МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

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

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель
д-р техн. наук, профессор В.Н. Клячкин

Ульяновск – 2020

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность работы

Для обеспечения безопасности и надежности технического объекта проводится его диагностика в условиях эксплуатации по результатам мониторинга показателей функционирования этого объекта. При этом часто диагностика сводится к распознаванию одного из двух состояний объекта: к разделению состояний на исправные или неисправные. Применению математических методов в задачах технической диагностики посвящены работы И.А. Биргера, В.И. Васильева, С.В. Жернакова, П.П. Пархоменко, H. 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 воды.

Работоспособность гидроагрегата зависит от уровня вибраций. Диагностика технического состояния производится по результатам непрерывного вибромониторинга. Процесс определяется десятью показателями: вибрациями нижнего и верхнего генераторного подшипника, 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 гидроагрегата.

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 другие. Существенной особенностью рассматриваемой задачи являются сравнительно небольшой объем выборки (как правило, сотни наблюдений, в отличие от «Big Data» в десятки и сотни тысяч в обычных задачах машинного обучения), а также несбалансированность классов в обучающей выборке: информации о показателях функционирования при неисправных состояниях объекта гораздо меньше, чем при исправных.

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 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 конференция «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 между-народной 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.

ГЛАВА 1. ОБЗОР ИССЛЕДОВАНИЙ В ОБЛАСТИ ДИАГНОСТИКИ ФУНКЦИОНИРОВАНИЯ ТЕХНИЧЕСКИХ ОБЪЕКТОВ

1.1. Основные задачи технической диагностики

В технической диагностике рассматриваются методы оценивания состояния технических объектов. Объектом диагностирования служит изделие и его составные части.

Техническая диагностика решает широкий круг задач [9], при этом главной задачей является распознавание состояния объекта при ограниченной информации об условиях его функционирования.

Получение информации о состоянии объекта существенно затруднено, поскольку распознавание проводится в условиях эксплуатации [9,105]. По таким данным часто трудно сделать однозначное заключение о состоянии объекта, используют статистические и интеллектуальные методы [9].

К задачам диагностирования относят контроль состояния объекта, определение места и причин отказа, а также прогнозирование возможности функционирования объекта [13,36].

Контроль технического состояния [98] - определение состояния объекта в текущий момент времени на основе соответствия между значениями показателей функционирования объекта и требованиями документации.

При эксплуатации технический объект может находиться в одном из следующих состояний [3]:

- исправном (все характеристики объекта соответствует требованиям нормативно- технической документации);

- неисправном (характеристики не соответствуют хотя бы одному из требований документации);

- работоспособном (значения всех показателей функционирования соответствуют требованиям нормативно-технической документации);

- неработоспособном (значение хотя бы одного показателя функционирования не соответствует требованиям нормативно-технической документации).

Количественными характеристиками состояния объекта обычно являются контролируемые показатели функционирования с заданными нормативами по допустимому изменению их значений.

Кроме показателей для оценки технического состояния объектов в диагностике используется понятие «признак состояния» - это качественная или количественная характеристика свойств объекта.

Под признаком состояния понимают значение (или интервал значений) какого-либо показателя, устанавливаемого для отличия данного состояния от других состояний. Признаки представляют собой менее полную информацию, в отличии от показателей, по которым можно судить о состоянии технического объекта не только в настоящий момент времени, но и делать прогноз на определенный период времени.

Техническое диагностирование объектов представляет собой процесс исследования, результатом которого является заключение о техническом состоянии объекта с указанием вида технического состояния [99].

По каждому показателю функционирования в документации указывается норматив, обычно соответствующий одному из состояний: функционирования, работоспособности или исправности [102]. Иногда указываются нормативные значения показателя для нефункционирующего, неработоспособного или неисправного состояний [42].

Объект диагностирования – технический объект (система, машина, прибор, узел и т.д.), для которого решается задача распознавания состояния. Объектом диагностирования могут выступать любые технические системы [13,91], удовлетворяющие следующим условиям:

- эти объекты могут находиться хотя бы в одном из двух взаимоисключающих состояниях;

- в них можно выделить элементы, каждый из которых также может находиться хотя бы в одном из двух взаимоисключающих состояниях.

1.2. Статистические методы технической диагностики

1.2.1. Наивный байесовский классификатор

Статистические методы диагностики характеризуются безразмерными величинами – оценками вероятности различных состояний системы, поэтому при диагностике могут одновременно учитываться признаки различной физической природы [4, 5].

Байесовский классификатор – простой вероятностный классификатор, основанный на применении формулы Байеса со строгими предположениями о независимости показателей функционирования объекта [108].

Выдвигается предположение о независимости влияния на результаты классификации различных атрибутов, это предположение резко упрощает сопутствующие вычисления [108]. В связи с этим данный метод называется наивным байесовским классификатором.

Каждый объект характеризуется определенным набором показателей функционирования X. Объект соответствует классу C_i при выполнении условия

$$P(C_i|X) > P\left(C_j\middle|X\right), \tag{1.1}$$

здесь $P(C_i|X)$ – апостериорная вероятность того, что объект с заданным набором показателей X принадлежит классу C_i; $P\left(C_j\middle|X\right)$ – апостериорная

вероятность того, что этот объект принадлежит произвольному классу C_j, не совпадающему с C_i [73].

Итак, объект относится к классу C_i тогда, когда апостериорная вероятность его принадлежности этому классу больше апостериорной вероятности того, что он принадлежит объекта любому другому классу.

Под апостериорной вероятностью понимают условную вероятность события после учета некоторой новой информации, связанной с этим событием [47]; по существу, это вероятность события A при условии, что произошло некоторое другое событие B.

Необходимо найти класс C, для которого выполняется условие [108]:

$$c = \arg\max_{C} P(C|X). \tag{1.2}$$

Таким образом, находится класс, для которого максимальна вероятность при заданных показателях функционирования X

$$\mathrm{P(C|X)} = \frac{\mathrm{P(X|C)\cdot P(C)}}{\mathrm{P(X)}} \tag{1.3}$$

Воспользовавшись «наивным» предположением о том, что показатели X не зависят друг от друга, получим:

$$\mathrm{c} = \arg\max_{c \in C} P(c) \prod_{j=1}^{p} P\left(x_j \middle| c\right),$$

где p – количество показателей функционирования системы.

При бинарной классификации (Y = 1 – объект исправен, Y = 0 – объект неисправен) получим:

$$P(Y=1|X) = \frac{P(Y=1)\prod_{j=1}^{p} P(x_j|Y=1)}{P(Y=1)\prod_{j=1}^{p} P(x_j|Y=1) + P(Y=0)\prod_{j=1}^{p} P(x_j|Y=0)}, \tag{1.4}$$

где $P(x_j \mid Y=1), P(x_j \mid Y=0)$ - вероятности наличия признаков $x_1,\ldots, x_p$ в классах $Y = 1$ и $Y = 0$ соответственно.

Рассматриваемый классификатор обеспечивает ошибку, сопоставимую с другими классификаторами, такими, как нейронные сети и дерево решений. Однако на практике его использование ограничено, так как предположение об условной независимости классов далеко не всегда выполняется.

Несмотря на свою простоту, метод Байеса часто используется в задачах классификации, например, в таких, как распознавание текста. Наивный байесовский классификатор также широко используется при обнаружении спама (sms-сообщения, сайты, почта и другие). Но чаще всего он применяется как эталон для сравнения различных моделей алгоритмов, или как часть в алгоритмических композициях.

1.2.2. Дискриминантный анализ

Статистический метод классификации многомерных наблюдений при наличии так называемых обучающих выборок ("классификация с учителем"), предложенный Р. Фишером, называется дискриминантным анализом.

В дискриминантном анализе сравниваются несколько совокупностей по среднему значению некоторой переменной, в дальнейшем эта переменная используется для предсказания, к какому классу относится набор новых данных.

Канонической дискриминантной функцией называется линейная функция:

$$d_i(x) = q_0 + q_1 \cdot x_1 + \cdots + q_p \cdot x_p, \qquad (1.5)$$

где d_i – значение канонической дискриминантной функции; x_p- значение дискриминантной переменной; $q_0 \dots q_p-$ коэффициенты дискриминантной функции [21].

Для геометрической интерпретации дискриминантного анализа можно рассмотреть объект, характеризующийся двумя переменными X_1 и X_2, тогда на рис. 1.1 представлено разбиение объектов на два различных класса.

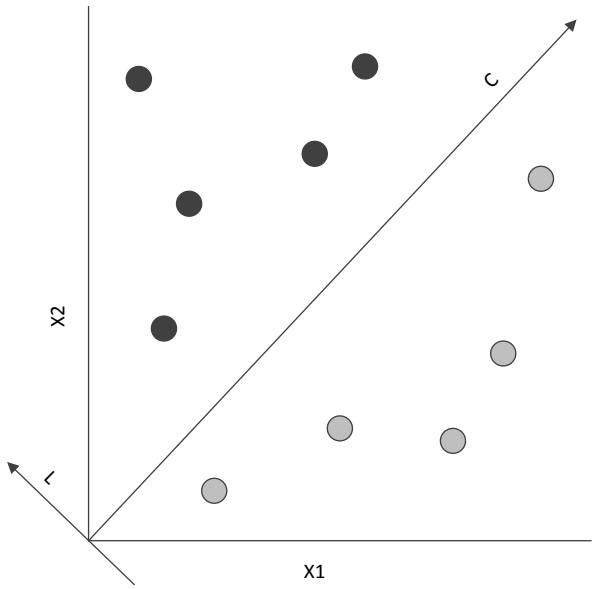


Рисунок 1.1. Геометрическая интерпретация дискриминантного анализа

Для наилучшего разделения двух рассматриваемых классов строится соответствующая линейная комбинация переменных x_1 и x_2. В двумерном случае эта задача сводится к определению новой системы координат так, чтобы проекции классов на новую ось *L* были как можно дальше друг от друга.

Ось *C,* перпендикулярная к оси *L,* разделяет два множества таким образом, чтобы точки оказались по разные стороны от этой прямой. При этом вероятность необходимо обеспечить минимум ошибки классификации.

Чтобы определить, к какому из двух классов относится объект, используются линейные функции [21]:

$$o_1(x) = q_0^1 + q_1^1 x_1 + \cdots + q_p^1 x_p,$$

$$o_2(x) = q_0^2 + q_1^2 x_1 + \cdots + q_p^2 x_p, \quad (1.6)$$

где *o(x)* – «счет», по которому делается вывод о принадлежности объекта к тому или иному классу.

В конечном итоге выбирается класс, у которого значение *o(x)* оказалось максимально.

$$c = \operatorname{argmax}_c(q_0^i + q_1^i x_1 + \cdots + q_n^i x_n) \quad (1.7)$$

В дискриминантном анализе используется теорема Байеса

$$P(Y = 1 \mid X) = \frac{P(Y = 1)G_1(X)}{P(Y = 1)G_1(X) + P(Y = 0)G_2(X)}, \quad (1.8)$$

и предположение о нормальности распределения исходных данных [21]:

$$G_1(X) = \frac{1}{(2\pi|\Sigma_1|)^{1/2}} \exp(-\frac{1}{2}(X - \mu_1)^T \Sigma_1^{-1}(X - \mu_1)), \quad (1.9)$$

$$G_2(X) = \frac{1}{(2\pi|\Sigma_2|)^{1/2}} \exp(-\frac{1}{2}(X - \mu_2)^T \Sigma_2^{-1}(X - \mu_2)),$$

где μ_1 и μ_2 - математические ожидания, $\sum_1$ и $\sum_2$ - ковариационные матрицы, $P(Y = 1|X)$ – условная вероятность исправного состояния технического объекта при наборе признаков *X*; $P(Y = 1)$, $P(Y = 0)$ – априорные вероятности принадлежности состояния технического объекта классам исправных и неисправных соответственно.

К недостаткам моделей дискриминантного анализа можно отнести то, что они плохо работают на несбалансированных классах, а также то,

что они не применимы для решения нелинейных задач. Кроме того, дискриминантный анализ склонен к переобучению [18].

1.2.3.Логистическая регрессия

В логистической регрессии, как и в любой множественной регрессии, анализируются связи между несколькими независимыми показателями (переменными или регрессорами) и зависимым откликом (переменной) [12,64], но при этом отклик является бинарным (например, исправен или неисправен) [86]. Логистическая регрессия оценивает вероятность наступления события для конкретного состояния объекта (возврат кредита/дефолт, исправный/неисправный объект, и т.д.).

Любая регрессионная модель могут быть представлена в виде:

$$y = F(x_1, x_2, \dots, x_p). \tag{1.10}$$

Линейная модель множественной регрессии предполагает, что зависимая переменная - линейная функция независимых:

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p. \tag{1.11}$$

Эту регрессию можно использовать для задачи оценки вероятности события, определив коэффициенты регрессии, например, методом наименьших квадратов [82]. Однако такая регрессия не учитывает бинарную природу отклика. В результате значения отклика могут оказаться большими 1 и меньшими 0. Но такие значения вообще недопустимы для оценки вероятности.

Один из наиболее распространенных приемов в этой ситуации – использование логистической регрессии [66]: вместо предсказания бинарной переменной отклик представляет непрерывную переменную со значениями на отрезке $[0, 1]$ при любых значениях независимых переменных. Это достигается применением следующего уравнения:

$$f(z) = \frac{1}{1+e^{-z}} \quad (1.12)$$

где

$$z = q_0 + q_1 x_1 + \ldots + q_p x_p$$

– стандартное уравнение регрессии.

Так как отклик Y принимает одно из двух значений, то вероятность исправности технического объекта $Y = 1$ равна [18]:

$$P\{Y = 1|X\} = f(z)\,, \quad (1.13)$$

а вероятность неисправности объекта $Y = 0$:

$$P\{Y = 0|X\} = 1 - f(z) \quad (1.14)$$

Таким образом, логистическая регрессия находится на основе следующего выражения:

$$\log \frac{P\{Y = 1|X\}}{P\{Y = 0|X\}} = \frac{f(z)}{1-f(z)} = q_0 + q_1 x_1 + \cdots + q_p x_p \quad (1.15)$$

Для нахождения параметров $q_0, \ldots, q_p$, обычно используют метод максимального правдоподобия: функции правдоподобия максимизируется с использованием различных вариантов метода градиентного спуска [2], метода Ньютона-Рафсона [16] или других методов.

Ограничение в использовании логистической регрессии связано с ее чувствительностью к корреляции между факторами [66], поэтому для корректной работы метода необходимо исключить сильно коррелированные показатели. К преимуществам этой модели можно отнести возможность взвешивания факторов, влияющих на результат.

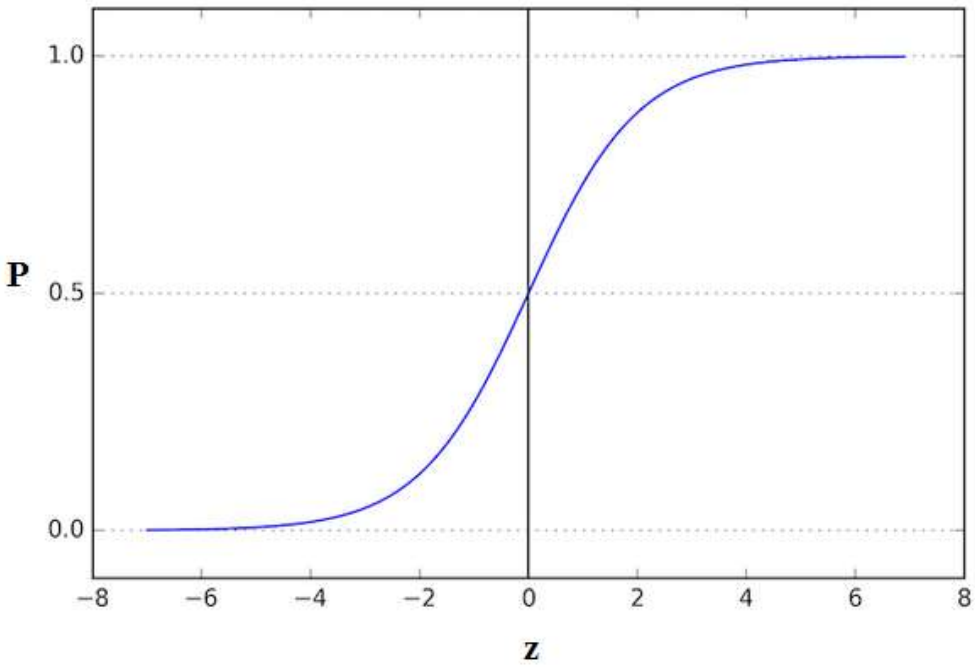


Рисунок 1.2. Кривая логистической регрессии

1.2.4. Метод опорных векторов

В методе опорных векторов объекты описываются p-мерными вещественными векторами: $X = \mathbb{R}^p, Y = \{-1, +1\}$. Это статистический метод, использующий методы оптимизации и аналитической геометрии для решения задачи бинарной классификации [76].

Необходимо построить гиперплоскость максимальной размерности [107], которая разделяет наблюдения на два класса: исправный и неисправный объекты.

Строится линейный классификатор

$$f(x) = sign(\sum_{j=1}^{p} w_j x_j - w_0), . \quad (1.18)$$

где x_j – показатели функционирования объекта, w_j – параметры алгоритма. Уравнение

$$\sum_{j=1}^{p} w_j x_j = w_0$$

описывает гиперплоскость, разделяющую классы.

На рис. 1.3 приведен пример с объектами двух типов. Разделяющая линия задает границу, справа от которой – все объекты типа принадлежат одному классу, а слева - другому.

Расстояние от оптимальной гиперплоскости до класса должно быть максимально, поэтому для нахождения параметров w_j решается задача оптимизации (квадратичного программирования) [78]: требуется минимизировать норму вектора w при наличии l ограничений (l – объем обучающей выборки, y_i – состояние объекта в i-ом наблюдении):

$$\frac{1}{2}\|w\|^2 \to \min, \tag{1.19}$$

$$y_i(\sum_{j=1}^{p} w_j x_j - w_0) \geq 1, \quad i = 1,...,l.$$

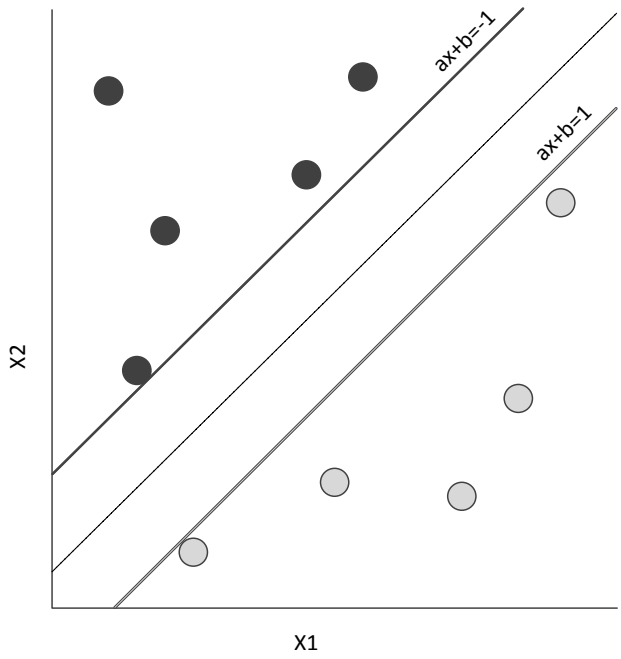


Рисунок 1.3. Метод опорных векторов

Задача решается с помощью метода множителей Лагранжа. Исправность технических объектов определяется таким образом: при $f(X) = 1$ технический объект считается исправным, при $f(X) = -1$ неисправным.

Одним из минусов метода является его чувствительности к шумам и стандартизации данных. Кроме того, далеко не всегда рассматриваемые классы линейно разделимы, однако задача может быть обобщена и на нелинейный случай.

Однако существует и ряд преимуществ рассматриваемого метода, например, так как задача оптимизации в данном случае – задача квадратичного программирования, таким образом задача имеет единственное решение.

При расчете вероятности нахождения объекта в исправном состоянии в качестве функции преобразования может быть использована логистическая функция:

$$P(Y = 1 | X) = \frac{1}{1 + \exp(-\sum_{j=1}^{p} w_j x_j + w_0)}. \qquad (1.20)$$

1.2.5. Методы принятия решения

Задача технической диагностики – определение состояния объекта по результатам измерений ряда косвенных показателей функционирования этого объекта. Одной из важных особенностей технической диагностики является распознавание при наличии лишь ограниченной информации, когда требуется руководствоваться определенными приемами и правилами для принятия обоснованного решения.

Для решения поставленной задачи могут быть использованы статистические методы принятия решений. При этом решающее правило в методах принятия решений подбирается так, чтобы выполнялось

определенное условие оптимальности, например, условие минимума риска [9].

В *методе минимального риска* вероятность принятия ошибочного решения находится как результат минимизации точки экстремума среднего риска ошибочных решений при максимуме правдоподобия (рис. 1.4).

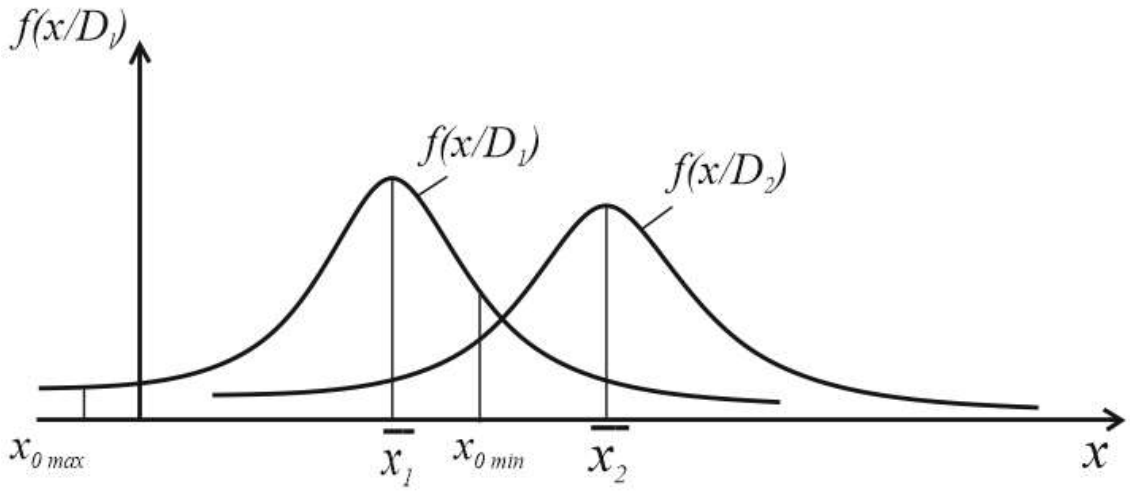


Рисунок 1.4. К оценке среднего риска

Отношение правдоподобия – это отношение плотностей вероятностей распределения x при двух состояниях: D_1 – исправном состоянии технического объекта, и D_2 — неисправном состоянии объекта; при этом учитываются C_{21} — цена ложной тревоги, и C_{12} — цена пропуска цели (первые индекс означает принятое состояние объекта, второй — действительное).

Вместо отношения правдоподобия часто используется логарифм этого отношения. Логарифмическая функция монотонно возрастает при возрастании аргумента, поэтому результат не изменяется. На практике расчет логарифма отношения правдоподобия при некоторых распределениях, например, нормальном, оказывается проще.

В *методе минимального числа ошибочных решений* вероятность ошибочного решения может быть найдена так:

$$P_{\text{ош}} = P_1 \int_{x_0}^{\infty} f(x/D_1)dx + P_2 \int_{-\infty}^{x_0} f(x/D_2)dx \qquad (1.21)$$

здесь P_1 и P_2 – априорные вероятности соответствующих диагнозов.

Этот метод применяется в случаях, когда цена пропуска дефекта примерно равна цене ложной тревоги.

Метод Неймана—Пирсона. Для значений P_i, расположенных в диапазоне от 0 до 1, риск принятия ошибочного решения возрастает и при $P_1 = P_1^*$ становится максимальным. Идея данного метода заключается в нахождении величины x_0, которая бы минимизировала потери, обусловленные ошибочными решениями при наиболее худших значениях P_i [3].

Далее, используя метод Неймана—Пирсона, минимизируется вероятность пропуска цели при выбранном допустимом уровне вероятности ложной тревоги [9].

Вероятность ложной тревоги:

$$P_1 \int_{x_0}^{\infty} f(x/D_1)dx \leq A \qquad (1.22)$$

где A- заданный допустимый уровень вероятности ложной тревоги; P_1 – вероятность того, что объект находится в исправном состоянии.

1.3.Интеллектуальные методы технической диагностики

1.3.1. Нейронные сети

Нейронные сети представляют математические модели, и их программные или аппаратные реализации, построенные по принципу организации биологических нейронных сетей – сетей нервных клеток живого организма [6]. Это понятие возникло при изучении процессов, протекающих в мозге при мышлении, и при попытке смоделировать эти процессы.

Нейронная сеть - это последовательность нейронов, соединенных между собой синапсами. Основная функция искусственного нейрона состоит в построении выходного сигнала, основываясь на входных сигналах.

В структуру нейрона входят синапсы, сумматор, нелинейный преобразователь [68]. Синапсы выполняют функции по передаче сигналов, а также их умножению на коэффициенты (веса), которые определяют силу связи. Сумматоры объединяют сигналы исходящих от синапсов. В последнем компоненте происходит преобразование сигнала, поступающего от сумматора функцией активации, после чего результат подается на выход.

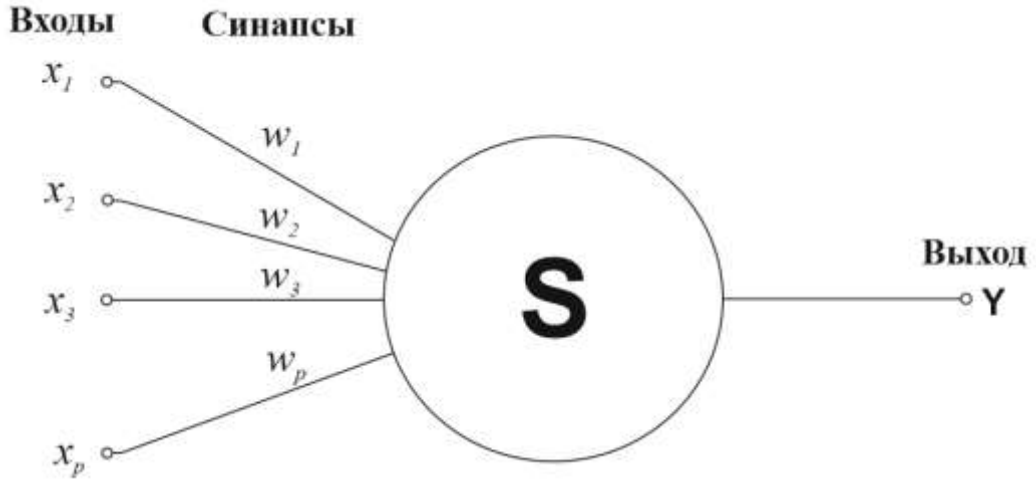


Рисунок 1.5. Структура нейрона

Состояние узла нейронной сети U определяется как результат линейной комбинации входных сигналов и описывается следующей моделью [68]:

$$U = \sum_{j=1}^{p} w_j x_j , \qquad (1.23)$$

где w_j – весовые коэффициенты, x_j – входной сигнал, p – число входов.

