сверточных нейронных сетей и методов глубокого обучения. Приведены результаты эксперимента на реальных данных, полученных со спутника Landsat-7. На этих данных система показала точность анализа примерно 85%. Дальнейшие исследования будут продолжены в области повышения точности работы системы анализа.

Система может быть использована для мониторинга вырубок и незаконного выжигания леса под сельскохозяйственные нужды, которые могут появиться в следствии действий людей; а также при изменениях лесного массива по природным причинам: ураганы, изменение климата, кислотные дожди, болезни леса, нашествие насекомых и т. д.

Список литературы

1. *Книжников Ю.Ф.* Аэрокосмические методы географических исследований / Книжников Ю.Ф., В.И. Кравцова, О.В. Тутубалина. / М.: «Академия», 2004. 336 с.
2. *Лурье И. К.* Геоинформационное картографирование. Методы геоинформатики и цифровой обработки космических снимков/ И. К. Лурье/ - М.: КДУ, 2008. 424 с.
3. *Vapnic V. Support-Vector Networks / Vapnic V., Cortex C. /* Mashine Learning 20. no. 3, 1995. 25 p.
4. *Vapnic V.A* Note on One Class of Perceptrons / Vapnic V., Chervonenkis A. // Automation and Remote Control. 1964. 25.
5. *Книжников Ю.Ф.* Аэрокосмические методы географических исследований: Учеб. для студ. вузов /Книжников Ю.Ф./ М.: Издательский центр «Академия», 2004. 336 с.
6. *Чандра А.М.* Дистанционное зондирование и географические информационные системы / Чандра А.М. Гош С.К. / М: Техносфера. 2008. 312 с.
7. *Лепский А.Е.* Математические методы распознавания образов /Лепский А.Е., Броневич А.Г./ Таганрог: ТТИ ЮФУ, 2009. 155 с.
8. *Гудфеллоу Я., Бенджио И., Курвилль А.* Глубокое обучение // Пер. с анг. А. А. Слинкина. – 2-е изд., испр. – М.:ДМК Пресс, 2018. 652 с.
9. *Осовский С.* Нейронные сети для обработки информации // Пер. с польск. И.Д. Рудинского. - 2-е изд., перераб. и доп. - М.: Горячая линия - Телеком, 2016. 448 с.
10. *Terrance DeVries, Graham W.* Taylor School of Engineering University of Guelph Guelph // ON N1G 2W1, Canada, 2017. 12 p.
11. *Olaf Ronneberger, Philipp Fischer, and Thomas Brox.* U-Net: Convolutional Networks for Biomedical Image Segmentation. // Computer Science Department and BIOSS Centre for Biological Signalling Studies, University of Freiburg, Germany, 2015. 8 p.

Голев Артём Владимирович *– инженер ИПУ РАН.*
E-mail:oiw23@mail.ru

DOI: 10.25728/avtprom.2021.01.09

Применение методов интеллектуального анализа эксплуатационных статистических данных для прогнозирования времени наработки на отказ промысловых трубопроводов

Д.П. Кармачев (АО «ТомскНИПИнефть»,
Национальный исследовательский Томский политехнический университет)

Представлены результаты исследования, направленного на создание экспертной системы выбора материального исполнения и способа внутренней антикоррозионной защиты промысловых трубопроводов в части применения различных интеллектуальных методов оценки статистических эксплуатационных данных об отказах промысловых трубопроводов. Приведены основные сведения об исходной выборке и проведенном разведочном анализе. Описаны процессы создания и оценки регрессоров, основанных на различных алгоритмах.

Ключевые слова: отказы промысловых трубопроводов, прогнозирование отказов, интеллектуальный анализ, случайный лес, градиентный бустинг, регрессионные модели.

Введение

Настоящая работа является продолжением исследований [1], направленных на создание прототипа программного обеспечения экспертной системы (далее «Система») выбора материального исполнения и способа внутренней антикоррозионной защиты промысловых трубопроводов (ПТ). Цель создания системы заключатся в снижении аварийности и повышении надежности ПТ на этапе эксплуатации, а также в повышении обоснованности выбора материального исполнения промысловых трубопроводов с учетом механизмов коррозионных процессов, протекающих на внутренней стенке трубопроводов. Целью текущего исследования является применение различных методов интеллектуального анализа статистических данных и выбор наиболее оптимального метода и достаточного числа признаков для последующего применения в модели представления знаний системы, предназначенной для определения времени наработки на отказ конкретного проектируемого участка ПТ.

