ТЕПЛОЭНЕРГЕТИКА, 2019, № 3, с. 49–56

АВТОМАТИЗАЦИЯ И ТЕПЛОВОЙ КОНТРОЛЬ В ЭНЕРГЕТИКЕ

МОДЕЛЬ РАННЕГО ОБНАРУЖЕНИЯ АВАРИЙНЫХ СИТУАЦИЙ НА ОБОРУДОВАНИИ ЭЛЕКТРОСТАНЦИЙ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

© 2019 г. А. А. Коршикова[a], *, А. Г. Трофимов[b]

[a]ООО "Инконтрол", 115280, Россия, Москва, ул. Ленинская Слобода, д. 23, стр. 2

[b]Национальный исследовательский ядерный университет "МИФИ", 115409, Россия, Москва, Каширское ш., д. 31

**e-mail: aakorshikova@gmail.com*

Поступила в редакцию 19.06.2018
После доработки 21.08.2018
Принята к публикации 29.08.2018

Рассматривается метод раннего обнаружения и предсказания аномальности функционирования технологического оборудования энергоблоков на примере питательного турбонасоса ПТН 1100-350-17-4 энергоблока 300 МВт. Определяется актуальность задачи предсказания возможных неисправностей технологического оборудования на ранней стадии их возникновения и объясняются особенности ее решения в энергетике. Очерчен круг дефектов технологического оборудования, определение которых может быть эффективно осуществлено методами предиктивной аналитики. Подчеркивается принципиальный тезис достаточности существующего в программно-техническом комплексе автоматической системы управления технологическими процессами парка аналоговых и дискретных измерений для применения методов предиктивной аналитики. Приводится краткий обзор современных методов предиктивной аналитики и особенностей обучающих модель алгоритмов. Отдельное внимание уделено проблемам подготовки исходных данных для обучения модели. Формулируется математическая задача моделирования показателя аномальности, принимающего значения от 0 (нормальное функционирование) до 1 (аномальное функционирование). Она, в свою очередь, сформулирована как задача бинарной классификации векторов признаков, характеризующих состояние оборудования в данный момент времени. Предложен оригинальный подход, сочетающий в себе метод MSET (Multivariate State Estimation Technique), в котором степень аномальности в техническом состоянии определяется по превышению критерием Хотеллинга порогового уровня, рассчитываемого алгоритмом автоматически, и методы машинного обучения, позволяющие избежать некоторых трудностей, присущих MSET. Для определения состава наиболее информативных признаков, по значениям которых может быть обнаружено раннее развитие аварийной ситуации использован ансамбль регрессионных моделей. Дано обоснование способа выбора моделируемой переменной и множества регрессоров. Метод расчета показателя аномальности основан на формировании ансамбля линейных регрессионных моделей. Продемонстрировано преимущество этого метода перед применением единственного классификатора. Предложен метод формирования сигнализации обнаружения аномальности работы технологического оборудования энергоблоков. Показано, что предложенная модель позволила выявить начало развития аварийной ситуации, в то время как отдельные параметры не выявили особенности функционирования насоса в предаварийном интервале времени.

Ключевые слова: технологическое оборудование, обнаружение аномалий, предиктивная аналитика, комитет классификаторов, логистическая регрессия

DOI: 10.1134/S0040363619030044

В период активного использования технологического оборудования неизбежно происходят события (аварии), которые могут оказать негативное влияние на его работу или привести к выходу из строя. Модель, предсказывающая будущую аварийную ситуацию, позволила бы своевременно принять меры для ее устранения и повысить тем самым эффективность использования технологического оборудования. Построением и исследованием таких моделей занимается предиктивная аналитика [1, 2].

На фоне широкого применения методов предиктивной аналитики в различных сферах человеческой деятельности (финансовые услуги, страхование, телекоммуникации, торговля, здравоохранение и др.) ее успехи в энергетике выглядят

весьма скромно. Этому есть простое объяснение. За почти вековую историю существования российской энергетики был сформирован объем измерений параметров состояния энергооборудования, позволяющий при соблюдении всех необходимых условий (регламентов) обеспечивать его безаварийную эксплуатацию. Предотвращение неисправности или аварии в подавляющем большинстве случаев не требует каких-то специальных методов, а осуществляется по регистрируемым измерениям (срабатывает предупредительная, аварийная сигнализация). Происходящие аварии вызваны, прежде всего, несоблюдением правил и регламентов эксплуатации оборудования или внешними воздействиями (аварии в энергосистеме, природные явления, человеческий фактор и т.п.), которые невозможно предвидеть.