Выходной сигнал τ:

$$\tau = C(U), \tag{1.24}$$

где $C(U)$ – функция активации, осуществляющая преобразование.

Рассматривая выходной сигнал, как оценку вероятности $P(Y=1|X)$ того, что рассматриваемый объект исправен при заданном наборе показателей функционирования X, и используя в качестве функции активации, например, логистическую функцию, получим:

$$P(Y=1|X) = \frac{1}{1+\exp(-\sum_{j=1}^{p} w_j x_j)}. \tag{1.25}$$

Обучение – одно из главных преимуществ нейронных сетей по сравнению с другими методами, которое является многопараметрической задачей нелинейной оптимизации. Решением этой задачи является определение коэффициентов, характеризующих связи между нейронами.

Нейронные сети могут находить связи различной степени сложности между входными и выходными данными, а также обобщать. Таким образом, используя успешно обученную модель, можно получить информацию о состоянии объекта при неполных и искаженных данных в обучающей выборке.

Наиболее часто используется архитектура многослойных нейронных сетей [97]. Такая нейронная сеть может моделировать функцию практически любой степени сложности Сложность функции определяется числом слоев и числом элементов в каждом слое. Определение числа промежуточных слоев и числа элементов в них является важной задачей проектирования сети.

Среди многослойных нейронных сетей можно выделить следующие наиболее значимые и важные классы нейронных сетей [68]:

- сети прямого распространения – искусственные нейронные сети, сигнал которых распространяется строго от входного слоя к выходному. Данные сети широко используются и успешно решают задачи: прогнозирование, кластеризация и распознание;

- рекуррентные нейронные сети или сети обратного распространения – отличие от предыдущего типа нейронных сетей в том, что в сетях такого типа сигнал может идти как от входного сигнала слоя к выходному, так и наоборот. Таким образом в рекуррентных сетях выход любого нейрона может определяться не только его весами и входным сигналом, но еще и предыдущими выходами;

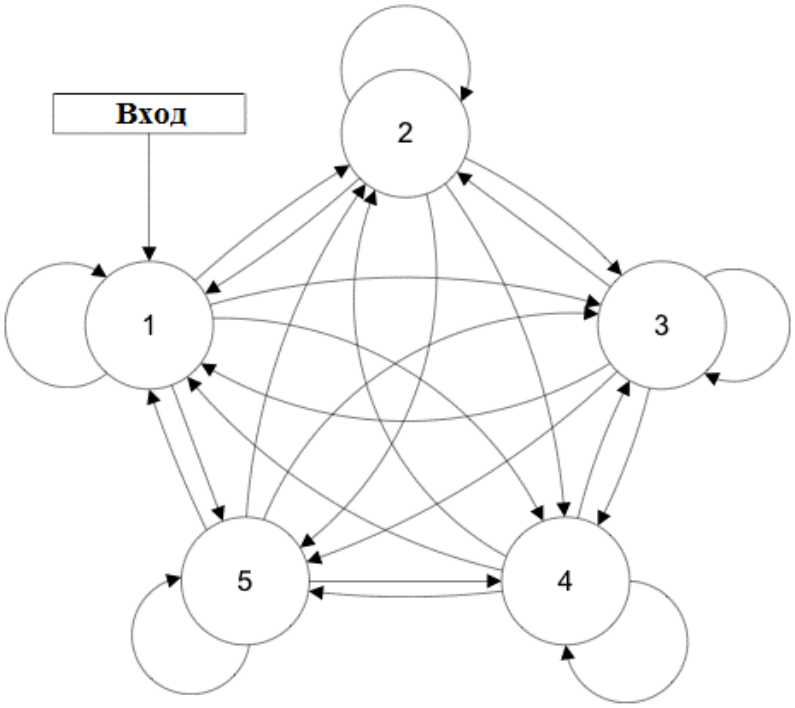


Рисунок 1.6. Рекуррентные нейронные сети

Как видно на рисунке 1.6, в рекуррентных нейронных сетях сигнал может идти в разных направлениях, как к любому слою до и после, так и обратно на слой, на котором только что находился.

На рисунке 1.7 представлена архитектура трехслойной нейронной сети прямого распространения.

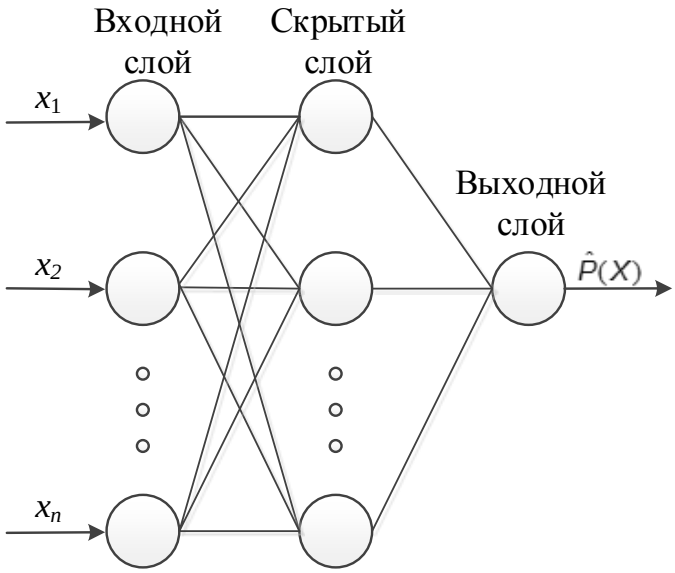


Рисунок 1.7. Архитектура нейронной сети

Количество нейронов во входном слое определяется количеством показателей технического объекта, а в выходном слое – размерностью вектора ответов. При бинарной классификации состояния технического объекта в выходном слое достаточно одного нейрона. По теореме Колмогорова количество нейронов в скрытом слое может быть найдено, как (2*n + 1), где n – количество показателей функционирования технического объекта. Для каждой исследуемой выборки проводится перебор количества нейронов в скрытом слое, начиная с двух и до (2*n + 1). При этом выбирается сеть с минимальной дисперсией ошибки.

Для обучения нейронной сети обычно используется метод обратного распространения ошибки, основная идея которого заключается в предположении о том, что ошибка распространяется по сети от выхода к входам.

1.3.2.Методы классификации, используемые в машинном обучении

Методы машинного обучения в настоящее время широко развиваются и активно применяются в различных областях деятельности [37]. Используется множество различных подходов к классификации. Это и классические статистические методы (наивный байесовский

классификатор, дискриминантный анализ, логистическая регрессия, метод опорных векторов), и методы, специально ориентированные на машинное обучение (нейронные сети), композиционные методы (бэггинг, бустинг) и другие.

Часто диагностика сводится к разделению состояний объекта на два класса: исправное и неисправное. При решении такой задачи могут быть использованы методы машинного обучения по прецедентам (с учителем), а именно методы, предназначенные для бинарной классификации.

В качестве исходных данных рассматриваются известные результаты (прецеденты) оценки состояния системы: исправна или неисправна техническая система при заданных значениях контролируемых показателей.

Таким образом, имеются множество ситуаций с заданными показателями и множество возможных состояний системы, которые в совокупности образуют исходную выборку.

Предполагается, что существует определенная связь между показателями функционирования объекта и его состояниями. На основе исходных данных необходимо восстановить эту зависимость, то есть построить алгоритм, способный для заданного набора показателей функционирования объекта выдать достаточно точный ответ о его состоянии.

Оценка качества модели с точки зрения возможности прогнозирования проводится по контрольной выборке. С этой целью исходная выборка из l опытов разделяется на два непересекающихся подмножества: обучающую выборку объёма $l_о$, с помощью которой и решается задача обучения, и контрольную (или тестовую) объёма $l_к = l - l_о$, не используемую для обучения [31].

Часто применяется кросс-валидация: выборка разбивается на N частей (на практике часто принимают $N = 5$ или $N = 10$). Части $(N - 1)$

используется для обучения, а оставшаяся – для контроля. Последовательно перебираются все варианты.

Для каждого варианта разбиения по выборке объемом $l_о$ проводится обучение и определяется функция ошибок $Q(a,X)$ на контрольной выборке $l_к$. Среднее значение этой функции по всем вариантам характеризует обобщающую способность алгоритма [28].

Методы машинного обучения применяются в различных областях, так наивный байесовский классификатор широко применяется в задачах классификации текста в различных информационных сферах.

Например, в статье С.В. Шанова [73] байесовский классификатор используется для определения тематики текста. Были получены результаты, по которым можно утверждать, что выбранный автором метод справился с поставленной задачей. Однако нет доказательств того, что данный метод превосходит по результатам другие методы и модели.

Ансамблевые (композиционные) методы могут быть использованы для прогнозирования оттока клиентов. Так в статье А.Б. Гольдштейна [19] производится анализ эффективности методов бустинга и бэггинга для этих целей.

Часто методы машинного обучения применяются в медицине. Например, в статье [53] для классификации патологий диссеминированного туберкулёза лёгких рассматриваются следующие методы: логистическая регрессия, метод опорных векторов, нейронные сети. Однако эти методы не показали достаточной точности для применения их в компьютерной системе медицинской диагностики.

Градиентный бустинг в статье [79] служит инструментом в биологической области, с его помощью проводилась классификация молекул. По результатам исследования градиентный бустинг показал наилучшую точность прогнозирования.

Применение машинного обучения, а именно нейронных сетей, в технической диагностике рассмотрено в работах И.В. Васильева и С.В.

Жернакова; например, в статье [15] рассматриваются нейронные методы решения задач контроля и диагностики технического состояния авиационного газотурбинного двигателя. Для решения поставленной задачи авторы разработали архитектуру нейросетевой экспертной системы «Эксперт Нейро».

1.3.3. Применение композиционных моделей

Рассмотренные алгоритмы классификации не всегда приводят к удовлетворительному решению задачи оценки состояния технического объекта. В этих случаях часто строят различные композиции из алгоритмов. При этом погрешности отдельных алгоритмов могут частично компенсироваться. Композиция или ансамбль алгоритмов, то есть набор моделей, используемых для решения одной и той же задачи, позволяет повысить точность классификации.

Два главных метода построения композиции – бэггинг и бустинг– дают обычно более точный результат, чем применение отдельного алгоритма на конкретном наборе данных.

Модели бустинга

Бустинг - один из эффективных методов машинного обучения, представляет последовательное построение ансамбля из «слабых» алгоритмов, при котором каждый следующий алгоритм компенсирует недостатки предыдущих (boosting – усиление). Основные причины широкого распространения бустинга – простота, универсальность, гибкость, а также высокая обобщающая способность [86].

Для разделения объектов на два класса могут быть применены различные разновидности бустинга: AdaBoost, GentleBoost, LogitBoost, RUSBoost, а также градиентный бустинг.

Классификатор строится путем взвешенного голосования композиции базовых правил. Используется информация об ошибках предыдущих правил: веса объектов выбираются так, чтобы новое правило точнее работало на тех объектах, на которых предыдущие правила чаще ошибались.

Классификатор представляют в виде знака функции

$$H(x) = sign\left[\sum_{t=1}^{T} \alpha_t h_t(x)\right], \qquad (1.26)$$

где $h_t(x)$ – базовые классификаторы, возвращающие один из двух результатов: -1 или +1; α_t – коэффициент взвешенного голосования для соответствующего классификатора.

Для расчета вероятности нахождения объекта в исправном состоянии при использовании логистической функции в качестве функции преобразования по аналогии с (1.20) получим:

$$P(Y=1|X) = \frac{1}{1+\exp(-\sum_{t=1}^{T} \alpha_t h_t)}. \qquad (1.27)$$

Качество композиции можно оценить по числу ошибок, допускаемых ею на заданной выборке:

$$Q = \sum_{i=1}^{N} L(y,H) = \sum_{i=1}^{N}\left[y^{(i)} \sum_{t=1}^{T} \alpha_t h_t(x^{(i)}) < 0\right], \qquad (1.28)$$

здесь $L(y,H)$ – функция потерь. Функционал Q необходимо минимизировать. Используются различные подходы. При применении алгоритма адаптивного усиления AdaBoost предполагается экспоненциальная аппроксимация функции потерь [110]:

$$L(y,H) = \exp(-yH). \qquad (1.29)$$

Последовательности действий в алгоритме AdaBoost такова: инициализируются веса наблюдений $w_i = 1/N$, организуется цикл $t = 1 \ldots T$, обучается простой классификатор $h_t(x)$ и определяется его ошибка (суммируются веса ошибочно классифицированных наблюдений)

$$\varepsilon_t = \sum_{i:h_t(x^{(i)}) \neq y^{(i)}} w_i(t), \tag{1.30}$$

коэффициент взвешенного голосования определяется по формуле:

$$\alpha_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t} \quad . \tag{1.31}$$

Далее производится перерасчет весов: если наблюдение классифицировано правильно, вес уменьшается, если неправильно – вес увеличивается:

$$w_i(t+1) = w_i(t) \exp(-\alpha_t y^{(i)} h_t(x^{(i)})), \tag{1.32}$$

при этом веса наблюдений нормируются; далее проводится переход к следующему t. По окончании работы алгоритма получаем итоговый классификатор.

Так как результаты выполнения методов бустинга представляются в отличной от других методов форме, вначале все приводятся к общему формату (вектор вероятностей от 0 до 1). Для методов GentleBoost, AdaBoost, LogitBoost результатом будет вектор оценок принадлежности к одному из классов, чем значение больше нуля, тем выше вероятность принадлежности к классу 1 (исправно) и наоборот, чем ближе к нулю, тем выше вероятность принадлежности к классу 0 (неисправно).

Чтобы привести к общему формату, проводится нормирование по максимальному по модулю значению.

$$M = \max(Sn_i) \, ; i = \overline{1..n}, \tag{1.33}$$

где n – количество элементов в контрольной выборке.

$$S = \frac{S_n}{2M} + 0.5, \tag{1.34}$$

где S_n – вектор, результат выполнения метода.

По результатам выполнения всех методов выводится ошибка классификации для каждого метода:

$$error = \frac{\sum_{i=1}^{n} |Y_{test} - Y_{pr}|}{n} \cdot 100, \tag{1.35}$$

где Y_{pr} – вектор прогнозируемых результатов.

Другой подход, используемый при наличии шумовых данных в исходной выборке, разновидность бустинга LogitBoost, основанная на идее логистической регрессии.

В алгоритме LogitBoost строится композиция бинарных классификаторов, использующая идеи логистической регрессии. Функция потерь имеет вид:

$$L(y, F) = \log(1 + \exp(-2yF), \tag{1.36}$$

где

$$F(x) = \frac{1}{2}\log\frac{P(y=1|x)}{P(y=-1|x)} \tag{1.37}$$

RUSBoost пытается смягчить проблему дисбаланса классов, регулируя распределение классов набора учебных данных. RUSBoost просто удаляет наугад примеры из класса, у которого большее количество примеров, пока не будет достигнуто требуемое распределение классов.

При использовании метода RUSBoost ошибка вычисляется по формуле

$$\epsilon_t = \sum_{(i,y):y_i \neq y} D_i(i)(1 - h_t(x_i, y_i) + h_i(x_t, y_i) \tag{1.38}$$

коэффициенты

$$\alpha_t = \frac{\epsilon_t}{1-\epsilon_t}. \quad (1.39)$$

Для обновления используется соотношение

$$D_{t+1}(i) = D_{t+1}(i) \cdot \alpha_t^{\frac{1}{2}\left(1+h_t(x_i,y_i)-h_t(x_i,y:y\neq y_i)\right)}. \quad (1.40)$$

Проводится нормировка

$$D_{t+1}(i) = \frac{D_{t+1}(i)}{\sum_i D_{t+1}(i)} \quad (1.41)$$

классификатор имеет вид:

$$H(x) = arg \max_{y\epsilon Y} \sum_{t=1}^{T} h_t(x,y) \cdot \log \frac{1}{\alpha_t}. \quad (1.42)$$

Для метода RUSBoost результатом выполнения будет два вектора оценок, в первом оценка принадлежности к классу 0, во втором оценка принадлежности к классу 1. Чтобы представить результат к формату вероятностей от 0 до 1, вычитаем из значений класса 1 значения класса 0, затем, как и для остальных методов бустинга, проводится нормировка по максимальному по модулю значению.

GentleBoost (также известный как GentleAdaBoost) сочетает в себе функции AdaBoost и LogitBoost. Как и AdaBoost, GentleBoost минимизирует экспоненциальную потерю. Но его числовая оптимизация настроена по-другому.

В методе GentleBoost

$$f_t(x) = arg \min_{f\epsilon\mathcal{F}} \epsilon(f) = \sum_{i=1}^{k} w_i (y_i - f(x_i))^2, \quad (1.43)$$

для обновления используется соотношение

$$F(x) = F(x) + f_t(x) \quad (1.44)$$

проводится перерасчет весов:

$$w_i = w_i e^{-y_i f_t(x_i)} \quad (1.45)$$

и получаем итоговый классификатор в виде:

$$F(x) = \sum_{t=1}^{T} f_t(x). \quad (1.46)$$

Для градиентного бустинга решается задача:

$$\sum_{i=1}^{m} L(y_i, a(x_i) + b(x_i)) \to min \quad (1.47)$$

Но данная задача не всегда решается аналитически (из-за достаточно сложных функций ошибок). Перепишем (1.46) в виде

$$F = \sum_{i=1}^{m} L(y_i, a(x_i) + b(x_i)) \to min_{(b_1,\dots,b_m)}, \quad (1.48)$$

если рассматривать эту задачу как задачу минимизации фукции $F(b_1, \dots, b_m)$, то, учитывая, что функция многих переменных максимально убывает в направление своего антиградиента:

$$-\left(L'\left(y_1, a(x_1)\right), \dots, L'\left(y_m, a(x_m)\right)\right), \quad (1.49)$$

получаем ответы алгоритма b

$$b_i = -L'\left(y_i, a(x_i)\right), i \in \{1, \dots, m\}, \quad (1.50)$$

который следует настраивать на обучающей выборке:

$$(x_i, -L'\left(y_i, a(x_i)\right))_{i=1}^{m} \quad (1.51)$$

Бэггинг деревьев решений

Метод бэггинга (bootstrap aggregation) был предложен Л. Брейманом в 1996 году. Главной особенностью рассматриваемого метода является использование бустрепа, при котором обучающая выборка формируется из исходной следующим образом: случайным образом из исходной выборки выбирается состояние объекта, после чего оно возвращается назад, так продолжается пока выборка не достигнет объема l [89].

Так как при построении композиции используются различные базовые алгоритмы, то ошибки каждого из них компенсируются при голосовании.

При достаточно малых выборках бэггинг также является эффективным методом, так как базовые алгоритмы значительно отличаются друг от друга даже, когда исключается небольшая доля объектов из обучающей выборки. При этом в некоторые выборки могут не попасть выбросы из исходной выборки, тем самым, при построении композиции, модели, построенные на таких выборках, сглаживают ошибки других моделей.

В тех случаях, когда исходных данных достаточно много, существует возможность построения подвыборок меньшей длины $l_0 \ll l$, при таком подходе появляется задача нахождения оптимального значения l_0.

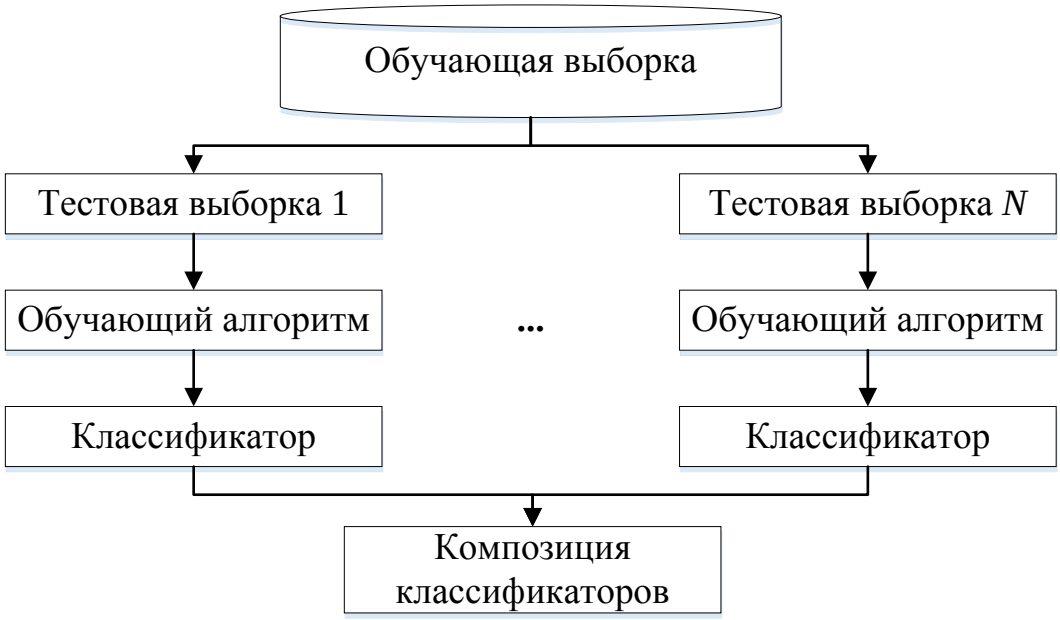


Рисунок 1.8. Блок-схема алгоритма бэггинга

На рисунке 1.8 представлена блок-схема алгоритма бэггинга, где с помощью бутстрэпа генерируются выборки $X_1,\ldots, X_M$, на каждой из которой проходит обучение и строятся классификаторы $a_i(x)$. Усреднением полученных классификаторов получается итоговый классификатор:

$$\alpha(x) = \frac{1}{M}\sum_{i=1}^{M} \alpha_i(x) \qquad (1.52)$$

Обучающим алгоритмом в бэггинге деревьев решений является дерево принятия решений, широко применяющееся в статистике и анализе данных.

Дерево решений состоит из «ветвей» и «листьев». Значение атрибутов записывается в ветвях (ребрах графа), от которых зависит целевая функция, значение которой хранится в листьях. При этом также существуют родительские узлы и потомки, по которым происходит разветвление [71].

Целью процесса построения дерева принятия решений является создание модели, с помощью которой можно произвести классификацию объекта, используя данные о нескольких переменных.

Рассмотрим общий алгоритм построения дерева:

- Вначале производится разделение по самому значимому фактору.
- Далее рекурсивный процесс продолжается, пока разбиение не приведёт к значимому различию между исправным и неисправным состояниями.
- На рисунке 1.9 показан вариант структуры дерева решений по исправности системы водоочистки. Как видно на этом рисунке, в листьях дерева хранятся данные об исправности объекта, а в ветвях дерева хранятся условия, при которых принимается значение о состоянии исправности системы водоисточника.

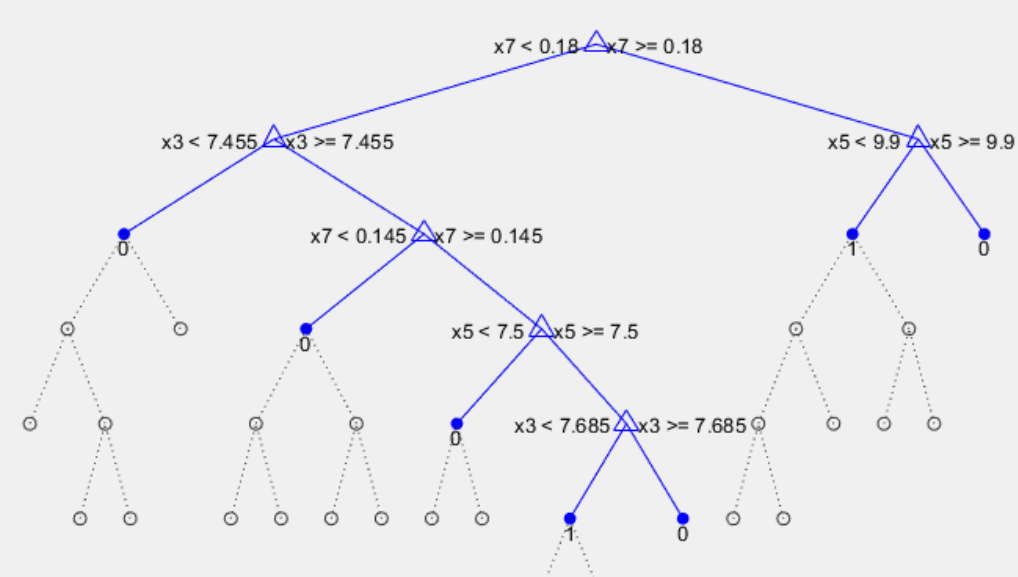


Рисунок 1.9. Дерево решения

Преимуществом бэггинга деревьев решений является возможность обработки исходной выборки большого объема, а также отсутствие чувствительности к масштабированию.

Рассмотренный метод позволяет производить построение модели классификации на данных с пропущенными значениями, а также способен эффективно работать как с непрерывными, так и с дискретными признаками.

Недостаток бэггинга деревьев решений – склонность к переобучению, т.е. высокая точность классификации на обучающей выборке, но низкая точность на тестовой выборке. К недостаткам также можно отнести достаточно большой размер получающихся моделей.

1.4. Постановка задач исследования

В соответствии с проведенным обзором для решения задачи технической диагностики могут быть использованы различные статистические методы и методы машинного обучения. Однако не существует универсальной модели, которая могла бы спрогнозировать состояние технического объекта с явным преимуществом.

Для повышения точности прогнозирования могут быть применены агрегированные методы классификации, которые используют набор базовых моделей.

Необходимо также учесть способ формирования контрольной и обучающей выборок, и влияние различных других факторов на качество диагностики.

Таким образом, возникают следующие основные задачи исследования, направленные на повышение точности прогнозирования состояния технического объекта по заданным показателям функционирования:

1. Анализ эффективности применения различных подходов машинного обучения для оперативной диагностики нарушений функционирования технического объекта при заданном наборе контролируемых показателей.

2. Разработка математических моделей и алгоритмов для диагностики состояния технического объекта с применением агрегированных классификаторов.

3. Оценка влияния различных факторов на качество диагностики.

4. Разработка программы для проведения испытаний по диагностике функционирования технического объекта

5. Разработка программного комплекса для автоматизированной оценки исправности технического объекта и прогнозирования его состояния с применением методов машинного обучения.

6. Оценка эффективности разработанных моделей и программных средств и численное исследование на реальных технических объектах.

ГЛАВА 2. РАЗРАБОТКА МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ДИАГНОСТИКИ ФУНКЦИОНИРОВАНИЯ ТЕХНИЧЕСКИХ ОБЪЕКТОВ НА ОСНОВЕ АГРЕГИРОВАННЫХ КЛАССИФИКАТОРОВ

2.1. Постановка задачи диагностирования

Исходные данные для диагностики состояния объекта представляются в виде матрицы X показателей функционирования системы, элементы которой x_{ij} – результат i-го наблюдения по j-му показателю; $i = 1, \ldots, l$, $j = 1, \ldots, p$ (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. Строке x_i матрицы показателей X соответствует определенное значение y_i вектора Y, характеризующее исправность объекта. Множество пар (x_i, y_i) дает выборку исходных данных – прецедентов.

Задача состоит в построении модели $a(x, w)$, которая предскажет ответ Y для любого заданного X. Обычно используются линейные модели [18]:

$$a(x, w) = w_0 + w_1 x_1 + \ldots + w_p x_p, \qquad (2.1)$$

где $w = (w_0\, w_1 \ldots w_p)$ – вектор параметров модели.

