Б

УДК 004.942

С. В. Беспалова, С. М. Романчук, Т. В. Ермоленко, В. И. Бондаренко
Государственное образовательное учреждение высшего профессионального образования
«Донецкий национальный университет»
83001, г. Донецк, ул. Университетская, 24

ПОСТРОЕНИЕ ПРЕДСКАЗАТЕЛЬНЫХ МОДЕЛЕЙ ПАРАМЕТРОВ ДАВЛЕНИЯ ВОДЫ В ВОДОРАСПРЕДЕЛИТЕЛЬНЫХ СЕТЯХ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

S. V. Bespalova, S. M. Romanchuk, T. V. Yermolenko, V. I. Bondarenko
State Educational Institution of Higher Education «Donetsk National University»
83001, c. Donetsk, University str., 24

CONSTRUCTION OF PREDICTIVE MODELS OF THE PARAMETERS OF WATER PRESSURE IN WATER DISTRIBUTION NETWORKS BY MEANS OF MACHINE TRAINING METHODS

С. В. Беспалова, С. М. Романчук, Т. В. Ермоленко, В. І. Бондаренко
Державна освітня установа вищої професійної освіти
«Донецький національний університет», м. Донецьк
83001, м. Донецьк, вул. Університетська, 24

ПОБУДОВА ПЕРЕДБАЧУВАНИХ МОДЕЛЕЙ ПАРАМЕТРІВ ТИСКУ ВОДИ У ВОДОРОЗПОДІЛЬНИХ МЕРЕЖАХ ЗА ДОПОМОГОЮ МЕТОДІВ МАШИННОГО НАВЧАННЯ

В статье исследуется эффективность методов машинного обучения для прогнозирования поведения водопроводной сети. Построены предсказательные модели значений давления воды в контрольных точках на сети. Проведен сравнительный анализ качества построенных моделей.
Ключевые слова: мультиколлинеарность, градиентный бустинг, нейросеть.

The effectiveness of machine learning methods for predicting the behavior of the water supply network is searched in the article. Predictive models of water pressure values at control points on the network are constructed. A comparative analysis of the quality of the constructed models is made.
Key words: multicollinearity, gradient boosting, neuronet.

У статті досліджується ефективність методів машинного навчання для прогнозування поведінки водопровідної мережі. Побудовані передбачувальні моделі значень тиску води в контрольних точках на мережі. Проведено порівняльний аналіз якості побудованих моделей.
Ключові слова: мультиколінеарність, градієнтний бустінг, нейромережа.

Введение

Задачи поддержания эффективных и оптимальных режимов работы объектов водораспределительной сети, позволяющих обеспечить экономию электроэнергии и воды, а также задачи распознавания внештатных режимов и аварийных ситуаций являются наиболее важными для предприятий водоснабжения. В настоящее время работой водопроводных узлов в большинстве случаев управляют ситуационно, принимая во внимание установленные режимы водоподачи, полученные с помощью гидравлических расчетов водопроводных сетей при проектировании. Использование этих методов при оперативном управлении затруднительно, что связано с необходимостью получения большого количества исходных данных, крайне высокой их неопределенностью и низкой достоверностью, чрезмерно большими затратами машинного времени на проведение расчетов даже при использовании мощных современных ЭВМ. Все чаще для поддержания эффективных и оптимальных режимов работы водораспределительных систем применяются методы интеллектуального анализа данных, с использованием информации базы данных о напорах на насосных станциях и значениях давления в диктующих точках водопроводных сетей.

В статье исследуется эффективность методов машинного обучения для прогнозирования поведения водопроводной сети с использованием онлайн-данных автоматизированной системы сбора и передачи информации. Выходные данные используются для оптимального управления режимами работы насосных станций и для обнаружения аномалий давления в сети вследствие прорывов.

1 Постановка задачи

Источником информации являются датчики давления, расположенные на насосных станциях. А также автономные системы измерения давления, расположенные в контрольных точках на водоводах сети. В рамках данной работы через x_i (i=1, …, 5) обозначены показатели давления воды на насосных станциях, через y_j (j=1, 2, 3) – на сетях.

Замеры производились автоматически каждые 30 мин. с передачей на сервер центральной диспетчерской по GPRS-каналу.

Временное окно полученных данных: с 14.02.2012 по 30.07.2012.

Необходимо построить предсказательную модель значений давления воды в контрольных точках на сети.

Цель работы – построить предсказательные модели для показателей давления воды в неактивных точках, провести анализ эффективности их работы на тестовой выборке.

Для достижения поставленной цели необходимо решить следующие ***задачи***:

1) провести разведочный анализ данных для выявления выбросов и пропущенных значений, а также наличия связи между переменными;

2) используя методы машинного обучения, построить предсказательные модели для показаний датчиков давления воды в неактивных точках водораспределительной сети;

3) оценить значимость предикторов для каждой модели;

4) сравнить эффективность построенных моделей, используя в качестве критерия коэффициент детерминации:

$$R^2 = 1 - \sum \hat{\varepsilon}_i^2 \Big/ \sum \left(y_i - \bar{y}\right)^2,$$

где $\hat{\varepsilon}_i$ – отклонения выборочных величин y_i переменной-отклика от значений $\hat{y}_i$, получаемых по модели, $\bar{y}$ – среднее значение переменной-отклика y

Объектом исследования являются предсказательные модели и методы их применения для реализации практических задач анализа работы водораспределительных сетей.

Предметом исследования являются методы и алгоритмы машинного обучения в задачах моделирования процессов в водораспределительных сетях.

Методы исследования – методы интеллектуального анализа данных, методы машинного обучения, методы математической статистики.

2 Разведочный анализ данных

Предварительный анализ данных проводится с целью выявления наиболее общих закономерностей и тенденций, характера и свойств анализируемых данных, законов распределения анализируемых величин и подготовки данных для дальнейшего изучения. На этапе разведочного анализа проводится выявление аномалий (точек-выбросов), заполнение пропущенных значений, проводится процедура анализа распределений переменных, просмотр корреляционных матриц с целью поиска коэффициентов, превосходящих по величине определенные пороговые значения, факторный и дискриминантный анализ, многомерное шкалирование, визуальный анализ гистограмм и т.п. [1].

Результаты разведочного анализа не используются для выработки управленческих решений. Их назначение – помощь в разработке наилучшей стратегии углубленного анализа, выдвижение гипотез, уточнение особенностей применения тех или иных математических методов и моделей. Без разведочного анализа углубленный анализ данных будет производиться практически «вслепую».