Тем не менее, можно с большой долей уверенности предположить, что существуют дефекты, приводящие в случае их несвоевременного устранения к аварии. Зарождение и развитие таких дефектов не контролируются отдельными измерениями, но они диагностируются методами предиктивной аналитики по поведению той или иной совокупности измеряемых параметров [3]. В связи с этим возникает необходимость разработки математических моделей, позволяющих предсказывать аварийные ситуации заранее.

Следует отметить, что применение методов предиктивной аналитики на оборудовании энергоблока не требует установки новых датчиков (объем необходимых измерений определяется заводами–изготовителями оборудования, проектировщиками энергоблока и технологами электростанции).

Основная идея предиктивной аналитики состоит в том, что возникновение аварии можно предсказать с некоторой вероятностью на основе непрерывного анализа данных, характеризующих функционирование диагностируемого оборудования. Предсказание можно считать состоявшимся, если оно произошло за несколько дней до аварии.

Современные тенденции в предиктивной аналитике сочетают в себе методы статистического и интеллектуального анализа данных с использованием обучающихся алгоритмов [4]. Они присущи всем существующим в настоящее время методам предиктивной аналитики, применяемым в теплоэнергетике, и предполагают предварительное “обучение” модели на основе имеющихся исходных данных. Исходными данными являются “исторические” значения измеряемых параметров, характеризующих работу конкретного технологического оборудования, которые берутся из архивов установленного на энергоблоке программно-технического комплекса (ПТК) за длительный период работы (обычно один–три года). Кроме того, используются сведения об обнаруженных за этот период дефектах (неисправностях), которые могут в случае их неустранения привести к аварии. Период работы оборудования с такими дефектами называют аномальным, без неисправностей – нормальным.

Обучающие алгоритмы определяются методами предиктивной аналитики. Например, метод регрессионных моделей использует обучающие интервалы для установления коэффициентов регрессии и порога вычисленного критерия (выходного параметра модели), классифицирующего период работы оборудования (нормальный/аномальный). Метод искусственных нейронных сетей на обучающих интервалах настраивает весовые коэффициенты нейронов.

Следует отметить, что критически важными для построения качественной прогностической модели являются адекватные данные по дефектам диагностируемого оборудования. К сожалению, как показывает опыт работы авторов с исходными данными энергоблоков различных электростанций, практически все журналы дефектов имеют одни и те же существенные недостатки: фиксируются не все дефекты и даты в лучшем случае соответствуют времени их обнаружения (а не возникновения!). При этом даты устранения обнаруженных дефектов также довольно часто отсутствуют.

Некоторые установленные на электростанциях современные ПТК автоматического управления могут включать в себя встроенные системы ранней диагностики отказа. В их основу, как правило, положены статистические модели обнаружения аномалий, работающие по принципу: если текущее состояние оборудования существенно отличается от состояния, характерного для нормального режима, то это признак аномального функционирования [5]. Недостатком этих систем является, с одной стороны, зачастую позднее обнаружение аномалии, когда времени на ее устранение уже нет, и, с другой стороны, значительное количество ложных предупреждений о возможной аварии.

В настоящей работе предложен метод обнаружения аномальностей функционирования объекта, использующий идею метода MSET[1] [6] совместно с моделями машинного обучения: линейной и логистической регрессией.

ПОСТАНОВКА ЗАДАЧИ

Пусть функционирование объекта в каждый момент времени t описывается вектором показа-

[1] MSET – Multivariate State Estimation Technique. В методе MSET степень аномальности в техническом состоянии определяется по превышению критерием Хотеллинга порогового уровня, рассчитываемого алгоритмом автоматически.

телей $x(t) = [x_1(t),\ldots, x_m(t)]^T$ (здесь m – число показателей; T – количество проведенных измерений). При его измерениях системой мониторинга с некоторым шагом Δt (например, 5 мин) образуется последовательность векторов, объединенных в матрицу $X = [x(1),\ldots, x(T)]$.

Каждый момент времени t, $t = \overline{1,T}$, отнесен экспертом к одному из двух классов: соответствующему нормальному функционированию или аномальному (предаварийному либо аварийному) состоянию объекта. После обозначения метки класса в момент времени t через $y(t)$ [$y(t) = 0$ для нормального состояния и $y(t) = 1$ для аварийного] вектор меток выглядит следующим образом: $y = [y(1),\ldots, y(T)]$.

