СОЗДАНИЕ СИСТЕМ ПРЕДИКТИВНОЙ АНАЛИТИКИ ДЛЯ ЭНЕРГЕТИЧЕСКИХ ОБЪЕКТОВ

Андрюшин А.В., Щербатов И.А., Цуриков Г.Н., Титов Ф.М.
Московский энергетический институт
AndriushinAV@mpei.ru, ShcherbatovIA@mpei.ru, grishatsurikov9826@yandex.ru,
fedor.titov.335@gmail.com

Аннотация: Показана необходимость прогнозирования значений технологических параметров в рамках решения задачи прогнозирования технического состояния энергетического оборудования для целей организации ремонтов по состоянию. Рассмотрена архитектура и методы предобработки больших объемов данных и прогнозирования значений параметров. Показана эффективность искусственных нейронных сетей при решении данных задач.

Ключевые слова: предиктивная аналитика, техническое состояние, предобработка данных, прогноз, искусственная нейронная сеть, энергетика.

Введение

Задача организации ремонтов оборудования энергетики по его фактическому состоянию является крайне актуальной, что косвенно характеризуется увеличением количества работ,

посвященных данной тематике. Для организации ремонтов оборудования при данном подходе требуется оценка его технического состояния. При этом актуальность приобретает необходимость заранее устанавливать необходимость ремонта. В этой связи имеет место задача прогнозирования технического состояния оборудования энергетики. Построение точного прогноза времени выхода из строя (отказа или аварии) оборудования позволит запланировать соответствующие мероприятия и ресурсы (финансовые, людские, материалы и машины) для приведения конкретной единицы оборудования в надлежащее техническое состояние.

Для прогнозирования в основном используются характеристики, основанные на экспертных знаниях, сформулированных в процессе эксплуатации оборудования эвристиках, а также предположении о том, что динамика изменения ряда параметров свидетельствует об отказе оборудования с определенной уверенностью через некоторый интервал времени. Такие системы называются предиктивными или прогнозными. Системы предиктивной аналитики (СПА) находят все большее применение в различных областях в связи с увеличением возможностей средств вычислительной техники по программной реализации математических методов обработки больших массивов данных. Неоднородность больших данных (Big Data) [1] и методы их обработки неразрывно связаны при реализации СПА. Проведем обзор и анализ существующих подходов к прогнозированию технического состояния оборудования энергетики, систем предиктивной аналитики, которые используются для этих целей, а также методов и алгоритмов обработки больших объемов данных.

1 Методы диагностики и прогнозирования технического состояния

В систему предиктивной аналитики поступает неструктурированная телеметрическая информация в реальном масштабе времени от различных информационных и автоматизированных система, таких как АСКУЭ и SCADA, а также учетных систем (например, фиксирующих дефекты, результаты диагностики и пр.), MES, ERP и EAM. Использование накопленной информации в скором будущем должно обеспечит энергетике качественный скачок вперед [3]. Наличие систем предиктивной аналитики может являться существенным конкурентным преимуществом современного предприятия энергетики.

Для большинства систем предиктивной аналитики характерна следующая структура [4]: импорт данных в систему; преобразование и подготовка данных; кластеризация для выявления тех данных, которые поступают затем в предиктивную (прогнозную) модель; прогнозирование трендов и отказов на основе предиктивной модели.

Для отнесения текущего технического состояния к одному из возможных состояний часто применяются различные алгоритмы кластеризации. Например, в системе предиктивной аналитики KNIME [4] для кластеризации данных используется алгоритм k-средних и авто регрессионная модель прогнозирования значений параметров. Алгоритм k-средних один из базовых алгоритмов кластеризации в системах данного класса. Но для него заранее необходимо знать число кластеров, а кроме того, данный алгоритм крайне чувствителен к выбору первоначальных центров кластеров. Авто регрессионные модели чувствительны к выбору порядка, что в свою очередь влияет на время (скорость вычислений снижается при повышении порядка регрессионной модели) получения прогноза.

