УДК 519.248:658.518.5

DOI: 10.15827/0236-235X.126.244-250

Дата подачи статьи: 21.11.18

2019. Т. 32. № 2. С. 244–250

Прогнозирование состояния технического объекта с применением методов машинного обучения

*В.Н. Клячкин¹, д.т.н., профессор, v_kl@mail.ru**Д.А. Жуков¹, аспирант, zh.dimka17@mail.ru*

*¹ Ульяновский государственный технический университет,
кафедра «Прикладная математика и информатика», г. Ульяновск, 432027, Россия*

Распознавание состояния технического объекта во время его функционирования обеспечивает раннее обнаружение неисправностей и их устранение в процессе обслуживания. Часто диагностика сводится к разделению состояний объекта на два класса: исправное и неисправное. При решении такой задачи могут быть использованы методы машинного обучения, предназначенные для бинарной классификации.

В данной статье в качестве исходных данных рассматриваются известные результаты (прецеденты) оценки состояния системы: при заданных значениях контролируемых показателей техническая система исправна или неисправна. Используется множество различных подходов к бинарной классификации: классические статистические модели, методы, специально ориентированные на машинное обучение, композиционные методы и другие. Для повышения качества прогнозирования может быть использован агрегированный подход – комбинация нескольких методов классификации.

Разработанная в среде Matlab программа обеспечивает прогнозирование состояния системы по заданным показателям ее функционирования. Пользователь имеет возможность выбрать объем контрольной выборки, метод обучения, критерии качества распознавания.

Было проведено численное исследование на двух примерах. Оценивалась исправность гидроагрегата по критерию стабильности вибраций по результатам мониторинга показаний датчиков, установленных в различных точках. Наилучшим оказался агрегированный классификатор, включающий градиентный бустинг и логистическую регрессию. При анализе исправности системы водоочистки по показателям качества питьевой воды максимальное значение F -критерия имело место при агрегировании нейронной сети и бэггинга деревьев решений.

Ключевые слова: техническая диагностика, бинарная классификация, агрегированный подход, Matlab, гидроагрегат, система водоочистки, F -критерий.

Обеспечение безопасности и надежности сложных и дорогостоящих технических систем обуславливает необходимость проведения диагностики во время их функционирования. Это дает возможность как можно раньше обнаружить неисправности и устранить их в процессе обслуживания. Например, состояние двигателя диагностируется по расходу топлива, температуре газов, уровню шума и вибрации, составу выпускных газов, зазору между цилиндром и поршнем, зазору между шейками коленчатого вала и подшипниками и по другим показателям [1]. Имея числовые значения этого набора показателей, необходимо оценить, исправен двигатель или необходима его остановка для обслуживания. При этом имеется риск ложной тревоги (когда исправный объект будет признан неисправным) или, наоборот, пропуска цели, при котором неисправный объект считается исправным.

Часто диагностика сводится к разделению состояний объекта на два класса: исправное и

неисправное. При решении такой задачи могут быть использованы методы машинного обучения, или 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. Эту выборку разбивают на две части: обучающую и контрольную. Обучающая часть предназначена для построения моделей, с помощью которых объекты разделяются на исправные и неисправные.

Предполагается, что существует некоторая зависимость между показателями функционирования объекта и его состояниями. На основе исходных данных требуется восстановить эту

зависимость, то есть построить алгоритм, способный для заданного набора показателей функционирования объекта выдать достаточно точный ответ о его состоянии. Качество классификации с помощью полученных моделей оценивается по контрольной выборке.

Постановка задачи

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, $i = 1, \dots, l$, $j = 1, \dots, p$ (l – количество строк, или число наблюдений, p – количество столбцов, или число показателей), и вектора-столбца ответов Y , состоящего из единиц для тех опытов, в которых объект исправен, и нулей при неисправном объекте. Каждой строке x_i матрицы X соответствует определенное значение y_i вектора Y . Совокупность пар (x_i, y_i) образует выборку исходных данных – прецедентов.