Далее для построения модели из состава имеющихся измерений необходимо определить выходную переменную и "объясняющие" ее входные переменные. Модели, построенные по методу MSET, относятся к классу "автоассоциативных", когда наборы входных и выходных параметров для построения регрессионной модели в нормальном режиме функционирования объекта совпадают [6, 7]. Такие модели особенно удобны в тех физических условиях, когда наблюдаемые параметры, с одной стороны, тесно взаимосвязаны, а с другой стороны, затруднительно или нецелесообразно выделять из них объясняющие и объясняемые параметры, как это делается при построении причинно-следственных моделей. От состава этих переменных во многом зависят предиктивные свойства отклонений ошибок модели от наблюдаемых значений. Разумно предположить, что выходные переменные модели зависят не только от наблюдаемых показателей объекта, но и еще от производных показателей, не измеряемых явно (например, относительных величин, производных величин и т.п.). Задача формирования нового, расширенного состава показателей $z_1,\ldots, z_M$ на основе исходных показателей $x_1,\ldots, x_m$ требует привлечения экспертов и понимания того, какие именно производные показатели могли бы быть чувствительны к развитию аварии.

Другая проблема метода MSET связана с расчетом степени аномальности функционирования объекта. В классическом варианте метода MSET [6] решение об аномальности принимается по результатам сравнения выходов построенной регрессионной модели с наблюдаемыми значениями моделируемой величины. На практике высокие невязки не всегда означают аварию и, наоборот, некоторые типы предаварийных ситуаций не всегда могут проявлять себя в увеличении невязок.

В настоящей работе ставится задача построения ансамблей регрессионных моделей [8, 9], использующих различные составы входных и выходных переменных, и моделей расчета показателя аномальности. Каждая модель в ансамбле рассчитывает свой показатель аномальности функционирования объекта $p_i(t)$ в момент времени t при $t = \overline{1,T}$, $i = \overline{1,N}$ (здесь N – число моделей в ансамбле). Итоговое решение $p(t)$ принимается в соответствии с решающим правилом ансамбля.

В каждый момент времени показатель аномальности $p(t)$ принимает значение из интервала (0; 1). Значения, близкие к 0, соответствуют нормальному функционированию объекта, близкие к 1 – аномальному.

МЕТОД РАСЧЕТА ПОКАЗАТЕЛЯ АНОМАЛЬНОСТИ

Предлагаемый алгоритм расчета показателя аномальности функционирования объекта состоит из следующих шагов.

1. Формирование признаков $z_1,\ldots, z_M$ на основе наблюдаемых показателей функционирования объекта $x_1,\ldots, x_m$ и расчет их значений $z(t) = [z_1(t),\ldots, z_M(t)]^T$ в каждый момент времени t, $t = \overline{1,T}$. В результате получаем последовательность векторов, объединенных в матрицу $Z = [z(1),\ldots, z(T)]$. Таким образом, исходные данные D для построения предиктивной модели – это матрица значений показателей Z размерности $M \times T$ и вектор меток y размерности T.

В настоящей работе в качестве исследуемого объекта мониторинга рассматривается питательный турбонасос ПТН 1100-350-17-4 энергоблока 300 МВт.

Исходя из экспертных мнений выбран следующий состав показателей $z_1(t),\ldots, z_M(t)$ насоса в момент времени t:

наблюдаемые значения $x_1(t),\ldots, x_m(t)$ (например, температура и давление смазочного масла, скорости вибрации подшипников турбин, температура и давление пара на входе и выходе ПТН и пр.);

значения $x_1(t),\ldots, x_m(t)$, нормированные на расход питательной воды $x^*(t)$, т.е. $\frac{x_1(t)}{x^*(t)},\ldots,\frac{x_m(t)}{x^*(t)}$, в данном случае $M = 2m$.

2. Разбиение исходной выборки D на непересекающиеся подмножества: две обучающие D_{tr1}, D_{tr2} и тестовую D_{tst} выборки, $D = D_{tr1} \cup D_{tr2} \cup D_{tst}$, случайным образом в заданном соотношении. Множества соответствующих моментов времени обозначим через T_{tr1}, T_{tr2} и T_{tst}, $T_{tr1} \cup T_{tr2} \cup T_{tst} = \{1,\ldots, T\}$. Обучающая выборка D_{tr1} используется для построения регрессионных моделей, выборка

D_{tr2} – для построения моделей расчета показателя аномальности функционирования объекта, выборка D_{tst} – для финального тестирования модели.

