УДК 519.254

Методика подготовки данных для интеллектуального анализа надежности вычислительных комплексов

В. С. Никулин

Подготовка данных для интеллектуального анализа надежности вычислительных комплексов занимает значимую часть времени в процессе извлечения данных из-за сложности механизмов ручной обработки накопленной информации. Поступающий объем данных из эксплуатации вычислительных комплексов является избыточным и разнородным. Существующие методики интеллектуального анализа данных имеют общее назначение, не предоставляя руководство для решения задач в исследуемой области. В данной работе предложена новая методика подготовки данных, ориентированная на анализ надежности вычислительных комплексов, с определением двух наборов задач: общей и специальной подготовки данных. Основным вкладом разработанной методики является выявление особенностей и потенциально значимых показателей исследуемого набора данных. Результат сравнительного анализа подтвердил сокращение времени подготовки данных при автоматизации специализированных задач без изменения качества подготовки данных.

Ключевые слова: вычислительные комплексы, подготовка данных, статистический анализ, извлечение данных, машинное обучение.

1. Введение

Подготовка данных [1] является важным этапом и занимает около 70–80 % общего времени в создании эффективных моделей интеллектуального анализа данных [2]. В связи с тем, что реальные данные могут быть неполными, зашумленными и противоречивыми [3], экспертам требуются систематизированные, подробные методики и автоматизированные инструменты для уменьшения усилий и временных затрат, необходимых для выполнения этого этапа [4].

Проанализированные работы носят ограниченный характер. В [5] описана методика процесса подготовки данных, применимая к различным областям исследований, но без детализации этапов решения задач. Методики [6, 7] ориентированы на конкретный этап подготовки данных, очистку или выбор данных. В [6] авторы предлагают разработку новой методики анализа данных в области энергетической безопасности из-за ограничений существующих методологий. В [8] авторы разрабатывают приложение на основе методики интеллектуального анализа данных.

В процессе выполнения данной работы был проведен предварительный анализ, на основании которого выявлены особенности и проблемы данных, поступающих в процессе эксплуатации вычислительных комплексов (ВК). Разработана методика подготовки данных, состоящая из двух наборов задач: общей и специальной части.

Общая часть предполагает возможность повторного использования в других областях исследований, специальная же часть ориентирована на данные, предоставляемые системой мониторинга (СМ) ВК [9]. Разработан процесс автоматизации предварительного анализа и специального набора задач, позволяющий более эффективно распределять ресурсы в разработке проекта, не влияя на качество подготовки данных.

2. Проблемы и особенности данных, полученных в процессе эксплуатации вычислительных комплексов

В данном разделе описан процесс получения исходных данных в разработанной СМ, схематично представленной на рис. 1. Проведен предварительный анализ для выявления проблем и особенностей данных, полученных из эксплуатации ВК.

2.1. Процесс получения исходных данных

Одной из задач СМ является сбор информации от элементов ВК (источников данных) и фиксация их в центре сбора и обработки данных (далее – ЦСОД) [10].

