

Методы оптимизации моделей машинного обучения в задаче предсказания результатов теста «Большая пятерка»

О. Р. Гавриленко

Санкт-Петербургский
Федеральный исследовательский
центр РАН

o.gavrilenko.r@gmail.com

В. Д. Олисеенко

Санкт-Петербургский
Федеральный исследовательский
центр РАН

vdo@dsos.pro

Т. В. Тулупьева

Санкт-Петербургский
Федеральный исследовательский
центр РАН;
РАНХуГС

tvt@dsos.pro

Аннотация. В данной статье исследуются методы оптимизации для подбора гиперпараметров в задаче предсказания результатов психологических особенностей пользователей социальной сети. Теоретическая значимость работы заключается в выработке методологии для оптимизации моделей машинного обучения в контексте предсказания результатов психологического теста по данным из социальных сетей.

Ключевые слова: анализ данных; большая пятерка; машинное обучение; нейронная сеть; предсказание; психологический тест

I. ЗАГОЛОВОК РАЗДЕЛА

Существующие исследования подтверждают связь между контентом, который пользователь размещает в своих социальных сетях и результатами различных психологических тестов, включая тест «Большая пятерка» [1, 2]. Данную связь также возможно предсказать при помощи методов искусственного интеллекта и машинного обучения [3, 4]. Несмотря на то, что область предсказания психологических особенностей пользователей социальных сетей стремительно развивается [5], ряд существенных вопросов остается открытым. Одним из таких вопросов является применение методов оптимизации при построении предсказательных моделей.

Цель работы заключается в исследовании методов оптимизации для подбора гиперпараметров в задаче предсказания результатов психологических особенностей пользователей социальной сети. Для достижения данной цели используется выборка респондентов из социальной сети VK.com с результатами прохождения теста. Теоретическая значимость работы заключается в выработке методологии для оптимизации моделей машинного обучения в контексте предсказания результатов психологического теста по данным из социальных сетей. Практическая значимость заключается в увеличении точности конечных метрик качества предсказания за счет использования релевантных оптимизаторов для подбора гиперпараметров и весов модели.

II. ОБЗОР ЛИТЕРАТУРЫ

Существует несколько подходов для решения задачи предсказания результатов психологического теста [6–11] на основе данных пользователя из социальных сетей. Первая группа [6–9] подходов относится к решению задач классификации, где авторы описывают задачу мультиклассовой классификации пользователя с результатами тестирования психологического теста. Вторая группа [10–12] относится к предсказанию психологических характеристик с использованием методов регрессии, где рассматриваются подходы к количественному предсказанию этих показателей.

В статье [6] авторы рассматривали вопрос взаимосвязи между личностными особенностями пользователя, оцениваемыми с помощью теста Р. Кеттелла и его подписками во «ВКонтакте». В работе решается задача мультиклассовой классификации для определения степени выраженности психологического фактора с помощью логистической регрессии, градиентного бустинга и случайного леса.

В статье [7] определяются признаки поведения русскоязычных пользователей социальной сети «ВКонтакте» в связи с уровнем выраженности у них признаков депрессии по опроснику депрессии А. Бека.

В исследовании участвовало 1268 пользователей, которые проходили опросник депрессии А. Бека через специальное приложение внутри социальной сети. В результате было сформировано 3 группы респондентов с разным уровнем выраженности признаков депрессии: 209 с низким, 780 со средним, 279 с высоким. Опросник описывает значимые симптомы депрессии, объединенные в 21 категорию. Каждая категория состоит из 4–5 пунктов, которые оцениваются от 0 до 3 баллов по нарастанию тяжести симптома.

Для выявления информативных признаков депрессии применялись методы опорных векторов и алгоритм случайного леса для классификации объектов. Те респонденты, у которых итоговый балл превышал 13, относились к группе здоровых пользователей. Если результат теста был меньше 29, пользователь попадал в класс депрессии. Для определения точности прогнозирования авторы использовали следующие

метрики: precision (точность), recall (полнота) и F1-мера (среднее взвешенное от precision и recall).

В работе [11] авторы рассматривают вопрос разработки методики предсказания подверженности пользователей социальных сетей деструктивным воздействиям по тесту Аммона на основе использования искусственных нейронных сетей.

В исследовании участвовало 460 студентов-медиков, которые проходили онлайн – тестирование Аммона. Результаты были помечены 18 шкалами, представляющими собой комбинации шести Я-функций и трех компонентов. Для каждой пары рассчитывалось значение Т-балла, где от 0 до 39 значение интерпретировалось как низкое, от 40 до 60 – среднее, от 61 до 110 – высокое.