3. Построение K линейных регрессионных моделей по данным обучающей выборки D_{tr1} и расчет ошибок моделей на данных выборки D_{tr2}. Пусть z_{i_k} и ζ_k – выходная (моделируемая) переменная и множество входов модели (регрессоров) соответственно для k-й регрессионной модели, $i_k \in \{1,\ldots,M\}$, ζ_k – некоторое подмножество множества признаков $\{z_1,\ldots,z_M\}$, не содержащее моделируемый признак z_{i_k}, $k = \overline{1,K}$. Тогда k-я регрессионная модель будет иметь вид

$$z_{i_k} = \varphi_k(\zeta_k) + e_k, \;\; k = \overline{1,K},$$

где φ_k – функция регрессии, линейная по признакам ζ_k; e_k – ошибка модели.

Моделируемый признак z_{i_k} и состав регрессоров ζ_k для каждой регрессионной модели выбираются случайным образом. Например, для одной модели в качестве выходной переменной было принято давление пара на выходе ПТН, в качестве входных – скорости горизонтальных и вертикальных вибраций подшипников, температура и давление пара на входе ПТН. Для другой модели выходной переменной являлась скорость горизонтальной вибрации первого подшипника, а входными – температуры упорных подшипников турбин, давление смазочного масла и температуры пара на входе и выходе ПТН.

Случайный способ выбора моделируемой переменной и множества регрессоров позволяет исключить на данном этапе привлечение экспертов для построения регрессионных моделей. В то же время если состав этих переменных выбран неудачно, т.е. точное моделирование выходной переменной невозможно, то такие модели отбрасываются и состав переменных разыгрывается заново. Точность модели оценивается по значению коэффициента детерминации на обучающей выборке D_{tr1}.

Для нахождения параметров линейных функций регрессии $\varphi_1,\ldots,\varphi_K$ был использован метод наименьших квадратов [10].

Вводится обозначение $e(t) = \left[e_1(t),\ldots,e_K(t)\right]^T$ (здесь $e_k(t) = z_{i_k}(t) - \varphi_k\left[\zeta_k(t)\right]$ – векторы отклонений выходов регрессионных моделей от наблюдаемых значений в момент времени t, $t \in T_{tr2}$). В результате получается матрица отклонений E размерности $K \times T_{tr2}$, составленная из векторов $e(t)$, $t \in T_{tr2}$.

4. Данные матрицы E и соответствующие метки классов в моменты времени t, $t \in T_{tr2}$, используются для обучения ансамбля из N бинарных классификаторов[2], в качестве которых выбраны логистические регрессии. Каждый бинарный классификатор работает в своем пространстве признаков, состоящем из случайно отобранных регрессионных остатков, полученных на предыдущем шаге. Таким образом, каждый бинарный классификатор решает свою задачу классификации, т.е. отнесение поступающего на вход вектора регрессионных остатков к одному из двух классов: нормальное или аномальное функционирование. Например, на вход первого классификатора поступают регрессионные остатки, полученные при моделировании давления пара на выходе ПТН и скорости горизонтальной вибрации первого подшипника, на вход второго – при моделировании температуры упорных подшипников турбин, температуры пара на входе и выходе ПТН и т.д.

Экспериментально показано, что точности отдельных классификаторов получаются неудовлетворительными – на основании только значений регрессионных остатков, поступающих на вход, нельзя сказать, находится ли система в нормальном или аварийном (предаварийном) режиме работы. Тем не менее, точности отдельных классификаторов получаются выше, чем при случайной классификации, что позволяет объединить их в ансамбль классификаторов. В [11] показано, что при определенных условиях точность работы ансамбля превосходит точности образующих его классификаторов (слабых классификаторов).

Математическая модель i-го слабого классификатора имеет вид

$$p_i = \frac{1}{1 + \exp\left[-\psi_i(\varepsilon_i)\right]}, \;\; i = \overline{1,N},$$

где p_i – выход i-го слабого классификатора, $p_i \in (0;1)$; ψ_i – функция регрессии, линейная по признакам ε_i; ε_i – некоторое подмножество множества признаков $\{e_1,\ldots,e_K\}$.

Состав регрессоров ε_i для каждого слабого классификатора выбирался случайным образом. Такой выбор не требует привлечения экспертных знаний и позволяет составить множество слабых классификаторов для объединения в комитет. Преимуществом комитета классификаторов перед единственным классификатором, на вход которого подавались бы все полученные остатки $e_1,\ldots,e_K$, является его большая устойчи-

[2] Классификатор – алгоритм, пытающийся предсказать по известным ему данным, к какому классу из заранее определенных будут относиться новые данные. Слабый классификатор производит классификацию с вероятностью ошибки меньшей, чем при простом угадывании (0.5 для бинарной классификации).

