ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ
РОССИЙСКОЙ
ФЕДЕРАЦИИ

ГОСТ (Р)
_____—_____.

Информационные технологии

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Качество данных для аналитики и машинного обучения. Часть 2. Показатели качества данных

ISO/IEC 5259-2 Information technology — Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures.

Настоящий проект стандарта не подлежит применению до его принятия

Москва
Стандартинформ
20XX

Предисловие

1 ПОДГОТОВЛЕН Федеральным государственным бюджетным образовательным учреждением высшего образования «Московский государственный университет имени М.В.Ломоносова» (МГУ имени М.В.Ломоносова) в лице Научно-образовательного центра компетенций в области цифровой экономики МГУ 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__ _____ 202_ г. № ____.

4 Настоящий стандарт идентичен разрабатываемому международному стандарту ISO/IEC 5259–2 Information technology — Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures, учитывающих специфику национальной стандартизации в области искусственного интеллекта и больших данных.

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 (www.rst.gov.ru)

© ISO, 20XX — Все права сохраняются
© Стандартинформ, оформление, 20XX

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

Предисловие II
1 Область применения 1
2 Нормативные ссылки 1
3 Термины и определения 1
4 Обозначения и сокращения 4
5 Модель качества данных для аналитики и машинного обучения 4
5.1 Измерение качества данных в жизненном цикле данных 4
5.2 Система измерения качества данных 6
5.3 Модель качества данных 7
5.4 Характеристики качества данных и меры качества 8
5.4.1 Общие положения 8
5.4.2 Переносимость 8
5.4.3. Понятность 10
5.4.4 Возможность аудита 11
5.4.5 Идентифицируемость 13
5.4.6 Актуальность 14
5.4.7. Достоверность 16
5.4.8 Полнота 18
Характеристики качества набора данных и меры качества 24
5.5.1 Общие положения 24
5.5.2 Масштабируемость данных 25
5.5.3 Обобщаемость 25
5.5.4 Эффективность 25
5.5.5 Представимость 27
5.5.6 Точность 31
5.5.7 Согласованность 31
5.5.8 Релевантность 33
5.5.9 Своевременность 36

5.5.10 Репрезентативность 36
5.5.11 Сбалансированность 37
5.5.12 Подобие 43
5.5.13 Разнообразие 45
6 Методы улучшения качества данных 46
6.1 Дополнение данных 46
6.2 Стандартизация данных 47
6.3 Нормализация данных 47
6.4 Вменение данных 48
6.5 Шифрование данных 48
7 Отчет о качестве данных 49
7.1 Информация о показателях качества данных 49
7.2. Руководство для организаций 50
Библиография 51
Приложение А (справочное) Сведения о соответствии ссылочных международных стандартов национальным стандартам 0
Разработка и документирование функции измерения 1
UML-модель системы измерения качества данных 3
Обзор характеристик качества данных и наборов данных 4

1 Область применения

В этом документе представлена модель качества данных, показатели качества данных и рекомендации по составлению отчетов о качестве данных в контексте аналитики и машинного обучения. Этот документ основан на серии стандартов ISO 8000, ISO/IEC 25012 и ISO/IEC 25024.

Цель этого документа — помочь организациям достичь своих целей в области качества данных и применима ко всем типам организаций.

2 Нормативные ссылки

Следующие документы упоминаются в тексте таким образом, что часть или все их содержание составляет требования настоящего документа. Для датированных ссылок применяется только цитируемое издание. Для недатированных ссылок применяется последнее издание ссылочного документа (включая любые поправки).

ГОСТ Р ИСО/МЭК 20546–2021 Информационные технологии. Большие данные. Обзор и словарь

ISO/IEC 22989:2022, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology (Информационные технологии — Искусственный интеллект — Понятия и терминология искусственного интеллекта)

ISO/IEC 5259-1:202X Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples (Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, терминология и примеры)

3 Термины и определения

ISO и IEC поддерживают терминологические базы данных для использования в стандартизации по следующим адресам:

- платформа просмотра ISO Online доступна по адресу https://www.iso.org/obp
- IEC Electropedia: доступно на http://www.electropedia.org/

Для целей настоящего документа применяются термины и определения, приведенные в ГОСТ Р ИСО/МЭК 20546–2021, ISO/IEC 22989:2022, ISO/IEC 5259–1:202X и следующие термины.

3.1 **показатель** Переменная, которой присваивается значение в результате измерения

Примечание — Форма множественного числа «показатели» используется для общего обозначения базовых показателей, производных показателей и индикаторов.

[ISO/IEC 15939:2017, 3.15]

3.2 **измерять:** Сделать измерение

[ISO/IEC 25000:2014]

3.3 **измеренное значение:** Значение количества, представляющее результат измерения

[JCGM 200:2012, 2.10]

3.4 **функция измерения:** алгоритм или расчет, выполняемый для объединения двух или более элементов измерения качества

[ISO/IEC 25021:2012, 4.7]

3.5 **элемент данных:** Наименьшая идентифицируемая единица данных в определенном контексте, для которой определение, идентификация, допустимые значения и другая информация задаются с помощью набора свойств

Примечания:

1 Поле считается синонимом элемента данных.

2 Элемент данных — это физический объект «контейнер» значений данных.

[ISO/IEC 25024:2015, 4.9]

3.6 **тип данных:** Категоризация абстрактного набора возможных значений, характеристик и набора операций для атрибута

Примечание- Примерами типов данных являются строки символов, тексты, даты, числа, изображения, звуки и т. д.

[ISO/IEC 25024:2015, 4.16]

3.7 **значение данных:** Содержание элемента данных (3.5)

Примечания:

1 В стандарте ISO/IEC 25012, 5.1.1 указано, что с «внутренней» точки зрения качество данных относится к самим данным, таким как значения предметной области и возможные ограничения.

2 Номер или категория, присвоенная атрибуту целевого объекта путем проведения измерения.

[ISO/IEC 25024:2015, 4.17]

3.8 **запись данных:** Набор связанных элементов данных (3.5), рассматриваемых как единое целое

[ISO/IEC 25024:2015, 4.15]

3.9 **кадр данных:** Набор записей данных (3.8), представленных определенной областью или целью с теми же элементами данных (3.4), которые используются для аналитики и машинного обучения

3.10 **требования к качеству данных:** Требование к атрибутам качества данных, которые удовлетворяют потребности в том, как данные должны использоваться

3.11 **целевые данные:** Данные, имеющие отношение к пользователю, качество которого необходимо измерить

3.12 П**ереобучение** <машинное обучение>:Создание модели, которая слишком точно соответствует обучающим данным и не может обобщать новые данные

Примечания

1 Переобучение может произойти из-за того, что обученная модель научилась на второстепенных признаках обучающих данных (т. е. значительное несоответствие между данными обучения и распределениями производственных данных или потому, что модель слишком сложна для данных обучения).

2 Переобучение можно определить, когда существует значительная разница между ошибками, измеренными на обучающих данных и на отдельных данных тестирования и проверки. На производительность переобученных моделей особенно влияет значительное несоответствие между обучающими данными и производственными данными.

[ISO/IEC 23052–3.2.4]

3.13 **кластер:** автоматически создаваемая категория элементов, которые являются частью набора данных и имеют общие атрибуты

Примечание — Кластеры не обязательно имеют имя.

[ISO/IEC 23052: 2022]

4 Обозначения и сокращения

ИИ	искусственный интеллект
ГО	глубинное обучение
МО	машинное обучение
CSV	Значения, разделенные запятыми
МК	Мера качества

5 Модель качества данных для аналитики и машинного обучения

5.1 Измерение качества данных в жизненном цикле данных

На рис. 1 показана роль качества данных в жизненном цикле данных для аналитики и машинного обучения. Модель качества данных состоит из характеристик качества данных, уровней целей качества и отчетов об оценке качества. Верхняя часть рисунка 1 (над пунктирной линией) основана на стандарте ISO/IEC 5259-1: — рисунок 3, на котором показана модель жизненного цикла данных и его различные

этапы. В нижней части рисунка 1 (под пунктирной линией) показаны этапы модели качества данных, определенной в этом документе. Этап требований к качеству данных наряду с требованиями к данным и планированием данных обеспечивает удовлетворение потребностей организации, которая использует и анализирует данные. Этап требований к качеству данных заключается в том, чтобы определить целевые данные, которыми нужно управлять для достижения качества, выбрать характеристики качества и указать меры качества и их критерии приемлемости для оценки, которые могут дать значимые результаты в аналитике и машинном обучении. Критерии приемлемости обеспечивают уровни качества, целевые измеренные значения и пороговые значения, которые соответствуют целям организации.

Требования к качеству данных могут противоречить друг другу. Когда одна характеристика качества влияет на другую, компромиссы могут быть достигнуты путем оценки каждого требования в отношении важности и воздействия. Кроме того, крайне важно сбалансировать стоимость управления качеством данных с приоритетом требований к качеству данных при определении характеристик и показателей данных. После определения характеристик, показателей и критериев качества данных выполняется этап оценки качества данных с целевыми данными во время подготовки и предоставления данных. Оценка качества данных реализуется действиями по профилированию данных и измерениям качества данных, как показано на рисунке 2:

- Профилирование данных — это действие, выполняемое для понимания распределения значений данных в целевых данных с использованием статистических методов для выборок или целых наборов данных;
- Измерение качества данных — это действие, которое производит меры качества данных и измеренные значения характеристик и профилей качества данных в отношении выявленных ключевых наборов данных.