Одним из наиболее часто применяемых способов прогнозирования технического состояния оборудования, в том числе в энергетике, являются моделирование с использованием временных рядов. Фактически производится анализ временных рядов параметров, которые характеризуют техническое состояние оборудования. Временной ряд – сумма двух составляющих (функции тренда и ошибки) [5].

Для прогнозирования на основе ретроспективных данных могут применяться системы машинного обучения на базе искусственных нейронных сетей (ИНС) [6].

В методологии IDARTS (Intelligent Data Analysis and Real-Time Supervision) используются решения, основанные на знаниях и правилах, применяемых для принятия решений, выявления неисправностей, потенциальных отклонений и иных критических для функционирования оборудования событий [7]. Практическая реализация IDARTS – мультиагентная система, где присутствует ряд агентов, которые отвечают за мониторинг оборудования, мониторинг подсистем, объединяющих несколько единиц оборудования и пр.

Для решения задач прогнозирования могут применяться методы глубокого (глубинного) обучения (Deep Learning) [8], например, многослойные нейронные сети различных видов. Эти технологии могут применяться для мониторинга условий эксплуатации, выявления зарождающихся

дефектов, диагностики первопричин сбоев [9], диагностики и классификации неисправностей оборудования [10].

Достоинства и недостатки классических методов построения моделей прогнозирования (без учета нейронных сетей и систем нечеткого логического вывода) на основе данных и временных рядов подробно изложены в источниках [11-12].

Для прогнозирования возникновения аварий и отказов в энергетике могут применяться различные модификации Байесовского подхода и Байесовских сетей. В [13] предложена динамическая байесовская сеть доверия, позволяющая прогнозировать значения вероятностей отказов и осуществлять поиск дефектов и неисправностей в соответствующих системах поддержки принятия решений.

К интеллектуальным методам, которые могут применяться в системах диагностики технического состояния оборудования и предиктивной аналитики энергетических объектов можно также отнести эволюционное программирование. Например, генетические алгоритмы могут применяться для реализации поиска в больших данных [14].

Проведенный критический анализ литературных источников в области обработки больших данных и временных рядов для целей реализации оценки состояния оборудования и прогнозирования его надежности позволяет сделать вывод о целесообразности разработки оригинального подхода создания систем предиктивной аналитики для предприятий энергетики.

2 Структура системы предиктивной аналитики

Качество прогнозирования технического состояния оборудования энергетики и эффективность функционирования других систем предприятия, использующих данные прогнозов, зависит от архитектуры и функциональности системы предиктивной аналитики. Процессы обработки данных, которые реализуются в современной СПА технологического оборудования для предприятий энергетики, представлены на рис. 1.

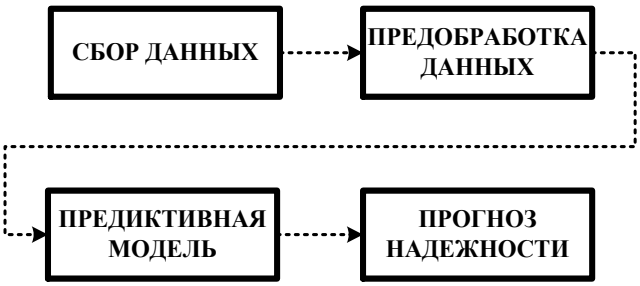


Рис. 1. Процессы обработки больших данных в СПА

Сбор данных большого объема сопряжен с рядом трудностей, т.к. они поступают от различных систем более низкого уровня, например, систем SCADA и/или АСКУЭ, а также более высокого уровня, таких как PLM, MES, EAM. При этом сами данные являются неоднородными, имеют отличающиеся типы, структуру передачи и способы хранения. Одна из ключевых подзадач получения прогноза технического состояния и других показателей, характеризующих функционирование оборудования энергетики в различных режимах, фактически должна решаться с использованием технологии промышленного интернета вещей (Industrial Internet of Things, IIoT) [15] и выходит за рамки рассмотрения в данной работе. Поэтому сделаем следующее упрощение.