При бинарной классификации вместо нуля и единицы часто используют множество ответов $Y = \{-1; +1\}$. В этом случае модель алгоритма примет вид:

$$a(x, w) = sign \sum_{j=0}^{p} w_j x_j \qquad (x_0 = 1). \qquad (2.2)$$

Процесс подбора оптимальных параметров алгоритма w_j по исходным данным, называется обучением алгоритма. Найденные параметры должны обеспечить оптимальное значение функционала качества. В рассматриваемой задаче чаще всего минимизируется функционал ошибок – это среднее количество несовпадений, где $L(a,x_i)$ называют функцией потерь:

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l} L(a, x_i) = \frac{1}{l}\sum_{i=1}^{l} [a(x_i) - y_i] \to \min. \qquad (2.3)$$

2.2. Критерии качества диагностики

Наиболее распространенным показателем, который может быть использован для оценки качества бинарной классификации, – доля правильных ответов на контрольной выборке

$$Accuracy = \frac{Q}{N}, \qquad (2.4)$$

где Q – количество правильно классифицированных объектов контрольной выборки, а N – общий размер контрольной выборки. Чаще используется противоположная характеристика - доля (или процент) ошибок на контрольной выборке.

Иногда для оценки качества классификации применяют средний квадрат отклонений истинной вероятности исправности в r-м наблюдении $P(Y_r)$ от её прогнозируемого значения $\hat{P}(X_r)$:

$$\sigma^2 = \frac{1}{N}\sum_{r=1}^{l} (P(Y_r) - \hat{P}(X_r))^2. \qquad (2.5)$$

При несбалансированных классах (когда исправных состояний объекта значительно больше, чем неисправных) доля ошибок не может

объективно оценивать качество классификации [65]. Гораздо более информативны точность

$$P = \frac{tp}{tp + fp}, \tag{2.6}$$

и полнота

$$R = \frac{tp}{tp + fn}, \tag{2.7}$$

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 критерий

$$F = \frac{2PR}{P + R}, \tag{2.8}$$

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

Функционалом качества также может быть выбрана площадь под ROC-кривой (receiver operating characteristics): *AUC* (area under the curve) [90]. ROC-кривая образуется, если по оси абсцисс брать значения $fp(c)$, а по оси ординат $tp(c)$, где c - порог.

Площадь под ROC-кривой позволяет оценить модель в целом, не привязываясь к конкретному порогу. Критерий AUC-ROC устойчив к несбалансированным классам и может быть интерпретирован как вероятность того, что случайно выбранный объект из класса 1 будет иметь значение вероятности ближе к 1, чем случайно выбранный объект из класса 0.

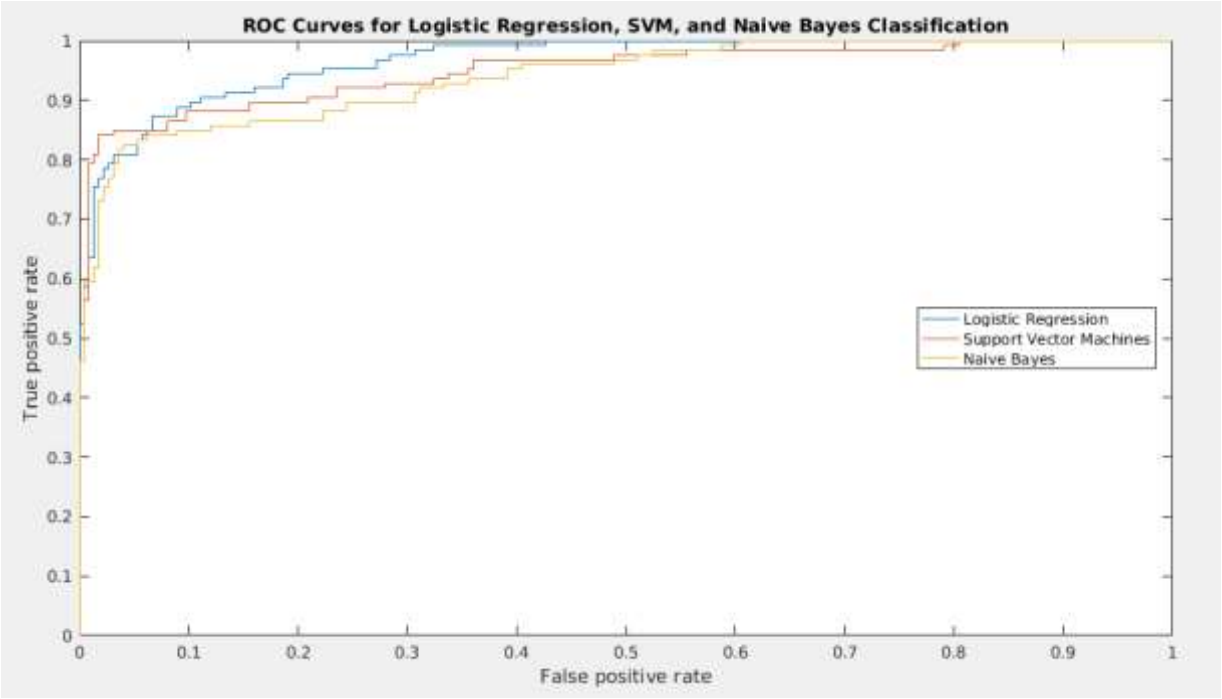


Рисунок 2.1. ROC- кривые

На рис. 2.1 показаны такие кривые, построенные в системе Matlab для рассмотренного ниже примера диагностики, при использовании трех методов бинарной классификации: логистической регрессии, метода опорных векторов и наивного байесовского классификатора.

Следует отметить то обстоятельство, что в зависимости от выбранного критерия качества решение задачи бинарной классификации различается весьма существенно.

На рис. 2.2 показаны значения различных критериев качества при использовании нескольких классификаторов при диагностике функционирования системы водоочистки.

Видно, что по критериям минимума ошибки по отдельной контрольной выборке и при кросс-валидации, а также максимуму полноты (2.7) и *F*-меры (2.8) лучшим методом диагностики оказался бэггинг деревьев решений БДР, по критерию минимума «дисперсии» по кросс-валидации – среднему квадрату отклонения истинной вероятности

исправности от её прогнозируемого значения (2.5) – дискриминантный анализ ДА, по максимуму точности (2.6) лучший результат показали сразу два метода – метод опорных векторов МОВ и градиентный бустинг GrB, последний метод оказался предпочтительным и по критерию максимума *AUC*.

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, применим площадь *AUC* под кривой ошибок.

	Метод	Ошибка по к/в	Ошибка кросс-вал	Дисперсия по крос-вал	Точнос...	Полнота	F-мера	AUC
1	ЛР	18.3908	13.7931	14.9734	0.9968	0.8646	0.9256	0.6727
2	ДА	21.8000	17.5287	3.8534	0.9215	0.8828	0.9007	0.6039
3	НС	18.3908	13.7931	11.4502	0.9799	0.8754	0.9245	0.6419
4	МОВ	18.4000	13.5057	6.4958	1	0.8649	0.9274	0.3828
5	БДР	14.9000	12.0690	19.8177	0.9763	0.8933	0.9329	0.7176
6	GrB	18.3908	13.5057	19.7076	1	0.8649	0.9271	0.7891

Рисунок 2.2. Критерии качества диагностики

2.3. Агрегированный подход

2.3.1 Агрегирование базовых методов обучения

Агрегированный подход был предложен для решения задач кредитного скоринга [74] и позднее применен для диагностики технического состояния систем. В композиционных методах (бэггинг, бустинг) для построения композиции 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. При этом для

достижения наилучшего результата надо решить вопросы о том, какие методы обучения использовать, как их объединить, и как принять единое решение об исправности объекта на основе решений отдельных методов?

Воспользуемся полным перебором наборов из H базовых методов. Тогда, например, при $H = 2$ получим три набора: два базовых и один агрегированный; при $H = 3$ наборов уже 7: три базовых, три агрегированных по два базовых и один агрегированный из всех трех 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.

Перечислим эти 11 методов, встроенных в систему Matlab:

-логистическая регрессия

-дискриминантный анализ

-наивный байесовский классификатор

-метод опорных векторов

-нейронные сети

-бэггинг деревьев решений

-градиентный бустинг

-AdaBoost

-LogitBoost

-GentleBoost

-RUSBoost

Такое большое количество моделей (2047 вариантов) может привести к неоправданно большому объему вычислений. Проведенные далее численные расчеты 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.

Для формирования единого решения об исправности объекта на основе решений отдельных методов классификации, рассмотрим три варианта агрегирования результатов: по среднему значению, по медиане, и с помощью процедуры голосования.

2.3.2 Методы агрегирования

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

$$\hat{P}_{АК\,ср}(X) = \frac{\sum_{K=1}^{H} \hat{P}_K(X)}{H}, \tag{2.9}$$

где $\hat{P}_{АК\,ср}(X)$ - вероятность того, что объект исправен.

При агрегировании *по медиане* вначале надо упорядочить ряд, содержащий результаты базовых методов в наборе. При нечетном числе базовых методов вероятность того, что объект исправен:

$$\hat{P}_{АК\,мед}(X) = \hat{P}_{\frac{H+1}{2}}(X). \tag{2.10}$$

В случае четного числа базовых методов, соответствующая вероятность находится как полусумма результатов срединных значений:

$$\hat{P}_{АК\,мед}(X) = \frac{\hat{P}_{\frac{H}{2}+1}(X) + \hat{P}_{\frac{H}{2}-1}(X)}{2}. \tag{2.11}$$

Результат агрегированного метода классификации *по голосованию* представляет собой среднее значение результатов базовых методов, которые определили исправность объекта с вероятностью *p* (вероятность

исправности объекта при исходных показателях функционирования объекта X $\hat{P}_K(X) \geq \text{p}$). В противном случае вероятность того, что объект исправен, считается нулевой.

Таким образом, значения вероятностей классификации, оказавшиеся ниже, чем *p*, приравниваются к нулю, а оставшиеся - к единице, и по этим значениям строятся агрегированные модели классификации.

Стоит отметить, что структуры агрегированных классификаторов могут отличаться друг от друга из-за того, что разбивка исходных данных на обучающую и контрольную выборки производится случайным образом. Отсюда вытекает вопрос, какую структуру агрегата выбрать для принятия окончательного решения об исправности объекта.

2.3.3 Математические модели агрегированных классификаторов

Построим математическую модель агрегированного метода классификации по среднему значению для двух базовых методов (как уже отмечено, увеличение числа компонентов в агрегате больше двух не приводит к значимому повышению точности). Пусть, например, при проведении бинарной классификации для диагностики функционирования конкретного технического объекта с применением кросс-валидации наилучшими (обеспечивающими максимум *F*-критерия (2.8)) оказались байесовский классификатор и логистическая регрессия.

В соответствии с формулой (1.4) при бинарной классификации ($Y = 1$ – объект исправен, $Y = 0$ – объект неисправен) с использованием наивного байесовского классификатора вероятность того, что при заданном наборе факторов X объект исправен, равна:

$$P(Y=1|X) = \frac{P(Y=1)\prod_{j=1}^{p} P(x_j|Y=1)}{P(Y=1)\prod_{j=1}^{p} P(x_j|Y=1) + P(Y=0)\prod_{j=1}^{p} P(x_j|Y=0)},$$

где x_j – показатели функционирования объекта, $P(Y=1)$ и $P(Y=0)$ – априорные вероятности исправности и неисправности объекта соответственно, $P(x_j \mid Y=1)$, $P(x_j \mid Y=0)$ – вероятности показателей $x_1,\ldots,x_p$ в классах $Y = 1$ и $Y = 0$.

При использовании логистической регрессии для проведения бинарной классификации с целью диагностики функционирования объекта из соотношений (1.12) и (1.13) следует зависимость для расчета вероятности того, что объект исправен

$$P(Y=1 \mid X) = \frac{1}{1+\exp(-(q_0 + \sum_{j=1}^{p} q_j x_j))},$$

где q_0, q_1 … q_p – параметры логистической регрессии, определяемые, например, методом максимального правдоподобия.

Тогда математическая модель агрегированного метода классификации по среднему значению для двух базовых методов – байесовского классификатора и логистической регрессии – примет вид:

$$\hat{P}_{AKcp}(X) = \frac{\frac{1}{2} P(Y=1) \prod_{j=1}^{p} P(x_j \mid Y=1)}{P(Y=1) \prod_{j=1}^{p} P(x_j \mid Y=1) + P(Y=0) \prod_{j=1}^{p} P(x_j \mid Y=0)} +$$

$$+ \frac{1/2}{1+\exp(-(q_0 + \sum_{j=1}^{p} q_j x_j))}. \qquad (2.12)$$

Предположим теперь, что при проведении диагностики функционирования рассматриваемого объекта наилучшими по *F*-критерию оказались метод опорных векторов и один из вариантов бустинга.

Для метода опорных векторов вероятности того, что объект исправен, определяется по формуле (1.20):

$$P(Y=1|X)=\frac{1}{1+\exp(-\sum_{j=1}^{p} w_j x_j + w_0)}.$$

где w_j– параметры метода опорных векторов, определяемые с использованием метода множителей Лагранжа из зависимостей (1.19).

Для методов бустинга (1.27) соответствующая вероятность

$$P(Y=1|X)=\frac{1}{1+\exp(-\sum_{t=1}^{T} \alpha_t h_t)}.$$

где $h_t(x)$ – 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.).

Тогда математическая модель агрегированного метода классификации по среднему значению для двух базовых методов – опорных векторов и бустинга – примет вид:

$$\hat{P}_{AKcp}(X)=\frac{1}{2}\left[\frac{1}{1+\exp(-\sum_{j=1}^{p} w_j x_j + w_0)}+\frac{1}{1+\exp(-\sum_{t=1}^{T} \alpha_t h_t)}\right]. \quad (2.13)$$

Подобным образом могут быть построены и модели агрегирования для других базовых методов.

Математическая модель агрегированного метода классификации *по медиане* для *двух* базовых методов совпадает с агрегированием по среднему значению (поскольку при четном числе членов ряда медиана – среднее двух центральных членов соответствующего вариационного ряда).

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:

$$\hat{P}_{АК\,мед}(X_r) = median\left(\frac{1}{1+\exp(-(q_0 + \sum_{j=1}^{p} q_j x_j))}; \right. \tag{2.14}$$

$$\left. \frac{P(Y=1)G_1(X)}{P(Y=1)G_1(X)+P(Y=0)G_2(X)}; \frac{1}{1+\exp(-\sum_{t=1}^{T} \alpha_t h_t)} \right).$$

Здесь для расчета вероятности исправности объекта с применением дискриминантного анализа использованы соотношения (1.8) и (1.9):

$$P(Y=1|\,X) = \frac{P(Y=1)G_1(X)}{P(Y=1)G_1(X)+P(Y=0)G_2(X)},$$

Предполагается справедливость гипотезы о нормальном распределении входных данных (с математическими ожиданиями μ_1 и μ_2 и ковариационными матрицами $\sum_1$ и $\sum_2$):

$$G_1(X) = \frac{1}{(2\pi|\Sigma_1|)^{1/2}} \exp(-\frac{1}{2}(X-\mu_1)^T \Sigma_1^{-1}(X-\mu_1)),$$

$$G_2(X) = \frac{1}{(2\pi|\Sigma_2|)^{1/2}} \exp(-\frac{1}{2}(X-\mu_2)^T \Sigma_2^{-1}(X-\mu_2)).$$

2.4. Обновление модели классификатора при поступлении новой информации о показателях функционирования объекта

2.4.1. Обновление параметров модели

Показатели функционирования технического объекта, как правило, по истечении некоторого времени изменяются, при этом поступает соответствующая новая информация о значениях этих показателей в процессе эксплуатации объекта.

Для адаптации к новым условиям функционирования объекта необходимо обновить параметры классификатора по вновь поступающим данным.

Эта адаптация агрегированных классификаторов может быть осуществлена на основе корректировки параметров математических моделей, входящих в состав агрегированных классификаторов.

Для адаптации может быть применена рекуррентная псевдоградиентная процедура [60, 94]. По сравнению с другими методами эта процедура позволяет находить оптимальные оценки параметров с быстрой сходимостью (порядок сходимости $O(\frac{1}{\sqrt{n}}))$) при минимальных вычислительных затратах. Уменьшение вычислительных затрат обусловлено тем, что в предложенной процедуре не считается градиент для каждого предыдущего состояния, а только для последнего.

Оптимизация математической модели состоит в оценке вектора её параметров. Минимизируется сумма квадратов отклонений «истинной» вероятности $P(Y_r)$ принадлежности r-го наблюдения классу от оценки вероятности принадлежности $\hat{P}(X_r, \overline{\alpha})$ r-го наблюдения классу, зависящая от показателей функционирования $\overline{X}$ объекта и вектора параметров модели $\overline{\alpha}$.

$$\Omega(\overline{\alpha}) = \sum_{r=1}^{l} (P(Y_r) - \hat{P}(X_r, \overline{\alpha}))^2. \qquad (2.15)$$

Суть псевдоградиентной процедуры заключается в постоянной корректировке вектора параметров модели при поступлении информации о каждом новом наблюдении на основе формулы:

здесь $\overline{\alpha}_{r+1}$ - следующее за $\overline{\alpha}_r$ приближение оптимального вектора $\overline{\alpha}$, υ_r – коэффициенты, определяющие величину шага,

$$\nabla J(\overline{\alpha}_r) = \nabla(P(Y_r) - \hat{P}(X_r, \overline{\alpha}_r))^2 \qquad (2.17)$$

– градиент отдельного слагаемого из формулы (2.15), который является псевдоградиентом для функции $\Omega(\overline{\alpha})$.

Путем добавления к старому вектору параметров $\overline{\alpha}_r$ поправки, получаемой в результате умножения числа υ_r на псевдоградиент $\nabla J(\overline{\alpha}_r)$ обеспечивается корректировка параметров модели.

При этом важно, что процедура отслеживает меняющуюся ситуацию, поскольку текущий вектор $\overline{\alpha}_r$ постоянно на неё реагирует в соответствии с новыми значениями показателей функционирования объекта.

Рассмотрим применение псевдоградиентной процедуры обновления коэффициентов модели на примере одного из базовых классификаторов – модели бустинга (1.27). Вероятность $\hat{P}(X_r, \overline{\alpha}_r)$ находится по формуле:

$$\hat{P}(X_r, \overline{\alpha}_r) = \frac{1}{1 + \exp(-\sum_{t=1}^{T} \alpha_{rt} h_t)}. \qquad (2.18)$$

Имеем:

$$\nabla J(\overline{\alpha}_r) = \frac{d}{d\overline{\alpha}_r} (P(Y_r) - \hat{P}(X_r, \overline{\alpha}_r))^2,$$

поэтому преобразуем (2.16):

$$\bar{\alpha}_{r+1} = \bar{\alpha}_r - \upsilon_r \nabla J(\bar{\alpha}_r) = \bar{\alpha}_r - \upsilon_r \frac{\partial}{\partial \bar{\alpha}_r} (P(Y_r) - \hat{P}(X_r, \bar{\alpha}_r))^2 =$$

$$= \bar{\alpha}_r + 2\upsilon_r (P(Y_r) - \hat{P}(X_r, \bar{\alpha}_r)) \frac{\partial \hat{P}(X_r, \bar{\alpha}_r)}{\partial \bar{\alpha}_r} = \bar{\alpha}_r + \eta_r (P(Y_r) - \hat{P}(X_r, \bar{\alpha}_r)) \frac{\partial \hat{P}(X_r, \bar{\alpha}_r)}{\partial \bar{\alpha}_r},$$

где $\eta_r = 2\upsilon_r$ - новый параметр шага процедуры.

Тогда корректировка коэффициентов модели бустинга выполняется по формуле:

$$\bar{\alpha}_{r+1} = \bar{\alpha}_r + \eta_r (P(Y_r) - \hat{P}(X_r, \bar{\alpha}_r)) \frac{\partial \hat{P}(X_r, \bar{\alpha}_r)}{\partial \bar{\alpha}_r}. \qquad (2.19)$$

Для обновления параметров математической модели классификации на примере бустинга предлагается следующий способ:

- В качестве начального вектора параметров бустинга принимается оптимальный вектор этой модели, построенный на «старых» данных.

- Для классификации каждого вновь поступающего набора данных о показателях функционирования объекта применяется модель бустинга с последним вектором её коэффициентов, имеющимся на момент поступления новой информации.

- По мере поступления значений $P(Y_r)$ выполняется процедура корректировки параметров модели на основе псевдоградиента (2.19).

2.4.2. Обновление структуры классификатора

Со временем применение псевдоградиентной процедуры, оптимизирующей только параметры модели, может оказаться недостаточным: начинает «устаревать» структура модели агрегированного классификатора. В этом случае необходимо произвести обновление структуры модели.

С целью обновления структуры модели классификации могут быть предложены следующие подходы:

При выборе наилучшей структуры запоминаются и две или более модели, значение *F*-меры которых показали, соответственно, второй, третий и так далее результат (они будут «запасными»).

Одна из запасных моделей по истечении некоторого времени может быть выбрана в качестве основной, если при поступлении новых данных о состоянии объекта она покажет результаты лучше текущей основной модели. Тем самым применяемая до этого модель, в свою очередь, становится запасной.

Другой подход: на основании опыта и с учётом рекомендаций специалистов производится регулярное обновление структуры модели через определённый период (например, один раз в неделю, в месяц и т.п.).

В этом случае проводится структурно-параметрическая идентификация модели по всем данным, полученным за этот период; в случае недостатка данных могут быть использованы данные за предшествующий период.

2.5. Выводы по главе

В главе рассмотрены вопросы построения математических моделей для адекватной диагностики функционирования технического объекта.

1) Разработаны модели для диагностики технического объекта, основанные на различных подходах к агрегированию базовых методов машинного обучения.

2) Проанализированы критерии качества методов бинарной классификации; показано, что для диагностики состояния технического объекта наилучшей мерой является *F*-критерий.

3) Предложены подходы к обновлению математических моделей классификации по мере обновления поступающей информации о показателях функционирования объекта.

Практическая реализация разработанных моделей и подходов для диагностики технического объекта предполагает разработку соответствующих алгоритмов и программ. Для оценки эффективности этих алгоритмов необходимо проведение экспериментальных исследований на реальных технических объектах.

ГЛАВА 3. ЧИСЛЕННОЕ ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ДИАГНОСТИКИ ФУНКЦИОНИРОВАНИЯ ТЕХНИЧЕСКИХ ОБЪЕКТОВ

3.1. Постановка задачи и объекты исследования

В качестве объекта исследования рассматривается сложная техническая система, о которой известны априорные данные о ее состоянии (показатели функционирования и оценка исправности). По этим данным необходимо разработать математические модели и алгоритмы для прогнозирования исправности объекта при новых показателях его функционирования.

Исследование проводилось на примере трех реальных технических объектов: системы водоочистки Санкт-Петербургского водоканала, системы управлении 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 pH, 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 (хотя бы один из показателей качества очищенной питьевой воды вышел за допустимые пределы или значения двух показателей приблизились к этим пределам).

На рис. 3.1 показана схема системы водоочистки, а на рис. 3.2 – часть файла данных о показателях водоисточника и исправности системы.

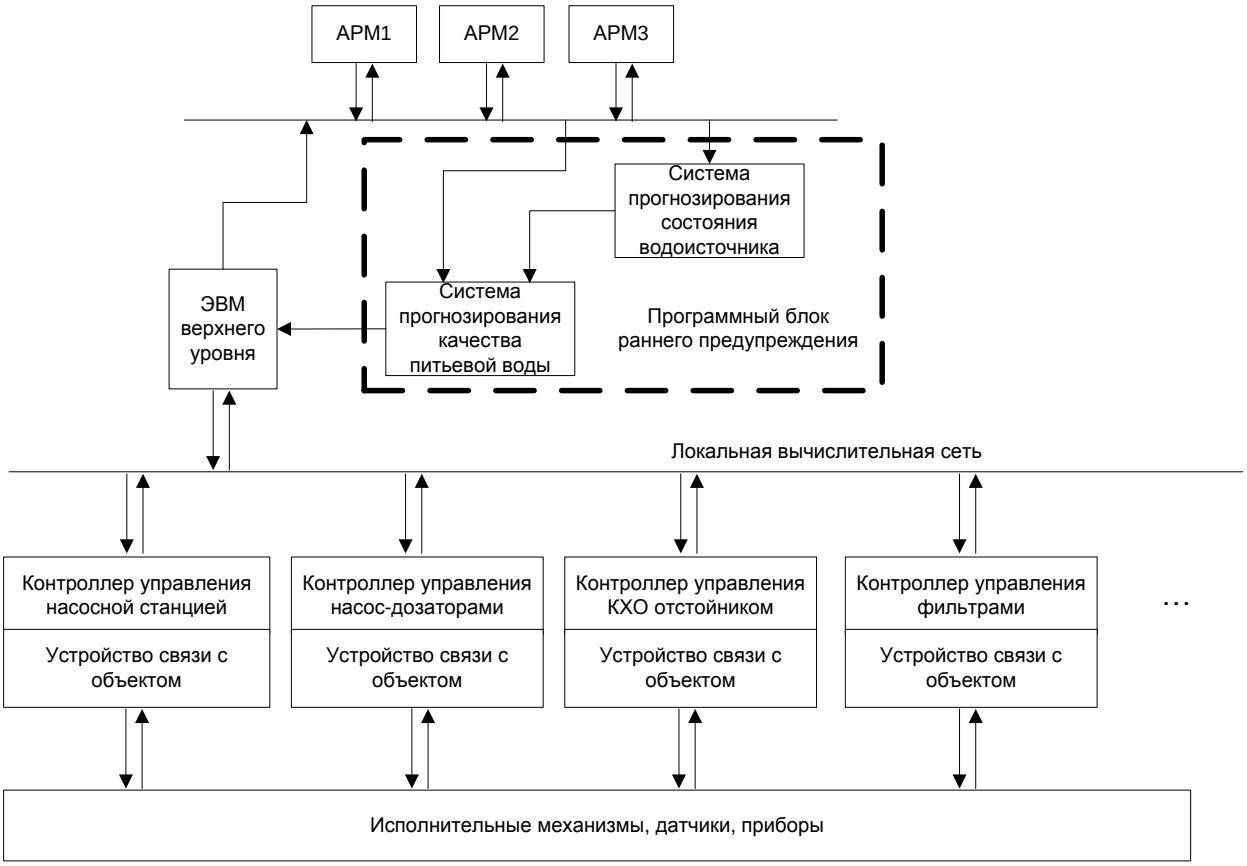


Рисунок 3.1. Схема системы водоочистки

	A	B	C	D	E	F	G	H	I
1	исправность У=1	X1	X2	X3	X4	X5	X6	X7	X8
2	1,00	1,20	36,00	13,00	7,52	0,55	7,1	7,15	0,22
3	1,00	1,10	36,00	12,00	7,62	0,57	8,60	7,15	0,22
4	1,00	0,60	41,00	9,20	7,59	0,57	8,60	7,63	0,22
5	1,00	0,40	40,00	3,10	7,41	0,58	8,60	7,54	0,20
6	1,00	0,40	37,00	5,00	7,60	0,62	8,20	7,25	0,20
7	1,00	0,30	39,00	4,20	7,54	0,55	7,80	7,44	0,20
8	1,00	0,70	40,00	2,80	7,49	0,63	8,60	7,54	0,20
9	1,00	0,60	36,00	1,50	7,55	0,55	8,40	7,15	0,20
10	1,00	0,30	35,00	1,50	7,43	0,55	7,90	7,05	0,20
11	1,00	0,20	38,00	1,70	7,50	0,60	7,90	7,35	0,20

Рисунок 3.2. Часть файла исходных данных о системе водоочистки

Задача: используя полученные данные – матрицу *X* показателей функционирования 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 водоочистки.