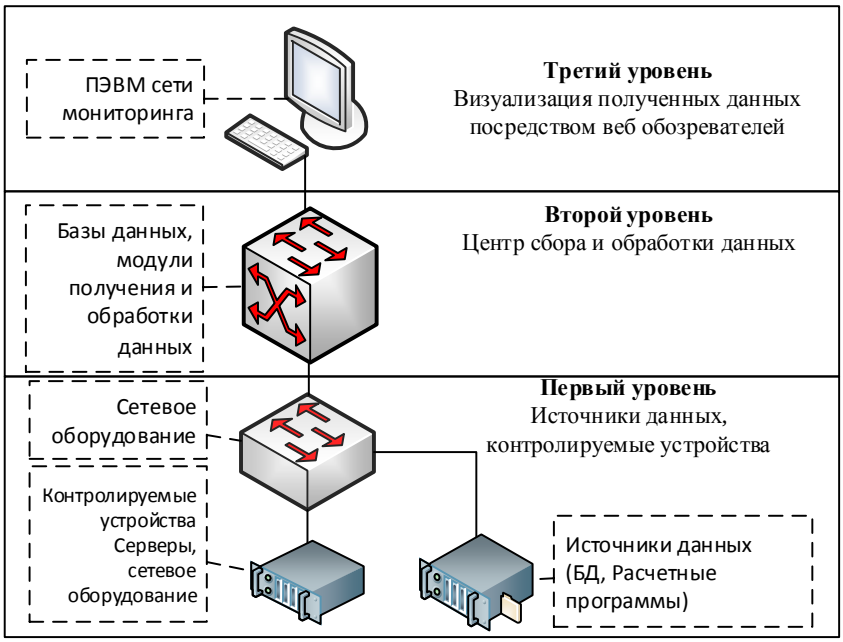

Рис. 1. Разработанная архитектура системы мониторинга

Как видно из рис. 1, СМ логически разделена на три уровня. Передача информации осуществляется между соседними уровнями, позволяя модифицировать и обновлять элементы ВК, не оказывая влияния на общую работоспособность СМ.

В процессе эксплуатации ВК СМ с первого уровня архитектуры (рис. 1) собирает информацию о контролируемых значениях, текущих процессах и отказах аппаратных компонентов. Собранный объем информации фиксируется в базе данных (БД) ЦСОД в виде таблиц, представленных на рис. 2.

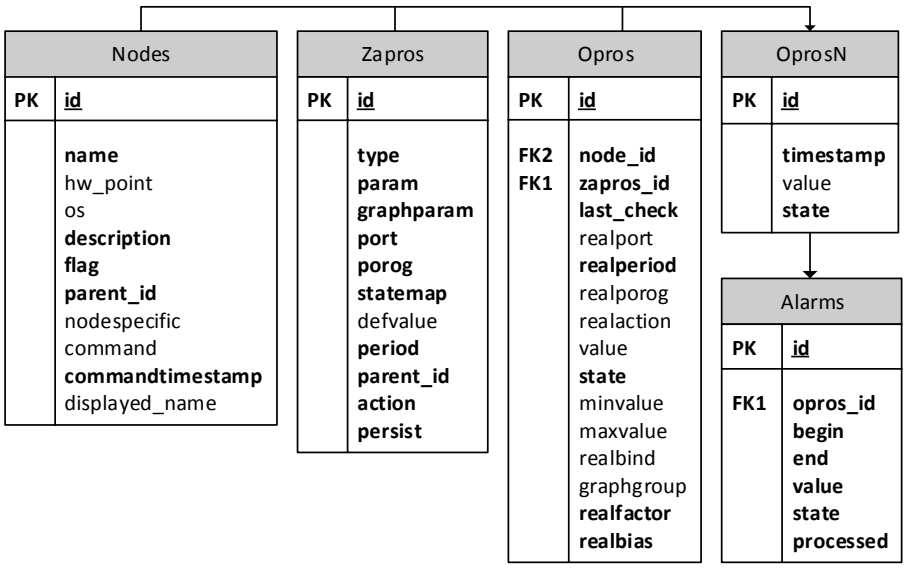

Рис. 2. Архитектура таблиц базы данных центра сбора и обработки данных

БД ЦСОД состоит из таблиц *nodes, zapros, opros, opros№* (где № – это *id*-номер из таблицы *opros*, которой соответствует таблица *opros№*).

Таблица узлов. Таблица *nodes* (узлы) предназначена для хранения перечня контролируемых элементов ВК.

Таблица запросов. Таблица *zapros* (запросы) предназначена для хранения информации о том, какие параметры способна опросить система, о методах их опроса, пороговых значениях.

Таблица опросов. Таблица *opros* (опросы) предназначена для хранения информации об узлах и запросах, а также для хранения максимальных, минимальных и последних полученных значений контролируемых параметров и некоторых других сведений.

Для хранения и архивирования данных в системе мониторинга предусмотрены однотипные таблицы, каждая из которых хранит данные для одного контролируемого параметра за всё время работы СМ. Каждому контролируемому параметру из таблицы *opros* будет соответствовать отдельная таблица *oprosN*, где *N* равен соответствующему *id* из таблицы *opros*.

Таблица нарушений. Таблица *alarms* предназначена для хранения информации о нарушениях пороговых условий значениями контролируемых параметров.

В решениях задач регрессионного анализа на основе реальных данных минимальный наблюдаемый период должен соответствовать циклу эксплуатации [3]. В нашем случае цикл эксплуатации равен одному году [10]. За данный период собрана исходная информация от 50 объектов, опрашиваемых по 31 параметру, среди которых 24 числовых и 7 категориальных. Полученный объем информации – 525000 событий.

