УДК 004.048
DOI: 10.55648/1998-6920-2022-16-4-69-79

К вопросу о подготовке данных при разработке модели нейронной сети

С. Б. Жанаева

Одной из особенностей работы операторов мобильной сети передачи данных является необходимость в постоянном мониторинге и обслуживании оборудования и каналов связи. Происходящие сбои в работе оборудования увеличивают стоимость эксплуатации и уменьшают лояльность клиентов. Возможность заблаговременного предсказания сбоев в работе сети послужило бы отличным решением для мобильных операторов. В текущей статье рассматривается вопрос предварительной подготовки собранных данных о работе мобильной сети оператора связи 4G+ для дальнейшего использования при разработке нейронной модели для предсказания сбоев. Приведены результаты проведенного анализа собранных данных, показаны характеристики, состав и структура данных, особенности которых в дальнейшем могут повлиять на обучение нейронной модели.

Ключевые слова: подготовка данных для нейронной модели, статистический анализ данных, мобильные сети, сбои в работе оборудования.

1. Введение

Подготовка данных для обучения модели нейронной сети является важным этапом разработки и создания модели. От качества подготовленных данных зачастую зависит производительность, точность и эффективность разработанной модели нейронной сети [1]. Для подготовки больших объемов данных как нельзя лучше подходят различные статистические методы, они важны при подготовке как обучающих, так и тестовых наборов данных, например, сюда входят следующие техники [2]:

- обнаружение выбросов;
- обработка пропущенного значения;
- выборка данных;
- масштабирование, нормализация данных;
- переменное кодирование;
- проверка корреляции переменных.

Предварительное понимание состава и распределения данных, анализ с помощью описательной статистики и визуализации необходимы для того, чтобы в дальнейшем определить наиболее подходящие методы и алгоритмы, которые следует выбрать при работе с данными [3].

В текущей статье описывается подготовка данных для создания модели нейронной сети, прогнозирующей возникновение сбоев в работе оператора мобильных данных. Сбор данных выполнялся в течение десяти месяцев в телекоммуникационной сети мобильного оператора 4G+. Таким образом, были собраны два типа данных: о состоянии базовых станций и об аварийных ситуациях на станциях.

При подготовке собранных данных было использовано несколько видов статистического анализа, и в результате были выявлены некоторые интересные ключевые моменты и критические проблемы.

2. Обзор литературы

При обучении нейронной сети подготовка данных играет большую, порой решающую, роль. От качества выполненной подготовки данных зависит точность и полнота предсказаний нейронной модели. Поэтому вопросу подготовки данных исследователи уделяют повышенное внимание.

Например, B. Schmidt и L. Wang [4] полагают, что одна из основных проблем при подготовке данных для прогнозирования сбоев заключается в том, что анализируются данные только из одного типа источников. Поэтому они использовали данные, полученные из различных функциональных департаментов компании, таких как служба эксплуатации, планирование, служба контроля качества и других. Исследователи в [5] также для получения наиболее полного объема информации о происходящих сбоях в работе оборудования выполняли сбор, подготовку и анализ данных, получаемых из различных источников. Собранные данные авторы [4–8] анализировали с помощью методов статистического анализа, проверяли состав и структуру данных.

Большинство исследователей [4–7] считают, что извлечение признаков — важный этап предварительной обработки, на котором необработанные данные, собранные из различных сигнальных каналов, преобразуются в набор статистических признаков в формате, поддерживаемом алгоритмами машинного обучения. Так, разработчики нейронной модели [4–7] определили статистические характеристики, которые затем передали в качестве входных данных для алгоритма машинного обучения. Набор статистических характеристик, извлеченных из сигнальных каналов, включал максимум, медиану, среднее значение и стандартное отклонение.

Исследователи [7, 8] выполнили подготовку материалов в несколько шагов: разделили и использовали базу данных (БД) об авариях в качестве маркированных данных; нормализовали и масштабировали данные, приведя их к единой размерности; разделили БД на обучающую, проверочную и тестовую. Во время предварительной обработки информации извлекли значения признаков корреляции временных рядов и в дальнейшем использовали их как входную информацию для глубоких нейронных сетей.