На этапе разведочного анализа данные были синхронизированы по времени, ширина временной окна составляла не более 900 с. Длина временных рядов составила 1932 значения. Было выявлено небольшое количество пропущенных значений и выбросов.

Для заполнения пропущенных значений и выбросов использовалась сплайновая интерполяция со сплайнами Акима. Этот метод дает хорошие результаты для значения аппроксимированной функции и требует наличия информации о точках в области интервала интерполяции для определения кубических полиномиальных коэффициентов.

При анализе нескольких количественных переменных очень удобным инструментом выявления характера связи между переменными являются матричные диаграммы рассеяния. На рис. 1 изображен матричный график для переменных x_i, y_j и *time* (время в секундах). На рисунке диаграммы рассеяния организованы в форме матрицы (значения переменной по столбцу используются в качестве координат X, а значения переменной по строке – в качестве координат Y). Гистограммы, изображающие распределение каждой переменной, располагаются на диагонали матрицы.

Визуальный анализ выявил мультиколлинеарность переменных. Так, очевидна попарная линейная зависимость показаний, снятых с активных точек, а также показаний неактивной точки y_1 от всех остальных. Поведение переменной y_2 отличается от общей тенденции для y_1 и y_3, y_2 имеет двумодальное распределение и большую дисперсию, в отличие от y_1 и y_3, значения которых сосредоточены у максимума.

Стоит обратить внимание на поведение переменной y_3, ее значения в основном сосредоточены в диапазоне 0.8-0.9, но наличие некоторой части распределенных по всему диапазону y_3 значений сказывается на величине дисперсии и будет влиять на поведение построенных моделей, внося в них шум.

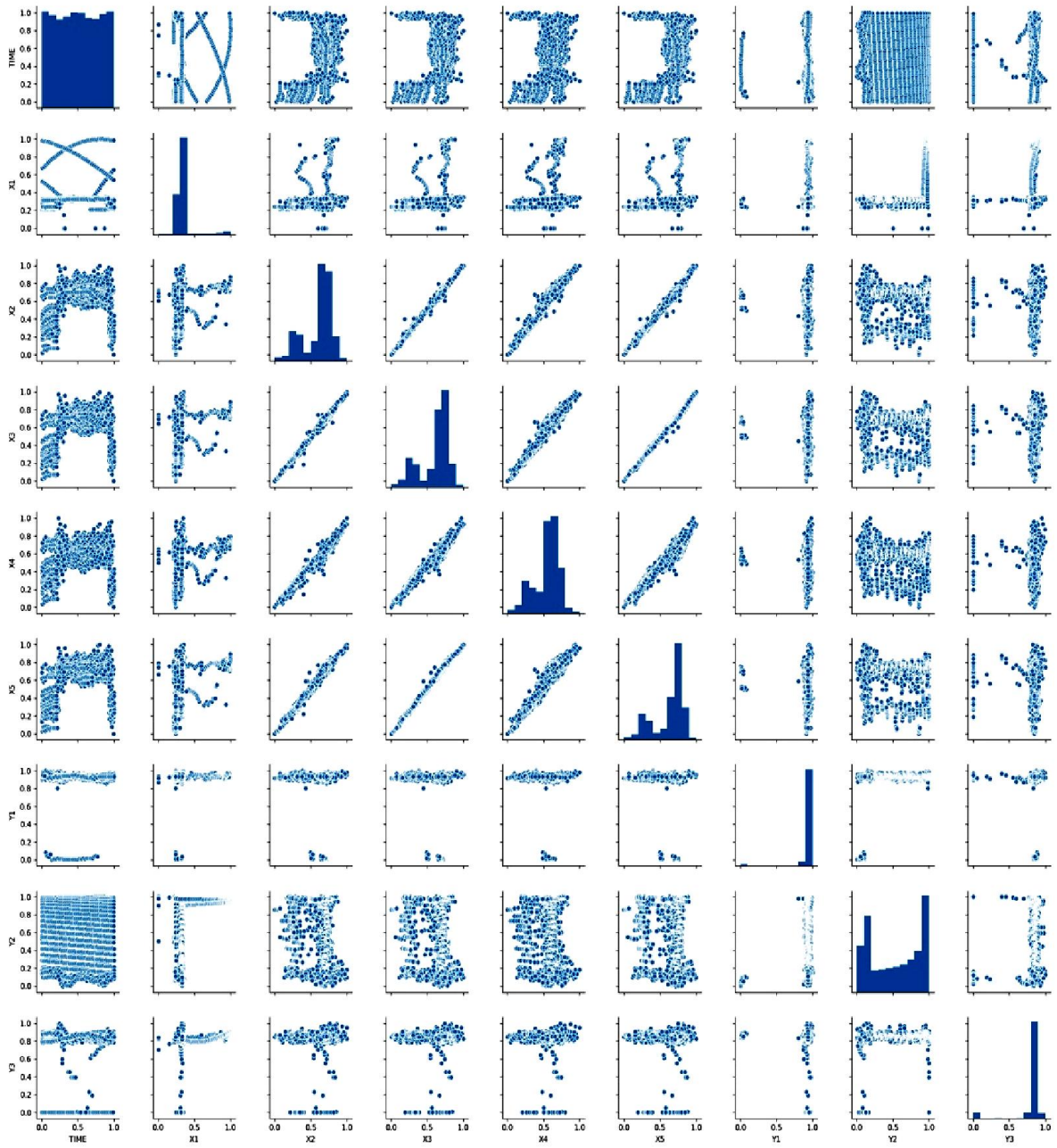

Рисунок 1 – Матричные диаграммы рассеяния,
полученные для данных с заполненными пропусками

Наличие мультиколлинеарности приводит к неустойчивости оценок параметров статистической модели, что выражается, в частности, в повышенной дисперсии этих оценок. Следовательно, необходимо использовать такие модели, которые не теряли бы своей эффективности для коллинеарных предикторов.

3 Сравнительный анализ моделей машинного обучения в условиях мультиколлинеарности переменных

Были использованы два подхода к построению предсказательных моделей: на основе ансамблей классификаторов и нейросетевого подхода.

Для построения ансамбля классификаторов была выбрана техника градиентного бустинга над решающими деревьями. Градиентный бустинг – одна из реализаций, в которой для минимизации функции потерь применяется градиентный спуск [2]. Этот выбор сделан из следующих соображений. Решающие деревья – достаточно слабые предсказывающие модели, но их преимущество в том, что мультиколлинеарность переменных не оказывает влияния на эффективность этих моделей. Значительно повысить качество предсказания позволяет градиентный бустинг. Таким образом, градиентный бустинг на основе регрессионных решающих деревьев – один из эффективных методов построения моделей регрессии в условиях мультиколлинеарности переменных.