Исследование системы управления гидроагрегатом проводилось по данным вибромониторинга с применением вибродатчиков, расположенных на различных участках гидроагрегата. На рис. 3.3 показана система управления гидроагрегатом, на рис. 3.4 часть исходных данных о вибрациях и состоянии гидроагрегата, а на рис. 3.5 места расположения вибродатчиков.

Процесс определялся десятью показателями: вибрациями нижнего X_1 и верхнего X_3 генераторного подшипника верхнего бьефа и на правом берегу X_2, X_4, боем вала гидротурбины на нижнем бьефе X_5 и правом берегу X_6, боем вала гидрогенератора X_7, X_8, а также вибрациями крышки гидротурбины X_9,X_{10}.

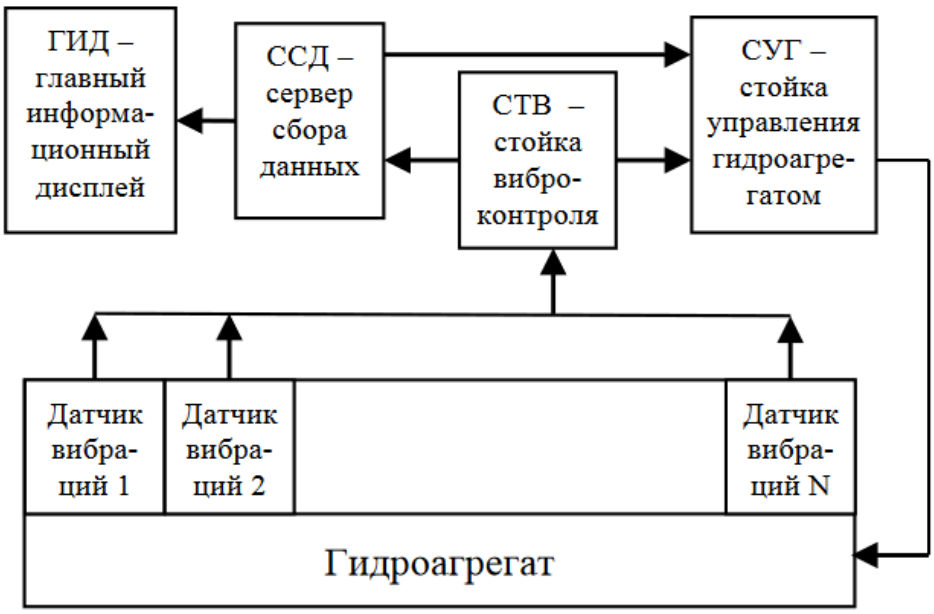


Рисунок 3. 3. Вибромониторинг в системе управления гидроагрегатом

Исходная выборка составила 1500 наблюдений, из которых 966 исправных состояний.

Задача: используя полученные данные – матрицу X показателей вибраций гидроагрегата размерностью 1500 строк и 10 столбцов и вектор-столбец ответов Y, разработать модель бинарного классификатора, которая

по вновь поступившим показаниям вибродатчиков обеспечила бы прогноз об исправности гидроагрегата.

У	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1	39,90	38,23	45,80	41,80	83,27	96,73	85,20	94,33	39,57	36,67
0	3,23	3,50	5,27	5,07	15,73	18,13	16,03	15,87	2,17	2,53
1	19,70	19,90	23,23	21,43	45,53	52,87	46,77	50,63	19,13	18,47
1	59,40	57,03	68,40	61,63	118,70	138,43	121,87	135,53	58,60	54,77
1	4,13	3,83	5,17	5,10	15,73	17,87	16,03	15,97	2,40	2,97
1	21,77	20,47	25,80	22,27	48,70	56,27	49,87	54,30	20,73	19,80
0	3,60	3,23	5,33	3,80	15,70	17,77	16,00	15,90	2,47	3,40
1	3,97	3,87	4,90	5,37	15,90	17,80	15,93	15,97	2,63	3,00
1	3,67	3,93	4,07	4,70	15,90	17,97	16,00	15,83	2,47	2,83
1	3,63	3,90	5,57	4,50	15,73	17,60	16,07	15,93	2,67	3,43
1	3,50	3,40	4,87	4,50	15,63	17,63	15,93	15,83	2,00	3,30
1	3,57	3,77	5,03	4,93	15,90	17,67	15,97	16,00	2,83	2,90
1	3,43	3,87	4,30	4,53	15,80	17,90	16,00	15,87	2,33	2,93

Рисунок 3.4. Часть файла исходных данных о вибрациях гидроагрегата

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).

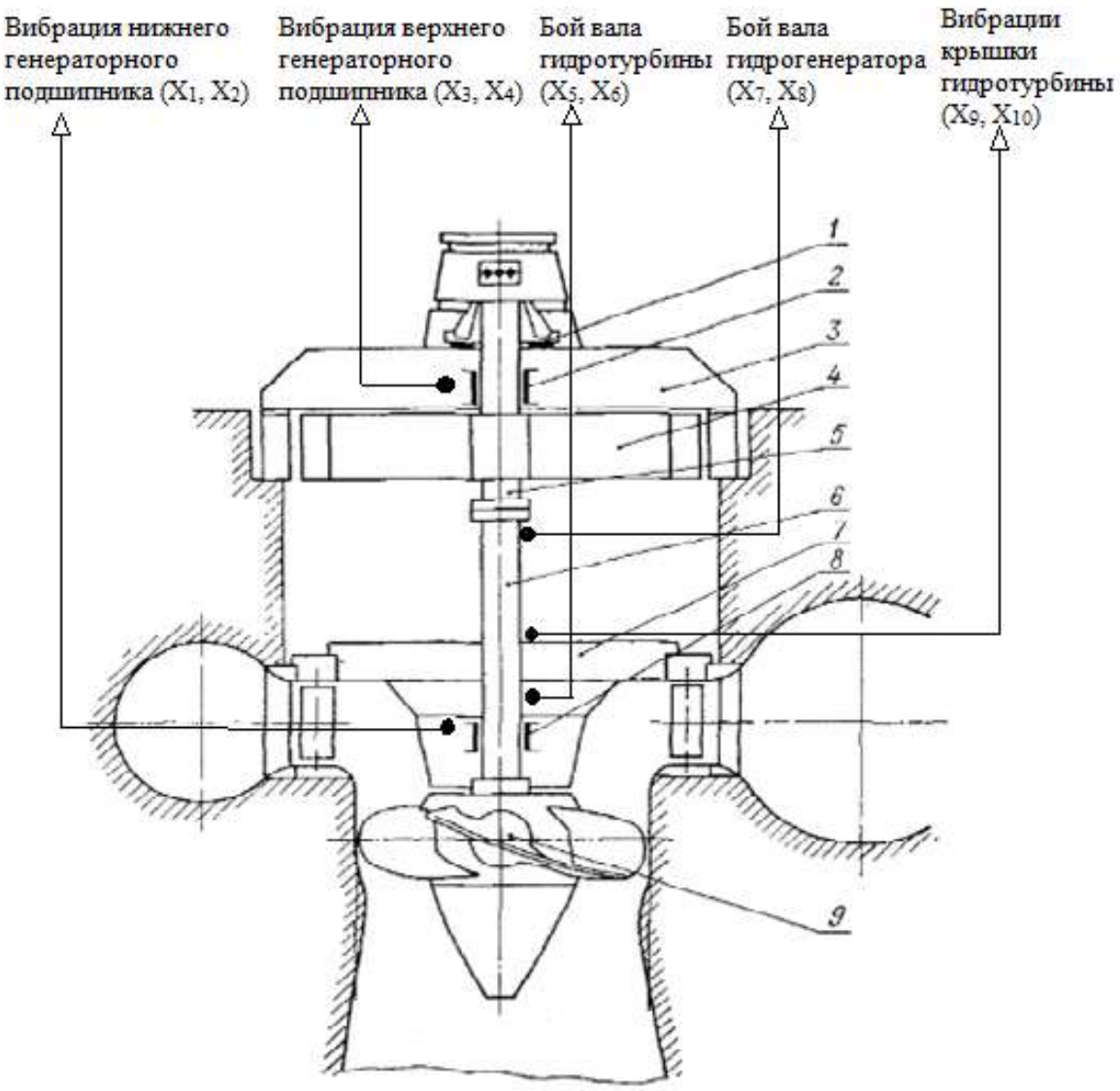


Рисунок 3.5. Схема гидроагрегата
и места установки датчиков вибрации
(1 – подпятник; 2 – подшипник направляющий гидрогенератора; 3 – крестовина гидрогенератора верхняя; 4 – ротор гидрогенератора; 5 – вал гидрогенератора; 6 – вал гидротурбины; 7 – крышка гидротурбины; 8 – подшипник направляющий гидротурбины; 9 – колесо рабочее)

На рис. 3.6. показана часть файла исходных данных о функционировании счетчиков системы водоснабжения горячей воды.

Задача: используя матрицу *X* показателей функционирования счетчиков размерностью 527 строк и 6 столбцов и вектор-столбец ответов

о наличии или отсутствии утечек в системе *Y*, разработать модель бинарного классификатора, которая по вновь поступившим показателям функционирования счетчиков даст прогноз об исправности системы водоснабжения.

У	X1	X2	X3	X4	X5	X6
1	56,83	5,75	5,39	2,46	3,55	0,016
1	57,31	5,41	5,18	2,32	2,35	0,01
0	57,67	5,21	5,09	2,43	4	0,005
1	57,9	5,25	5,12	2,48	4,17	0,005
1	66,17	5,6	5,38	2,3	2,56	0,01
1	65,68	5,94	5,72	2,49	1,96	0,01
0	66,26	7,99	6,49	6	2,52	0,084
1	64,98	7,38	6,33	2,45	2,55	0,056
1	65,73	8,24	6,78	2,26	5,69	0,081
1	65,84	8,92	6,91	2,47	2,99	0,113
1	65,86	8,82	6,86	2,45	6,76	0,111
1	65,83	8,44	6,67	2,15	1,82	0,099
1	66,04	8,18	6,46	2,31	3,18	0,096
1	66,3	7,41	6,41	2,44	3,09	0,055
1	66,31	7,92	6,35	2,35	2,41	0,087

Рисунок 3.6. Часть файла исходных данных о функционировании счетчиков системы горячего водоснабжения

Все три рассмотренные проблемы диагностики и прогнозирования состояния сложных технических систем представляют задачи бинарной классификации при наличии обучающей выборки, которые могут быть решены методами машинного обучения.

Для корректного решения поставленных задач необходимо проведение следующих исследований:

- определение влияния объема контрольной выборки на качество классификации,
- определение влияния метода отбора значимых признаков,
- выбор показателя качества классификации,
- определение наилучшего метода классификации,

- разработка программного комплекса для диагностики технического объекта.

Для проведения перечисленных испытаний необходима разработка специальной программы. Такая программа была разработана в среде Matlab.

3.2. Разработка программы для проведения испытаний «Оценка исправности технического состояния объекта с применением машинного обучения»

Практическая реализация методов машинного обучения возможна на базе библиотеки инструментов Statistics and Machine Learning Toolbox в пакете Matlab. Matlab – пакет прикладных программ для решения задач технических вычислений и одноименный язык программирования, используемый в этом пакете. Machine Learning Toolbox содержит алгоритмы и инструменты для организации, анализа и моделирования данных.

В настоящее время Matlab развивается в направлении искусственного интеллекта – совершенствуются уже разработанные технологии и методы, а также создаются новые.

Для реализации алгоритмов машинного обучения и агрегированных методов также может быть использован алгоритмический язык Python, который, как и Matlab, хорошо подходит для экспериментов и выбора наилучших алгоритмов. Главным достоинством Matlab в сравнении с Python является то, что в данной среде удобно производить исследование эффективности различных методов машинного обучения, так как к каждому из них можно легко обратиться.

При разработке на языке Python используются библиотеки общего назначения: Pandas, Scikit-learn, numPy и другие. Это привело к тому, что их интерфейс поддерживает большинство специализированных библиотек,

но, если это не так, приходится самостоятельно писать коннекторы или выбирать другие библиотеки.

Преимущества Matlab:

- Matlab является единым комплексом, таким образом можно обращаться к любым встроенным функциям,

- удобный и понятный интерфейс,

- реализовано большинство методов машинного обучения,

- множество встроенных математических и статистических функций, позволяющих упростить и ускорить исследование.

Например, при написании кода программы использовались функции:

corrcoef(*x, y*) – функция, которая рассчитывает матрицу парных коэффициентов корреляции векторов *x* и *y*.

regress(*y, x*) – функция, предназначенная для расчета точечных оценок коэффициентов линейного уравнения регрессии.

median(*X*) – функция, возвращающая значение среднего элемента одномерного массива.

nchoosek(*V, k*) – функция, возвращающая массив всех сочетаний элементов вектора *V* размером *k*.

perfcurve(*labels, scores, posclass*) – функция, возвращающая значение площади под ROC-кривой.

- удобство работы с матрицами и векторами.

К недостаткам Matlab можно отнести:

- небольшой встроенный набор инструментов среды GUIDE, которая используется для создания приложений с графическим интерфейсом пользователя,

- относительно медленный в работе, например, в циклах и ООП,

- код, написанный на Matlab, может использовать только тот, у кого установлен Matlab,

- алгоритмы являются проприетарными, таким образом невозможно увидеть код большинства используемых встроенных алгоритмов.

С учетом целей исследования была разработана программа, обеспечивающая:

- использование различных базовых методов (включая композиционные), а также построение агрегированных классификаторов,

- применение различных критериев качества классификации: доли ошибок на контрольной выборке, *F*-критерия, площади AUC под ROC-кривой и других,

- изменение объема контрольной выборки (в статье [20] показано, что, варьируя объем контрольной выборки, можно существенно повысить качество классификации),

- различные методы отбора значимых показателей,

- проведение кросс-валидации,

- прогнозирование новых состояний технического объекта.

Файл исходных данных представляет таблицу Excel (формат .xls или .xlsx), в которой в первом столбце приведены значения *y*, а в остальных *p* столбцах – значения показателей *x* функционирования объекта для каждого из *l* наблюдений. После загрузки файла данных (рис. 3.7) вводится объем контрольной выборки в процентах от общего числа наблюдений (по умолчанию 10%). Нажав кнопку *Разделить выборку*, можно просмотреть на экране обучающую и контрольную (тестовую) части выборки. Разделение исходной выборки производится случайным образом.

При необходимости пользователь проводит отбор значимых показателей с учетом их коррелированности или по значимым регрессорам (кнопка *Выполнить*). Без нажатия этой кнопки программа проводит расчет по всей совокупности показателей.

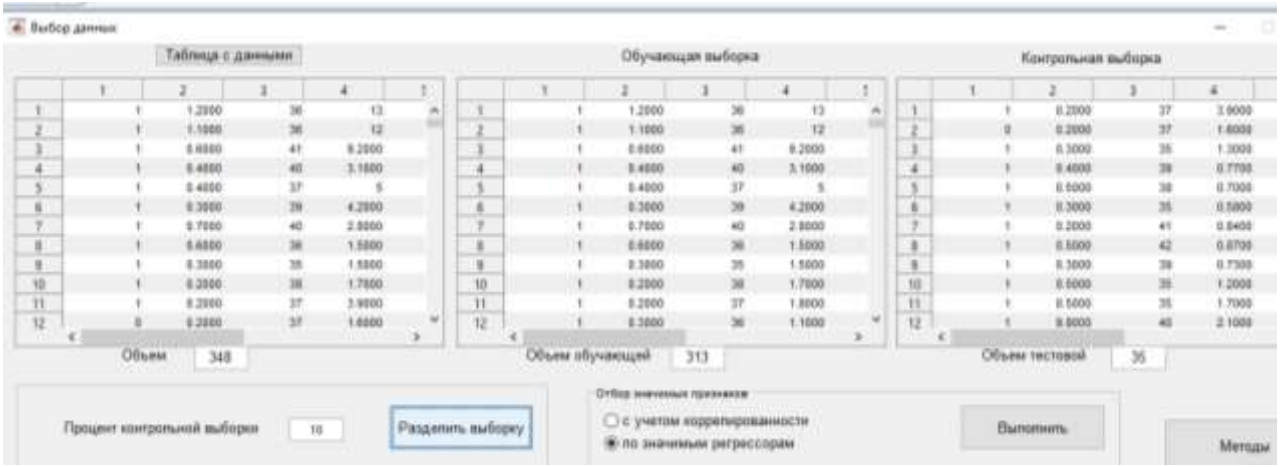
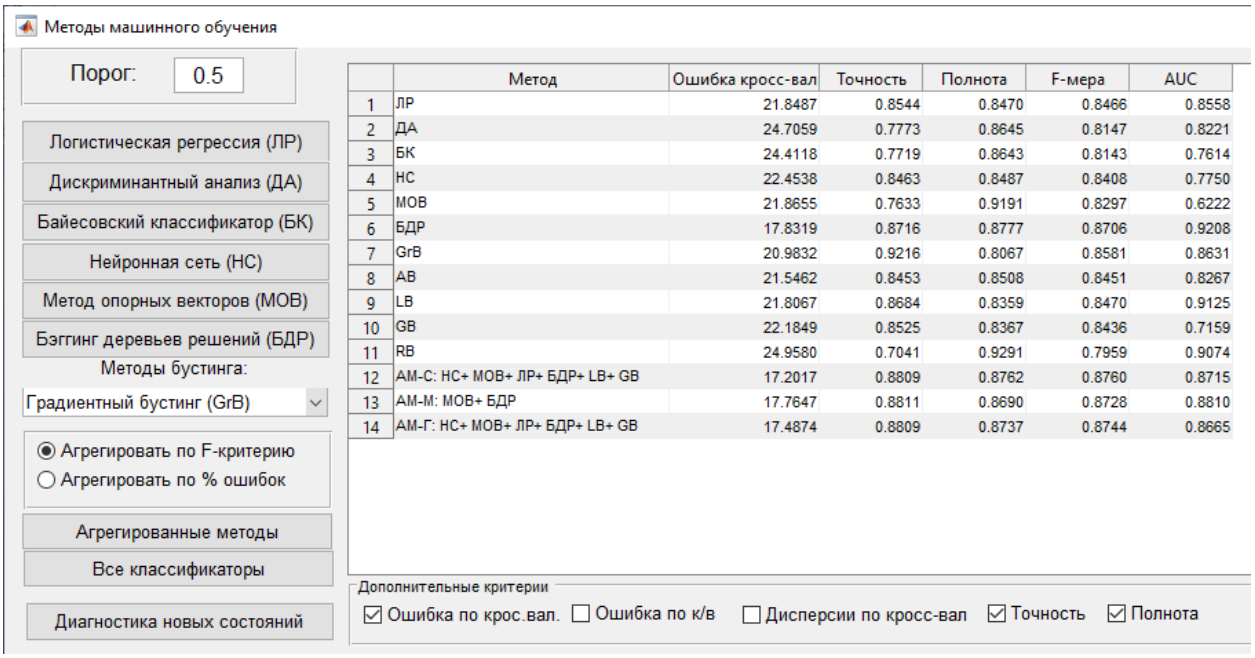


Рисунок 3.7. Загрузка файла данных и разделение выборки на обучающую и контрольную (тестовую)

После нажатия кнопки *Методы* открывается окно с перечнем используемых методов в левой части и формой для вывода результатов в правой части окна (рис. 3.8). Устанавливается порог, определяющий, при каких значениях вероятности того, что объект исправен, его следует относить к действительно исправным (по умолчанию при $p>x_0$ y= 1, в противном случае y = 0, где x_0 рассчитывается методом Неймана-Пирсона, исходя из формулы 1.22). Пользователь выбирает интересующие его методы классификации (11 базовых и три агрегированных). По мере нажатия кнопок с выбранным методом в правой части окна выводятся характеристики качества классификации.



	Метод	Ошибка кросс-вал	Точность	Полнота	F-мера	AUC
1	ЛР	21.8487	0.8544	0.8470	0.8466	0.8558
2	ДА	24.7059	0.7773	0.8645	0.8147	0.8221
3	БК	24.4118	0.7719	0.8643	0.8143	0.7614
4	НС	22.4538	0.8463	0.8487	0.8408	0.7750
5	МОВ	21.8655	0.7633	0.9191	0.8297	0.6222
6	БДР	17.8319	0.8716	0.8777	0.8706	0.9208
7	GrB	20.9832	0.9216	0.8067	0.8581	0.8631
8	AB	21.5462	0.8453	0.8508	0.8451	0.8267
9	LB	21.8067	0.8684	0.8359	0.8470	0.9125
10	GB	22.1849	0.8525	0.8367	0.8436	0.7159
11	RB	24.9580	0.7041	0.9291	0.7959	0.9074
12	АМ-С: НС+ МОВ+ ЛР+ БДР+ LB+ GB	17.2017	0.8809	0.8762	0.8760	0.8715
13	АМ-М: МОВ+ БДР	17.7647	0.8811	0.8690	0.8728	0.8810
14	АМ-Г: НС+ МОВ+ ЛР+ БДР+ LB+ GB	17.4874	0.8809	0.8737	0.8744	0.8665

Рисунок 3.8. Вывод результатов расчета

Для удобства просмотра полученных результатов пользователю предлагается выбор *Дополнительных критериев*, которые отображаются в итоговой таблице: ошибка по кросс-валидации, ошибка по одной контрольной выборке (без использования кросс-валидации), дисперсия по кросс-валидации, точность и полнота. Три столбца –используемые методы и два критерия оценки качества моделей: *F*-мера и площадь AUC под кривой ошибок в таблице отображаются по умолчанию.

Также пользователю предоставляется выбор, по какому критерию производить агрегирование: по минимуму процента ошибок на контрольной выборке или по максимальному значению *F*-критерия.

Пользователь выбирает метод машинного обучения, который в наилучшей степени соответствует поставленной задаче (например, на рис. 3.8 по максимуму *F*-критерия выбирается агрегированный метод по среднему значению АМ-С; запись АМ-С = НС + МОВ + ЛР + БДР + LB + GB означает, что в состав агрегата входят 6 базовых методов: нейронная сеть НС, метод опорных векторов МОВ, логистическая регрессия ЛР, бэггинг деревьев решений БДР и два метода бустинга: LogitBoost LB и GentleBoost GB).

На рис. 3.7 – 3.8 в качестве примера представлены некоторые данные и результаты расчета системы водоочистки. Видно, что объем контрольной выборки – 10%, т.е. кросс-валидация проводилась с разбивкой исходной выборки на 10 частей. Отбор значимых показателей в этом опыте не проводился.

Из базовых классификаторов лучшим и по критерию ошибки кросс-валидации, и по *F*-мере, и по AUC оказался бэггинг деревьев решений (БДР). В то же время все три агрегированных классификатора (АМ-С – по среднему, АМ-М – по медиане и АМ-Г – по голосованию) показали несколько лучшие результаты по критерию ошибки кросс-валидации и по *F*-мере.

На рис. 3. 9 показана блок-схема программы.

При необходимости *Диагностики новых состояний* по вновь поступившим показателям функционирования используется соответствующая кнопка и вводится файл с новыми данными (рис. 3.10).

Таким образом, разработанная программа позволяет пользователю проводить анализ исходных данных, самостоятельно подбирать параметры, способствующие улучшению классификации, а также производить прогноз нового состояния.

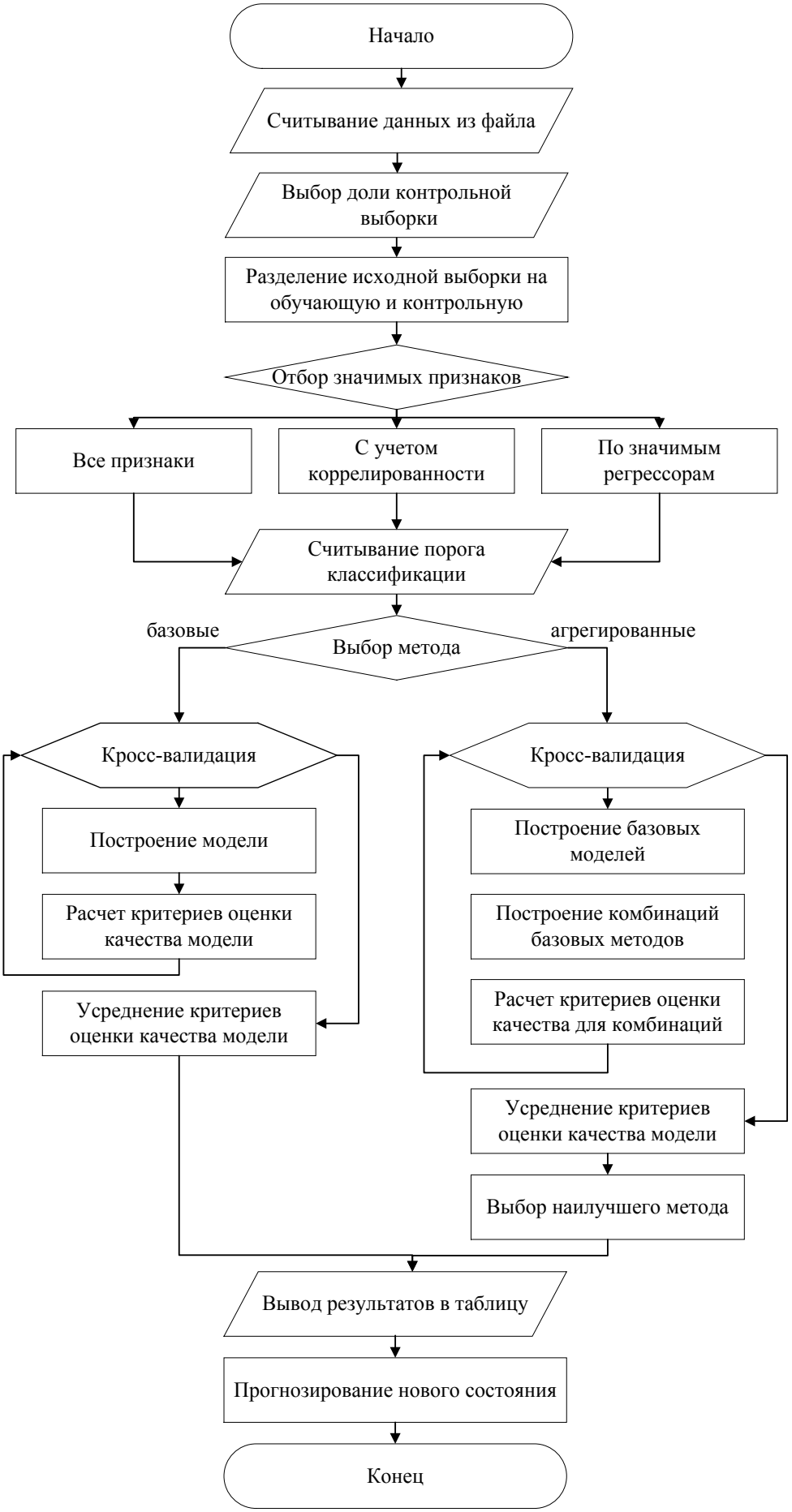


Рисунок 3.9. Блок-схема алгоритма

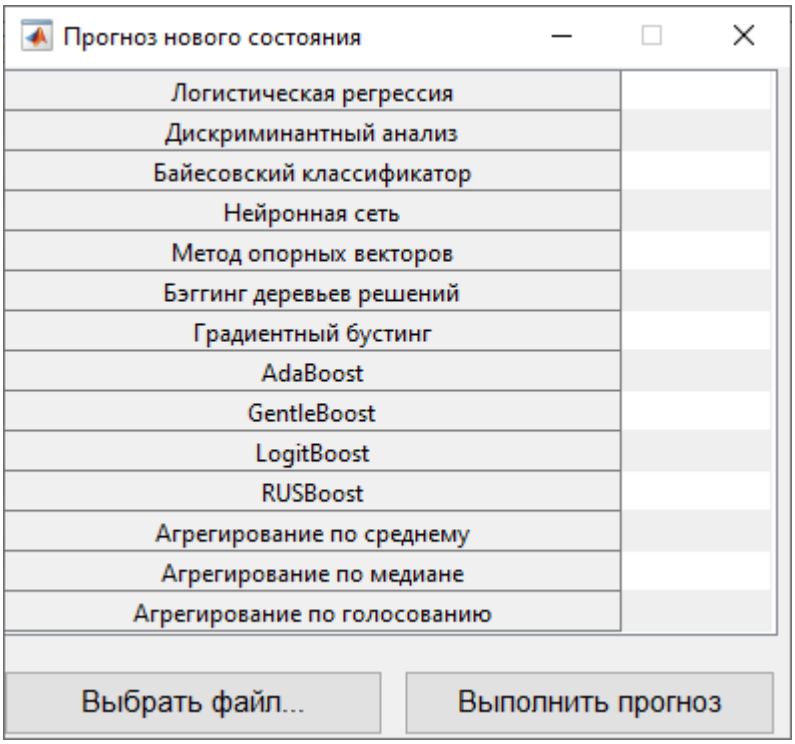


Рисунок 3.10. Окно прогнозирования нового состояния объекта

3.3. Исследование статистических свойств критериев качества диагностики

Исследование проводилось на данных системы водоочистки. Все испытания повторялись многократно (от 30 до 100 раз), так как разбиение на контрольную и обучающую части выборки происходит случайным образом, и значения критериев качества диагностики также являются случайными величинами.

