

На правах рукописи

ЖУКОВ Дмитрий Анатольевич

**РАЗРАБОТКА МОДЕЛЕЙ, АЛГОРИТМОВ И ПРОГРАММ
ДИАГНОСТИКИ ФУНКЦИОНИРОВАНИЯ ТЕХНИЧЕСКИХ ОБЪЕКТОВ
С ИСПОЛЬЗОВАНИЕМ АГРЕГИРОВАННЫХ КЛАССИФИКАТОРОВ**

Специальность 05.13.18 – Математическое моделирование,
численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Ульяновск - 2020

Работа выполнена на кафедре «Прикладная математика и информатика» в федеральном государственном бюджетном образовательном учреждении высшего образования «Ульяновский государственный технический университет».

Научный руководитель: доктор технических наук, профессор
Клячкин Владимир Николаевич

Официальные оппоненты: **Иванов Александр Куприянович**,
доктор технических наук, Федеральный научно-производственный центр АО «НПО «Марс» (г. Ульяновск), комплексное научно-исследовательское отделение - 2, главный научный сотрудник

Капитанчук Василий Вячеславович,
кандидат технических наук, доцент, ФГБОУ ВО «Ульяновский институт гражданской авиации имени Главного маршала авиации Б.П. Бугаева», кафедра организации аэропортовой деятельности и информационных технологий, доцент кафедры

Ведущая организация: Акционерное общество «Концерн
«Моринформсистема-Агат» (г. Москва)

Защита состоится «01» апреля 2020 года в 12:00 часов на заседании диссертационного совета Д 212.278.02 при ФГБОУ ВО «Ульяновский государственный университет», расположенном по адресу: Ульяновск, ул. Набережная р. Свияги, 106, корп. 1, ауд. 703.

С диссертацией и авторефератом можно ознакомиться в научной библиотеке Ульяновского государственного университета и на сайте вуза <http://www.ulsu.ru>, с авторефератом – на сайте Высшей аттестационной комиссии при Минобрнауки России <https://vak.minobrnauki.gov.ru>.

Отзывы на автореферат в двух экземплярах, заверенные печатью организации, просим направлять по адресу: 432970, г. Ульяновск, ул. Л. Толстого, д. 42, УлГУ, Отдел подготовки кадров высшей квалификации.

Автореферат разослан «...» _____ 2020 г.

Ученый секретарь
диссертационного совета

Волков М.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы

Для обеспечения безопасности и надежности технического объекта проводится его диагностика по результатам мониторинга показателей функционирования этого объекта. При этом часто диагностика сводится к распознаванию одного из двух состояний: к разделению состояний на исправные или неисправные¹. Применению математических методов в задачах технической диагностики посвящены работы И.А. Биргера, В.И. Васильева, С.В. Жернакова, П.П. Пархоменко, Н. Czichos, D.C. Montgomery и других специалистов. Актуальность проблемы обусловлена постоянно растущими требованиями к обеспечению безопасности и надежности техники, а современные компьютерные технологии, в частности, используемые в диссертационном исследовании методы машинного обучения, позволяют более точно диагностировать исправное или неисправное состояние рассматриваемого объекта.

Например, исправность функционирования системы водоочистки оценивалась по показателям качества питьевой воды в зависимости от физико-химических показателей водоисточника: температуры, цветности, мутности, щёлочности и других показателей. При прогнозировании неисправности системы (когда показатели качества питьевой воды могут не соответствовать требованиям) осуществляется изменение доз реагентов или долив чистой воды².

Работоспособность гидроагрегата зависит от уровня вибраций. Диагностика технического состояния производится по результатам непрерывного вибромониторинга. Процесс определяется десятью показателями: вибрациями генераторных подшипников, боем вала гидротурбины и другими показателями. Данные в режиме реального времени поступают на стойку управления гидроагрегатом. По этим данным необходимо оценить работоспособность агрегата. При слишком больших вибрациях снижается нагрузка, а если вибрации достигают критических значений, производится останов гидроагрегата³.

При решении вопроса об исправности технического объекта требуется оценить его состояние по заданным показателям функционирования. При этом могут использоваться методы многомерной классификации, как стандартные статистические, так и методы машинного обучения, к которым относятся нейронные сети, ансамбли моделей и другие⁴. Существенной особенностью рассматриваемой задачи являются сравнительно небольшой объем выборки (как правило, сотни наблюдений, в отличие от «BigData» в обычных задачах машинного обучения), а также несбалансированность классов в обучающей выборке: информации о неисправных состояниях объекта гораздо меньше, чем об исправных.

¹ Биргер И.А. Техническая диагностика. М.: Машиностроение, 1978. 240 с. (2-е изд.: М.: URSS, 2019)

² Бубырь Д.С., Клячкин В.Н. Управление дозами реагентов при очистке питьевой воды // Экологические системы и приборы, 2018. №3. С. 3-10.

³ Клячкин В.Н., Алексеева А.В. Оценка стабильности функционирования гидроагрегата по результатам мониторинга вибраций. // Приборы и системы. Управление, контроль, диагностика, 2019. № . С. 11-18.

⁴ Теория и практика машинного обучения / В.В. Воронина, А.В. Михеев, Н.Г. Ярушкина, К.В. Святков. Ульяновск: УлГТУ, 2017. 290 с.

Диагностика исправности технического объекта может рассматриваться как задача бинарной классификации. Точность решения этой задачи зависит от объема и качества выборки с данными по результатам мониторинга в процессе эксплуатации, метода классификации, критериев качества диагностики, способа разделения выборочных данных на обучающую и контрольную части, значимости контролируемых показателей и других факторов. Выбор этих факторов, обеспечивающих необходимую точность диагностики, является актуальной задачей. При этом необходима разработка программного комплекса, который в автоматическом режиме, анализируя исходные данные о результатах предшествующей эксплуатации, давал бы заключение об исправности объекта и прогнозировал его состояние по заданным показателям функционирования.

Объектом исследования в диссертационной работе являются сложные технические объекты, в частности, рассматривается приложение предложенных методов и моделей к анализу исправности системы водоочистки, гидроагрегата, счетчиков системы горячего водоснабжения.

Предметом исследования являются математические модели, алгоритмы и программы для диагностики состояния технического объекта.

Цель работы

– повышение точности диагностики состояния технического объекта за счет агрегирования базовых методов классификации на основе машинного обучения и выбора факторов, оказывающих влияние на качество диагностики, путем использования специально разработанных программных средств.