Решающее дерево делит признаковое пространство (пространство предикторов) объектов на конечное семейство областей, каждому его листу соответствует своя область. При каждой итерации градиентного бустинга к уже имеющимся деревьям добавляется новое не просто, используя антиградиент функции ошибок, а проводится попытка определить, с каким коэффициентом его лучше добавить. То есть итоговый алгоритм является не суммой базовых, а их линейной комбинацией. Этот коэффициент называют скоростью (темпом) обучения.

Используемые в работе параметры алгоритма градиентного бустинга следующие:

- функция потерь – квадратичная;
- скорость обучения 0.1;
- максимальная глубина дерева 3;
- число классификаторов 500;
- минимальная сумма весов объектов в листе 1.2.

Увеличение этого параметра направлено на борьбу с переобучением.

Данные были разбиты на обучающую и тестовую выборки (доля обучающей выборки – 0.8). После чего были построены три модели, в качестве переменной-отклика поочередно брались y_j (j =1, 2, 3). Оценка эффективности каждой модели проводилась с помощью коэффициента детерминации R^2, в табл. 1 приведены значения R^2, полученные по тестовой и обучающей выборке для трех моделей.

Таблица 1 – Оценка эффективности моделей градиентного бустинга

Отклик / Выборка	y_1	y_2	y_3
Обучающая	0.997222081366871	0.9959418046593616	0.960156341153223
Тестовая	0.9955615264454241	0.9923361152018203	0.9037186382663568

По рис. 2 – 4 можно визуально оценить результаты прогнозирования каждой из моделей, на рисунках отображены предсказанные (pred) и реальные (test) нормированные значения тестовой выборки для переменных-откликов y_1-y_3, взятые за одни сутки.

Как и ожидалось (рис. 4), на предсказанные значения переменной y_3 оказала влияние небольшая часть значений в обучающей выборке, равномерно распределенных по всему диапазону изменения y_3.

Для каждой модели оценивалась важность предикторов, используя алгоритм Random Forest (случайный лес). Random Forest представляет собой ансамбль многочисленных деревьев решений [3]. Каждый из этих классификаторов строится на случайном подмножестве объектов обучающей выборки. Случайный лес имеет некоторые преимущества для использования его в качестве алгоритма оценки важ-

ности предикторов: он имеет очень мало настраиваемых параметров, относительно быстро и эффективно работает, что позволяет находить информативность признаков без значительных вычислительных затрат.

Алгоритм обучают на выборке и во время построения модели Random Forest для каждого элемента обучающей выборки вычисляется out-of-bag-ошибка – усредненная оценка функции потерь базовых алгоритмов на тех данных, на которых они не обучались. Чтобы оценить важность предиктора, его значения перемешиваются для всех объектов обучающей выборки и out-of-bag-ошибка считается снова. Важность предиктора оценивается путем усреднения по всем деревьям разности показателей out-of-bag-ошибок до и после перемешивания значений.

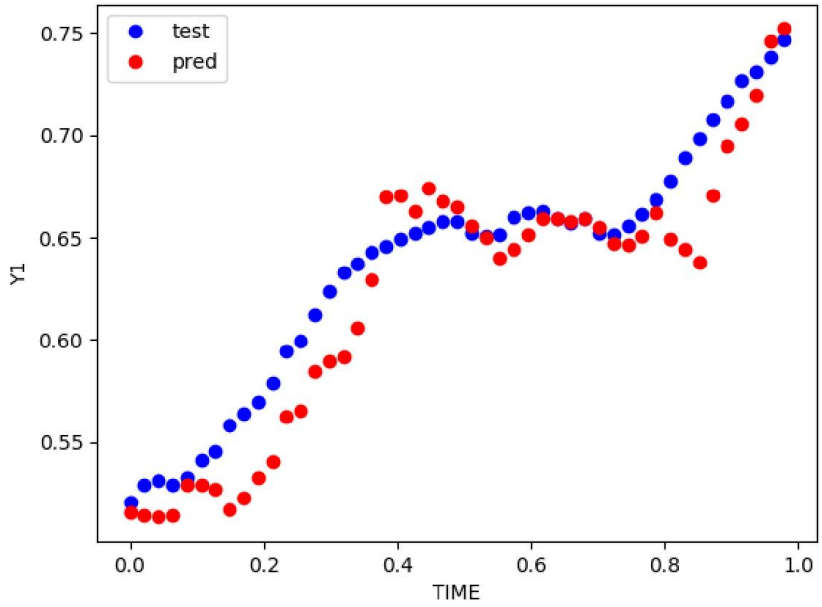

Рисунок 2 – Диаграмма рассеяния между нормированными переменными *time* и y_1, построенная для реальных и предсказанных значений переменной-отклика y_1

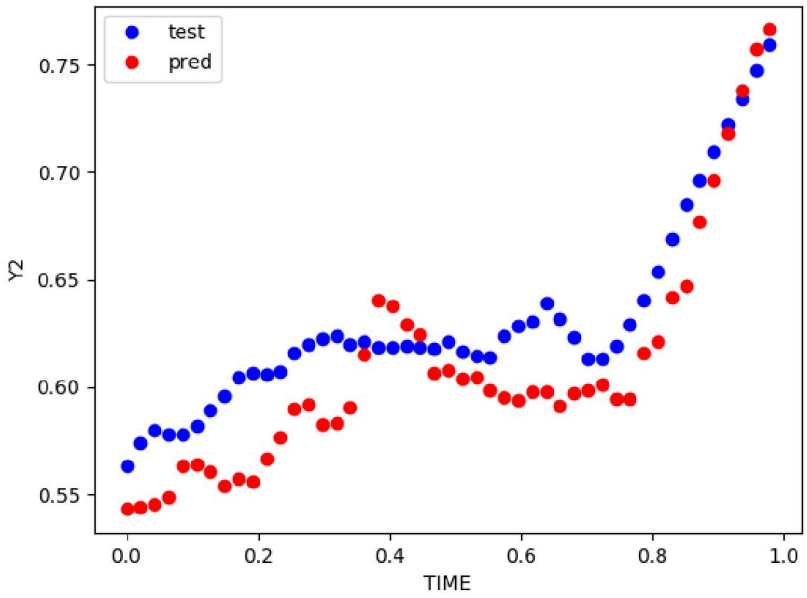

Рисунок 3 – Диаграмма рассеяния между нормированными переменными *time* и y_2, построенная для реальных и предсказанных значений переменной-отклика y_2