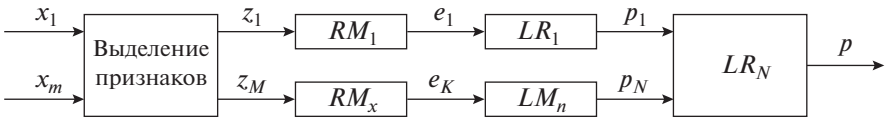

Рис. 1. Схема модели расчета показателя аномальности функционирования объекта. *RM* – регрессионная модель; *LR* – логистическая регрессия

вость, что уменьшает настройку на конкретный состав входных признаков, которые также были получены в результате случайного отбора регрессоров на шаге 3.

Если пространство входных переменных для слабого классификатора выбрано неудачно, т.е. разделимость данных двух классов в нем практически не отличима от случайной, то такие слабые классификаторы отбрасываются и состав входных переменных разыгрывается заново. В качестве меры разделимости классов использовали значения показателя AUC ROC на обучающей выборке D_{tr2} [12].

Вектор выходов слабых классификаторов в момент времени t обозначается как $p(t) = [p_1(t),\ldots,p_N(t)]^T$ [здесь $p_i(t)$ – выход i-го слабого классификатора в момент времени t, $t \in T_{tr2}$]. В результате получается матрица показателей аномальности P размерности $N \times T_{tr2}$, составленная из векторов $p(t)$, $t \in T_{tr2}$.

5. Данные матрицы P и соответствующие метки в моменты времени t, $t \in T_{tr2}$, используются для построения решающего правила ансамбля Ф:

$$p = \Phi(p_1,\ldots,p_N),$$

где p – итоговый показатель аномальности.

В простейшем случае в качестве решающего правила может быть использовано простое усреднение выходов слабых классификаторов. Однако, учитывая возможный неравный вклад отдельных слабых классификаторов в итоговое решение, а также их различные показатели точности и обобщающие способности, в качестве решающего правила Ф применяем логистическую регрессионную модель:

$$\Phi(p_1,\ldots,p_N) = \frac{1}{1+\exp\left[-(\beta_0+\beta_1 p_1+\ldots+\beta_N p_N)\right]},$$

где $\beta_0,\ldots,\beta_N$ – настраиваемые параметры модели, вычисляемые методом максимального правдоподобия [13].

На вход логистической регрессии подаются выходы слабых классификаторов $p_1,\ldots,p_N$, на выходе моделируется итоговый показатель p.

Схема предложенной модели приведена на рис. 1.

После построения ансамбля (обучение всех регрессионных моделей, всех слабых классификаторов и решающего правила) проводится его тестирование на данных тестовой выборки D_{tst} и визуально оценивается поведение рассчитанного показателя аномальности $p(t)$ в предаварийные интервалы времени.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

Экспериментальные исследования проводили на исторических данных функционирования питательного турбонасоса энергоблока 300 МВт, записанные в течение трех лет. В каждый момент времени с интервалом $\Delta t = 5$ мин регистрировали $m = 45$ показателей функционирования насоса (например, давление масла, скорость вибрации подшипников, температуру пара и др.). После исключения моментов времени, когда насос находился в выключенном состоянии, число временных отсчетов составило $T = 39629$. На протяжении рассматриваемого периода времени были зарегистрированы четыре дефекта в работе насоса (перегрев упорных колодок приводной турбины, течь в крышку со стороны подшипника, течь конденсата в накидную гайку и торцевую крышку со стороны пускового устройства). Для каждой неисправности известен момент ее фиксации (с точностью до дня). Вектор меток классов y формировали следующим образом: значения $y(t)$ полагали равными 1 для всех моментов времени, начиная от полудня второго дня до фиксации дефекта и заканчивая полуднем второго дня после, все остальные метки полагали равными 0. На рис. 2 показан фрагмент исходных данных, предшествующий одной из аварий.

Множество всех моментов времени разбито на подмножества T_{tr1}, T_{tr2} и T_{tst} в соотношении 40/40/20, т.е. в обучающие выборки T_{tr1} и T_{tr2} попадают случайным образом отобранные 40% всех имеющихся временных отсчетов, в тестовую выборку T_{tst} – 20%.

На основе наблюдаемых показателей насоса $x_1,\ldots,x_m$ были рассчитаны производные показатели $z_1,\ldots,z_M$, $M = 89$, которые были использованы для построения $K = 50$ регрессионных моделей. Моделируемую переменную и состав регрессоров в каждой модели выбирали случайным образом.