Сбор данных с различных подсистем и от IIoT производится требуемым образом, данные поступают в режиме реального времени с дискретизацией, обеспечивающей необходимые показатели качества и скорости реализации процессов предиктивного анализа оборудования энергетики.

Вторая актуальная подзадача – предобработка данных, целью которой является получение генеральной совокупности данных как основы для синтеза или коррекции предиктивной модели, а также получения непосредственно прогноза. Поступающие данные должны быть по возможности отфильтрованы от помех, шумов и ошибок, а также привязаны к реальному масштабу времени. Совокупности данных должны быть равномерными, нормализованными, а в ряде случаев их требуется получить с использованием специальных методов, например, когда данных не достаточно для реализации предиктивной модели или ее корректировки.

Для реализации СПА разработана структура, содержащая основные модули, реализующие один из ключевых подходов Индустрии 4.0 кибер-физические объекты (рис. 2).

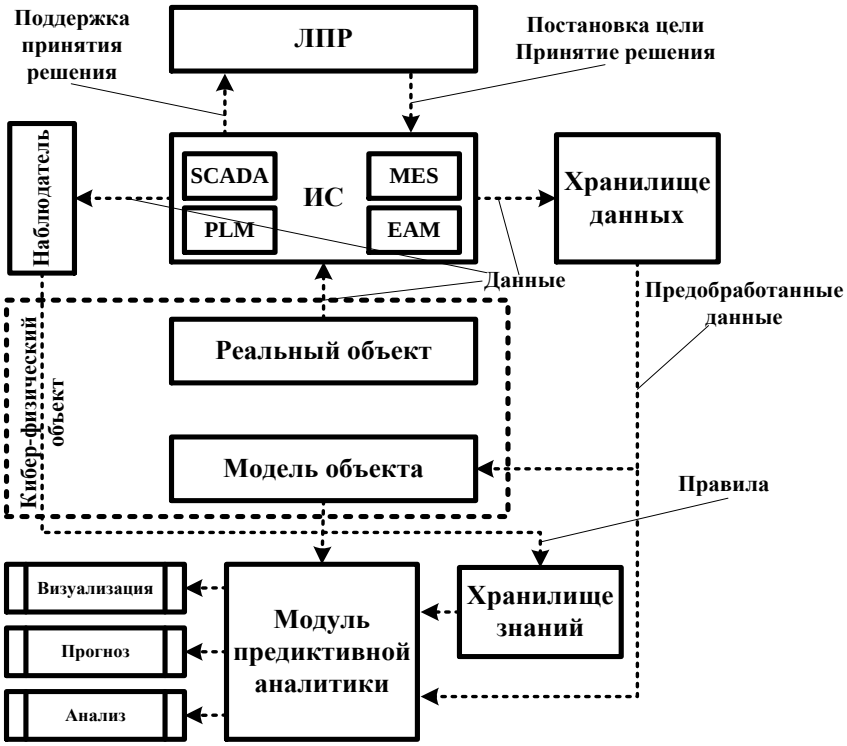


Рис. 2. Структура системы предиктивной аналитики

На структурной схеме представлены основные системы и подсистемы, а также информационные потоки и потоки данных, обеспечивающие прогнозирование трендов ключевых параметров, технического состояния оборудования энергетики и принятия решений.

Согласно концепции кибер-физических систем данные от датчиков, установленных на оборудовании, поступают в соответствующие информационные системы (ИС), к которым можно отнести SCADA, MES, PLM, EAM. Эти данные доступны для ЛПР, решающего задачу постановки глобальной цели системы энергетического оборудования. Данные информационные системы обеспечивают поддержку принятия решений при управлении жизненным циклом производственных активов предприятий энергетики, оптимизации генерации или транспорта энергии и т.д.

Таким образом, разрабатываемая СПА является частью общего подхода по цифровизации предприятий энергетики, которая в свою очередь направлена на повышение эффективности их функционирования.