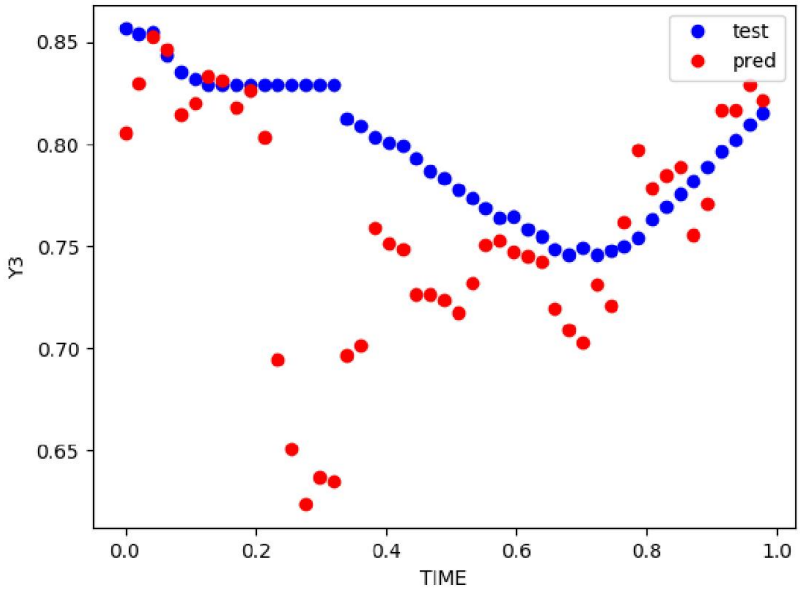

Рисунок 4 – Диаграмма рассеяния между нормированными переменными *time* и y_3, построенная для реальных и предсказанных значений переменной-отклика y_3

На рис. 5 – 7 приведены результаты оценки важности предикторов. Предикторы расположены по убыванию степени важности.

Для построения нейросетевой модели использовались рекуррентные нейросети (Recurrent Neural Networks, RNN). Это сети, содержащие обратные связи и позволяющие сохранять информацию. Такие сети тесно связаны с последовательностями и списками [4].

В RNN связи между узлами образуют ориентированный граф вдоль последовательности данных и выходы активации от нейронов распространятся в обоих направлениях (от входов к выходам и от выходов ко входам) [5]. Благодаря чему создаются петли в архитектуре нейронной сети («состояние памяти» нейронов). Это состояние позволяет нейронам запоминать набор последовательностей. То есть на каждом шаге сеть создает несколько копий самой себя. Каждая из этих копий принимает на вход текущее окно в определенный момент времени (определенную часть последовательности) и значение, полученное из предыдущей копии, затем их комбинирует и передает получившийся результат в следующий элемент. Таким образом, на каждом шаге мы фактически обучаем глубокую нейронную сеть, в которой столько слоев, сколько элементов в последовательности мы уже видели. Основное ее отличие состоит в том, что веса на каждом слое одинаковые, все слои делят одни и те же переменные между собой (shared weights).

Преимуществами RNN являются использование одной матрицы весов, что позволяет снизить затраты памяти, а также использование общих весов позволяет градиентам по весам не затухать до нуля сразу же.

Проблемами RNN является взрыв градиентов: если матрица весов такова, что заметно увеличивает норму вектора градиента при проходе через один «виртуальный слой» обратного распространения, получится, что при проходе через T-слоёв эта норма возрастет экспоненциально от T, т.к. используются общие веса. Эта проблема решается за счёт использования архитектуры с долгой кратковременной памятью (Long Short Term Memory, LSTM).

В LSTM-сетях внутренние нейроны «оборудованы» сложной системой так называемых ворот (gates), а также концепцией клеточного состояния (cell state), которая и представляет собой некий вид долгосрочной памяти. Ворота с помощью фильтров (входного, выходного и забывающего) определяют, какая информация попадет в клеточное состояние, какая сотрется из него и какая повлияет на результат, который выдаст RNN на данном шаге. Входной фильтр определяет, сколько информации из предыдущего слоя будет храниться в клетке. Выходной фильтр определяет, сколько информации получат следующие слои. Забывающий фильтр контролирует меру сохранения значения в памяти. Таким образом, нейроны LSTM-сети хорошо «помнят» недавно полученную информацию, но не имеют возможности надолго сохранить в памяти что-то, что обработали много циклов назад, какой бы важной та информация ни была.