Задача состоит в построении модели $a(x, w)$, которая предскажет ответ Y для любого заданного X [2–4]. Обычно используются линейные модели:

$$a(x, w) = w_0 + w_1x_1 + \dots + w_px_p, \quad (1)$$

где $w = (w_0 \ w_1 \ \dots \ w_p)$ – вектор параметров модели. В задачах бинарной классификации часто вместо нуля и единицы используют множество ответов $Y = \{-1; +1\}$. В этом случае модель алгоритма примет вид:

$$a(x, w) = \text{sign} \sum_{j=0}^p w_j x_j \quad (x_0 = 1). \quad (2)$$

Параметры w_j подбираются по исходным данным; процесс подбора параметров называется обучением алгоритма. Найденные параметры должны обеспечить оптимальное значение некоторого функционала качества. Часто минимизируется функционал ошибок (это среднее количество несовпадений фактического состояния i -го объекта y_i и прогнозируемого $a(x_i)$ по модели (2)):

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^l L(a, x_i) = \frac{1}{l} \sum_{i=1}^l [a(x_i) - y_i] \rightarrow \min. \quad (3)$$

Здесь $L(a, x_i)$ называют функцией потерь, она фиксирует наличие несовпадения опытного значения состояния объекта для заданного множества показателей функционирования x_i (строки матрицы X) со значением, прогнозируемым по построенному алгоритму $a(x_i)$.

Используются и другие функционалы для оценки качества классификации. В частности,

F -критерий является гармоническим средним точности и полноты [5]:

$$F = 2PR/(P + R), \quad (4)$$

где точность

$$P = tp/(tp + fp) \quad (5)$$

и полнота

$$R = tp/(tp + fn) \quad (6)$$

оцениваются по количеству правильно классифицированных исправных состояний tp , количеству неправильно классифицированных исправных состояний fp и количеству неправильно классифицированных неисправных состояний fn . F -критерий в отличие от функционала ошибок объективно оценивает качество классификации при несбалансированных классах (преобладание количества значений одного из классов над другим). Эта ситуация характерна для технических систем: неисправных состояний в исходной выборке обычно значительно меньше исправных.

Еще одним функционалом качества может быть площадь под ROC-кривой (receiver operating characteristics) – AUC (area under the curve) [6]. ROC-кривая образуется, если по оси абсцисс отложить значения $fp(c)$, а по оси ординат $tp(c)$, где c – некоторый порог. Площадь под ROC-кривой позволяет оценить модель в целом, не привязываясь к конкретному порогу. Критерий AUC, как и F -критерий, 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.

Методы машинного обучения

Данные методы применяются в самых разных областях деятельности. Используется множество различных подходов к классификации, в частности, к бинарной. Это и классические статистические модели (наивный байесовский классификатор, дискриминантный анализ, логистическая регрессия и другие) [1–4], и методы, специально ориентированные на машинное обучение (например, нейронные сети, метод опорных векторов), композиционные методы (бэггинг, бустинг в различных вариантах [7–10]) и другие.

Проблема состоит в том, что нельзя заранее определить, какой из выбранных методов обеспечит решение задачи с необходимой точностью, поэтому часто используются различные методы или их комбинации, а решение о применении принимается по результатам исследо-

вания функционала качества для контрольной выборки. В статье [11] для повышения качества прогнозирования предложен агрегированный подход – использование комбинации нескольких методов классификации. Эти результаты были подтверждены экспериментально и для задач технической диагностики [12–14].

В агрегированном подходе в отличие от композиционных методов совместно используются различные методы классификации, построенные на обучающей выборке. Для достижения наилучшего результата используется полный перебор наборов из всех H используемых базовых методов. Тогда, например, при $H = 2$ получим три набора: два базовых и один агрегированный; при $H = 3$ наборов уже семь: три базовых, три агрегированных по два базовых и один агрегированный из всех трех базовых методов. Нетрудно заметить, что в общем случае число наборов равно $2^H - 1$. Для формирования единого решения об исправности объекта рассмотрим агрегирование результатов по среднему значению, по медиане и с помощью процедуры голосования.