В роли классификаторов авторы использовали многослойную нейронную сеть с тремя скрытыми слоями нейронов, метод опорных векторов с радиально-базисным ядром, линейная регрессия, а также сверточная нейронная сеть. Для оценки каждой модели использовалась метрика точности. Можно увидеть, что основная часть исследований о предсказании результатов психологического результата говорят об анализе данных и выявлении значимых признаков, однако вопрос улучшения метрик качества за счет использования оптимизаторов ранее не рассматривался в литературе.

III. ПОСТАНОВКА ЗАДАЧИ

В данной работе рассматривается психологический тест «Большая пятерка», который состоит из 5 факторов: экстраверсия, открытость, доброжелательность, добросовестность и невротизм [12]. Тест состоит из 80 вопросов, каждый вопрос содержит в себе 5 вариантов ответов, которые определяют насколько человек соотносит себя с высказыванием в вопросе: -2 («Почти никогда»), -1 («Редко»), 0 («Иногда»), 1 («Часто»), 2 («Почти всегда»). Результаты тестирования представляются в виде оценки по шкале от 0 до 100 для каждого фактора, которая формируется на основе расшифровки ответов теста [12]. задачу предсказания результатов теста «Большая пятерка» к задаче регрессии, где результаты каждого фактора предсказываются в рамках двух интервалов на интервалы от 0 до 50 (для определения положительного фактора) и от 50 до 100 (для определения отрицательного фактора). Математическая постановка задачи определяется следующим образом: пусть X_0 – множество характеристик пользователя из социальных сетей, y – результаты теста «Большая пятерка» по каждому фактору. Тогда для каждого пользователя необходимо построить по 5 моделей предсказания, которые получают на вход информацию с личной страницы пользователя, а на выходе получаем результат теста по одному из факторов.

IV. ПОСТАНОВКА ЗАДАЧИ

В работе используются открытые данные по тем пользователям, которые использовали приложение VK. Всего респондентов 6987 человек, из которых 70,8 % – женщины, 29,2 % – мужчин. Полная выборка состоит 6978 строк и 48 столбцов (id пользователя, наименование теста, дата выполнения теста, результат признака

«Экстраверсия», результат признака «Открытость», «Доброжелательность», «Добросовестность», «Невротизм», время выполнения теста, ссылка на другие социальные сети, флаг наличия контента, личная информация: образование, город, в котором пользователь живет сейчас, родной город, школа, пол, дата рождения, любимые книги, компания, должность, количество подписчиков, хобби, посты, группы, видео, фотографии, альбомы, аудиозаписи, список друзей, количество интересных страниц, подписки, количество видеозаписей, количество плейлистов видеозаписей, количество постов, посты, количество подарков, количество фотографий, количество групп, количество клипов, количество подписчиков на клипы, количество видов клипов, пожелания). Эксперимент выполнялся с использованием языка программирования Python 3.12 совместно со следующими библиотеками: Numpy 2.2.3, Pandas 2.2.3, Scikit-learn 1.6.1, Optuna 4.2.1.

Подготовка данных для построения моделей предсказания результатов психологического теста началась с анализа пользовательской информации. В исследовании участвовало 6987 пользователей, из которых 357 с закрытой страницей, поэтому они были удалены из общей выборки. При сборе пользовательских данных было определено 25 признаков с числовыми значениями и 23 признака с текстовыми значениями. Для предсказания результатов психологического теста в работе рассматриваются только числовые признаки, оставшаяся часть выборки была удалена. Следующим шагом необходимо обработать пустые или нулевые значения признаков. Для этого шага были рассчитаны стандартное отклонение и квантили для определения рассеивания данных и наличия выбросов. Обработка таких значений выполнялась заменой на среднее значение у следующих признаков: количество видеоплейстивов, альбомы, клипы, подписчики, количество друзей, количество лайков на клипы, моду: аудиозаписи, видеозаписи, подарки, посты, количество подписок пользователя или медиану: количество интересных страниц, количество фотографий, количество подписчиков на клипы, количество видов клипов, количество групп, в зависимости от стандартного отклонения признака. Выбросы обработаны с помощью удаления данных на 1–2 % у каждого признака.

Для дальнейшего предсказания были рассмотрены следующие алгоритмы машинного обучения с различными гиперпараметрами:

1. Random Forest Regressor: количество деревьев от 50 до 200 с шагом 15, глубина деревьев от 5 до 30, минимум объектов для разбиения от 2 до 20, минимум объектов в листе от 1 до 10, количество признаков при разбиении от 0,4 до 1, функция ошибки: средняя абсолютная ошибка, средняя квадратическая ошибка.