Исходная выборка содержит 109989 примеров отказов и охватывает 641 месторождение. Сводная информация об исходной выборке в разрезе причин и природы возникновения отказов представлена в табл. 1.

Таблица 1. Сводная информация об исходной выборке

Наименование признака	Категория	Число, ед.
Причина отказа	Внутренняя коррозия	67777
	Внешняя коррозия	22881
	Не определено (NaN)	*14356*
	Прочее – производственный и строительный брак, механические повреждения и т.д.	4975
Природа отказа (отверстия, разрывы, свищи, расслоения, потеря герметичности и пр.)	Тело	80527
	Не определено (NaN)	*19398*
	Прочее	6257
	Сварной шов	3807

Текущие исследования базируются на результатах работы [1], в качестве целевых значений определены: среднее время наработки на отказ участка ПТ, время наработки на первый отказ участка ПТ. Целевые значения соответствуют каждому отдельно взятому уникальному участку ПТ, следовательно, в текущей работе рассматривается выборка, содержащая 18344 уникальных эксплуатируемых ПТ (с учетом импутации [2] пропущенных значений непрерывных признаков), на каждом из которых от начала эксплуатации зафиксировано n отказов (где $n \geqslant 1$) по причинам внутренних коррозий.

Построение и оценка регрессионных моделей

В качестве одного из подходов в рамках настоящих исследований применяется алгоритм случайный лес (random forest, RF) [3, 4].

Итоговое предсказание определяется выражением:

$$a(x) = \frac{1}{N}\sum_{i=1}^{N} b_i(x),$$

где N — число деревьев, b — решающее дерево.

В текущей задаче регрессии при каждом разбиении выбиралось $m = \frac{n}{3}$ случайных векторов признаков из общего числа n.

При построении регрессора в качестве критерия расщепления выполнялась минимизация MSE (среднеквадратичной ошибки) [5].

В качестве дополнительного алгоритма восстановления регрессии был применен алгоритм градиентного бустинга — Light Gradient Boosted Machine (LGBM), подробно описанный в [6].

Оценка качества моделей выполнялась с применением коэффициента детерминации R^2 [7]. Принимая во внимание факт, что значение R^2 возрастает при повышении числа векторов-признаков, на которых базируется модель, для оценки моделей и их сравнения также был применен скорректированный коэффициент детерминации, в котором задействуются несмещенные оценки дисперсий:

$$R^2_{adj} = 1-(1-R^2)\frac{(n-1)}{(n-k)},$$

где n — число образцов (примеров) данных, а k — число векторов-признаков.

Кодирование категориальных признаков осуществляется методом числового кодирования [2]. Определение параметров обучения моделей выполняется с применением кросс-валидации [2]. В табл. 2 представлены результаты работы наилучших полученных регрессоров относительно экспериментов №№ 4...7, где в качестве целевого значения принято среднее время наработки на отказ. Подготовленные выборки разбиты случайным образом в соотношениях 80% (обучение) к 20% (тест) для всех рассматриваемых экспериментов. Реализация и оценка работы регрессоров выполняется в программной среде Python.

Исключение признака «Год ввода в эксплуатацию» существенно снижает качество регрессоров (эксперимент № 4), данный признак является ключевым для всех рассматриваемых регрессоров, условная важность данного признака превышает 90%. Исклю-

Таблица 2. Признаковые пространства и показатели качества регрессоров (по экспериментам)