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

$$\hat{P}_{AK \text{ ср.}}(X_r) = \frac{\sum_{K=1}^H \hat{P}_K(X_r)}{H}, \quad (7)$$

где $\hat{P}_{AK \text{ ср.}}(X_r)$ – вероятность того, что r -й объект исправен.

При агрегировании по медиане вначале следует ранжировать ряд, содержащий результаты базовых методов в наборе. При нечетном числе базовых методов вероятность того, что r -й объект исправен:

$$\hat{P}_{AK \text{ мед.}}(X_r) = \hat{P}_{\frac{H+1}{2}}(X_r). \quad (8)$$

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

Результат агрегированного метода классификации по голосованию представляет собой среднее значение результатов базовых методов, которые определили исправность объекта с вероятностью, например, не ниже 0,1:

$$\hat{P}_K(X_r) \geq 0,1, \quad (9)$$

в противном случае вероятность того, что r -й объект исправен, равняется нулю.

Для оценки качества полученной модели с использованием кросс-валидации исходная выборка разбивается на N частей: $(N - 1)$ часть ис-

пользуется для обучения, одна – для контроля. Последовательно перебираются все варианты. Для каждого разбиения вычисляется критерий качества (3)–(6).

Программа диагностики состояния технического объекта методами машинного обучения

Практическая реализация методов машинного обучения возможна на базе библиотеки инструментов Statistics and Machine Learning Toolbox в пакете Matlab. 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 [12] показано, что, варьируя объем контрольной выборки, можно существенно повысить качество классификации).

Файл исходных данных представляет собой таблицу, 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 (рис. 1) 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.

После нажатия кнопки *Методы* открывается окно с перечнем используемых методов в левой части окна и формой для вывода результатов в правой (рис. 2). Устанавливается порог, определяющий, при каких значениях вероятности того, что объект исправен, его следует относить к действительно исправным (по умолчанию при $p > 0,5$ $y = 1$, в противном случае $y = 0$). Пользователь выбирает интересующие его методы классификации (11 базовых и 3 агрегированных (7)–(9)). По мере нажатия кнопок с выбранным методом в правой части окна выводятся характеристики качества классификации. Кроме F -критерия и значения AUC, можно вывести процент ошибок и дисперсию по кросс-валидации, процент ошибок по контрольной выборке, значения точности и полноты.

Рис. 1. Загрузка файла данных и разделение выборки на обучающую и контрольную (тестовую)

Fig. 1. Data file download and sample division into training and control (test)

Рис. 2. Вывод результатов расчета

Fig. 2. The output of the calculation results

Пользователь выбирает метод машинного обучения, который в наилучшей степени соответствует поставленной задаче (например, по максимуму F -критерия (рис. 2) выбирается AM-C = МОВ + GrB + БДР + LB: агрегированный метод по среднему значению, включающий сочетание метода опорных векторов, градиентного бустинга, бэггинга деревьев решений и LogitBoost).

Для диагностики состояния объекта при новых значениях контролируемых показателей эти значения вводятся из файла, программа рассчитывает прогнозируемую исправность объекта по выбранному методу машинного обучения.

На рисунке 3 показана блок-схема программы.

Разработанный программный продукт имеет следующие характеристики: тип операционной системы – Windows7 и выше; среда разработки – MatlabR2016a; размер программы – 152 Кб.

Численное исследование

Оценивалась исправность гидроагрегата по критерию стабильности вибраций [15] на основании результатов мониторинга показаний 10 датчиков, установленных в различных точках. Исходная выборка включала результаты

5 000 наблюдений. Наилучшим оказался агрегированный классификатор по медиане, включающий градиентный бустинг и логистическую регрессию, при этом объем контрольной выборки был равен 20 %, значение F -критерия составило 0,904.

В другом опыте при анализе исправности системы водоочистки по восьми показателям качества питьевой воды использовались результаты 1 557 наблюдений (исправное состояние в 1 204 случаях). Максимальное значение F -критерия 0,881 имело место при агрегировании нейронной сети и бэггинга деревьев реше-

ний, при этом объем контрольной выборки составил 10 %.

