



Д.А. Жуков, В.Н. Клячкин

ДИАГНОСТИКИ ИСПРАВНОСТИ ТЕХНИЧЕСКОГО ОБЪЕКТА С ИСПОЛЬЗОВАНИЕМ ПАКЕТА MATLAB

(Ульяновский государственный технический университет)

Для распознавания исправности состояния технического объекта могут быть использованы методы машинного обучения, с помощью которых проводится бинарная классификация: объекты подразделяются на исправные и неисправные [1-3]. Практическая реализация этих методов удобна на базе библиотеки инструментов Statistics and Machine Learning Toolbox в пакете Matlab [4]. Matlab представляет собой высокоуровневый язык для технических расчетов, интерактивную среду разработки алгоритмов и мощный инструмент анализа данных.

Распознавание состояния объекта проводится по результатам измерений показателей функционирования. В качестве исходных данных рассматривается множество N прецедентов $(x^{(i)}, y^{(i)})$, $i = 1 \dots N$: объекты с заданными d показателями функционирования $x = (x_1, x_2, \dots, x_d)$ и соответствующими состояниями y , принимающими одно из двух значений $(0,1)$: $y = 1$ соответствует исправному состоянию объекта, $y = 0$ – неисправному.

На основе этих данных надо восстановить зависимость между показателями функционирования и состоянием объекта, то есть построить алгоритм, способный для любого набора показателей объекта выдать достаточно точный ответ о его состоянии. Это частный случай одной из задач машинного обучения – классификации при обучении по прецедентам.

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

Задача классификации решается по следующему принципу: объект считается исправным $Y = 1$, если модель спрогнозировала вероятность $P\{Y = 1 | X\} > 0,5$, и неисправным $Y = 0$ – в противном случае. При этом в качестве порога может быть использовано число, отличное от 0,5.

Качество классификации зависит от ряда факторов: объема и качества исходной выборки, метода машинного обучения, способа разделения исходной выборки на обучающую и контрольную части. Предварительно производится кодирование нечисловых показателей, обработка выбросов и пропущенных наблюдений.

Важным является выбор показателей, характеризующих состояние объекта [8-9]. Таких показателей обычно достаточно много, и их влияние на состоя-



ние объекта может существенно отличаться. Использование при классификации малозначимых показателей может усложнить расчет и привести к некорректным результатам.

Для оценки качества модели с точки зрения возможности прогнозирования исходную обучающую выборку разбивают на два непересекающихся подмножества: собственно обучающую выборку (с помощью которой и решается задача обучения) и контрольную (или тестовую), не используемую для обучения. При использовании кросс-валидации выборка разбивается на k частей. $(k - 1)$ часть используется для обучения, одна – для контроля. Последовательно перебираются все варианты. Для каждого разбиения решается задача обучения и вычисляется функция ошибок на контрольной выборке.