Для достижения поставленной цели решаются задачи:

- анализ эффективности применения различных подходов машинного обучения для оперативной диагностики нарушений функционирования технического объекта при заданном наборе контролируемых показателей;
- разработка математических моделей и алгоритмов для диагностики состояния технического объекта с использованием агрегированных классификаторов;
- оценка влияния различных факторов на качество диагностики;
- разработка программы для проведения испытаний по диагностике функционирования технического объекта;
- разработка программного комплекса для автоматизированной оценки исправности технического объекта с применением методов машинного обучения;
- оценка эффективности разработанных моделей и программных средств и численное исследование на реальных технических объектах.

Методы исследования

При решении задач исследования применялись методы теории вероятности, математической статистики, методы машинного обучения и численные методы. При разработке программного комплекса использовались методы объектно-ориентированного программирования.

Научной новизной обладают:

- впервые предложенные математические модели и алгоритмы диагностики состояния технического объекта на основе применения агрегированных классификаторов в машинном обучении;
- разработанный алгоритм бинарной классификации, обеспечивающий выбор наилучшего способа разбиения выборки на обучающую и контрольную части при кросс-валидации, способа отбора значимых показателей функционирования, а также формирование структуры агрегированного классификатора;
- предложенные численные методы обновления моделей технической диагностики (как параметров, так и структуры математической модели) при поступлении новой информации о показателях функционирования объекта;
- полученные с использованием предложенных математических моделей и алгоритмов результаты численного исследования исправности реальных технических объектов, показывающих значения критериев качества диагностики выше, чем при существующих методиках;
- программный комплекс автоматизированной диагностики, позволяющий производить оперативный анализ поступающих данных о состоянии технического объекта для обнаружения его неисправного состояния.

Достоверность проведенного исследования обеспечивается корректным применением методов теории вероятности, математической статистики, численных методов, методов объектно-ориентированного программирования, а также подтверждается результатами проведенных испытаний.

Теоретическая значимость работы состоит в разработке новых математических моделей и алгоритмов диагностики состояния технического объекта, а также в исследовании влияния различных факторов, которые непосредственно определяют качество диагностики.

Практическая значимость работы заключается в том, что, использование разработанного программного комплекса автоматизированной диагностики на основе предложенных моделей и алгоритмов обеспечивает повышение безопасности и надежности работы технических объектов.

Основные научные положения, выносимые на защиту:

1) Математические модели, разработанные на основе методов машинного обучения с применением агрегированных классификаторов, обеспечивающие повышение качества диагностирования технического состояния объекта.

2) Алгоритм бинарной классификации, значимо повышающий точность идентификации состояния технического объекта за счет выбора наилучшего объема контрольной выборки, отбора значимых показателей функционирования, а также формирования структуры агрегированного классификатора.

3) Адаптация численных методов на основе псевдоградиентной процедуры для корректировки параметров моделей агрегированных классификаторов при поступлении новой информации о показателях функционирования, позволяющая оперативно диагностировать возможные неисправности объекта.

4) Разработанный на основе предложенных моделей и алгоритмов программный комплекс, обеспечивающий анализ состояния технического объекта и оценку его исправности по результатам мониторинга показателей функционирования.

Реализация и внедрение результатов работы.

Диссертационная работа выполнялась при поддержке грантов Российского фонда фундаментальных исследований и Правительства Ульяновской области по проектам №16-48-732002 и №18-48-730001.

Результаты исследования внедрены в ЗАО «Системы водоочистки» (г. Ульяновск) при анализе работы системы водоочистки на водоканале Санкт-Петербурга, источник водоснабжения – «Западный Кронштадт».

Результаты диссертационной работы также используются в учебном процессе Ульяновского государственного технического университета в дисциплинах «Теория надежности», «Статистический контроль и управление процессами», «Статистические методы прогнозирования», читаемых студентам, обучающихся в бакалавриате и магистратуре по направлению «Прикладная математика», а также «Статистические методы в управлении качеством» по направлению «Управление качеством».

Апробация работы. Результаты исследования докладывались на научно-технических конференциях Ульяновского государственного технического университета в 2015 – 2019 г.г., на международной конференции и молодежной школе «Информационные технологии и нанотехнологии» (Самарский национальный исследовательский университет имени академика С.П. Королева, 2017 и 2019 г.г.), Международной научной конференции «FarEastCon» (Дальневосточный федеральный университет, Владивосток, 2019), на Национальной конференции по искусственному интеллекту (Ульяновск, 2019), на XV ежегодной Международной научно-технической конференции «IT-технологии: развитие и приложения» (Владикавказ, 2018 г.), на международной научно-технической конференции «Перспективные информационные технологии» (Самарский научный центр РАН, 2017 и 2018 г.г.), на научно-практической международной конференции (школе-семинаре) молодых ученых «Прикладная математика и информатика: современные исследования в области естественных и технических наук» (Тольятти, 2017-2019 г.г.) и других.

Публикация результатов работы. По результатам диссертации опубликованы 22 научные работы (из них пять статей без соавторов), в том числе семь статей в журналах по перечню ВАК и три статьи в изданиях, индексируемых Scopus. Получены два свидетельства о государственной регистрации программ для ЭВМ.

Личный вклад автора. Постановка задач исследования осуществлялась совместно с научным руководителем. Все основные теоретические и практические исследования проведены автором диссертационной работы самостоятельно.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав, заключения и списка использованных источников из 111 наименований. Объем диссертации составляет 133 страницы, включая 34 рисунка, 17 таблиц и приложения, в которых представлены копии документов о внедрении результатов

проведенных исследований и свидетельств о государственной регистрации разработанных программ.

СОДЕРЖАНИЕ РАБОТЫ

Введение содержит обоснование актуальности темы работы, сформулированы цель и задачи исследования, научная новизна, теоретическая и практическая значимость полученных результатов, приведены сведения об использовании, реализации и апробации результатов работы.

В первой главе проведен краткий обзор исследований в области диагностики функционирования технических объектов и сформулированы задачи работы.

Основной задачей технической диагностики является распознавание состояния технической системы в условиях ограниченной информации. Предполагается, что существует некоторая зависимость между показателями функционирования объекта и его состояниями. На основе исходных данных требуется восстановить эту зависимость, то есть построить алгоритм, способный для заданного набора показателей выдать достаточно точный ответ о его состоянии. Качество классификации оценивается по контрольной выборке.

Среди статистических методов диагностики получили распространение: байесовский классификатор (основан на применении формулы Байеса со строгими – «наивными» – предположениями о независимости влияния показателей функционирования на результаты классификации), дискриминантный анализ (метод многомерной классификации при наличии обучающих выборок), логистическая регрессия – разновидность множественной регрессии, учитывающая бинарность зависимой переменной (объект исправен или неисправен), метод опорных векторов: ищется разделяющая полоса максимальной ширины между классами исправных и неисправных объектов.

