DOI: 10.25728/avtprom.2020.11.02

Определение признаковых пространств в рамках разведочного анализа эксплуатационных статистических данных об отказах и условиях эксплуатации простых участков промысловых трубопроводов

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

Представлены результаты исследования, направленного на создание экспертных систем, предназначенных для прогнозирования времени наработки на отказ простых участков промысловых трубопроводов на этапах проектирования и эксплуатации. В первой части работы даны основные сведения об исходной выборке и проведенном разведочном анализе. Вторая часть работы посвящена определению долей верных ответов (accuracy) реализованных классификаторов, а также оценке важностей признаков и анализу полученных результатов. Приведены перспективы дальнейших исследований с применением методов интеллектуального анализа статистических данных по отношению к различным признаковым пространствам, полученным по результатам текущей работы.

Ключевые слова: отказы промысловых трубопроводов, интеллектуальный анализ, случайный лес, разведочный анализ, признаки, факторы эксплуатации.

Введение

В процессах добычи, первичной подготовки, транспортировки нефти и газа нефтегазовые компании применяют различные по конструктивным характеристикам промысловые трубопроводы (ПТ), которые эксплуатируются под воздействием различных внутренних и внешних факторов. К основным внутренним факторам эксплуатации относятся: гидродинамические параметры перекачки, обводненность, физико-химические свойства перекачиваемых сред. К основным внешним факторам эксплуатации относятся: климатические условия эксплуатации ПТ, способы прокладки, физико-химические характеристики среды заложения участков ПТ. Совокупность внутренних и внешних факторов оказывает непосредственное воздействие на образование и развитие внутренних и внешних коррозионных процессов участков ПТ [1].

Настоящая работа связана с разработкой прототипов программного обеспечения экспертных систем (далее «системы») предназначенных для прогнозирования времени наработки на отказ простых участков ПТ на этапе эксплуатации, а также для выбора материального исполнения и способа внутренней антикоррозионной защиты ПТ на этапе проектирования. Цель создания систем заключатся в снижении аварийности и повышении эксплуатационной надежности ПТ, а также в повышении обоснованности выбора материального исполнения промысловых трубопроводов с учетом механизмов коррозионных процессов, протекающих на внутренней стенке трубопроводов. Целью текущего исследования является выполнение предварительного разведочного анализа данных для определения перечня экспериментов (признаковых пространств). В рамках определенных экспериментов в дальнейшем будет реализовано применение различных методов интеллектуального анализа статистических данных и выбор наиболее оптимального метода и достаточного числа признаков для последующего применения в модели представления знаний систем.

Разведочный анализ исходной выборки

Исходные эксплуатационная информация представляет собой статистические данные о характеристиках, отказах и условиях эксплуатации различных участков ПТ. Исходные данные получены путем выгрузки необходимой информации из системы OisPipe. Выборка охватывает 641 месторождение и включает 109989 примеров отказов (строк образцов) и 86 признаков (столбцов). В исследование вовлекаются признаки, которые, по предварительной оценке, косвенно либо напрямую влияют на надежность участков ПТ.

Предварительная оценка, в результате которой в исследование вовлекаются либо игнорируются те или иные признаки, базируется на следующих правилах вовлечения:

1) признак определяем на этапе проектирования и известен до факта возникновения отказа;

2) признак характеризует объект исследования (участок ПТ) относительно эксплуатационной надежности объекта характеризующейся отказами, произошедшими по причинам внутренних коррозий;

3) признак характеризует внутренние факторы эксплуатации объекта — гидродинамические параме-

тры перекачиваемых сред, физико-химические свойства (ФХС) перекачиваемых сред.

В соответствии с п. 1 правил вовлечения не рассматриваются следующие признаки: «Причина отказа», «Характер отказа», «Координата отказа», «Завод-изготовитель», «Подрядчик-строитель». С целью сравнительного анализа в исследование принимаются признаки, характеризующие внешние факторы: «Тип наружного покрытия», «Тип теплоизоляционного покрытия», «Способ прокладки», «Тип грунта», «Глубина заложения». Данные признаки оказывают влияние на температуру перекачиваемых сред и, как следствие, влияют на образование и развитие коррозионных процессов [2].

В эксперименты не вовлекаются признаки, характеризующие организационное и экономические показатели, так как данные признаки не связаны с эксплуатационной надежностью.

Отметим, что в качестве примеров категориального признака «Материал участка» рассматриваются только те материалы (марки стали), которые будут доступны пользователю Системы при проектировании участков ПТ. Данный признак содержит множество категорий (> 30 ед.). При этом наиболее частые категории соответствуют обозначенным в системе маркам стали, что, по предварительной оценке, должно привести к повышению точности конечной модели.

