УДК 519.248:681.518.5

Д.А. Жуков, В.Н. Клячкин

ВЛИЯНИЕ ОБЪЕМА КОНТРОЛЬНОЙ ВЫБОРКИ НА КАЧЕСТВО ДИАГНОСТИКИ СОСТОЯНИЯ ТЕХНИЧЕСКОГО ОБЪЕКТА[1]

Жуков Дмитрий Анатольевич, *окончил факультет информационных систем и технологий Ульяновского государственного технического университета, аспирант кафедры «Прикладная математика и информатика» УлГТУ. Имеет научные труды в области статистических методов и машинного обучения. [e-mail: zh.dimka17@mail.ru].*

Клячкин Владимир Николаевич, *доктор технических наук, профессор, окончил механический факультет Ульяновского политехнического института. Профессор кафедры «Прикладная математика и информатика» УлГТУ. Имеет научные труды в области надежности и статистических методов. [e-mail: v_kl@mail.ru].*

Аннотация

Рассматривается задача прогнозирования исправности технического объекта по известным показателям его функционирования. Исходными данными являются известные результаты оценки состояния объекта по информации о предшествующей эксплуатации: при заданных значениях контролируемых показателей техническая система исправна или неисправна. Такая задача может быть решена методами машинного обучения, она сводится к бинарной классификации состояний объекта. Качество диагностики может существенно зависеть от множества факторов: метода обучения, правильного выделения факторов, характеризующих работу объекта, объема выборки и других. В работе проводится исследование влияния объема контрольной выборки на качество диагностики, оцениваемое по количеству неверно спрогнозированных состояний методом кросс-валидации. Испытания проводились в пакете Matlab, использовано десять различных методов машинного обучения: логистическая регрессия, метод опорных векторов, бэггинг деревьев решений и другие. Показано, что при правильном выборе доли контрольной выборки можно повысить качество диагностики на 5–7%.

Ключевые слова: техническая диагностика, исправность, показатели функционирования, машинное обучение, контрольная выборка, кросс-валидация.

THE EFFECT OF THE CONTROL SAMPLE VOLUME ON THE QUALITY OF DIAGNOSTICS OF THE TECHNICAL OBJECT STATE

Dmitrii Anatolevich Zhukov, *graduated from the Faculty of Information System and Technologies of Ulyanovsk State Technical University; Postgraduate Student of the Department of Applied Mathematics and Computer Science; an author of proceedings in the field of the statistical methods and machine learning. e-mail: zh.dimka17@mail.ru.*

Vladimir Nikolaevich Kliachkin, *Doctor of Engineering; graduated from the Mechanical Faculty of Ulyanovsk Polytechnic Institute; Professor at the Department of Applied Mathematics and Informatics of Ulyanovsk State Technical University; an author of scientific papers in the field of reliability issues and statistical methods. e-mail: v_kl@mail.ru.*

Abstract

The problem of predicting the serviceability of a technical object in terms of its performance is considered. The source data are the known results of the evaluation of the state of an object based on the results of the previous operation: If the specified values of controlled indicators technical system intact or defective. Such a problem can be solved by methods of machine learning, it reduces to a binary classification of the states of the object. The quality of diagnostics can significantly depend on many factors: the method of training, the correct allocation of factors characterizing the operation of the object, the volume of the sample, and others. The work studies the effect of the control sample volume on the quality of diagnosis, estimated by the number of mis-predicted states using the cross-validation method. The tests were carried out

1 Исследование выполнено при финансовой поддержке РФФИ, проект № 16-48-732002.

in the Matlab package, ten different training methods were used: logistic regression, support vector method, decision tree bugging, and others. It is shown that the correct choice of the proportion of the control sample can improve the diagnostic quality to 5–7%.

Key words: technical diagnostics, serviceability of the indicator of functioning, machine learning, control sample, cross-validation.

Введение

Диагностика состояния технического объекта проводится с целью повышения его надежности. Решение задачи диагностики сводится к распознаванию состояния объекта: как правило, к разделению состояний объекта на исправные, то есть способные выполнять заданные функции, или неисправные [1].

Часто диагностика проводится в процессе эксплуатации объекта по результатам измерений косвенных показателей его функционирования. При этом имеется риск ложной тревоги (когда исправный объект будет признан неисправным) или наоборот – пропуска цели, при котором неисправный объект считается исправным.