На примере одного контролируемого объекта «Supermicro A+ Server-1041M-T2» проведена подготовка данных по разработанной методологии. Параметры, характеризующие условия эксплуатации контролируемого объекта, и технические характеристики представлены в табл. 1.

Таблица 1. Основные технические характеристики контролируемого объекта

Техническая характеристика	Значение
Общие технические характеристики объекта: – модель используемого сервера – модель используемой материнской платы – максимальная потребляемая мощность	 Supermicro A+Server-1041M-T2 Supermicro H8QME-2+ 1000Вт
Допустимые для работы значения параметров воздуха: – температура воздуха – относительная влажность воздуха	 От 10 °С до 35 °С От 8 % до 90 %
Характеристики процессоров: – тип процессоров – разрядность процессоров – количество ядер в одном процессоре – количество процессоров в каждом объекте – производительность каждого объекта	 AMD Opteron 8218 64 2 шт. 4 шт. 41.6 Гфлопс
Подсистема оперативной памяти:	
– объём оперативной памяти – количество модулей памяти – тип модулей памяти	16 Гбайт 8 шт. DDR2 667 ECC REG
Подсистема дисковой памяти:	
– количество HDD	4 шт.
– режим отказоустойчивости	RAID1
Мониторинг:	
– модель платы мониторинга	Supermicro AOC-SIMSO
– тип протокола	IPMI 2.0

2.2. Предварительный анализ исходных данных

Согласно [2] предварительный анализ состоит из последовательных этапов, необходимых для определения целевой переменной, выявления отклонений и особенностей исходных данных, с целью формирования основных задач подготовки данных.

Исходя из основной цели, расчета параметров эксплуатационной надежности ВК, сформированы *требования по наполнению набора данных*:

- Структурными единицами в расчете надежности являются заменяемые элементы [11]. Учитывая архитектуру СМ и идеологию фиксации контролируемых значений и отказов контролируемых устройств, необходимо формировать наборы данных по каждому контролируемому устройству.
- Каждый набор данных должен проходить все этапы подготовки данных.
- Применяемые методы подготовки данных должны иметь возможность автоматизации процессов.

На основании требований по формированию набора данных для контролируемого объекта Supermicro A+ Server-1041M-T2 посредством SQL запроса из БД ЦСОД сформирован массив X_{it}, представленный в табл. 2, где i – количество контролируемых параметров, а t – временной интервал (количество событий).

Таблица 2. Массив данных по одному из контролируемых устройств

ID	Время фиксации	Парам. 1	Парам. 2	Парам. 3	Парам. …	Парам. 30	Парам. 31
1	10:24:20 20.11.19	23.5	24.4	2.96	…	1	NORMAL
2	10:25:20 20.11.19	23.7	25.2	2.99	…	1	NORMAL
…	…	…	…		…	…	
199	10:26:24 27.11.19	23.9	26.3	2.86	…	0	CRITICAL
200	10:28:40 27.11.19	23.5	24.6	2.89	…	1	NORMAL

В сформированном массиве данных (выборочной совокупности), представленном в табл. 2, t = 200 последовательных записей, этот объем информации равен 24 часам эксплуатации ВК и является достаточным для проведения предварительного анализа [12].

В Парам. 1 – Парам. 30 отражены такие характеристики, как температура каждого из процессоров, скорость вращения вентиляторов, доступность по сети Ethernet, напряжение на процессорах и т.д. Тип опрашиваемых характеристик зависит от модели платы мониторинга, представленной в табл. 1.

Выбор целевой переменной. Целевая переменная является фактором с некоторым количеством уровней (классов). Большинство моделей дают лучшие результаты при обучении на целевой переменной с двумя классами. При наличии большего количества классов принимаются специальные дополнительные меры для решения таких задач. Целевая переменная при подготовке данных для обучения кодируется, а после предсказания – декодируется [13].

В сформированном наборе данных в качестве целевой переменной выступает Парам. 31, описывающий, является ли контролируемое устройство работоспособным на момент фиксации, в виде значений «NORMAL/CRITICAL/FAILURE».

Определение особенностей набора данных. Для определения возможных выбросов и пропусков информации на основании выборки (табл. 2) на примере Парам. 3 (Напряжение на CPU 1) построена гистограмма плотности распределения (рис. 3).