3. Подготовка данных о функциональном состоянии базовых станций

Во время сбора данных при наблюдении и анализе работы базовых станций (БС) было замечено, что только на 30–35 % станций были зафиксированы аварийные сообщения. Большинство БС во время наблюдения работали без нареканий, чего и следовало ожидать. Таким образом, база данных, содержащая информацию о функциональном состоянии оборудования, содержит информацию как о станциях, на которых не возникало аварий, так и информацию по станциям с авариями.

Для обучения нейронной модели прогнозированию аварий и причин аварий понадобятся данные со станций, на которых были зафиксированы аварии и на которых аварии не появлялись. Поэтому была выполнена выборка данных и сформированы две отдельные базы данных о функциональном состоянии базовых станций с авариями и без них.

В качестве одного из этапов подготовки данных была выполнена проверка собранных данных. По результатам проверки выяснилось, что некоторые строки (около 1.7 %) содержали пустые значения, кроме того, в некоторых ячейках (около 3.8 %) стояли не числовые значения,

а буквенные. Буквенные значения содержали аббревиатуру INF (Information Not Found), что означает потерю и пропуск собранных данных.

На следующем этапе была выполнена очистка данных: строки с пропущенными значениями были удалены; поля, заполненные буквенными значениями, были заменены на среднее арифметическое от соседних значений в столбце.

В каждой из собранных баз данных о функциональном состоянии базовых станций Average KPI и Channel KPI содержится информация о 2700 базовых станциях по 42 показателям. С помощью библиотек Python Numpy и Pandas [9] были определены количество собранных данных, размерность. После очистки баз данных и удаления отсутствующих значений базы данных приобрели следующую размерность:

– база данных Average KPI содержит 878793 строки записей по 42 показателям, из них 347556 строк содержат информацию по базовым станциям с авариями и 531237 строк содержат информацию по базовым станциям без аварий;
– база данных Channel KPI содержит 878629 строк записей по 42 показателям, из них 347556 строк содержат информацию по базовым станциям с авариями и 531073 строки содержат информацию по базовым станциям без аварий.

После выборки, проверки и очистки данных были заданы переменные: столбцам с 42 показателями были последовательно присвоены имена переменных от D1 до D42.

Для лучшего понимания состава и структуры собранных данных использовались методы описательной статистики, эти методы могут дать хорошее представление о свойствах каждого атрибута данных [2, 10, 11]. Так, с помощью функции description() библиотеки Pandas [9] определили 8 основных статистических свойств каждого атрибута: количество, математическое ожидание, среднеквадратичное отклонение, минимальное значение, 25-, 50-, 75-процентные значения, максимальное значение. Ниже в табл. 1 показаны статистические характеристики атрибутов собранных данных о состоянии БС по выборочным показателям D4–D9.

Таблица 1. Статистические характеристики атрибутов D4–D9

Параметр	D4	D5	D6	D7	D8	D9
count	878593	878217	877929	878684	878744	878649
mean	0.972533	0.873245	0.928842	0.942802	0.03580	0.02416
std	0.166	0.01347	0.0107	0.0108	0.01824	0.01559
min	0.0000	0.0000	0.25103	0.00043	0.0049	0.0049
25 %	1.0000	0.0000	0.92847	0.94143	0.0154	0.0081
50 %	1.0000	0.47453	0.93131	0.94346	0.0194	0.0186
75 %	1.0000	0.73461	0.93245	0.94542	0.0227	0.0592
max	1.0000	1.0000	0.99753	0.94551	0.09451	0.0956

Проверка и анализ базы данных о состоянии базовых станций показали, что некоторые атрибуты этой базы данных обладают большой разрежённостью, содержат большое количество нулей в значениях. Например, атрибут D9 (DROP_RATE_LTE), который отображает процент соединений LTE, завершившихся обрывом, содержит около 94 % нулевых значений. Следовательно, во время наблюдения на 94 % базовых станций обрывы соединений LTE не были зафиксированы. Как известно, большая разрежённость базы данных привносит проблемы при разработке нейронной модели, так как увеличивает требуемое время для обучения модели.