Исходными данными являются известные результаты оценки состояния объекта по результатам предшествующей эксплуатации: при заданных значениях контролируемых показателей техническая система исправна или неисправна. Например, исправность 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 другим показателям.

Предполагается, что существует некоторая неизвестная 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. Это задача машинного обучения, или обучения по прецедентам (с учителем): частным случаем этой задачи является бинарная классификация, т. е. разделение состояний объекта на два класса [2–4].

Для оценки качества построенного алгоритма с точки зрения возможности прогнозирования исходную выборку разбивают на два непересекающихся подмножества. Первое подмножество – это собственно обучающая выборка, с помощью которой и решается задача обучения (которая, как правило, сводится к оценке параметров модели соответствующего алгоритма). Второе подмножество является контрольной (или тестовой) выборкой, не используемой для обучения. По этой части выборки оценивается ошибка прогнозирования, которая и характеризует качество обучения. Соотношение объемов обучающей и контрольной выборок может быть различным.

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

1 Диагностика состояния технического объекта с использованием методов машинного обучения

Исходные данные для диагностики состояния технического объекта представляются в виде матрицы X показателей функционирования системы, элементы которой x_{ij} – результат i-го наблюдения по j-му показателю; $i = 1, \ldots, l, \quad j = 1, \ldots, p$ (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. 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 функции (модели алгоритма) $a : X \to Y$, которая предскажет ответ Y для любого заданного X [5]. Обычно используются линейные модели:

$$a(x, w) = w_0 + w_1 x_1 + \ldots + w_p x_p, \quad (1)$$

где $w = (w_0\ w_1\ \ldots\ w_p)$ – 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:

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

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

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

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 $a(x_i)$.

Качество классификации можно оценить и по дисперсии ошибок – среднему квадрату отклонения истинной вероятности $P(Y_i)$ исправности объекта для i-го объекта контрольной выборки от ее прогнозируемого значения $\tilde{P}(Y_i)$:

$$\sigma^2 = \frac{1}{l}\sum_{i=1}^{l}\left[P(Y)_i - \tilde{P}(Y_i)\right]^2 \quad (4)$$

(для исправного объекта $P(Y_i)=1$, для неисправного $P(Y_i)=0$).

Методы машинного обучения активно используются в самых разных областях деятельности. Используется множество различных подходов к классификации. Это и классические статистические методы [6, 7] (байесовские классификаторы, дискриминантный анализ, логистическая регрессия) и методы, специально ориентированные на машинное обучение [4, 5] (метод опорных векторов, нейронные сети), композиционные методы (бэггинг, бустинг), агрегированный подход [8, 9] и другие.

Нельзя заранее сказать, какой из выбранных методов обеспечит корректное решение задачи, поэтому часто используются множество различных методов или их комбинации, а решение о применении для выполнения конкретной технической задачи принимается по результатам исследования функционала качества (3) для контрольной выборки.

Например, как один из методов бинарной классификации может использоваться метод опорных векторов (SupportVectorMachine – SVM), основанный на работах по построению оптимальной разделяющей гиперплоскости (эти работы проводились группой математиков в Институте проблем управления Академии наук СССР под руководством В.Н. Вапника в 60–70 годах прошлого века [5]). Оптимальность состояла в максимальном удалении разделяемых объектов от разделяющей поверхности. Параллельные границы гиперплоскости, разделяющей классы, описываются уравнениями:

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

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

с вектором нормали w (p – количество показателей, характеризующих условия функционирования объекта). Это задача квадратичного программирования: минимизировать норму вектора w при наличии l ограничений (l – количество наблюдений в обучающей выборке):

$$\|w\|^2 \to \min,$$

$$y_i\left(\sum_{j=1}^{p} w_j x_j - w_0\right) \geq 1;$$

здесь $i = 1,\ldots, l$. Задача решается методом множителей Лагранжа и имеет единственное решение.

Используется также композиция алгоритмов [5], при этом погрешности отдельных алгоритмов взаимно компенсируются. Опыт показывает, что два основных метода построения композиции – бэггинг и бустинг – часто дают более точный результат, чем применение отдельного алгоритма на конкретном наборе данных.