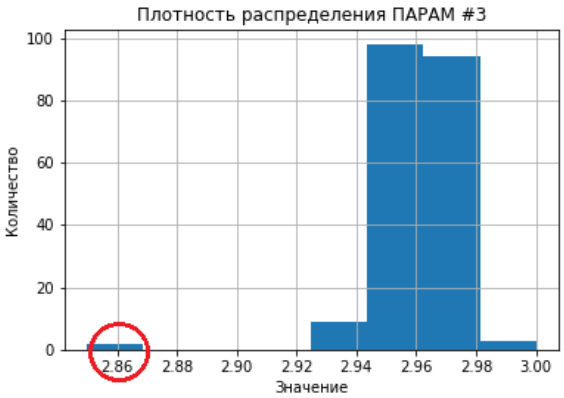

Рис. 3. Плотность распределения контролируемого значения

Как видно из рис. 3, основной объем контролируемых значений (*ось X*) располагается в диапазоне от 2.92 до 3.00. При этом есть некоторое количество значений на оси X, расположенных в диапазоне от 0 до 2.86. Данные значения могут свидетельствовать о наличии шумов и пропусков информации.

Диаграмма размаха (boxplot) позволяет проводить предварительный анализ данных по одиночной или групповой оценке параметров. Она показывает распределение количественных данных, снижая затраты на сравнение между переменными или по уровням категориальной переменной.

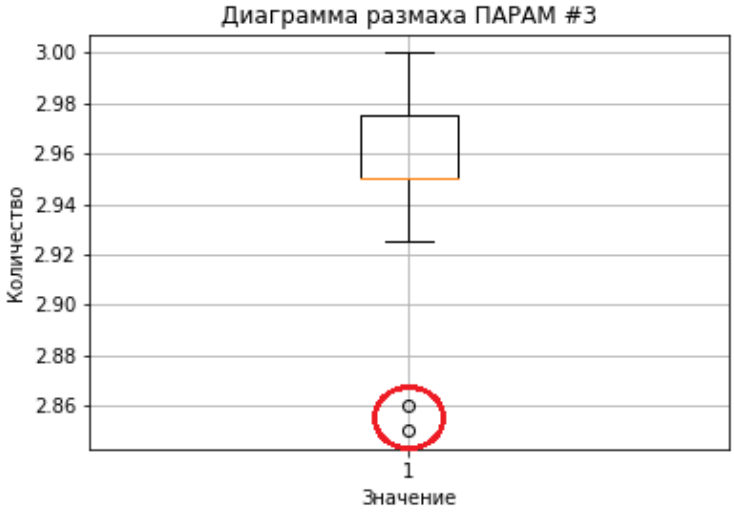

Рис. 4. Распределение контролируемых значений

На представленной диаграмме (рис. 4) «окно» показывает квартиль набора данных, а «усы» – остальное распределение, за исключением точек, которые определены как «выбросы» [14]. Данный показатель может отрицательно сказаться на итоговой регрессии. Для принятия окончательного решения, проведено сравнение с граничными значениями, зафиксированными в СМ. По результатам сравнения сделан вывод, что значения, находящиеся в диапазоне от 0 до 2.86 (рис. 3), а также отраженные в виде точек на рис. 4, являются «выбросами» информации.

2.3. Выводы

Проведенный предварительный анализ показал, что информация, получаемая от СМ, содержит пропуски, выбросы, а также является избыточной и разнородной для расчета показателей надежности ВК. Переменные характеризуются наличием пропущенных значений. Одна переменная является пустой, то есть в каждом из наблюдений не имеет значений.

Наличие пропусков и выбросов связано с разнотипностью применяемых в эксплуатации датчиков опроса состояния ВК, в связи с этим скорость ответа и запись в БД происходит в разные моменты времени. В соответствии с [15] такие пропуски характеризуются как пропуски 2-го типа.

Пропущенные значения могут повлиять на общие результаты. Если игнорировать наличие пропусков в данных или полагать, что достаточно исключить из анализа данные с пропущенными значениями, то существенно вырастает риск получения неверных или незначимых результатов. Данные особенности необходимо учитывать при решении задачи подготовки и нормализации данных для расчета надежности.

Применение диаграммы размаха позволяет сократить временные затраты на предварительный анализ данных за счет возможности комплексной оценки параметров без потери качества.