Для защищаемых участков ПТ признаки «Тип ингибитора» и «Способ подачи ингибитора» рассматриваются в качестве категориальных, а признак «Расход ингибитора» в качестве непрерывного. Для незащищаемых трубопроводов данные признаки принимают нулевые значения.

Исходная выборка содержит расширенную информацию о физико-химических свойствах перекачиваемых сред — Ca^{2+} в воде (мг/л), CO_2 в воде (мг/л), растворенный O_2 (мг/л) H_2S в воде и нефти (мг/л), Mg^{2+} в воде (мг/л), взвешенные частицы (мг/л), общая минерализация (мг/л) и т.д. Учет всех данных признаков приводит к сокращению размерности выборки до 2143 примеров отказов при исключении строк данных с пропущенными значениями. В связи с этим расширенная информация о физико-химических свойствах рассматривается только в рамках одного из экспериментов с целью проведения сравнительного анализа результатов.

С целью вовлечения в исследования наибольшего числа примеров отказов, произошедших по причинам внутренних коррозий, выдвигается гипотеза о том, что косвенно физико-химические свойства перекачиваемых сред могут быть учтены через признак «Месторождение» либо совместно через признак «Месторождение» и признак «Группа коррозионных контуров (ГКК)». Здесь «ГКК» является вычисляемым параметром, которое принимает значение в диапазоне 1...4 ед. и указывает на общую условную агрессивность перекачиваемых сред в разрезе детализированных ФХС (алгоритм расчета ГКК в текущей работе не рассматривается). При этом для расчета ГКК используются некоторые признаки из всей группы ФХС, следовательно, вовлечение «ГКК» не приводит к сокращению выборки в рамках эксперимента, который учитывает данный признак. Выдвигаемая гипотеза основывается на предположении о том, что физико-химические свойства перекачиваемых сред для различных участков ПТ в рамках какого-либо месторождения могут быть приняты усредненными на определенном временном интервале. Принимая во внимание возможную важность определенного временного интервала, также предлагается в рамках некоторых экспериментов задействовать признак «Год ввода в эксплуатацию». Таким образом, модели, в составе которых задействован данный признак, могут быть использованы в системе, предназначенной для прогнозирования времен наработки на отказ участков ПТ на этапе эксплуатации, где данный признак должен принимать значение из интервала значений признака «Год ввода в эксплуатацию», задействованных в наборе примеров отказов для данной модели.

Принятие в дальнейшее исследование других признаков, не описанных в рамках настоящей работы, выполнено по аналогии с ранее проводимыми работами по разведочному анализу данных [3].

С целью дальнейшего сравнительного анализа результатов в отношении экспериментов с учтенными признаками, характеризующими внешние факторы, и учтенным признаком «Дата ввода участка ПТ», а также с целью проверки выдвигаемой гипотезы о повышении качества прогнозирования посредством косвенного учета ФХС через признаки «Месторождение» и «ГКК» дальнейшие исследования базируются на пяти основных (1...5) и двух дополнительных (6, 7) экспериментах, признаковые пространства которых отражены в таблице 1. Комбинаторика признаков нацелена на оценку количественных показателей регрессоров и классификаторов в рамках задачи определения достаточного количества признаков для решения задачи по построению прогнозирующих моделей с наилучшими показателями. Дополнительные эксперименты 6 и 7 направлены на сравнительный анализ важности учета признаков, характеризующих внешние факторы, а также не эксплуатационных нормированных признаков «Рабочее давление» и «Скорость коррозии» соответственно.

Целевое значение — время наработки на отказ. Значение вычисляется из исходных баз данных как разность между значениями «дата ввода» и «дата отказа». В качестве единицы измерения для данной переменной принимается мера «год». В качестве отдельно взятых примеров таблицы рассматриваются отказы ПТ, в том числе в качестве различных отказов ПТ рассматриваются отказы одного простого участка ПТ. Данное решение обусловлено тем, что участок ПТ может быть рассмотрен как континуум-система [4]. Это означает, что на протяжении всего срока эксплуатации участок ПТ может отказать n раз (при $n \to \infty$), где

Таблица 1. Признаковые пространства (по экспериментам)