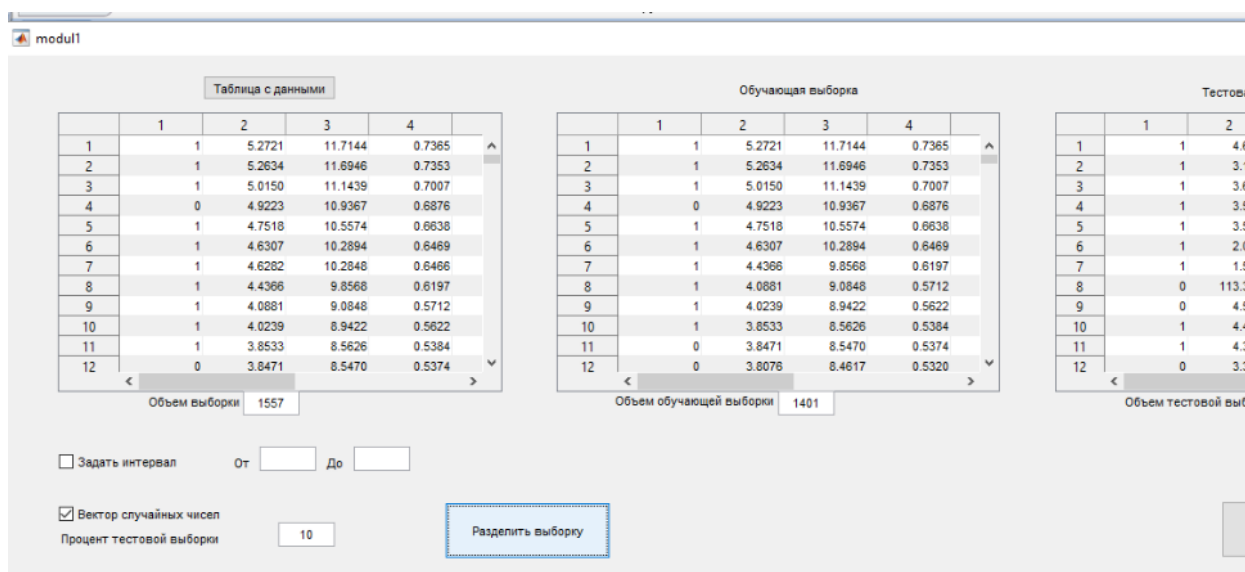


Рис. 1. Разделение выборки на обучающую и тестовую части

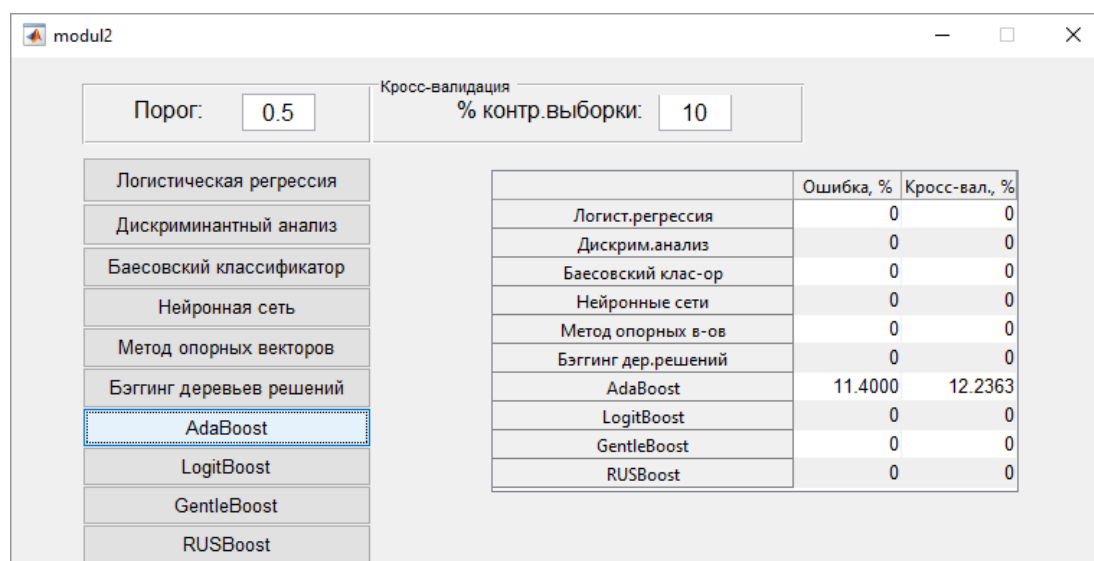


Рис. 2. Выбор метода обучения и результаты расчета



Разработанная в Matlab программа обеспечивает ввод исходных данных из файла электронной таблицы Excel (в первом столбце – значения Y , далее – значения отобранных показателей функционирования), позволяет заданным образом разделять выборку на обучающую и тестовую части: предусмотрен как случайный отбор, так и выбор определенного диапазона данных для тестирования (рис. 1).

Задается пороговое значение вероятности и объем выборки для кросс-валидации. Выбирается один или несколько методов машинного обучения, реализованных в Matlab. Результатом расчета является процент ошибок для выбранного метода в двух вариантах: по тестовой выборке и при кросс-валидации. На рис. 2 задано пороговое значение 0,5, объем контрольной выборки при кросс-валидации 10%, выбран один из встроенных методов бустинга - Adaboost и показаны полученные результаты: процент ошибок по тестовой выборке 11,4%, при кросс-валидации 12,2%.

Литература

1. Воронина В.В. Теория и практика машинного обучения : учебное пособие / В. В. Воронина, А. В. Михеев, Н. Г. Ярушкина, К. В. Святков. – Ульяновск : УлГТУ, 2017. – 290 с.
2. Witten, I. H. Data mining : practical machine learning tools and techniques / Ian H. Witten, Eibe Frank. – 2nd ed. – San Francisco: Morgan Kaufmann Publishers, 2005. – 525p.
3. Мерков, А. Б. Распознавание образов. Введение в методы статистического обучения / А.Б. Мерков. – М.: Едиториал УРСС, 2011. – 256 с.
4. 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.
5. Клячкин В.Н., Кувайскова Ю.Е., Алексеева В.А. Статистические методы анализа данных. – М.: Финансы и статистика, 2016. – 240 с.
6. Klyachkin V.N., Kuvayskova Yu.E., Zhukov D.A. The use of aggregate classifiers in technical diagnostics, based on machine learning // CEUR Workshop Proceedings, vol.1903. Data Science. Information Technology and Nanotechnology. – 2017. – P. 32-35.
7. Жуков Д.А., Клячкин В.Н. Алгоритмы бустинга в задачах технической диагностики // Перспективные информационные технологии (ПИТ 2017): труды Международной научно-технической конференции / под ред. С.А. Прохорова. – Самара: Издательство Самарского научного центра РАН, 2017. – С. 787-790.
8. Клячкин В.Н. Статистические методы в управлении качеством: компьютерные технологии. – М.: Финансы и статистика, ИНФРА-М, 2009. – 304 с.
9. 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.