При обучении нейронной сети одним из приемов, которые ускоряют обучение модели, является нормализация входных данных [1, 11]. Нормализация данных позволяет масштабировать числовые значения в указанном диапазоне. Нормализация данных была выполнена методом масштабирования данных на основе математического ожидания и среднеквадратичного отклонения: деление разницы между данными и математическим ожиданием на величину среднеквадратичного отклонения [1, 11, 12] по следующим формулам:

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i,$$

$$\sigma^2 = \mu\left[x - \mu(x)\right]^2, \tag{1}$$

$$x_{in} = \frac{x_i - \mu(x_i)}{\sigma},$$

где μ – математическое ожидание, σ – среднеквадратичное отклонение, x_i – i-ое значение показателя x, x_{in} – новое нормализованное значение показателя x. Нормализация была выполнена для всех показателей по столбцам, кроме, столбцов времени, даты и уникального идентификатора соты. Всего нормализация данных была выполнена по 39 показателям.

4. Подготовка данных об авариях на базовых станциях

С помощью библиотек Python Numpy и Pandas [9] были определены количество собранных данных по авариям и размерность. Так, база данных по аварийным ситуациям содержит 3847653 строки записей и 5 столбцов. Каждая строка отображает сообщение об аварийном состоянии оборудования мобильной сети. Столбцы содержат следующие данные: уровень критичности аварии, наименование и ID базовой станции, время фиксирования аварии, предположительная причина аварии, наименование трансмиссии. При подготовке базы данных по авариям часть строк записей были удалены, ниже перечислены причины.

1. Удаление дублирующих сообщений. При мониторинге работы базовых станций и сборе данных было отмечено, что в зависимости от типа аварии на станции сообщения об одной и той же проблеме могут появляться в системе мониторинга несколько раз. Дублирующие сообщения могут отображаться в системе мониторинга с разницей в несколько секунд или минут, создавая избыточность в базе данных. При выполнении проверки собранных данных подобные дублирующие сообщения были удалены.

2. Удаление пропущенных значений. При выполнении проверки собранных данных оказалось, что часть полей (около 1.2 %) содержали пустые значения, некоторые поля (ок. 2.6 %) были заполнены буквенной аббревиатурой INF (Information Not Found), т.е. часть данных была потеряна. Данные были очищены от пропущенных значений следующим образом: так как пустых и буквенных значений было немного, а общего количества данных по авариям было достаточно, то строки с пустыми полями были удалены.

3. Удаление форс-мажорных аварийных ситуаций. При создании модели нейронной сети, прогнозирующей возникновение сбоев и аварийных ситуаций в работе оборудования, необходимо учесть тот факт, что, несмотря на все вычислительные способности разрабатываемой модели, некоторые форс-мажорные аварийные ситуации предсказать невозможно, так как причины их возникновения не зависят от состояния оборудования или показателей характеристик радиоканалов. Например, к таким аварийным ситуациям относятся сбои с внешней причиной возникновения, такие как отключение электропитания на подстанции или срабатывание датчика открытия дверей при посещении места установки базовой станции персоналом. Чтобы не обучить разрабатываемую модель выводу ложных случайных зависимостей, была проведена выборка данных, и данные об авариях отключения питания, открывания дверей, срабатывания датчиков движения были удалены.

После очистки баз данных, выполнения выборки, удаления отсутствующих полей, избыточности дублирующих сообщений база данных по авариям приобрела следующую размерность: 347556 строк записей и 5 столбцов.

Показатели базы данных по авариям носят категориальный характер, например, это уровни критичности, наименования аварий и трансмиссия.

Для дальнейшей работы с базой данных необходимо было выполнить кодирование переменных, поэтому показателям в заголовках столбцов в базе данных по авариям были присвоены переменные от BDA1 до BDA5 по порядку следования. В табл. 2 показана выборка из базы данных по авариям с присвоенными переменными.

Таблица 2. Выборка из базы данных по аварийным ситуациям в сети

Уровень	Наименование	Время	Предполагаемая причина	Трансмиссия
BDA1	BDA2	BDA3	BDA4	BDA5
Minor	37323_AKT-Beyneu	2021-02-15 11:32:28	Abis control link broken	Aktau-BSC-2
Minor	51247_KZL-Merke	2021-02-15 11:32:28	Abis control link broken	Kzl-BSC-1
Major	71422_ALM-Issyk	2021-02-15 11:32:28	Rectifier failure	Alm-BSC-2
Critical	37375_AKT-Karasu	2021-02-15 11:32:29	POWER OFF	Aktau-BSC-2
Minor	71458_ALM-Issyk	2021-02-15 11:32:29	Abis control link broken	Alm-BSC-1

Атрибутам наименований аварий (Abis control link broken, Rectifier failure, SCTP association is broken и т.д.) показателя BDA4 были присвоены переменные от A1 до A12 в зависимости от типа аварийного сообщения. Атрибутам уровней критичности присвоены переменные от AL1 до AL4 (Critical, Major, Minor, Warning) соответственно. Атрибуты показателей времени, даты аварии и уникального идентификатора соты не изменялись.

