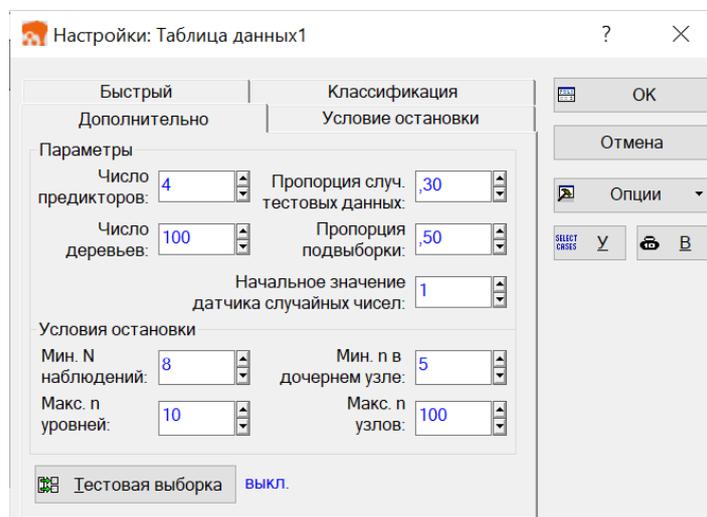


Рис. 1. Настройки «случайного леса»

На рис. 2 показан ход процесса построения «случайного леса»: по горизонтальной оси отложено количество деревьев, по вертикальной – доля ошибок классификации на обучающей (синяя линия) и тестовой (красная линия) выборках. Видно, что процесс стабилизировался уже при 40 деревьях (хотя в настройках указано 100 деревьев, при необходимости для повышения точности можно добавить заданное количество деревьев): при этом на обучающей выборке ошибок нет, на тестовой выборке доля ошибок 0,01.

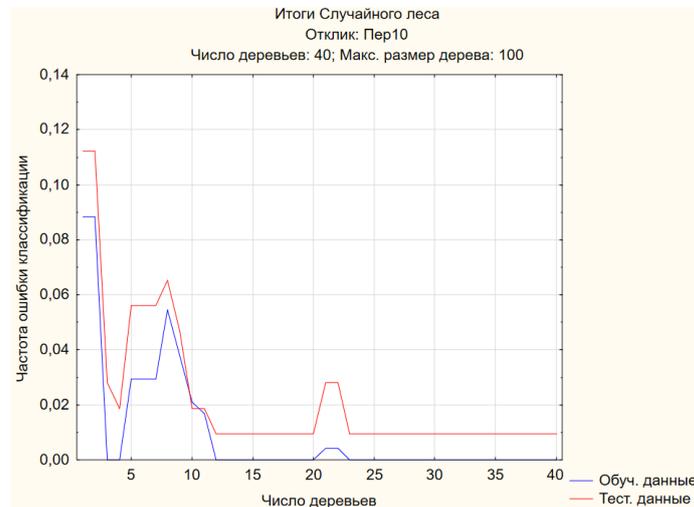


Рис. 2. Процесс построения «случайного леса»

Матрица классификации для тестовой выборки показана на рис. 3: наблюдение 3-го класса ошибочно отнесено к первому; остальные наблюдения классифицированы верно: это 57 наблюдений класса 1, когда компьютер работал исправно, 15 наблюдений 2-го класса, 14 – 4-го класса и 10 наблюдений 5-го класса.

Матрица классификации (к)					
Отклик: Пер10					
Тестовая выборка; Число деревьев: 40					
	Класс Предск. 1	Класс Предск. 2	Класс Предск. 3	Класс Предск. 4	Класс Предск. 5
Наблюд. 1	57,00000				
Наблюд. 2		15,00000			
Наблюд. 3	1,00000		10,00000		
Наблюд. 4				14,00000	
Наблюд. 5					10,00000

Рис. 3. Матрица классификации

Программа автоматически ранжирует показатели функционирования по значимости (рис. 4), что позволяет при большом их количестве сократить наименее значимые показатели. В нашем исследовании наиболее значимыми оказались показатели X9 и X5, наименее значимым – X4. Из графика дерева (рис. 5) видно, что вначале разделение проводится по показателю X9, затем – X5 и т.д.