Наименование признака	№ эксперимента *число* примеров отказов						
	1 *2143*	**2** *19623*	**3** *19623*	**4** *19623*	**5** *19623*	**6** *19623*	**7** *19623*
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	+	+	+	+	+	-	+
Общая минерализация воды, мг/л	+	-	-	-	-	-	-
Ca^{2+} в воде, мг/л	+	-	-	-	-	-	-
Mg^{2+} в воде, мг/л	+	-	-	-	-	-	-
K^{2+}, Na^{+} в воде, мг/л	+	-	-	-	-	-	-
HCO_3^- в воде, мг/л	+	-	-	-	-	-	-
Cl^- в воде, мг/л	+	-	-	-	-	-	-
SO_4^{2-} в воде, мг/л	+	-	-	-	-	-	-
pH среды	+	-	-	-	-	-	-
CO_2 в воде, мг/л	+	-	-	-	-	-	-
Растворенный O_2, мг/л	+	-	-	-	-	-	-
H_2S в воде, мг/л	+	-	-	-	-	-	-
Взвешенные частицы, мг/л	+	-	-	-	-	-	-
Сера, мг/л	+	-	-	-	-	-	-
H_2S в нефти, мг/л	+	-	-	-	-	-	-
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	+	+	+	+	+	-	+

каждый отказ обусловлен причинами, которые учитываются в рамках настоящих исследований (признаки отказа). Следует отметить, что в качестве целевых значений в дальнейших исследованиях в рамках различных экспериментов будут приняты значения времени наработки на первый отказ, а также среднее время наработки на отказ для каждого уникального участка ПТ.

В процессе предварительной обработки данных выполняется процедура исключения примеров отказов, значения векторов-признаков которых неизвестны (NaN) и не могут быть однозначно определены. По результатам проведения процедур исключения число примеров отказов сокращается с 109989 до 19623 для экспериментов 2...7 и до 2143 для эксперимента 1 соответственно. При этом в рамках импутации[1] [5] вос-

[1] Импутация - замещение ошибочных, противоречивых и отсутствующих ответов в процессе редактирования данных другими ответами - значениями показателей.

Таблица 2. Значения долей верных ответов классификатора

№ эксперимента	1	2	3	4	5	6	7
Доля верных ответов (accuracy)	0,54	0,85	0,85	0,85	0,87	0,86	0,87

станавливаются пропущенные значения вектора-признака «Содержание воды» для водоводов высокого и низкого давления равные 100%, обозначающие абсолютное содержание воды в перекачиваемых средах.

Оценка важностей признаков

С целью охвата полного перечня признаков при оценке их важностей, кодирование категориальных признаков осуществляется числовым методом [6], так как не представляется возможным присвоить каждой категории признака числовые значения, которые бы верно отражали количественную корреляцию между категориями в рамках категориальных векторов-признаков в отношении эксплуатационной надежности участков ПТ. С целью проверки адекватности данного подхода оценка важностей выполняется по отношению к стандартизированным [7] и исходным непрерывным признакам. При этом обе полученные диаграммы важностей являются полностью идентичными. На рис. 1 представлена диаграмма важностей признаков для эксперимента № 5, как наиболее полного в части числа вовлеченных признаков и наиболее адекватного с точки зрения интерпретируемости результатов, так как эксперимент № 1, охватывающий все признаки за исключением «ГКК», содержит 2143 примера отказов. Построение диаграммы и вычисление значений важностей выполняется в программной среде Python с применением классифицирующей модели, функционирующей на основе ансамблевого алгоритма случайный лес (random forest) [8, 9].

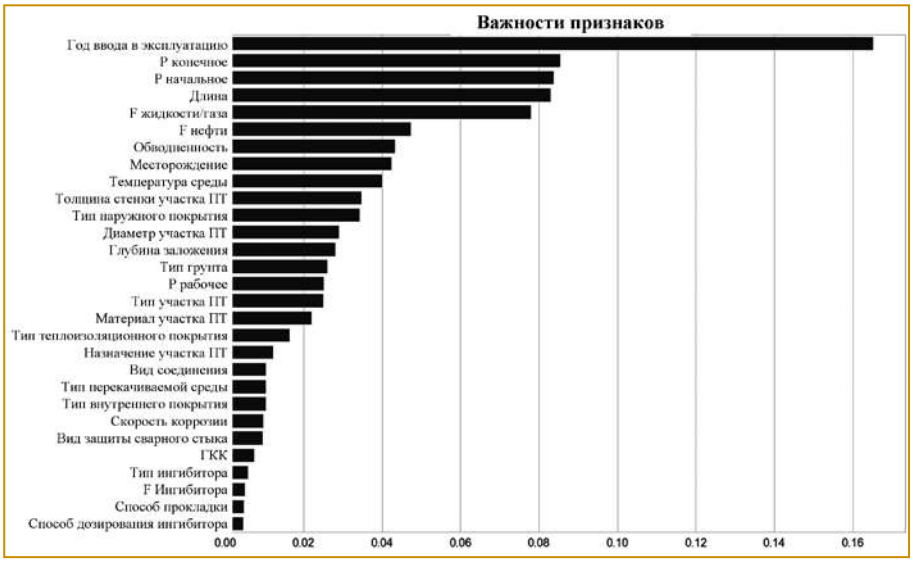

Рис. 1. Диаграмма важностей признаков относительно среднего времени наработки на отказ (для эксперимента №5)