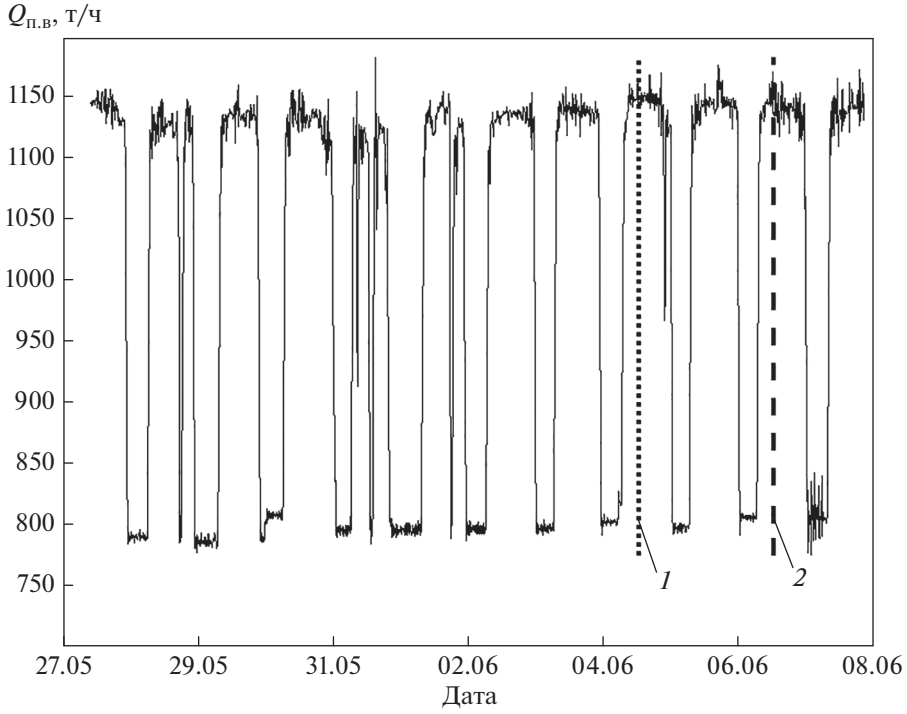

Рис. 2. Зависимость расхода питательной воды $Q_{п.в}$ от интервала времени, предшествующего аварии (течь конденсата в накидную гайку) (2014 г.).
1 – начало предаварийного интервала; *2* – авария

Если коэффициент детерминации R^2 (интерпретируется как соответствие модели данным) регрессионной модели на обучающей выборке D_{tr1} оказывался меньше 0.7, то такую модель исключали. Среднее значение коэффициентов детерминации R^2 построенных моделей на обучающей выборке D_{tr1} равно 0.98 (среднеквадратичное отклонение 0.04), на выборке D_{tr2} (которая являлась для них тестовой) – 0.97 (среднеквадратичное отклонение 0.07). Полученные значения свидетельствуют о том, что построенные регрессионные модели в среднем обладают хорошей обобщающей способностью, т.е. вероятность ошибки на тестовой выборке не сильно отличается от ошибки на обучающей выборке.

Для каждой регрессионной модели в каждый момент времени t, $t \in T_{tr2}$, были рассчитаны ошибки моделирования $e_k(t)$, $k = \overline{1, K}$, которые использовали для построения $N = 20$ слабых классификаторов (логистических регрессий). Число и состав регрессоров для каждой логистической регрессии выбирали случайным образом. Если количественная интерпретация качества слабого классификатора (показатель AUC ROC) на обучающей выборке D_{tr1} оказывалась меньше 0.6, то такой классификатор исключали. Среднее значение показателей AUC ROC построенных классификаторов на обучающей выборке D_{tr2} равно 0.95 (среднеквадратичное отклонение 0.08), на тестовой выборке D_{tst} – 0.94 (среднеквадратичное отклонение 0.11). Полученные значения показывают, что построенные слабые классификаторы в среднем обладают хорошей обобщающей способностью.

Далее выходы слабых классификаторов на обучающей выборке D_{tr2} использовали для обучения логистической регрессии. Достигнутое значение показателя AUC ROC обученной регрессионной модели на обучающей выборке D_{tr2} равно 0.99, на тестовой выборке D_{tst} – 0.98.

На рис. 3 приведен график зависимости наблюдаемого на выходе логистической регрессионной модели показателя аномальности p функционирования насоса в течение интервала времени, предшествующего аварии.

Из графика видно, что устойчивый рост показателя аномальности начинается уже утром 3 июня, в то время как аварийный интервал в обучающей выборке начинался только с полудня 4 июня. Ранняя тенденция к возрастанию значений показателя свидетельствует о наличии предиктивной способности у рассчитанного показателя.