Так как показатели данных по авариям категориальные, то методы описательной статистики, рассчитывающие математическое ожидание, среднеквадратичное отклонение и т.д., предназначенные для числовых значений, в данном случае не применимы. Поэтому для анализа данных воспользовались методами распределительной статистики с целью проверки, каких значений больше, как они распределены по общему объему данных [2, 12]. Для того чтобы проанализировать распределение атрибутов классов, воспользовались инструментами библиотеки Pandas. Некоторые результаты анализа данных приведены в табл. 3.

Таблица 3. Количество сообщений по уровням аварийных ситуаций

№ п.п.	Заданные переменные	Уровень критичности	Количество аварийных сообщений
1	AL4	Warning	97868
2	AL3	Minor	99866
3	AL2	Major	86889
4	AL1	Critical	62933

При решении категориальных задач нужно представлять, насколько сбалансированы значения классов. Если значения классов довольно сильно несбалансированы (т.е. присутствует намного больше наблюдений для одного класса, чем для другого), то это приведет к дополнительным сложностям при обучении модели и может потребоваться специальная обработка на этапе подготовки данных [2, 12]. Так как в текущем исследовании именно предполагаемый тип аварийной ситуации планируется использовать в качестве ответа, выдаваемого нейронной моделью, то важно было определить, как количественно распределены сообщения о

различных типах аварий на станциях. Это было выполнено с помощью инструментов библиотеки Pandas. Ниже в табл. 4 приведены результаты.

Таблица 4. Количество аварийных сообщений по видам аварии

№ п.п	Вид аварийного сообщения	Количество сообщений
1	A1	24 329
2	A2	17 378
3	A3	35 104
4	A4	28 085
5	A5	41 707
6	A6	34 408
7	A7	10 427
8	A8	24 610
9	A9	17 659
10	A10	52 133
11	A11	27 804
12	A12	33 913

Как видно из результатов анализа данных, наименьшее количество сообщений приходится на аварийные ситуации вида A7. Количественное распределение сообщений по остальным видам аварийных ситуаций A1–A6, A8–A12 относительно равномерно. Следовательно, необходимо определить, достаточно ли будет собранных данных – 10427 записей – для прогнозирования возникновения аварийной ситуации вида A7. Сделать это необходимо эмпирическим путем, проверив на тестовых данных, с какой точностью нейронная модель сможет спрогнозировать аварии вида A7.

5. Подготовка данных по первопричинам аварий

Несмотря на то, что система мониторинга работы базовых станций при обнаружении сбоев в работе оборудования выдает сигнал аварии и причину возникновения аварийной ситуации, в ряде случаев бывает так, что выданная системой мониторинга причина не в полной мере соответствует действительности. Выехавшая на место инженерная бригада обнаруживает другую причину сбоев или совокупность причин. При наблюдении за системой мониторинга и сбоями в сети было замечено, что около 7.4 % причин аварий, выданных системой мониторинга, отличается от выявленных и подтвержденных причин, определенных на месте установки оборудования.

Поэтому для повышения точности прогнозирования было принято решение учитывать скорректированные инженерами первопричины аварийных ситуаций. Скорректированные данные по первопричинам аварий сохранялись в *xls*-файле, затем при загрузке в Python они были преобразованы в формат *csv*-файла. С помощью библиотек Python Numpy и Pandas были определены количество собранных данных и размерность. База данных по первопричинам аварий имела 68596 строк и 4 столбца записей.

Как видно из результатов вычислений, информации по первопричинам аварий, которую удалось собрать вручную, оказалось намного меньше количества данных, собранных с помощью автоматической системы мониторинга. Но, несмотря на малый объем, собранные вручную данные репрезентативны. Информация в этой базе данных была получена путем визита на место установки оборудования и проверена инженерной бригадой.

С целью дальнейшего обучения нейронной модели база данных по первопричинам аварий была объединена с базой данных по авариям, зафиксированным с помощью системы

мониторинга. При объединении баз данных причины аварий, соответствующие по времени, месту и дате, были заменены первопричинами аварий, выявленных инженерами.