Данные поступающие от внешних по отношению к системе предиктивной аналитики ИС помещаются в хранилище данных. В указанную подсистему поступают только те данные, которые в дальнейшем могут быть использованы для оценивания и прогнозирования ТС оборудования, отказов и аварий. Те данные, которые фактически используются в конкретный момент времени для построения прогнозов, имеют статус предобработанных и поступают на вход СПА, а также подаются на цифровую модель объекта (единицы энергетического оборудования). Выход, рассчитанный по модели объекта, поступает на вход системы предиктивной аналитики.

Отдельный аспект функционирования СПА – использование знаний экспертов предметной области для совершенствования механизмов прогнозирования и анализа, а также увеличения точности оценки и прогнозирования технического состояния оборудования энергетики. Часть данных может использоваться в специальной подсистеме «наблюдателе», которая обеспечивает сбор данных для коррекции формализованных знаний экспертов, например, на основе продукционных правил. Система предиктивной аналитики на основе поступающих данных и имеющихся экспертных знаний обеспечивает прогнозирование ТС оборудования, характерных показателей и трендов, их интерпретацию и анализ, а также визуализацию происходящих в оборудовании процессов.

3 Сбор данных в системе предиктивной аналитики

Данные, поступающие в систему предиктивной аналитики, влияют на ряд аспектов, обуславливающих ключевые свойства ее применения: эффективность прогнозирования как мера минимизации затрат на поддержание оборудования энергетики в надлежащем техническом состоянии (обеспечении заданной надежности); качество прогнозирования как объективная характеристика возможности использования полученного прогноза для управления оборудованием или принятия решений на уровне предприятия.

Сформулируем общие требования к сбору данных в системах предиктивной аналитики для предприятий энергетики:

- привязка ко времени – данные, поступающие в систему, должны быть привязаны к реальному времени, при этом данные учетных систем, куда сведения вносятся по результатам осмотра (например, определение дефектов при визуальном осмотре) должны вносятся максимально оперативно для наиболее точной их привязки к моменту времени возникновения;
- требуемая частота дискретизации – данные поступают в систему с такой частотой дискретизации, которая обеспечивает возможность прогнозирования трендов технологических параметров и технического состояния оборудования;
- достаточный объем – объем данных, поступающих в систему, является необходимым и достаточным для прогнозирования технического состояния оборудования;
- целостность – поступающие данные не должны быть повреждены, изменены и пр. при передаче в систему;
- непротиворечивость – отсутствие данных, которые являются противоречивыми и, как следствие, ухудшающими прогностические возможности системы предиктивной аналитики.

Структурная схема сбора данных системы предиктивной аналитики представлена на рис. 3 (ЕО – единица оборудования энергетики, КО – компонента крупномасштабной энергетической системы [16]).

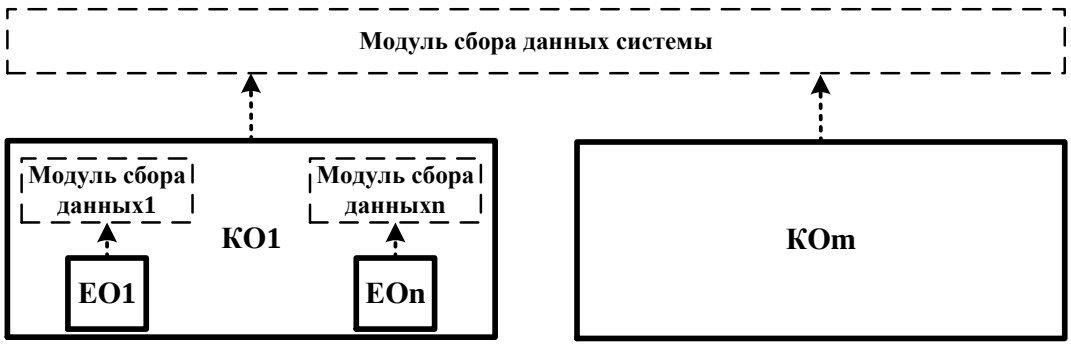


Рис. 3. Сбор данных системы предиктивной аналитики

Необходимо отметить, что согласно концепции промышленного интернета вещей целесообразно обеспечить максимальную унификацию аппаратных средств и протоколов передачи данных.