3. Методика подготовки данных полученных в процессе эксплуатации вычислительных комплексов

Методика подготовки данных для интеллектуального анализа надежности вычислительных комплексов разрабатывалась с учётом особенностей, выявленных в процессе предварительного анализа. Были определены последовательные задачи для каждого этапа подготовки данных, представленные в табл. 3.

Таблица 3. Наборы задач по подготовке данных

1. Общий набор задач		2. Специальный набор задач		
1.1. Очистка данных: – Обнаружение – Устранение – Исправление	1.2. Выбор данных: – Выбор параметров – Выбор записей	2.1. Форматирование: – Тип данных – Структура данных	2.2. Формирование: – Выявление зависимостей – Получение – Создание атрибутов	2.3. Интеграция: – Выявление – Коррекция – Установление отношений

Разделение этапов подготовки данных на общий и специальный набор задач позволяет автоматизировать часть задач процесса подготовки данных, а также применять разработанную методику в других областях сложных технических систем.

3.1. Общая подготовка данных

К общей подготовке данных относятся этапы, не зависящие от конкретной цели и применимые в других областях интеллектуального анализа. Данные этапы направлены на очистку, а также выбор данных из источников для перехода к решению специального набора задач.

Очистка данных. Степень преобразования и очистки исходных данных зависит от степени их загрязненности. Использование метода удаления пропусков из набора данных может привести к потере зависимостей [16].

Проведенный предварительный анализ параметров (табл. 2) показал, что наблюдаемые пропуски в Парам. 3 (Напряжение на CPU 1) связанны с асинхронностью фиксирования показателей контролируемых параметров. Это свидетельствует о том, что данные пропущены не случайно, а ввиду некоторых закономерностей.

Наиболее эффективным является использование методов восстановления пропущенных данных. В соответствии с проведенным сравнительным анализом [13] в нашем случае целесообразно использование метода «подстановки среднего» для заполнения пропущенных значений в массивах данных, содержащих информацию о процессах, происходящих в сложных динамических системах [12].

В основе метода подстановки среднего по выборке используется замена пропусков на среднее значение по данному параметру. Метод видоизменяет изначальное распределение, делая его более сконцентрированным около среднего значения и уменьшая дисперсию.

Расчет среднего значения по выборке производится по формуле:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i , \qquad (1)$$

где $\bar{X}$ – выборочное среднее, n – объем выборки, X_i – i-й элемент выборки.

Исходя из (1), среднее значение Парам. 3 составляет 2.95. Соответственно, все пропущенные значения будут заменены на него. Этот метод дает лучшие результаты, чем методы, основанные на исключении наблюдений.

Выбор данных. Данный этап решает задачи выбора параметров и событий, необходимых для формирования набора данных [17]. Выбор данных осуществляется на основе предвари-

тельного анализа на стадии подготовки данных, а также в процессе обучения модели при применении *L1* регуляризации и расчете коэффициента средней абсолютной ошибки MAE [17]. Предварительный анализ показал, что 12 параметров не содержат информативных показателей и подлежат удалению из выборки. Размерность конечного набора данных составила 19 параметров (в том числе один целевой) и 53000 событий.

3.2. Специальная подготовка данных

Специальная подготовка данных конкретизирует этапы, имеющие более тесную связь с конечной целью интеллектуального анализа. Этапы включают в себя следующие задачи.

Форматирование данных. Изменение типа данных и синтаксической структуры данных атрибутов и значений (при необходимости).

В связи с тем, что алгоритмы машинного обучения не работают с текстовыми данными в явном виде, для Парам. 31, отражающего текущее состояние в виде значений «NORMAL/CRITICAL/FAILURE», изменен тип данных на десятичное представление в соответствии: 1 – NORMAL, 2 – CRITICAL, 3 – FAILURE.

Формирование данных. Выявление, получение и создание новых параметров (при необходимости) или заполнение пропущенных значений с помощью арифметических операций.

Для некоторых параметров требуется формирование производных признаков с дальнейшим преобразованием в векторы для моделей машинного обучения, а также трансформация для повышения точности алгоритмов интеллектуального анализа данных.