Качество данных следует оценивать на основе требований организации, а любые проблемы, связанные с качеством данных, следует смягчать или решать иным образом. Модели контролируемого машинного обучения строятся с использованием обучающих данных, а соответствие требованиям определяется данными тестирования и проверки. Когда производительность обученной модели соответствует требованиям организации, ее можно рассматривать для развертывания и использования. Для каждого развертывания можно выполнить точную настройку, чтобы убедиться, что развернутая модель соответствует

дополнительным конкретным требованиям в дополнение к общему набору требований. Систему ИИ с моделью машинного обучения следует контролировать, чтобы убедиться, что она продолжает соответствовать требованиям, и при необходимости настраивать, переобучать или заменять. Жизненный цикл системы ИИ описан в ISO/IEC 22989, 6.

5.2 Система измерения качества данных

Приложение С описывает основу для измерения характеристик и показателей качества данных на основе эталонной модели измерения качества, определенной в ISO/IEC 25020. Модель качества данных содержит набор характеристик качества данных, которые могут быть далее подразделены на подхарактеристики. Характеристика качества данных связана с одним или несколькими показателями качества данных. Каждому показателю качества данных присваивается измеренное значение с помощью функции измерения двух или более элементов показателя качества. Элемент меры качества данных представляет собой базовую меру, определяемую в терминах атрибута метода измерения для его количественной оценки.

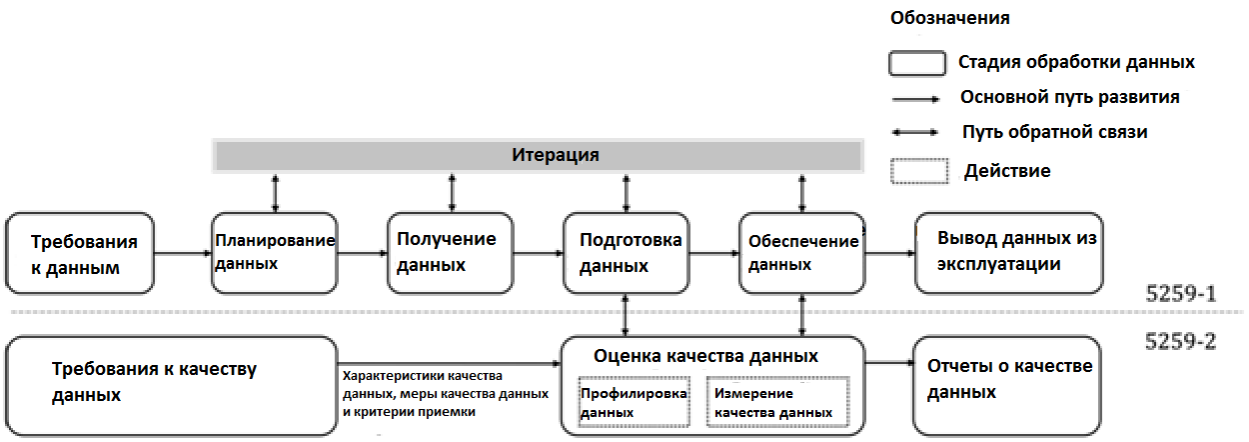

Рисунок 1 — Роль модели качества данных в жизненном цикле данных для аналитики и машинного обучения

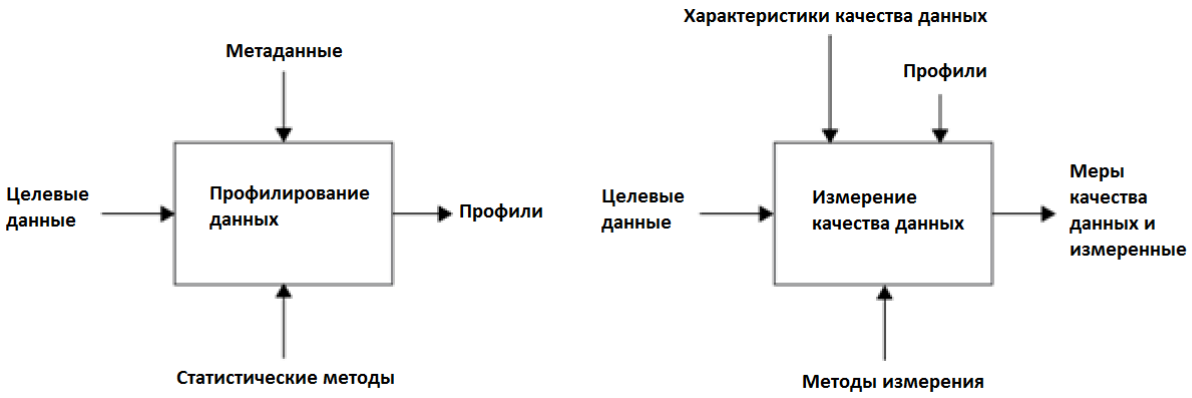

Рисунок 2 — Оценка качества данных

5.3 Модель качества данных

Стандарт ISO/IEC 5259–1 определяет модель качества данных как набор характеристик, который обеспечивает основу для определения требований к качеству данных и оценки качества данных. На практике модель качества данных состоит из характеристик качества данных и показателей качества данных, соответствующих контексту использования данных. Требования к качеству данных могут включать приемлемые критерии для выбранных показателей качества данных.

Модель качества задается определенным набором характеристик качества и взаимосвязей между ними. Модель качества данных в этом документе определяет, как различные аспекты качества данных влияют на жизненный цикл данных аналитики и машинного обучения. ISO 8000–8 и ISO/IEC 25012 описывают модели качества данных. ISO 8000–8 определяет три характеристики качества: синтаксическую (формат), семантическую (значение) и прагматическую (полезность) для общей поддержки промышленных данных как продукта бизнес-процессов и производственных процессов. ISO/IEC 25012 определяет общую модель качества данных для данных, сохраняемых в структурированном формате в компьютерной системе как часть программного продукта. ISO/IEC 25012 учитывает все типы данных символов, строк, текстов, дат, чисел, изображений, звуков и т. д. ISO/IEC 25012 обеспечивает пятнадцать характеристик: точность, полнота, непротиворечивость, достоверность, актуальность, доступность, соответствие требованиям,

конфиденциальность, эффективность, точность, прослеживаемость, понятность, доступность, переносимость и возможность восстановления.

Серия ISO 8000 касается различных аспектов качества данных, таких как управление данными, управление качеством данных (включая обработку) и оценку зрелости. Серия ISO/IEC 25000 посвящена системам и требованиям к качеству программных продуктов и их оценке. Эти международные стандарты выходят за рамки серии ISO 8000 (см. ISO 8000–1:2022, 4). Серия ISO/IEC 25000 дает практическое представление о качестве данных в контексте отдельных систем. Этот вид дополняет серию стандартов ISO 8000 (см. ISO 8000–1:2022, 8.4).

5.4 Характеристики качества данных и меры качества

5.4.1 Общие положения

Характеристики и меры качества данных используются для определения и проверки требований к качеству данных для определенных атрибутов целевых данных. Каждая характеристика качества связана с более чем одной мерой качества данных для количественной оценки. Показатели качества данных можно классифицировать несколькими способами:

- по характеристикам требований к качеству;
- по стадиям жизненного цикла данных;
- по свойствам (значение данных, элемент данных, запись данных и фрейм данных);
- по целевым данные (данные для обучения, данные проверки, данные испытаний, данные производства и данные прогнозирования);
- в соответствии с другими аспектами (например, расширение данных, стандартизация данных, нормализация данных, вменение данных и кодирование данных).

5.4.2 Переносимость

Переносимость данных — это возможность легко переносить данные из одной системы в другую без необходимости повторного ввода данных.

Примечание — Суть здесь заключается в простоте перемещения данных. Это может быть достигнуто путем предоставления исходной системой данных именно в том формате, который принимает целевая система. Но даже если форматы не совпадают, преобразование между ними может быть простым и понятным с помощью общедоступных инструментов. С другой стороны, процесс распечатки данных и повторного ввода их для целевой системы нельзя назвать «простым».

5.4.2.1 Общие положения

ISO/IEC 25024 описывает характеристику качества переносимости данных с точки зрения возможности перемещать данные из одной системы в другую в рамках определенного контекста, сохраняя при этом их качество.

Данные, используемые в аналитике и машинном обучении, могут обрабатываться в нескольких системах. Например, данные для машинного обучения могут быть получены в одной системе, процессы качества данных могут выполняться для данных с использованием второй системы, а затем данные передаются в третью систему для обучения модели машинного обучения.

Если качество данных (то есть соответствие требованиям) не поддерживается при передаче данных из одной системы в другую, то сама обученная модель машинного обучения может не соответствовать требованиям.

5.4.2.2 МК для переносимости

В таблице 1 представлены показатели качества данных для переносимости в контексте аналитики и машинного обучения.

Таблица 1 — Меры переносимости

ID	Имя	Описание	Функция измерения	Целевые категории данных
Por-D-1	Коэффициент переносимости данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования

Por-D-2	Перспективная переносимость данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования

5.4.3. Понятность

5.4.3.1 Общие положения

ISO/IEC 25012 описывает понятность с точки зрения возможности пользователей читать и интерпретировать данные. Кроме того, понятность включает использование соответствующих символов, единиц измерения и языков. Модели машинного обучения — это математические конструкции, учитывающие величину чисел. Модели машинного обучения могут не соответствовать требованиям, если единицы измерения функций используются ненадлежащим образом.

Для задач обработки естественного языка неправильное использование человеческих языков и символов может привести к сбою таких задач, как понимание языка и преобразование текста в речь.

Хотя меры качества данных, описанные в этом документе, являются количественными, люди, использующие данные для машинного обучения, также проводят качественную оценку данных. Надлежащее использование символов, единиц измерения и языков может помочь в качественной оценке.

5.4.3.2 МК для понятности

В таблице 2 представлены показатели качества данных для понятности в конкретном контексте использования аналитики и машинного обучения.