Персонал, эксплуатирующий оборудование энергоблока, не обнаружил особенностей в поведении отдельных показателей функционирования насоса в предаварийном интервале времени (в частности, для показателя, изображенного на

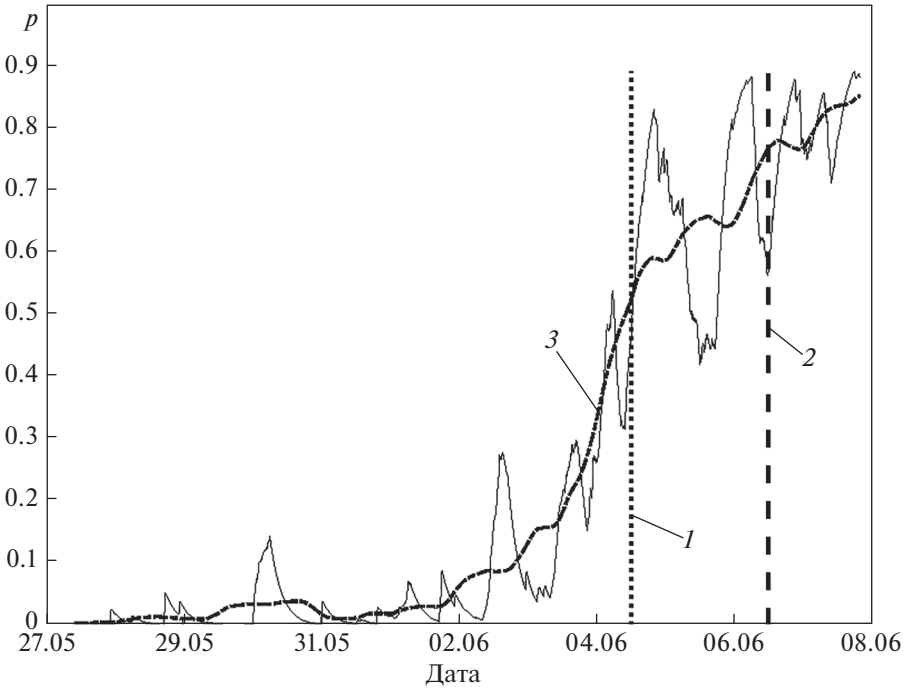

Рис. 3. Зависимость показателя аномальности функционирования насоса от времени перед аварией (2014 г.).
1 – начало предаварийного интервала; *2* – авария; *3* – "отфильтрованный"/сглаженный показатель аномальности

рис. 2, ни авария, ни предаварийное состояние никак не проявляются).

При превышении показателем аномальности установленного предела ($p > p_{пр}$) должна быть сформирована сигнализация. По поводу этой сигнализации необходимо сделать следующие замечания:

значение $p_{пр}$ должно быть установлено в процессе опытной эксплуатации оборудования и должно обеспечивать почти гарантированное отсутствие ложных срабатываний;

для снижения вероятности ложных срабатываний дискретный сигнал $p > p_{пр}$ должен вызывать сигнализацию с задержкой по переднему фронту с временем задержки примерно несколько часов или показатель аномальности должен быть "сглажен" применением фильтров (апериодическое звено, фильтр скользящего среднего);

дискретный сигнал $p > p_{пр}$ должен вызывать только сигнализацию. Решение об отключениях, остановах принимает персонал электростанции.

ВЫВОДЫ

1. Для решения ключевой проблемы метода MSET, связанной с выбором состава входных и выходных переменных для построения регрессионной модели, использован ансамбль регрессионных моделей. Для расчета итогового показателя аномальности применена логистическая регрессионная модель.

2. В результате экспериментальных исследований предложенного метода на реальных данных функционирования питательного насоса ПТН 1100-350-17-4 построена модель, обладающая хорошими обобщающими и предиктивными способностями (AUC ROC ≈ 0.98), что позволяет использовать ее в качестве основы для создания системы мониторинга состояния объекта и предсказания будущих аварий.

3. Предложенный метод требует дополнительного апробирования на других архивных выборках, в режимах офлайн и онлайн на текущих данных, совершенствования в части прогнозирования сроков выхода на предаварийное и аварийное техническое состояние, определения конкретных причин возникновения дефекта.