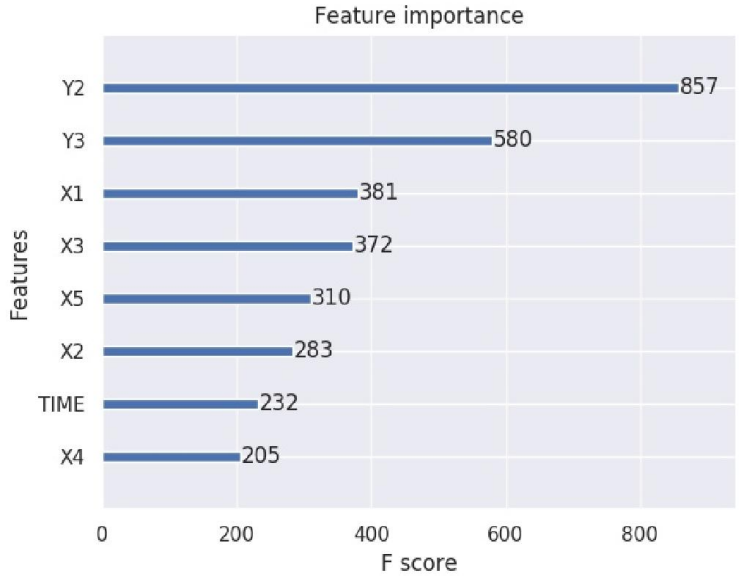

Рисунок 5 – Оценка важности предикторов для модели с переменной-откликом y_1

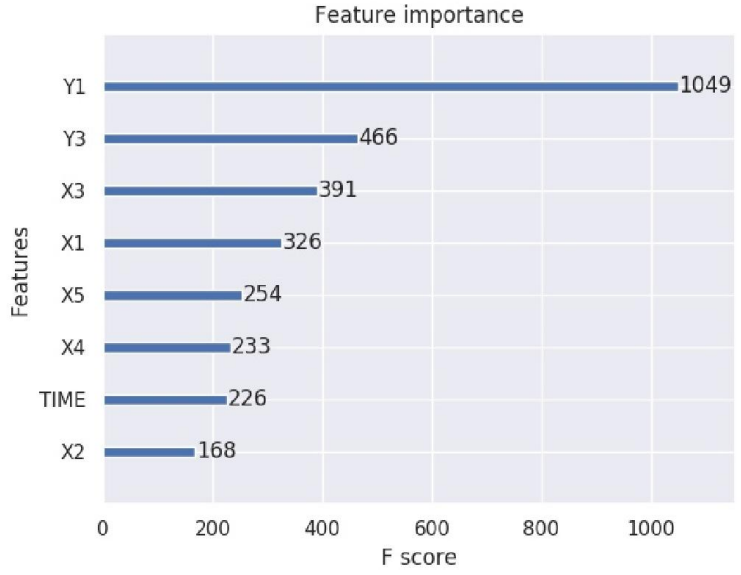

Рисунок 6 – Оценка важности предикторов для модели с переменной-откликом y_2

Рисунок 7 – Оценка важности предикторов для модели с переменной-откликом y_3

Хотя градиенты не затухают совсем, влияние текущего входа или текущего состояния сети, тем не менее, обычно не может распространяться слишком далеко. Влияние текущего входа затухает экспоненциально по мере удаления. Это серьезная проблема, которая не позволяет «обычным» рекуррентным сетям обучаться распознавать далекие зависимости в данных. То есть RNN к концу последовательности уже забывают, с чего там начиналось (последние элементы последовательности всегда будут гораздо важнее первых). Поэтому часто рассматривают так называемые двунаправленные рекуррентные сети (bidirectional RNN, BRNN).

BRNN представляет собой обычную RNN с двумя разными слоями: один слой читает последовательность слева направо, а другой – справа налево. Матрицы весов абсолютно независимы, между ними нет взаимодействия, просто для каждого элемента последовательности получатся два состояния: слева направо и справа налево.

Главная цель BRNN состоит в том, чтобы получить состояние, отражающее контекст и слева, и справа для каждого элемента последовательности. Поэтому основные примеры задач, где двунаправленные сети лучше обычных – это ряд задач, где важно учитывать всю входящую последовательность целиком. К таким задачам относится анализ временных рядов.

В данной работе, поскольку анализируемые данные представляют собой временные ряды, использовалась BLSTM-сеть, архитектура которой показана на рис. 8. Входная последовательность $(t, x_1(t), x_2(t), \ldots, x_5(t))$, где $x_i(t)$ – показания датчика активной i-й точки в момент времени t. Выходная последовательность – $(y_1(t), y_2(t), y_3(t))$, где $y_j(t)$ – прогнозируемые показания датчика неактивной j-й точки.

Предварительно данные масштабировались с помощью алгоритма min-max. Масштабированные значения для временного ряда $\{x_t\}$ имеют следующий вид:

$$\frac{x_t - \min(x)}{\max(x) - \min(x)}.$$

Этот алгоритм применяется для случаев, когда распределение значений входных данных не является нормальным (в случае наших данных распределение далеко от нормального (см. диагональные элементы рис. 1)) и в данных нет выбросов.

Рисунок 8 демонстрирует окончательную архитектуру нейросети, которая содержит 9 скрытых слоев. В табл. 2 приведены значения коэффициента детерминации R^2, полученные по тестовой и обучающей выборке для нейросетевых моделей с различным числом скрытых слоев.