6. Корреляция между показателями

Перед началом обучения нейронной сети были проверены и определены показатели, наиболее коррелирующие с возникновением сбоев на станциях. В качестве предварительной оценки корреляции изменений 39 показателей при возникновении или отсутствии аварий было выполнено сравнение статистических характеристик показателей в случае отсутствия аварийных ситуаций и в случае наличия сбоев на станции. Для этого воспользовались базами данных о состоянии базовых станций, описанными в разделе 3, разделенными на базы данных без аварий и с авариями.

Были вычислены математические ожидания и стандартные отклонения [1, 2] по 39 показателям состояния базовых станций для аварийных ситуаций и для случаев, когда аварии на станциях не были зафиксированы. Вычисленные значения были нормализованы. Показатели были разбиты на группы по 10 показателей, и проверка выполнялась поочередно. На рис. 1 показаны значения нормализованного математического ожидания показателей состояния базовых станций для выборочных атрибутов D4–D9, вычисленных для трех различных случаев.

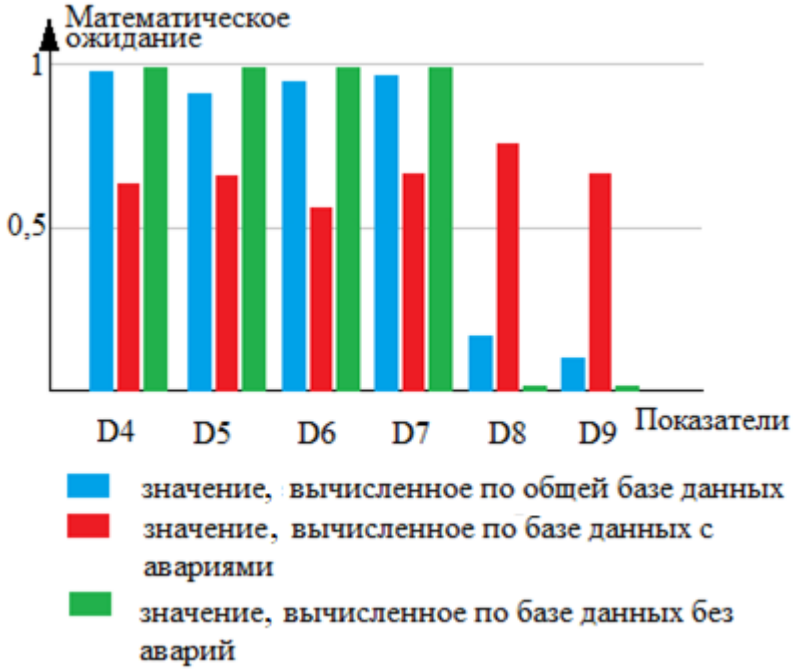

Рис. 1. Значения математического ожидания показателей D4–D9.

Как видно из рис. 1, значения атрибутов D4–D9 коррелируют с появлением аварийных ситуаций, причем они имеют как положительную, так и отрицательную корреляцию.

Для количественной оценки взаимосвязи между различными атрибутами и сообщениями об аварийных ситуациях воспользовались вычислением коэффициента корреляции Пирсона с помощью функции pearsonr() библиотеки NumPy [9, 12]. В табл. 5 приведены результаты вычисления коэффициента корреляции Пирсона для 39 атрибутов D4–D42. Атрибуты D1, D2, D3 содержат значения даты, времени возникновения аварии и уникальный идентификатор соты. По атрибутам D1–D3 вычисление коэффициента корреляции Пирсона не проводилось.

Таблица 5. Коэффициент корреляции Пирсона
между значениями атрибутов базы данных и сообщениями об авариях

№ п.п.	Атрибуты	Значения коэффициента корреляции Пирсона
1	D4	- 0.58
2	D5	- 0.62
3	D6	- 0.69
4	D7	- 0.73
5	D8	0.71
6	D9	0.88
7	D10	0.23
8	D11	0.12
9	D12	0.34
10	D13	0.28
11	D14	- 0.63
12	D15	- 0.77
13	D16	- 0.81
14	D17	- 0.76
15	D18	0.71
16	D19	0.86
17	D20	0.74
18	D21	0.82
19	D22	-0.37
20	D23	-0.32
21	D24	-0.28
22	D25	-0.23
23	D26	-0.27
24	D27	-0.25
25	D28	0.87
26	D29	0.84
27	D30	- 0.75
28	D31	- 0.82
29	D32	- 0.86
30	D33	- 0.83
31	D34	0.08
32	D35	0.21
33	D36	0.23
34	D37	0.15
35	D38	0.17
36	D39	- 0.86
37	D40	- 0.87
38	D41	- 0.73
39	D42	- 0.77

