УДК 004

ПРОБЛЕМЫ БОЛЬШИХ ВЫБОРОК ДАННЫХ

Рогов Д.Е., студент гр. ИИм-221, 2 курс
Кузбасский государственный технический университет имени Т.Ф. Горбачева, г. Кемерово

Выборки больших данных представляют собой важный аспект современного анализа данных, но они также сталкиваются с несколькими проблемами. Одной из основных сложностей является обработка и хранение огромных объемов информации [1-4]. Это требует высокой вычислительной мощности и эффективных методов обработки данных [5].

Еще одной проблемой является качество данных. В больших объемах информации могут присутствовать ошибки [6], выбросы [7] или неполные записи (рисунок 1) [8], что может исказить результаты анализа. Гарантировать чистоту данных становится сложной задачей.

0,246	0	0	0
	0,010822511	0,002662338	0,377358491
	0	0	0
	0,002164502	0,000532468	0,075471698
	0	0	0
	0	0	0
	0	0	0
	0,001623377	0,000399351	0,056603774
	0	0	0
	0,001082251	0,000266234	0,037735849
	0,000541126	0,000133117	0,018867925
	0	0	0
	0	0	0
	0,002164502	0,000532468	0,075471698
	0,002705628	0,000665584	0,094339623
	0	0	0
	0,002705628	0,000665584	0,094339623
	0,001082251	0,000266234	0,037735849
	0,002705628	0,000665584	0,094339623
	0	0	0
	0,001082251	0,000266234	0,037735849

Рисунок 1 – Пример неполных данных

Сложности возникают также при необходимости выбора подходящих алгоритмов для анализа больших данных. Некоторые методы, эффективные на небольших выборках, могут оказаться неэффективными или даже неприменимыми при работе с огромными объемами данных [9].

Вопросы конфиденциальности и безопасности также становятся более актуальными при работе с большими данными. Необходимо разработать эффективные механизмы для защиты личных данных и предотвращения возможных утечек информации [10].

Наконец, важно учитывать вычислительные затраты при анализе больших данных. Это может потребовать значительных инвестиций в вычислительное оборудование и технологии, чтобы обеспечить эффективную обработку и хранение данных [11].

Все эти проблемы подчеркивают важность разработки инновационных подходов к работе с большими данными, чтобы извлекать максимальную пользу из их потенциала, минимизируя при этом возможные негативные.

Список литературы:

1. Свидетельство о государственной регистрации программы для ЭВМ № 2023680124 Российская Федерация. BrainPower : № 2023669010 : заявл. 16.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN QXBJIM.
2. Математические и программные методы построения моделей глубокого обучения : Учебное пособие / А. В. Протодьяконов и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 176 с. – ISBN 978-5-9729-1484-5. – EDN PZLUAH.
3. Свидетельство о государственной регистрации программы для ЭВМ № 2023680335 Российская Федерация. Maitak Intelligence Natural Language Processing Module : № 2023669704 : заявл. 27.09.2023 : опубл. 28.09.2023 / Р. В. Майтак.
4. Методы восстановления непараметрической регрессии в условиях несбалансированных данных / А. Д. Салычева и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2024. – 192 с. – ISBN 978-5-9729-1856-0. – EDN AAJATW.
5. Свидетельство о государственной регистрации программы для ЭВМ № 2023684619 Российская Федерация. Efficient Network: № 2023684038: заявл. 14.11.2023: опубл. 16.11.2023 / П. А. Пылов.
6. Свидетельство о государственной регистрации программы для ЭВМ № 2023680070 Российская Федерация. Модернизированная модель DBSCAN для определения скрытых взаимосвязей : № 2023668841 : заявл. 13.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN KQUUKF.
7. Асимптотический анализ поведения прикладных моделей машинного обучения : Учебное пособие / А. В. Протодьяконов и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 144 с. – ISBN 978-5-9729-1455-5. – EDN APHQME.

8. Свидетельство о государственной регистрации программы для ЭВМ № 2023684621 Российская Федерация. Destructed Deep Random Forest: № 2023684050: заявл. 14.11.2023: опубл. 16.11.2023 / П. А. Пылов.
9. Свидетельство о государственной регистрации программы для ЭВМ № 2023684622 Российская Федерация. Mask Made AI: № 2023684042: заявл. 14.11.2023: опубл. 16.11.2023 / П. А. Пылов.
10. Свидетельство о государственной регистрации программы для ЭВМ № 2023680103 Российская Федерация. Cognitive Solution : № 2023669189 : заявл. 19.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN QEMFJA.
11. Свидетельство о государственной регистрации программы для ЭВМ № 2023684624 Российская Федерация. Программа автоматического распознавания лиц в видеопотоке: № 2023684236: заявл. 15.11.2023: опубл. 16.11.2023 / П. А. Пылов.