4 Предобработка данных

Ранее мы условились, что в данных, поступающих на вход синтезированного алгоритма, отсутствуют пропуски и выбросы. Рассмотрим случай, когда указанные аномалии присутствуют в измерительных данных систем предиктивной аналитики. Первый случай – наличие пропусков в получаемых данных. В литературных источниках описано большое число методов и алгоритмов восстановления данных при наличии пропусков [17-19]. Общая черта данных методов – заполнение пропущенного одного или нескольких значений.

Пусть $\hat{D}=\left(\hat{D}_1,\ldots,\hat{D}_n\right)$ – измеренные значения технологических параметров и характеристик оборудования (векторы), взятые в моменты времени $t_1, \ldots, t_p$, где p – количество значений (выборка или выгрузка из базы данных) или число ретроспективных периодов, а матрица $M_D=\left\|\hat{d}_{ij}\right\|$, $i=\overline{1,n}$, описывает пространство решаемой задачи восстановления пропущенных данных. Обозначим $\hat{d}_{ij}^{*}$

восстановленное значение пропущенного (отсутствующего) в выборке значения $\hat{d}_{ij}$. Восстановление единственного пропущенного значения в выборке по параметру D_i (столбцу M_D^i матрицы M_D) есть нахождение функциональной зависимости $FP_i(D_i)$ для каждого $\hat{d}_{ij}^*$, т.ч. $\arg\min\left\|\hat{d}_{ij}^* - \hat{d}_{ij}\right\|$. Тогда задача восстановления единственного пропущенного значения в выборке по одному технологическому параметру может быть формализована в общем виде следующим образом:

$$\forall \hat{d}_{ij}^* \exists FP_i(D_i): \arg\min\left\|\hat{d}_{ij}^* - \hat{d}_{ij}\right\| \quad (1)$$

Для каждого столбца M_D^i существует не более одного пропущенного значения, иначе выборка разбивается на несколько выборок так, чтобы в каждой из них было не более одного пропущенного значения. Рассмотрим пример решения задачи восстановления пропущенных данных. Решение проводилось для нескольких методов (Cubic / Liner SVM и искусственные нейронные сети с алгоритмом обучения Левенберга-Марквардта). Результаты представлены в табл. 1.

Таблица 1. Результаты экспериментов по применению алгоритмов восстановления пропусков

Пропуски	Регрессионные методы		Нейронная сеть	
	Алгоритм машинного обучения	СКО (RMSE)	Алгоритм машинного обучения	СКО (RMSE)
1 столбец (10%)	Cubic SVM	0.27517	Levenberg-Marquardt	0.82636
1 строка (10%)	Liner SVM	0.37958	Levenberg-Marquardt	1.64428
3 столбца (30%)	Liner SVM	0.31582	Levenberg-Marquardt	1.1475
3 строки (30%)	Liner SVM	0.19736	Levenberg-Marquardt	4.8081
5 столбцов (50%)	Liner SVM	0.35404	Levenberg-Marquardt	2.3039
5 строк (50%)	Linear Regression	0.38985	Levenberg-Marquardt	2.6446

Результаты свидетельствуют о целесообразности применения регрессионных моделей для решения задачи восстановления пропусков в данных.

Следующей не менее важной задачей по предобработке больших массивов данных для СПА является необходимость выявления и устранения выбросов (аномально больших или малых значений технологических параметров, присутствующих в генеральной совокупности). Выброс – элемент, существенно отличающийся от других элементов выборки.

Пусть имеется $D = \{d_0,...,d_i,...\}, i \in \mathrm{N}^+$ – измеренные значения технологических параметров и характеристик. В системе предиктивной аналитики обрабатывается только часть (один сегмент) бесконечно большого (в общем виде, с учетом дискретности измерений и времени функционирования единицы оборудования) ряда измеренных значений $D_{\Delta t} = \{d_t\}, t \in \Delta t$. Тогда задача восстановления выброса или аномального значения в выборке по одному технологическому параметру может быть формализована в общем виде следующим образом:

$$\forall d_t \in D_{\Delta t} \exists FA(D_{\Delta t}) = \begin{cases} 0 \\ 1 \end{cases} \quad (2)$$

где $FA(D_{\Delta t})$ – функциональная зависимость, обеспечивающая восстановление выброса в выборке значений конкретного параметра, значение 0 – отсутствие аномалии (выброса), а 1 означает ее наличие.