Заключение

Разработанная программа диагностики состояния технического объекта методами машинного обучения обеспечивает прогнозирование исправности системы по заданным показателям ее функционирования. Пользователь имеет возможность выбрать объем контрольной выборки, метод обучения, критерии качества распознавания. При этом необходимо

предварительно сформировать выборку исходных данных по результатам предшествующей работы объекта.

Исследование выполнено при финансовой поддержке РФФИ и Правительства Ульяновской области, грант № 18-48-730001.

Литература

1. Биргер И.А. Техническая диагностика. М.: Машиностроение, 1978. 240 с.
2. Witten I.H., Frank E. Data mining: practical machine learning tools and techniques. SF: Morgan Kaufmann Publ., 2005, 525 p.
3. Мерков А.Б. Распознавание образов. Введение в методы статистического обучения. М.: Едиториал УРСС, 2011. 256 с.
4. Воронина В.В., Михеев А.В., Ярушкина Н.Г., Святлов К.В. Теория и практика машинного обучения. Ульяновск: Изд-во УлГТУ, 2017. 290 с.
5. Соколов Е.А. Линейная классификация. URL: <https://github.com/esokolov/ml-course-hse/blob/master/2018-fall/lecture-notes/lecture04-linear-class.pdf> (дата обращения: 01.11.2018).
6. Дьяконов А.М. AUC ROC (площадь под кривой ошибок). URL: <https://dyakonov.org/2017/07/28/auc-gos-ploshad-pod-krivoy-oshibok/> (дата обращения: 01.11.2018).
7. Воронцов К.В. Машинное обучение. Композиция классификаторов. URL: <https://yadi.sk/i/FtIu6V0beBmF> (дата обращения: 01.11.2018).
8. Neykov M., Jun S. Liu, Tianxi Cai. On the characterization of a class of fisher-consistent loss functions and its application to boosting. JMLR, 2016, no. 17, pp. 1–32.
9. Wyner A.J., Olson M., Bleich J., Mease D. Explaining the success of adaboost and random forests as interpolating classifiers. JMLR, 2017, no. 18, pp. 1–33.
10. Chen T., Guestrin C. XGBoost: a scalable tree boosting system. Proc. 22nd ACM SIGKDD, 2016, pp. 765–794.
11. Клячкин В.Н., Кувайскова Ю.Е., Жуков Д.А. Диагностика технического состояния аппаратуры с использованием агрегированных классификаторов // Радиотехника. 2018. № 6. С. 46–49.
12. Жуков Д.А., Клячкин В.Н. Влияние объема контрольной выборки на качество диагностики состояния технического объекта // Автоматизация процессов управления. 2018. № 2. С. 90–95.
13. Repp P.V. The system of technical diagnostics of the industrial safety information network. JP: CS, 2017, vol. 803, art. 012127.
14. Kiselev M.I., Pronyakin V.I. and Tulekbaeva A.K. Technical diagnostics functioning machines and Mechanisms. IOP Conf. Ser.: MSE, 2018, vol. 312. DOI: 10.1088/1757-899X/312/1/012012.
15. Клячкин В.Н., Кувайскова Ю.Е., Иванова А.В. Система статистического анализа и контроля стабильности вибраций гидроагрегата // Программные продукты и системы. 2018. № 3. С. 600–625. DOI: 10.15827/0236-235X.123.620-625.

Software & Systems
DOI: 10.15827/0236-235X.126.244-250

Received 21.11.18
2019, vol. 32, no. 2, pp. 244–250

Forecasting the state of a technical object using machine learning methods

V.N. Klyachkin¹, Dr.Sc. (Engineering), Professor, v_kl@mail.ru

D.A. Zhukov¹, Postgraduate Student, zh.dimka17@mail.ru

¹ Ulyanovsk State Technical University, Ulyanovsk, 432027, Russian Federation

Abstract. State identification of a technical object during its operation enables early detection of malfunction and in-service repair. The diagnostics is frequently confined to splitting object states into two classes: a healthy and failure state. When solving this problem, it is possible to use machine learning methods for binary classification.