Наиболее распространенным из интеллектуальных методов технической диагностики являются нейронные сети. Обучение нейронной сети представляет многопараметрическую задачу нелинейной оптимизации. Многослойная сеть может моделировать функцию практически любой сложности, причем число слоев и число элементов в каждом слое определяют сложность этой функции. Нейронные сети успешно применяются при диагностике авиационных газотурбинных двигателей⁵.

Перечисленные методы классификации не всегда приводят к удовлетворительному решению задачи оценки состояния технического объекта. В этих случаях можно попытаться построить композиции алгоритмов. При этом часто погрешности отдельных алгоритмов взаимно компенсируются. Опыт показывает, что два главных метода построения композиции – бэггинг и бустинг – дают значительно более точный результат, чем применение отдельного алгоритма.

⁵ Васильев В.И., Жернаков С.В. Контроль и диагностика технического состояния авиационных двигателей на основе интеллектуального анализа данных. Вестник УГАТУ. 2006. Т.7. №2(15). С. 71-81.

Рассмотренные методы используются в машинном обучении для классификации объектов. При использовании кросс-валидации исходная выборка разбивается на несколько частей. Одна часть используется для контроля, а оставшиеся – для обучения. Последовательно перебираются все варианты. Среднее значение критерия качества диагностики характеризует обобщающую способность алгоритма.

Проведенный обзор показал, что для решения задачи технической диагностики могут использоваться различные методы, однако ни один из них не имеет явного преимущества. Необходима апробация разных методов; при этом следует учесть множество факторов: критерий качества диагностики, способ формирования контрольной и обучающей выборок, способ отбора значимых показателей функционирования и другие. Таким образом, возникают сформулированные выше основные задачи исследования, направленные на повышение качества диагностики.

Во второй главе разработаны математические модели для диагностики функционирования технических объектов. Исходные данные о состоянии объекта представляются 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 вектор-столбец ответов Y , состоящий из единиц (для тех опытов, в которых объект исправен) и нулей при неисправном объекте. 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 предскажет ответ Y для любого заданного X . Параметры модели подбираются по исходным данным; процесс подбора называется обучением алгоритма. Найденные параметры должны обеспечить оптимальное значение функционала качества. Наиболее распространенным показателем, который может быть использован для оценки качества бинарной классификации, является доля правильных ответов на контрольной выборке.

При несбалансированных классах (когда исправных состояний объекта значительно больше, чем неисправных) доля правильных ответов или доля ошибок не могут объективно оценивать качество классификации. Гораздо более информативны точность $P = tp/(tp + fp)$ и полнота $R = tp/(tp + fn)$, где tp – количество правильно классифицированных исправных состояний, fp – количество неправильно классифицированных исправных состояний, fn – количество неправильно классифицированных неисправных состояний. На основе этих двух показателей может быть сформирован единый критерий⁶

$$F = \frac{2PR}{P + R}, \quad (1)$$

– это гармоническое среднее точности и полноты (F -мера): чем ближе значение F к единице, тем качество классификации выше.

⁶ 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.

Функционалом качества также может быть выбрана площадь AUC (area under the curve) под кривой ошибок. Эта кривая образуется, если по оси абсцисс брать значения $fp(c)$, а по оси ординат $tp(c)$, где c - порог. Критерий AUC также устойчив к несбалансированным классам.

Учитывая, что, как правило, для сложных технических объектов количество прецедентов с исправными состояниями гораздо больше, чем с неисправными, основным критерием качества будем считать F -меру; иногда, как дополнительный критерий качества, используем площадь AUC под кривой ошибок.

Для решения задач кредитного скоринга в машинном обучении был предложен агрегированный подход⁷. В композиционных классификаторах для построения ансамбля используется один и тот же метод классификации, построенный или на разных подмножествах выборки или ориентированный на компенсацию ошибки предыдущей итерации. Представляет интерес совместное использование *различных* методов классификации, построенных на обучающей выборке.

Воспользуемся полным перебором наборов из H базовых методов. Нетрудно видеть, что в общем случае число наборов равно $2^H - 1$. При используемых в дальнейшем исследовании одиннадцати базовых классификаторов общее число наборов составит $2^{11} - 1 = 2047$ вариантов. Такое большое количество моделей (2047 вариантов) может привести к неоправданно большому объему вычислений. Проведенные далее численные исследования реальных объектов показали, что увеличение числа компонентов в агрегате больше двух не приводит к значимому повышению точности. Таким образом, на практике возможен перебор лишь десяти агрегированных моделей: сочетание лучшей из базовых с одной из остальных.

Пусть $\hat{P}_K(X)$ – вероятность исправности объекта, найденная с помощью K -го базового метода, $K = 1, \dots, H$. Тогда при агрегировании *по среднему значению*:

$$\hat{P}_{AMC}(X) = \frac{1}{H} \sum_{K=1}^H \hat{P}_K(X), \quad (2)$$

где $\hat{P}_{AMC}(X)$ - вероятность того, что объект исправен. При агрегировании *по медиане* вначале следует ранжировать ряд, содержащий результаты базовых методов в наборе. Результат агрегированного метода классификации *по голосованию* представляет собой среднее значение результатов базовых методов, которые определили исправность объекта с вероятностью, не ниже некоторого p . Таким образом, значения вероятностей классификации, оказавшиеся ниже, чем p , приравниваются к нулю, а оставшиеся - к единице, и по этим значениям строятся агрегированные модели классификации.

При этом, поскольку, как уже отмечалось, разбивка исходных данных на обучающую и контрольную выборки производится случайным образом, структуры агрегированных классификаторов оказываются различными, и возникает вопрос, какую структуру предпочесть для принятия решения об исправности объекта.

⁷ Клячкин В.Н., Шунина Ю.С. Система оценки кредитоспособности заемщиков и прогнозирования возврата кредитов // Вестник компьютерных и информационных технологий, 2015. №11. С. 45-51.

Построим математическую модель агрегированного классификатора по среднему значению для двух базовых методов. Пусть, например, при проведении диагностики конкретного объекта с применением кросс-валидации наилучшими (обеспечивающими максимум F -критерия (1)) оказались байесовский классификатор и логистическая регрессия. Тогда математическая модель агрегированного классификатора по среднему значению для этих двух базовых методов примет вид:

$$\hat{P}_{AMC}(X) = \frac{\frac{1}{2} P(Y=1) \prod_{j=1}^p P(x_j | Y=1)}{P(Y=1) \prod_{j=1}^p P(x_j | Y=1) + P(Y=0) \prod_{j=1}^p P(x_j | Y=0)} + \frac{1/2}{1 + \exp(-(q_0 + \sum_{j=1}^p q_j x_j))}, \quad (3)$$

где x_j - показатели функционирования объекта, $P(Y=1)$ и $P(Y=0)$ - априорные вероятности исправности и неисправности объекта соответственно, $P(x_j | Y=1)$, $P(x_j | Y=0)$ - вероятности показателей x_1, \dots, x_p в классах $Y=1$ и $Y=0$, $q_0, q_1 \dots q_p$ – параметры логистической регрессии, определяемые методом максимального правдоподобия.