СПИСОК ЛИТЕРАТУРЫ

1. **Siegel E.** Predictive analytics: The power to predict who will click, buy, lie, or die. John Wiley & Sons Incorporated, 2016.
2. **Waller M.A., Fawcett S.E.** Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management // J. Business Logistics. 2013. № 34(2). P. 77–84.
3. **Громак Е.В., Наумов С.А., Шишов В.А.** Система удаленного мониторинга и прогностики АО "РОТЕК" как элемент энергетической безопасности // Новое

в российской электроэнергетике. 2016. № 6. С. 36–46.

4. **Random** forest as a predictive analytics alternative to regression in institutional research / L. He, R.A. Levine, J. Fan, J. Beemer, J. Stronach // Practical Assessment, Research & Evaluation. 2018. № 23(1). P. 1–16.
5. **Липатов М.** Первый в России комплекс предиктивной аналитики для энергетического и промышленного оборудования // Экспозиция. Нефть. Газ. 2016. № 3 (49). С. 82–83.
6. **Cheng S., Pecht M.** Multivariate state estimation technique for remaining useful life prediction of electronic products // AAAI Fall Symposium on Artificial Intelligence for Prognostics. Arlington, 2007. P. 26–32.
7. **Zavaljevski N., Gross K.C.** Sensor fault detection in nuclear power plants using multivariate state estimation technique and support vector machines. Argonne National Lab. (USA), 2000. ANL/RA/CP-103000.
8. **Dietterich T.G.** Ensemble learning // The handbook of brain theory and neural networks. 2002. Issue 2. P. 110–125.
9. **Tresp V.** Committee machines // Handbook for neural network signal processing / Ed. by Y.H. Hu, J.-N. Hwang. CRC Press, 2001.
10. **Draper N.R., Smith H.** Applied regression analysis. John Wiley & Sons, 2014.
11. **Zhang Cha, Yunqian Ma.** Ensemble machine learning: methods and applications. Springer Science & Business Media, 2012.
12. **Hanley J.A., McNeil B.J.** The meaning and use of the area under a receiver operating characteristic (ROC) curve // Radiology. 1982. № 143 (1). P. 29–36.
13. **Hosmer D.W., Jr., Lemeshow S., Sturdivant R.X.** Applied logistic regression. John Wiley & Sons, 2013. ISBN: 9780470582473.

Model for Early Detection of Emergency Conditions in Power Plant Equipment Based on Machine Learning Methods

A. A. Korshikova[a, *] and A. G. Trofimov[b]

[a] *OOO Incontrol, Moscow, 115280 Russia*

[b] *National Research Nuclear University Moscow Engineering Physics Institute, Moscow, 115409 Russia*

**e-mail: aakorshikova@gmail.com*

Received June 19, 2018; in final form, August 21, 2018; accepted August 29, 2018

Abstract—The article discusses a method for early detection and prediction of abnormality in operation of power-unit process equipment taking as an example the PTN 110-350-17-4 turbine driven feedwater pump of a 300 MW power unit. The importance of the problem of predicting possible process equipment malfunctions at an early state of their occurrence is determined, and the specific features of solving it in the power industry are explained. The range of process equipment defects that can be efficiently detected using the predictive analytics methods is outlined. The fundamental assertion stating that the scope of analog and discrete measurements available in the process control system's set of computerized automation tools is sufficient for applying the predictive analytics methods is emphasized. Modern predictive analytics methods are briefly reviewed, and the specific features of model training algorithms are mentioned. Separate attention is paid to the problems of preparing initial data for training the model. The mathematical problem of modeling an abnormality indicator taking the values from 0 (normal operation) to 1 (abnormal operation) is formulated. In turn, this problem is formulated as the binary classification problem of attribute vectors characterizing the equipment state at the given moment of time. An original approach is suggested, which combines the multivariate state estimation technique (MSET), in which the degree of abnormality in a technical state is determined from the extent to which the Hotelling criterion exceeds a threshold level (which is automatically calculated in the algorithm), and machine learning methods, the use of which makes it possible to overcome a number of difficulties inherent in the MSET. For solving the problem of determining the composition of the most informative attributes from the values of which early development of an emergency can be detected, it is proposed to use an ensemble of regression models. A method for selecting the modeled variable and the set of regressors is substantiated. An abnormality indicator calculation method based on composing an ensemble of linear regression models is proposed, and the advantage of using an ensemble over a single classifier is shown. A method for producing an alarm in response to detected abnormality in the operation of power unit process equipment is suggested. It is shown that it became possible by using the proposed model to detect the onset of the emergency development process, whereas individual indicators failed to reveal pump operation singularities in the preemergency interval of time.

Keywords: process equipment, detection of abnormalities, predictive analytics, committee of classifiers, logistic regression