The basic data in this paper are the known results (precedents) of a system state evaluation: the technical system is nonfaulty or faulty with predetermined values of monitored indicators. There are many different approaches to binary classification. They are classical statistical models, methods focusing on machine learn-

ing, composite methods and others. In order to improve quality of forecasting, it is appropriate to use an aggregated approach that is a combination of several classification methods.

The program developed in Matlab allows forecasting a system state by its predetermined operation indicators. The user may select a validation set volume, a learning method, and recognition quality criteria.

The authors have conducted a numerical study on two examples. The evaluation of a hydraulic unit good condition has taken into account a vibration stability criterion according to the results of monitoring sensors installed in various places. The aggregated classifier which includes gradient boosting and logistic regression showed the best result. In analysis of water treatment system in respect to drinking water quality, the maximum F -criterion value was when aggregating a neural network and bagging of decision trees.

Keywords: technical diagnostics, binary classification, aggregated approach, Matlab, hydraulic unit, water treatment system, F -criterion.

Acknowledgements. *The work has been supported by the Russian Foundation for basic research and the Government of the Ulyanovsk region, grant no. 18-48-730001.*

References

1. Birger I.A. *Technical Diagnostics*. Moscow, Mashinostroenie Publ., 1978, 240 p.
2. Witten I.H., Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco, Morgan Kaufmann Publ., 2005, 525 p.
3. Merkov A.B. *Pattern Recognition. Introduction to Statistical Training Methods*. Moscow, Editorial URSS Publ., 2011, 256 p.
4. Voronina V.V., Mikheev A.V., Yarushkina N.G., Svyatov K.V. *Machine Learning Theory and Practice*. Ulyanovsk, UIGTU Publ., 2017, 290 p.
5. Sokolov E.A. *Linear Classification*. Available at: <https://github.com/esokolov/ml-course-hse/blob/master/2018-fall/lecture-notes/lecture04-linclass.pdf> (accessed November 1, 2018).
6. Dyakonov A.M. *AUC ROC*. Available at: <https://dyakonov.org/2017/07/28/auc-roc-ploshchad'-pod-krivoj-oshibok/#more-5362> (accessed November 1, 2018).
7. Vorontsov K.V. *Machine Learning. Classifier Composition*. Available at: <https://yadi.sk/i/FItlu6V0beBmF> (accessed November 1, 2018).
8. 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, pp. 1–32.
9. Wyner A.J., Olson M., Bleich J., Mease D. Explaining the success of AdaBoost and random forests as interpolating classifiers. *J. of Machine Learning Research*. 2017, no. 18, pp. 1–33.
10. Chen T., Guestrin C. XGBoost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining*. 2016, pp. 765–794.
11. Klyachkin V.N., Kuvayskova Yu.E., Zhukov D.A. Diagnostics of technical state of the equipment using aggregated classifiers. *J. Radioengineering*. 2018, no. 6, pp. 46–49 (in Russ.).
12. Zhukov D.A., Klyachkin V.N. The Effect of the Control Sample Volume on the Quality of Diagnostics of the Technical Object State. *Automation of Control Processes*. 2018, no. 2, pp. 90–95 (in Russ.).
13. Repp P.V. The system of technical diagnostics of the industrial safety information network. *J. of Physics: Conf. Series*. 2017, vol. 803m, art. 012127.
14. Kiselev M.I., Pronyakin V.I., Tulekbaeva A.K. Technical diagnostics functioning machines and Mechanisms. *IOP Conf. Series: Materials Science and Engineering*. 2018, vol. 312. DOI: 10.1088/1757-899X/312/1/012012.
15. Klyachkin V.N., Kuvayskova Yu.E., Ivanova A.V. The system of statistical analysis and control of hydraulic unit vibration stability. *Software & Systems*. 2018, no. 3, pp. 620–625. DOI: 10.15827/0236-235X.123.620-625.