При использовании трех базовых методов в качестве агрегированного значения по медиане принимается центральный член вариационного ряда. Например, при использовании в качестве базовых методов логистической регрессии, дискриминантного анализа и бустинга:

$$\hat{P}_{AMM}(X) = median \left(\frac{1}{1 + \exp(-(q_0 + \sum_{j=1}^p q_j x_j))}; \frac{P(Y=1)G_1(X)}{P(Y=1)G_1(X) + P(Y=0)G_2(X)}; \frac{1}{1 + \exp(-\sum_{t=1}^r \alpha_t h_t)} \right), \quad (4)$$

где $G_1(X)$ и $G_2(X)$ – оценки плотности нормального распределения для объекта в исправном и неисправном состояниях, $h_t(x)$ – базовые классификаторы бустинга; α_t – коэффициент взвешенного голосования для соответствующего классификатора $h_t(x)$, определяемый в зависимости от конкретного метода бустинга (AdaBoost, LogitBoost, GentleBoost, и др.).

Подобным образом строятся и модели агрегирования для других базовых методов.

По истечении некоторого времени показатели функционирования объекта изменяются, необходимо обновить параметры классификатора для адаптации к изменяющимся условиям. Адаптация может быть осуществлена путем корректировки параметров моделей, входящих в состав агрегированных классификаторов, с использованием рекуррентной псевдоградиентной процедуры. Эта численная процедура позволяет находить оценки параметров с достаточно быстрой сходимостью (порядок сходимости $O(\frac{1}{\sqrt{n}})$) при минимальных вычислительных затратах по срав-

нению с другими методами^{8,9}. Уменьшение вычислительных затрат обусловлено тем, что в предложенной процедуре не считается градиент для каждого предыдущего состояния, а только для последнего. Оптимизация модели состоит в оценке вектора её параметров, минимизирующих сумму квадратов отклонений «истинной» вероятности $P(Y_r)$ принадлежности r -го наблюдения классу от оценки вероятности принадлежности $\hat{P}(X_r, \bar{\alpha})$ r -го наблюдения классу, зависящая от показателей функционирования объекта и вектора параметров модели $\bar{\alpha}$.

$$\Omega(\bar{\alpha}) = \sum_{r=1}^l (P(Y_r) - \hat{P}(X_r, \bar{\alpha}))^2. \quad (5)$$

Псевдоградиентная адаптация заключается в корректировке вектора параметров модели при поступлении информации о каждом новом наблюдении на основе процедуры:

$$\bar{\alpha}_{r+1} = \bar{\alpha}_r - \nu_r \nabla J(\bar{\alpha}_r), \quad (6)$$

где $\bar{\alpha}_{r+1}$ – следующее за $\bar{\alpha}_r$ приближение оптимального вектора $\bar{\alpha}$; ν_r – коэффициенты, влияющие на величину шага,

$$\nabla J(\bar{\alpha}_r) = \nabla (P(Y_r) - \hat{P}(X_r, \bar{\alpha}_r))^2 \quad (7)$$

– градиент отдельного слагаемого из формулы (5), который является псевдоградиентом функции $\Omega(\bar{\alpha})$. Корректировка параметров модели достигается путем добавления к старому вектору параметров $\bar{\alpha}_r$ поправки, получаемой в результате умножения числа ν_r на псевдоградиент $\nabla J(\bar{\alpha}_r)$.

Например, в случае одного из базовых классификаторов – модели бустинга – псевдоградиентная процедура обновления модели имеет вид:

$$\bar{\alpha}_{r+1} = \bar{\alpha}_r + \eta_r (P(Y_r) - \hat{P}(X_r, \bar{\alpha}_r)) \frac{\partial \hat{P}(X_r, \bar{\alpha}_r)}{\partial \bar{\alpha}_r}, \quad (8)$$

где $\eta_r = 2\nu_r$ – новый параметр шага процедуры.

По истечении некоторого времени структура модели классификации может «устареть», и применение процедуры, которая оптимизирует только параметры модели, может оказаться недостаточным. В этом случае производится обновление структуры модели, например, по такой схеме. При построении исходной структуры выбираются две или более конкурирующих моделей: не только та, для которой F -мера оказалась максимальна, но и предыдущие по качеству модели (они будут запасными). Вначале применяется лучшая модель, а качество запасных оценивается виртуально. С течением времени более хорошие результаты может показывать одна из запасных моделей, тогда она становится основной, то есть применяющейся на практике, а применяемая до этого модель переводится в запасные.

⁸ Поляк Б.Т., Цыпкин Я.З. Оптимальные псевдоградиентные алгоритмы адаптации // Автоматика и телемеханика. 1980. №8. С. 74-84.

⁹ Крашенинников В.Р., Кувайскова Ю.Е., Клячкин В.Н., Шулнина Ю.С. Обновление моделей прогнозирования состояния объектов в виде систем временных рядов и многомерных классификаторов // Вестник компьютерных и информационных технологий. 2017. №6. С.11-16

Практическая реализация предложенных моделей предполагает разработку соответствующих алгоритмов и программ. Для оценки их эффективности необходимо проведение экспериментальных исследований на реальных объектах.

В третьей главе описаны результаты численного исследования эффективности машинного обучения для диагностики функционирования технических объектов. Исследование проводилось на примере трех реальных объектов: системы водоочистки Санкт-Петербургского водоканала, системы управления гидроагрегатом на Краснополянской ГЭС и счетчиков горячего водоснабжения системы водоснабжения города Ульяновска.

Исправность системы водоочистки Y оценивалась по показателям качества питьевой воды в зависимости от физико-химических показателей водоисточника (Западный Кронштадт): X_1 – температуры, X_2 – цветности, X_3 – мутности, X_4 – значения рН, X_5 – щёлочности, X_6 – окисляемости, и доз добавляемых реагентов: X_7 – коагулянта и X_8 – флокулянта. Получены результаты 348 наблюдений за показателями функционирования (8 показателей), в 47 случаях состояние системы признано неисправным. Задача: используя полученные данные – матрицу X размерностью 348 строк и 8 столбцов и вектор-столбец ответов об исправности системы Y , разработать модель бинарного классификатора, которая по вновь полученным показателям функционирования водоисточника обеспечила бы оперативную диагностику и дала бы прогноз об исправности (или неисправности) системы.