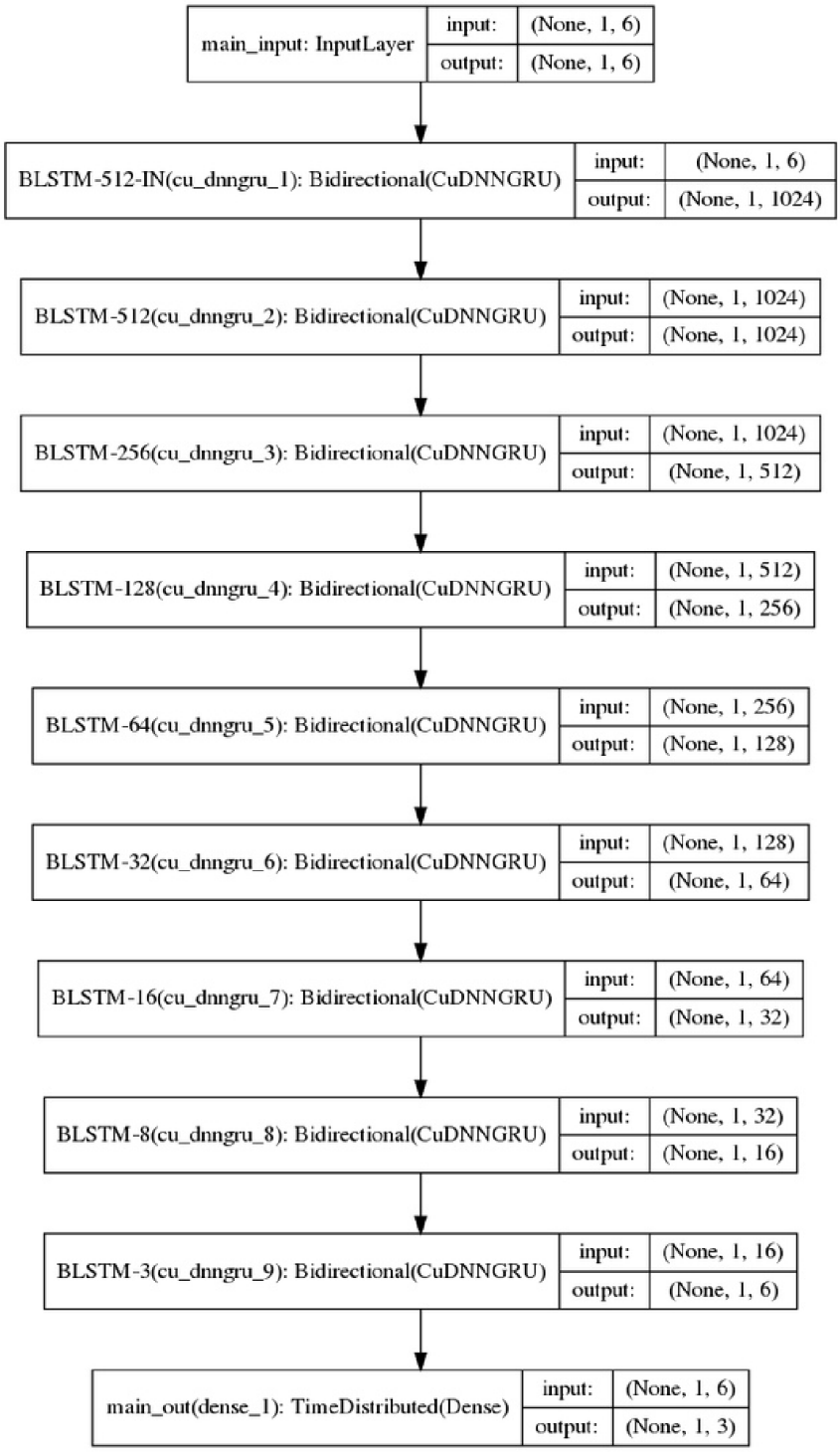

Рисунок 8 – Архитектура BLSTM-сети, используемая для прогноза показаний датчиков в неактивных точках

Таблица 2 – Оценка эффективности нейросетевых моделей

выборка	4 скрытых слоя	9 скрытых слоев
обучающая	0.976436197757721	0.9995164275169373
тестовая	0.9750288128852844	0.9972876906394958

Как видно из таблицы, увеличение слоев повышает эффективность модели.

Значения предсказанных (pred) и реальных (test) нормированных значений y_1-y_3 в зависимости от нормированной переменной *time* приведены на рис. 9 – 11.

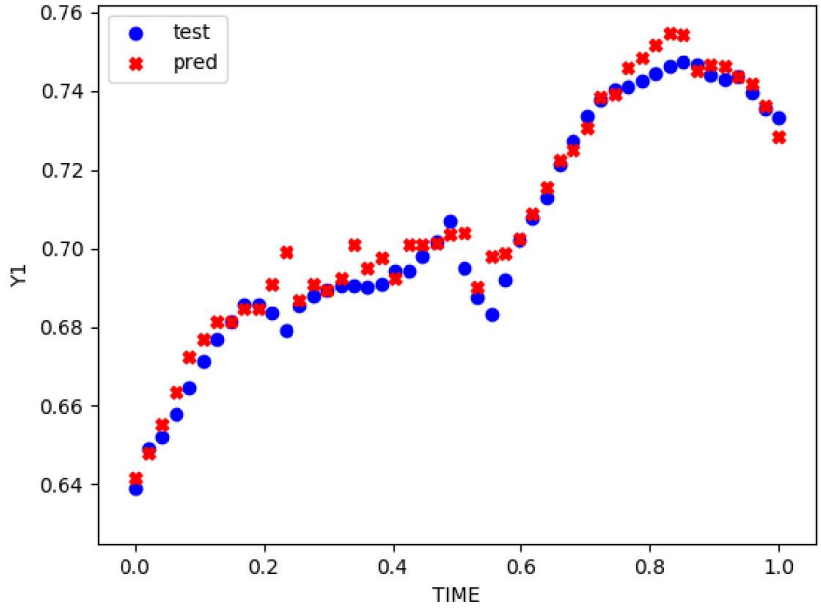

Рисунок 9 – Нормированные значения из тестовой выборки за сутки и значения, полученные нейросетевой моделью, для переменной y_1

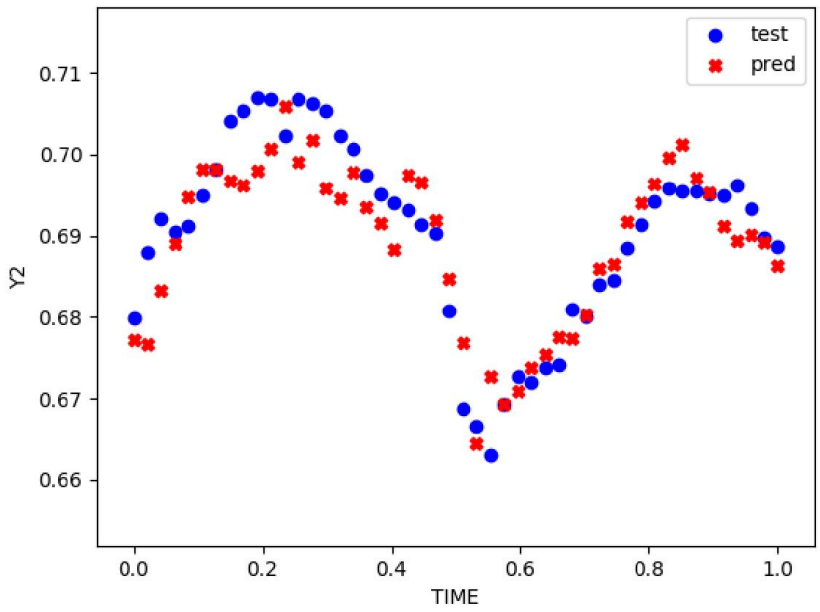

Рисунок 10 – Нормированные значения из тестовой выборки за сутки и значения, полученные нейросетевой моделью, для переменной y_2

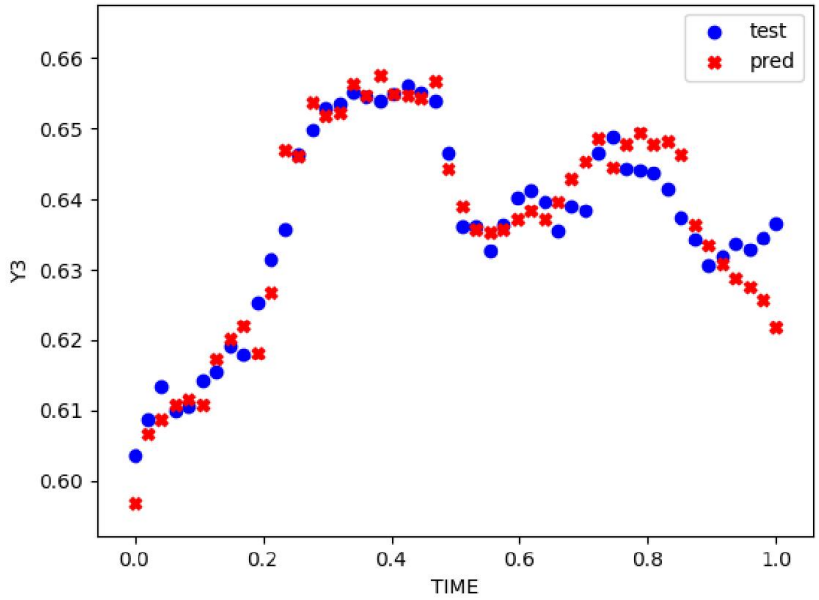

Рисунок 11 – Нормированные значения из тестовой выборки за сутки и значения, полученные нейросетевой моделью, для переменной y_3.

