К ПРОБЛЕМЕ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ПРИ РАБОТЕ С БОЛЬШИМИ МАССИВАМИ ДАННЫХ

Титова А.В.

Белорусский государственный университет информатики и радиоэлектроники г. Минск, Республика Беларусь

Яшин К.Д. – к. ф.-м. н., доцент

Анализируются характеристики производительности алгоритмов корреляционного анализа, реализованных в библиотеках Python.

В качестве предмета исследования выбран корреляционный анализ. Исследования и решение поставленных задач выполнялось по схеме (рисунок 1).

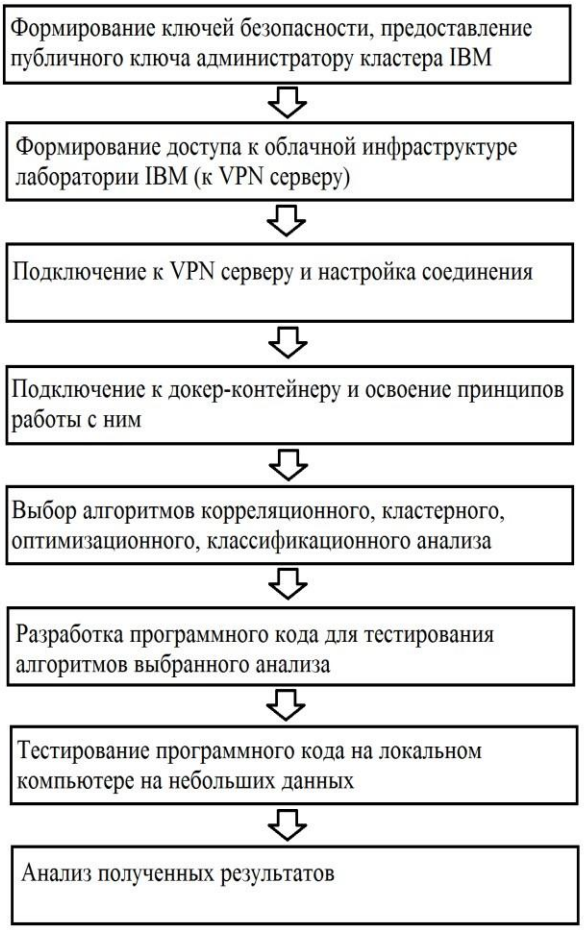

Рисунок 1 – Схема алгоритма выполнения проекта

Для исследования выбраны два вида корреляционных зависимостей [1]:

коэффициент корреляции Пирсона

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}},$$

коэффициент корреляции Спирмана

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}.$$

Основной задачей явилось установление различий в эффективности работы алгоритмов. Это осуществлялось путем тестирования алгоритмов на наборах информационных данных. Последние отличаются как по размерам, так и по структуре. На этапе эксперимента предполагается, что все информационные данные, используемые для тестов эффективности, уже обработаны и не имеют структурных ошибок, таких как отсутствие значений, значения неверного формата и т.д.

Как известно [2], процесс бенчмаркинга собирает и анализирует информационные данные по характеристикам производительности. Рекомендатор содержит пять вложенных статистических моделей измерения производительности. Они прогнозируют объем компьютерных ресурсов, требуемых каждым алгоритмом. Это позволяет пользователям рекомендатора изменять размеры и вес набора информационных данных по пяти показателям производительности, учитывая бизнес-требования. Программа рекомендатор генерирует оценку для каждого из алгоритмов.

В настоящем исследовании рассмотрены вопросы актуальности анализа эффективности алгоритмов машинного обучения, используемых для задач бизнеса и промышленности. Представлена методология выполнения поставленных в проекте задач, описаны результаты выполнения проекта. Подробно описаны этапы настройки необходимой среды и начальные тестовые данные выбранных алгоритмов корреляционного анализа [3].

Автор благодарит Б. Зибицкера, профессора Чикагского университета (США), за оказание технической помощи и консультаций при выполнении работы.

Список использованных источников:

1.Корреляция: Материал из Википедии — свободной энциклопедии: Версия 93206762, сохранённая в 10:05 UTC 10 июня 2018 / Википедия, свободная энциклопедия. — Электрон. дан. — Сан-Франциско: Фонд Викимедиа, 2018. — Режим доступа: https://ru.wikipedia.org/?oldid=93206762

2.Okallau B., Shebik B., Wishart H. Recommender Development for Selection of Appropriate Regression Algorithms in Pyspark.ML Library (материал предоставлен профессором Чикагского университета Б. Зибицкером).

3.Титова, А. В. Анализ эффективности машинных алгоритмов при работе с большими объемами данных / А. В. Титова, К. Д. Яшин // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня : сборник материалов V Международной научно-практической конференции, Минск, 13–14 марта 2019 г. В 2 ч. Ч. 2 / Белорусский государственный университет информатики и радиоэлектроники; редкол. : В. А. Богуш [и др.]. – Минск, 2019. – С. 229 – 241.