По значениям коэффициентов корреляции Пирсона из табл. 5 видно, что часть показателей состояния базовых станций показывают высокую степень изменчивости в случаях аварий на станциях. Например, атрибуты показателей D8, D9, D19, D20, D21 имеют высокую степень положительной корреляции с возникновением аварий. Показатели D7, D15, D16, D17, D30, D31, D32, D33, напротив, имеют отрицательную корреляцию.

Положительная корреляция с аварийными ситуациями наблюдалась у показателей функционального состояния БС, отображающих наличие каких-либо проблем при выполнении голосового соединения или передачи данных. Например, при увеличении количества аварийных ситуаций на станциях, увеличивалось и количество блокированных соединений и соединений, завершившихся обрывом. Соответственно, отрицательная корреляция наблюдалась у показателей, характеризующих успешно установленные и завершенные голосовые соединения, пакетные передачи, объем переданного трафика и т.д.

7. Заключение

Качественно выполненная работа по подготовке данных является залогом высокой точности прогнозирования разрабатываемой нейронной модели. Поэтому на данном этапе исследовательской работы по прогнозированию сбоев в сети мобильного оператора были выполнены анализ и подготовка данных, которые включали в себя проверку и выборку данных, удаление и замену пропущенных значений, нахождение статистических характеристик, определение взаимозависимости атрибутов данных, нахождение коэффициентов корреляции. Проведенный анализ и подготовка собранных данных о функциональном состоянии и аварийных ситуациях на базовых станциях помогли определить основные характеристики данных, их состав, структуру, выявили закономерности в поведении данных, кроме того, были обнаружены некоторые проблемные моменты. Результаты проведенного анализа данных могут послужить хорошей отправной точкой в дальнейшем при обучении нейронной сети.

Литература

1. *Brink H., Richards J. W., Fetherolf M.* Real world Machine Learning. US: Manning Publications Co., 2017. 266 p.
2. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer Series in Statistics / 2nd Edition. NY: Springer, 2017. 764 p.
3. *Грас Д.* Data science. Наука о данных с нуля: пер. с англ. СПб: БХВ-Петербург, 2017. 336 с.
4. *Schmidt B., Wang L.* Cloud-enhanced predictive maintenance // Int J Adv Manuf Technol. 2018. № 99. P. 5–13. DOI 10.1007/s00170-016-8983-8.
5. *Hu C., Youn B. D., Kim T.* Semi-supervised learning with co-training for data-driven prognostics // Proc. IEEE Int. Conf. on Prognostics and Health Management: Enhancing Safety, Efficiency, Availability, and Effectiveness of Systems Through PHM Technology and Application. Denver, CO, USA. 18–21 June, 2012. P. 1297–1306. DOI: 10.1109/ICPHM.2012.6299526.
6. *Wu D., Jennings C., Terpenny J., Kumara S.* Cloud-Based Machine Learning for Predictive Analytics: Tool Wear Prediction in Milling // Proc. IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December, 2016. P. 2062–2069.
7. *Yin-Hsin Liu, Yao-Chung Tu, Chang-Yu Hsu, Hsin-Chieh Chao*. Predicting malfunction of mobile network base station using machine learning approach // Proc. 20th Asia-Pacific Network Operations and Management Symposium (APNOMS), Matsue, Japan, 18–20 September, 2019. P. 1–4. DOI: 10.23919/APNOMS.2019.8892894.
8. *Xu M., Baraldi P., Al-Dahidi S., Zio E*. Fault prognostics by an ensemble of Echo State Networks in presence of event based measurements // Engineering Applications of Artificial Intelligence. V. 87. 2020. P. e103346. DOI: 10.1016/j.engappai.2019.103346.
9. *McKinney W.* Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter. Boston: O'Reilly Media, 2022. 579 p.
10. *Hoaglin D. C., Mosteller F., Tukey J. W.* Understanding Robust and Exploratory Data Analysis. NY: Wiley-Interscience, 2000. 447 p.