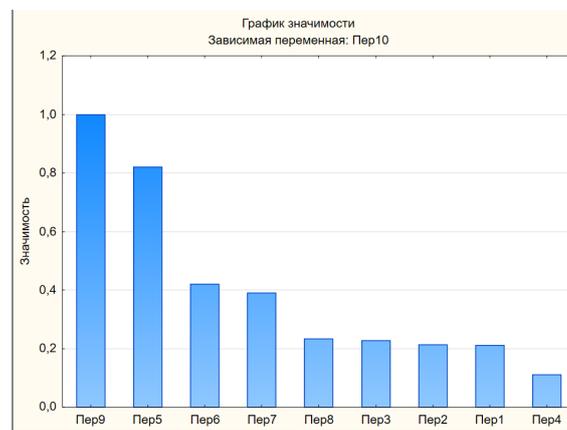


Рис. 4. Гистограмма значимости показателей

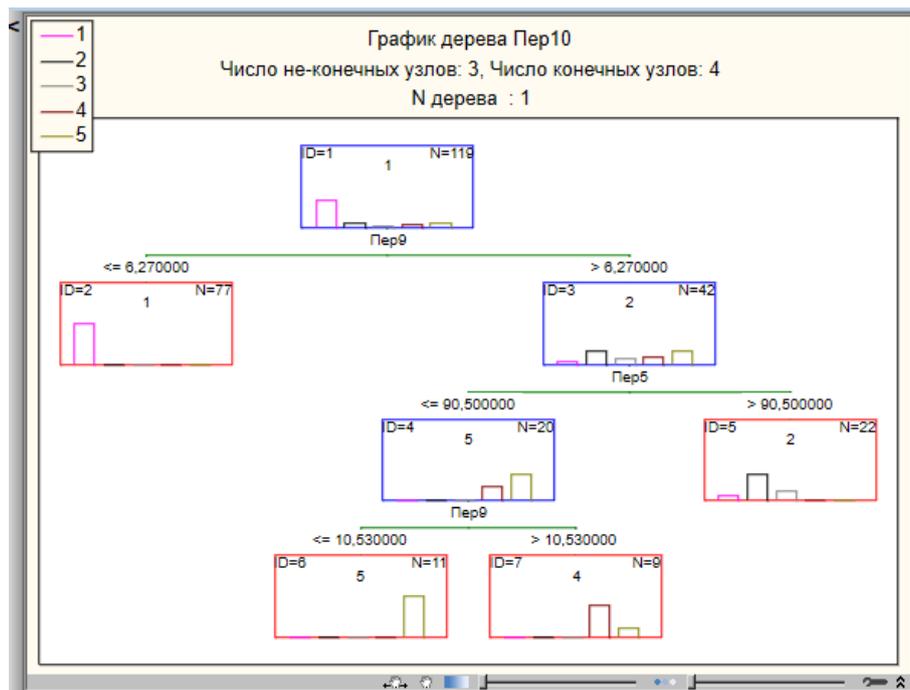


Рис. 5. График дерева

### Заключение

Разработана технология мультиклассовой диагностики технических систем с применением «случайного леса» в системе Statistica. На примере диагностики компьютерной системы показано, что использование этого метода обеспечивает достаточно высокую точность классификации: доля ошибок на тестовой выборке составила 0,01. В случае несбалансированности классов в качестве критерия классификации вместо доли ошибок используется  $F$ -мера. Точность классификации повышается путем добавления деревьев. При необходимости количество показателей функционирования может быть сокращено с учетом их значимости.

### Список литературы

1. Биргер И. А. Техническая диагностика. М. : Машиностроение, 1978. 240 с.
2. Клячкин В. Н., Крашенинников В. Р., Кувайскова Ю. Е. Прогнозирование и диагностика стабильности функционирования технических объектов. М. : Русайнс, 2020. 200 с.
3. Северцев Н. А., Бецов А. В., Дарьина А. Н. Методы и модели создания автоматизированных средств контроля для повышения безопасности функционирования технических систем // Надежность и качество сложных систем. 2019. № 2. С. 19–26.
4. Юрков Н. К. Риски отказов сложных систем // Надежность и качество сложных систем. 2014. № 1. С. 18–24.
5. Воронина В. В., Михеев А. В., Ярушкина Н. Г., Святов К. В. Теория и практика машинного обучения : учеб. пособие. Ульяновск : УлГТУ, 2017. 290 с.
6. Васильев В. И., Жернаков С. В. Классификация режимов работы ГТД с использованием технологии нейронных сетей // Вестник Уфимского государственного авиационного технического университета. 2009. Т. 12. С. 53–56.
7. Жернаков С. В., Гильманшин А. Т. Применение интеллектуальных алгоритмов на основе нечеткой логики и нейронных сетей для решения задач диагностики отказов авиационного ГТД // Intelligent Technologies for Information Processing and Management (ITIPM'2014) : proceedings of the 2<sup>nd</sup> International Conference. 2014. P. 112–115.
8. Жуков Д. А. Анализ критериев качества классификации при диагностике функционирования технического объекта // Автоматизация процессов управления. 2019. № 3. С. 112–117.
9. Breiman L. Random Forests // Machine Learning. 2001. Vol. 45, № 1. P. 5–32.
10. Neykov M., Liu J. S., Cai T. On the Characterization of a Class of Fisher-Consistent Loss Functions and its Application to Boosting // Journal of Machine Learning Research. 2016. № 17. P. 1–32.
11. Wyner A. J., Olson M., Bleich J., Mease D. Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers // Journal of Machine Learning Research. 2017. № 18. P. 1–33.

12. Davis J., Goadrich M. The relationship between Precision-Recall and ROC curves // Proceedings of the 23rd international conference on Machine learning. Pittsburgh. 2006. P. 233–240.
13. Wang B., Zou H. A Multicategory Kernel Distance Weighted Discrimination Method for Multiclass Classification // *Technometrics*. 2019. Vol. 61. P. 396–408. doi: 10.1080/00401706.2018.1529629
14. Клячкин В. Н., Донцова Ю. С. Сравнительный анализ точности нелинейных моделей при прогнозировании состояния системы на основе марковской цепи // *Известия Самарского научного центра Российской академии наук*. 2013. Т. 15, № 4-4. С. 924–927.
15. Боровиков В. П. *STATISTICA. Искусство анализа данных на компьютере*. 2-е изд. СПб. : Питер, 2003. 700 с.