Например, в нашем случае Парам. 3 (Напряжение на CPU 1), Парам. 4 (Напряжение на CPU 2), Парам. 5 (Напряжение на CPU 3), Парам. 6 (Напряжение на CPU 4) мультиколлинеарные, т.е. линейно зависимые [16]. Как известно, линейная зависимость признаков снижает качество итогового обучения. Чтобы не потерять информативность и избавиться от линейной зависимости, был создан новый параметр – Парам. 3–6:

$$P = \frac{1}{n}\sum_{i=1}^{n} x_i , \qquad (2)$$

где P – новый параметр контролируемого объекта, n – количество объединяемых параметров, x_i – значение параметров.

Интеграция данных. Выявление, исправление конфликтов интеграции, установление отношений данных и выбор схемы интеграции данных [17].

Интеграция с хранилищем данных БД ЦСОД выполнена по схеме «звезда». Создана таблица с атрибутами: ключ, время события, параметры, состояние контролируемого устройства. Окончательная интеграция набора данных заключается в формировании итогового набора данных с содержанием всех событий за указанный период эксплуатации.

После формирования итогового набора данных необходимо провести *нормализацию параметров* [17].

Алгоритмы машинного обучения, основанные на градиентных методах, сильно чувствительны к «масштабу» значений данных. Для последовательного масштабирования данных временных рядов используются методы нормализации и стандартизации. Нормализация предполагает замену номинальных значений параметров так, чтобы каждый из них лежал в диапазоне от 0 до 1. Стандартизация же подразумевает такую предобработку данных, после которой каждый признак имеет среднее, равное 0, и дисперсию, равную 1. В библиотеках Scikit-Learn уже есть готовые для реализации этих методов функции. В связи с отсутствием информации о законе распределения величин целесообразно проводить нормализацию данных [5].

После проведения всех этапов подготовки данных, а также нормализации параметров итоговый набор данных имеет вид, представленный в табл. 4.

Таблица 4. Итоговый набор данных по одному из контролируемых устройств

Парам. 1	Парам. 2	Парам. 3–6	Парам. 7	Парам. 8	Парам. …	Парам. 31
0.235	0.244	0.0296	0.244	0.244	…	1
0.237	0.252	0.0299	0.252	0.252	…	1
…	…	…	…	…	…	…
0.239	0.263	0.0286	0.263	0.263	…	2
0.235	0.246	0.0289	0.246	0.246	…	1

3.3. Автоматизация этапов подготовки данных

В связи с периодичностью повторения задачи обработки данных, а также выявленными особенностями работы СМ целесообразно выполнить переконфигурацию СМ и автоматизацию специального набора задач подготовки данных для сокращения временных затрат в будущем.

Конфигурация СМ необходима для синхронизации ответов от применяемых в эксплуатации датчиков опроса состояния ВК с целью минимизации пропусков 2-го типа, выявленных в ходе предварительного анализа. Данный шаг позволит снизить нагрузку на метод автоматизации и увеличить точность восстанавливаемых данных.

Подсистема автоматической подготовки данных построена на основе разработанного метода автоматизации, представленного на рис. 5. В качестве инструментов для разработки использованы язык программирования общего назначения Python 3.5 с библиотеками sclearn, pandas, numpy и язык SQL для установления связи между приложением и хранилищем данных.