Наименование признака		№ эксперимента				
		4	5	6	7	8
Непрерывные						
L (длина) простого участка, км		+	+	+	+	+
S (толщина) простого участка, мм		+	+	+	+	+
D (диаметр) простого участка, мм		+	+	+	+	+
$F_ж$ (расход) жидкости/газа в среде, $м^3$/сут.		+	+	+	+	+
$F_н$,(расход) нефти, т/сут.		+	+	+	+	+
Содержание воды (обводненность), %		+	+	+	+	+
$P_{нач}$, атм.		+	+	+	+	+
$P_{кон}$, атм.		+	+	+	+	+
$P_{раб}$, атм.		+	+	+	-	+
Температура транспортируемой среды, °С		+	+	+	+	+
$F_и$ (расход) ингибитора, $г/м^3$		+	+	+	+	+
Глубина заложения, м		+	+	-	+	+
Год ввода в эксплуатацию		-	+	+	+	+
ГКК		+	+	+	+	+
Скорость коррозии, мм/год		+	+	+	-	-
Категориальные						
Месторождение		+	+	+	+	+
Назначение участка ПТ		+	+	+	+	+
Тип участка ПТ		+	+	+	+	+
Тип перекачиваемой среды		+	+	+	+	+
Материал		+	+	+	+	+
Вид соединения		+	+	+	+	+
Тип внутреннего покрытия		+	+	+	+	+
Вид защиты сварного стыка		+	+	+	+	+
Тип ингибитора		+	+	+	+	+
Способ дозирования ингибитора		+	+	+	+	+
Способ прокладки		+	+	-	+	-
Тип грунта		+	+	-	+	+
Тип наружного покрытия		+	+	-	+	+
Тип теплоизоляционного покрытия		+	+	-	+	+
Качественные характеристики регрессоров						
LGBM	R^2_{adj} обучение	0,864	0,982	0,983	0,983	0,983
	R^2_{adj} тест	0,541	0,941	0,942	0,940	0,942
RF	R^2_{adj} обучение	0,933	0,990	0,939	0,991	0,990
	R^2_{adj} тест	0,557	0,936	0,939	0,936	0,940

Таблица 3. Качественные характеристики регрессионных моделей в отношении тестовых данных (для LGBM, RD, CBR – набор данных эксперимента №8, для AutoML – полный набор данных эксперимента №5)

Наименование модели	Среднее время наработки на отказ		Время наработки на первый отказ	
	MSE	MAE	MSE	MAE
LGBM	5,613	1,853	7,229	2,097
RF	5,983	1,899	7,538	2,142
CBR	6,063	1,931	7,292	2,106
Модели H_2O AutoML				
StackedEnsemble_BestOfFamily	5,946	1,937	7,341	2,127
StackedEnsemble_AllModels	5,936	1,936	7,359	2,124
GBM_model_10	6,227	1,961	7,454	2,146
DeepLearning_model	41,072	4,213	43,890	4,734

чение нормированных признаков «Рабочее давление (Рраб)» и «Скорость коррозии» практически не оказывает влияние на качественные характеристики моделей (эксперимент № 7). Наилучшими показателями качества обладают регрессоры, не учитывающие признаки, характеризующие внешние факторы эксплуатации (эксперимент № 6).

Признаковые пространства для экспериментов 4, 5, 6, 7 определялись эмпирически. Для определения признакового пространства эксперимента №8, обладающего наилучшими качественными характеристиками, применяются алгоритмы прямого перебора возможных комбинаций признаков [8, 9] с параллельной оценкой метрики сходимости в отношении тестовых данных. В частности, применяется алгоритм Sequential Forward Floating Selection (SFFS), подробно рассмотренный и также примененный в работе [10]. Для обеспечения гарантированности результатов применения SFFS метрика сходимости вычислялась 10-ти кратно в отношении каждого признакового пространства.

Дополнительно для проведения сравнительного анализа в качестве интеллектуальных средств оценки статистических данных применяются: библиотека H2O, в частности, метод AutoML [11] для автоматизации процесса подбора моделей, а также библиотека градиентного бустинга CatBoost [12, 13] для построения регрессора CatBoostRegressor (CBR). Отметим, что применение данных библиотек не требует обязательного кодирования категориальных признаков.

Результаты работы моделей представлены в табл. 3 в отношении тестовых данных, которые во всех случаях составляют 20% от исходной выборки. Исследования проводятся относительно целевых значений: среднее время наработки на отказ и время наработки на первый отказ. В качестве показателей качества представлены значения среднеквадратичной ошибки (MSE) и абсолютной ошибки (MAE) [5]. В табл. 3 приведены усредненные показатели по результатам 10-кратно повторяемых процессов моделирования и оценки, где в рамках каждой итерации обучения выборка предварительно перемешивалась и разделялась случайным образом на обучающую (80%) и тестовую (20%).

Наилучшими качественными характеристиками обладает регрессор на основе градиентного бустинга LGBM. Результаты автоматизированного поиска наилучших моделей также свидетельствуют о приоритете применения моделей, базирующихся на градиентом бустинге, результаты по наиболее приоритетной (GBM_model_10) представлены в табл. 3.

Также с помощью AutoML выполняется подбор моделей, основанных на применении искусственных нейронных сетей (ИНС) и глубоком обучении [14]. Для сравнения качественные характеристики данного вида регрессора (DeepLearning_model) представлены в табл. 3.