Исследование системы управления гидроагрегатом проводилось по данным вибромониторинга с применением вибродатчиков, расположенных на различных участках гидроагрегата. Процесс определялся десятью показателями: вибрациями нижнего и верхнего генераторного подшипника верхнего бьефа и на правом берегу, боем вала гидротурбины на нижнем бьефе и правом берегу, боем вала гидрогенератора, а также вибрациями крышки гидротурбины. Исходная выборка составила 1500 наблюдений, из которых 966 – в исправном состоянии.

В качестве третьего объекта исследования использовалась система горячего водоснабжения в городе Ульяновске: контроль проводился по данным, снятым со счетчиков водоснабжения. Контролировалось функционирование системы водоснабжения горячей воды на наличие утечек ($Y = 0$) по параметрам: X_1 – температура воды в подающем трубопроводе, X_2 – расход рабочей жидкости в трубопроводе подачи, X_3 – расход рабочей жидкости в трубопроводе «обратки», X_4 – подаваемое давление, X_5 – обратное давление, X_6 – количество тепловой энергии. Объем исходной выборки составил 527 наблюдений (423 исправных состояния).

Все три рассмотренные проблемы диагностики состояния сложных технических систем представляют задачи бинарной классификации при наличии обучающей выборки, которые могут быть решены методами машинного обучения. Для проведения испытаний необходима разработка специальной программы. Такая программа была разработана в среде Matlab на базе библиотеки инструментов Statistics and Machine Learning Toolbox, в которую входят рассмотренные базовые методы классификации.

С учетом целей исследования программа обеспечивает:

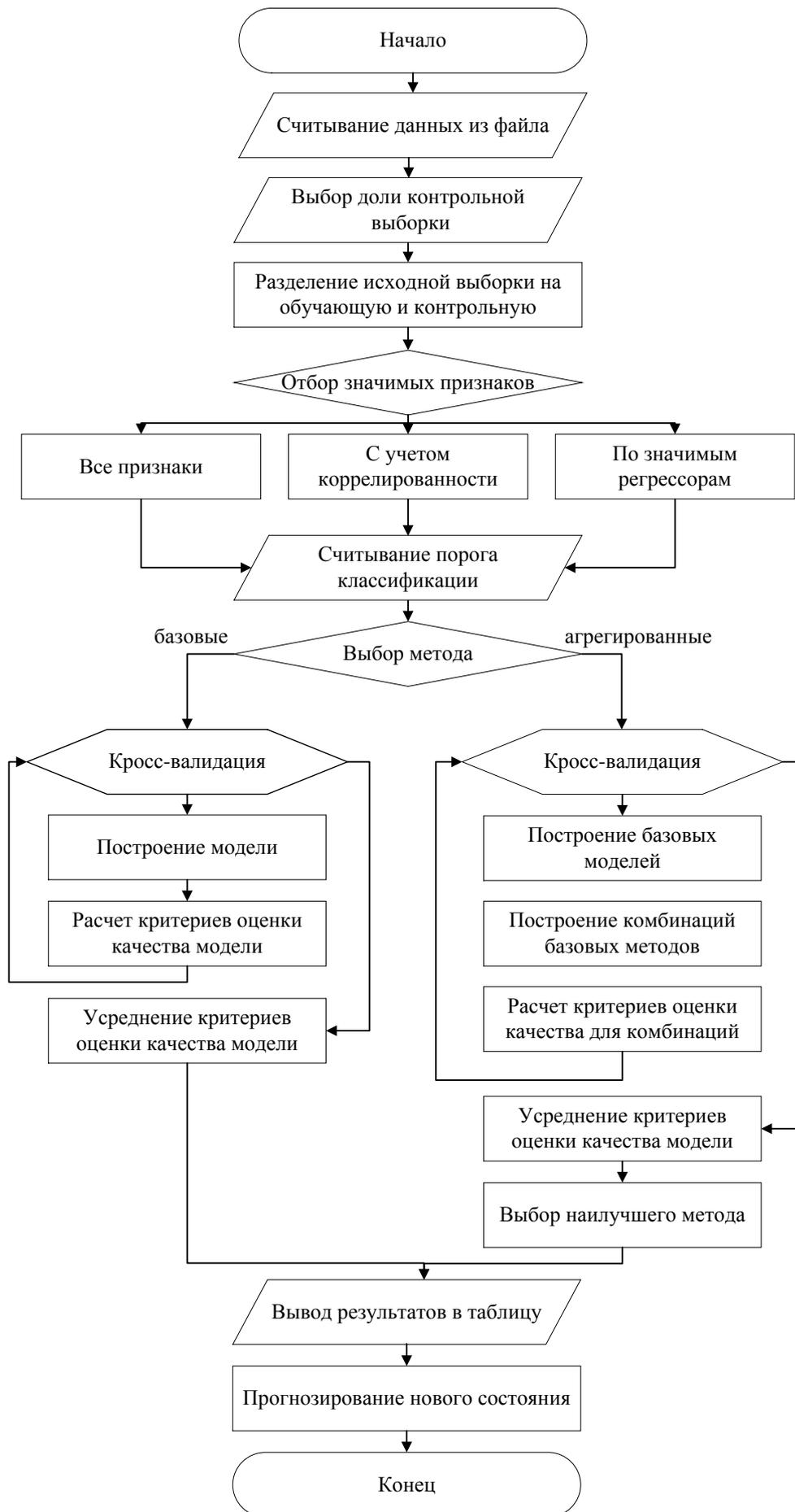


Рисунок 1 – Блок-схема алгоритма программы для проведения испытаний

- использование различных базовых методов, а также построение агрегированных классификаторов,
- применение различных критериев качества классификации: доли ошибок на контрольной выборке, F -критерия, площади AUC под кривой ошибок и других,
- изменение объема контрольной выборки,
- различные методы отбора значимых показателей,
- проведение кросс-валидации,
- прогнозирование новых состояний технического объекта.

На рис. 1 показана блок-схема алгоритма, а на рис. 2 – пример расчета по базовым методам машинного обучения. Видно, что расхождение по F -критерию между различными методами достигает 5,2%.