При наличии небольших обучающих выборок используют бэггинг: из имеющейся выборки исходных данных случайным образом с возвратом формируется несколько подмножеств такого же объема, как и исходная выборка. При этом некоторые объекты попадут в эти подмножества по несколько раз, а некоторые не попадут вообще (бутстреп-выборки). На основе каждого подмножества строится классификатор, и результаты комбинируются путем голосования или усреднения. В качестве базовых алгоритмов часто используются деревья решений: в отличие от других методов классификации, деревья решений не сводятся к построению функциональной зависимости, а последовательно разделяют данные на классы. При этом на первом шаге построения дерева разделение производится по самому значимому фактору.

В бустинге итоговое правило также строится путем взвешенного голосования композиции базовых правил. При этом используется информация об ошибках предыдущих правил: веса объектов выбираются таким образом, чтобы новое правило точнее работало на тех объектах, на которых предыдущие правила чаще ошибались. В различных модификациях используются разные аппроксимации функции потерь в формуле (3): в наиболее распространенном методе AdaBoost – экспонента, в LogitBoost (этот метод используется при наличии шумовых данных) – аппроксимация, основанная на логистической регрессии, и другие.

2 Оценка качества диагностики

Алгоритм a, который минимизирует функционал (3), может не обеспечивать хорошее прогнозирование исправности объекта. Ситуация, когда качество работы алгоритма на новых объектах значительно хуже, чем на исходной выборке, свидетельствует о переобучении: алгоритм слишком хорошо подогнан под обучающую выборку и не способен к обобщению на другие выборки. Таким образом, построенный алгоритм не сможет предсказывать состояние исследуемого объекта при новых параметрах функционирования.

Для оценки качества модели с точки зрения возможности прогнозирования исходную выборку из l опытов разбивают на два непересекающихся подмножества: собственно обучающую выборку объемом l_o (с помощью которой и решается задача обучения (3)) и контрольную (или тестовую) выборку объемом $l_к = l - l_o$, не используемую для обучения.

При использовании кросс-валидации выборка разбивается на N частей (блоков). $(N-1)$ часть используется для обучения, а одна – для контроля. Последовательно перебираются все варианты: процесс повторяется N раз так, чтобы каждый из блоков использовался один раз как тестовый набор. Для каждого разбиения решается задача обучения по выборке l_o и вычисляется функция ошибок $Q(a, X)$ на контрольной выборке $l_к$. Среднее значение этой функции по всем вариантам разбиения

и характеризует обобщающую способность алгоритма, и по существу является оценкой качества диагностики технического объекта по рассматриваемому алгоритму.

3 Результаты испытаний

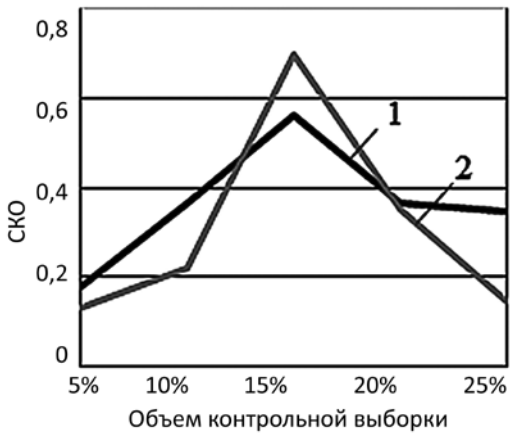
Для проведения испытаний использовались результаты наблюдений за технической системой, функционирование которой определяется восемью показателями [10–12]. Строились пять вариантов алгоритмов на основе базовых методов машинного обучения – наивного байесовского классификатора, дискриминантного анализа, логистической регрессии, метода опорных векторов и нейронных сетей, а также пять – на основе композиционных методов: бэггинга деревьев решений и различных вариантов бустинга: AdaBoost, LogitBoost, GentleBoost, RUSBoost. Модули для реализации всех этих методов встроены в пакет Matlab, в котором и проводились испытания.

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 1 приведены усредненные результаты: процент ошибок по каждому из используемых методов при каждом значении объема контрольной выборки. Графическое представление полученных результатов показано на рисунке 1, а.

Видно, что лучшие результаты (минимальный усредненный процент ошибок на контрольной выборке по результатам кросс-валидации) показали методы бустинга LogitBoost и GentleBoost – от 8,29% до 9,87%.