В таблице 3.1 приведены усредненные значения *F*-критерия и площадь AUC под кривой ошибок для тех пяти методов машинного обучения, где эти величины оказались максимальными. Как показывают расчеты, корреляция между этими двумя показателями незначима на уровне значимости 0,05. При совпадении *F*-критерия по отдельным классификаторам значения AUC могут быть использованы для выбора лучшего метода классификации.

Видно, что наилучшие результаты показал бэггинг деревьев решений. Расхождение по *F*-критерию между наилучшим и наихудшим (0,801 для метода RUSBoost) результатами составило 8,7%, по *AUC* – 21,5%.

Таблица 3.1. Результаты базовых методов

	F-критерий	*AUC*
Нейронная сеть	0,836	0,822
Бэггинг деревьев решений	0,871	0,893
Градиентный бустинг	0,860	0,862
AdaBoost	0,852	0,854
Логистическая регрессия	0,844	0,870

Исследовалось распределение *F*-критерия и площади под кривой ошибок *AUC*. Результаты представлены на рис. 3.11-3.14, построенных в системе Statistica, где показаны гистограммы распределения и числовые характеристики этих показателей качества классификации.

Видно, что распределение обоих показателей не противоречит гипотезе о нормальности (по всем трем рассмотренным критериям нормальности: Колмогорова – Смирнова, Лиллиефорса, Шапиро - Уилка). Это обстоятельство позволяет при необходимости строить доверительный интервал для математического ожидания *F*-критерия или площади *AUC* под кривой ошибок, а также использовать стандартный алгоритм проверки гипотез.

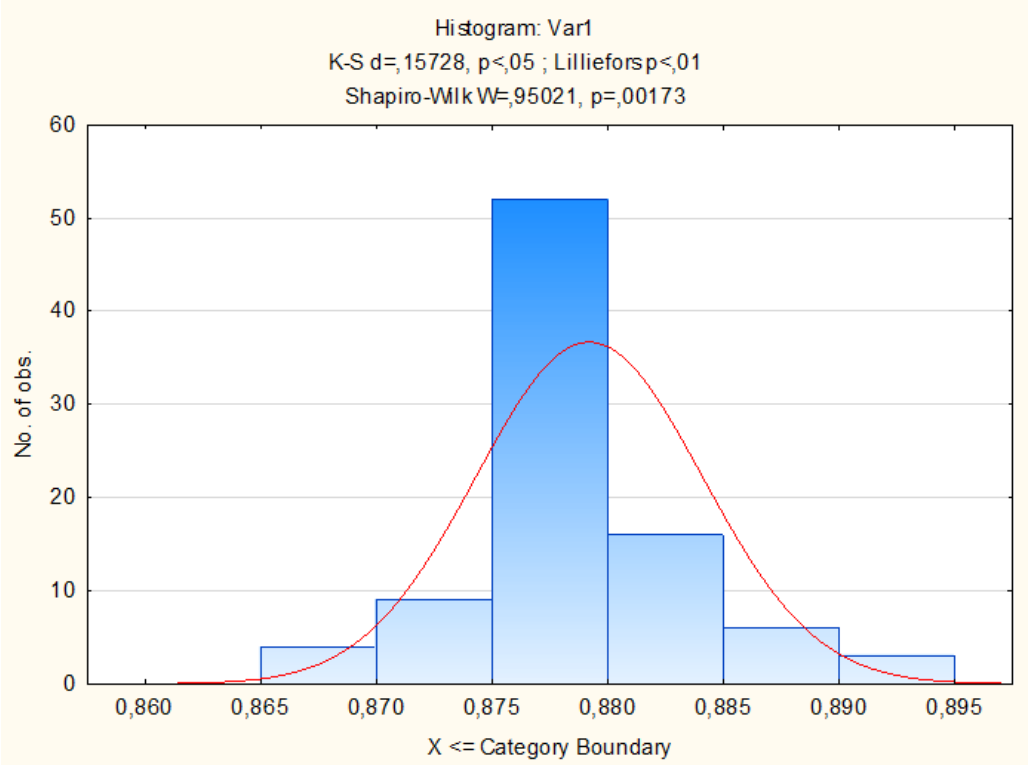


Рисунок 3.11.Гистограмма распределения *F*-критерия

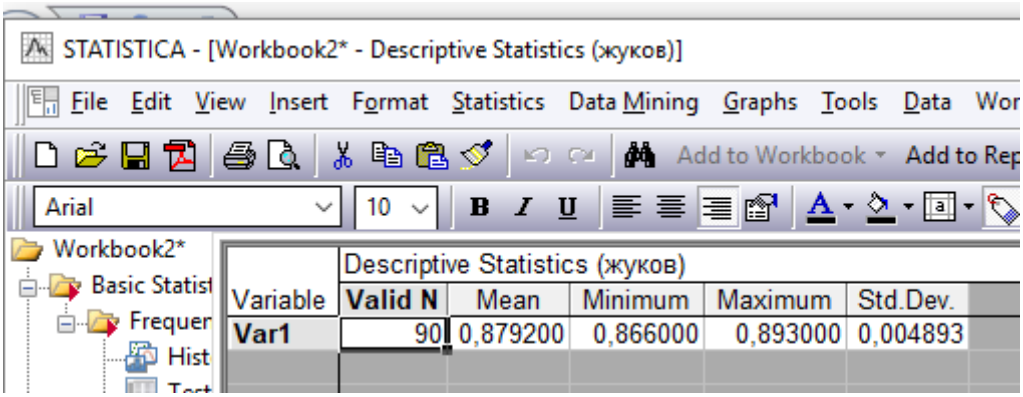


Рисунок 3.12. Числовые характеристики *F*-критерия

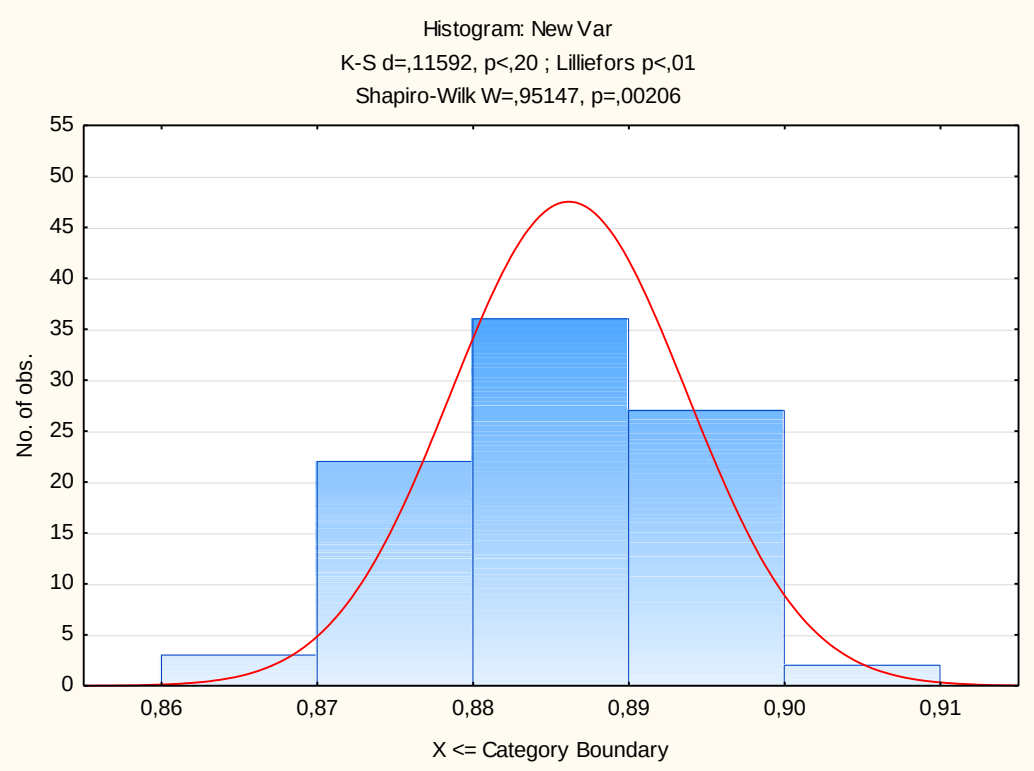


Рисунок 3.13. Гистограмма распределения *AUC*

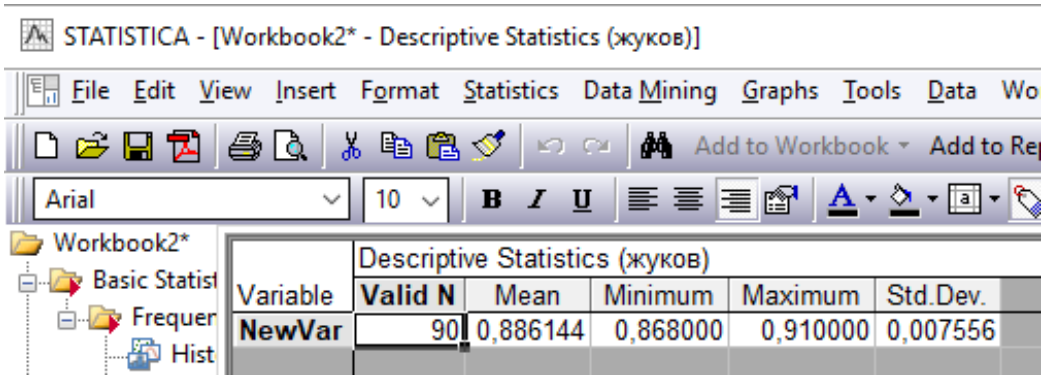


Рисунок 3.14. Числовые характеристики *AUC*

3.4. Влияние объема контрольной выборки на качество диагностики

Цель этого этапа исследования – изучить влияние доли контрольной выборки на качество алгоритмов машинного обучения при анализе исправности технического объекта.

Для оценки качества построенных алгоритмов использовалась процедура кросс-валидации, при этом объем контрольной выборки

варьировался от 5 до 25%. Каждое испытание повторялось пятикратно. В таблице 3.2 приведены усредненные результаты: значение *F*-критерия по каждому из используемых методов при каждом значении объема контрольной выборки. Графическое представление полученных результатов показано на рисунке 3.15, *а*.

Видно, что лучшие результаты (максимальное значение *F*-критерия на контрольной выборке по результатам кросс-валидации) показал бэггинг деревьев решений.

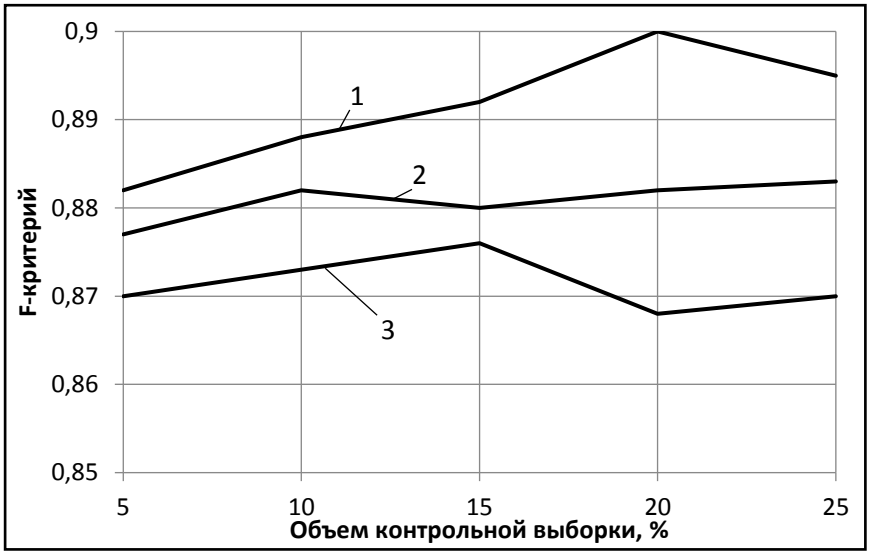
Также был произведен расчет рассеяния значений - среднеквадратичное отклонение по результатам пяти испытаний при заданном объеме контрольной выборки (рис. 3.15, *б*).

Таблица 3.2. Оценка качества диагностики

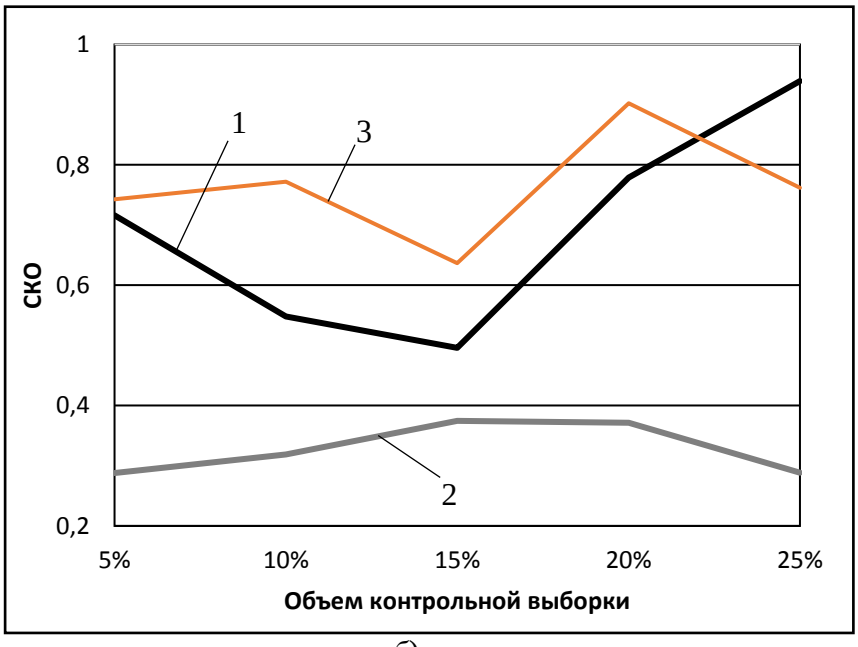
Метод	Объем контрольной выборки				
	5%	**10%**	**15%**	**20%**	**25%**
Логистическая регрессия	0,8433	0,8404	0,8442	0,8425	0,8429
Дискриминантный анализ	0,8080	0,8115	0,8128	0,8143	0,8138
Наивный байесовский классификатор	0,8119	0,8142	0,8162	0,8169	0,8175
Нейронные сети	0,8298	0,8307	0,8357	0,8419	0,8375
Метод опорных векторов	0,8195	0,8238	0,8269	0,8285	0,8266
Бэггинг деревьев решений	0,8674	0,8681	0,8702	0,8719	0,8716
Градиентный бустинг	0,8590	0,8602	0,8595	0,8615	0,8635
AdaBoost	0,8536	0,8549	0,8549	0,8499	0,8519
LogitBoost	0,8375	0,8355	0,8398	0,8408	0,8438
GentleBoost	0,8264	0,8319	0,8329	0,8279	0,8279
RUSBoost	0,8024	0,8009	0,8024	0,8033	0,8014

Исследование показало неоднозначный характер влияния объема контрольной выборки на качество диагностики: для нейронной сети

лучший результат (*F*-критерий, равный 0,8419) оказался при объеме контрольной выборки 20%, а худший (0,8298) – при объеме в 5%; для AdaBoost самое низкое значение критерия выявлено при объеме контрольной выборки 20%, а самое высокое – при 10-15%. Для метода GentleBoost лучшим оказался объем контрольной выборки 15%.



а)



б)

Рисунок 3.15. Зависимость *F*-критерия
от объема контрольной выборки
(1 – Бэггинг деревьев решений, 2 – Градиентный бустинг, 3 – AdaBoost):
а) среднее значение, б) среднеквадратичное отклонение (СКО)

Проведенное исследование показало, что изменение объема контрольной выборки влияет на различные методы машинного обучения по-разному, но за счет правильного выбора данного параметра возможно увеличить точность диагностики до 5%.

Основываясь на данном результате, можно рекомендовать производить соответствующие испытания до начала проведения технической диагностики конкретного объекта. Именно такой подход обеспечит прогноз технического состояния объекта с минимальной ошибкой. Следует заметить, что при объеме контрольной выборки 5% практически все методы показали наименьшее значение *F*-критерия.

Для данных, полученных по системе водоочистки Санкт-Петербургского водоканала – наилучшим вариантом оказался бэггинг деревьев решений при объеме контрольной выборки 20% от объема исходной выборки.

Аналогичное исследование проводилось по результатам вибромониторинга гидроагрегата и для счетчиков воды в системе горячего водоснабжения. Результаты оказались идентичными: различные методы обучения по-разному реагируют на изменение объема контрольной выборки.

3.5.Влияние способа отбора значимых показателей

Цель этого этапа исследования – изучить влияние способа отбора значимых факторов на эффективность машинного обучения при анализе исправности технического объекта.

Отбор значимых показателей по корреляционной матрице.

В табл. 3.3 показана корреляционная матрица, характеризующая степень тесноты линейной связи как между самими показателями функционирования $X_1 – X_8$, так между ними и состоянием системы Y.

Анализ этой матрицы показывает:

- показатели X_3 (мутность) и X_4 (значение pH) оказывают слабое влияние на исправность системы: соответствующие значения коэффициентов корреляции $r(X_3, Y) = -0{,}071$ и $r(X_4, Y) = -0{,}009$,

- существует практически линейная связь показателя X_1 (температуры) с X_8 (дозой флокулянта): коэффициент корреляции $r(X_1, X_8) = -0{,}964$, а также показателя X_2 (цветности) с X_7 (дозой коагулянта): коэффициент корреляции $r(X_2, X_7) = 0{,}999$. Действительно, начальная доза коагулянта определяется по цветности, а начальная доза флокулянта подбирается с учётом температуры воды.

Таблица 3.3. Корреляционная матрица

Показатель	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
X_1	1,000	– 0,545	0,226	0,220	– 0,328	0,270	– 0,561	– 0,964	– 0,566
X_2	– 0,545	1,000	– 0,057	– 0,024	0,318	0,069	0,999	0,578	0,224
X_3	0,226	– 0,057	1,000	0,190	– 0,128	0,078	– 0,061	– 0,045	– 0,071
X_4	0,220	– 0,024	0,190	1,000	– 0,097	0,273	– 0,035	– 0,189	– 0,009
X_5	– 0,328	0,318	– 0,128	– 0,097	1,000	– 0,022	0,320	0,287	0,222
X_6	0,270	0,069	0,078	0,273	– 0,022	1,000	0,062	– 0,245	– 0,144
X_7	– 0,561	0,999	– 0,061	– 0,035	0,320	0,062	1,000	0,594	0,236
X_8	– 0,964	0,578	– 0,045	– 0,189	0,287	– 0,245	0,594	1,000	0,563
Y	– 0,566	0,224	– 0,071	– 0,009	0,222	– 0,144	0,236	0,563	1,000

Отбор значимых показателей по регрессионной модели.

На рисунке 3.16 представлены результаты расчёта, полученные в системе Statistica. Видно, что наименее значимыми в линейной регрессионной модели

$$Y = b_0 + b_1 * X_1 + b_2 * X_2 + \ldots + b_8 * X_8$$

являются показатели X_3 (мутность), для которого p-значение по критерию Стьюдента составило 0,90, и X_6 (окисляемость)– его p-значение 0,88: с позиций регрессионного анализа эти факторы не оказывают значимого влияния на отклик Y (исправность системы).

При проведении машинного обучения сопоставим результаты оценки исправности системы как по всей совокупности показателей функционирования, так и при частичном или полном удалении рассмотренных показателей.

N=348	Regression Summary for Dependent Variable: Var9 (348) R= ,60223681 R?= ,36268918 Adjusted R?= ,34764939 F(8,339)=24,115 p<0,0000 Std.Error of estimate: ,36817					
	b*	Std.Err. of b*	b	Std.Err. of b	t(339)	p-value
Intercept			-9,72953	4,912543	-1,98055	0,048450
Var1	-0,29027	0,231398	-0,02006	0,015993	-1,25443	0,210551
Var2	-1,49252	1,182069	-0,14666	0,116151	-1,26263	0,207590
Var3	-0,00791	0,062264	-0,00151	0,011870	-0,12702	0,899002
Var4	0,14294	0,045897	0,62994	0,202265	3,11443	0,002000
Var5	0,08618	0,047443	1,04414	0,574785	1,81657	0,070166
Var6	-0,00682	0,044844	-0,00061	0,004029	-0,15204	0,879246
Var7	1,32890	1,198312	1,34293	1,210965	1,10898	0,268226
Var8	0,35794	0,232992	5,51432	3,589385	1,53629	0,125401

Рисунок 3.16. Результаты регрессионного анализа

Для проведения расчётов использовалась разработанная программа «Оценка исправности технического объекта с применением методов машинного обучения». Контрольная выборка формировалась случайным образом в объёме 10% от исходной выборки. Прогнозируемые по полученной модели вероятности исправного состояния объекта для контрольной выборки сравнивали с опытными значениями и рассчитывали значения F-критерия, приведенные в табл. 3.4–3.6. Эти значения и характеризуют качество диагностики: чем ближе значение к 1, тем качественнее проведена диагностика объекта.

В таблицах 3.4– 3.6 приведены значения F-критерия при расчётах с использованием различных методов машинного обучения как с участием всех факторов, так и при удалении некоторых из них.

Таблица 3.4. Результаты исследования (F-критерий)
с исключением факторов, слабо коррелированных с состоянием системы

Метод	Все факторы	Без X_3	Без X_4	Без X_3 и X_4
Логистическая регрессия	0,8404	0,8487	0,8331	0,8324
Дискриминантный анализ	0,8115	0,8160	0,8122	0,8157
Байесовский классификатор	0,8142	0,8121	0,8135	0,8084
Нейронные сети	0,8307	0,8426	0,8196	0,8167
Метод опорных векторов	0,8238	0,8290	0,8190	0,8108
Бэггинг деревьев решений	0,8681	0,8679	0,8552	0,8569
Градиентный бустинг	0,8602	0,8593	0,8609	0,8623
AdaBoost	0,8549	0,8569	0,8492	0,8584
LogitBoost	0,8355	0,8466	0,8387	0,8408
GentleBoost	0,8319	0,8350	0,8265	0,8304
RusBoost	0,8009	0,8058	0,8032	0,8062

Таблица 3.5. Результаты исследования (F-критерий)
с исключением сильно коррелированных факторов

Метод	Все факторы	Без X_7	Без X_8	Без X_7 и X_8
Логистическая регрессия	0,8404	0,8424	0,8429	0,8425
Дискриминантный анализ	0,8115	0,8114	0,8156	0,8123
Байесовский классификатор	0,8142	0,8174	0,8244	0,8261
Нейронные сети	0,8307	0,8446	0,8307	0,8388
Метод опорных векторов	0,8238	0,8257	0,8301	0,8274
Бэггинг деревьев решений	0,8681	0,8702	0,8659	0,8658
Градиентный бустинг	0,8602	0,8625	0,8633	0,8632
AdaBoost	0,8549	0,8527	0,8523	0,8565
LogitBoost	0,8355	0,8405	0,8456	0,8420
GentleBoost	0,8319	0,8302	0,8302	0,8346
RusBoost	0,8009	0,8020	0,8027	0,8045

Из таблицы 3.4 видно, что эффект от удаления рассмотренных факторов не получен: в некоторых случаях при удалении рассматриваемых показателей *F*-критерий уменьшился. Однако, при удалении фактора X_3 у восьми методов значения *F*-критерия увеличилось, хотя и незначительно.

Из таблицы 3.5 видно, что при удалении факторов X_7 и X_8, улучшение значения *F*-критерия происходит по всем методам, кроме бэггинга деревьев решений.

Из таблицы 3.6 следует, что у метода LogitBoost произошло увеличение значения *F*-критерия с 0,8355 до 0,8471 (на 1,4 %), некоторое улучшение результатов имеет место для других методов, но не для всех: таким образом, отбор значимых показателей может существенно отличаться для конкретных исследуемых объектов по различным критериям.

Таблица 3.6. Результаты исследования (*F*-критерий) с исключением факторов, незначимых по критерию Стьюдента

Метод	Все факторы	Без X_3 и X_6
Логистическая регрессия	0,8404	0,8434
Дискриминантный анализ	0,8115	0,8194
Байесовский классификатор	0,8142	0,8199
Нейронные сети	0,8307	0,8326
Метод опорных векторов	0,8238	0,8181
Бэггинг деревьев решений	0,8681	0,8658
Градиентный бустинг	0,8602	0,8642
AdaBoost	0,8549	0,8533
LogitBoost	0,8355	0,8471
GentleBoost	0,8319	0,8397
RusBoost	0,8009	0,8023

Аналогичные результаты по анализу влияния способа отбора значимых факторов на эффективность машинного обучения были получены и при анализе исправности двух других технических объектов – системы вибромониторинга гидроагрегата и счетчика системы горячего водоснабжения.