11. *Флах П.* Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А. А. Слинкина. М.: ДМК Пресс, 2015. 400 с.
12. *Шолле Ф.* Глубокое обучение на Python. Серия «Библиотека программиста». СПб.: Питер, 2018. 400 с.

Статья поступила в редакцию 15.06.2022;
переработанный вариант – 01.10.2022.

Жанаева Сауле Бактыкереевна
аспирант специальности 09.06.01, кафедра прикладной математики и кибернетики СибГУТИ (630102, Новосибирск, ул. Кирова, 86), e-mail: szhanayeva@gmail.com.

Data preparation for a neural network model

Saule B. Zhanayeva
Postgraduate student, Siberian State University of Telecommunications and Information Science (SibSUTIS, Novosibirsk, Russia), szhanayeva@gmail.com.

One of the features of the mobile data network operators is the need for continuous monitoring and maintenance of equipment and communication channels. The equipment failures that sometimes occur increase the cost of operation and reduce customer loyalty. The ability to predict network malfunctions in advance would be a great solution for mobile operators. The paper discusses the issue of preliminary data preparation of 4G+ mobile network for further use in the development of a neural network model for predicting malfunctions. The results of the analysis of the collected data are presented, the characteristics, composition and data structure that may affect the training of the neural network model later are shown.

Keywords: data preparation for neural network, statistical data analytics, mobile networks, equipment malfunctions.

References

1. Brink H., Richards J. W., Fetherolf M. *Real world Machine Learning*. US, Manning Publications Co., 2017. 266 p.
2. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer Series in Statistics*. 2nd Edition. NY, Springer, 2017. 764 p.
3. Grus J. Data science. *Nauka o dannyh s nulya* [Fata Science from Scratch]. Transl. from Eng. SPb, BHV-Peterburg, 2017. 336 p.
4. Schmidt B., Wang L. Cloud-enhanced predictive maintenance. *Int J Adv Manuf Technol*., 2018, no 99, pp. 5–13. DOI 10.1007/s00170-016-8983-8
5. Hu, C., Youn, B. D., Kim, T. Semi-supervised learning with co-training for data-driven prognostics. Proceedings of *IEEE Int. Conf. on Prognostics and Health Management: Enhancing Safety, Efficiency, Availability, and Effectiveness of Systems Through PHM Technology and Application.* Denver, CO, USA. 18–21 June, 2012, pp. 1297-1306. DOI: 10.1109/ICPHM.2012.6299526
6. Wu D., Jennings C., Terpenny J., Kumara S. Cloud-Based Machine Learning for Predictive Analytics: Tool Wear Prediction in Milling. *IEEE International Conference on Big Data (Big Data)*. Washington, DC, USA, 5-8 December, 2016, pp. 2062-2069.
7. Yin-Hsin Liu, Yao-Chung Tu, Chang-Yu Hsu, Hsin-Chieh Chao. Predicting malfunction of mobile network base station using machine learning approach. *Proceedings of the 20th Asia-Pacific Network Operations and Management Symposium (APNOMS).* Matsue, Japan, 18-20 September, 2019, pp. 1-4. DOI: 10.23919/APNOMS.2019.8892894

8. Xu M., Baraldi P., Al-Dahidi S., Zio E. Fault prognostics by an ensemble of Echo State Networks in presence of event based measurements. *Engineering Applications of Artificial Intelligence,* vol. 87, 2020, pp. e103346. DOI: 10.1016/j.engappai.2019.103346
9. McKinney W. *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*. 3d Ed. Boston, O'Reilly Media, 2022. 579 p.
10. Hoaglin D. C., Mosteller F., Tukey J. W. *Understanding Robust and Exploratory Data Analysis.* NY, Wiley-Interscience, 2000. 447 p.
11. Flach P. *Machinnoe obuchenie. Nauka i isskustvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannyh* [Machine Learning. The Art and Science of Algorithms that Make Sense of Data]. Transl. from Eng. by A.A. Slinkin. Moscow, DMK Press, 2015. 400 p.
12. Chollet F. *Glubokoe obuchenie na Python* [Deep Learning with Python]. Seriya «Biblioteka programmista». SPb., Piter, 2018. 400 p.