Качественные характеристики моделей LGBM, RF, CBR, GBM приблизительно равны. При этом категориальные признаки для CBR и GBM (в AutoML) не были предварительно закодированы. Следовательно, применение числового кодирования категориальных признаков для моделей LGBM и RF в рамках настоящих экспериментов допустимо.

По результатам анализа полученных качественных характеристик можно сделать вывод, что ансамблевые алгоритмы в большей степени применимы для решения текущих задач в сравнении с моделями, основанными на ИНС.

Значение средней абсолютной ошибки прогноза среднего времени наработки на отказ для наилучшего регрессора LGBM составляет 1,853 года. При этом график остатков свидетельствует о наличии существенных отклонений прогнозов в диапазоне ±

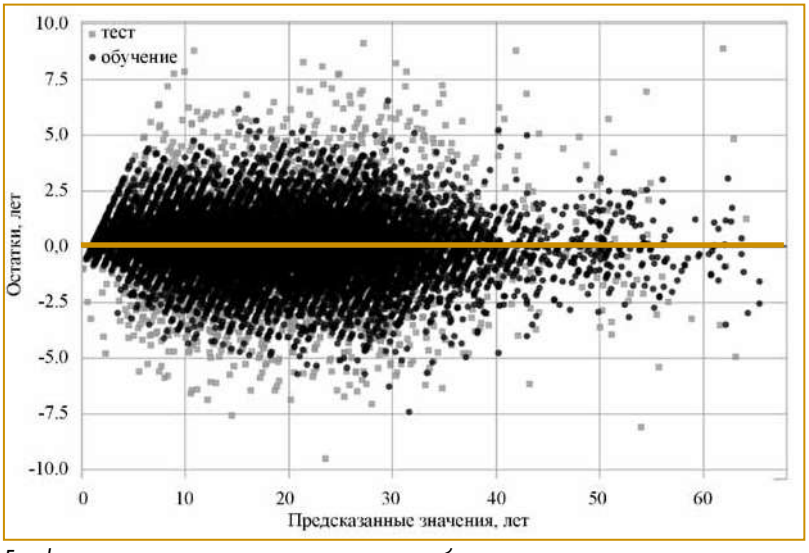

График остатков среднего времени наработки на отказ для регрессора LGBM (эксперимент №8)

10 лет. Следовательно, применение текущей модели в экспертной системе должно сопровождаться ограничениями, накладываемыми на значения признаков примеров эксплуатируемых участков ПТ. Ограничения должны быть направлены на недопустимость ошибок прогноза, которые не удовлетворяют требованиям, выдвигаемым экспертами. Отметим, что наложение ограничений в применении данного регрессора приведет к повышению средней абсолютной ошибки прогноза регрессора.

Заключение

По результатам проведенного интеллектуального анализа статистических данных разработаны регрессоры, предназначенные для прогнозирования среднего времени наработки на отказ и времени наработки на первый отказ проектируемых и эксплуатируемых участков ПТ на основе исторических эксплуатационных данных. Все применяемые модели, базирующиеся на алгоритмах случайного леса и градиентного бустинга, показали хорошие результаты, что подтверждают рассчитанные качественные показатели. Результаты применения AutoML в H2O свидетельствуют о том, что ИНС в меньшей степени применимы для решения данных задач. Вовлечение в исследования признака «Дата ввода» существенно повышает качество регрессоров. Наилучшими показателями обладает модель LGBM, которая на основании данных исследований позволяет с точностью 2 года определять среднее время наработки на отказ и время наработки на первый отказ проектируемых и эксплуатируемых участков ПТ.

Дальнейшие исследования будут направлены на установление ограничений, накладываемых на примеры данных, вводимых в экспертную систему с целью повышения качества прогноза и нивелирования прогнозов, ошибка которых не удовлетворяет требованиям системы.