Рассмотрим примеры решения задачи определения выбросов в данных, поступающих от средств измерений объектов энергетики в систему предиктивной аналитики. В качестве базового математического аппарата выберем искусственные нейронные сети. Обучающая выборка содержит информацию о температуре окружающего воздуха (X_1), давлении в конденсаторе паровой турбины (X_2) и значениях мощности парогазовой установки при данных параметрах. Для выборки, после

представления данных в трехмерном пространстве, отобрано 100 элементов, среди которых визуально легко определить значения, заметно отличающиеся от остальных. Для анализа точности определения выбросов с использованием разработанного способа первоначально выбросы были определены на основании методики из ГОСТ Р ИСО 16269-4-2017 (значения параметра мощность, определенные как выброс кодировались единицей, в противном случае нулем).

Выбросы значений для параметра мощность (Y) идентифицируются путем нахождения стьюдентизированной ошибки и ее последующим сравнением с показателем распределения Стьюдента (при доверительной вероятности 0.995, числе степеней свободы $k=n-p-2$, где n - количество элементов в выборке, p - количество переменных, влияющих на значения Y).

Для определения потенциальных выбросов по переменным X используется параметр h, характеризующий расстояние между значением, принимаемым случайной величиной X в i-й точке и средним арифметическим всех n значений, принимаемых X. Элемент выборки считается выбросом, если значение данного параметра меньше, чем результат выражения $h \leq 2(p+1)/n$.

Потенциальный выброс считается реальным если расстояние Кука и показатель DFFITS принимают определенные значения. Например, расстояние Кука сравнивается с показателем F-распределения (процентилем уровня, равным 50%, значением $p+1$, значением $n-p-1$), а показатель DFFITS с результатом выражения $2\sqrt{(p+1)/n}$.

Подготовленные таким образом данные использовались для обучения и тестирования искусственной нейронной сети. Результат идентификации выбросов составил 100%, т.е. все выбросы были определены верно (рис. 4). Среди 100 элементов выборки 9 выбросов идентифицированы верно, также как и 91 элемент, который к выбросам не относится.

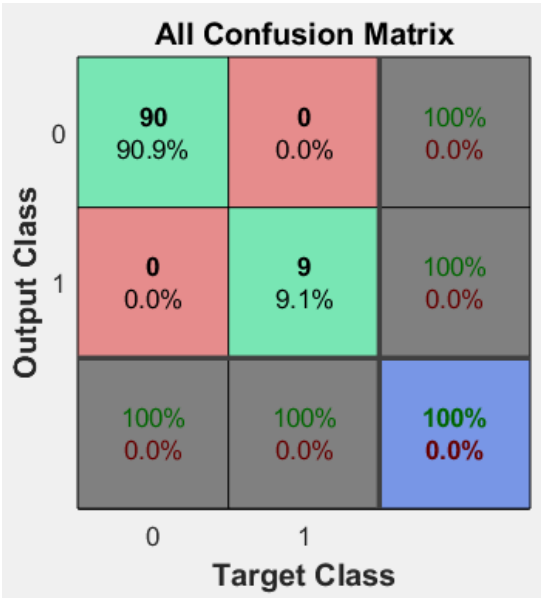
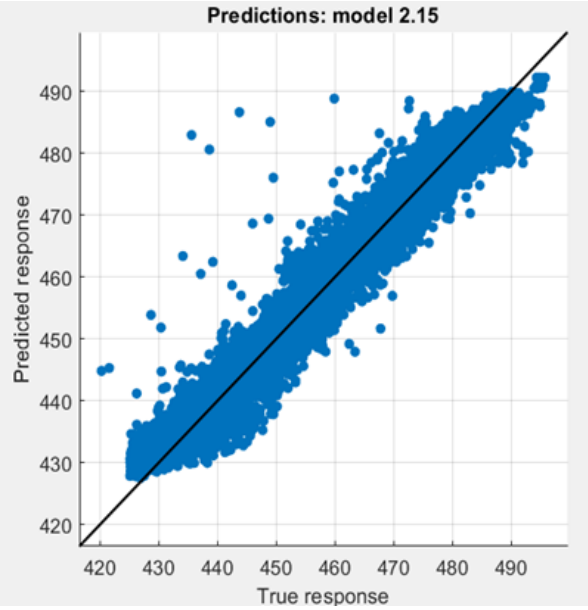