Сравнивая рис. 4 и рис. 11, можно сделать вывод о большей точности нейросетевой модели и о верном выборе типа нейросети – BLSTM, способной, в отличие от деревьев решений, учитывать для предсказания текущих значений временных последовательностей их контекст и слева, и справа.

Общее количество параметров для сети, архитектура которой приведена на рис. 7, равно 8 949 423, время обучения 5 – 10 минут 1000 эпох. Зависимость значения коэффициента детерминации от количества эпох обучения показана на рис. 12.

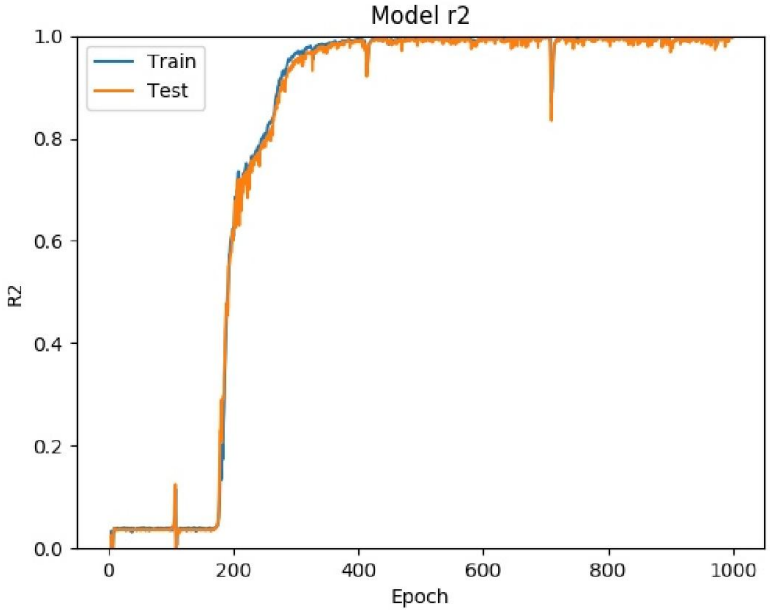

Рисунок 12 – График зависимости R^2 от количества эпох обучения BLSTM-сети

На рис. 13, 14 показана зависимость значений различных функций потерь от количества эпох обучения, полученных на тестовой выборке. Обозначения на рисунках: y' – предсказанные значения выхода сети, y – значения тестовой выборки, n – объем тестовой выборки.

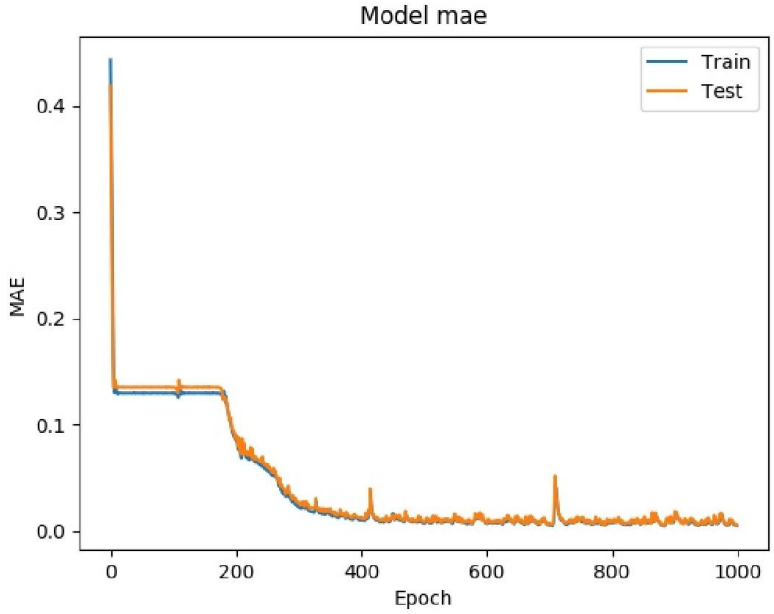

Рисунок 13 – График зависимости функции потерь MAE $L(y',y)=\frac{1}{n}\sum_i|y'_i-y_i|$ от количества эпох обучения BLSTM-сети

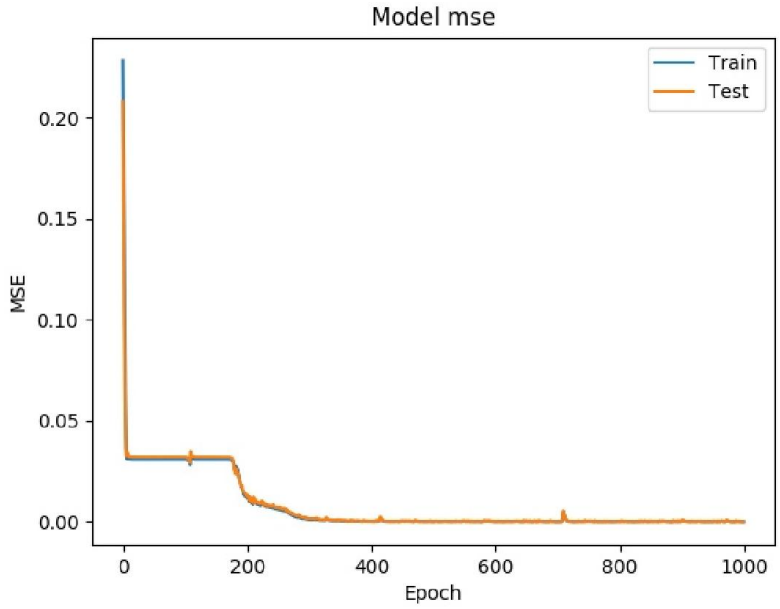

Рисунок 14 – График зависимости функции потерь MSE $L(y',y)=\frac{1}{n}\sum_i(y'_i-y_i)^2$ от количества эпох обучения BLSTM-сети

Было проведено исследование важности предикторов с помощью нейросети по следующему алгоритму:

1. Обучение нейросети.
2. Тестирование нейросети на тестовой выборке.
3. Вычисление среднего значения для каждой входной переменной.
4. Поочередное значение целевой функции при фиксировании каждого входа тестовой выборки.
5. Анализ значений целевой функции.

