ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р 71484.4— 2024 (ИСО/МЭК 5259-4:2024)

Искусственный интеллект

КАЧЕСТВО ДАННЫХ ДЛЯ АНАЛИТИКИ И МАШИННОГО ОБУЧЕНИЯ

Часть 4

Структура процесса управления качеством данных

(ISO/IEC 5259-4:2024, MOD)

Издание официальное

Москва
Российский институт стандартизации
2024

gost.e-ecolog.ru

Предисловие

1 ПОДГОТОВЛЕН Научно-образовательным центром компетенций в области цифровой экономики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В.Ломоносова» (МГУ имени М.В.Ломоносова) и Обществом с ограниченной ответственностью «Институт развития информационного общества» (ИРИО) на основе собственного перевода на русский язык англоязычной версии стандарта, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 28 октября 2024 г. № 1552-ст

4 Настоящий стандарт является модифицированным по отношению к международному стандарту ИСО/МЭК 5259-4:2024 «Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 4. Структура процесса управления качеством данных» [ISO/IEC 5259-4:2024 «Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework», MOD] путем изменения отдельных фраз (слов, значений, показателей, ссылок), которые выделены в тексте курсивом.

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном стандарте, приведены в дополнительном приложении ДА

5 ВВЕДЕН ВПЕРВЫЕ

*Правила применения настоящего стандарта установлены в статье 26 Федерального закона от 29 июня 2015 г. № 162-ФЗ «О стандартизации в Российской Федерации». Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (*gost.e-ecolog.ru

© ISO, 2024
© IEC, 2024
© Оформление. ФГБУ «Институт стандартизации», 2024

Настоящий стандарт не может быть полностью или частично воспроизведен, тиражирован и распространен в качестве официального издания без разрешения Федерального агентства по техническому регулированию и метрологии

Содержание

1 Область применения 1
2 Нормативные ссылки 1
3 Термины и определения 2
4 Сокращения 3
5 Принципы процесса управления качеством данных 3
6 Структура процесса управления качеством данных 3
6.1 Общие положения 3
6.2 Планирование качества данных 6
6.3 Оценка качества данных 6
6.4 Повышение качества данных 7
6.5 Валидация процесса управления качеством данных 7
6.6 Применение структуры процесса управления качеством данных 8
7 Процесс управления качеством данных для машинного обучения 8
7.1 Общие положения 8
7.2 Формирование требований к данным 9
7.3 Планирование работы с данными 11
7.4 Комплектование данных 11
7.5 Подготовка наборов данных 12
7.6 Предоставление данных 17
7.7 Вывод данных из эксплуатации 17
8 Методы и процесс разметки данных 17
8.1 Общие положения 17
8.2 Принципы разметки данных 17
8.3 Методы разметки данных 18
8.4 Процесс разметки данных 18
9 Роли участников 21
9.1 Общие положения 21
9.2 Планировщик данных 21
9.3 Создатель данных 21
9.4 Сборщик данных 22
9.5 Инженер данных 22
9.6 Распорядитель данных 22
9.7 Пользователь данных 22
10 Процесс управления качеством данных для машинного обучения с частичным привлечением учителя 22
10.1 Общие положения 22
10.2 Формирование требований к данным 22
10.3 Планирование работы с данными 22
10.4 Комплектование данных 23
10.5 Подготовка данных 23
10.6 Предоставление данных 23
10.7 Вывод данных из эксплуатации 23
11 Процесс управления качеством данных для обучения с подкреплением 23
11.1 Общие положения 23
11.2 Формирование требований к данным 23
11.3 Планирование работы с данными 23
11.4 Комплектование данных 23
11.5 Подготовка данных 24
11.6 Предоставление данных 24
11.7 Вывод данных из эксплуатации 24
12 Процесс управления качеством данных для аналитики 25
12.1 Общие положения 25
12.2 Формирование требований к данным 25
12.3 Планирование работы с данными 25

12.4 Комплектование данных 25
12.5 Подготовка данных 25
12.6 Предоставление данных 27
12.7 Вывод данных из эксплуатации 27
Приложение ДА (справочное) Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном стандарте 28
Библиография 29

Введение

Количество продуктов, систем или решений, связанных с искусственным интеллектом, в последние годы быстро растет. Одна из общих характеристик систем искусственного интеллекта, особенно для систем, использующих машинное обучение с учителем, заключается в том, может ли система искусственного интеллекта обучаться на наборе данных перед развертыванием или ее можно обучать динамически в процессе использования системы.

Общепризнано, что данные играют существенную роль в системах искусственного интеллекта на основе машинного обучения. Для всех подходов к машинному обучению с учителем, без учителя, с частичным привлечением учителя, с подкреплением качество данных может быть главной проблемой при создании и использовании данных для обучения и оценки систем машинного обучения. Как правило, при использовании более точных и богатых данных результаты аналитики и машинного обучения могут быть более полезными и надежными. Кроме того, для разработки систем искусственного интеллекта на основе обучения с учителем необходимы большие объемы размеченных данных для конкретных задач. Это делает аккуратно размеченные данные одним из самых важных ресурсов в сфере искусственного интеллекта. В настоящее время существует проверенный рынок промышленных сервисов и инструментов для разметки обучающих данных. Сегодня этот рынок достигает уровня зрелости, который оправдывает разработку международных стандартов в интересах поставщиков и пользователей этих услуг и инструментов для обеспечения высокого качества размеченных данных.

В настоящем стандарте описывается внедрение единой стандартизированной процедуры обработки данных в отношении качества данных для аналитики и машинного обучения.

В разделе 5 описываются принципы процесса управления качеством данных, в разделе 6 описывается структура процесса управления качеством данных. В разделе 7 описывается процесс управления качеством данных для машинного обучения, в разделе 8 описываются методы и процессы маркировки данных, в разделе 9 описаны роли участников в процессах управления качеством данных, в разделах 10 и 11 описаны особенности процессов управления качеством данных для машинного обучения с частичным привлечением учителя и для обучения с подкреплением. В разделе 12 описывается, как структура процессов управления качеством данных применяется к аналитике.

Настоящий стандарт подробно описывает структуру процессов, которая может быть использована для выполнения требований, указанных в *ГОСТ Р 71484.3*. Он также показывает связь с процессами, которые отображены в модели жизненного цикла данных в стандарте *ГОСТ Р 71484.1*.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Искусственный интеллект

КАЧЕСТВО ДАННЫХ ДЛЯ АНАЛИТИКИ И МАШИННОГО ОБУЧЕНИЯ

Часть 4

Структура процесса управления качеством данных

Artificial intelligence.
Data quality for analytics and machine learning.
Part 4. Data quality process framework

Дата введения — 2025—01—01

1 Область применения

Настоящий стандарт устанавливает общие организационные подходы, используемые независимо от типа, размера или характера организации, для обеспечения качества данных для обучения и оценки в области аналитики и машинного обучения. Стандарт включает в себя руководство по процессу управления качеством данных для:

- машинного обучения с учителем;
- машинного обучения без учителя;
- машинного обучения с частичным привлечением учителя;
- аналитики.

Настоящий стандарт применим к обучающим и тестовым данным, которые поступают из различных источников, включая сбор и комплектование данных, подготовку данных, разметку данных, оценку и использование данных. Настоящий стандарт не определяет конкретные сервисы, платформы или инструменты.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ГОСТ Р 54995 Телевидение вещательное цифровое. Требования к кодированию аудио и видеосигналов для приложений вещания, основанных на транспортных потоках MPEG-2

ГОСТ Р 59926—2021 (ISO/IEC TR 20547-2:2018) Информационные технологии. Эталонная архитектура больших данных. Часть 2. Варианты использования и производные требования

ГОСТ Р 71476 (ИСО/МЭК 22989:2022) Искусственный интеллект. Концепции и терминология искусственного интеллекта

ГОСТ Р 71484.1 (ИСО/МЭК 5259-1:2024) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, термины и примеры

ГОСТ Р 71484.2 (ИСО/МЭК 5259-2:2024) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 2. Показатели качества данных

ГОСТ Р 71484.3 (ИСО/МЭК 5259-3:2024) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 3. Требования и рекомендации по управлению качеством данных

Издание официальное

ГОСТ Р ИСО 2859-1 Статистические методы. Процедуры выборочного контроля по альтернативному признаку. Часть 1. Планы выборочного контроля последовательных партий на основе приемлемого уровня качества

ГОСТ Р ИСО/МЭК 17826 Информационные технологии. Интерфейс управления облачными данными (CDMI)

ГОСТ Р ИСО/МЭК 19794-5 Автоматическая идентификация. Идентификация биометрическая. Форматы обмена биометрическими данными. Часть 5. Данные изображения лица

Примечание — При пользовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указателю «Национальные стандарты», который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя «Национальные стандарты» за текущий год. Если заменен ссылочный стандарт, на который дана недатированная ссылка, то рекомендуется использовать действующую версию этого стандарта с учетом всех внесенных в данную версию изменений. Если заменен ссылочный стандарт, на который дана датированная ссылка, то рекомендуется использовать версию этого стандарта с указанным выше годом утверждения (принятия). Если после утверждения настоящего стандарта в ссылочный стандарт, на который дана датированная ссылка, внесено изменение, затрагивающая положение, на которое дана ссылка, то это положение рекомендуется применять без учета данного изменения. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, рекомендуется применять в части, не затрагивающей эту ссылку.

3 Термины и определения

В настоящем стандарте применены следующие термины с соответствующими определениями:

3.1

аутсорсинг (outsourcing): Любая работа (или процесс), передаваемая организацией для исполнения внешней организации.

[*ГОСТ Р 56398—2015, пункт 3.14*]

3.2 **автономная аннотация** (stand-off annotation): Аннотация, охватывающая различные слои первичных данных и сериализуемая в документе, отделенном от документа, который содержит первичные данные.

Примечание — См. [1], пункт 2.7.

3.3

служба облачных вычислений (cloud service): Одна или более возможностей, предоставляемых через облачные вычисления, вызываемых посредством определенного интерфейса.

[*ГОСТ ISO/IEC 17788—2016, пункт 3.2.8*]

3.4 **создатель данных** (data originator): Сторона, которая создала данные и может обладать правами на них.