Список литературы

1. *Кармачев Д.П.* Определение признаковых пространств в рамках разведочного анализа эксплуатационных статистических данных об отказах и условиях эксплуатации простых участков промысловых трубопроводов // Автоматизация в промышленности. 2020. №11. с. 12-16.
2. *Рашка С.* Python и машинное обучение. Перевод А.В. Логунова. М: ДМК Пресс, 2017. 418 с.
3. *Leo Breiman.* Random Forest // Machine Learning (journal): journal. 2001, Vol.45. no 1. P. 5-32.
4. *Hastie T., Tibshirani R., Friedman J. Chapter 15.* Random Forest//The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer-Verlag. 2009. 746p.
5. *Терехов В.А., Ефимов Д.В., Тюкин И.Ю.* Нейросетевые системы управления. М.: Высшая школа, 2002. 184 с.
6. *Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree // Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3149-3157.
7. *Кинякин В.Н., Милевская Ю.С.* Некоторые предостережения по проверке качества модели регрессии с помощью коэффициента детерминации // Вестник Московского университета МВД России. 2014. №8. с. 200-204.
8. *Ferri F.J., Pudil P., Hatef M., Kittler J.* Comparative study of techniques for large-scale feature selection // Pattern Recognition in Practice IV : 403-413.
9. *Pudil P., Novovičová J., & Kittler J.* Floating search methods in feature selection // Pattern recognition letters 15.11 (1994): 1119-1125.
10. *Joe Bemister-Buffington, Alex J. Wolf, Sebastian Raschka, and Leslie A. Kuhn.* Machine Learning to Identify Flexibility Signatures of Class A GPCR Inhibition // Biomolecules 2020, 10, 454.
11. *Erin LeDell and Sebastien Poirier.* H2O AutoML: Scalable Automatic Machine Learning // 7th ICML Workshop on Automated Machine Learning, 2020. https://www.automl.org
12. *Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A.* CatBoost: unbiased boosting with categorical features//NeurIPS, 2018, arXiv:1706.09516. https://arxiv.org/abs/1706.09516.
13. *Dorogush A.V., Ershov V., Gulin A.* CatBoost: gradient boosting with categorical features support // Workshop on ML Systems at NIPS. 2017. http://learningsys.org.
14. *Созыкин А.В.* Обзор методов обучения глубоких нейронных сетей // Вестник ЮУрГУ. Серия «Вычислительная математика и информатика». 2017. Т. 6. №3.

Кармачев Денис Павлович *– ведущий инженер отдела АСУТП АО «ТомскНИПИнефть», Национальный исследовательский Томский политехнический университет.*
E-mail: karmachevd@mail.ru

Корпоративный акселератор ОАО «РЖД» определил лучшие стартапы

Корпоративный акселератор ОАО «РЖД» в структуре АО «ВНИИЖТ» совместно с GenerationS подвели итоги акселерационной программы по направлению «Подвижной состав», которая была запущена в июле 2020 г.

Ее финалистами стали 20 технологических проектов в области повышения эффективности и безопасности железнодорожного транспорта, которые прошли многоэтапный отбор из более чем 600 заявок. В течение двух месяцев финалисты дорабатывали свои решения вместе с функциональными заказчиками со стороны ОАО «РЖД», а также официальными партнерами акселератора, изучали технические требования к продукции и экономике проектов, методику испытаний и другие аспекты взаимодействия с компанией.

В рамках Демо дня 14 команд-финалистов представили свои решения и ожидаемые эффекты от их внедрения руководству ОАО «РЖД» и партнерам акселератора в лице крупных компаний транспортной отрасли, таких как ООО «Газпромтранс», ПАО «Первая Грузовая Компания», АО «Синара-Транспортные Машины» и др. Члены жюри отобрали семь самых перспективных стартапов – с ними будут подписаны Дорожные карты проведения дальнейших испытаний, по итогам которых будет принято решение о масштабировании проектов на сети железных дорог и организации долгосрочного сотрудничества с ОАО «РЖД».

Победители корпоративного акселератора ОАО «РЖД».

1. Апейрон – система мониторинга подвижного состава.
2. Titan Power Solution – гибридные системы рекуперации энергии.
3. Твейджер – беспроводная сеть обмена короткими сообщениями со сверхдальней зоной покрытия.
4. Система управления ТОиР по фактическому состоянию техники – оптимальное управление техническим состоянием подвижного состава.
5. Гудвилл – комплексные универсальные защитные покрытия.
6. SmartMaintenance – прогнозная аналитика для подвижного состава.
7. Imredi – digital-трансформация процессов управления и развития персонала.

В акселерационной программе ОАО «РЖД» в 2020 г. приняли участие не только российские, но и зарубежные стартапы из Казахстана, Израиля, Беларуси, Латвии, Германии, США, Сингапура, Узбекистана, Франции и Эстонии.

Https://accelerator.rzd.ru