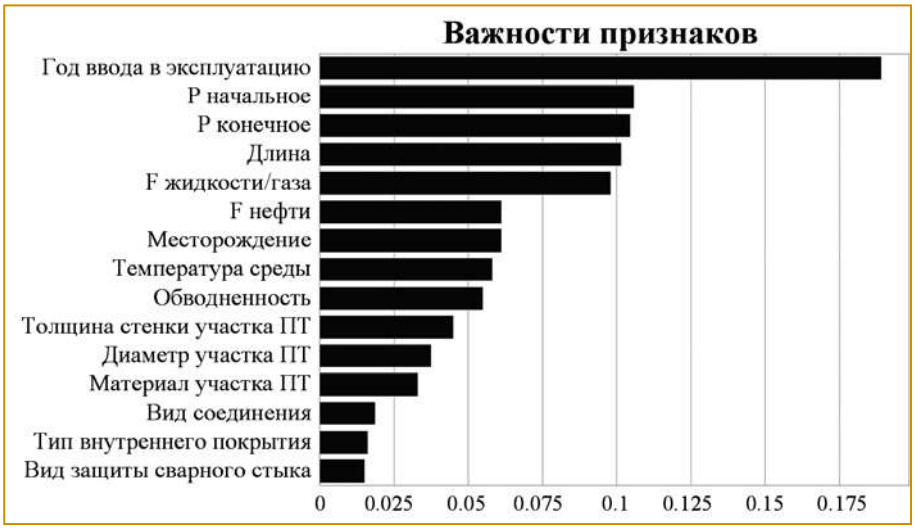

Рис. 2. Диаграмма важностей признаков относительно среднего времени наработки на отказ (для контрольного эксперимента)

Дополнительно для каждого эксперимента в программной среде Python выполняется расчет доли верных ответов (accuracy) [10], результаты расчетов представлены в табл. 2. Расчет долей верных ответов для всех экспериментов выполняется в равных условиях: настроечные параметры модели принимаются по умолчанию, в качестве критерия расщепления принимается индекс Джини (Gini coefficient) [11]. Задача классификации решается относительно классов, определенных следующими интервалами значений среднего времени наработки на отказ (в годах): [1...4], (4...7], (7...8], (8...10], (10...12], (12...14], 15...16], (16...19], (19...22], (22...25], (26...29], (29...56]. Классы обозначены таким образом, что каждый из классов включает приблизительно равное число примеров отказов (около 1600 с допущением ± 5%). Оценка классификаторов выполняется по отношению к тестовым выборкам, которые составляют 30% от общего числа примеров для эксперимента № 1 и 20% от общего числа примеров для других экспериментов.

Доля верных ответов классификатора для эксперимента № 1 свидетельствует о недостаточном числе примеров отказов,

содержащих расширенную информацию о ФХС перекачиваемых сред. Вовлечение признаков «ГКК» и «Месторождение» не приводит к повышению точности в рамках текущей модели, базирующейся на алгоритме случайного леса. При этом модель, реализованная в рамках эксперимента № 5, обладает большей точностью в сравнении с моделями экспериментов № 3 и № 4, что также подтверждается важностью признака «Год ввода в эксплуатацию» в соответствии с диаграммой (рис. 1). Исключение признаков, характеризующих внешние факторы эксплуатации, снижает долю верных ответов классификатора. Вероятно, это связано с тем, что сформированные выборки также включают в себя некоторые отказы, произошедшие по причинам внешних коррозий, либо исключенные внешние факторы оказывают существенное влияние на температуру перекачиваемых сред и, как следствие, на образование и протекание коррозионных процессов, в том числе на внутренней стенке простых участков ПТ. Исключение нормированных признаков «Рабочее давление» и «Скорость коррозии» не приводит к снижению точности.