Примечания

1 Создателем данных может быть физическое лицо.

2 Создатель данных может отличаться от физического или юридического лица, упомянутого в данных, описанного ими, либо явно или неявно связанного с ними. Например, создателем данных могут быть собраны персональные данные, идентифицирующие других физических лиц. Эти субъекты персональных данных также могут обладать правами в отношении такого набора данных.

3 Права могут включать право на публичное использование, право на отображение имени, право на идентичность, право запрещать использование данных оскорбительным образом.

4 См. [2], пункт 3.2.

3.5

ограничивающий прямоугольник (bounding box): Прямоугольная область, содержащая аннотируемый объект.

[*ГОСТ Р 70268.2—2022, пункт 3.3*]

3.6 **сегментация** (segmentation): Процесс отделения интересующих объектов от их окружения.

Примечания

1 Сегментация может применяться к двумерным, трехмерным, растровым или векторным данным.

2. См. [3], пункт 3.1.13

3.7 **ключевая точка** (key-point): Точка на объекте, представляющая интерес.

4 Сокращения

В настоящем стандарте применены следующие сокращения:

ИИ — искусственный интеллект;

МО — машинное обучение;

DLC — жизненный цикл данных (data life cycle);

DQPF — инструментарий управления качеством данных (data quality process framework).

5 Принципы процесса управления качеством данных

ГОСТ Р 71484.1 определяет качество данных как характеристику того, что данные соответствуют требованиям организации в конкретных условиях.

Независимо от данных и методологии оценки процесс управления качеством данных для аналитики и машинного обучения должен основываться на общих принципах, которые применяются во всей модели жизненного цикла данных. Организациям следует определять и документировать общие принципы качества данных, принимая во внимание следующие аспекты:

- данные и наборы данных соответствуют конкретной задаче МО или аналитики;
- используется модель качества данных, основанная на характеристиках качества данных;
- валидируется качество данных в соответствии с требованиями к качеству данных на основе измерения показателей качества данных и заданных целевых значений;
- на каждой стадии верифицируется соответствие процесса заданным целевым значениям и другим требованиям;
- тестируется корректность и робастность, в том числе такими методами, как состязательное тестирование, предназначенными для выявления ошибок;
- соответствие требованиям организации в области безопасности, защиты персональных данных, справедливости и этики;
- защищается здоровье и благополучие аннотаторов и других лиц, участвующих в процессе управления качеством данных;
- документируется прогресс и соблюдение заданных принципов и требований.

6 Структура процесса управления качеством данных

6.1 Общие положения

Структура процесса управления качеством данных (data quality process framework), основанная на принципах из раздела 5, предназначена для того, чтобы предоставить организациям возможность управлять качеством данных так, чтобы они соответствовали заданным требованиям.

Применение структуры процесса управления качеством данных может способствовать формированию:

- стратегии управления качеством данных;
- плана по управлению качеством данных;
- требований к качеству данных, включая модель качества данных, показатели качества данных и их целевые значения;
- результатов управления качеством данных (таких, как итоги измерения качества данных, отчеты об ошибках, применяемые методы улучшения и аугментации данных);
- рекомендаций по улучшению процессов управления качеством данных;
- разрешений об использовании наборов данных в проектах аналитики или машинного обучения.

На рисунке 1 представлена общая структура процесса управления качеством данных, включающая следующие компоненты:

- планирование качества данных: разработка планов управления качеством данных посредством анализа требований к качеству данных и жизненного цикла данных и определения методов управления качеством данных;
- оценка качества данных: измерение и мониторинг качества данных в модели жизненного цикла данных и предоставление результатов для формирования плана управления качеством данных;
- повышение качества данных: внедрение процессов улучшения качества данных (таких, как очистка, преобразование, аугментация, масштабирование);
- валидация процесса управления качеством данных: оценка показателей качества данных и процессов подтверждения того, что данные соответствуют требованиям, и при необходимости предоставление обратной связи для процесса улучшения качества данных.



Рисунок 1 — Общая структура процесса управления качеством данных

На рисунке 2 показана взаимосвязь между моделью жизненного цикла данных в *ГОСТ Р 71484.1* и структурой процесса управления качеством данных, которую можно использовать во время всего жизненного цикла данных для управления качеством данных.

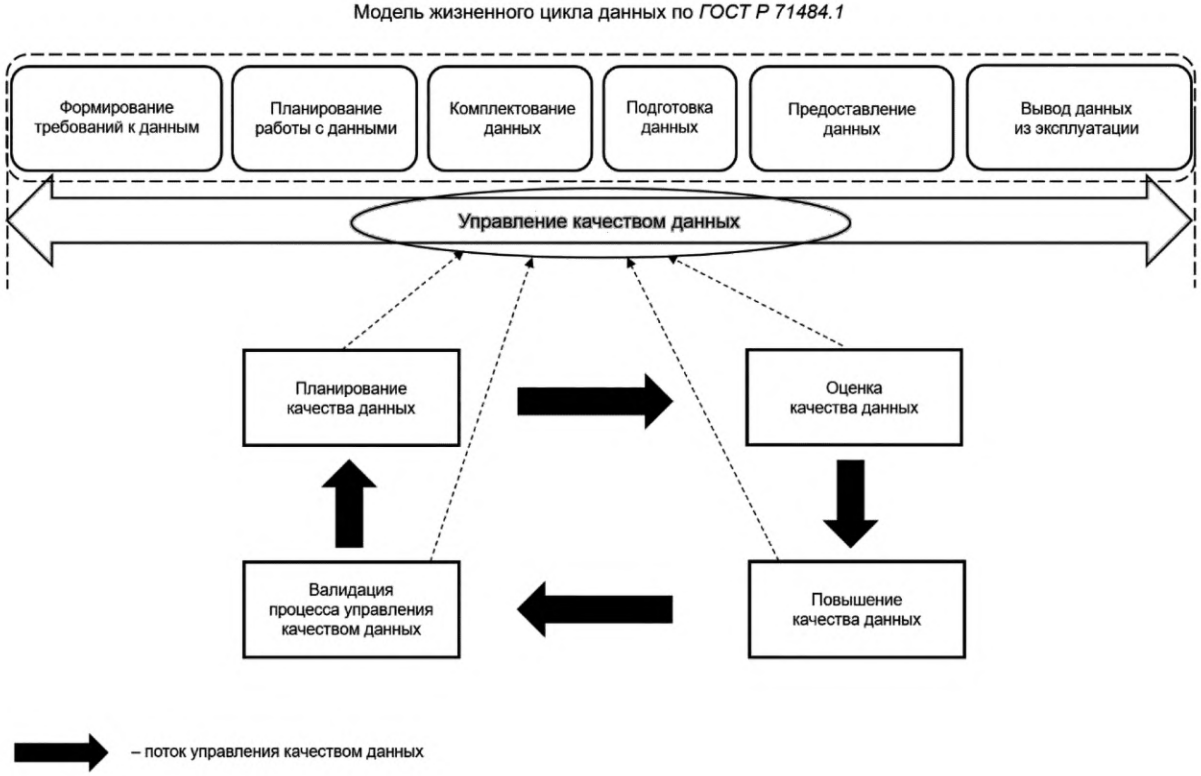

Рисунок 2 — Связь между моделью жизненного цикла данных и структурой процесса управления качеством данных

6.2 Планирование качества данных

Виды деятельности и результаты процесса планирования качества данных включают:

- виды деятельности:
  - анализ требований к качеству данных от заинтересованных сторон в модели жизненного цикла данных;
  - построение модели качества данных согласно *ГОСТ Р 71484.1* и *ГОСТ Р 71484.2;*
  - определение соответствующих показателей качества данных и их целевых значений согласно *ГОСТ Р 71484.2*;
- результаты:
  - модель качества данных;
  - показатели качества данных;
  - целевые значения показателей качества данных;
  - результаты управления качеством данных, такие как итоги измерения качества данных, отчеты об ошибках, применяемые методы улучшения и аугментации данных.

6.3 Оценка качества данных

Виды деятельности и результаты процесса оценки качества данных включают:

- виды деятельности:
  - применение показателей качества данных;
  - сравнение итогов измерения качества данных с установленными целевыми значениями показателей;

- проверка выполнимости требований к данным;
- результаты:
  - документирование различий между итогами измерения показателей качества данных и их целевыми значениями и анализа оказываемого воздействия;
  - документирование результатов оценки качества данных.

6.4 Повышение качества данных

Виды деятельности и результаты процесса повышения качества данных включают:
- виды деятельности:
  - применение методов повышения качества данных, описанных в 7.5.9;
  - повторение процессов оценки качества данных;
- результаты:
  - документирование использованных методов улучшения качества данных;
  - документирование результатов оценки качества данных.

6.5 Валидация процесса управления качеством данных

Виды деятельности и результаты процесса валидации качества данных включают:
- виды деятельности:
  - оценка соответствия данных заданным требованиям;
  - если данные не соответствуют требованиям, анализируются и реализуются изменения в отдельных процессах управления качеством данных;
  - заполнение записей о происхождении данных (если применимо);
- результаты:
  - документирование результатов оценки качества данных;
  - отчеты об ошибках;
  - рекомендации по улучшению процессов управления качеством данных;
  - разрешения об использовании данных при определенных обстоятельствах, выданные соответствующими заинтересованными сторонами.

Примечания

1 Валидация процесса управления качеством данных осуществляется экспертами и другими заинтересованными сторонами в области качества данных.

2 На рисунке 3 показано, как можно валидировать процесс управления качеством данных.