Рис. 4. Результаты работы искусственной нейронной сети

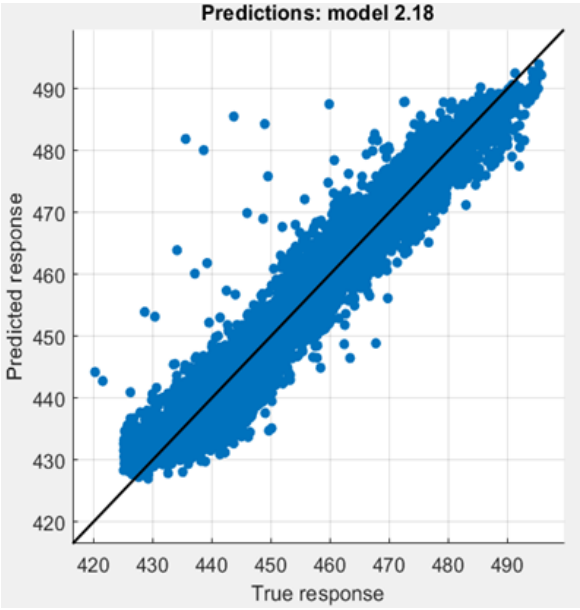
И, наконец, заключительным этапом является нормализация данных. Нормализация позволяет привести значения технологических параметров и вычисляемых показателей к одному интервалу значений, например, [0;1]. Это крайне необходимая процедура в связи с тем, что, например, в предиктивной модели могут присутствовать параметры, абсолютные значения которых (без учета размерностей) отличаются на несколько порядков. В работе применяется минимаксное масштабирование $D_i^{'} = \frac{D_i - D_{i,\min}}{D_{i,\max} - D_{i,\min}}$, позволяющее получить значения в интервале [0;1], где 1 соответствует максимальному значению в выборке, а 0 минимальному (для случая, когда значения выборки достаточно равномерно заполняют интервал).

5 Прогнозирование значений параметров

В качестве примера прогнозирования значений параметров покажем прогнозирование мощности Y в зависимости от значений X_1 и X_2. Для выбора наиболее подходящего метода прогнозирования были исследованы искусственные нейронные сети, SVM и алгоритм бэггинга деревьев принятия решений. На рис. 5 представлены результаты прогнозирования.



а) метод бэггинга деревьев решений



б) SVM-метод

Рис. 5. Результаты прогноза с помощью различных методов

При использовании алгоритма бэггинга деревьев решений и SVM-алгоритма, среднеквадратичное отклонение прогнозируемой величины составило 3.4487 и 3.5648 МВт соответственно, а применение искусственных нейронных сетей показало результат 3.4271 МВт. При этом именно нейронные сети показали наименьшее время получения результата.

Заключение

Разработана структура системы предиктивной аналитики для крупномасштабных энергетических систем. Прогностические возможности системы могут использоваться для прогнозирования значений различных технологических параметров, которые являются характеристиками технического состояния оборудования энергетики. Решена задача подготовки данных для использования в системах предиктивной аналитики. В качестве базового математического аппарата применяется высокоэффективные методы искусственных нейронных сетей, которые обеспечивают высокую точность прогнозирования при требуемой скорости получения прогноза.

