УДК 004.891.2

РАЗРАБОТКА ИНСТРУМЕНТОВ ПРОМЫШЛЕННОГО АНАЛИЗА ТАБЛИЧНЫХ ДАННЫХ НА ОСНОВЕ ГИБРИДНЫХ АЛГОРИТМОВ

Кирпа Д.П. (ИТМО)

Научный руководитель – кандидат технических наук, доцент Кугаевских А.В.
(ИТМО)

Введение. Анализ больших потоков данных в условиях сегодняшнего дня является ключевым процессом для успешного осуществления работы производств и крупных компаний [1]. Эти данные в общем случае имеют абсолютно разную природу и применение (от показателей датчиков на химическом предприятии до истории покупок пользователей в интернет-магазине) и обрабатываются широким кругом специалистов. Во многих производственных и бизнес-процессах работой с данными вынуждены заниматься люди, не имеющие образования и опыта в области статистики и аналитики. Это приводит к неэффективным затратам рабочего времени специалистов и ошибкам в принятии решений. Разработке системы, предоставляющей набор инструментов для эффективной работы с табличными данными любого происхождения и посвящена данная работа. Задачи сервиса - генерация релевантных задаче пользователя описаний таблиц на естественном языке, а также семантический поиск таблиц и данных из таблиц.

Основная часть. Система разбивается на две независимые компоненты:

1) Подсистема генерации решающих задачу пользователя текстовых описаний таблиц.
2) Подсистема семантического поиска данных.

По итогу система комбинирует две компоненты, генерируя ответы на запросы пользователя по наиболее релевантным данным из коллекции.

Работа первой подсистемы основана на моделировании процесса мышления аналитика данных при работе с таблицами и реализация ключевых моментов этого процесса в виде программного алгоритма. Начальная стадия работы состоит в понимании общего контекста данных и получения закономерностей в них. Понимание общего контекста данных возлагается на генеративную текстовую нейросетевую модель. Она формирует предположения о наиболее соответствующих проблеме пользователя колонках и методах их обработки в специальном формате. С использованием этой информации применяются классические методы статистики, выводятся зависимости и закономерности в данных. От применения нейросетевых подходов на этом этапе было решено отказаться, так как использование существующих подходов глубинного обучения пока что не даёт значительного преимущества [2]. Завершается обработка ответом генеративной текстовой нейросети, которая на основании данных с предыдущих шагов делает выводы о данных в виде текста на естественном языке и графиков. Архитектура была спроектирована таким образом, что конкретные используемые модели не были зафиксированы. Это позволило провести опыты с различными моделями.

Вторая подсистема основана на индексе для поиска данных, извлекающего семантические свойства из табличных данных и агрегирующего связи между ними. Для разработки такого индекса были использованы гибридные классические и нейросетевые подходы.

Выводы. По итогам работы была разработана и протестирована система генерации текста по табличным данным в рамках заданной пользователем задачи. Было показано, что комбинация различных нейросетевых и классических подходов машинного обучения позволяет моделировать процессы работы специалиста определённого профиля при решении типовых задач. Предложенный принцип работы может быть положен в основу корпоративных инструментов для работы с табличными данными и программ-тренажёров для обучения аналитике данных и повышения корпоративной культуры работы с большими данными.

Список использованных источников:

1. Шпилькина Т.А., Ляшкова О.В. Роль Big Data в деятельности корпораций // Экономика и бизнес: теория и практика. 2020. №4-3.
2. Gorishniy Y., Rubachev I., Kartashev N., Shlenskii D., Kotelnikov A., Babenko A. TabR: Tabular Deep Learning Meets Nearest Neighbors in 2023 //arXiv preprint arXiv:2307.14338. – 2023.