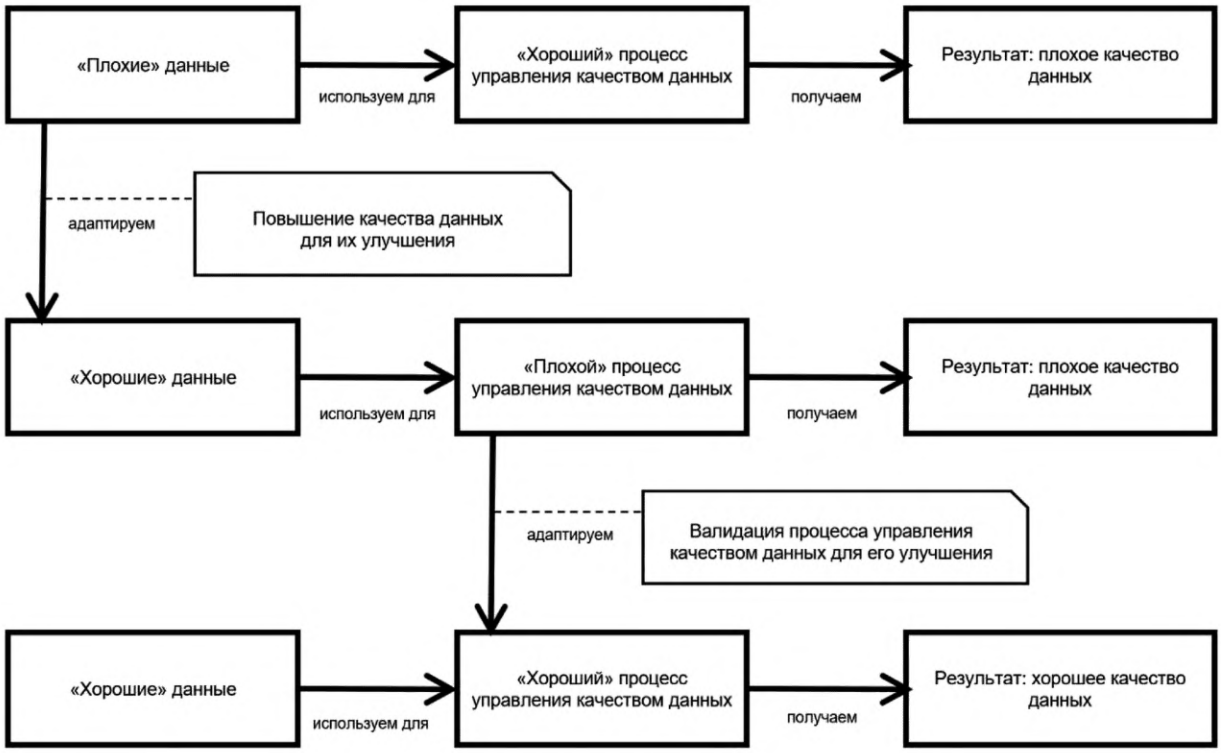

Рисунок 3 — Связь между качеством данных и процессами управления качеством данных

6.6 Применение структуры процесса управления качеством данных

Структура процесса управления качеством данных предоставляет дополнительную информацию для управления качеством данных в модели жизненного цикла данных. Пользователи должны применять данную структуру в сочетании с отдельными отобранными процессами управления качеством данных, описанными в настоящем стандарте.

Например, планирование качества данных в структуре процесса управления качеством данных может отображаться на детально описанные процессы на стадии планирования данных в модели жизненного цикла данных. Аналогичным образом процесс оценки качества данных может отображаться на процесс оценки качества данных на стадии подготовки данных в модели жизненного цикла данных.

Данные, соответствующие требованиям, необходимы для обучения, тестирования и валидации моделей машинного обучения, а также для задач анализа данных. Если обучающие, тестовые или валидационные данные для машинного обучения не соответствуют требованиям, то и выходные данные обученных моделей также могут не соответствовать требованиям. Полученные данные часто бывают несовершенны (например, недостаточно точны, основаны на недостаточном количестве выборок, устаревшие). Процесс управления качеством данных можно использовать для улучшения и оптимизации данных в той мере, в какой они будут соответствовать требованиям организации. Кроме того, сам процесс управления качеством данных можно улучшить и оптимизировать.

На рисунке 3 показана взаимосвязь между качеством данных и процессами управления качеством данных.

7 Процесс управления качеством данных для машинного обучения

7.1 Общие положения

Целью процесса управления качеством данных, описанного в настоящем стандарте, является предоставление рекомендаций и передовых методов, которые организации могут использовать для обеспечения соответствия данных, используемых для МО, предъявляемым к ним требованиям. Про-

цесс управления качеством данных основан на применении структуры процесса управления качеством данных согласно разделу 6. Особенности этого процесса для конкретной задачи МО будут зависеть:

- от самой задачи МО (например, обработка изображений, прогнозирование, обработка естественного языка);
- подхода к МО;
- процессов МО;
- области применения;
- типов данных;
- требований к качеству данных.

Процесс управления качеством данных может быть связан со следующими стадиями (см. рисунок 4):

- формирование требований к данным;
- планирование работы с данными;
- комплектование наборов данных;
- подготовка наборов данных;
- предоставление данных;
- вывод данных из эксплуатации.

7.2 Формирование требований к данным

Требования к данным основаны на контексте задачи, приложения и подхода МО и закладывают основу для остальных стадий процесса управления качеством данных. Для достижения качества данных при формировании требований к данным должны быть заданы и задокументированы как минимум следующие аспекты:

- необходимые признаки данных;
- необходимый объем данных;
- происхождение;
- приемлемая смещенность;
- статистические характеристики;
- репрезентативность с точки зрения поведения, демографии и географического местоположения субъектов в модели МО;
- модель качества данных, основанная на выбранных характеристиках качества данных;
- соответствующие показатели качества данных;
- целевые значения показателей качества данных;
- требования законодательства.

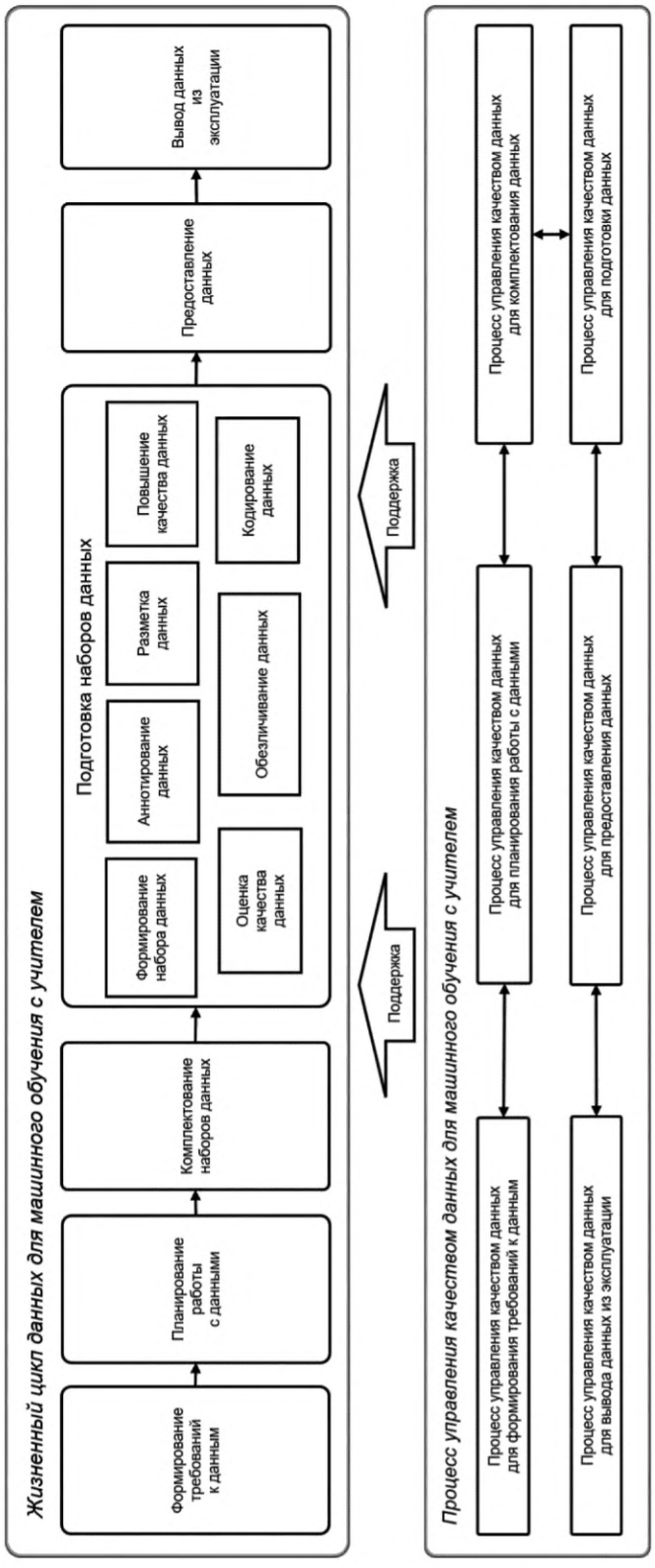

Рисунок 4 — Пример процесса управления качеством данных для МО с учителем

7.3 Планирование работы с данными

Планирование работы с данными основывается на процессах формирования требований к данным и может гарантировать наличие планов и ресурсов для успешного выполнения процесса управления качеством данных. Процесс планирования данных должен учитывать, по крайней мере, следующие элементы:

- модель данных или архитектуру данных, необходимые для удовлетворения требований к данным;
- план сбора необходимых данных в соответствии с предъявляемыми к ним требованиями;
- план комплектования достаточного количества данных, пригодных для решения задач аналитики или МО;
- роли, навыки и персонал, необходимые для реализации процесса управления качеством данных;
- ИТ и другие ресурсы, необходимые для реализации процесса управления качеством данных;
- время и бюджет, необходимые для реализации процесса управления качеством данных;
- план выполнения мероприятий по управления качеством данных в соответствии с моделью качества данных;
- план соблюдения требований законодательства;
- план соблюдения принципов процесса управления качеством данных;
- план выявления и устранения любых пропусков или недостатков, обнаруженных в полученных данных.

7.4 Комплектование данных

Данные, используемые для разработки модели МО, могут поступать из разных источников (например, из систем Интернета вещей: обработка транзакций, опросов, статичных изображений, видео-, звукозаписи, веб-форм, генераторов синтетических данных) с различными типами данных (например, числа, текст, двоичный код), в файлах с различными форматами данных (например, XML [*4*], JSON *по ГОСТ Р ИСО/МЭК 17826*, CSV [*5*], JPEG *по ГОСТ Р ИСО/МЭК 19794-5*, MPEG *по ГОСТ Р 54995*) и схемами. Организация уже может иметь данные, обозначенные в требованиях, но может и собирать новые данные. В некоторых случаях данные могут поступать из потоковых источников или источников, работающих в режиме, близком к реальному времени (например, ленты социальных сетей, поисковые системы), и могут использоваться для постоянного улучшения модели ИИ.