Таблица 2 — Показатели понятности

ID	Имя	Описание	Функция измерения	Целевые категории данных
Und-I-1	Разборчивость символов	См. ISO/IEC 25024		Тренировочные данные Данные проверки

				Данные тестирования
Und-I-2	Семантическая понятность	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования
Und-I-4	Понятность значений данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования
Und-D-2	Понятность представления данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования

5.4.4 Возможность аудита

5.4.4.1 Общие положения

Для целей настоящего документа доступность относится к характеристике набора данных, состоящей в том, что весь набор данных или его часть прошел аудит или что данные доступны соответствующим заинтересованным сторонам для целей проведения аудита. Аудит наборов данных, используемых для аналитики и машинного обучения, может повысить достоверность данных и может потребоваться для соответствия требованиям.

Пример. Набор данных изображений используется для распознавания изображений и был помечен сторонним подрядчиком. Чтобы убедиться, что

изображения помечены должным образом, организация использует другую третью сторону для аудита подмножества помеченных изображений.

5.4.4.3 МК для возможности аудита

В таблице 3 представлены показатели качества данных для возможности аудита в конкретном контексте использования аналитики и машинного обучения.

Таблица 3 — Показатели возможности аудита

ID	Имя	Описание	Функция измерения	Целевые категории данных
Aud-I-1	Проверенные записи	Соотношение записей в наборе данных, которые были проверены	X=A/B A=Количество записей в наборе данных, которые были проверены. B=Общее количество записей в наборе данных	Тренировочные данные Данные проверки Данные тестирования Производственные данные Выходные данные

Окончание таблицы 3

Aud-I-2	Проверяемые записи	Доля записей в наборе данных, доступных для проверки	X=A/B A= Количество записей в наборе данных, которые доступных для проверки. B= Общее количество записей в наборе данных	Тренировочные данные Данные проверки Данные тестирования Производственные данные Выходные данные

5.4.5 Идентифицируемость

5.4.5.1 Общие положения

ISO/IEC 29100 описывает идентифицируемость как возможность идентифицировать субъекта PII на основе PII. Важно понимать, можно ли использовать любую PII в наборе данных для идентификации принципа PII, поскольку юридические требования в некоторых юрисдикциях могут ограничивать такую деятельность. Процессы деидентификации могут применяться к данным обучения, проверки, тестирования и производства, чтобы уменьшить возможность идентификации. В Таблице 4 представлены QM для идентифицируемости.

Пример. Модель машинного обучения обучается на поисковых запросах с целью таргетированной рекламы. Набор данных включает IP-адрес пользователя, который в некоторых юрисдикциях считается PII. К набору данных применяется анонимизация для удаления IP-адреса перед разделением набора данных на наборы данных для обучения, проверки и тестирования.

5.4.5.2 МК для идентифицируемости

В таблице 4 представлены меры качества данных для идентификации в конкретном контексте использования аналитики и машинного обучения.

Таблица 4 — Показатели идентифицируемости

ID	Имя	Описание	Функция измерения	Целевые категории данных
Iden-I-1	Коэффициент идентифицируемости	Отношение записей данных в наборе данных, которые можно использовать для идентификации	X=A/B A=Количество записей в наборе данных, которые можно использовать для идентификации B= количество записей в наборе данных	Тренировочные данные Данные проверки Данные тестирования

5.4.6 Актуальность

5.4.6.1 Общие положения

ISO/IEC 25012 описывает актуальность данных с точки зрения правильного возраста по отношению к использованию данных. Для МО актуальность может выражаться в возрастном диапазоне, подходящем для задачи МО. Например, данные о людях могут быть неполными для недопредставленных групп населения до изменений в правилах и социальных нормах. Модели машинного обучения, основанные на экономических данных, собранных за несколько десятилетий, могут быть неверными, если данные не скорректированы с учетом инфляции, обменных курсов и других факторов, меняющихся со временем.

Актуальность набора данных можно описать с точки зрения общего периода времени, который охватывает набор данных (например, изображения или предложения, собранные с 2010 по 2021 год), периода времени между последней датой элемента данных и текущей датой (например, 8 месяцев), что такое цикл обновления (например, каждые 6 месяцев). Актуальность следует рассматривать как составную метрику, основанную на этих трех характеристиках.

Пример 1. Модель машинного обучения, используемая для прогнозирования будущих продаж, постоянно не учитывает фактическую сумму продаж. В результате расследования было установлено, что для создания модели использовались данные обучения за десять лет сделок купли-продажи, но значения данных не были скорректированы с учетом инфляции.

Пример 2. Модель МО используется для классификации кандидатов, за которых будет голосовать тот или иной человек. Модель была обучена на протоколах голосования за предыдущие двадцать лет. Некоторые недопредставленные группы не голосовали регулярно до последних десяти лет. Модель ML постоянно не может правильно предсказать выбор, сделанный недопредставленными избирателями.

5.4.6.2 МК для актуальности

В таблице 5 представлены показатели качества данных для актуальности в конкретном контексте использования аналитики и машинного обучения.

Таблица 5 — Показатели актуальности

ID	Имя	Описание	Функция измерения	Целевые категории данных
Curr-I-1	Актуальность свойства	Соотношение элементов данных в наборе данных, попадающих в требуемый возрастной диапазон	X=A/B A=количество элементов данных для свойства, попадающих в требуемый возрастной диапазон B=общее количество элементов данных для свойства	

Окончание таблицы 5

Curr-I-2	Актуальность записи	Соотношение записей данных в наборе данных, при котором все элементы данных в записи попадают в требуемый возрастной диапазон	X=A/B A=Количество записей данных, которые попадают в требуемый возрастной диапазон. B= общее количество элементов данных в наборе данных.	

5.4.7. Достоверность

5.4.7.1 Общие положения

ISO/IEC 25012 определяет достоверность с точки зрения степени, в которой данные имеют атрибуты, которые пользователи считают истинными и правдоподобными в конкретном контексте использования. Достоверность применима к отдельным элементам данных, к связанным элементам данных в записи данных и ко всему набору данных. Контекст, в котором используются данные, может повлиять на их воспринимаемую достоверность и достоверность. Данные могут быть искажены во время обработки (например, при передаче, хранении, вычислении) уполномоченными и неуполномоченными сторонами. Возникающая проблема машинного обучения заключается в том, что несанкционированные стороны вмешиваются в данные обучения, проверки, тестирования и производства, чтобы преднамеренно сделать обученные модели непригодными для использования или манипулировать выводами, сделанными обученной моделью.

Процессы, используемые при подготовке данных, могут изменять данные без изменения их значения (например, нормализация, ампутация, разделение или объединение признаков). В этих случаях данные сохраняют свою достоверность.

Пример 1. Набор данных используется для обучения, проверки и тестирования модели машинного обучения, которая затем не может обеспечить требуемую производительность на производственных данных. Аудит безопасности показывает, что неавторизованная сторона случайным образом изменила значения данных в обучающем наборе данных.

Пример 2.Набор обучающих данных содержит числовые признаки, диапазоны которых сильно различаются. Исследователь данных решает нормализовать значения данных для этих функций, чтобы сделать их более сопоставимыми. Хотя значения данных изменятся после нормализации, они по-прежнему заслуживают доверия, поскольку их значение не изменилось в контексте ML.

5.4.7.2 МК для достоверности

В таблице 6 представлены показатели качества данных для достоверности в конкретном контексте использования аналитики и машинного обучения.

Таблица 6 — Показатели достоверности

ID	Имя	Описание	Функция измерения	Целевые категории данных
Cre-I-1	Достоверность значений	См. ISO/IEC25024		Тренировочные данные Данные проверки Данные тестирования
Cre-I-2	Достоверность источника	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования

Окончание таблицы 6

Cre-I-3	Достоверность словаря данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования
Cre-I-4	Достоверность модели данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования

5.4.8 Полнота

5.4.8.1 Общие положения

ISO/IEC 25012 описывает полноту с точки зрения данных, имеющих значения для всех ожидаемых атрибутов и экземпляров объекта. В некоторых случаях алгоритмы машинного обучения могут дать сбой, когда они сталкиваются с одним или несколькими пустыми значениями данных при обучении, проверке или тестировании наборов данных. Кроме того, обученные модели МО также могут дать сбой, если производственные данные содержат пустые значения данных.

Показатели полноты могут помочь специалистам по МО выполнить свои требования к данным и указать, следует ли предпринять дополнительные шаги, как описано в пункте 6.4.

Полнота значения может быть измерена как отношение отсутствующих значений набора данных в наборе данных. Его следует измерять для каждого элемента схемы набора данных = столбца (например, значения X отсутствуют в дате, значения Y отсутствуют в источнике изображения и т. д.), сколько значений из общего числа значений (отсутствуют в каждом столбце) и в целом (например, В наборе

данных отсутствуют общие значения X+Y из общего количества значений (столбцы X строк).

Характеристика полноты помеченных данных в наборе данных является относительной. В разных сценариях значение полноты может быть разным, и его следует рассматривать в конкретных контекстах использования. Факторы, которые следует учитывать для измерения полноты набора данных:

- Полнота набора данных, используемого для классификации изображений на основе МО, должна проверять немаркированные образцы в наборе данных, которые нельзя напрямую использовать в контролируемом МО;
- Полнота набора данных, используемого для обнаружения объектов на основе МО, должна проверять неполноту помеченных ограничивающих рамок на объектах.

В частности, в реальной жизни часто используется тестовый набор с несколькими объектами разных категорий, поскольку сложно захватить сцену с одним изолированным объектом, занимающим все пространство просмотра. В этом случае, чтобы измерить полноту набора данных для распознавания изображений на основе МО, следует учитывать:

- в выборке существует какой-либо целевой объект;
- все целевые объекты категоризированы;
- все обнаруженные целевые объекты помечаются ограничивающими прямоугольниками или другими способами.

Пример 1. Мера полноты для набора данных указывает, что в наборе данных отсутствует более половины значений данных для функции почтового индекса. Исследователь данных решает, что функция почтового индекса не является необходимым предиктором для их задачи классификации, и решает удалить функцию почтового индекса из наборов данных для обучения, проверки, тестирования и производства.