На рис. 3.17 приведены результаты испытаний для системы водоочистки: а) с учетом всех показателей, б) с исключением коррелированных факторов, в) с исключением факторов, незначимых по регрессионной модели.

	Метод	Ошибка по крос-вал.	F-критерий	AUC
1	ЛР	21.5487	0.8467	0.8179
2	ДА	24.6957	0.8151	0.7847
3	БК	23.8344	0.8217	0.7738
4	НС	20.9358	0.8599	0.8720
5	МОВ	23.5487	0.8163	0.8547
6	БДР	18.6874	0.8686	0.8606
7	GrB	21.2878	0.8577	0.8181
8	AB	20.4141	0.8591	0.8472
9	LB	21.2671	0.8482	0.8806
10	GB	23.0021	0.8423	0.7938
11	RB	23.2505	0.8115	0.9195
12	АМ-С: НС+ GrB+ БДР	16.9275	0.8826	0.8764
13	АМ-М: НС+ БДР	16.9275	0.8817	0.8785
14	АМ-Г: НС+ GrB+ БДР+ GB	16.9275	0.8820	0.8625

а)

	Метод	Ошибка по крос-вал.	F-критерий	AUC
1	ЛР	22.1201	0.8447	0.8491
2	ДА	24.7205	0.8144	0.7485
3	БК	24.1449	0.8196	0.6804
4	НС	22.3975	0.8325	0.8294
5	МОВ	22.1615	0.8274	0.7630
6	БДР	18.3892	0.8687	0.8943
7	GrB	20.1118	0.8667	0.8904
8	AB	18.0911	0.8735	0.8037
9	LB	25.0186	0.8245	0.7708
10	GB	24.1615	0.8291	0.7972
11	RB	24.4306	0.7995	0.8676
12	АМ-С: МОВ+ БДР	16.3768	0.8861	0.8841
13	АМ-М: МОВ+ GrB+ БДР+ RB	15.2257	0.8920	0.8840
14	АМ-Г: МОВ+ БДР	16.3768	0.8861	0.8848

б)

	Метод	Ошибка по крос-вал.	F-критерий	AUC
1	ЛР	23.0062	0.8370	0.8787
2	ДА	25.2795	0.8116	0.7466
3	БК	24.7205	0.8158	0.7391
4	НС	25.2671	0.8155	0.8325
5	МОВ	22.3851	0.8269	0.9048
6	БДР	20.6874	0.8560	0.8790
7	GrB	20.9938	0.8608	0.9043
8	AB	22.9814	0.8401	0.7716
9	LB	23.8427	0.8318	0.7954
10	GB	27.5611	0.8047	0.8358
11	RB	23.2878	0.8102	0.8328
12	АМ-С: МОВ+ GrB+ БДР	16.6418	0.8817	0.8814
13	АМ-М: GrB+ БДР	16.6501	0.8817	0.8785
14	АМ-Г: GrB+ БДР+ AB	16.6460	0.8811	0.8782

в)

Рисунок 3.17. Результаты расчета с различными вариантами отбора показателей функционирования системы водоочистки

3.6.Применение агрегированного подхода к диагностике реальных объектов

3.6.1.Система водоочистки

Был произведен расчет значения *F*-критерия для всех базовых методов при объеме контрольной выборки 10%. Все испытания повторялись 30 раз. Полученные результаты представлены в таблице 3.7.

По полученным результатам, можно сделать вывод, что наилучшим методом для исходных данных оказался бэггинг деревьев решений, для которого значение *F*-критерия оказалось 0,8681.

Из таблицы 3.8 видно, что по значению *F*-критерия метод агрегирования по медиане оказался лучше остальных. Если сравнивать агрегированные методы с базовыми, то видно, что у агрегированных результаты лучше. Так наилучший базовый метод примерно на 1,5% хуже, чем агрегированный метод по медиане.

Таблица 3.7. Базовые методы

Метод	*F*-критерий
Логистическая регрессия	0,840
Дискриминантный анализ	0,812
Байесовский классификатор	0,814
Нейронные сети	0,831
Метод опорных векторов	0,824
Бэггинг деревьев решений	0,868
Градиентный бустинг	0,860
AdaBoost	0,855
LogitBoost	0,8355
GentleBoost	0,8319
RusBoost	0,8009

Таблица 3.8. Агрегированные методы

Метод	*F*-критерий
По среднему	0,881
По медиане	0,882
По голосованию	0,879

3.6.2. Вибромониторинг гидроагрегата

Из таблицы 3.9 можно сделать вывод, что наилучшими методами при вибромониторинге оказались методы агрегирования (результаты всех трех методов агрегирования одинаковы), но улучшение относительно остальных базовых методов незначительно.

Таблица 3.9. Методы машинного обучения

Метод	F-критерий
Логистическая регрессия	0,9522
Дискриминантный анализ	0,9367
Байесовский классификатор	0,9361
Нейронные сети	0,9449
Метод опорных векторов	0,9374
Бэггинг деревьев решений	0,9946
Градиентный бустинг	0,9493
AdaBoost	0,9980
LogitBoost	0,9981
GentleBoost	0,9979
RusBoost	0,9371
АМ-С	0,9984
АМ-М	0,9984
АМ-Г	0,9984

3.6.3. Счетчики горячего водоснабжения

По полученным результатам (таблица 3.10) можно сделать вывод, что у агрегированных методов по среднему значению и по медиане 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%.

Таблица 3.10. Методы машинного обучения

Метод	*F*-критерий
Логистическая регрессия	0,8798
Дискриминантный анализ	0,8836
Байесовский классификатор	0,8765
Нейронные сети	0,8977
Метод опорных векторов	0,8857
Бэггинг деревьев решений	0,9957
Градиентный бустинг	0,9646
AdaBoost	0,9947
LogitBoost	0,9949
GentleBoost	0,9957
RusBoost	0,9855
АМ-С	0,9966
АМ-М	0,9966
АМ-Г	0,9957

3.6.4. К вопросу о выборе метода агрегирования и структуры агрегата

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 окончательного решения об исправности объекта.

Испытания на примере системы водоочистки проводились при одном и том же объеме контрольной выборки с использованием всех показателей функционирования. В таблице 3.11 представлены соответствующие результаты по *F*-критерию для пяти вариантов каждого типа агрегирования.

Таблица 3.11. *F*-критерий при агрегировании

Структура агрегата	*F*-критерий
Агрегирование по среднему значению	
GrB+БДР+AB	0,891
GrB+ БДР	0,889
БДР+AB	0,889
MOB+БДР+AB+LB	0,889
MOB+БДР	0,879
Агрегирование по медиане	
ДА+MOB+GrB+БДР+AB+GB+RB	0,892
MOB+БДР	0,881
MOB+БДР+AB+RB	0,891
MOB+БДР+LB	0,888
GrB+БДР	0,887
Агрегирование по голосованию	
НС+MOB+БДР+AB+RB	0,887
GrB+БДР+AB	0,889
БДР+AB	0,889
GrB+БДР	0,885
MOB+GrB+БДР+LB+GB+RB	0,887

Представляет интерес проанализировать, какие из базовых методов чаще всего входят в состав агрегированных. Распределение методов показано на рис. 3.18. Заметим еще раз, что полученные результаты описывают конкретный процесс - функционирование системы водоочистки, особенностью которого в данном случае является слишком малое для машинного обучения число опытов. Отметим лишь, что бэггинг деревьев решений вошел во все агрегированные классификаторы для всех трех исследованных объектов.

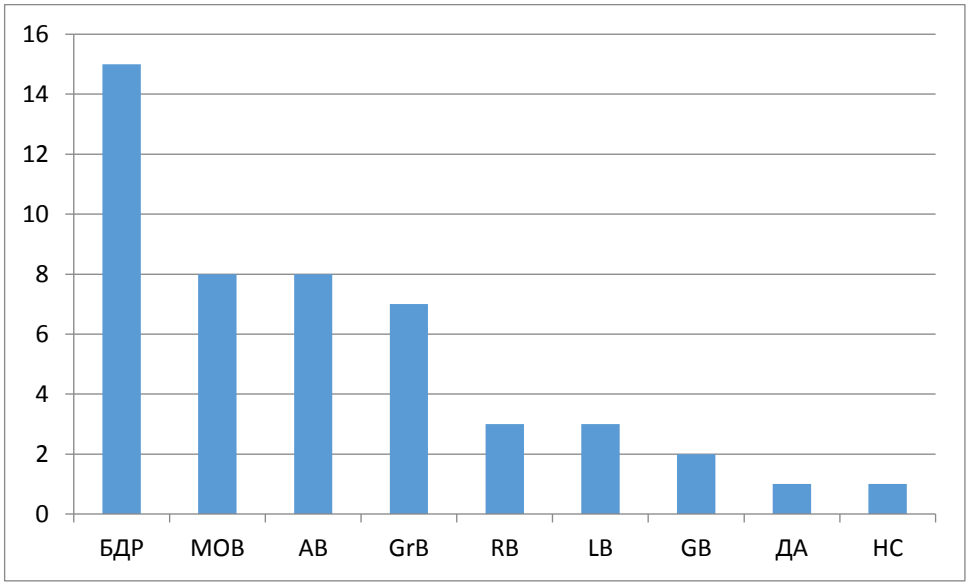


Рисунок 3.18. Распределение базовых методов при формировании агрегированных классификаторов для системы водоочистки

3.7.Проверка статистических гипотез

Нормальность распределения показателей качества бинарной классификации, показанная в пункте 3.3, позволяет использовать стандартный подход для проверки гипотезы о том, что агрегирование действительно приводит в рассмотренном примере к повышению качества диагностики.

Проверялась нулевая гипотеза о равенстве средних значений *F*-критерия при агрегировании и при применении базовых методов классификации (сравнивались данные с бэггингом деревьев решений, как показавшим лучший результат из базовых методов). Как альтернативная, рассматривалась гипотеза о превышении среднего значения при агрегировании.

Таблица 3.12. Проверка гипотезы о равенстве средних для двух методов обучения БДР и АМС: двухвыборочный тест Фишера для дисперсии

	АМС	*БДР*
Среднее	0,9027	0,8937
Дисперсия	0,59E-5	1,455E-5
Наблюдения	30	30
Число степеней свободы	29	29
Выборочное значение статистики Фишера	1,09	
Квантиль распределения Фишера (критическое значение)	1,86	

Таблица 3.13. Двухвыборочный *t*-тест с одинаковыми дисперсиями

	АМС	*БДР*
Среднее	0,9027	0,8937
Дисперсия	1,59E-5	1,45E-5
Наблюдения	30	30
Объединенная дисперсия	1,5E-05	
Гипотетическая разность средних	0	
Число степеней свободы	58	
Выборочное значение *t*-статистики	8,93	
t критическое одностороннее	1,67	
t критическое двухстороннее	2	

Вначале (таблица 3.12) сравнивались дисперсии двух выборок по критерию Фишера (разница оказалась статистически незначима: выборочное значение статистики Фишера 1,09 оказалось меньше критического 1,86 и попало в область принятия решения о равенстве

дисперсий). Затем по критерию Стьюдента с одинаковыми дисперсиями (таблица 3.13) был сделан вывод о том, что нулевая гипотеза отвергается (выборочное значение статистики Стьюдента 8,93 превысило критическое как по одностороннему, так и по двустороннему критериям, и попало в критическую область): среднее значение *F*-критерия при агрегировании выше, чем при применении базовых классификаторов.

Таким образом, агрегирование значимо повышает качество классификации. Такой же вывод был получен и для двух других объектов исследования.

Как уже отмечено, значения *F*-критерия различаются при различных агрегированных методах не слишком существенно. Проверим гипотезу о том, что увеличение в структуре агрегата числа базовых классификаторов больше двух несущественно влияет на значения *F*-критерия. Этот факт требуют проверки для каждого конкретного объекта.

Поскольку контрольная выборка формируется случайным образом, результаты классификации при повторении испытаний каждый раз несколько отличаются друг от друга. В частности, получаются и разные агрегаты, максимизирующие *F*-меру.

В таблице 3.14 частично приведены результаты испытаний системы водоочистки: отобраны десять первых испытаний, в которых агрегат включает два базовых компонента (правая часть таблицы) и такое же число испытаний с б*о*льшим количеством компонент (левая часть таблицы: от трех до пяти компонент).

Анализируя полученные данные, видим, что во все агрегаты вошел, как компонент, бэггинг деревьев решений (этот классификатор оказался лучшим из базовых методов для рассматриваемого объекта).

Таблица 3.14. Структуры классификаторов и *F*-мера

Структура классификатора	*F*-мера (F_1)	Структура классификатора	*F*-мера (F_2)
АМ-С: GrB+ БДР+ LB	0,9077	АМ-С: БДР+ GB	0,9061
АМ-С: НС+ GrB+ БДР+ LB+ GB	0,9028	АМ-С: БДР+ LB	0,9016
АМ-С: ЛР+ БДР+ AB	0,9067	АМ-С: БДР+ AB	0,9040
АМ-С: МОВ+ ЛР+ БДР+ LB+ GB	0,9015	АМ-М: БДР+ GB	0,9061
АМ-М: GrB+ БДР+ LB	0,9049	АМ-М: БДР+ AB	0,9056
АМ-М: ЛР+ БДР+ GB+ RB	0,8960	АМ-М: БДР+ LB	0,9075
АМ-М: ЛР+ БДР+ LB+ RB	0,9037	АМ-Г: БДР+ GB	0,9061
АМ-Г: GrB+ БДР+ LB	0,9044	АМ-Г: БДР+ LB	0,9004
АМ-Г: ЛР+ БДР+ AB	0,9070	АМ-Г: БДР+ AB	0,9040
АМ-Г: МОВ+ GrB+ БДР	0,8962	АМ-Г: БДР+ LB	0,9064

Обозначения: АМ-С, АМ-М, АМ-Г – агрегирование по среднему, по медиане и голосованию соответственно, GrB – градиентный бустинг, БДР – бэггинг деревьев решений, AB – AdaBoost, МОВ – метод опорных векторов, LB – LogitBoost, ДА – дискриминантный анализ, RB – RUSBoost, GB – GentleBoost, ЛР – логистическая регрессия, НС – нейронная сеть.

Значения *F*-меры как по первому, так и по второму столбцу не противоречат предположению о нормальности распределения этой характеристики. Это предположение проверялось по критерию Шапиро-Уилка (для малых выборок).

Проверим гипотезу о равенстве значений *F*-меры в левой и правой частях таблицы 3.14. Вначале проверяется гипотеза о равенстве дисперсий по критерию Фишера (таблица 3.15). Видим, что эта гипотеза отвергается: выборочное значение критерия 3,34 больше критического 3,18.

Поэтому проверяем гипотезу о равенстве средних значений *F*-меры, используя двухвыборочный тест Стьюдента с различными дисперсиями (таблица 3.16). Выборочное значение статистики Стьюдента – 1,13 по модулю меньше критического значения (как для одностороннего, так и двухстороннего критериев), что подтверждает выдвинутую гипотезу: значение *F*-меры при двух компонентах агрегированного классификатора отличается от агрегата с большим числом компонент незначимо.

Таблица 3.15.Двухвыборочный тест Фишера для дисперсии

	F_1	F_2
Среднее	0,9031	0,9048
Дисперсия	2E-05	5,15956E-06
Наблюдения	10	10
Число степеней свободы	9	9
Выборочное значение статистики Фишера	3,34	
Квантиль распределения Фишера (критическое значение)	3,18	

Таблица 3.16. Двухвыборочный *t*-тест с различными дисперсиями

	F_1	F_2
Среднее	0,9031	0,9048
Дисперсия	1,73E-05	5,15956E-06
Наблюдения	10	10
Гипотетическая разность средних	0	
Число степеней свободы	14	
Выборочное значение *t*-статистики	-1,13	
t критическое одностороннее	1,76	
t критическое двухстороннее	2,14	

Заметим, что использование агрегированного классификатора для системы водоочистки в наилучшем варианте (агрегирование по среднему значению, сочетание градиентного бустинга, бэггинга деревьев решений и метода LogitBoost) привело к повышению *F*-меры с значения 0,8578 до 0,9077 (на 5,8 %).

Подобный результат получен при построении агрегированных классификаторов для оценки вибростойкости гидроагрегата: значение *F*-меры повысилось на 3,1 %: с 0,8923 до 0,9204 (агрегирование по медиане, сочетание градиентного бустинга и логистической регрессии).

На двух рассмотренных примерах реальных технических объектов были проведены проверки гипотез о том, что, во-первых, агрегирование значимо увеличивает *F*-меру, во-вторых, увеличение числа компонент агрегата больше двух не дает существенного улучшения качества диагностики. Эти гипотезы не противоречат результатам опытов.

Отсюда вытекает важный вывод о возможности резкого сокращения времени на вычисления. Вместо перебора всех вариантов агрегирования для поиска максимального значения *F*-критерия (при трех методах агрегирования и 11 базовых методах классификации, использованных в пакете Matlab, $3*(2^{11} - 1) = 6141$ вариант); достаточно перебрать только варианты, включающие по два базовых метода (3*11!/2!9! = 165).

Можно учесть и еще одно обстоятельство. Во всех опытах в состав агрегата вошел лучший из базовых методов – бэггинг деревьев решений. Учет этого факта позволяет сократить количество перебираемых вариантов до 30.

3.8.Выводы по главе: предлагаемая методика диагностики

Проведенные численные исследования влияния различных факторов на эффективность машинного обучения при диагностике технических объектов с использованием специально разработанной программы

«Оценка исправности технического состояния объекта с применением машинного обучения», реализованной в среде Matlab, позволили установить следующее.

1.Такие факторы, как доля контрольной выборки в общем объеме исходных данных (или количество блоков разбиения выборки при кросс-валидации), а также метод отбора значимых показателей оказывают неоднозначное влияние на качество классификации: для каждого конкретного технического объекта необходимо оценивать эти факторы индивидуально.

2.Учитывая несбалансированность классов при исследовании функционирования технических объектов (количество наблюдений с исправными объектами, как правило, значительно больше, чем с неисправными), 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, при совпадении значений этой меры для различных методов, можно дополнительно использовать в качестве критерия площадь под кривой ошибок.

3.Агрегированные классификаторы в исследованных примерах оказались значимо лучшими, чем базовые классификаторы. Как показали испытания для агрегированных методов достаточно использовать два, максимум три базовых метода, которые будут обеспечивать достаточную точность классификации. Лучшими считаются модели тех методов, у которых значения *F*-критерия оказалось выше, чем у остальных. Таким образом, сокращается время выполнения программы, за счет того, что не нужно перебирать все возможные комбинации 11 методов ($3*(2^{11} - 1) = 6141$), достаточно перебрать 30 вариантов.

4.Для практического применения необходима разработка специального программного комплекса для автоматизированной диагностики технических объектов, который:

-обеспечивает ввод исходных данных заданного вида,

-оценивает необходимый объем контрольной выборки (количество блоков разбиения заданной выборки при кросс-валидации), обеспечивающий наибольшую эффективность классификации,

-выбирает значимые показатели функционирования объекта и оценивает их влияние на качество диагностики,

-способен производить построение моделей бинарной классификации как базовых, так и агрегированных методов, самостоятельно выбирая наилучшие из них,

-обеспечивает прогноз нового состояния по введенным новым значениям показателей функционирования объекта.

ГЛАВА 4. РАЗРАБОТКА АЛГОРИТМОВ И КОМПЛЕКСА ПРОГРАММ «ДИАГНОСТИКА СОСТОЯНИЯ ТЕХНИЧЕСКОГО ОБЪЕКТА С ИСПОЛЬЗОВАНИЕМ АГРЕГИРОВАННЫХ КЛАССИФИКАТОРОВ»

4.1. Структура программного комплекса

Цель разработки – создание алгоритма и комплекса программ для диагностики функционирования технического объекта, позволяющих в автоматическом режиме построить наилучшую модель для прогнозирования состояния технического 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.

С учетом целей исследования был разработан комплекс программ, включающий:

1) Блок подготовки исходных данных.

Реализация считывания исходных данных:

```
[filename, pathname]=uigetfile('*.xlsx'); % Получение файла с исходными данными
a=xlsread(filename); % считывание данных
set(handles.uitable1,'data',a); % вывод исходной выборки на форму
n=size(a,1); % Размер исходной выборки
n=num2str(n); % Преобразование размера в формат строки
set(handles.edit1,'string',n); % Вывод размера исходных данных
a = get(handles.uitable1, 'Data');
Y = a(:, 1); % Запоминаем вектор-столбец ответов
N=size(a,2);
X = a(:, 2 : n); % Запоминаем матрицу параметров
```

2) Блок построения моделей базовых методов.

В данном блоке происходит построение моделей для каждого базового метода, а также вычисляются их значения критериев качества.

Пример реализации расчета *F*-критерия для метода опорных векторов:

```
globalkk; % Значение порога
SVModel = fitcsvm(xtrain,ytrain,'Standardize',true); % Обучение модели опорных векторов
[SVMModel,ScoreParameters] = predict(SVModel,xtest); % Прогноз по значениям контрольной выборки
p = ScoreParameters(:,2); % Считываются вероятности принадлежности к классам
y1 = (p>= kk); % Сравнение полученных вероятностей с порогом
[~, precision, recall] = property(ytest, y1); % Расчет критериев качества: полнота и точность
F = 2 * precision * recall / (precision + recall); % Расчет F-критерия
```

Пример реализации расчета критериев качества классификации по кросс-валидации для метода Градиентный бустинг:

```
fork = 1 : K% начало кросс-валидации
trainIdx = cv.training(k); % индексы массива обучающей выборки
testIdx = cv.test(k); % индексы массива обучающей выборки
brtModel = brtTrain(X(trainIdx, :), Y(trainIdx, :), 2, 25, 0.1); % модель Градиентного бустинга
s = sum(testIdx); % объем контрольной выборки
tmpX = X(testIdx, :); % тестовая выборка
p = zeros(s, 1); % Заполнение нулевыми значениями вектор для прогнозированных значений
forj = 1:s % Перебор по элемент контрольной выборки
p(j) = brtTest(tmpX(j, :), brtModel); % прогнозирование для контрольной выборки
end;
```

```
BRT = brtModel; % запоминаем модель для использования в агрегированных классификаторах
y1 = (p>= kk); % сравниваем с порогом
mse(k) = mean(abs(Y(testIdx) - y1)) * 100; % расчет процента ошибки на контрольной выборке
[accuracy(k), precision(k), recall(k)] = property(Y(testIdx), y1);%вызов функции для расчета точности, полноты
F(k) = 2 * precision(k) * recall(k) / (precision(k) + recall(k));% расчет F-критерия
yt = Y(testIdx); % Выбираем из исходных данных значения для контрольной выборки
FPR_nn, TPR_nn, ~, AUC(k)] = perfcurve(yt, p, '1');% расчет AUC
end;
if (sizeUI> 1) % Проверяем пустая ли таблица для вывода результата
data = get(handles.uitable8, 'Data');  % Записываются значения в таблицу
name = (data(1 : sizeUI - 1, 1));
end; % Конец кросс-валидации
str = ['GrB']; % значение столбца «Метод»
name(sizeUI, 1) = cellstr(str);
%  расчет средних значений критериев оценки качества модели логистической регрессии
ac = mean(accuracy); %
pr = mean(precision); % Точность
rec = mean(recall); % Полнота
Fm = mean(F); % F-критерий
a = mean(AUC(k)); % AUC-ROC
```

Пример реализации расчета критериев качества классификации по кросс-валидации для метода Логистическая регрессия:

```
for k=1:K% начало кросс-валидации
trainIdx = cv.training(k); % индексы массива обучающей выборки
testIdx = cv.test(k); % индексы массива обучающей выборки
mdl = GeneralizedLinearModel.fit(X(trainIdx,:), Y(trainIdx,:), 'Distribution', 'binomial'); % расчет модель логистической регрессии
```

```
score = mdl.predict(X(testIdx, :)); % прогнозирование для контрольной выборки
    y1=(score>=kk); % сравниваем с порогом
mse(k) = mean(abs(Y(testIdx) - y1)) * 100; % расчет процента ошибки на контрольной выборке
    [accuracy(k), precision(k), recall(k)] = property(Y(testIdx), y1); % вызов функции для расчета точности, полноты
F(k) = 2 * precision(k) * recall(k) / (precision(k) + recall(k)); % расчет F-критерия
yt = Y(testIdx);
    [FPR_nn,TPR_nn,~,AUC(k)] = perfcurve(yt,score,'1'); % расчетAUC
end;
str = ['ЛР']; % Значение столбца «Метод» в таблице результатов
name(sizeUI, 1) = cellstr(str);
%  расчет средних значений критериев оценки качества модели логистической регрессии
ac = mean(accuracy);
pr = mean(precision);
rec = mean(recall);
Fm = mean(F);
a = mean(AUC(k));
```

Пример реализации алгоритма функции для вычисления критериев качества классификации:

```
function [a, p, r] = property(x1, x2) % Объявление функции для расчета критериев качества
TP = 0; % Обнуляем значения для TruePositive
TN = 0; % Обнуляем значения для TrueNegative
FP = 0; % Обнуляем значения для FalsePositive
FN = 0; % Обнуляем значения для FalseNegative
fork = 1 : size(x1)
    if (x1(k) == 1) % Если объект является исправным
        if (x1(k) == x2(k))
            TP = TP + 1;
```

```
                else
                FP = FP + 1;
                end;
        else % Если объект является неисправным
                if (x1(k) == x2(k))
        TN = TN + 1;
                else
                        FN = FN + 1;
                end;
        end;
end;
a = (TP + TN) / (TP + TN + FP + FN);
p = TP / (TP + FP);
r = TP / (TP + FN);
```

3) Блок выбора объема контрольной выборки служит для определения оптимального объема контрольной выборки.

Расчет значений *F*-критерия для объема контрольной выборки 25%:

```
for kcv=1:V(1);
trainCVx = X(cv.training(kcv),:); % обучающая выборка X
trainCVy = Y(cv.training(kcv),:); % обучающая выборка Y
testCVx = X(cv.test(kcv),:); % тестовая выборка X
testCVy = Y(cv.test(kcv));    % тестовая выборка Y
for j = 1 : 11
f(kcv, j) = AllM(j, trainCVx, trainCVy, testCVx, testCVy); % Расчет F-критерия для всех базовых методов
end;
end;
FMean = mean(f); % усреднение результатов
maxValue = zeros(1, 3);
```

4) Блок отбора значимых показателей функционирования.