Литература

1. *Wang J., Zhang W., Shi Y., Duan S., Liu J.* Industrial Big Data Analytics: Challenges, Methodologies, and Applications. 2018 : [Электронный ресурс] Режим доступа : https://arxiv.org/ftp/arxiv/papers/1807/1807.01016.pdf.
2. *Bi Z.M., Cochran D.S.* Big data analytics with applications // Journal of Management Analytics. 2014. Vol. 1, No. 4. P. 249–265.
3. *Chidambaram V., Evans H., Etheredge K.* Big Data: Is the Energy Industry Starting to See Real Applications? // Supply Chain Management Review. 2015. No. 12. P. 62-64.
4. *Silipo R., Winters P.* Big Data, Smart Energy, and Predictive Analytics Time Series Prediction of Smart Energy Data. 2013. 2009. [Электронный ресурс]. - Режим доступа: https://files.knime.com/sites/default/files/inline-images/knime_bigdata_energy_timeseries_whitepaper.pdf.
5. *Шевцов Ю.Д., Дудник Л.Н., Арефьева С.А., Фадеев Е.Д.* Прогнозирование параметров технического состояния двигателей энергетических установок // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2017. № 132. С. 508-517.
6. *Shin S.J., Meilanitasari P.* Developing a big data analytics platform for manufacturing systems: architecture, method, and implementation // International Journal of Advanced Manufacturing Technology. 2018. P.1-42.
7. *Peres R.S., Rocha A.D., Leitao P., Barata J.* IDARTS - Towards Intelligent Data Analysis and Real-Time Supervision for Industry 4.0 // Computers in Industry. 2018. P. 1-12.

8. *Wang J., Ma Y., Zhang L., Gao R.X., Wu D.* Deep learning for smart manufacturing: Methods and applications // Journal of Manufacturing Systems. 2018. Vol. 48. P. 144-156.
9. *Park J.K., Kwon B.K., Park J.H., Kang D.J.* Machine learning-based imaging system for surface defect inspection. International Journal of Precision Engineering and Manufacturing-Green Technologies. 2016. Vol. 3. No. 3. P. 303–310.
10. *Zhao R., Yan R., Chen Z., Chen Z., Mao K., Wang P., et al.* Deep learning and its applications to machine health monitoring: a survey. 2016 : [Электронный ресурс] Режим доступа : https://arxiv. org/pdf/1612.07640.pdf.
11. *Букреев В.Г., Колесникова С.И., Янковская А.Е.* Выявление закономерностей во временных рядах в задачах распознавания состояний динамических объектов: монография. – Томск: Изд-во Томского политехнического университета. 2010. – 254 с.
12. *Рыбалко В.В*. Математические модели контроля надёжности объектов энергетики. Монография. - ГОУВПО СПбГТУРП. СПб., 2010. - 151 с.
13. *Вычужанин В.В., Шибаева Н.О.* Информатизация прогнозирования риска структурно сложных технических систем с помощью моделей байесовских сетей доверия // Информатика и математические методы в моделировании. 2016. Том. 6. №. 3. С. 205-214.
14. *Goldberg D.E.* Genetic algorithms in search, optimization, and machine learning. – Reading, MA: Addison-Wesley, 1989. 372 p.
15. *Цуриков Г.Н., Щербатов И.А.* Применение промышленного интернета вещей на объектах энергетики // Мехатроника, автоматика и робототехника. 2018. № 2. С. 97-100.
16. *Щербатов И.А.* Управление сложными слабоформализуемыми многокомпонентными системами. -Ростов н/Д: ЮНЦ РАН, 2015. - 288. с.
17. *Литтл Р.Дж.А., Рубин Д.Б.* Статистический анализ данных с пропусками. - М.: Финансы и статистика, 1990. – 336 с.
18. *Злоба Е., Яцкие И.* Статистические методы восстановления пропущенных данных // Computer Modelling & New Technologies. -2002. - Vol. 6.- № 1. - Pp. 51-61.
19. *Абраменкова И.В., Круглов В.В.* Методы восстановления пропусков в массивах данных // Программные продукты и системы. 2005. № 2. С. 18-22.