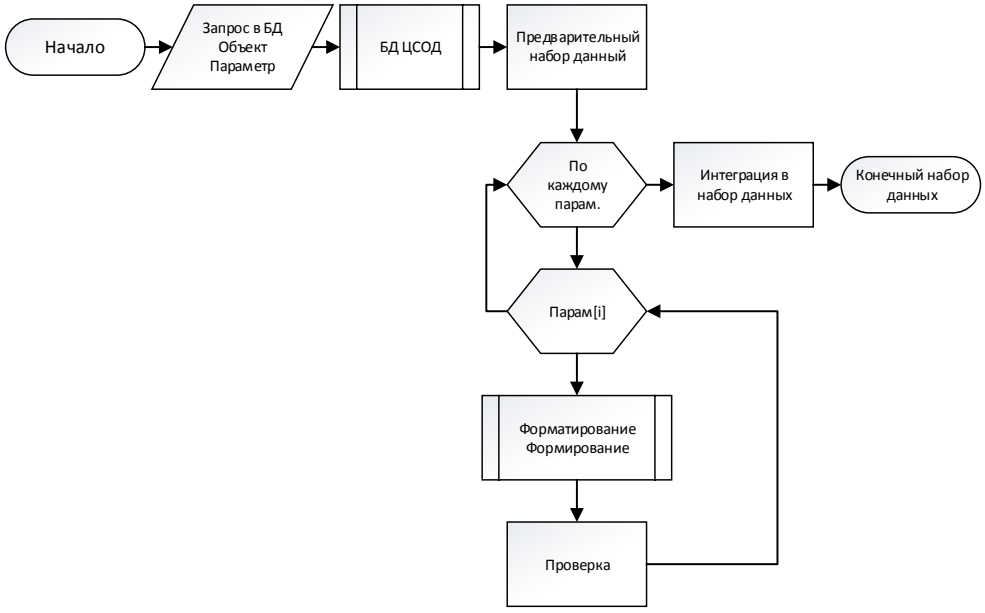

Рис. 5. Общая схема метода автоматизации процесса подготовки данных

Метод (рис. 5) предполагает автоматизацию специального набора задач, а именно этапов форматирования, формирования и интеграции данных. Входными параметрами данного метода являются БД, объект и параметры. В нашем случае в качестве объекта выступает контролируемое устройство ВК, а в качестве параметров – опрашиваемые СМ контролируемые значения данного объекта. Процесс реализации алгоритма автоматизации подготовки данных состоит из следующих этапов:

– выполнение запросов в БД с использованием набора правил Python DB-API и языка запросов SQL;

– формирование данных в предварительные выборки посредством библиотек matplotlib.pandas;

– отбор параметров и объектов с учетом их релевантности, качества и технических ограничений (объема и типа);

– очистка данных, удаление некорректных значений (Missing values или NA), исключение дублей и однотипных параметров объекта, восстановление уникальности, целостности и логических связей гибридным адаптационным методом;

– создание производных признаков с дальнейшим преобразованием в векторы для моделей машинного обучения, а также трансформация для повышения точности алгоритмов машинного обучения;

– интеграция, слияние данных из различных источников (информационных систем, таблиц, протоколов и пр.), включая их агрегацию, вычисление новых значений путем суммирования информации из множества существующих записей;

– форматирование и формирование набора данных.

3.4. Проверка метода автоматизации этапов подготовки данных

Проверка метода автоматизации проведена на данных, подготовленных ручным способом. Из наборов случайным образом удалялось некоторое количество значений параметров. Качество восстановленных значений методом автоматизации оценивалось по доле пропущенных значений (точности) и средней квадратичной ошибки (погрешности) [16]. Тестирование методов проводилось на двух наборах данных:

- Набор 1. Выборка значений показателей работы системы электропитания ВК. Число показателей – 18, число событий – 2330;
- Набор 2. Выборка значений показателей системы хранения данных ВК. Число показателей – 17, число событий – 1250.

Основываясь на результатах, представленных в табл. 5, метод автоматизации дает точную оценку пропущенных значений с допустимой погрешностью.

Таблица 5. Сравнение качественных показателей подготовки данных

	Набор данных № 1		Набор данных № 2	
	Точность	Погрешность	Точность	Погрешность
Форматирование данных	90.7 %	12 %	94.7 %	9 %

В табл. 6 показано сравнение времени выполнения этапов подготовки данных при выполнении вручную и автоматически.

Таблица 6. Сравнение временных показателей подготовки данных

Задача	Ручная подготовка (мин)	Автомат. подготовка (мин)	Снижение временных затрат на:
Форматирование данных	33.53	0.058	99.83 %
Формирование признаков	5.16	0.33	99.61 %
Интеграция данных	40.5	1.4	96.55 %

Результаты сравнения временных показателей подготовки данных доказывают значительное сокращение времени при автоматизации этапов подготовки данных.

4. Заключение

Проведение предварительного анализа данных позволяет выявить особенности и потенциально значимые показатели исследуемого набора данных.

Разделение этапов подготовки данных позволяет автоматизировать специальный набор задач, а также применять предложенную методику не только в области расчета показателей эксплуатационной надежности ВК, но и других сложных технических систем. По результатам тестирования метода автоматизации этапов подготовки данных подтверждено сокращение времени подготовки данных более чем на 95 %, при этом качество подготовки данных осталось на должном уровне.

Практическим результатом разработанной методики является сформированный набор данных, необходимый для дальнейшего расчета параметров эксплуатационной надежности ВК с применением интеллектуального анализа данных.