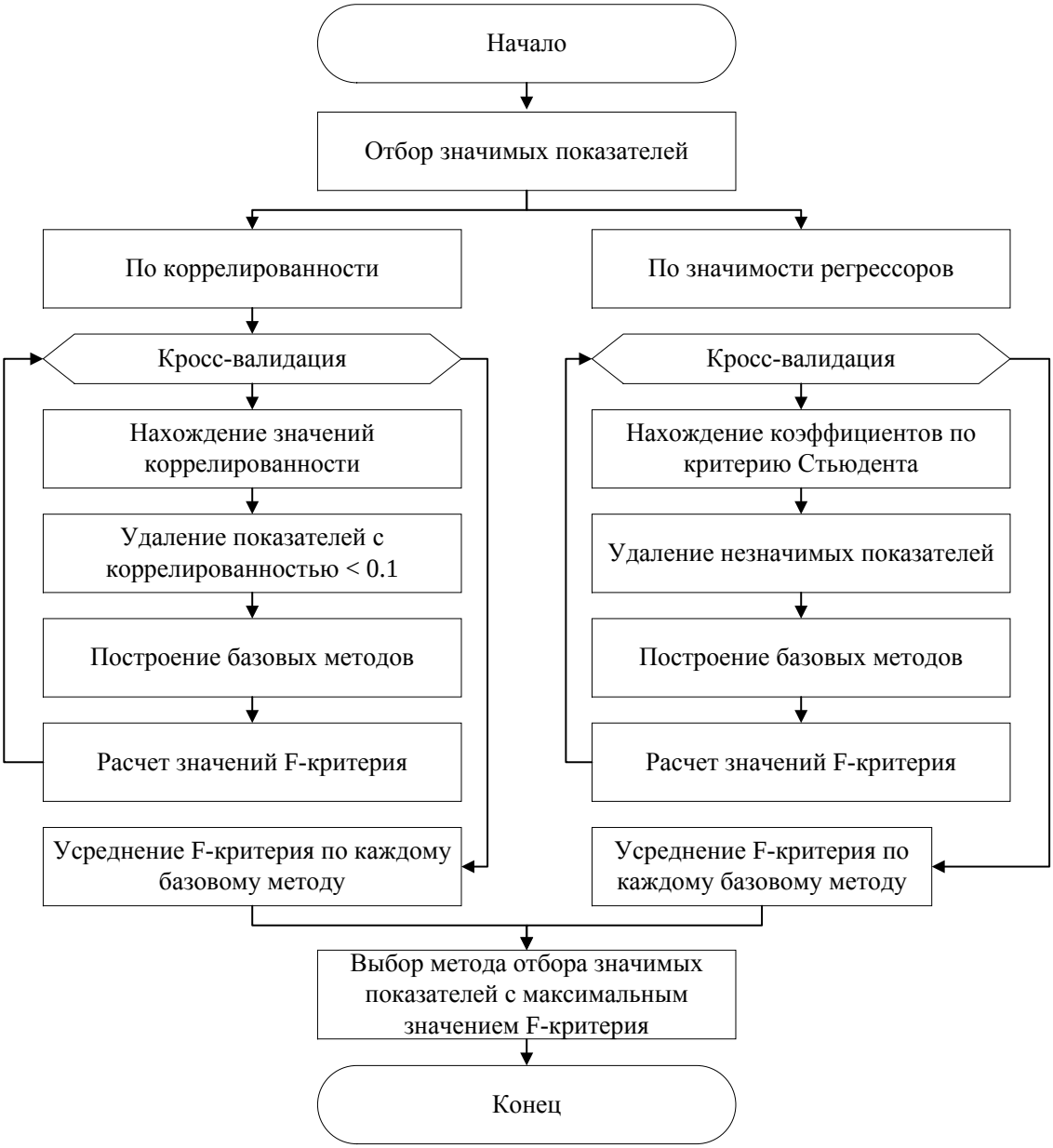


Рисунок 4.1. Блок-схема алгоритма отбора значимых показателей

Пример реализации алгоритма отбора значимых признаков по коррелированности:

```
cv = cvpartition(numel(Y), 'kfold', V(maxV)); % Получение значений индексов элементов для контрольной и обучающей выборок
for kcv = 1 : V(maxV); % Начало кросс-валидации
trainCVx = X(cv.training(kcv),:); % Выбираем элементы для обучающей выборки из матрицы параметров
trainCVy = Y(cv.training(kcv),:); % Выбираем элементы для обучающей выборки из вектора-столбца отклика
```

testCVx = X(cv.test(kcv),:); % Выбираем элементы для контрольной выборки из матрицы параметров

testCVy = Y(cv.test(kcv)); % Выбираем элементы для контрольной выборки из вектора-столбца отклика

[b,bint,r,rint,stats] = regress(trainCVy,trainCVx); % Построение регрессионной модели

R = zeros(size(b)); % Создаем вектор, заполненный нулями

for i = 1 : size(b)

tmpR = corrcoef(X(:,i),Y); % Получаем коэффициенты корреляционной матрицы

R(i) = tmpR(1, 2);

end;

trainCVx = trainCVx(:, abs(R) > 0.1); % Выбираем параметры, которые показали значения больше 0.1 для обучающей выборки

testCVx = testCVx(:, abs(R) > 0.1); % Выбираем параметры, которые показали значения больше 0.1 для контрольной выборки

for j = 1 : 3; % Цикл по трем лучшим методам

tmp(kcv, j) = AllM(Imax(j), trainCVx, trainCVy, testCVx, testCVy); % Получаем значения F-критерия для методов

end;

end;

xOtb(2) = max(mean(tmp)); % Запоминаем максимальное значение

Пример реализации алгоритма отбора значимых признаков по значимости регрессоров:

tmp = zeros(V(maxV), 3); % Получение значений индексов элементов для контрольной и обучающей выборок

forkcv = 1 : V(maxV); % Начало кросс-валидации

trainCVx = X(cv.training(kcv), :); % Выбираем элементы для обучающей выборки из матрицы параметров

trainCVy = Y(cv.training(kcv), :); % Выбираем элементы для обучающей выборки из вектора-столбца отклика

testCVx = X(cv.test(kcv),:); % Выбираем элементы для контрольной выборки из матрицы параметров

```
testCVy = Y(cv.test(kcv)); % Выбираем элементы для контрольной выборки из вектора-столбца отклика
[b, bint, r, rint, stats] = regress(trainCVy, trainCVx);
YFIT = zeros(size(trainCVy));
for i = 1 : size(trainCVy);
    for j = 1 : size(b);
        YFIT(i) = YFIT(i) + b(j) * trainCVx(i, j);
    end;
end;
YY = sum((trainCVy - YFIT) .^ 2);
f = size(trainCVy) - 2;
XX = zeros(size(b));
fori = 1 : size(b);
    XX(i) = sum((trainCVx(:, i) - mean(trainCVx(:, i))) .^ 2);
end;
s = zeros(size(b));
Tst = zeros(size(b));
ss = size(trainCVy) - 1 - 1; % Степени свободы
znazh = zeros(size(b)); % Создаем вектор, заполненный нулями
Tkr = tinv(0.95, ss(1));
fori = 1 : size(b); % Расчет значения по тесту Стьюдента
    s(i) = sqrt(YY / (f(1) * XX(i)));
    Tst(i) = abs(b(i)) / s(i);
    if (Tst(i) >Tkr)
        znach(i) = 1;
    end;
end;
trainCVx = trainCVx(:, znach> 0);
testCVx = testCVx(:, znach> 0);
for j = 1 : 3; % % Цикл по трем лучшим методам
    tmp(kcv, j) = AllM(Imax(j), trainCVx, trainCVy, testCVx, testCVy); % Получаем значения F-критерия для методов
end;
end;
xOtb(3) = max(mean(tmp)); % Запоминаем максимальное значение
```

5) Блок агрегирования методов машинного обучения.

На рисунке 4.2 представлена блок-схема расчета *F*-критерия для агрегированного метода по среднему значению.

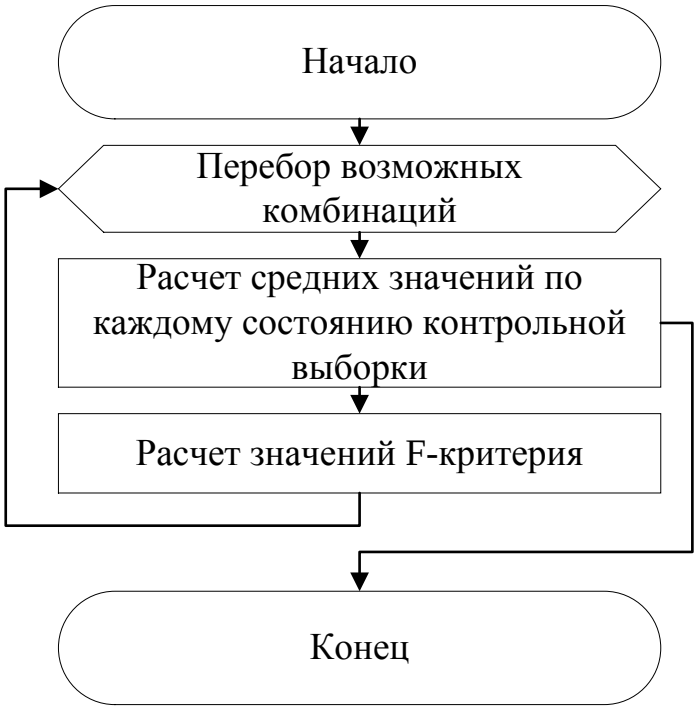


Рисунок 4.2. Блок-схема алгоритма АМ-С

6) Блок вывода получившихся результатов в таблицу предназначен для визуализации результатов.

Реализация вывода результатов в таблицу:

```
tableR(4,1) = {t1}; % Заполнение столбцов
tableR(4,2) = {t2};
tableR(5,1) = {t3};
tableR(5,2) = {t4};
tableR(6,1) = {t5};
tableR(6,2) = {t6};
set(handles.uitable6, 'Data', tableR); % Вывод в таблицу на форме программы
```

7) Блок прогнозирования состояния технического объекта.

Прогнозирование состояния технического объекта наивным байесовским методом:

```
NB = NaiveBayes.fit(X,Y,'Distribution','normal'); % Обучение модели наивного байесовского классификатора
score = posterior(NB, Xnew'); % прогнозирование по полученной модели
res = (score(2) >= kk); % сравнение вероятностей с порогом
data(3) = {res}; % подготовка результатов для вывода в таблицу
table(3) = res;
```

Прогнозирование состояния технического объекта методом дискриминантного анализа:

```
try
da = ClassificationDiscriminant.fit(xtrain,ytrain, 'discrimType', 'diagLinear'); % Обучение модели дискриминантного анализа
catch
    %Если ошибка – ничего не происходит
end
   [~, score] = da.predict(xtest); % прогнозирование по полученной модели
p = score(:, 2); % записываем вектор столбец результатов
y1 = (p>= kk); % сравниваем полученные вероятности с порогом
[~, precision, recall] = property(ytest, y1);% получение значений точности и полноты
F = 2 * precision * recall / (precision + recall);% расчет значения F-критерия
```

Прогнозирование состояния технического объекта методом бустинга GentleBoost:

```
gb = fitensemble(xtrain, ytrain,'GentleBoost',160,'Tree'); % Обучение модели GentleBoost
[labels, s1] = predict(gb, xtest); % прогнозирование по полученной модели
M = max(abs(s1(:, 2))); % поиск максимального значения в столбце полученного прогноза
p = ((s1(:, 2) / M) + 1) / 2; % нормируются значения
y1=(p>=kk); % сравнение вероятностей с порогом
```

```
[~, precision, recall] = property(ytest, y1); % получение значений точности и полноты
F = 2 * precision * recall / (precision + recall); % расчет значения F-критерия
```

Вначале пользователю предлагается выбор файла, в котором хранятся исходные данные. Данный файл должен быть в формате .xls или .xlsx, а содержимое файла: 1 столбец – отклик, далее – столбцы показателей функционирования технического объекта.

После считывания файла исходных данных задается объем контрольной выборки. Для сокращения вычислений принимается начальное значение объема контроля 25%, т.е. при кросс-валидации выборка разбивается на $H = 4$ части, три из которых используются для обучения.

Далее проводится обучение по всем 11 методам с использованием всех заданных показателей функционирования объекта (без отбора значимых показателей) на базе библиотеки инструментов Statistics and Machine Learning Toolbox в пакете Matlab с расчетом *F*-меры. Выбираются заданное пользователем количество методов (исследования показывают, что достаточно два-три, хотя возможен и полный перебор), показавших наибольшее значение *F*-меры.

При оценке исправности объекта возможно и изменение порога - значения, которое определяет границы классов: например, если порог = 0.6, то значения вероятностей от 0.6 до 1 будут относиться к классу исправных объектов, а меньше 0.6 к классу неисправных.

Для отобранных методов исследуется вначале влияние объема контрольной выборки, а затем – способа отбора значимых показателей. Объем контрольной выборки последовательно снижается с 25% ($H = 4$) до 5% (соответственно $H = 20$) с шагом 5%. Для вариантов, в которых *F*-мера оказалась максимальной, проводится отбор значимых показателей. При

этом используются два подхода. Один – удаление показателей, для которых корреляция показателей функционирования объекта с его исправностью незначима; второй – по незначимости регрессоров (по критерию Стьюдента) в линейной регрессионной модели *Y* от *X*.

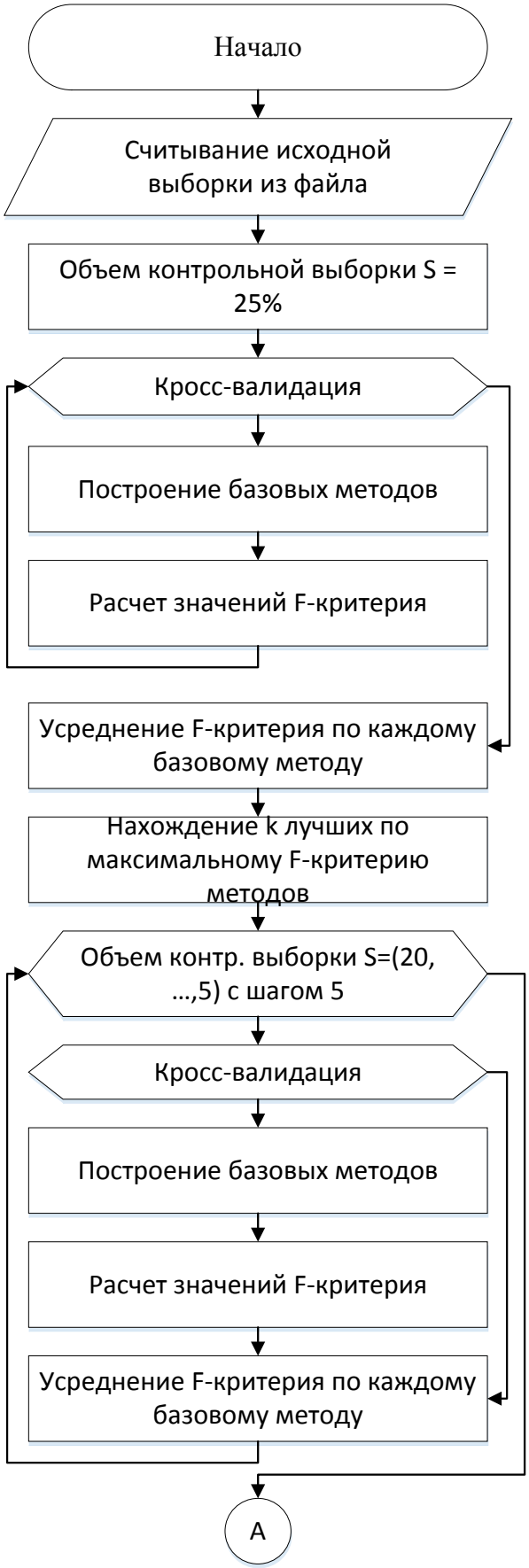
По результатам проведенных испытаний принимается значение объема контрольной выборки и показатели, которые будут использованы для продолжения машинного обучения. С учетом этих двух факторов строятся агрегированные классификаторы трех типов - по среднему значению, по медиане, и с помощью процедуры голосования. Выбирается модель (включающая вид агрегирования и компоненты агрегата: несколько базовых классификаторов), которая в дальнейшем будет использована для прогнозирования состояния технического объекта.

Разработанный программный комплекс по данному алгоритму имеет следующие характеристики: тип операционной системы – Windows7 и выше; среда разработки – Matlab R2016a; общий размер программы – 1,87 Мбайт.

После загрузки файла с исходными данными (файл формата .xls или .xlsx) на экране программы отображается выбранные данные. Далее пользователь может выполнить диагностику, нажав на кнопку «Выполнить расчет».

После выполнения расчета, можно произвести прогноз нового состояния, нажав соответствующую кнопку на экране программы.

4.2. Блок-схема алгоритма



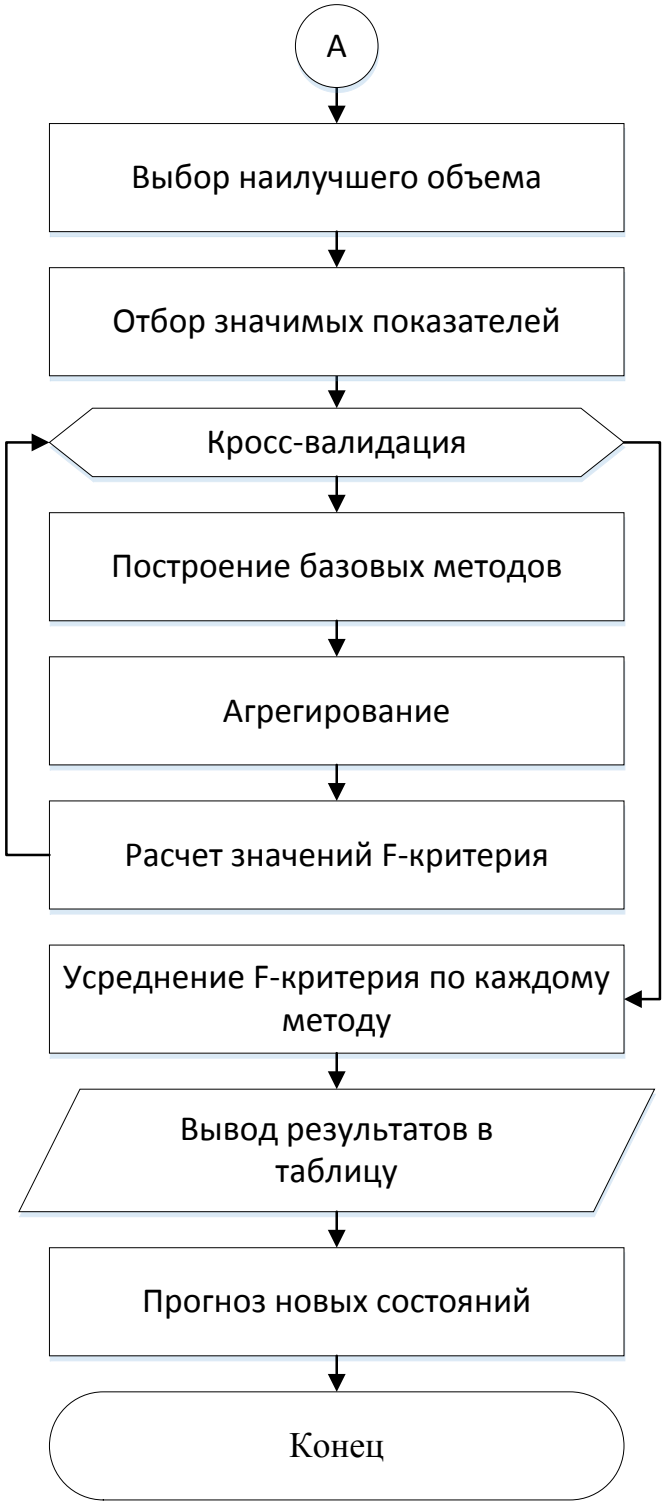


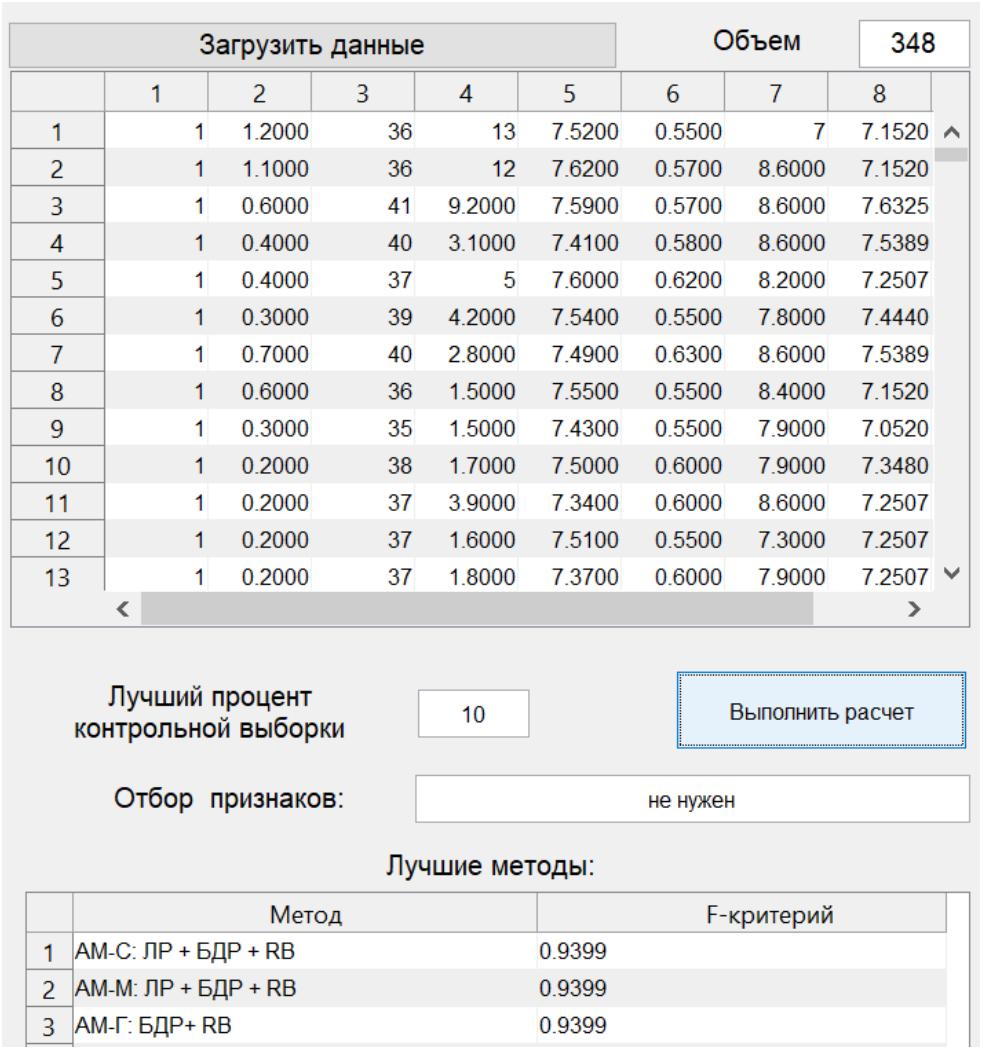
Рисунок 4.3. Блок-схема алгоритма программного комплекса «Диагностика состояния технического объекта с использованием агрегированных классификаторов»

4.3. Интерфейс программы

При запуске программного комплекса «Диагностика состояния технического объекта с использованием агрегированных классификаторов» на экране появляются соответствующие окно (рис. 4.4.).

Кнопка «Загрузить данные» позволяют пользователю выбрать файл, содержащий исходную выборку. В окне «Объем» появится объем исходной выборки, а в таблице появятся выбранные пользователем исходные данные – в первом столбце отклик (0 – неисправный, 1 - исправный), а в последующих столбцах выводятся показатели функционирования.

Нажатие на кнопку «Выполнить расчет» запускает все расчеты данного программного комплекса. После выполнения алгоритма программы в окне «Отбор признаков» появляется одно из трех значений: «по коррелированности», «по значимости регрессоров» или «не нужен»; также в таблицу «Лучшие методы» выводятся три лучших базовых методов, а также три метода агрегирования: по среднему, по медиане и по голосованию.



Загрузить данные

Объем 348

	1	2	3	4	5	6	7	8
1	1	1.2000	36	13	7.5200	0.5500	7	7.1520
2	1	1.1000	36	12	7.6200	0.5700	8.6000	7.1520
3	1	0.6000	41	9.2000	7.5900	0.5700	8.6000	7.6325
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

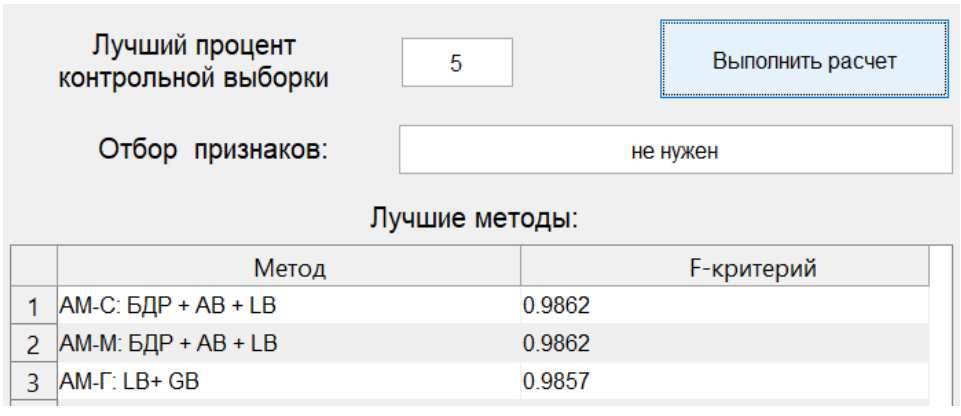
Выполнить расчет

Отбор признаков: не нужен

Лучшие методы:

	Метод	F-критерий
1	АМ-С: ЛР + БДР + RB	0.9399
2	АМ-М: ЛР + БДР + RB	0.9399
3	АМ-Г: БДР+ RB	0.9399

а)



Лучший процент контрольной выборки 5

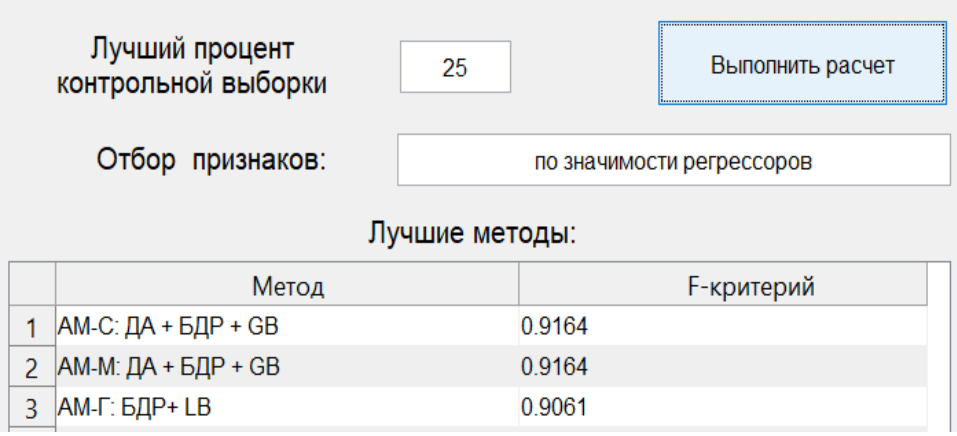
Выполнить расчет

Отбор признаков: не нужен

Лучшие методы:

	Метод	F-критерий
1	АМ-С: БДР + АВ + LB	0.9862
2	АМ-М: БДР + АВ + LB	0.9862
3	АМ-Г: LB+ GB	0.9857

б)



в)

Рисунок 4.4. Результаты автоматизированного расчета:

а) для системы водоочистки, *б*) для гидроагрегата, *в*) для счетчиков

4.4. Прогнозирование состояния объекта

В программном комплексе «Диагностика состояния технического объекта с использованием агрегированных классификаторов» возможно произвести прогнозирование нового технического состояния. Прогнозирование происходит по лучшему методу, полученному при выполнении основного алгоритма программного комплекса.