Исходя из общей картины распределения важностей и долей верных ответов, выполняется оценка классификатора, основанного на дополнительном контрольном эксперименте, признаки которого характеризуют только фактические гидродинамические свойства перекачиваемых сред, параметры участков, материал, сварной стык и внутреннее покрытие. Также данный эксперимент включает признак «Год ввода в эксплуатацию». Доля верных ответов данного классификатора составляет 0,87, при этом исключение признака «Год ввода» сокращает значение метрики до 0,85. На рис. 2 представлена диаграмма распределения важностей признаков контрольного эксперимента.

Признаков, учтенных в рамках контрольного эксперимента, достаточно для обеспечения максимального значения доли правильных ответов классификатора, основанного на алгоритме случайного леса. Отметим, что контрольный эксперимент не включает признаки, характеризующие внешнюю стенку участков ПТ, исключение которых приводит к снижению значения метрики в отношении сравнительной оценки 5-го и 6-го экспериментов при прочих равных условиях. Вероятно, исключение других признаков, обеспечивающих переход от 6-го эксперимента к контрольному, нивелирует переобучение модели, происходящее по причинам установления ошибочных зависимостей моделью.

Заключение

Выдвинутая гипотеза о повышении качества прогнозирования посредством косвенного учета ФХС перекачиваемых сред с помощью вовлечения в пространство признаков «ГКК» и «Месторождение» на текущий момент не подтверждается. Требуется дальнейшее проработка данного вопроса на этапе применения других интеллектуальных методов. Тем не менее реализованные классификаторы демонстрируют удовлетворительные результаты в соответствии с рассчитанными долями верных ответов. Принимая во внимание результаты 1-го, 5-го и контрольного экспериментов, а также применение в рамках текущей предварительной оценки только алгоритма случайного леса, дальнейшие исследования должны базироваться на пространстве признаков 5-го эксперимента и включать процедуры перекрестной проверки [5] векторов-признаков в рамках каждого применяемого интеллектуального метода в задачах регрессии и классификации, а также базироваться в том числе на прямом кодировании категориальных признаков [5] для методов, которые в этом нуждаются.

Список литературы

1. *Владимирова А.И.* Промышленная безопасность и надежность магистральных трубопроводов / А.И. Владимирова, В.Я. Кершенбаума/ М: Нац. ин-т нефти и газа. 2009. – 695 с.
2. *Калимуллин А.А.* Химические методы защиты от коррозии нефтепромысловых трубопроводов / А.А. Калимуллин, Е.Н. Сафонов, К.Р.Низамов, И.Ш. Гарифуллин /. Сб. научн. тр. ООО «Башгеопроект». Уфа, 2007. Вып. 119. Ч. 1. С. 132-139.
3. *Кармачев Д.П.* Анализ статистических данных об отказах промысловых трубопроводов ПАО «НК «Роснефть» /Д.П. Кармачев// Автоматизация, телемеханизация и связь в нефтяной промышленности. 2019. №9 (554). с. 15-21.
4. *Острейковский В.А.* Статистический анализ надежности нефтепромысловых трубопроводов / В.А. Острейковский, Я.В. Силин // Нефтегазовое дело. 2008. №1.
5. *Рашка С.* Python и машинное обучение / С. Рашка / Пер. с английского А.В. Логунова. – М: ДМК Пресс, 2017. – 418 с.
6. *Терехов В.А.* Нейросетевые системы управления / В.А. Терехов, Д.В. Ефимов, И.Ю. Тюкин / М.: Высшая школа, 2002. – 184 с.
7. *Дьяконов А.Г.* Методы решения задач классификации с категориальными признаками / А.Г. Дьяконов // Прикладная математика и информатика. Тр. факультета Вычислительной математики и кибернетики МГУ им. М.В. Ломоносова. 2014. № 46. – с. 103-127.
8. *Breiman L.* Random Forest / L. Breiman // Machine Learning (journal): journal. – 2001, Vol.45, no 1 – P. 5-32.
9. *Hastie T.* Chapter 15. Random Forest/T. Hastie, R. Tibshirani, J. Friedman//The Elements of Statistical Learning: Data Mining, Inference, and Prediction — 2nd ed. — Springer-Verlag, 2009-746p.
10. *Дудченко П.В.* Метрики оценки классификаторов в задачах медицинской диагностики/П. В. Дудченко//Молодежь и современные информационные технологии: тр. XVI международной научно-практической конф. студентов, аспирантов и молодых ученых. 2018. — Томск: Изд-во ТПУ, 2018. — с. 164-165.
11. *Красс М.С., Чупрынов Б.П.* Основы математики и ее приложения в экономическом образовании. 4-е изд., испр. – М.: Дело, 2003 – 688 с.

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