Также был произведен расчет рассеяния значений процента ошибок, как среднеквадратичное отклонение по результатам пяти испытаний при заданном объеме контрольной выборки (рис. 1, б).

а) среднее значение



б) среднеквадратичное отклонение (СКО)

Рис. 1. Зависимость процента ошибок от объема контрольной выборки для методов бустинга (1 – LogitBoost, 2 – GentleBoost)

Таблица 1

Оценка качества диагностики

Метод	Объем контрольной выборки				
	5%	10%	15%	20%	25%
Логистическая регрессия	16,48	16,49	16,67	16,64	16,48
Дискриминантный анализ	18,32	18,46	18,36	18,45	18,66
Наивный байесовский классификатор	18,35	18,32	18,45	18,41	18,24
Нейронные сети	16,70	16,95	16,94	16,02	17,07
Метод опорных векторов	17,74	17,59	17,72	17,90	17,82
Бэггинг деревьев решений	17,15	17,63	17,71	18,04	18,26
AdaBoost	12,37	12,56	12,98	12,78	12,90
LogitBoost	9,60	9,60	9,68	9,85	9,87
GentleBoost	8,53	8,29	8,65	8,53	8,91
RUSBoost	21,32	21,28	21,01	20,82	20,83

Исследование показало неоднозначный характер влияния объема контрольной выборки на качество диагностики: для нейронной сети лучший результат (процент ошибок 16,02%) оказался при объеме контрольной выборки 20%, а худший (17,07%) – при объеме в 25%; для AdaBoost самый низкий процент ошибок выявлен при объеме контрольной выборки 5%, а самый высокий – при 15%. Для метода GentleBoost лучшим оказался объем контрольной выборки 10% (рис. 1), при этом снижение процента ошибок составило 7%.

Заключение

Проведенное исследование показало, что различные методы машинного обучения по-разному реагируют на изменение объема контрольной выборки. При этом за счет правильного выбора этой величины возможно повышение точности диагностики на 5–7%. В связи с этим можно рекомендовать на стадии выбора алгоритма для проведения технической диагностики конкретного объекта провести соответствующие испытания. Для рассмотренного объекта наилучшим вариантом оказался GentleBoost при объеме контрольной выборки 10% от исходной. Именно такой подход обеспечит прогноз технического состояния объекта с минимальной ошибкой.

СПИСОК ЛИТЕРАТУРЫ

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. 2nd ed. – San Francisco : Morgan Kaufmann Publishers, 2005. – 525 p.

3. Мерков А.Б. Распознавание образов. Введение в методы статистического обучения. – М. : Едиториал УРСС, 2011. – 256 с.

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

5. Воронцов К.В. Машинное обучение. Композиция классификаторов. – URL: https://yadi.sk/i/FItIu6V0beBmF.

6. Клячкин В.Н. Статистические методы в управлении качеством: компьютерные технологии. – М. : Финансы и статистика, ИНФРА-М, 2009. – 304 с.

7. Клячкин В.Н., Кувайскова Ю.Е., Алексеева В.А. Статистические методы анализа данных. – М. : Финансы и статистика, 2016. – 240 с.

8. Шунина Ю.С., Клячкин В.Н. Прогнозирование платежеспособности клиентов банка на основе методов машинного обучения и марковских цепей // Программные продукты и системы. – 2016. – № 2. – С. 105–112.

9. Клячкин В.Н., Кувайскова Ю.Е., Жуков Д.А. Использование агрегированных классификаторов при технической диагностике на базе машинного обучения // Информационные технологии и нанотехнологии (ИТНТ-2017) : сб. тр. III междунар. конф. и молодеж. школы / Самарский национальный исследовательский университет им. акад. С.П. Королева. – Самара, 2017. – С. 1770–1773.

10. Жуков Д.А., Клячкин В.Н. Задачи обеспечения эффективности машинного обучения при диагностике технических объектов // Современные проблемы проектирования, производства и эксплуатации радиотехнических систем. – 2016. – № 1 (10). – С. 172–174.

11. Жуков Д.А., Клячкин В.Н. Алгоритмы бустинга в задачах технической диагностики // Перспективные информационные технологии : тр. Междунар. науч.-техн. конф. – Самара : Издательство Самарского научного центра РАН, 2017. – С. 787–790.