### References

1. Birger I.A. *Tekhnicheskaya diagnostika = Technical diagnostics*. Moscow: Mashinostroenie, 1978:240. (In Russ.)
2. Klyachkin V.N., Krashennnikov V.R., Kuvayskova Yu.E. *Prognozirovanie i diagnostika stabil'nosti funkcionirovaniya tekhnicheskikh ob"ektov = Forecasting and diagnostics of the stability of the functioning of technical facilities*. Moscow: Rusayns, 2020:200. (In Russ.)
3. Severtsev N.A., Betskov A.V., Dar'ina A.N. Methods and models of creation of automated control tools to improve the safety of functioning of technical systems. *Nadezhnost' i kachestvo slozhnykh system = Reliability and quality of complex systems*. 2019;(2):19–26. (In Russ.)
4. Yurkov N.K. Risks of failures of complex systems. *Nadezhnost' i kachestvo slozhnykh system = Reliability and quality of complex systems*. 2014;(1):18–24. (In Russ.)
5. Voronina V.V., Mikheev A.V., Yarushkina N.G., Svyatov K.V. *Teoriya i praktika mashinnogo obucheniya: ucheb. posobie = Theory and practice of machine learning : textbook*. Ulyanovsk: UIGTU, 2017:290. (In Russ.)
6. Vasil'ev V.I., Zhernakov S.V. Classification of GTE operation modes using neural network technology. *Vestnik Ufimskogo gosudarstvennogo aviatsionnogo tekhnicheskogo universiteta = Bulletin of the Ufa State Aviation Technical University*. 2009;12:53–56. (In Russ.)
7. Zhernakov S.V., Gil'manshin A.T. Application of intelligent algorithms based on fuzzy logic and neural networks for solving problems of failure diagnostics of aviation GTE. *Intelligent Technologies for Information Processing and Management (ITIPM'2014): proceedings of the 2nd International Conference*. 2014:112–115. (In Russ.)
8. Zhukov D.A. Analysis of classification quality criteria in the diagnosis of the functioning of a technical object. *Avtomatizatsiya protsessov upravleniya = Automation of control processes*. 2019;(3):112–117. (In Russ.)
9. Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5–32.
10. Neykov M., Liu J.S., Cai T. On the Characterization of a Class of Fisher-Consistent Loss Functions and its Application to Boosting. *Journal of Machine Learning Research*. 2016;(17):1–32.
11. Wyner A.J., Olson M., Bleich J., Mease D. Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers. *Journal of Machine Learning Research*. 2017;(18):1–33.
12. Davis J., Goadrich M. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*. Pittsburgh. 2006:233–240.
13. Wang B., Zou H. A Multicategory Kernel Distance Weighted Discrimination Method for Multiclass Classification. *Technometrics*. 2019;61:396–408. doi:10.1080/00401706.2018.1529629
14. Klyachkin V.N., Dontsova Yu.S. Comparative analysis of the accuracy of nonlinear models in predicting the state of a system based on a Markov chain. *Izvestiya Samarskogo nauchnogo tsentra Rossiyskoy akademii nauk = Proceedings of the Samara Scientific Center of the Russian Academy of Sciences*. 2013;15(4-4):924–927. (In Russ.)
15. Borovikov V.P. *STATISTICA. Искусство анализа данных на компьютере = The art of data analysis on a computer*. 2nd ed. Saint Petersburg: Piter, 2003:700. (In Russ.)

### Информация об авторах / Information about the authors

#### Владимир Николаевич Клячкин

доктор технических наук, доцент,  
профессор кафедры прикладной  
математики и информатики,  
Ульяновский государственный  
технический университет  
(Россия, г. Ульяновск, ул. Северный Венец, 32)  
E-mail: v\_kl@mail.ru

#### Vladimir N. Klyachkin

Doctor of technical sciences, associate professor,  
professor of the sub-department  
of applied mathematics and informatics,  
Ulyanovsk State Technical University  
(32 Severny Venec street, Ulyanovsk, Russia)

**Ирина Николаевна Карпунина**

кандидат технических наук,  
доцент кафедры общепрофессиональных дисциплин,  
Ульяновский институт гражданской авиации  
(Россия, г. Ульяновск, ул. Можайского, 8/8)  
E-mail: karpunina531@yandex.ru

**Irina N. Karpunina**

Candidate of technical sciences,  
associate professor of the sub-department of OPD,  
Ulyanovsk Institute of Civil Aviation  
(8/8 Mozhaiskiy street, Ulyanovsk, Russia)

**Авторы заявляют об отсутствии конфликта интересов /  
The authors declare no conflicts of interests.**

**Поступила в редакцию/Received 15.12.2021**

**Поступила после рецензирования/Revised 10.01.2022**

**Принята к публикации/Accepted 15.02.2022**