	Метод	F-мера	AUC
1	ЛР	0.8833	0.8501
2	ДА	0.8602	0.7000
3	БК	0.8623	0.7059
4	НС	0.8805	0.3708
5	МОВ	0.8592	0.7783
6	БДР	0.9041	0.8952
7	АВ	0.8810	0.8580
8	LB	0.8800	0.9302
9	GB	0.8684	0.7726
10	RB	0.8644	0.8375
11	GrB	0.8854	0.8748

Рисунок 2 – Представление результатов исследований; GrB, АВ, LB, GB, RB - методы бустинга: градиентный, AdaBoost, LogitBoost, GentleBoost, RUSBoost

Предварительное исследование проводилось на данных системы водоочистки. Все испытания повторялись многократно, так как разбиение на контрольную и обучающую части выборки происходит случайным образом, и критерии качества диагностики также являются случайными величинами.

Исследовалось распределение F -критерия и площади под кривой ошибок AUC в системе Statistica. Распределение обоих показателей не противоречит гипотезе о нормальности. Это обстоятельство позволяет при необходимости строить доверительный интервал для математического ожидания F -критерия или площади AUC под кривой ошибок, а также использовать стандартную методику проверки статистических гипотез.

При изучении влияния доли контрольной выборки на качество алгоритмов машинного обучения при анализе исправности технического объекта использовалась процедура кросс-валидации, при этом объем контрольной выборки варьировался от 5 до 25%. Каждое испытание повторялось пятикратно.

Исследование показало, что объем контрольной выборки оказывает влияние на результаты расчета методами машинного обучения. В связи с этим можно рекомендовать на стадии выбора алгоритма для проведения диагностики конкретного

объекта провести соответствующие испытания. Для рассмотренного объекта – системы водоочистки – наилучшим вариантом оказался бэггинг деревьев решений при объеме контрольной выборки 20% от исходной (рис. 3). Аналогичное исследование проводилось по результатам вибромониторинга гидроагрегата и для счетчиков воды. Результаты оказались идентичными: различные методы обучения по-разному реагируют на изменение объема контрольной выборки.

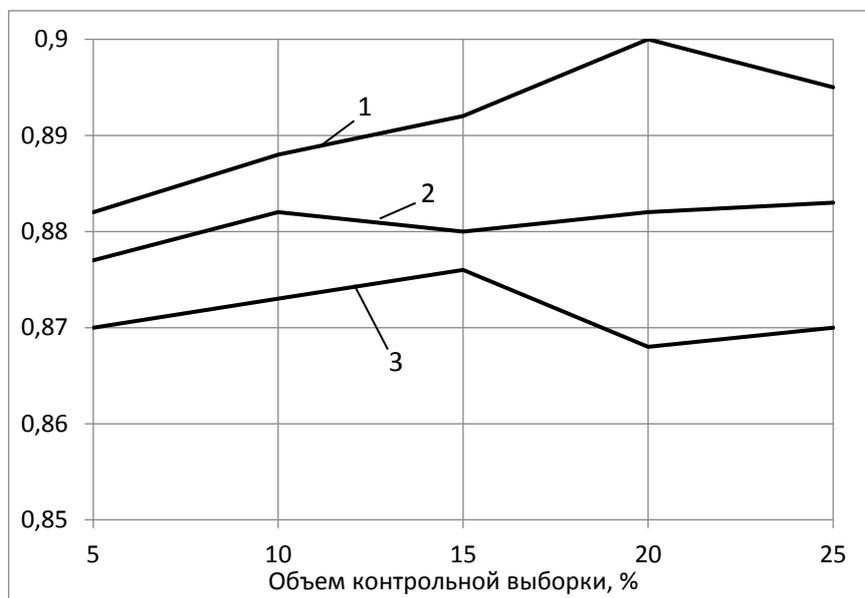


Рисунок 3 – Зависимость F -критерия от объема контрольной выборки (1 – бэггинг деревьев решений, 2 – градиентный бустинг, 3 – AdaBoost)

Также проводилось изучение влияния способа отбора значимых факторов на эффективность машинного обучения. В одном испытании отбор проводился по корреляционной матрице: последовательно исключались из обучения показатели, сильно коррелированные между собой, и показатели, практически некоррелированные с откликом – исправностью объекта. В другом опыте отбор значимых показателей проводился по критерию Стьюдента для соответствующей регрессионной модели. Исследование показало, что удаление незначимых факторов оказывает неоднозначное влияние на эффективность машинного обучения, и для каждого метода подбор значимых факторов индивидуален. В системе диагностики состояния объекта необходимо предусмотреть подсистему поиска метода обучения и отбора значимых показателей, обеспечивающую минимальную погрешность.

На рис. 4 приведены результаты испытаний для системы водоочистки: а) с исключением коррелированных факторов, б) с исключением факторов, незначимых по регрессионной модели. Расчет с учетом всех показателей показан на рис.2.

Метод	F-мера	AUC
ЛР	0.8772	0.8049
ДА	0.8599	0.6307
БК	0.8674	0.7042
НС	0.8679	0.9053
МОВ	0.8636	0.8873
БДР	0.9020	0.9479
АВ	0.8862	0.9078
GrB	0.8883	0.9143
RB	0.8654	0.8304
GB	0.8814	0.9290
LB	0.8767	0.9467

а)

Метод	F-мера	AUC
ЛР	0.8808	0.8955
ДА	0.8596	0.7959
БК	0.8676	0.6875
НС	0.8856	0.8402
МОВ	0.8607	0.8833
БДР	0.9026	0.8229
АВ	0.8852	0.8320
RB	0.8616	0.8054
GB	0.8793	0.8225
LB	0.8908	0.9312
GrB	0.8854	0.9381

б)

Рисунок 4 – Результаты расчета с вариантами отбора показателей в системе водоочистки

	Метод	F-мера	AUC
1	AM-C: НБК+ МОВ+ БДР	0.9316	0.6712
2	AM-M: НС+ ДА+ НБК+ МОВ+ GrB+ БДР	0.9321	0.6747
3	AM-G: НБК+ ЛР+ БДР+ GB	0.9314	0.6732

Рисунок 5 – Построение агрегированных классификаторов (AM-C – по среднему, AM-M – по медиане, AM-G – по голосованию)

При оценке эффективности агрегированного подхода проводился расчет F -критерия при заданном объеме выборки. Для системы водоочистки наилучшим из базовых методов оказался бэггинг деревьев решений $F = 0,9041$; агрегирование (рисунок 5) повысило этот показатель до 0,9321 (лучшим оказался агрегированный классификатор по медиане AM-M, включающий 6 методов: НС – нейронную сеть, ДА – дискриминантный анализ, НБК – байесовский классификатор, МОВ – метод опорных векторов, GrB – градиентный бустинг и БДР – бэггинг деревьев решений).