Процесс комплектования данных должен учитывать, по крайней мере, следующее:

- использование именно тех элементов, которые определены в процессе планирования работы с данными;
- соблюдение принципов процесса управления качеством данных;
- ключевые свойства данных, определенные процессом формирования требований к данным, такие как:
  - происхождение;
  - смещение;
  - надежность;
  - валидность;
  - типы данных;
  - схема;
  - формат;
- контекст использования данных при разработке модели МО, в том числе:
  - обучение;
  - валидация;
  - тестирование;
  - эксплуатация;
  - охват (например, демографические данные, поведение, географическое местоположение).

Для статичных изображений и видео ключевые свойства данных включают:

- разрешение;
- четкость;
- освещенность;

- цвет;
- фоновый шум.

После комплектования качество данных должно быть дополнительно оценено в соответствии с рекомендациями 7.5.8.

7.5 Подготовка наборов данных

7.5.1 Общие положения

Цель процесса подготовки данных — привести данные в состояние, при котором их можно будет успешно использовать для разработки модели МО, производительность которой соответствует требованиям организации.

В процессе подготовки данных должны учитываться как минимум следующие элементы:

- формирование набора данных;
- разметка данных;
- аннотация данных;
- оценка качества данных относительно целевых значений показателей качества данных, установленных в процессе формирования требований к данным;
- повышение качества данных:
  - очистка данных;
  - стандартизация данных;
  - нормализация данных;
  - заполнение данных;
- обезличивание данных;
- кодирование данных.

7.5.2 Машинное обучение с учителем

Машинное обучение с учителем может использовать все элементы, описанные в 7.5.1.

7.5.3 Машинное обучение без учителя

Машинное обучение без учителя не использует размеченные данные, но может использовать другие элементы, описанные в 7.5.1.

7.5.4 Машинное обучение с частичным привлечением учителя

Машинное обучение с частичным привлечением учителя, при котором в процессе обучения используются как размеченные, так и неразмеченные данные, может использовать все элементы, описанные в 7.5.1.

7.5.5 Формирование наборов данных

Стандарт [*6*] описывает комплектование наборов данных как процесс отбора и объединения данных в единый набор данных, который затем используется для обучения или оценки модели МО. Может возникнуть необходимость объединить, реорганизовать или дополнить полученные данные, чтобы создать набор данных, подходящий для конкретной задачи, приложения или подхода на основе машинного обучения. Примерами процессов формирования набора данных могут служить:

- объединение данных из нескольких источников;
- выбор из полученных данных конкретных объектов;
- разделение данных об объекте (например, на день, месяц и год);
- добавление синтетических данных;
- повторный выбор из полученных данных (например, сузить полосу пропускания с 16 кГц до 8 кГц, выбор случайным образом 12 студентов на курсе из 100 вместо выбора одного из тех, у кого день рождения в определенном месяце);
- случайный выбор: каждая выборка в наборе данных имеет равные шансы быть отобранной;
- стратифицированный выбор: данные делятся на подгруппы на основе интересующих признаков, например пол и возраст. Такой подход используется для обеспечения репрезентативного представления каждой подгруппы.

Наборы данных могут иметь разные форматы, определяемые поставщиками данных. Для обеспечения согласованной обработки наборов данных разных форматов они могут быть преобразованы, сериализованы с помощью определенного инструмента машинного обучения и сохранены в специальной форме, где метаданные, выборки данных и их разметка сделаны таким образом, что это позволяет улучшить качество данных при использовании.

Похожие компоненты и их семантика затем могут быть обобщены и типизированы. Такой подход может облегчить повторное использование, обмен, хранение данных, доступ к ним и сравнение наборов данных. Общие компоненты набора данных в табличной форме включают:

- индекс: компонент, который можно использовать для облегчения просмотра набора данных и доступа к нему. Он содержит информацию, относящуюся к каталогам, именам и файлам меток;
- заголовок: компонент, который можно использовать для записи информации об организации данных внутри набора. Он содержит информацию, относящуюся к объему и расположению скалярных и блочных данных, а также информацию, относящуюся к их соответствиям, статистике, разделам, типам данных и измерениям;
- страница: страница относится к определенному сегменту внутри файла данных, в котором хранятся фактические скалярные данные (например, данные целочисленного, строкового, плавающего и других типов) или блочные данные (такие, как изображение, видео и аудио) выборок и меток.

Качество подготовленного набора данных следует оценивать в соответствии с 7.5.8.

7.5.6 Разметка данных

См. раздел 8, содержащий рекомендации по разметке данных.

7.5.7 Аннотирование данных

Данные могут быть аннотированы с помощью метаданных, которые предоставляют описательную информацию о наборе данных. Метаданные могут использоваться заинтересованными сторонами для каталогизации данных, поисковых и рекомендательных инструментов, отслеживания данных и обмена данными. Метаданные могут создаваться обладателями данных и пользователями данных в ходе их деятельности. Прежде чем выбирать и использовать набор данных, заинтересованные стороны могут использовать метаданные, чтобы убедиться, что набор данных соответствует их требованиям.

Примеры метаданных включают:

- метаданные, связанные с комплектованием наборов данных (такие, как источники данных; персонал, занимающийся аннотированием данных; дата и время операций обработки данных или передачи прав владения при совместном использовании данных);
- метаданные, связанные с контентом (например, сферы деятельности и технические области, форматы данных, объем данных, количество категорий данных, примеры выборок, атрибуты данных, статистическая информация, связанная с распределением данных, ограничивающие прямоугольники, сегментации, ключевые точки и файлы);
- метаданные, связанные с качеством (например, результаты измерения качества данных).

Примеры

1 Для обучения модели МО распознаванию транспортных средств пользователь находит подходящие наборы данных на основе нескольких доступных примеров изображений. Обычной ситуацией является то, что у пользователя есть только несколько изображений транспортных средств, сфотографированных в реальных условиях, и ему необходимо детально проверить каждый потенциальный набор данных в условиях ограниченных ресурсов. В этом случае метаданные можно использовать для выбора набора данных для обучения модели МО в сочетании со сравнением на основе подобия признаков. Сравнение метаданных (таких, как сфера деятельности, формат данных, атрибуты данных) может помочь пользователю быстро определить наиболее подходящие наборы данных.

Важно, что метаданные можно использовать как обобщение содержимого каждой выборки в наборе данных. В этом случае метаданные могут быть представлены в определенной форме, например в виде схемы для описания метаинформации файла выборки и соответствующих меток.

2 Схема набора данных для классификации изображений может включать:

- ***поле «имя файла» строкового типа;***
- ***поле «метка» целочисленного типа;***
- ***поле «данные» типа байт;***
- ***поле «метка времени» типа datetime.***

Для подробного описания могут быть использованы метаданные, содержащие описание поля. Например, такое семантическое описание может включать:

- имя поля;
- тип данных поля (например, float32);
- размерность поля.

7.5.8 Оценка качества данных

Цель оценки качества данных — определить, соответствуют ли данные или набор данных требованиям к качеству данных. Может возникнуть необходимость повторять оценку качества данных каждый раз, когда происходит преобразование данных или смещение данных.

Оценка качества данных должна включать как минимум следующее:

- статистические характеристики полученных данных, включая оценку их влияния на требования к данным;
- использование характеристик качества данных и показателей качества данных в соответствии с целями, установленными в процессе формирования требований к данным (характеристики качества данных и показатели качества данных описаны в *ГОСТ Р 71484.2*);
- документирование процесса и результатов оценки.

Если набор данных не соответствует требованиям к данным, организации следует рассмотреть следующие возможные действия:

- улучшить набор данных;
- прекратить использование набора данных;
- получить новый набор данных.

7.5.9 Повышение качества данных

7.5.9.1 Основные положения

Результаты оценок качества данных часто показывают, что данные не соответствуют требованиям к качеству данных. Во многих случаях набор данных может быть улучшен. Процессы, которые используют для улучшения данных, зависят как от самих данных, так и от обнаруженных недостатков. Примеры процессов улучшения данных включают:

- очистку;
- фильтрацию;
- нормализацию;
- приведение к стандартному виду;
- масштабирование;
- заполнение;
- аугментацию;
- кодирование.

7.5.9.2 Очистка данных

Очистка данных включает в себя исправление или удаление неполных, неправильных или нерелевантных данных. Очистка данных может выполняться на уровне признака, записи или элемента в зависимости от обстоятельств.

Методы очистки данных могут включать в себя:

- удаление повторяющихся записей (например, повторяющихся записей в результате объединения наборов данных);
- удаление или исправление записей данных с неправильными элементами данных (например, исправление форматов дат, удаление записей с неправильными элементами данных из набора данных);
- удаление или заполнение отсутствующих, или нулевых элементов данных (например, заполнение данных для пустых или нулевых элементов данных; см. «Заполнение данных» в 7.5.9.3).

7.5.9.3 Нормализация, приведение к стандартному виду и заполнение

Нормализация данных. Наборы данных, которые имеют очень большие диапазоны или смесь малых и больших диапазонов значений, могут привести к тому, что алгоритмы МО будут рассчитывать большие веса или переоценивать признаки, связанные с большими диапазонами значений.

Для установления общего диапазона для всего набора данных можно использовать нормализацию путем масштабирования значений до единичной нормы, т. е. от 0 до 1.

Приведение данных к стандартному виду. Для достижения заданной производительности некоторых алгоритмов МО данные должны иметь определенное распределение. Для получения требуемого распределения можно использовать стандартные преобразования данных.

Примеры стандартных преобразований включают следующее [*7*]:

- Standard Scalar: преобразование, которое создает набор данных со средним значением, равным нулю, и стандартным отклонением, равным единице. Использование стандартного скаляра обычно не рекомендуется для разреженных данных;

- MinMax Scaler: преобразование, которое масштабирует каждое значение так, чтобы оно находилось в заданном диапазоне от 0 до 1;

- MaxAbs Scalar: преобразование, которое масштабирует каждое значение в диапазоне от 1 до −1. Это преобразование можно использовать для разреженных данных;

- Robust Scalar: преобразование, которое масштабирует каждое значение так, чтобы оно находилось в пределах межквартильного диапазона, т. е. между первым и третьим квартилями. Это преобразование можно использовать для наборов данных с большим количеством выбросов.

Заполнение данных. Некоторые алгоритмы МО предполагают, что все данные будут полными и имеющими смысл. Однако в наборах данных часто отсутствуют значения, которые могут быть представлены пробелами, а не числами (not-a-number, NaN) или какими-либо другими значениями.

Для структурированных данных (например, табличных данных со значениями, разделенными запятыми) одним из методов борьбы с отсутствующими значениями является удаление всей строки или столбца с отсутствующим значением. Это может привести к удалению важных признаков или созданию разреженного набора данных, которого недостаточно для обучения, валидации и тестирования модели МО.

Другой подход заключается в использовании статистических методов для вычисления соответствующих значений. Основные статистические методы заполнения включают использование среднего, медианного, наиболее часто встречающегося или заданного постоянного значения, т. е. фиксированного значения.

Преобразования для статистического подхода к заполнению данных включают:

- Simple Imputer: преобразование, которое можно запрограммировать для замены отсутствующих значений средним, медианным, наиболее частым или постоянным значением. Среднее значение и медиану можно использовать только для числовых значений, тогда как наиболее часто встречающиеся значения и константы можно использовать как для строк, так и для числовых значений;

- Iterative Imputer: многовариантное преобразование, которое осуществляет замены отсутствующих значений путем итеративного анализа значений всех признаков в наборе данных.

7.5.9.4 Аугментация данных

В некоторых случаях набор данных, используемый для МО, не может адекватно представлять генеральную совокупность данных, используемую для анализа и прогнозирования; типичным примером является выборка с недостаточным или избыточным количеством определенных объектов и, соответственно, несбалансированными данными. Это может быть результатом исторической ситуации, обусловленной усилиями по защите персональных данных или небольшим количеством встречающихся в природе объектов с заданной характеристикой.

Неспособность набора данных представлять генеральную совокупность при использовании МО может привести к тому, что модель МО не сможет хорошо обобщать при использовании эксплуатационных данных или получит неоднородную точность классификации для разных классов данных [*8*], что приведет к соответствующим социальным, правовым и этическим проблемам, если объектами классификаций являются люди [*9*].

Аугментация данных увеличивает количество или разнообразие данных за счет применения к исходным данным операций модификации, добавления, преобразования, заполнения или их объединения.

Типичные методы аугментации текстовых данных включают:

- замену сущностей (например, замена слов в предложении словами из тех же категорий);

- обратный перевод (например, предложение переводится на другой язык, а затем это новое предложение переводится обратно на исходный язык);

- замену синонимов (например, замена слов в предложении синонимами);

- случайную вставку (например, поиск случайного синонима случайного слова в предложении и вставка этого синонима в случайную позицию в предложении);

- случайную перестановку (например, случайный выбор двух слов в предложении и перестановка их местами);

- случайное удаление (например, случайное удаление произвольного слова в предложении);

- нарушение порядка следования предложений (например, изменение порядка предложений в абзаце для создания нового абзаца);

- создание предложений с использованием генеративной модели;

- изменение дерева зависимостей;

- искажение и восстановление.

Типичные методы аугментации наборов изображений включают:

- масштабирование: изображение увеличивается или уменьшается в размерах для создания нового изображения;
- обрезку: выбирается часть изображения, обрезается, а затем изменяется до исходного размера изображения;
- отражение: изображение зеркально отражается по горизонтали, вертикали или по обоим направлениям;
- вращение: изображение поворачивается на угол от 0° до 360°;
- трансфокацию;
- внесение шума.

Методы аугментации для наборов голосовых данных могут включать:

- добавление шума;
- изменение скорости речи;
- реверберацию.

7.5.10 Обезличивание данных

Полученные данные могут содержать персональные данные, потенциально угрожающие нарушению конфиденциальности субъектов данных.

Примерами персональных данных могут служить:

- имя;
- адрес;
- адрес сетевого подключения;
- местоположение;
- биометрическая информация;
- демографическая информация;
- уникальный идентификатор, включая номера паспортов или данные кредитной карты.

Обучающие, валидационные и тестовые данные для проекта МО, в которых содержатся персональные данные, должны быть обезличены в максимально возможной степени. Если эксплуатационные данные используются для того, чтобы сделать выводы об отдельных людях, может возникнуть необходимость использовать персональные данные.

Пример — Организация розничной торговли хочет открыть три новых магазина в мегаполисе в зависимости от местоположения своих нынешних клиентов. Организация имеет базу данных о нескольких тысячах клиентов и планирует использовать алгоритм кластеризации для определения лучших мест для магазинов. Для этого организация выполняет запрос к базе данных клиентов по почтовым индексам и не использует никакие другие сведения.

Обезличивание — это процесс удаления, изменения или ограничения доступа к персональным данным таким образом, что они больше не могут быть связаны с одним или несколькими людьми. Примеры методов обезличивания могут включать:

- анонимизацию;
- псевдонимизацию;
- удаление связей;
- агрегирование;
- дифференциальную приватность.

Дополнительные сведения об обезличивании данных см. в [*10*].

7.5.11 Кодирование данных

В алгоритмах МО обычно предполагается, что данные должны быть в числовой форме, в то же время категориальные данные часто представлены текстовыми строками. Категориальные данные можно преобразовать в числовые данные с помощью кодирования.

Преобразования для кодирования категориальных данных включают [*11*]; [*12*]:

- OrdinalEncoder: преобразование набора категориальных значений в порядковый набор целых чисел, т. е. от 1 до *n*. Некоторые алгоритмы машинного обучения при вычислении весов в модели МО могут придавать слишком большое значение порядковому номеру результата;
- LabelEncoder: преобразование нечисловых меток для целевой переменной в числовые категории;
- OneHotEncoding: преобразование набора значений категорий в массив с категориальными метками, представленными в виде признаков и двоичных значений, которые указывают, соответствует ли запись заданной характеристике;

- Get Dummies: преобразование категориальных значений в массив функций и двоичных индикаторов. В этом случае функции называются фиктивными переменными.

7.6 Предоставление данных

7.6.1 Общие положения

Как описано в *ГОСТ Р 71484.1*, предоставление данных обеспечивает возможность использования подготовленных данных для машинного обучения. Перед предоставлением данных организация должна убедиться, что они соответствуют всем заданным требованиям. Предоставление данных может включать в себя передачу или перемещение данных из одной системы в другую. Организация должна обеспечить сохранение качества данных, если такая передача или перемещение необходимы.

7.6.2 Машинное обучение с учителем

При обучении с учителем обучающие данные используются алгоритмом МО для создания модели МО. Затем модель МО применяется к эксплуатационным данным для получения соответствующих логических выводов.

7.6.3 Машинное обучение без учителя

Несмотря на то, что некоторые методы машинного обучения без учителя используют обучение (например, при кластеризации K-средних), модель, как правило, создается на основе эксплуатационных данных, а затем с ее помощью делаются логические выводы об эксплуатационных данных (например, определяется принадлежность записи данных к центроиду).

7.6.4 Машинное обучение с частичным привлечением учителя

Машинное обучение, которое представляет собой гибрид машинного обучения без учителя и с учителем и использует неразмеченные обучающие данные в дополнение к размеченным обучающим данным.

7.7 Вывод данных из эксплуатации

Набор данных может быть выведен из эксплуатации, когда в нем больше нет необходимости. Вывод данных из эксплуатации может включать сохранение данных для повторного использования в будущем, возврат данных их распорядителю или уничтожение данных.

Если есть намерение повторно использовать набор данных в будущем, организация должна обеспечить сохранение качества набора данных, включая его безопасность и конфиденциальность любых субъектов данных. Кроме того, необходимо создать и проверить полную резервную копию набора данных.

План вывода данных из эксплуатации должен быть составлен для определения участвующего в нем персонала с распределением ролей и обязанностей и должен быть рассмотрен соответствующими заинтересованными сторонами.

8 Методы и процесс разметки данных

8.1 Общие положения

Размеченные данные используются в качестве обучающих данных для задач машинного обучения с учителем или с частичным привлечением учителя, таких как классификация и регрессия. *ГОСТ Р 71476* определяет метку как значение целевой переменной, присвоенное неделимому элементу размеченных входных данных, а разметку данных — как процесс присоединения к данным описательной информации без внесения каких-либо изменений в сами данные. Другие формы разметки данных включают разметку частей изображения или фрагментов речи.

В некоторых случаях обучающие данные не содержат метки для целевых переменных (например, изображения, образцы аудио), тогда как в других случаях метки включаются или могут быть определены на основе полученных данных (например, годовые продажи на основе данных о транзакциях, история неплатежей по кредитам на основе данных о финансовых операциях).

8.2 Принципы разметки данных

Должна быть разработана схема обучения разметке данных, чтобы аннотаторы-люди и автоматические системы разметки могли эффективно научиться применять ее с приемлемой точностью.

Понятия и терминология, используемые в схеме обучения разметке данных, должны использоваться постоянно.

Метки должны иметь четко определенную семантику, чтобы разметка была понятна как аннотаторам-людям, так и автоматическим системам разметки.

В процессе разметки данных должны быть предприняты шаги для минимизации, насколько это возможно, любых фактических или предполагаемых предубеждений в отношении отдельных лиц или групп лиц.

8.3 Методы разметки данных

Методы разметки данных зависят от типа данных и могут включать разметку объектов, ограничивающий прямоугольник, разметку ключевых точек, сегментацию экземпляров, семантическую сегментацию и разметку последовательностей. К объектам разметки данных относятся следующие:

- разметка изображений: широко применяется в настоящее время. Основные методы разметки включают разметку точек, разметку рамок, разметку областей, трехмерную разметку и классификационную разметку. Существует также множество сценариев использования, например, для обеспечения безопасности, в образовании или при автоматическом вождении;
- разметка голосовых данных: применяется для распознавания голоса, распознавания голосовых отпечатков и синтеза речи. Данные могут содержать разметку, соответствующую ролям говорящих, характеристикам окружающей среды, а также многоязычную разметку, просодическую разметку, системную разметку, разметку эмоций, разметку шума и т. д.;
- разметка текста: важный аспект обработки естественного языка. Для обеспечения высокой точности прогнозирования текстовые данные могут иметь разметку сегментации предложений, разметку семантических суждений, разметку перевода текста, разметку эмоций, разметку полифонических слов, разметку цифровых символов и т. д.

Несколько меток данных с разным содержимым могут находиться на разных уровнях разметки, что объясняется автономным форматом аннотаций. Слои разметки над объектом могут сосуществовать как единообразный способ перекрестных ссылок между слоями. Это позволяет не только использовать несовместимые друг с другом слои, но и использовать альтернативные метки, в которых используются разные схемы разметки для одного и того же объекта.

Примеры

1 При разметке речевых данных можно использовать один слой для аннотирования событий, времени и пространства, а также информации о говорящем; второй слой можно использовать для аннотирования содержимого речевых сигналов; третий слой можно использовать для аннотирования недопустимых речевых сигналов.

2 При разметке видеоданных:

- ***персонажи и объекты в видео выделяются через ключевой кадр;***
- ***предусмотрены метки для поведения, типа объекта, пола и т. д.;***
- ***указывается время начала и окончания.***

Данные могут быть размечены вручную аннотатором-человеком или автоматически с использованием псевдоразметки. Псевдоразметка — это процесс использования обученной на размеченных данных модели МО для предсказания меток в неразмеченных данных. Это простой и быстрый способ разметки данных, а псевдоразмеченные данные можно использовать в качестве справочной информации для аннотатора-человека, что может помочь повысить эффективность решения задачи разметки.

8.4 Процесс разметки данных

8.4.1 Общие положения

На рисунке 5 показан типичный процесс разметки данных, который содержит стадии подготовки, выполнения и вывода. Для обеспечения качества разметки данных описание каждой стадии приведено в 8.4.2—8.4.9.



Рисунок 5 — Процесс разметки данных

8.4.2 Спецификации разметки

При внедрении разметки данных подробная спецификация разметки данных должна включать четкие инструкции, примеры и примечания к разметке.

1 Инструкции используются для пояснения определений меток и указания компонентов меток, типов меток и всех операций, используемых инструментами или платформами для разметки.

2 Должны быть приведены примеры для иллюстрации, как правильно выполнять разметку для нестандартных или сложных случаев.

3 В примечаниях к маркировке указываются ошибки, которых следует избегать, особенности методов разметки и требуемые дополнительные методы обработки. Ошибки можно разделить на несколько типов, например:

- ложноотрицательная разметка (например, объекты, которые должны быть маркированы в процессе проверки, не найдены);
- ложноположительная разметка (например, маркируются объекты, которые следует исключить в процессе проверки);
- выход за границы диапазона (например, метки, которые не соответствуют требованиям или выходят за границы разметки).

Кроме того, спецификации разметки должны включать определение соответствующих ролей и их обязанностей, ожидаемого времени разметки данных, требований к точности, а также требований к безопасности и конфиденциальности данных.

8.4.3 Роли участников при разметке

К процессу разметки данных относятся следующие роли:

- аннотатор-человек выполняет фактическую разметку данных;
- контролер проверяет все метки или выборку меток для каждой порции данных, размеченных аннотатором-человеком;
- менеджер — ответственное лицо, выполняющее роль распределителя работ по разметке, руководителя группы контролеров по качеству и руководителя аннотаторов и контролеров.

8.4.4 Инструменты или платформы для разметки

Выбор инструментов и платформ для разметки может осуществляться до обучения модели МО и может включать использование облачных сервисов. Инструменты и платформы для разметки должны соответствовать всем требованиям к данным. Инструменты и платформы для разметки должны поддерживать основные функции, в том числе:

- обеспечение требуемой производительности платформ и инструментов, а также сокращение ошибок, совершаемых аннотаторами-людьми;
- управление записями и отслеживаемость каждого аннотатора-человека;
- создание соответствующих руководств пользователя и руководств по эксплуатации.

8.4.5 Постановка задачи разметки

Для различных задач машинного обучения и требований к обучению модели МО задача разметки данных может включать классификацию, распознавание и сегментацию. Прежде чем приступить к разметке данных, необходимо четко определить основные задачи разметки данных, включая описание сценариев, модели МО и ее использования, а также информацию о данных.

Необходимо определить цель задачи маркировки данных и руководствоваться ею для обеспечения эффективности результирующей модели МО.

8.4.6 Распределение задач по разметке

Задачи по разметке данных следует распределить между разными командами в соответствии с разными организационными подходами. Как правило, выбор организационного подхода к разметке данных зависит от сложности и размера обучающих данных, необходимых знаний в предметной области и степени понимания задач машинного обучения командами, назначенными для разметки данных. Для разных задач разметки данных можно использовать разные организационные подходы, как показано в таблице 1.

Таблица 1 — Применимость различных организационных подходов

Организационные подходы	Задачи разметки данных
Внутренняя разметка	Задачи, требующие значительного знания контекста задачи МО, алгоритмов и данных; задачи, в которых используются высокочувствительные конфиденциальные данные (например, финансы, здравоохранение, национальная безопасность); задачи повышенного риска; задачи, требующие быстрого выполнения; задачи, которые могут быть решены за счет имеющихся внутренних ресурсов
Аутсорсинговая разметка	Задачи, в которых используются большие наборы обучающих данных, требующие ресурсов, превосходящих те, которые доступны для внутренней разметки; задачи, где данные не являются высокочувствительными; задачи с умеренным и низким риском
Краудсорсинговая разметка	Задачи, в которых допустимо публичное использование набора данных; задачи, в которых большое разнообразие аннотаторов-людей полезно для проекта; задачи, которые решают организации с ограниченным бюджетом развития; задачи, требующие большого количества аннотаторов-людей для соблюдения ограничений по времени; задачи с низким уровнем риска

8.4.7 Контроль процесса разметки

Контроль качества процесса разметки данных можно выполнить в два этапа: при первоначальной разметке данных и при последующей проверке качества данных.

На начальном этапе разметки данных специалисты должны проаннотировать данные в строгом соответствии с инструкциями по разметке и представить отчет о качестве результатов разметки после самооценки.

На этапе проверки качества данных контролеры выборочно проверяют образцы размеченных данных, чтобы убедиться, что они соответствуют требованиям. Если выборка не соответствует требованиям, данные могут быть возвращены на доработку.

8.4.8 Проверка качества результатов разметки

Цель проверки качества разметки — убедиться в том, что результаты разметки данных соответствуют требованиям. Следует учитывать такие характеристики качества данных, как аккуратность, точность, полнота и согласованность. Проверка качества результатов разметки данных может включать:

- сплошную проверку;
- выборочную проверку результатов;
- экспертную оценку и оценку со стороны менеджера;
- проверку несколькими контролерами с последующей экспертной проверкой и оценкой менеджера.

Выбор способа проверки меток данных должен определяться в соответствии с используемым организационным подходом и рисками, связанными с задачей МО.

Внутренняя разметка: в данном случае организация в максимальной степени контролирует процесс разметки и квалификацию аннотаторов. Задачи МО с высоким уровнем риска могут требовать тщательной проверки результатов разметки, тогда как проверка результатов разметки для задач МО с низким уровнем риска может включать проверку от 5 % до 10 % размеченных данных.

Аутсорсинговая разметка: в данном случае организация не может напрямую контролировать процесс разметки или квалификацию аннотаторов, но может устанавливать требования к исполнителям.

Если разметка данных передается третьим лицам, то организации следует применять более строгий процесс проверки результатов разметки даже для проектов с низким уровнем риска.

Краудсорсинговая разметка: в данном случае аннотаторами могут быть волонтеры, квалификация которых неизвестна организации. Если краудсорсинговая разметка используется для задач с низким уровнем риска, то организации следует рассмотреть возможность строгой проверки результатов разметки данных.

Подходы к формированию выборки для проверки результатов разметки могут включать:

- случайную выборку: проверяется случайно отобранная часть размеченных данных;
- стратифицированную выборку: проверяется случайно отобранная часть размеченных данных от каждого аннотатора;
- сплошная проверка: проверяется весь размеченный набор данных.

Примечание — Информацию о выборе соответствующей части образцов для проверки см. в *ГОСТ Р ИСО 2859-1*.

Независимо от того, какой подход к проверке результатов разметки данных используется, стандарты проверки должны быть согласованными и соответствовать спецификациям и требованиям разметки.

8.4.9 Пересмотр результатов разметки

Размеченные данные, предоставленные аннотаторами, могут быть пересмотрены из-за проблем с качеством. Существует множество методов обнаружения экземпляров с неправильной разметкой в наборе данных. Общий метод называется перекрестной проверкой.

Первый шаг — разделить обучающие данные на n частей. Для каждой из n частей модель МО обучается на остальных $n–1$ частях. Затем эта обученная модель используется для прогнозирования меток исключенных частей. Если эта обученная модель предсказывает экземпляр как принадлежащий к другой метке, отличной от метки обучающих данных, то этот экземпляр с высокой вероятностью будет неправильно размечен аннотаторами. Затем тот же процесс выполняется еще $n–1$ раз, при этом $n–1$ модели машинного обучения обучаются на других исключенных частях, и каждая модель используется для прогнозирования меток на исключенных частях. Затем можно проверить все экземпляры в размеченных обучающих данных и отфильтровать экземпляры с несовпадающими метками. Эти экземпляры можно вернуть аннотаторам для проверки и доработки.

Для управления изменениями в данных, при пересмотре размеченных данных следует сохранять как исходные, так и исправленные данные. Кроме того, должны быть записаны временная метка и имя специалиста, внесшего изменения при каждом пересмотре.

9 Роли участников

9.1 Общие положения

Согласно модели жизненного цикла данных участники процесса управления качеством данных для МО могут выполнять разные роли, включая роли планировщика данных, создателя данных, сборщика данных, специалиста по подготовке данных, распорядителя данных и пользователя данных. Роли, описанные в настоящем стандарте, не исключают друг друга, т. е. одна сторона может играть несколько ролей в данном процессе. Например, создатель данных также может быть распорядителем данных.

9.2 Планировщик данных

Планировщик данных — сторона, ответственная за процесс планирования данных (см. в 7.3).

9.3 Создатель данных

Создатель данных — сторона, которая формирует данные и может иметь на них определенные права (например, доступ, использование, изменение, совместное использование). Создателем данных может быть физическое лицо. Создатель данных может напрямую передавать данные пользователю данных или промежуточной стороне, например распорядителю данных.

9.4 Сборщик данных

Сборщик данных — сторона, которая комплектует наборы данных для выполнения задач МО или аналитики в соответствии с требованиями к данным. Сборщик данных может быть внутренним или внешним. Сборщик данных может отвечать за предоставление данных в определенной форме.

9.5 Инженер данных

Инженер данных — сторона, которая отвечает за подготовку полученных данных, чтобы удовлетворить требованиям задачи МО или аналитики. Инженеры данных могут выполнять разметку данных, использовать показатели качества данных и применять процессы улучшения качества данных. Инженеры данных обычно используют инструменты и наборы инструментов для выполнения своих обязанностей.

9.6 Распорядитель данных

Распорядитель данных — сторона, имеющая юридические полномочия разрешать другим сторонам обрабатывать данные. Распорядитель данных может использовать соглашения о совместном использовании данных, как описано в [*2*], для доведения требований до пользователей данных. Распорядитель данных может выполнять другие роли, например: создатель данных, сборщик данных или инженер данных.

9.7 Пользователь данных

Пользователь данных — сторона, уполномоченная использовать данные под юридическим контролем распорядителя данных. Пользователи данных могут выполнять другие роли, например планировщик данных или инженер данных. В некоторых случаях роли создателя данных, распорядителя данных и пользователя данных выполняются в одной организации.

10 Процесс управления качеством данных для машинного обучения с частичным привлечением учителя

10.1 Общие положения

Обучающие данные, используемые для машинного обучения с учителем, требуют разметки, что может оказаться дорогостоящим процессом при работе с большими объемами обучающих данных. Недостатком обучения с частичным привлечением учителя является ограниченность спектра его применения. При обучении с частичным привлечением учителя необходимые обучающие данные включают данные для машинного обучения как без учителя, так и с учителем. Как правило, они содержат небольшое количество размеченных данных и большое количество неразмеченных данных. Размеченные данные можно использовать для разметки неразмеченных данных.

10.2 Формирование требований к данным

Процесс формирования требований к данным для машинного обучения с частичным привлечением учителя должен включать определение характеристик и пропорций размеченных и неразмеченных данных. В дополнение к рекомендациям из 7.2 важны следующие аспекты:

- качество размеченных данных;
- сбалансированность меток данных между различными категориями;
- распределение неразмеченных данных.

10.3 Планирование работы с данными

Планирование работы с данными для машинного обучения с частичным привлечением учителя должно охватывать как размеченные, так и неразмеченные данные. Для машинного обучения с частичным привлечением учителя процесс планирования описан в 7.3. Кроме того, важны следующие аспекты:

- планирование источников комплектования наборов данных;
- планирование проверки качества размеченных данных;
- планирование пропорции размеченных и неразмеченных данных.

10.4 Комплектование данных

Для машинного обучения с частичным привлечением учителя процесс комплектования данных описан в 7.4.

10.5 Подготовка данных

Для машинного обучения с частичным привлечением учителя процесс подготовки данных включает формирование набора данных и аннотирование данных, как описано в 7.5. Очистка данных, обезличивание данных и выборка данных для обучения с частичным привлечением учителя зависят от того, какие данные обрабатываются: размеченные или неразмеченные.

Кроме того, для неразмеченных данных можно вычислить сходство выборки неразмеченных данных с размеченными данными, чтобы удалить нерелевантные неразмеченные данные.

Когда выборка данных используется для обучения с частичным привлечением учителя, обычно имеется меньше размеченных данных и больше неразмеченных данных. Важно контролировать количество используемых неразмеченных данных. Следовательно, если количество неразмеченных данных слишком велико, может потребоваться отобрать неразмеченные данные и сформировать из них более репрезентативную выборку.

10.6 Предоставление данных

Для обучения с частичным привлечением учителя применяется процесс предоставления данных для размеченных и неразмеченных данных, описанный в 7.6.

10.7 Вывод данных из эксплуатации

Процесс вывода из эксплуатации данных, используемых для обучения с частичным привлечением учителя, описан в 7.7.

11 Процесс управления качеством данных для обучения с подкреплением

11.1 Общие положения

Для обучения с подкреплением данные генерируются в процессе обучения. В этом разделе используется пример беспилотного автомобиля, чтобы проиллюстрировать процесс управления качеством данных для обучения с подкреплением.

11.2 Формирование требований к данным

Процесс формирования требований к данным для обучения с подкреплением должен включать следующее:

- объект обучения с подкреплением;
- необходимое количество генерируемых данных;
- правила генерации данных.

11.3 Планирование работы с данными

В дополнение к рекомендациям из 7.3 важно спланировать создание окружения для генерации данных, которое охватывает правила, характеристики агента и окружения, функцию вознаграждения для обучения с подкреплением.

При проведении обучения с подкреплением процесс планирования данных отсутствует, поскольку данные генерируются в ходе обучения.

11.4 Комплектование данных

Задачи обучения с подкреплением, как правило, не требуют предварительного комплектования данных, они генерируются во время обучения динамически посредством непрерывного взаимодействия между агентом и окружением. Генерируемые данные включают в себя изменение состояния окружения, множество предпринятых агентом действий и соответствующие значения вознаграждений. Сгенерированные данные обычно не требуют дополнительной ручной разметки.

11.5 Подготовка данных

11.5.1 Общий процесс

Процесс подготовки данных для обучения с подкреплением состоит из четырех этапов, которые должны быть специально разработаны для различных задач обучения с подкреплением.

a) Проектирование действий агента. Должен быть спроектирован набор действий, предпринимаемых при взаимодействии агента и окружения. Например, набор действий для беспилотного автомобиля включает ускорение, замедление, торможение, ручное управление и другие варианты контроля поведения транспортного средства в поперечном или продольном направлении.

b) Определение формата (входных) данных о состоянии агента. Для разных задач входные данные агента различаются. Например, для беспилотного автомобиля входными данными являются изменяющиеся во времени сигналы, получаемые датчиками транспортного средства, включая изображение с камеры и сигналы лидара.

c) Определение формы комплектования данных. Данные для обучения с подкреплением комплектуются последовательно, поэтому необходимо определить количество эпизодов и длину каждой последовательности. Например, для беспилотного автомобиля необходимо определить время сбора данных.

d) Оценка и определение вознаграждения. После того как агент совершит ряд действий, он определяет вознаграждение, которое будет получено за этот набор действий. Например, для беспилотного автомобиля агент получает вознаграждение в зависимости от того, благополучно ли автомобиль прибудет, от количества затраченного времени, от оптимальности выбранного маршрута и от того, произошла ли авария.

11.5.2 Запись данных

Основываясь на четырех этапах подготовки данных из 11.5.1, агент может непрерывно взаимодействовать со средой в процессе обучения с подкреплением, при этом записывается весь процесс взаимодействия, включая:

- действия, совершаемые агентом;
- обратную связь, полученную по каждому действию; например, в сценарии беспилотного автомобиля обратной связью является изменение дорожной обстановки, воспринимаемое датчиками;
- вознаграждение за каждый шаг, поскольку общее вознаграждение агента от окружения сильно варьируется для разных типов задач.

Данные, записанные в процессе взаимодействия, затем сохраняются в виде последовательности для дальнейшего обучения модели, причем каждый процесс взаимодействия в последовательности содержит все три этих элемента.

11.6 Предоставление данных

Предоставление данных в процессе обучения с подкреплением — это предоставление данных в реальном времени, генерируемых в результате взаимодействия между агентом и окружением. Для выполнения некоторых конкретных задач агент может предпринять непреднамеренные или необоснованные действия, которые приносят высокую награду, но приводят к нарушениям безопасности. В этом случае необходимо вмешательство человека для корректировки поведения агента, например путем тонкой настройки функции вознаграждения для предотвращения подобных случаев.

В процессе взаимодействия с окружением агент постоянно исследует новые действия, чтобы найти наилучшую стратегию принятия решений для этих действий. Таким образом, агент случайным образом исследует некоторые действия, а не использует изученную стратегию. Кроме того, при непрерывной итеративной эволюции агента данные о действиях и стратегиях принятия решений агента должны регулярно обновляться.

11.7 Вывод данных из эксплуатации

Процесс вывода данных из эксплуатации для данных, используемых для обучения с подкреплением, описан в 7.7.

12 Процесс управления качеством данных для аналитики

12.1 Общие положения

Для аналитики процесс управления качеством данных может использоваться для управления и поддержки качества традиционных подходов к анализу данных, для обнаружения полезной информации, обоснования выводов и поддержки принятия решений.

12.2 Формирование требований к данным

В дополнение к рекомендациям из 7.2 важны следующие аспекты:
- источники данных и сценарии использования для анализа;
- методы обезличивания данных.

12.3 Планирование работы с данными

При проведении анализа необходимо следовать процессу планирования работы с данными, описанному в 7.3. Кроме того, важно запланировать удаление персональных данных.

12.4 Комплектование данных

12.4.1 Общие положения

Как правило, перед стадией подготовки данных необходимо определить цели анализа данных, чтобы определить источник комплектования данных и атрибуты, которые необходимо документировать. Комплектование данных может включать загрузку и хранение данных. Данные для отдельных действий варианта использования должны быть собраны, сам процесс сбора должен быть стандартизирован, а данные должны храниться надлежащим образом для последующего анализа.

12.4.2 Загрузка данных

Первым шагом является определение целей анализа данных. Сбор данных в соответствии с целями анализа данных является основой для управления качеством используемых данных. Во-вторых, определение методов комплектования данных, таких как сбор скрытых сведений через интернет или сбор данных с помощью таких аппаратных средств, как камеры и микрофоны, также является основой для обеспечения качества собранных данных. Наконец, методы комплектования данных должны быть стандартизированы, что может уменьшить различия между данными из разных пакетов и улучшить возможность их повторного использования.