Пример 2. Мера полноты для набора данных, используемого для задачи регрессии МО, указывает, что один процент значений данных для функции, которая является хорошим предиктором, пуст. Остальные данные имеют нормальное распределение. Исследователь данных решает заполнить пустые значения данных средним статистическим значением доступных значений данных.

Пример 3. Мера полноты для набора данных, используемого для задачи кластеризации машинного обучения, указывает на то, что небольшое количество записей имеет одно или несколько отсутствующих значений данных. Исследователь данных решает удалить эти записи из обучающих данных.

Пример 4. Полнота элементов — это соотношение отсутствующих записей данных по сравнению с целевыми номерами элементов, ожидаемыми для правильной точности набора данных. Если в наборе данных утверждается, что есть изображения всех растений в США, но отсутствуют 10 растений с северо-востока.

5.4.8.2 МК для полноты

В таблице 7 представлены меры качества данных для полноты в конкретном контексте использования аналитики и машинного обучения.

Таблица 7 — Показатели полноты

ID	Имя	Описание	Функция измерения	Целевые категории данных
Comp-I-1	Полнота значения	Полнота записей данных в наборе данных	X=A/B A= количество записей данных в наборе данных, не имеющих нулевых значений данных B= количество записей данных в наборе данных	Тренировочные данные Данные проверки Данные тестирования

Продолжение таблицы 7

Comp-I-2	Полнота элемента	Полнота записей в наборе данных	N = Всего строк в наборе данных M = Всего столбцов в наборе данных A = sum (A1,..., AM) / N ==> доля элементов с отсутствующими значениями (на уровне набора данных) A1 = % элементов у которых отсутствует 1 значение для каждого элемента (на уровне элемента) A2 = % элементов у которых отсутствует 2 значения для каждого элемента (на уровне элемента) AM = % элементов у которых отсутствует M значений для каждого элемента (на	Тренировочные данные Данные проверки Данные тестирования

			уровне элемента)	

Продолжение таблицы 7

Comp-I-3	Полнота свойства	Полнота элементов данных для данного свойства в наборе данных	X=A/B A= количество элементов данных, связанных с данным свойством, со связанным значением данных, отличным от нуля B= количество элементов данных, связанных с данным свойством	Тренировочные данные Данные проверки Данные тестирования Производственные данные
Comp-I-4	Полнота записи	Полнота записей в наборе данных	X=A/B A= количество записей данных в наборе данных, не имеющих нулевых значений данных B= количество записей данных в наборе данных	Тренировочные данные Данные проверки Данные тестирования Производственные данные

Продолжение таблицы 7

Comp-I-5	Полнота меток категории	В наборе данных соотношение образцов с немаркированными или неполностью маркированными категориями	X = 1- A/B A = количество образцов с немаркированными или неполностью маркированными категориями B = общее число образцов в наборе данных	Тренировочные данные Данные проверки Данные тестирования
Comp-I-6	Полнота метки ограничительной рамки	В наборе данных соотношение отсутствующих ограничительных рамок к истинным.	X = 1 – A/B A = количество неполных либо отсутствующих ограничительных рамок B = количество истинных ограничительных рамок во всем наборе	Тренировочные данные Данные проверки Данные тестирования

Характеристики качества набора данных и меры качества

5.5.1 Общие положения

В контексте аналитики и машинного обучения общее качество набора данных для обучения, набора данных для проверки или набора тестовых данных может быть столь же важным, как и качество отдельного значения данных. Несмотря на то, что каждое значение данных в наборе данных является точным, набор данных, который неправильно отражает лежащее в основе распределение данных, может привести к

неправильному результату анализа или созданию ошибочной модели машинного обучения.

Примечание. Качество наборов данных может быть измерено статистическими характеристиками, определенными государственным учреждением, например. Статистический отдел Организации Объединенных Наций (СООН) и Европейская статистика (ЕВРОСТАТ), такие как: доступность для авторизованных пользователей, точность, последовательность, актуальность, понятность, актуальность, своевременность [4].

5.5.2 Масштабируемость данных

Масштабируемость данных — это степень, в которой качество данных поддерживается по мере увеличения объема и скорости передачи данных.

5.5.3 Обобщаемость

Обобщаемость — это степень, в которой данные могут использоваться как в определенных контекстах алгоритма ML, так и в контекстах, выходящих за рамки тех, которые изначально были явно определены.

5.5.4 Эффективность

5.5.4.1 Общие положения

Эффективность набора данных означает, в какой степени набор данных соответствует группе требований для использования в конкретном процессе МО:

- Для системы компьютерного зрения на основе машинного обучения эффективность набора данных может быть наименьшим приемлемым коэффициентом, при котором количество изображений с яркостью или разрешением ниже требуемого порога по всем образцам в наборе данных;
- Для системы классификации изображений на основе МО эффективность набора данных может относиться к наименьшему приемлемому соотношению количества изображений, принадлежащих классу, к общему количеству выборок в наборе данных;

- ля системы обнаружения объектов на основе МО эффективность набора данных может относиться к наименьшему приемлемому соотношению, при котором количество изображений с номерами или областями ограничивающих рамок ниже требуемого порога по всем образцам в наборе данных.

5.5.4.2 МК для эффективности

В таблице 8 представлены показатели качества данных для эффективности в конкретном контексте использования аналитики и машинного обучения.

Т а б л и ц а — Показатели эффективности

ID	Имя	Описание	Функция измерения	Целевые категории данных
Eff-I-1	Эффективность яркости	Соотношение образцов с недопустимой яркостью в наборе данных	X = A/B A = количество образцов с недопустимой яркостью B = общее количество образцов в наборе данных	Тренировочные данные Данные проверки Данные тестирования
Eff-I-2	Эффективность разрешения	Соотношение образцов с неприемлемым разрешением в наборе данных	X = A/B A = количество образцов с неприемлемым разрешением B = общее количество образцов в наборе данных	Тренировочные данные Данные проверки Данные тестирования

Окончание таблицы 8

Eff-I-3	Эффективность размера категории	Соотношение категорий, в которых количество категоризированных образцов ниже порогового значения	X = A/B A = количество категорий, в которых количество категоризированных образцов ниже порогового значения B = общее количество категорий	Тренировочные данные Данные проверки Данные тестирования
Eff-I-4	Эффективность области ограничительной рамки	Соотношение ограничивающих рамок, площадь которых меньше порога приемлемости	X = A/B A = количество ограничивающих рамок, площадь которых меньше порога приемлемости B = количество ограничивающих рамок	Тренировочные данные Данные проверки Данные тестирования

5.5.5 Представимость

5.5.5.1 Общие положения

ISO/IEC 25012 описывает представимость как степень, в которой значения данных имеют атрибуты, которые правильно представляют истинное значение предполагаемых атрибутов. ISO/IEC 25012 далее описывает точность с точки зрения:

- синтаксическая представимость, которая учитывает близость значений данных к набору синтаксически правильных значений данных в соответствующей области;
- семантическая представимость, которая учитывает близость значений данных к набору семантически правильных значений данных в соответствующей области.

Модели машинного обучения представляют собой математические конструкции, что означает, что низкая синтаксическая или семантическая т представимость значений данных в наборах данных для обучения, проверки, тестирования или производства может привести к тому, что сама модель будет неверной или что выводы, сделанные моделью, могут быть неверными.

Точность данных в наборе данных может иметь разное значение в разных приложениях. Для контролируемой системы классификации на основе машинного обучения правильность меток может повлиять на точность вывода обученной модели на основе данных. Факторы, которые следует учитывать для измерения точности маркировки, включают:

- корректность меток категорий;
- корректность помеченных ограничивающих рамок.

Пример 1. Если фраза «ленивая собака» введена как «lzy dg», система понимания естественного языка на основе машинного обучения может неправильно интерпретировать фразу.

Пример 2. Если число 100 введено как 1000 в обучающих данных, регрессионная модель может неправильно рассчитать вес связанного признака, а если запись была сделана в производственных данных, выводы могут быть неверными.

5.5.5.2 МК для представимости

В таблице 9 представлены показатели качества данных для точности в конкретном контексте использования аналитики и машинного обучения.

Таблица 9 — Показатели точности

ID	Имя	Описание	Функция измерения	Целевые категории данных
Acc-I-1	Синтаксическая представимость данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования Производственные данные
Acc-I-2	Семантическая представимость данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования Производственные данные
Acc-I-3	Обеспечение представимости данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования Производственные данные
Acc-1-4	Риск непредставимости набора данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования Производственные данные

Продолжение таблицы 9

Acc-I-5	Представимость модели данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования Производственные данные
Acc-I-6	Представимость метаданных	См. ISO/IEC 25024		Метаданные
Acc-I-7	Диапазон представимости данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования Производственные данные
Acc-I-8	Представимость маркировки ограничительной рамки	Соотношение образцов с неправильными ограничивающими рамками в наборе данных	X = A/B A = количество образцов, в каждом из которых есть неправильные ограничивающие рамки B = общее количество образцов в наборе	Тренировочные данные Данные проверки Данные тестирования

Окончание таблицы 9

Acc-I-9	Представимость маркировки класса	Соотношение образцов с неправильными классами в наборе данных	X = A/B A = количество образцов, в каждом из которых указан неправильный класс B = общее количество образцов в наборе	Тренировочные данные Данные проверки Данные тестирования

5.5.6 Точность

Точность описана в ISO/IEC 25012:2008, 5.3.2.5.

Точность относится к близости повторных измерений одного и того же явления друг к другу, то есть к степени, в которой случайная ошибка влияет на измеренные значения.

5.5.7 Согласованность

5.5.7.1 Общие положения

ISO/IEC 25012 описывает согласованность в терминах согласованности данных с другими данными и отсутствии противоречий. Согласованность — это ключевой аспект данных, используемых для машинного обучения, поскольку функции, используемые в обучающих данных, должны вместе обеспечивать модель, которая делает правильные выводы о производственных данных. Кроме того, МО может быть буквальным в интерпретации значений данных. Повторяющиеся записи могут привести к переоценке некоторых функций. Противоречия между функциями в обучающих данных могут привести к тому, что обученная модель будет работать ниже требований.

Распределение данных по каждому признаку можно рассматривать как меру согласованности. В некоторых случаях моделям машинного обучения требуются данные с нормальным распределением для удовлетворения требований к производительности.

Пример. Веб-форма используется для сбора информации о предпочтениях избирателей в отношении политических кандидатов. Организованная группа людей заполняет сайт записями своего любимого кандидата. При использовании для обучения модели машинного обучения эти повторяющиеся данные могут привести к тому, что модель переоценит конкретного кандидата при выводе для людей, чьи характеристики схожи с теми, кто заполнил веб-форму.

5.5.7.2 МК для согласованности

В таблице 10 представлены показатели качества данных для согласованности в конкретном контексте использования аналитики и машинного обучения.

Таблица 10 — Показатели согласованности

ID	Имя	Описание	Функция измерения	Целевые категории данных
Cons-I-1	Согласованность записи данных	Доля повторяющихся записей в наборе данных	X=A/B X= число повторяющихся записей в наборе данных B=Общее число записей в наборе данных	Тренировочные данные Данные проверки Данные тестирования Производственные данные

Окончание таблицы 10

Cons-I-2	Распределение значений данных	Статистическое распределение значений данных для данного свойства в наборе данных	Соответствующая мера распределения и функция измерения должны быть определены в соответствии с задачей МО.	Тренировочные данные Данные проверки Данные тестирования Производственные данные
Cons-I-3	Согласованность формата данных	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования Производственные данные
Cons-I-4	Семантическая согласованность	См. ISO/IEC 25024		Тренировочные данные Данные проверки Данные тестирования Производственные данные

5.5.8 Релевантность

5.5.8.1 Общие положения

Для целей настоящего документа релевантность относится к степени, в которой набор данных (при условии, что он является точным, полным, непротиворечивым, актуальным и т. д.) подходит для данного контекста.

Для МО релевантность может означать, что выбранные функции в обучающих данных и их значения данных являются хорошими предикторами для целевой переменной. Репрезентативность может быть связанной характеристикой набора данных, указывающей на его релевантность.

Пример 1. Модель МО используется для определения кредитоспособности людей. Обучающие данные являются репрезентативными для выборки генеральной совокупности, которая, как ожидается, появится в производственных данных. Данные обучения включают в себя соответствующие характеристики, такие как предыдущая кредитная история, доход, стаж работы и собственный капитал, которые являются хорошими предикторами кредитоспособности. Тренировочные данные также включают рост и вес каждого человека. Статистические тесты не показывают никакой связи роста и веса с предыдущей кредитной историей и считаются плохими предикторами будущей кредитной эффективности. Чтобы повысить общую релевантность набора данных, функции роста и веса удалены.

5.5.8.2 МК для релевантности

В таблице 11 представлены показатели качества данных для релевантности в конкретном контексте использования аналитики и машинного обучения.

Таблица 11 — Показатели релевантности

ID	Имя	Описание	Функция измерения	Целевые категории данных
Rel-I-1	Релевантность свойства	Соотношение свойств в наборе данных, релевантных данному контексту.	X=A/B A= Количество свойств в наборе данных, которые считаются релевантными в контексте использования данных. B=общее количество	Тренировочные данные Данные проверки Данные тестирования Производственные данные

			свойств в наборе данных	

Окончание таблицы 11

Rel-I-2	Релевантность записи	Соотношение записей в наборе данных, релевантных данному контексту.	X=A/B A= Количество записей в наборе данных, которые считаются релевантными в контексте использования данных. B=общее количество записей в наборе данных	Тренировочные данные Данные проверки Данные тестирования Производственные данные

Релевантность относится к степени, в которой набор данных (при условии, что он является точным, полным, непротиворечивым, своевременным и т. д.) подходит для выполнения поставленной задачи.

5.5.9 Своевременность

Своевременность возникает, когда скорость передачи данных достаточно высока, чтобы данные были актуальными.

Своевременность относится к задержке между временем, когда происходит явление, и временем, когда данные, записанные для этого явления, доступны для использования; эта характеристика качества данных особенно важна, когда набор данных представляет собой непрерывный поток данных.

5.5.10 Репрезентативность

5.5.10.1 Общие положения

ISO/IEC 20252 определяет репрезентативность как степень, в которой выборка отражает изучаемую целевую совокупность. Для контролируемого машинного

обучения набор обучающих данных можно рассматривать как выборку, а производственные данные — как изучаемую совокупность. Когда обучающие данные не в достаточной степени представляют производственные данные, обученная модель машинного обучения может не работать должным образом.

Пример 1. Система распознавания лиц, обученная только на изображениях людей со светлым оттенком кожи, может неправильно идентифицировать людей при применении к изображениям людей с темным оттенком кожи.

Пример 2. Система профилактического обслуживания, обученная только данным от электродвигателей, может неправильно предсказать необходимое обслуживание применительно к двигателям внутреннего сгорания.

5.5.10.2 МК для репрезентативности

В таблице 12 представлены показатели качества данных для репрезентативности в конкретном контексте использования аналитики и машинного обучения.

Таблица 12 — Показатели репрезентативности

ID	Имя	Описание	Функция измерения	Целевые категории данных
Rep-I-1	Коэффициент репрезентативности	Отношение релевантных признаков, обнаруженных у субъектов совокупности, к признакам, обнаруженным в выборке	X = A/B A= количество целевых атрибутов в образцах B= количество атрибутов в совокупности	Тренировочные данные Данные проверки Данные тестирования Производственные данные

5.5.11 Сбалансированность

5.5.11.1 Общие положения

Для набора данных баланс относится к распределению выборок по всем аспектам набора данных. Например, если набор данных представляет X категорий элементов, количество выборок в каждой категории должно быть равномерно распределено, чтобы этот набор данных был сбалансированным. Для набора данных изображения такие аспекты могут включать метки категорий, значимые для бизнес-логики, разрешение фигуры, яркость фигуры, отношение ширины к высоте помеченных ограничивающих рамок, размер помеченных ограничивающих рамок и любые другие параметры, которые потенциально влияют на производительность модели МО.

Баланс набора данных может частично повлиять на общую производительность модели машинного обучения. Для системы компьютерного зрения на основе МО следует учитывать баланс набора данных:

- Когда существуют значительные различия в яркости или разрешении между образцами обучающего набора данных и реальными данными, модели машинного обучения могут дать сбой из-за зашумленных данных, вызванных тусклостью или нечеткостью;
- В системе классификации на основе МО наличие несбалансированной категории выборки может привести к сбою обнаружения и классификации редких экземпляров; такие экземпляры могут быть даже ошибочно классифицированы или идентифицированы как зашумленные данные;
- В системе обнаружения объектов на основе машинного обучения значительные различия в отношении ширины к высоте или размере ограничивающих рамок могут привести к несоответствию размеров обнаруживаемых объектов при фиксированном размере рецептивного поля; следовательно, это также может привести к потере обобщаемости, если не применяются дополнительные подходы к проверке или корректировке многоразмерных объектов.

5.5.11.2 МК для сбалансированности

В таблице 13 представлены показатели качества данных для баланса в конкретном контексте использования аналитики и машинного обучения.

Таблица 13 — Показатели сбалансированности

ID	Имя	Описание	Функция измерения	Целевые категории данных
Bal-I-1	Сбалансированность яркости	Обратная величина максимального отношения разницы яркости образца изображения к усредненной яркости образцов в наборе данных.	$\mathrm{X} = \mathrm{MAX}(\frac{\lvert A - B \rvert}{B})^{-1}$ MAX (•) = функция, находящая максимальное значение A = яркость изображения B = средняя яркость	Тренировочные данные Данные проверки Данные тестирования
Bal-I-2	Сбалансированность разрешения	Обратная величина максимального отношения разницы разрешения образца изображения к усредненному разрешению в наборе данных.	$\mathrm{X} = \mathrm{MAX}(\frac{\lvert A - B \rvert}{B})^{-1}$ MAX (•) = функция, находящая максимальное значение A = разрешение изображения B = среднее разрешение	Тренировочные данные Данные проверки Данные тестирования

Продолжение таблицы 13

Bal-I-3	Сбалансированность изображений между категориями	Обратная величина максимального отношения разницы размера категории (количества содержащихся выборок) к усредненному размеру категории набора данных.	$X = \mathrm{MAX}(\frac{\lvert A - B \rvert}{B})^{-1}$ MAX (•) = функция, находящая максимальное значение A = размер категории в наборе данных B = средний размер категории в наборе данных	Тренировочные данные Данные проверки Данные тестирования
Bal-I-4	Сбалансированность соотношения высоты и ширины ограничительной рамки	Обратная величина максимального отношения разницы отношения высоты и ширины ограничительной рамки к усредненному отношению высоты ограничительной рамки к ширине в образцах в наборе данных.	$X = \mathrm{MAX}(\frac{\lvert A - B \rvert}{B})^{-1}$ MAX (•) = функция, находящая максимальное значение A = ограничивающая рамка с отношением высоты к ширине выборки в наборе данных B = ограничивающая рамка с отношением высоты к ширине во	Тренировочные данные Данные проверки Данные тестирования

			всем наборе данных	

Окончание таблицы 13

Bal-I-5	Сбалансированность площади ограниченной рамками категории	Обратная величина максимального отношения усредненной площади ограничительной рамки категории к усредненной площади ограничительной рамки всех выборок в наборе данных.	$X = \mathrm{MAX}(\frac{\lvert A - B \rvert}{B})^{-1}$ MAX (•) = функция, находящая максимальное значение A = усредненная площадь ограничительной рамки категории в наборе данных B = усредненная площадь ограничительной рамки всех элементов в наборе данных	Тренировочные данные Данные проверки Данные тестирования
Bal-I-6	Сбалансированность площади ограничительной рамки образца	Обратная величина максимального отношения площади ограничивающей рамки образца к усредненной площади ограничивающей рамки всех образцов в наборе данных	$X = \mathrm{MAX}(\frac{\lvert A - B \rvert}{B})^{-1}$ MAX (•) = функция, находящая максимальное значение A = площадь ограничивающей рамки образца B = усредненная площадь ограничивающей рамки всех образцов в наборе данных	Тренировочные данные Данные проверки Данные тестирования

5.5.12 Подобие

5.5.12.1 Общие положения

Сходство набора данных относится к сходству между образцами с точки зрения интересующих функций. Это относится к задачам классификации (см. ISO/IEC 23053 —, 6.2.3 [16]), которые обычно реализуются с использованием контролируемого обучения (см. ISO/IEC 23053 —, 7.2). Это также относится к задачам кластеризации (см. ISO/IEC 23053 —, 6.2.4), которые обычно реализуются с использованием обучения без учителя (см. ISO/IEC 23053 —, 7.3). Как задачи классификации, так и задачи кластеризации требуют адекватного уровня различий между образцами для успешного выполнения (см. ISO/IEC TR 24027:2022, 5.2 [17]).

Модель машинного обучения, обученная на наборе данных, содержащем очень похожие изображения (например, созданные путем небольшого сдвига пикселей на основе нескольких исходных изображений), может иметь риск переобучения и, следовательно, меньшую обобщаемость. В этом случае можно рассмотреть применение подходов улучшения данных, таких как вращение, сдвиг и т. д., которые могут улучшить обобщаемость модели МО. Эти подходы не могут работать, если количество исходных изображений ограничено. В этом случае следует проверить долю однородных образцов. Другой подход заключается в рассмотрении алгоритмов кластеризации с методами уменьшения дрейфа темы [7].

Дальнейшие меры определяют сходство данных с помощью геометрического подхода: т. е. набор данных из N кортежей и M атрибутов может быть представлен, например, как N векторов в M-мерном пространстве, поэтому его можно анализировать и сравнивать с использованием инструментов геометрии; в частности, подобие может быть связано с взаимным положением векторов в пространстве.

Примечания

1 Для кортежа см. определение в ISO/IEC 20944–1:2013 «часть отношения, однозначно описывающая экземпляр объекта и его атрибуты».

2 Кортеж может быть представлен одной строкой таблицы отношений.

3 Набор векторов, являющийся ортогональным базисом, рассматривается как набор данных с наименьшим сходством.

5.5.12.2 МК для подобия

Таблица 14 представляет набор мер для сходства в конкретных приложениях аналитики и машинного обучения.

Таблица 14 —Показатели подобия

ID	Имя	Описание	Функция измерения	Целевые категории данных
Sim-I-1	Подобие элементов	Соотношение похожих элементов в наборе данных. Чем ниже, тем лучше.	X = (A - B) / A A = количество всех элементов в наборе данных B = количество кластеров элементов, обработанных алгоритмом кластеризации	Тренировочные данные Данные проверки Данные тестирования
Sim-I-2	Плотность элементов	Плотность нормализованного набора данных	X= A-B= λ_{max} - λ_{min} λ_{max}, λ_{min} дают max, min eigenvalue значения G(MxM) где G = $\Phi_{norm}{}^{T}\,\Phi_{norm}$ и Φ_{norm} (NxM) нормализованный набор данных (Примечание 2)	Набор данных с N векторами и M атрибутами Значение данных
	Примечания 1 X=0 означает наименьшее сходство. 2. Фnorm вычисляется из Ф(NxM) после вычитания из каждой строки его среднего значения и нормализации до 1. Визуально нормализованные данные опираются на поверхность гиперсферы радиуса = 1 с центром в начале координат (M<=N) [6].			

Окончание таблицы 14

Sim-I-3	Независимость образцов	Отношение результата анализа метода главных компонент к размерности набора данных	X= 1-A\B= 1-K\N K<=N размер, найденный методом PCA (Примечание 2)	Набор данных Φ с N векторами и M атрибутами (Примечание 5) Значение данных
	Примечания 1 X=0 означает наименьшее сходство. 2 K — число главных компонент. K<=M — наименьшее число собственных значений матрицы C (ПРИМЕЧАНИЕ 3), начиная с наибольшего, выбранное таким образом, чтобы представлять 95% их суммы [5]. 3 C(MxM) = ΦmeanT Φmean, где Φmean вычисляется из Φ(NxM) после вычитания из каждой строки его среднего значения. Визуально нормализованные данные Φmean соответствуют (гипер)эллипсоиду с собственными векторами в качестве оси и с центром в начале координат. 4 Главные компоненты могут быть выбраны с критериями и/или процентным соотношением, отличными от указанных выше; см. Приложение B для изменения меры. 5 Набор данных из M изображений в градациях серого, каждое из которых имеет nxn пикселей, может быть представлен как N=nxn кортежей (векторов) с M атрибутами (размерами).			

5.5.13 Разнообразие

5.5.13.1 Общие положения

Разнообразие набора данных относится к разнице между образцами с точки зрения целевых данных. Это относится к задачам классификации (см. ISO/IEC 23053, 6.2.3), которые обычно реализуются с использованием контролируемого обучения (см. ISO/IEC 23053, 7.2). Это также относится к задачам кластеризации (см. ISO/IEC 23053, 6.2.4), которые обычно реализуются с использованием обучения без учителя

(см. ISO/IEC 23053, 7.3). И классификация, и задачи кластеризации требуют адекватного уровня различий между образцами для успешного выполнения (см. ISO/IEC TR 24027:2022, 5.2).

В наборе данных, используемом для модели МО, важна адекватная разница между выборками. Если все или большинство записей данных в наборе данных одинаковы, модель МО, полученная из этого набора данных, может иметь риск переобучения и, следовательно, быть менее обобщаемой. Подходы к улучшению данных могут улучшить обобщаемость модели МО, но эти подходы могут потерпеть неудачу, если разнообразие исходного набора данных ограничено. Разнообразие тесно связано с репрезентативностью и сбалансированностью. Это характеристика качества набора данных, которую можно использовать для оценки точности набора данных.

Измерение разнообразия может быть выполнено в контексте конкретных целевых данных, как это определено требованиями задачи ML.

6 Методы улучшения качества данных

6.1 Дополнение данных

В некоторых случаях набор данных, используемый для машинного обучения, не может адекватно представить совокупность, подвергающуюся анализу и прогнозированию: типичным примером является недостаточная или избыточная выборка определенных меток с последующими несбалансированными данными. Это может быть результатом исторической дискриминации, усилий по защите конфиденциальности или естественного небольшого количества образцов с заданной характеристикой.

Неспособность набора данных представить совокупность при использовании машинного обучения может привести к тому, что модель машинного обучения плохо обобщает производственные данные или получает разброс классификаций по классам данных [1] с соответствующими социальными, юридическими и этическими факторами особенно, когда объектами классификаций являются физические лица [2].

Расширение данных увеличивает количество или разнообразие данных за счет манипулирования исходными данными, таких как изменение, добавление,

преобразование, вменение или операции объединения. Подробные методы и процессы улучшения качества данных за счет расширения данных описаны в ISO/IEC 5259–4, пункт 7.5.1.

6.2 Стандартизация данных

Для достижения заданной производительности некоторых алгоритмов машинного обучения данные должны иметь нормальное распределение. Стандартизированные преобразования могут изменить форму набора данных, чтобы он выглядел как имеющий нормальное распределение.

Примеры стандартизированных преобразований включают:

- Стандартный скаляр — это преобразование создает набор данных со средним значением, равным нулю, и стандартным отклонением, равным единице; использование стандартного скаляра обычно не рекомендуется для разреженных данных;
- MinMax Scaler — это преобразование масштабирует каждое значение, чтобы оно находилось в заданном диапазоне от 0 до 1;
- MaxAbs Scalar — это преобразование масштабирует каждое значение в диапазоне от 1 до -1; это преобразование можно использовать с разреженными данными;
- Надежный скаляр — это преобразование масштабирует каждое значение так, чтобы оно находилось в межквартильном диапазоне, т. е. между первым квартилем и третьим квартилем. Это преобразование можно использовать для наборов данных со многими выбросами.

6.3 Нормализация данных

Наборы данных с очень большими диапазонами или сочетанием малых и больших диапазонов могут привести к тому, что алгоритмы машинного обучения будут вычислять большие веса или чрезмерно выделять признаки с большими диапазонами.

Нормализованные преобразования можно использовать для установления общего диапазона в наборе данных путем масштабирования значений до единичной нормы, то есть от 0 до 1.

6.4 Вменение данных

Некоторые алгоритмы машинного обучения предполагают, что все данные должны быть полными и иметь смысл. Однако в наборах данных часто есть отсутствующие значения, которые могут быть представлены пробелами, NaN (не числом) или каким-либо другим заполнителем.

Для структурированных данных (например, табличных данных в формате CSV) одним из способов устранения пропущенных значений является удаление всей строки или столбца с пропущенным значением. Это может привести к удалению важных функций или созданию разреженного набора данных, которого недостаточно для обучения, проверки и тестирования модели машинного обучения.

Другой подход заключается в использовании статистических методов для условного исчисления соответствующих значений. Основные статистические методы вменения включают использование среднего, медианы, наиболее часто встречающегося значения или определенной константы, т. е. фиксированного значения.

Преобразования для статистического вменения включают:

- SimpleImputer — это преобразование можно запрограммировать на замену отсутствующих значений средним, медианным, наиболее частым или постоянным значением. Среднее значение и медиана могут использоваться только с числовыми значениями, в то время как наиболее частое и постоянное можно использовать как со строками, так и с числовыми значениями;
- IterativeImputer — это многомерное преобразование вменения аппроксимирует замену отсутствующих значений путем итеративного анализа значений для всех признаков в наборе данных.

6.5 Шифрование данных

Алгоритмы машинного обучения обычно предполагают, что данные будут представлены в числовой форме, тогда как категориальные данные часто выражаются с помощью текстовых строк. Категориальные данные могут быть преобразованы в числовые данные с помощью кодирования.

Преобразования для кодирования категориальных данных включают:

- OrdinalEncoder — это преобразование преобразует набор категориальных значений в порядковый набор целых чисел, то есть от 1 до n. Некоторые алгоритмы машинного обучения чрезмерно подчеркивают порядковый номер результата при вычислении весов в модели ML;
- LabelEncoder — это преобразование можно использовать для преобразования нечисловых меток целевой переменной в числовые категории;
- OneHotEncoding — это преобразование преобразует набор категориальных значений в массив с категориальными метками, представленными в виде функций, и двоичными значениями, указывающими, подходит ли запись под данную функцию;
- Get Dummies — это преобразование также преобразует категориальные значения в массив функций и двоичных индикаторов. В этом случае функции называются фиктивными переменными.

7 Отчет о качестве данных

7.1 Информация о показателях качества данных

ISO 8000-8:2015, Приложение D, описывает модель деятельности для измерения качества информации и данных. Модель деятельности была изменена для аналитики и машинного обучения и приведена на рис. 3

Рисунок 3 — Информация о показателях качества данных в источнике данных

7.2. Руководство для организаций

Отчеты о качестве описывают, как организации применяют каждую меру качества данных для своих требований к качеству, и предоставляют оценку с конкретными измерениями. Любая организация, заявляющая о соответствии этому документу для выбранного экземпляра реализации предполагает, что:

а) выбраны характеристики качества данных, которые должны быть указаны или оценены, из этого документа или выбраны альтернативные характеристики качества данных, если это целесообразно;

б) указан один или несколько целевых данных, для которых измеряются выбранные характеристики качества данных;

в) выбраны соответствующие меры качества данных из этого документа или выбраны альтернативные меры качества данных, соответствующие каждой характеристике качества данных, относящейся к указанным целевым данным;

г) содержит обоснование любых изменений в выбранных характеристиках или показателях качества данных;

д) перечислены любые дополнительные показатели качества или используемые элементы показателей качества, которые не включены в настоящий документ.

Порядок пунктов а) и б) может применяться в обратном порядке.

При использовании модифицированных или новых показателей качества данных пользователь должен указать целевые данные, функцию измерения и соответствующие характеристики качества данных. В этом документе не приводится полный список показателей качества данных в течение жизненного цикла качества данных. Организация может определить другие меры качества данных в зависимости от контекста.

Библиография

[1]	ГОСТ Р ИСО/МЭК 20546-2021	Информационные технологии. Большие данные. Обзор и словарь
[2]	ИСО/МЭК 22989:2022	Информационные технологии — Искусственный интеллект — Понятия и терминология искусственного интеллекта
	ISO/IEC 22989:2022	Information technology — Artificial intelligence — Artificial intelligence concepts and terminology
[3]	ИСО/МЭК 5259–1:202X	Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, терминология и примеры
	ISO/IEC 5259–1:202X	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples
[4]		

Библиография оригинала

[1] Barocas, S. and Selbst, A. D. Big data's Disparate Impact. *California Law Review*. 2016, **104**(3), 671-732. doi:10.2139/ssrn.2477899

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

[3] Vetrò, A., Torchiano, M. and Mecati, M. A data quality approach to the identification of discrimination risk in automated decision making systems. *Government Information Quarterly*. 2021, **38**(4), 101619. doi :10.1016/j.giq.2021.101619

[4] D. Natale, M.C. Paoletti, A Simonetta. Data Quality and Statistical Information. *Journal of accidents and occupational diseases* [online]. 2021 [viewed 2022-04-01]. Available at: https://www.inail.it/cs/internet/docs/alg-ass-stat-la-qualita-dei-dati-e-l-informazione-statistica.pdf

[5] Trenta A. ISO/IEC 25000 quality measures for A.I.: a geometrical approach. in Proceedings *APSEC IWESQ 2020* [online]. 2020, 20-21 [viewed 2022-04-01]. Available at: http://ceur-ws.org/Vol-2800/paper-05.pdf

[6] Trenta, A. Data bias measurement: a geometrical approach through frames. in Proceedings *APSEC IWESQ 2021* [online]. 2021, 11-16 [viewed 2022-04-01]. Available at: http://ceur-ws.org/Vol-3114/paper-03.pdf

[7] Webb, G.I., Lee, L.K., Goethals, B. et al. Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery*. 2018, **32**, 1179–1199. doi:10.1007/s10618-018-0554-1

[8] ISO/IEC 5259-3: —, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines*

[9] ISO/IEC 5259-4: —, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework*

[10] ISO 8000 (all parts), *Data quality*

[11] ISO 8000-1:2022, *Data quality — Part 1: Overview*

[12] ISO 8000-2:2020, *Data quality — Part 2: Vocabulary*

[13] ISO 8000-8:2015, *Data quality — Part 8: Information and data quality: Concepts and measuring*

[14] ISO/IEC/IEEE 15939:2017, *Systems and software engineering — Measurement process*

[15] ISO/IEC 19505-1:2012, *Information technology — Object Management Group Unified Modeling Language (OMG UML) — Part 1: Infrastructure*

[16] ISO/IEC 23053:2022, *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

[17] ISO/IEC TR 24027:2022, *Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI-aided decision making*

[18] ISO/IEC 25000 (all parts), *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE)*

[19] ISO/IEC 25010:2011, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*

[20] ISO/IEC 25012:2008, *Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model*

[21] ISO/IEC 25020:2019, *Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Quality measurement framework*

[22] ISO/IEC 25021:2012, *Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Quality measure elements*

[23] ISO/IEC 25024:2015, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality*

[24] ISO/IEC 42001: —, *Information Technology — Artificial intelligence — Management system*

[25] JCGM 200:2012, *International vocabulary of metrology — Basic and general concepts and associated terms (VIM)*

[26] Aroyo, L., Lease, M., Paritosh, P. and Schaekermann, M. Data Excellence for AI: Why should you care? *ACM Interactions*. XXIX.2, 66. Available from: https://interactions.acm.org/archive/ view/march-april-2022/data-excellence-for-ai

Приложение А
(справочное)
Сведения о соответствии ссылочных международных стандартов национальным стандартам

Таблица А

Обозначение ссылочного международного стандарта	Степень соответствия	Обозначение и наименование соответствующего национального стандарта
ISO/IEC 22989:2022, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology	NEQ*	ПНСТ 553–2021** Информационные технологии. Искусственный интеллект. Концепции искусственного интеллекта и терминология (ISO/IEC DIS 22989, NEQ)
*NEQ — неэквивалентный стандарт. ** В России принят предварительный национальный стандарт, основанный на ранней редакции международного стандарта		

ПРИЛОЖЕНИЕ В

(справочное)

Разработка и документирование функции измерения

<Примечание редактора N1440(IT-06): убедитесь в наличии авторских прав>

Некоторые из мер, перечисленных в серии ISO 25000 и в этом документе, определены с общей точки зрения. Для практического применения в соответствии со стандартом ISO 25020 требуется подробный проект функции измерения и ее контекстная информация.

В этом примере показано, как спроектировать и задокументировать функцию измерения с мерой Acc-I-1 из ISO 25024 для синтаксической точности, соответствующей этому документу.

Таблица В.1 определяет показатель Acc-I-1 из ИСО/МЭК 25024:

Таблица В.1 — Синтаксическая точность Acc-I-1

ID	*Имя*	*Описание*	*Функция измерения*	*Жизненный цикл данных (DLC)* *Целевые объекты* *Характеристики*
Acc-I-1	Синтаксическая точность	Отношение близости значений данных к набору значений, определенных в предметной области	X=A/B A=количество элементов данных, соответствующие значения которых синтаксически точные B=количество элементов данных, для которых требуется	Все DLC, кроме Проектирование данной Файл данных Элемент данных, значение данных

			синтаксическая точность	
ПРИМЕЧАНИЕ 1. Одно значение считается «синтаксически точным», если оно совпадает со значением из идентифицированного источника проверенной информации: результат «да» или «нет». ПРИМЕЧАНИЕ 2. Примером низкой степени синтаксической точности является случай, когда слово Mary хранится как Marj. ПРИМЕЧАНИЕ 3. В статистическом анализе результат представляет собой близость значений данных к истинным значениям, определенным в области. ПРИМЕЧАНИЕ 4. См. ИСО/МЭК 25012:2008, 5.3.1.1. ПРИМЕЧАНИЕ 5. Для A и B могут быть выбраны различные варианты, например. A как сумма расстояний Джаро-Винклера от домена, B как размерность массива строк.				

Для определенных целей (например, когда для оценки необходимо сравнение) определения меры Acc-I-1 на рисунке B.1 недостаточно, и необходимо получить представление о «синтаксически точных значениях» для A и «количестве данных». элементы, для которых требуется синтаксическая точность» для B, то есть выбрать метод измерения расстояния между строками для A и определить домен для B. Существует несколько возможностей для определения как A, так и B, и среди них следующие:

— A, так как количество вхождений условия расстояние равно нулю от строки против всех строк в домене B;

— B как количество допустимых строк.