Литература

1. *Hellerstein J., Carreras C., Rattenbury T., Kandel S., Heer J.* Principles of Data Wrangling: Practical Techniques for Data Preparation. 1st ed. California: O'Reilly Media, 2017. P. 50–62.
2. *Порутчиков М. А.* Анализ данных. Самара: Изд-во Самарского университета, 2016. 29 с.
3. *Zhang S., Zhang C., Yang Q.* Data preparation for data mining // Appl. Artif. Intell. 2003. P. 375–381.
4. *Захаров Д. Н., Никулин В. С.* Анализ методов статистической оценки эксплуатационной надежности вычислительных комплексов // Наукоемкие технологии в космических исследованиях Земли. 2020. Т. 12, № 1. С. 64–69. DOI: 10.36724/2409-5419-2020-12-1-64-69.
5. *Chapman P., Clinton J., Kerber R.* CRISP-DM 1.0 Step-by-step data mining guide. // CRISP-DM Consortium. 2000.
6. *Береснева Н. М, Курганская О. В.* Методология подготовки данных для вычислительных экспериментов в исследованиях энергетической безопасности России // Вестник Иркутского государственного технического университета. 2017. Т. 21, № 9. С. 45–57.
7. *Amir R. Razavi.* A Data Pre-processing Method to Increase Efficiency and Accuracy in Data Mining. DOI:10.1007/11527770 59. 2005.
8. *Wei, C. K., Su, S., and Yang, M. C.* Application of data mining on the development of a disease distribution map of screened community residents of Taipei county in Taiwan // J. Med. Syst. 2012. № 36. P. 2021–2027. DOI:10.1007/s10916-011-9664-7.
9. *Никулин В. С., Павлова А. И.* Создание автоматизированной системы сбора сведений о качестве функционирования вычислительных комплексов // Наука молодых. 2017. № 5. С. 540–544.
10. *Никулин В. С.* Сравнительный анализ СУБД для реализации подсистемы хранения событий мониторинга вычислительных комплексов // Сборник научных трудов «Наука. Технологии. Иновации». 2019. Т. 2. С. 46–48.
11. *Матвеевский В. Р.* Надежность технических систем: учебное пособие. М.: Московский государственный институт электроники и математики, 2002. 113 с.
12. *Карлов И. А.* Восстановление пропущенных данных при численном моделировании сложных динамических систем // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика. Телекоммуникации. Управление. 2013. № 186. С. 137–144.
13. *Карлов И.* Методы восстановления пропущенных значений с использованием инструментария Data Mining // Вестник Сибирского гос. аэрокосмического ун-та им. Академика М. Ф. Решетнева. 2011. № 7 (40). С. 29–33.

14. *Кузовлев В. И.* Метод выявления аномалий в исходных данных при построении прогнозной модели решающего дерева в системах поддержки принятия решений // Наука и образование: науч. изд. МГТУ им. Н. Э. Баумана. 2012. № 9. С. 16.
15. *Schafer J. L., Graham J. W.* Missing data: Our view to the state of the art // Psychological methods. 2002. P. 51–61.
16. *Литтл Р. Д. А., Рубин Д.* Статистический анализ данных с пропусками. М.: Финансы и статистика, 1991. 336 с.
17. *Чубукова И. А.* Data Mining: учебное пособие. 2-е-е изд. М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2008.

Статья поступила в редакцию 24.03.2020;
переработанный вариант – 03.06.2020.

Никулин Владимир Сергеевич
аспирант кафедры информационных технологий НГУЭУ «НИНХ» (630099, Новосибирск, ул. Каменская, 52/1), e-mail: nikulin-94@inbox.ru.
Область научных интересов: вычислительные комплексы, машинное обучение, надежность технических систем.

Methods of data preparation data for intelligent analysis of the computer systems reliability

V. Nikulin

Data preparation for intelligent analysis of the computer systems reliability takes a significant part of time in the process of data extraction. The incoming data from the computer systems operation is redundant and heterogeneous. Existing data mining techniques have a general purpose. In this paper, we propose a new methodology for data preparation. The methodology focuses on the analysis of computer systems reliability with two tasks setting: general and special data preparation. The main contribution of the developed methodology is the identification of features and potentially significant indicators in the data set. The comparative analysis results confirmed time reduction for data preparation when automating specialized tasks.

Keywords: computer systems, data preparation, statistical analysis, data mining, machine learning.