При вводе данных, например, из таблицы 4.1,

Таблица 4.1. Новые данные для системы водоочистки

X1	X2	X3	X4	X5	X6	X7	X8
0,20	37,00	1,10	7,52	0,55	8,50	6,95	0,20

получаем результат (рис. 4.5): У = 1 – объект исправен.

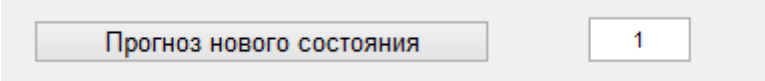


Рисунок 4.5. Прогнозирование нового технического состояния для системы водоочистки

4.5. Экономическое обоснование

Затраты, связанные с нарушением исправности системы водоочистки, можно представить в виде суммы двух слагаемых:

Z_1 – затраты, связанные с устранением обнаруженного нарушения (изменение доз коагулянта и флокулянта), средний расход на одно нарушение – $z_1 = 20$ тыс. руб.,

Z_2 – затраты, связанные с остановкой работы системы вследствие ложной тревоги, стоимость одного часа простоя – $Z_2 = 30$ тыс. руб.

Затраты Z_1 определяются в зависимости от количества нарушений за заданный период. Пусть λ – среднее количество нарушений за один день, P_1 - вероятность пропуска нарушения, тогда за месяц (30 дней) количество нарушений окажется равным $\lambda P_1 * 30$, а соответствующие затраты:

$$Z_1 = \lambda * P_1 * 365 * z_1.$$

Затраты Z_2 определяются количеством часов простоя, связанного с ложной тревогой. Пусть P_2 – вероятность ложной тревоги, тогда за месяц количество ложных тревог окажется $P_2 * 30$, если время простоя, связанное с одной ложной тревогой, составляет t, то затраты, связанные с остановкой системы из-за ложной тревоги составят:

$$Z_2 = P_2 * 365 * t * z_2.$$

В диссертационной работе Бубыря Д.С. [11] приведены оценки этих показателей при существующей системе обнаружения нарушений на примере критерия нарушения цветности на водоканале Санкт-Петербурга: $P_1 = 0{,}0012$ (т.е. раз в 83 дня происходит пропуск нарушения), $P_2 = 0{,}026$ (т.е. раз в 38 дней происходит одна «ложная тревога»), $\lambda = 0{,}33$ (одно нарушение в 3 дня), $t = 2$ часа. Тогда суммарные затраты за месяц:

$$0{,}33 * 0{,}0012 * 365 * 20 + 0{,}026 * 365 * 2 * 30 = 572 \text{ тыс. руб.}$$

Предлагаемый подход с использованием агрегированных классификаторов, поиском наилучшего объема контрольной выборки и оценкой значимости обеспечивает повышение F-меры на 15%, что

приводит к соответствующему снижению как вероятности ложной тревоги, так и вероятности пропуска нарушения, в результате и суммарные затраты сокращаются на 15% и составят 486 тыс. руб. таким образом годовой экономический эффект составит 86 тыс. руб. только при учете одного из 6 показателей. Предполагая аналогичный порядок значений для всех контролируемых показателей, получим ориентировочный экономический эффект в 516 тыс. руб.

4.6. Выводы по главе

Разработан алгоритм и программный комплекс «Диагностика состояния технического объекта с использованием агрегированных классификаторов» для диагностики и прогнозирования состояния технического объекта на основе агрегированных методов классификации, включающий в себя: разбиение исходных данных на контрольную и обучающую выборки, отбор значимых признаков, построение базовых и агрегированных классификаторов, поиск наилучшего по *F*-критерию метода, а также возможность прогнозирования состояния технического объекта. Все перечисленные функции рассчитываются в автоматизированном режиме, тем самым для осуществления диагностики технического объекта достаточно загрузить данные о его предшествующих состояниях и программный комплекс подберет оптимальные параметры для получения наиболее точного результата.

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.

6) Проведенное численное исследование на реальных данных системы водоочистки 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.

СПИСОК ЛИТЕРАТУРЫ

1. Абрамов О. В. Контроль и прогнозирование технического состояния систем ответственного назначения // Надежность и качество сложных систем. – 2018. – №. 4 (24). – С. 108–115.
2. Алексеева, Е. В. Численные методы оптимизации: учеб. пособие / Е.В. Алексеева, О. А. Кутненко, А. В. Плясунов. – Новосиб. ун-т. Новосибирск, 2008. – 128 с.
3. Амренов С. А. Методы контроля и диагностики систем и сетей связи / - Астана, Казахский государственный агротехнический университет, 2005 г. – URL: https://www.studmed.ru/amrenov-sa-konspekt-lekcii-metody-kontrolya-i-diagnostiki-sistem-i-setey-svyazi-chast-2_779971cca99.html (дата обращения: 21.02.2019).
4. Андерсон, Т. Введение в многомерный статистический анализ / Т. Андерсон. – М. : Физматгиз, 1963. – 500 с.
5. Балдин, А. В. Прикладной статистический анализ данных / А. В. Балдин, В. В. Криницин. – М. : ПРИОР, 1998. – 261 с.
6. Барский А. Б. Нейронные сети: распознавание, управление, принятие решений. — М.: Финансы и статистика, 2004. – 176 с.
7. Бендат, Дж. Прикладной анализ случайных данных / Дж. Бендат, А. Пирсол – М. : Мир, 1989. – 540 с.
8. Бигус Г.А. ,Даниев Ю.Ф. , Быстрова Н.А. , Галкин Д.И. Основы диагностики технических устройств и сооружений. М.: Изд-во МВТУ им. Н.Э.Баумана. 2015. – 448 с.
9. Биргер И.А. Техническая диагностика. М.: Машиностроение, 1978.– 240 с. (2-е изд.: М. : URSS, 2019).
10. Боровиков, В. П. Прогнозирование в системе STATISTICA в среде Windows / В. П. Боровиков, Г. И. Ивченко. – М. : Финансы и статистика, 1999. – 384 с.

11. Бубырь, Д.С. Разработка моделей, алгоритмов и программ прогнозирования показателей качества питьевой воды в системе водоочистки : дис. канд. тех. наук: 05.13.18 . – Ульяновск, 2017. – 145 с.
12. Валеев С.Г. Регрессионное моделирование при обработке наблюдений. М.: Наука, 1991. – 272 с.
13. Валеев С.Г., Булыжев Е.М. Системы раннего предупреждения аномальной ситуации при анализе состояния СОЖ // Справочник. Инженерный журнал. – 2011. – №10. – С. 39–42.
14. Васильев В.И., Жернаков С.В. Классификация режимов работы ГТД с использованием технологии нейронных сетей / Вестник Уфимского государственного авиационного технического университета. – 2009. – Т. 12(1). – С. 53–56.
15. 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.
16. Васильев, Н.П. Опыт расчета параметров логистической регрессии методом Ньютона-Рафсона для оценки зимостойкости растений / Н. П. Васильев, А. А. Егоров // Математическая биология и биоинформатика. – 2011. – Т. 6. – № 2. – С.190–199.
17. Вьюгин В.В. Математические основы машинного обучения и прогнозирования. – М. : МЦНМО, 2014. – 304 с.
18. Воронцов К.В. Машинное обучение. Композиции классификаторов. – URL: http://www.intuit.ru/studies/courses/13844/1241/lecture/27000 (дата обращения: 01.03.2019).
19. Гольдштейн А. Б., Поздняков В. А., Скоринов М. Ю. Анализ эффективности ансамблевых методов прогнозирования оттока клиентов // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2018). – 2018. – С. 243–247.

20. Грешилов А. А. Математические методы принятия решений. – Издательство МГТУ им. НЭ Баумана, 2014. –647 с.
21. Дубров, А. М. Многомерные статистические методы: Учебник / А.М. Дубров, В. С. Мхитарян, Л. И. Трошин. – М. : Финансы и статистика, 2003. – 352 с.
22. Жернаков С. В. Применение технологии нейронных сетей для диагностики технического состояния авиационных двигателей / Интеллектуальные системы в производстве. – 2006. – № 2 (8). – С. 70–83.
23. Жернаков С.В., Гильманшин А.Т. Применение интеллектуальных алгоритмов на основе нечеткой логики и нейронных сетей для решения задач диагностики отказов авиационного ГТД // В сборнике: Intelligent Technologies for Information Processing and Management (ITIPM'2014) Proceedings of the 2nd International Conference. 2014. С. 112–115.
24. Жуков Д.А. Анализ критериев качества классификации при диагностике функционирования технического объекта. // Автоматизация процессов управления. – 2019. – № 3 (57). – С. 112–117.
25. 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.
26. 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.

27. 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.
28. 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.
29. Жуков Д. А., Клячкин В. Н. Алгоритмы бустинга в задачах технической диагностики //Перспективные информационные технологии (ПИТ 2017). – 2017. – С. 787–790.
30. 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.
31. 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.
32. 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.
33. Жуков Д.А., Клячкин В.Н. Использование агрегированных классификаторов при машинном обучении в задачах технической диагностики // Информационные технологии моделирования и управления. – 2019. – С.75–80.

34. Жуков Д.А., Клячкин В.Н. Критерии качества диагностики функционирования технических объектов методами машинного обучения // Информатика, моделирование, автоматизация проектирования" (ИМАП - 2018). – 2018. – С. 87–90.
35. 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.
36. 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.
37. Жуков Д.А., Клячкин В.Н., Кувайскова Ю.Е. Сравнительный анализ методов машинного обучения при прогнозировании состояния технического объекта // Радиоэлектронная техника. – 2017. – №. 1 (10). – С. 189–192.
38. Жуков Д. А., Клячкин В. Н. Диагностика исправности технического объекта с использованием пакета MATLAB // Перспективные информационные технологии: труды Международной научно-технической конференции. – Самарский научный центр РАН, 2018. – С. 55–57.
39. 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.

40. Замятин А. В. и др. Введение в интеллектуальный анализ данных: учебное пособие. – 2016. – 119 с.
41. Калиткин Н.Н. Численные методы. – М: Наука, 1978. – 512 с.
42. Киселевич В.П., Клячкин В.Н., Сухов В.В. Прогнозирование ресурса вычислительной системы по результатам испытаний // Автоматизация процессов управления. – 2014. – №. 1 (35). – С. 55–60.
43. Клячкин В. Н. Статистические методы в управлении качеством: компьютерные технологии. М.: Финансы и статистика, ИНФРА-М, 2009. – 304с.
44. 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.
45. 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.
46. Клячкин В. Н., Кувайскова Ю.Е., Жуков Д.А. Диагностика технического состояния аппаратуры с использованием агрегированных классификаторов // Радиотехника. – 2018. – №. 6. – С.46–49.
47. Клячкин В.Н. Диагностика состояния технического объекта с использованием агрегированных классификаторов / В.Н. Клячкин, Д.А. Жуков // Свидетельство о регистрации программы для ЭВМ. №2019611560. – 2019.
48. Клячкин В.Н. Модели и методы статистического контроля многопараметрического технологического процесса. М.: ФИЗМАТЛИТ, 2011. – 196 с.
49. Клячкин В.Н. Оценка исправности технического объекта с применением машинного обучения / В.Н. Клячкин, Ю.Е. Кувайскова,

Д.А. Жуков // Свидетельство о регистрации программы для ЭВМ. №2019611562. – 2019.

50. Клячкин В.Н., Жуков Д.А. Алгоритм диагностики функционирования технического объекта с использованием агрегированных классификаторов // Автоматизация процессов управления. – 2019. – №. 2 (56). – С. 37–43.
51. Клячкин В.Н., Жуков Д.А. Прогнозирование состояния технического объекта с применением методов машинного обучения //Программные продукты и системы. – 2019. – №. 2. – С.244–250.
52. Клячкин В.Н., Кувайскова Ю.Е., Алексеева В.А.. Статистические методы анализа данных. М. : Финансы и статистика, 2016. – 240с.
53. Костин К. А., Ламонова Т. С. Классификация патологий диссеминированного туберкулёза лёгких с помощью методов машинного обучения //Сборник избранных статей научной сессии ТУСУР. – 2018. – Т. 1. – №. 3. – С. 129–132.
54. Костин Д. В., Шелухин О. И. Сравнительный анализ алгоритмов машинного обучения для проведения классификации сетевого зашифрованного трафика // T-Comm-Телекоммуникации и Транспорт. – 2016. – Т. 10. – №. 9. – С. 43–52.
55. Кузьмина С.В., Ефимов А.И. Актуальные методы машинного обучения в области классификации // Актуальные проблемы современной науки и производства. – 2018. – С. 34–38.
56. Лепский А. Е. Математические методы распознавания образов: Курс лекций / А. Е. Лепский, А. Г. Броневич. – Таганрог: Изд-во ТТИЮФУ, 2009. – 155 с.
57. Мерков А. Б. Распознавание образов. Введение в методы статистического обучения //М.: Едиториал УРСС. – 2011. – 256 с.
58. Пархоменко П. П. Определение технического состояния многопроцессорных вычислительных систем путем анализа графа синдромов //Автоматика и телемеханика. – 1999. – №. 5. – С. 126–134.

59. Платонов Ю. М., Уткин Ю. Г. Диагностика, ремонт и профилактика персональных компьютеров. – М.: Горячая линия - Телеком, 2003. – 312 с.
60. Поляк Б.Т., Цыпкин Я.З. Псевдоградиентные алгоритмы адаптации и обучения // Автоматика и телемеханика. – 1973. – №3. – С. 45–68.
61. Поляк Б.Т. Оптимальные псевдоградиентные алгоритмы адаптации / Б. Т. Поляк, Я. З. Цыпкин // Автоматика и телемеханика. – 1980. – №. 8. – С. 74–84.
62. СанПиН 2.1.4.1074-01 «Питьевая вода. Гигиенические требования к качеству воды централизованных систем питьевого водоснабжения. Контроль качества».
63. Санталов А. А., Жуков Д. А. Диагностика технического состояния системы с применением нейросетевых методов // Перспективные информационные технологии (ПИТ 2018). – 2018. – С. 202–205.
64. Себер, Дж. Линейный регрессионный анализ / Дж. Себер. – М. : Мир, 1980. – 456 с.
65. Соколов Е.А. ФКН ВШЭ. Лекция 4. Линейная классификация. – URL: https://github.com/esokolov/ml-course-hse/blob/master/2018-fall/lecture-notes/lecture04-linclass.pdf (дата обращения: 01.03.2019).
66. Теория и практика машинного обучения: учеб. пособие / В.В. Воронина, А.В. Михеев, Н.Г. Ярушкина, К.В. Святов. – Ульяновск: УлГТУ, 2017. – 290 с.
67. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. – Litres, 2017. – 392 с.
68. Хайкин Саймон. Нейронные сети: полный курс, 2-е издание.: Пер. с англ. – М. : Издательский дом “Вильям”, 2006. – 1104 с.
69. Халафян А.А. STATISTICA 6. Статистический анализ данных. 3-е изд. М.: ООО «Бином-Пресс», 2007. – 512 с.
70. Чернышова Г. Ю., Красильникова Е. Ю. Применение методов

интеллектуального анализа данных для оценки времени простоя оборудования в процессе технического обслуживания и ремонта // Вестник Саратовского государственного социально-экономического университета. – 2017. – №. 3 (67). – С. 161–166.

71. Черкасов Д. Ю., Иванов В. В. МАШИННОЕ ОБУЧЕНИЕ //Наука, техника и образование. – 2018. – №. 5 (46).
72. Четыркин Е. М. Статистические методы прогнозирования. М.: Статистика, 1977. – 200 с.
73. Шанов С. В., Чупин П. Г., Афонин А. Ю. Применение байесовского классификатора для определения тематики текста //Моделирование, оптимизация и информационные технологии. – 2018. – Т. 6. – №. 1. – С. 131–139.
74. Шунина Ю.С. Прогнозирование платежеспособности клиентов банка на основе методов машинного обучения и марковских цепей / Ю.С. Шунина, В.Н. Клячкин // Программные продукты и системы. – 2016. – №2. – С. 105–112.
75. Юлдашев М.Н., Адамов А.П., Адамова А.А. Классификация состояний беспроводной сенсорной сети с использованием методов машинного обучения // Проблемы разработки перспективных микро- и наноэлектронных систем (МЭС). – 2016.– №. 2. – С. 248–251.
76. Ясницкий Л.Н. Искусственный интеллект. Элективный курс: Учебное пособие. – М.: БИНОМ. Лаборатория знаний. – 2011. – 240с.
77. Ясницкий Л. Н. Интеллектуальные системы: учебник //М.: Лаборатория знаний. – 2016. – 221с.
78. Alexey Nefedov. Support Vector Machines: A Simple Tutorial. – 2016. – 34p.
79. Babajide Mustapha I., Saeed F. Bioactive molecule prediction using extreme gradient boosting // Molecules. – 2016. – Vol. 21. – №. 8. – P. 983.

80. Bersimis S., Psarakis S., Panaretos J. MultivariateStatistical Process Control Charts: An Overview //Quality and reliability Engineering International. – 2007. – V. 23. – P. 517–543.
81. Breiman, L. Bagging predictors / L. Breiman // Machine Learning. – 1996. – Vol. 26. – №. 2. – P. 123–140.
82. Cameron, A.A., Trivedi, P.K. Regression Analysis of Count Data. — Cambridge: Cambridge University Press, 2013. – 553 p.
83. Chen T., Guestrin C. Xgboost: A scalable tree boosting system // Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining. – ACM, 2016. – P. 785–794.
84. Czichos, H. Technical Diagnostics: Principles, Methods, and Applications / Horst Czichos // NCSLI Measure. – 2014. – Vol.9. – P.32–40. – DOI: 10.1080/19315775.2014.11721681
85. 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.
86. De Menezes F. S. et al. Data classification with binary response through the Boosting algorithm and logistic regression // Expert Systems with Applications. – 2017. – Vol. 69. – P. 62-73.
87. De Sá J. P. M. Applied statistics using SPSS, Statistica, MatLab and R. – Springer Science & Business Media, 2007. – 520p.
88. Duer S., DuerR., Mazuru S. Determination of the expert knowledge base on the basis of a functional and diagnostic analysis of a technical object // Nonconventional Technologies Review / Revista de Tehnologii Neconventionale. – 2016. – Vol. 20. – №. 2. – P. 23–29.
89. Gaikwad D. P., Thool R. C. Intrusion detection system using bagging ensemble method of machine learning // 2015 International Conference on Computing Communication Control and Automation. – IEEE, 2015. – P. 291–295.–doi: 10.1109/ICCUBEA.2015.61.
90. Hand D. J., Till R. J. A simple generalisation of the area under the ROC

curve for multiple class classification problems // Machine learning. – 2001. – Vol. 45. – №. 2. – P. 171–186.

91. Kiselev M. I., Pronyakin V. I., Tulekbaeva A. K. Technical diagnostics functioning machines and mechanisms //IOP Conference Series: Materials Science and Engineering. – IOP Publishing, 2018. – Vol. 312. – №. 1. – P. 012012.
92. Klyachkin V.N., Kuvayskova Yu.E., Zhukov D.A. The use of aggregate classifiers in technical diagnostics, based on machine learning / CEUR Work-shop Proceedings. – 2017. – V. 1903. – P. 32–35.
93. 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.
94. Krasheninnikov V. R., Klyachkin V. N., Kuvayskova Y. E. Models Updating for Technical Objects State Forecasting // 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC). – IEEE, 2018. – P. 1–4.
95. Kuravsky L. S., Baranov S. N. Technical diagnostics and monitoring based on capabilities of wavelet transforms and relaxation neural networks // Insight-Non-Destructive Testing and Condition Monitoring. – 2008. – T. 50. – №. 3. – P. 127-132.
96. Kuvayskova Y.E. The prediction algorithm of the technical state of an object by means of fuzzy logic inference models // Procedia Engineering. «3rd International Conference «Information Technology and Nanotechnology», ITNT 2017». – 2017. – P. 767–772.
97. Mark Hudson Beale, Martin T. Hagan, Howard B. Demuth Neural NetworkToolbox. User’s Guide. – 2017. – URL: https://www.academia.edu/34938587/Neural_Network_Toolbox_Users_Guide (дата обращения: 14.04.2018).

98. Menčík J. Diagnostics // Concise Reliability for Engineers. – Intech Open, 2016. – URL: https://www.intechopen.com/books/concise-reliability-for-engineers/diagnostics (дата обращения: 11.02.2019).
99. Montgomery D.C. Introduction to statistical quality control. – New York: John Wiley and Sons, 2009. – 754 p.
100. Neykov M., Liu J.S., Cai T. On the characterization of a class of fisher-consistent loss functions and its application to boosting // J. of Machine Learning Research.–2016. – No. 17. – P. 1–32.
101. Oulladji L. et al. Arabic text detection using ensemble machine learning //International Journal of Hybrid Intelligent Systems. – 2018. – Vol. 14. – №. 4. – P. 233–238.
102. Repp P. V. The system of technical diagnostics of the industrial safety information network //Journal of Physics: Conference Series. – IOP Publishing, 2017. – Т. 803. – №. 1. – P. 012127.
103. Ryan T. P. Statistical methods for quality improvement. – John Wiley & Sons, 2011. – 687 p.
104. Šoltésová S., Baron P. The operation monitoring condition of the production machinery and facilities using the tools of technical diagnostics //Applied Mechanics and Materials. – Trans Tech Publications, 2013. – Т. 308. – P. 105-109.
105. Tao H. et al. Real-time driver fatigue detection based on face alignment // Ninth International Conference on Digital Image Processing (ICDIP 2017). – International Society for Optics and Photonics, 2017. – Vol. 10420. – P. 1042003.
106. Torlay L. et al. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy //Brain informatics. – 2017. – Т. 4. – №3. – P. 159.
107. Umer Khan, Lars Schmidt-Thieme, Alexandros Nanopoulos Collaborative SVM classification in scale-free peer-to-peer networks / Expert Systems with Applications. – 2017. – Vol. 69. – P. 74–86.

108. Vijayarani S., Dhayanand S. Liver disease prediction using SVM and Naïve Bayes algorithms // International Journal of Science, Engineering and Technology Research (IJSETR). – 2015. – T. 4. – №. 4. – P. 816–820.
109. Witten I.H., Frank E. Data mining: practical machine learning tools and techniques. SF: Morgan Kaufmann Publ., 2005. – 525 p.
110. Wyner A. J. et al. Explaining the success of adaboost and random forests as interpolating classifiers // The Journal of Machine Learning Research. – 2017. – T. 18. – №. 1. – P. 1558-1590.
111. Zhukov, D.A., Klyachkin, V.N., Krasheninnikov, 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. 19–25.

ПРИЛОЖЕНИЯ.

Приложение 1. Акт о внедрении

Закрытое акционерное общество
«Системы водоочистки»

Ул. Гончарова, 32а, г. Ульяновск, 432063
т/ф (8422) 65-50-82
ИНН 7325071536, КПП 732501001, ОГРН 1077325007561
http://www.ochistka-voda.ru/
E-mail:ecovita@nm.ru

УТВЕРЖДАЮ
Генеральный директор
ЗАО Системы водоочистки
Э.Е. Булыжев
_____
_____ г.

А К Т
о внедрении результатов
кандидатской диссертационной работы

Комиссия в составе:
председатель – Булыжев Евгений Михайлович, д.т.н., генеральный конструктор ЗАО «Системы водоочистки»,
члены комиссии:
Рябов Георгий Константинович – к.т.н., доцент, главный специалист по очистке жидкостей ЗАО «Системы водоочистки»;
Аксененко Павел Владимирович – ведущий инженер-конструктор ЗАО «Системы водоочистки»,

составили настоящий акт о том, что результаты диссертационной работы Жукова Дмитрия Анатольевича «Разработка моделей, алгоритмов и программ диагностики функционирования технических объектов с использованием агрегированных классификаторов», представленной на соискание ученой степени кандидата технических наук, а именно: модели, алгоритмы и программный комплекс диагностики и прогнозирования состояния водоисточника и показателей питьевой воды на основе агрегированных классификаторов внедрены в ЗАО «Системы водоочистки» при разработке технологии своевременного предупреждения о нарушении исправности системы водоочистки на станции очистки природной поверхностной воды и подготовки питьевой воды в г. Санкт-Петербурге.

Использование указанных результатов позволяет:
- повысить точность диагностики функционирования станции водоочистки до 15% за счет применения агрегированных классификаторов при машинном обучении и учета факторов, оказывающих влияние на качество бинарной классификации,
- обеспечить своевременное оперативное обновление моделей диагностирования путем использования псевдоградиентного алгоритма,
- автоматизировать процесс обнаружения нарушений в работе станции водоочистки путем использования разработанного программного комплекса с целью своевременной корректировки доз реагентов, тем самым достичь экономического эффекта в размере 516 тыс. руб. в год.

Председатель комиссии _____ Е.М. Булыжев
Члены комиссии: _____ Г.К. Рябов
_____ П.В. Аксененко

Приложение 2. Справка о внедрении в учебный процесс

УТВЕРЖДАЮ»
Первый проректор УлГТУ,
проректор по учебной работе
Е.В. Суркова

« 14 » октября 2019 г.

СПРАВКА

о внедрении в учебный процесс

Ульяновского государственного технического университета

результатов диссертационной работы Д.А. Жукова

Результаты диссертации Жукова Дмитрия Анатольевича «Разработка моделей, алгоритмов и программ диагностики функционирования технических объектов с использованием агрегированных классификаторов», представленной на соискание ученой степени кандидата технических наук, а именно математические модели и алгоритмы диагностики состояния технических объектов, рассматриваемые при изучении дисциплин «Теория надёжности», «Статистический контроль и управление процессами», «Статистические методы прогнозирования», читаемых студентам, обучающимся в бакалавриате и магистратуре по направлению «Прикладная математика», а также «Статистические методы в управлении качеством» по направлению «Управление качеством», и разработанное программное обеспечение, используемое в лабораторном практикуме, внедрены в учебный процесс Ульяновского государственного технического университета.

Использование указанных результатов позволило повысить эффективность обучения студентов за счет усвоения современных математических методов и компьютерных технологий диагностики функционирования сложных технических систем.

Зав. кафедрой
«Прикладная математика и информатика»
д-р техн. наук, профессор В.Р. Крашенинников

Приложение 3. Свидетельство о государственной регистрации программы для ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019611562

Оценка исправности технического объекта с применением машинного обучения

Правообладатель: ***федеральное государственное бюджетное образовательное учреждение высшего образования «Ульяновский государственный технический университет» (RU)***

Авторы: ***Клячкин Владимир Николаевич (RU), Кувайскова Юлия Евгеньевна (RU), Жуков Дмитрий Анатольевич (RU)***

Заявка № **2019610219**
Дата поступления **10 января 2019 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***29 января 2019 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Приложение 4. Свидетельство о государственной регистрации программы для ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019611560

Диагностика состояния технического объекта с использованием агрегированных классификаторов

Правообладатель: ***федеральное государственное бюджетное образовательное учреждение высшего образования «Ульяновский государственный технический университет» (RU)***

Авторы: ***Клячкин Владимир Николаевич (RU), Жуков Дмитрий Анатольевич (RU)***

Заявка № **2019610217**
Дата поступления **10 января 2019 г.**
Дата государственной регистрации в Реестре программ для ЭВМ ***29 января 2019 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев