Коробейников В.С., Малахов С.В. Определение характеристик качества данных и степень их применимости // Международный журнал информационных технологий и энергоэффективности. – 2024. – Т. 9 № 8(46) с. 5-9

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.6

ОПРЕДЕЛЕНИЕ ХАРАКТЕРИСТИК КАЧЕСТВА ДАННЫХ И СТЕПЕНЬ ИХ ПРИМЕНИМОСТИ

[1]Коробейников В.С., Малахов С.В.
ФГБОУ ВО «ПОВОЛЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТЕЛЕКОММУНИКАЦИЙ И ИНФОРМАТИКИ», Самара, Россия (443010, г. Самара ул. Льва Толстого, 23), e-mail: [1]vlad.k.k78@gmail.com

Оценка качества данных в разрезе характеристик является эффективным подходом к мониторингу за качеством данных организации в рамках всего набора данных, а также их атрибутов.

Ключевые слова: Управление качеством данных; качество данных; большие данные; характеристики качества данных; оценка качества данных.

DETERMINING DATA QUALITY CHARACTERISTICS AND THE EXTENT OF THEIR APPLICABILITY

[1]Korobeynikov V.S., Malakhov S.V.
VOLGA REGION STATE UNIVERSITY OF TELECOMMUNICATIONS AND INFORMATICS, Samara, Russia (443010, Samara, Lva Tolstogo str., 23), e-mail: [1]vlad.k.k78@gmail.com

Assessing data quality by characteristics is an effective approach to monitoring the quality of an organization's data across its entire data set, as well as its attributes.

Keywords: Data quality management; data quality; big data; data quality characteristics; data quality assessment.

Концепция искусственного интеллекта впервые появилась в 1950 году в работе Алана Тьюринга "Вычислительная техника и интеллект", но, несмотря на это алгоритмы машинного обучения повсеместно стали применяться относительно недавно. Машинное обучение основывается на работе с большими данными, которые должны соответствовать требованиям к качеству данных для поставленных целей. Проблема качества данных является достаточно актуальной, поскольку в машинном обучении и аналитике качество полученного решения напрямую зависит от входных данных.

Существуют международные стандарты, содержащие информацию, относящуюся к качеству данных. Международная серия стандартов ISO 8000 [3] была впервые предложена в 2002 году, а часть про качество данных была утверждена в 2017 году, однако этот стандарт разработан с точки зрения управления и производства данных и соответствует лишь характеристикам необходимым для определенной цели, а данные исходя из этого могут требовать предварительной обработки для повторного использования в целях аналитики и моделирования. Стандарт ISO\IEC 11179 [4] описывает принципы и процедуры для описания данных чтобы обеспечить единообразие и совместимость данных, ISO\IEC 20943 [3] описывает возможность обмена информацией между системами и установления

семантической согласованности данных, ISO/IEC 25012 [3] определяет меры качества данных для количественного измерения качества данных с точки зрения их характеристик. Также есть разрабатываемая серия международных стандартов ISO/IEC WD 5259 [4], в которой определены характеристики и показатели качества данных, а также требования к управлению качеством данных в течение жизненного цикла данных.

Управлением качества данных в крупных организациях занимается директор по данным (Chief Data Officer - CDO). Как правило, это заключается в организации методов и обеспечении контроля за качеством данных промышленных источников, а также развитии культуры качества данных. Исходя из описанных стандартов серии ISO/IEC можно выделить следующие этапы по управлению качеством данных:

- Модель качества данных – это набор характеристик, обеспечивающий основу для требований к качеству данных и ее оценке на всем жизненном цикле данных. При влиянии характеристик качества данных друг на друга могут возникать противоречия, в этом случае требования к качеству данных следует ранжировать по важности и уровню воздействия на данные;
- Меры качества данных – это критерии оценки характеристик модели качества данных и определение уровней их приемлемости;
- Оценка качества данных – это результаты оценки качества данных, а также решение о соответствии или несоответствии уровня качества. Включает профилирование данных, состоящее в анализе распределений атрибутов данных в том числе с использованием статистических методов, а также измерение качества данных в рамках выбранных мер качества данных;
- Улучшение качества данных – подходы к повышению уровня качества данных;
- Отчетность – составление отчетности с информированием о результатах управления качеством данных.
- Характеристики качества данных могут определяться как для атрибутов, так и для всего набора данных. При этом характеристики можно объединить в группы [2] такие, как:
- Сопровождаемость – характеристики, касающиеся описания и соответствия подходам к разработке данных;
- Надежность – характеристики, рассматривающие данные в качестве внутренней достоверности данных исходя из их согласованности и воспроизводимости. Связана с сопровождаемостью поскольку плохо поддерживаемые данные может быть трудно воспроизвести и проверить на согласованность;
- Верность – характеристики, рассматривающие данные в плане полноты и репрезентативности. Также коррелирует с сопровождаемостью с точки зрения актуальности и своевременности;
- Действительность – характеристики, рассматривающие данные в рамках целей для которых они предназначены.
- В Таблице 1 ниже приведен пример разбивки по группам характеристик качества данных и их применимости.

Таблица 1 – Разбивка по группам характеристик качества данных и их применимости

Группа	**Характеристика**	**Применимость**	
		атрибуты данных	**набор данных**
Сопровождаемость	Переносимость	да	да
	Понятность	да	да
	Контролируемость	да	да
	Идентифицируемость	да	нет
	Доступность	нет	да
	Масштабируемость	нет	да
Надежность	Восстанавливаемость	нет	да
	Точность	нет	да
	Достоверность	да	да
	Согласованность	нет	да
	Представимость	нет	да
Верность	Актуальность	нет	да
	Полнота	да	нет
	Своевременность	нет	да
	Репрезентативность	нет	да
	Сбалансированность	нет	да
	Подобие	нет	да
	Разнообразие	нет	да
Действительность	Эффективность	нет	да
	Релевантность	да	да
	Обобщаемость	нет	да

Приведенные характеристики качества данных могут зависеть друг от друга, иметь несколько другие трактовки и описание мер качества данных в зависимости от различных стандартов и источников.

В модели качества данных и при анализе необязательно использовать все приведенные характеристики, кроме того некоторые характеристики качества данных зависят друг от друга и могут быть разделены на подхарактеристики. Такие характеристики, как репрезентативность, подобие, разнообразие, а также эффективность и релевантность наиболее распространены и применимы при решении конкретных задач машинного обучения. Итоговые результаты в разрезе характеристик и проверок можно интерпретировать на попадание в зоны материальности по принципу сигналов, где красный - критичный уровень, желтый - приемлемый уровень, зеленый - хороший уровень качества данных. Критерии качества данных и их материальность могут варьироваться в зависимости от данных и цели их применения. Итоговым показателем качества данных является интегральный показатель качества данных (ИПКД), который является обобщенной количественной мерой того, насколько данные соответствуют требуемым стандартам в рамках рассмотренных характеристик качества данных. На основе рассчитанного ИПКД проводится интерпретация, и анализ полученных результатов для принятия решений в части улучшения и контроля процедуры управления качеством данных.

Управление качеством данных особенно актуально для банков, в которых генерируется большой объем данных, а также разрабатываются модели машинного обучения для различных направлений деятельности и проектов банка. Таким примером может являться подход на основе внутренних рейтингов (ПВР) – это продвинутый способ оценки кредитного риска банка. Описание этого подхода содержится в положении 483-П Банка России «О порядке расчета величины кредитного риска на основе внутренних рейтингов» [1], которое также определяет требования к валидации качества данных, согласно которым Банк определяет во внутренних документах методику и порядок обеспечения качества данных в разрезе характеристик, включая актуальность, согласованность, доступность, контролируемость, восстанавливаемость, точность, полноту, достоверность. Также стоит отметить, что используемые в компаниях витрины данных для целей моделирования и аналитики должны проходить этапы приемо-сдаточных испытаний (ПСИ), иметь документацию, содержащую бизнес-требования, потоки данных, тесты на качество данных, а также своевременно обновляться и проходить соответствующий мониторинг качества данных. В то же время проверки могут быть простыми, например, проверка на пропущенные значения, а также более сложными с проверкой на соответствие методологии или с использованием различных алгоритмов и статистических методов.

Список литературы

1. Приложение 3 к Положению Банка России от 6 августа 2015 г. N 483-П «О порядке расчета величины кредитного риска на основе внутренних рейтингов»: [Электронный ресурс]. URL: https://www.cbr.ru.
2. Центр компетенций НТИ по большим данным МГУ: [Электронный ресурс]. URL: https://bigdata.msu.ru/standards/
3. International Organization for Standardization: [Электронный ресурс]. URL: https://www.iso.org/

4. International Electrotechnical Commission: [Электронный ресурс]. URL: https://www.iecee.org/

References

1. Appendix 3 to the Regulation of the Bank of Russia dated August 6, 2015 N 483-P "On the procedure for calculating the amount of credit risk based on internal ratings": [Electronic resource]. URL: https://www.cbr.ru.
2. NTI Competence Center for Big Data of Moscow State University: [Electronic resource]. URL: https://bigdata.msu.ru/standards/
3. International Organization for Standardization: [Electronic resource]. URL: https://www.iso.org /
4. International Electrotechnical Commission: [Electronic resource]. URL: https://www.iecee.org/