С этими определениями пересмотренная мера соответствия Acc-I-1-IT-2 определена на рисунке B.2:

Таблица B.2 — Синтаксическая точность Acc-I-1-IT-2

ID	*Имя*	*Описание*	*Функция измерения*	*Жизненный цикл данных (DLC)* *Целевые объекты* *Характеристики*

Acc-I-1-IT-2	Синтаксическая точность	Отношение близости значений данных к набору значений, определенных в области	X=1-A/B A=количество значений данных, для которых расстояние от области равно нулю B= количество значений в области	Все DLC, кроме Проектирование данной Файл данных Элемент данных, значение данных
ПРИМЕЧАНИЕ 1. Лучшее сходство между сравниваемыми строками и их областью будет тогда, когда расстояние меньше, поэтому X ниже — лучше. ПРИМЕЧАНИЕ 2. ID включает дополнительную часть «IT-2», см. ИСО/МЭК 25020:2019, приложение С.				

В таблице B.3 подробно описано применение меры Acc-I-1-IT-2 для сравнения точности двух баз данных, каждая из которых содержит 3 имени. Сравнение производится с синтаксисом, состоящим из имен длины 4. Имена длины 4 являются подмножеством всех возможных строк длины 4.

Из результатов X мы имеем, что значения в базе данных R более точны, чем значения в базе данных W, так как имя Marj не принадлежит синтаксису имен.

Таблица B.3 — Пример показателей синтаксической точности Acc-I-1-IT-2

	База данных R	База данных W
	John	John
	Jack	Jack
	Mary	Marj
X=1-A/B	$1-\frac{3}{26^4-Not_adm}$	$1-\frac{2}{26^4-Not_adm}$
где: Not_adm = набор значений, не принадлежащих синтаксису 26^4- Not_adm = набор значений, принадлежащих синтаксису		

ПРИЛОЖЕНИЕ С

(справочное)

UML-модель системы измерения качества данных

Унифицированный язык моделирования (UML) — это стандартизированный графический язык для визуализации, спецификации, построения и документирования

артефактов программных систем, описанный в ISO 19505–1. На рисунке С.1 показана диаграмма классов структуры показателей качества данных, чтобы обеспечить общий словарь и отношения между моделями качества данных и показателями в этом документе. Подробную информацию о нотации UML см. в ISO 19505–1.

Каждый отдельный класс здесь описан в разделе 5.2.

ПРИЛОЖЕНИЕ D

(справочное)

Обзор характеристик качества данных и наборов данных

<Примечание редактора N1928(US-02): необходимо пересмотреть, что является более подходящим, «уровень данных» или «уровень элемента данных», а также категории верхнего уровня «ремонтопригодность», «валидность», «надежность», и «верность».>

<Примечание редактора N1976(IT02): - текст не соответствует стандартам ISOIEC 25010, 25012, 25024 и 5259-1, 5259-2.

- предлагаемые «категории качества данных» не встречаются в документе US16; в документе ремонтопригодность, валидность, надежность и точность определяются как «ключевые свойства», а не как категории; документ, по-видимому, ориентирован на качество процессов и управление, а не на качество продукта, которое вместо этого включено в сферу действия 5259-2. ;

- предложение US-02 в основном предполагает агрегирование характеристик качества данных. Но, например, категория «сопровождаемость», используемая в стандартах разработки программного обеспечения, связана с программным обеспечением, а сопровождаемость в 25010 требует некоторых подхарактеристик продукта, которые облегчают и позволяют выполнять

ПРИМЕЧАНИЕ — Сопровождаемость, достоверность, надежность и точность не являются характеристиками сами по себе и не связаны с теми же или разными терминами, используемыми в ISO 25010. /IEC 25010), но это связано с прозрачностью данных (соответствие, подотчетность, документация, идентификация происхождения).

В предложении US-02 неясно, как (суб)характеристики качества данных могут облегчить возможность их обслуживания (действия обслуживания).

Поддержание и улучшение качества данных является общей целью руководства, управления и процесса, а не характеристикой данных как продукта.

Предлагается найти модифицированный термин сопровождаемость (с новым определением) или обсудить вклад в другой стандарт или часть 5259, отличную от 5259-2, относящуюся к процессу, управлению управлением данными.

Либо предлагается описать агрегирование или сопоставление характеристик в новом информативном Приложении 5259-2.>

Чтобы представить взаимосвязь между элементами данных и общими характеристиками качества набора данных, полезно сгруппировать их по четырем точкам зрения: Сопровождаемость, достоверность, надежность и точность.

ПРИМЕЧАНИЕ — Сопровождаемость, достоверность, надежность и точность не являются характеристиками сами по себе и не связаны с теми же или разными терминами, используемыми в ISO 25010. /IEC 25010), но это связано с прозрачностью данных (соответствие, подотчетность, документация, идентификация происхождения).

— Сопровождаемость: Сопровождаемость относится к свойству набора данных (на основе передового опыта жизненного цикла разработки программного обеспечения), которое указывает на уровень соответствия методам разработки набора данных и уровень прозрачности для поддержки принятия решений и подотчетности с набором данных. Это относится к существованию и актуальности документации, определяющей происхождение версий набора данных и право собственности, а также к механизму контроля версий. Это группа характеристик качества для переносимости, понятности, аудируемости и идентифицируемости целевых данных как элементов данных и наборов данных, а также для масштабируемости всего набора данных.
— Достоверность: Достоверность набора данных относится к эксплуатационной достоверности набора данных, т. е. к тому, насколько хорошо он описывает (в целом или на уровне элементов) явления, для которых он предназначен, и какова его актуальность. Например, учитывает ли набор данных всю сложность и спектр субъективности явлений, на которые он нацелен, допускает ли он обобщение в рамках явлений и каковы справочные материалы для этой части состояния? Искусство и доступность. Достоверность является всеобъемлющей группой как для элемента данных, так и для набора данных актуальности.
— Надежность: Надежность набора данных относится к внутренней достоверности набора данных, т. е. к его достоверности, вытекающей из его непротиворечивости, воспроизводимости и воспроизводимости в целевых данных элементов данных и наборов данных. Надежность сильно коррелирует с сопровождаемостью, потому что наборы данных, которые не поддерживаются должным образом, часто невозможно воспроизвести или проверить на согласованность. Кроме того, показатели точности и

прецизионности могут использоваться для определения аспектов надежности набора данных.

— Точность: точность набора данных относится к характеристикам полноты, непротиворечивости, сбалансированности и репрезентативности набора данных по отношению к явлению, на которое он нацелен. Как правило, способы выборки данных из более крупных источников или выбор самих источников могут повлиять на точность результирующего набора данных с точки зрения его разнообразия и включения возможных систематических ошибок, например. временные, географические, культурные и другие социальные предубеждения. Точность также связана с надежностью с точки зрения таких свойств, как актуальность и своевременность.

В таблице D.1 показана группа характеристик качества данных. Качество набора данных можно измерить как сочетание двух аспектов: (1) качества отдельных элементов и (2) совокупного показателя качества всего набора данных. Отсюда и разделение характеристик качества (в таблице 1) в элементе данных и наборе данных. Каждая характеристика качества может применяться к целевым данным элемента данных или набора данных.

Таблица D.1 — Группа характеристик качества данных и наборов данных

Группы характеристик	Качественные характеристики	Целевые данные	
		Элемент данных	Набор данных
Сопровождаемость	Портируемость	✓	✓
	Понятность	✓	✓
	Возможность аудита	✓	✓
	Идентифицируемость	✓	x
	Масштабируемость данных	x	✓
Действительность	Современность	✓	✓
	Обобщаемость	x	✓
	Эффективность	x	✓
Надежность	Достоверность	✓	✓
	Правильность	x	✓
	Точность	x	✓
Верность	Полнота	✓	x

	Последовательность Актуальность Своевременность Репрезентативность Баланс Сходство Разнообразие	x x x x x x x	✓ ✓ ✓ ✓ ✓ ✓ ✓

ПРИЛОЖЕНИЕ Е

(справочное)

Меры качества для синтетических данных

Чтобы быть полезными, синтетические данные должны походить на реальные данные с точки зрения математических и статистических свойств, а также иметь разумный уровень баланса в появляемости классов их атрибутов. Весь набор данных может состоять из синтетических данных, или синтетические данные могут использоваться для дополнения набора данных, состоящего в основном из реальных данных.

Одним из примеров синтетических данных является создание изображений людей с зелеными глазами из фотографий людей с голубыми глазами. При добавлении в набор данных, используемый для обучения моделей распознавания изображений, синтетические данные сделают исходный набор данных более сбалансированным и помогут модели лучше распознавать людей с зелеными глазами при прогнозировании производственных данных. Уровень достигнутого баланса можно измерить с помощью существующих метрик из научной литературы.

Учитывая контекст Приложения А и его роль в машинном обучении, его можно рассматривать как потенциальный дополнительный элемент «внутренней точки зрения» ISO 25012 на качество данных. В частности, меры связаны с качественными характеристиками точности и полноты атрибутов данных, важных аспектов, подчеркнутых в ISO 8000–2, ISO 25012, ISO 25024. Эти меры также связаны с качеством. характеристика репрезентативности.

Примеры специализированных мер качества данных для больших данных и машинного обучения, которые полезны для синтетических данных, описаны в приложении В

ГОСТ (Р) _____—_____.
(проект)

УДК 004.01:006.354 ОКС 35.020

Ключевые слова: информационные технологии, искусственный интеллект, рамочная структура, качество данных, характеристики качества данных, большие данные, набор данных, методы определения качества данных, модель данных, подготовка наборов данных, вывод данных из эксплуатации.

Председатель совета директоров ООО «Институт развития информационного общества», руководитель разработки стандарта		Ю.Е. Хохлов
Руководитель дирекции развития цифровых технологий ООО «Институт развития информационного общества», профессор, Санкт-Петербургский государственный университет, разработчик стандарта		А.В. Богданов