В качестве целевой функции использовалась функция потерь MSE. Чем больше значение целевой функции, тем важнее предиктор. Результаты исследования сведены в табл. 3, где указаны значения MSE при фиксированном входе.

Таблица 3 – Значения функции потерь при фиксированных входах нейросети

time	x_1	x_2	x_3	x_4	x_5
2.2289617	1.7885027	2.1235294	1.9561572	1.9372189	0.097451

Анализируя полученные результаты, можно сделать вывод, что время (time) для BLSTM-сети наиболее важный признак.

Выводы

С целью построения предсказательной модели значений давления воды в контрольных точках водопроводной сети проведен разведочный анализ данных для выявления выбросов и пропущенных значений, а также наличия связи между переменными. При этом была выявлена мультиколлинеарность переменных.

Построены предсказательные модели для показаний датчиков давления воды в неактивных точках водораспределительной сети на основе ансамблей классификаторов и нейросетевого подхода.

Произведена оценка значимости предикторов для каждой модели. Для модели на основе ансамблей классификаторов использовали алгоритм Random Forest, для нейросети – анализ значений целевой функции. Высокая корреляция между показаниями датчиков в неактивных точках отразилась на оценке важности предикторов, полученных для моделей бустинга.

Проведен сравнительный анализ качества построенных моделей с использованием коэффициента детерминации. При этом метод градиентного бустинга и BLSTM-сети дают ошибку, близкую к нулю (менее 0,01).

Заключение

Таким образом, разработанные модели позволяют прогнозировать процессы, происходящие в водопроводных сетях. При этом нейросетевые модели показывают более лучший результат, но являются более ресурсоемкими.

Список литературы

1. Джон Тьюки. Анализ результатов наблюдений. Разведочный анализ [Текст] / Джон Тьюки. – М. : Мир, 1981. – 696 с.
2. Zhi-Hua Zhou. Ensemble Methods: Foundations and Algorithms [Текст] / Zhi-Hua Zhou. – New York : Chapman & Hall, 2012. – 222 p.
3. Введение в статистическое обучение с примерами на языке R [Текст] / Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. ; пер. С. Э. Мастицкого. – М. : ДМК Пресс, 2016. – 450 с.
4. Пикалёв Я. С. Глубинное обучение в задаче автоматического распознавания речи. [Текст] / Я. С. Пикалёв // Интеллектуальные технологии и проблемы математического моделирования: материалы Всерос. науч. конф. (Дивноморское, 24 – 26 сентября 2018 г.) / под общ. ред. Б. В. Соболя; Донской гос. техн. ун-т. – Ростов-на-Дону: ДГТУ, 2018. – с. 16-17.
5. Николенко С. Глубокое обучение [Текст] / С. Николенко, А. Кадурин, Е. Архангельская – СПб. : Питер, 2018. – 480 с.

References

1. John Tukey. *Analiz rezul'tatov nablyudeniy. Razvedochnyy analiz* [Analysis of the results of observations. Exploration analysis], M ., World, 1981, 696 p.
2. Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms.* New York, Chapman & Hall, 2012, 222 p.
3. James G., Whitton D., Hasti T., Tibshirani R. *Vedeniye v statisticheskoye obucheniye s primerami na yazyke R* [Introduction to statistical learning with examples in R language], Per. S.E. Mastitsky,. M ., DMK Press, 2016, 450 p.

4. Pikalyov Y.S. Glubinnoye obucheniye v zadache avtomaticheskogo raspoznavaniya rechi [Deep learning in the task of automatic speech recognition]. *Intellektual'nyye tekhnologii i problemy matematicheskogo modelirovaniya: materialy Vseros. nauch. konf.* [Intellectual technologies and problems of mathematical modeling: materials of All-Russia. scientific conf.] (Divnomorskoe, September 24 - 26, 2018) / under total. ed. B.V. Sable; Don State tech. un-t - Rostov-on-Don: DGTU, 2018. p. 16-17.
5. Nikolenko S., Kadurin A., Arkhangelskaya E. *Glubokoye obucheniye* [Deep Learning], SPb., Peter, 2018, 480 p.

RESUME

S. V. Bespalova, S. M. Romanchuk, T. V. Yermolenko, V. I. Bondarenko

Construction of predictive models of the parameters of water pressure in water distribution networks by means of machine training methods

The support`s problems of effective and definitive working conditions of objects of water distribution network, providing an opportunity to electricity and water saving and also the problems of identification of inappropriate conditions and emergencies are the most important for water supply organizations. This article examines the effectiveness of autoregressive methods for forecasting the behavior of the water supply network using online data of Computer-Assisted Acquisition System.

The source data are represented by time series of water pressure values at control points of the water supply network. An exploratory analysis has been carried out, predictive models have been constructed for the registration of water pressure sensors in inert points of the water distribution network based on the ridged and lasso regression models, and the regression on the main components.

The statistical significance of the three models was evaluated and the efficiency of the models constructed was compared with the prediction.

The conclusion: The developed models allow to predict the processes occurring in water supply networks. It is indicated that for two variables the optimal models for the forecast were chosen. While for the third variable, none of the models allows them to be used for prediction. As applicable the use of non-linear regression models is assumed.

РЕЗЮМЕ

С. В. Беспалова, С. М. Романчук, Т. В. Ермоленко, В. И. Бондаренко

Построение предсказательных моделей параметров давления воды в водораспределительных сетях с помощью методов машинного обучения

Задачи поддержания эффективных и оптимальных режимов работы объектов водораспределительной сети, позволяющих обеспечить экономию электроэнергии и воды, а также задачи распознавания внештатных режимов и аварийных ситуаций являются наиболее важными для предприятий водоснабжения. В этой статье исследуется эффективность методов машинного обучения для прогнозирования поведения водопроводной сети с использованием онлайн-данных автоматизированной системы сбора и передачи информации.

Исходные данные представлены временными рядами значений давления воды в контрольных точках водопроводной сети. Проведен разведочный анализ, построены предсказательные модели для показаний датчиков давления воды в неактивных точках водораспределительной сети на основе ансамблей классификаторов и нейросетевого подхода. Произведена оценка значимости предикторов для каждой модели с использованием алгоритма Random Forest и анализа значений целевой функции.

Проведен сравнительный анализ качества построенных моделей с использованием коэффициента детерминации.

Разработанные модели позволяют прогнозировать процессы, происходящие в водопроводных сетях. При этом нейросетевые модели показывают более лучший результат, но являются более ресурсоемкими.

Статья поступила в редакцию 25.02.2019.