12. Клячкин В.Н., Кравцов Ю.А., Жуков Д.А. Оценка эффективности диагностики состояния объекта по наличию неслучайных структур на карте Хотеллинга // Автоматизация процессов управления. – 2015. – № 1. – С. 50–56.

REFERENCES

1. Birger I.A. *Tekhnicheskaia diagnostika* [Engineering Diagnostics]. Moscow, Mashinostroenie Publ., 1978. 240 p.

2. Witten I.H., Frank E. *Data Mining: Practical Machine Learning Tools and Techniques.* 2nd Edition. San Francisco, Morgan Kaufmann Publishers, 2005. 525 p.

3. Merkov A.B. *Raspoznavanie obrazov. Vvedenie v metody statisticheskogo obucheniia* [Image Identification. Overview of Techniques of Statistical Learning]. Moscow, Editorial URSS Publ., 2011. 256 p.

4. Voronina V.V., Mikheev A.V., Yarushkina N.G., Sviatov K.V. *Teoriia i praktika mashinnogo obucheniia. Ucheb. posobie* [Theory and Practice of Machine Learning. Study Guide]. Ulyanovsk, UlSTU Publ., 2017. 290 p.

5. Vorontsov K.V. *Mashinnoe obuchenie. Kompozitsiia klassifikatorov* [Machine Learning. Classifier Ensemble]. Available at: https://yadi.sk/i/FItIu6V0beBmF.

6. Kliachkin V.N. *Statisticheskie metody v upravlenii kachestvom: kompiuternye tekhnologii* [Statistical Methods in Quality Management: Computer-Aided Technologies]. Moscow, Finansy i statistika Publ., INFRA-M, 2009. 304 p.

7. Kliachkin V.N., Kuvaiskova Iu.E., Alekseeva V.A. *Statisticheskie metody analiza dannykh* [Statistical Methods for Data Analysis]. Moscow, Finansy i statistika Publ., 2016. 240 p.

8. Shunina Iu.S., Kliachkin V.N. Prognozirovanie platezhesposobnosti klientov banka na osnove metodov mashinnogo obucheniia i markovskikh tsepei [Bank Clients' solvency Forecasting based on Machine Learning Methods ND Markov Chains]. *Programmnye produkty i sistemy* [Software and Systems], 2016, no. 2, pp. 105–112.

9. Kliachkin V.N., Kuvaiskova Iu.E., Zhukov D.A. Ispolzovanie agregirovannykh klassifikatorov pri tekhnicheskoi diagnostike na baze mashinnogo obucheniia [Aggregative Classifiers when Technical Diagnostics based on Machine Learning]. *Informatsionnye tekhnologii i nanotekhnologii (ITNT-2017): sb. tr. III mezhdunar. konf. i molodezh. shkoly* [Information Technologies and Nanotechnologies (ITNT-2017). Proc. of the Third Int.

Conference and Youth School]. Samara University Publ., Samara, 2017, pp. 1770–1773.

10. Zhukov D.A., Kliachkin V.N. Zadachi obespecheniia effektivnosti mashinnogo obucheniia pri diagnostike tekhnicheskikh obektov [Problems of Machine Learning Efficiency in Diagnostics of Technical Objects]. *Sovremennye problemy proektirovaniia, proizvodstva i ekspluatatsii radiotekhnicheskikh system* [Current Issues on Radio System Design, Production and Operation], 2016, no. 1 (10), pp. 172–174.

11. Zhukov D.A., Kliachkin V.N. Algoritmy bustinga v zadachakh tekhnicheskoi diagnostiki [Boosting Algorithms in Engineering Diagnostics Problems]. *Perspektivnye informatsionnye tekhnologii. Tr. Mezhdunar. nauch.-tekhn. konf.* [Advanced Information Technologies. Proc. of International Sci. and Tech. Conf.]. Samara, Izdatelstvo Samarskogo nauchnogo tsentra RAN Publ., 2017, pp. 787–790.

12. Kliachkin V.N., Kravtsov Iu.A., Zhukov D.A. Otsenka effektivnosti diagnostiki sostoianiia obekta po nalichiiu nesluchainykh struktur na karte Khotellinga [Evaluation of Object Status Diagnosing Efficiency to Non-Random Structures Existence on the Hotelling's Chart]. *Avtomatizatsiia protsessov upravleniia* [Automation of Control Processes], 2015, no. 1, pp. 50–56.