Для разных сценариев аналитики загружаемые типы данных и объемы данных могут быть совершенно разными. Поэтому в случае подготовки мультимодальных данных необходимо заранее учитывать, например, какие типы данных могут быть востребованы, может ли точность графических данных в различных форматах повлиять на их последующую обработку, сохранять ли аудиоданные в виде фреймов или фонем.

12.4.3 Хранение данных

Способы хранения данных могут быть разработаны в зависимости от объемов данных, типов данных и требований к их обработке при анализе. Например, когда объем данных очень велик, полученные данные можно хранить в распределенной файловой системе. К тому же можно выбрать различные типы баз данных для хранения, такие как реляционные и нереляционные (например, графовую базу данных).

12.5 Подготовка данных

12.5.1 Общие положения

Данные, полученные по разным сценариям, могут содержать неверные или некорректные данные, например, из-за нестандартных устройств сбора, физического повреждения устройства хранения или ограничений реальных сценариев. Поэтому полученные данные следует очистить для повышения качества данных. Кроме того, для различных задач необходимо использовать такие преобразования данных, чтобы удовлетворить требованиям моделей анализа данных. Для упрощения анализа полученные данные следует нормализовать посредством очистки, преобразования и агрегирования.

12.5.2 Очистка данных

Полученные данные могут содержать ошибки или отсутствующие значения из-за различий в источниках данных и уровне качества данных. В этом случае очистка данных (см. 7.5.9.2, 7.5.9.3) может

использоваться для проверки согласованности данных в соответствии с единым стандартным форматом, для обработки неверных или недостающих данных во время хранения и для исправления идентифицированных ошибок в файле из хранилища данных. Например, из текстовых данных можно удалить бессмысленные слова в соответствии с требованиями анализа текста, а изображения можно обрезать и сохранить в стандартном размере и формате.

В случае большого объема данных можно сформировать и очистить выборку в соответствии с требованиями к качеству анализа данных.

12.5.3 Преобразование данных

При росте объема данных типы получаемых данных могут сильно различаться. С увеличением разнообразия данных первоначально используемый формат хранения и размеры хранилища могут перестать соответствовать требованиям решаемых задач. Для того чтобы реализовать большое количество задач аналитики, необходимо преобразовать данные в единый формат для хранения. Например, для изображений с разным разрешением можно применить преобразование масштаба, чтобы привести все изображения к единому масштабу.

Правильное преобразование данных (см. 7.5.9.3) может иметь важное значение для эффективности анализа данных. Например, если строки «собака» или «кошка» не могут быть обработаны аналитической моделью напрямую, то строки необходимо преобразовать в числовые значения 0 и 1, чтобы обеспечить возможность вычислений.

12.5.4 Агрегирование данных

Когда количество полученных данных становится очень большим, может возникнуть необходимость сгруппировать данные, прежде чем их можно будет проанализировать.

Агрегация данных объединяет тесно связанные данные, которые могут быть разбросаны по разным наборам данных для получения более полного описания данных. Группировка данных — это форма агрегирования данных, которая делит исходные данные на различные группы на основе определенных характеристик для удовлетворения требований аналитики.

Основной целью группировки данных является наблюдение за характеристиками распределения данных. После группировки данных полезно построить таблицу распределения частот путем расчета частоты данных в каждой группе для наблюдения за основным распределением данных.

12.5.5 Оценка качества данных

Оценка качества данных используется для исследования и оценки процессов комплектования и хранения данных, а также для создания системы мониторинга и оценки в режиме реального времени. После выявления каких-либо проблем с качеством данных их можно со временем устранить, чтобы снизить риски неправильного использования данных.

Для оценки качества данных следует выбрать соответствующий показатель. Визуализация данных может использоваться для изучения данных в графическом формате и получения дополнительного понимания данных. Дополнительную информацию о показателях качества данных см. в *ГОСТ Р 71484.2*. Дополнительную информацию о визуализации качества данных см. в [*13*].

12.5.6 Повышение качества данных

Повышение качества данных может включать аугментацию данных и извлечение данных.

Аугментация данных используется для увеличения объема данных на основе существующих данных. Аугментация данных для аналитики аналогична описанной в 7.5.9.4. Методы аугментации данных позволяют расширить набор данных без изменения атрибутов в наборе данных или их статистического распределения.

Применение методов извлечения данных позволяет обнаружить скрытые знания и выявить ранее неизвестные закономерности в данных. Таким образом, качество данных также можно проверить путем их внедрения в процесс извлечения данных и интерпретируемости полученного заключения в рамках рассматриваемого вида деятельности. Постоянная корректировка направления извлечения данных также полезна для повышения качества данных.

Извлечение данных может включать корреляционный анализ, анализ временных рядов и кластерный анализ:

- корреляционный анализ определяет корреляцию между различными количественно определяемыми событиями, т. е. когда происходит одно событие, часто происходит и другое событие. Например, проводится корреляционный анализ нескольких параметров из наборов климатических данных. Дополнительную информацию см. в 5.9.5 *ГОСТ Р 59926—2021*;
- анализ временных рядов определяет изменения полученных данных с течением времени посредством серии записей точек данных с согласованными временными интервалами, например анали-

тика на основе данных для временных рядов из киберфизических систем. Дополнительную информацию см. в 5.10.1 *ГОСТ Р 59926—2021*;

- кластерный анализ — это процесс группировки элементов данных в разные кластеры, так чтобы элементы в каждом кластере были больше похожи друг на друга, чем на элементы в других кластерах, например группировка документов по тематике. Дополнительную информацию см. в 5.3.4 *ГОСТ Р 59926—2021*.

Результаты интеллектуального анализа данных можно отобразить с помощью визуализации данных. Визуализация данных может помочь оценить качество данных. Таким образом, визуальное представление данных также помогает выявить проблемы внутри данных и дает рекомендации по улучшению процесса управления качеством данных.

12.6 Предоставление данных

Процесс предоставления данных для аналитики описан в 7.6.

12.7 Вывод данных из эксплуатации

Процесс вывода из эксплуатации данных для аналитики описан в 7.7.

Приложение ДА
(справочное)

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном стандарте

Таблица ДА.1

Обозначение национального стандарта	Степень соответствия	Обозначение и наименование ссылочного международного стандарта
ГОСТ Р 59926—2021 (ISO/IEC TR 20547-2:2018)	MOD	ISO/IEC TR 20547-2:2018 «Информационные технологии. Эталонная архитектура больших данных. Часть 2. Варианты использования и производные требования»
ГОСТ Р 71476—2024 (ИСО/МЭК 22989:2022)	MOD	ISO/IEC 22989:2022 «Искусственный интеллект. Концепции и терминология искусственного интеллекта»
ГОСТ Р ИСО 2859-1—2007	IDT	ISO 2859-1:1999 «Статистические методы. Процедуры выборочного контроля по альтернативному признаку. Часть 1. Планы выборочного контроля последовательных партий на основе приемлемого уровня качества»
Примечание — В настоящей таблице использованы следующие условные обозначения степени соответствия стандартов: - IDT — идентичный стандарт; - MOD — модифицированные стандарты.		

Библиография

[1] ИСО 24612:2012 Управление языковыми ресурсами. Лингвистическая аннотационная система [Language resource management — Linguistic annotation framework (LAF)]

[2] ИСО/МЭК 23751:2022 Информационные технологии. Облачные вычисления и распределенные платформы. Рамочное соглашение об обмене данными [Information technology — Cloud computing and distributed platforms — Data sharing agreement (DSA) framework]

[3] ИСО/МЭК 3532-1:2023 Информационная технология. Моделирование медицинских изображений на основе 3D-печати. Часть 1. Общие требования (Information technology — Medical image-based modelling for 3D printing — Part 1: General requirements)

[4] *XML 1.0, выпуск 5. Рекомендации консорциума W3C. Расширяемый язык разметки текста (XML 1.0, issue 5. W3C Recommendation. 26.11.2008. Extensible Markup Language* gost.e-ecolog.ru

[5] RFC 4180 Общий формат и тип MIME для файлов с разделителями значений запятыми (CSV) [RFC 4180 Common Format and MIME Type for Comma-Separated Values (CSV) Files]

[6] *ПНСТ 838—2023/ИСО/МЭК 23053:2022 Искусственный интеллект. Структура описания систем искусственного интеллекта, использующих машинное обучение*

[7] de Amorim L.B.V., Cavalcantia G.D.C., Cruz R.M.O., The choice of scaling technique matters for classification performance. arXiv preprint. 2022. arXiv: 2212.12343v1

[8] Barocas S, Selbst A.D., Big data's Disparate Impact. California Law Review. 2016, 104(3), 671—732. doi: 10.2139/ssrn.247789

[9] Japkowicz N., Stephen S., The class imbalance problem: A systematic study. Intelligent data analysis. 2002, 6(5), 429—449. doi: 10.5555/1293951.1293954

[10] ИСО/МЭК 27559:2022 Информационная безопасность, кибербезопасность и защита конфиденциальности. Система деидентификации данных, повышающая конфиденциальность *(Information security, cybersecurity and privacy protection — Privacy enhancing data de-identification framework)*

[11] Biswas S., Rajan H., Fair Preprocessing: Towards Understanding Compositional Fairness of Data Transformers in Machine Learning Pipeline. arXiv preprint. 2021. arXiv: 2106.06054v5

[12] Temple P., Acher M., Perrouin G., Biggio B., Jézéquel J.-M., Roli F., Towards Quality Assurance of Software Product Lines with Adversarial Configurations. arXiv preprint. 2019. arXiv: 1909.07283v1

[13] ISO/IEC CD TR 5259-6 Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 6. Структура визуализации качества данных (Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 6: Visualization framework for data quality)

УДК 004.01:006.354 ОКС 35.020

Ключевые слова: искусственный интеллект, качество данных, аналитика, машинное обучение, структура, управление качеством данных

Технический редактор *В.Н. Прусакова*
Корректор *Л.С. Лысенко*
Компьютерная верстка *М.В. Малеевой*

Сдано в набор 19.11.2024. Подписано в печать 03.12.2024. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 4,18. Уч.-изд. л. 3,35.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «Институт стандартизации»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.

gost.e-ecolog.ru gost.e-ecolog.ru

gost.e-ecolog.ru