При исследовании вибромониторинга гидроагрегата лучшим оказался метод бустинга AdaBoost с значением $F = 0,998$, при этом агрегированные методы превысили эту величину незначительно. При оценке исправности счетчиков горячего водоснабжения у агрегированных методов значение F -критерия больше, чем у остальных методов, например, F -критерий агрегированных методов относительно метода опорных векторов выше на 12%, а относительно градиентного бустинга более чем на 3%.

Испытания повторялись по 30 раз. Проверялась нулевая гипотеза о равенстве средних значений F -критерия при агрегировании и при применении базовых методов (сравнивались агрегированные данные с бэггингом деревьев решений, как показавшим лучший результат). Как альтернативная, рассматривалась гипотеза о превышении среднего значения при агрегировании. Вначале сравнивались дисперсии двух выборок по критерию Фишера, затем по критерию Стьюдента о равенстве средних был сделан вывод о том, что нулевая гипотеза отвергается: среднее значение F -критерия при агрегировании выше, чем при применении базовых классификаторов; агрегирование значительно повышает качество классификации.

Далее проверялась гипотеза о том, что увеличение в структуре агрегата числа базовых классификаторов больше двух несущественно влияет на значения F -критерия. С этой целью вся выборка из 30 испытаний разбивалась на два подмножества. В первое включались данные по агрегатам, состоящим только из двух компонент, во второе – все остальные. Проверка гипотезы о равенстве средних в этих подмножествах показывает справедливость этой гипотезы: среднее значение F -критерия не изменяется при увеличении количества базовых классификаторов в структуре агрегата. Это обстоятельство позволило резко сократить объем вычислений при разработке комплекса программ для технической диагностики.

В четвертой главе разработаны алгоритм и комплекс программ для диагностики функционирования технического объекта, позволяющие в автоматическом режиме построить наилучшую модель для прогнозирования состояния объекта. При этом для расчетов применяется кросс-валидация на всех этапах: при выборе наилучшего объема контрольной выборки, при отборе значимых показателей, при поиске методов обучения, обеспечивающих максимальное значение F -критерия на контрольной выборке.

Разработанный комплекс программ включает блоки подготовки исходных данных, построения моделей базовых методов, выбора объема контрольной выборки, отбора значимых показателей функционирования, блок агрегирования методов машинного обучения (для уточнения результатов принято решение об использовании трех (а не двух) базовых классификаторов в составе агрегата), блок вывода полученных результатов в таблицу и блок прогнозирования состояния технического объекта.

На рисунке 6,а показаны результаты автоматизированного расчета исправности системы водоочистки. Лучшими оказались методы агрегирования, включающие три базовых классификатора: логистическую регрессию ЛР, бэггинг деревьев решений БДР и RUSBoost RB ($F = 0,9399$). При этом отбор значимых показателей не улучшил результат, наилучший объем контрольной выборки оказался 10% (кросс-валидация с разбивкой выборки на 10 частей).

Аналогичные результаты получены по диагностике 1500 наблюдений по результатам вибромониторинга системы управления гидроагрегатом (рис. 6,б): максимальное значение F -критерия составило 0,9862 при объеме контрольной выборки 5% и использовании в расчете всех десяти показателей вибродатчиков, при этом агрегированный классификатор включает бэггинг деревьев решений и два метода бустинга: AdaBoost и LogitBoost.

Для счетчиков горячего водоснабжения по результатам 527 прецедентов наилучшим оказался результат $F = 0,9164$ (рис. 6,в) при объеме контрольной выборки 25% и отборе показателей функционирования по значимости регрессоров.

Загрузить данные		Объем								348
	1	2	3	4	5	6	7	8		
1	1	1.2000	36	13	7.5200	0.5500		7	2.1520	
2	1	1.1000	36	12	7.6200	0.5700	8.6000		7.1520	
3	1	0.8000	41	9.2000	7.5800	0.5700	8.6000		7.8325	
4	1	0.4000	40	3.1000	7.4100	0.5800	8.6000		7.5389	
5	1	0.4000	37	5	7.6000	0.6200	8.2000		7.2507	
6	1	0.3000	39	4.2000	7.5400	0.5500	7.8000		7.4440	
7	1	0.7000	40	2.8000	7.4900	0.6300	8.6000		7.5389	
8	1	0.6000	36	1.5000	7.5500	0.5500	8.4000		7.1520	
9	1	0.3000	35	1.5000	7.4300	0.5500	7.9000		7.0520	
10	1	0.2000	38	1.7000	7.5000	0.6000	7.9000		7.3480	
11	1	0.2000	37	3.9000	7.3400	0.6000	8.6000		7.2507	
12	1	0.2000	37	1.6000	7.5100	0.5500	7.3000		7.2507	
13	1	0.2000	37	1.8000	7.3700	0.6000	7.9000		7.2507	

Лучший процент контрольной выборки		10	Выполнить расчет
Отбор признаков:		не нужен	
Лучшие методы:			
Метод	F-критерий		
1 AM-C ПР + БДР + RB	0.9399		
2 AM-M ЛР + БДР + RB	0.9399		
3 AM-G БДР + RB	0.9399		

а)

Лучший процент контрольной выборки		5	Выполнить расчет
Отбор признаков:		не нужен	
Лучшие методы:			
Метод	F-критерий		
1 AM-C БДР + AB + LB	0.9862		
2 AM-M БДР + AB + LB	0.9862		
3 AM-G LB + GB	0.9857		

б)

Лучший процент контрольной выборки		25	Выполнить расчет
Отбор признаков:		по значимости регрессоров	
Лучшие методы:			
Метод	F-критерий		
1 AM-C ДА + БДР + GB	0.9164		
2 AM-M ДА + БДР + GB	0.9164		
3 AM-G БДР + LB	0.9061		

в)

Рисунок 6 – Результаты автоматизированного расчета: а) для системы водоочистки, б) для гидроагрегата, в) для счетчиков

Заключение

Поставленная цель работы – повышение точности диагностики состояния технического объекта за счет агрегирования базовых методов классификации на основе машинного обучения и выбора факторов, оказывающих влияние на качество диагностики, путем использования специально разработанных программных средств – достигнута.

Получены следующие основные результаты:

1) Предложены математические модели агрегированных классификаторов для оценки исправности технических объектов на основе методов машинного обучения, которые повышают точность диагностирования состояния объекта.

2) Опираясь на статистические исследования, доказано влияние объема контрольной выборки и способа отбора значимых показателей на качество диагностики технического объекта.

3) Путем проведения статистических испытаний показана эффективность разработанных моделей и алгоритмов, при этом значение F -критерия на исследуемых выборках за счет применения агрегирования, выбора объема контроля и отбора значимых показателей увеличилось до 15% относительно базовых методов.

4) Предложенные численные методы корректировки параметров на основе псевдоградиентной процедуры, адаптированной к агрегированным классификаторами, а также методы обновления структуры моделей при поступлении новой информации о показателях функционирования объекта позволяют оперативно обновлять результаты диагностики и выявлять неисправное состояние объекта.

5) Программный комплекс, разработанный на основе предложенных моделей и алгоритмов для диагностики состояния технического объекта, обеспечивает поддержку принятия решений в условиях эксплуатации.

б) Проведенное численное исследование на реальных данных системы водочистки показало значимое повышение F -критерия при диагностике исправности системы при применении разработанного программного комплекса. Аналогичные результаты, свидетельствующие о повышении качества диагностики, получены при анализе системы вибромониторинга гидроагрегата и счетчиков горячей воды в системе водоснабжения.

Основные положения диссертационной работы изложены в следующих работах:

Статьи, опубликованные в журналах, рекомендованных ВАК РФ:

1. Жуков Д.А. Анализ критериев качества классификации при диагностике функционирования технического объекта // Автоматизация процессов управления. – 2019. – № 3 (57). – С. 112-117.
2. Клячкин В.Н., Жуков Д.А. Алгоритм диагностики функционирования технического объекта с использованием агрегированных классификаторов // Автоматизация процессов управления. – 2019. – № 2 (56). – С. 37-43.
3. Клячкин В.Н., Жуков Д.А. Прогнозирование состояния технического объекта с применением методов машинного обучения // Программные продукты и системы. – 2019. – №2. – С.244-250.
4. Клячкин В. Н., Кувайскова Ю.Е., Жуков Д.А. Влияние способа отбора значимых показателей на качество диагностики состояния технического объекта // Автоматизация. Современные технологии. – 2019. – Т. 73. № 1. – С.32-36.
5. Жуков Д.А., Клячкин В.Н. Влияние объема контрольной выборки на качество диагностики состояния технического объекта // Автоматизация процессов управления. – 2018. – № 2 (52). – С. 90-95.
6. Клячкин В. Н., Кувайскова Ю.Е., Жуков Д.А. Выбор метода бинарной классификации при технической диагностике с применением машинного обучения // Известия Самарского научного центра Российской академии наук. – 2018. – Т. 20. – № 4-3. – С. 494-497.
7. Клячкин В. Н., Кувайскова Ю.Е., Жуков Д.А. Диагностика технического состояния аппаратуры с использованием агрегированных классификаторов // Радиотехника. – 2018. – № 9. – С.46-49.

Публикации в изданиях, индексируемых в Scopus:

8. Zhukov, D.A., Klyachkin, V.N., Krashennnikov, V.R., Kuvayskova, Yu.E. Selection of aggregated classifiers for the prediction of the state of technical objects / CEUR Workshop Proceedings. – 2019. – V. 2416. – P. 361-367
9. Klyachkin V.N., Zhukov D.A., Zentsova E.A. Analysis of stable functioning of objects using machine learning / CEUR Workshop Proceedings. – 2019. – V. 2416. – P. 19-25.
10. Klyachkin V.N., Kuvayskova Yu.E., Zhukov D.A. The use of aggregate classifiers in technical diagnostics, based on machine learning // CEUR Workshop Proceedings. – 2017. – V. 1903. – P. 32-35.

Публикации в других изданиях:

11. Жуков Д.А. Особенности диагностики функционирования технического объекта методами машинного обучения // Современные проблемы проектирования, производства и эксплуатации радиотехнических систем. Сборник научных трудов. – Ульяновск, 2019. – С. 214-216.
12. Жуков Д.А. Оценка качества диагностики функционирования технического объекта методами машинного обучения по различным критериям // Вестник Ульяновского государственного технического университета. – 2018. – № 4 (84). – С. 40-43.

13. Жуков Д.А., Клячкин В.Н. Диагностика исправности технического объекта с использованием пакета МАТЛАВ // Перспективные информационные технологии: труды Международной научно-технической конференции. – Самарский научный центр РАН, 2018. – С. 55-57.
14. Жуков Д.А., Клячкин В.Н. Отбор значимых показателей при диагностике технического объекта с применением машинного обучения // IT-технологии: развитие и приложения. XV Ежегодная Международная научно-техническая конференция. – Владикавказ, 2018. – С. 261-266.
15. Жуков Д.А., Клячкин В.Н. Применение метода главных компонент при диагностике состояния технического объекта // Прикладная математика и информатика: современные исследования в области естественных и технических наук. Материалы научно-практической международной конференции (школы-семинара) молодых ученых. – Тольятти, 2018. – С. 109-112.
16. Жуков Д.А. Использование методов машинного обучения при исследовании состояния технического объекта // Современные проблемы проектирования, производства и эксплуатации радиотехнических систем. Сборник научных трудов. – Ульяновск, 2017. – С. 179-182.
17. Жуков Д.А., Клячкин В.Н. Анализ эффективности алгоритмов бустинга при диагностике функционирования технических объектов // Прикладная математика и информатика: современные исследования в области естественных и технических наук Материалы научно-практической всероссийской конференции (школы-семинара) молодых ученых. – Тольятти, 2017. – С. 185-188.
18. Жуков Д.А., Хорева А.С., Кувайскова Ю.Е., Клячкин В.Н. Формирование контрольных выборок при технической диагностике объекта с применением машинного обучения // Математические методы и модели: теория, приложения и роль в образовании. Международная научно-техническая конференция. – Ульяновск, 2016. – С. 44-48.
19. Жуков Д.А., Клячкин В.Н. Задачи обеспечения эффективности машинного обучения при диагностике технических объектов // Современные проблемы проектирования, производства и эксплуатации радиотехнических систем. Сборник научных трудов. – Ульяновск, 2016. – № 10. – С. 172-174.
20. Жуков Д.А. Повышение эффективности машинного обучения при решении задач технической диагностики // В сборнике: IN MEMORIAM: Султан Галимзянович Валеев / сборник памяти С. Г. Валеева. – Ульяновск, 2016. – С. 139-143.

Свидетельства о регистрации программ:

21. Клячкин В.Н., Жуков Д.А. Диагностика состояния технического объекта с использованием агрегированных классификаторов // Свидетельство о регистрации программы для ЭВМ. №2019611560.
22. Клячкин В.Н., Кувайскова Ю.Е., Жуков Д.А. Оценка исправности технического объекта с применением машинного обучения // Свидетельство о регистрации программы для ЭВМ. №2019611562.

Жуков Дмитрий Анатольевич
Разработка моделей, алгоритмов и программ диагностики функционирования технических объектов
с использованием агрегированных классификаторов
Автореферат