2. Gradient Boosting Regressor: скорость обучения от 0,005 до 0,2, глубина деревьев от 3 до 10, минимум объектов для разбиения от 2 до 20, минимум объектов в листе от 1 до 10, количество признаков при разбиении от 0,4 до 1, функция ошибки: средняя квадратическая ошибка, средняя абсолютная ошибка, доля данных для обучения каждого дерева от 0,5 до 1,0, количество деревьев решений от 50 до 200 с шагом 50.

3. Feedforward Neural Network: количество слоев: 1,2,3 слоя с 64 нейронами, 4,5 с 32 нейронами, скорость обучения – 0,001, регуляризация весов – 0,01, функция потерь: средняя абсолютная ошибка, количество эпох – 400, размер батча – 64, функция активации для скрытых слоев – LeakyReLU с $\text{negative_slope}=0.1$

Для подбора гиперпараметров использовался метод GridSearchCV [14] из библиотеки Scikit-learn для моделей случайного леса и градиентного бустинга, который позволяет перебрать все заданные комбинации и оценить их при помощи перекрёстной проверки на пяти блоках тестовой выборки, а также оптимизатор Optuna [15], который использует несколько алгоритмов оптимизации (TPE, CMA-ES) для систематического поиска наилучших гиперпараметров. Для нейронной сети используется оптимизатор Lion [16], который подбирает оптимальные шаги для обновления весов, учитывая как локальные, так и глобальные особенности градиентов.

Для оценки результатов предсказания в работе использовались следующие метрики: MAE (средняя абсолютная ошибка), RMSE (среднеквадратичная ошибка), которые являются релевантными для задачи регрессии:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|,$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2},$$

где y_i – значение из тестовой выборки, $\hat{y}$ – предсказанное значение, N – общее количество экземпляров.

V. ПОСТАНОВКА ЗАДАЧИ

В табл. I представлены результаты эксперимента, где модели предсказывают значения каждого признака психологического теста в диапазоне от 0 до 100.

ТАБЛИЦА I. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

Фактор	Модель	Без оптимизации		С оптимизацией	
		MAE	RMSE	MAE	RMSE
Открытость	Baseline	9,608	12,038	-	-
	Случайный лес	6,800	8,727	6,659	8,609
	Градиентный бустинг	6,789	8,712	6,676	8,610
	Нейронная сеть	-	-	6,886	8,843
Невротизм	Baseline	14,949	18,760	-	-
	Случайный лес	10,399	12,791	10,124	12,512
	Градиентный бустинг	10,415	12,821	10,181	12,443
	Нейронная сеть	-	-	10,651	13,162
Доброжелательность	Baseline	13,196	12,038	-	-
	Случайный лес	9,096	11,364	9,220	11,423
	Градиентный бустинг	9,154	11,405	9,290	11,452
	Нейронная сеть	-	-	9,189	11,517
Добросовестность	Baseline	13,463	16,841	-	-
	Случайный лес	9,339	11,561	9,301	11,474
	Градиентный бустинг	9,296	11,544	9,329	11,485

Фактор	Модель	Без оптимизации		С оптимизацией	
		MAE	RMSE	MAE	RMSE
Экстраверсия	Нейронная сеть	-	-	9,396	11,638
	Baseline	12,809	16,028	-	-
	Случайный лес	8,880	11,124	8,696	10,942
	Градиентный бустинг	8,877	11,120	8,688	10,920
	Нейронная сеть	-	-	9,336	11,612

Затраченное время расчета каждой модели, рассчитанное с помощью операционной системы macOS, версия python: 3.12.7:

- случайный лес: 3 часа 17 минут;
- градиентный бустинг: 3 минуты 13 секунд;
- нейронная сеть с пятью скрытыми слоями: 1 минута 12 секунд.

По результатам эксперимента видно, что все модели показывают более точный результат в сравнении с базовой моделью, которая была построена с помощью Т-распределения, так как считается, что каждый признак имеет нормальное распределение. Помимо этого видно, что результаты с использованием методов оптимизации показывают наилучшие значения метрик. Среди моделей машинного обучения наилучшую точность показывает метод случайный лес, однако затраченное на расчет время значительно больше (на 3 часа 14 минут), чем у градиентного бустинга, при этом наибольшее отличие метрики MSE у случайного леса от градиентного бустинга составляет 0,00702 у признака «Доброжелательность», что говорит о незначительном перевесе точности. Таким образом, учитывая все обстоятельства, наилучшей моделью необходимо считать градиентный бустинг, показывающей значения лучше базовой модели на 3–4 единицы, а также которая считается довольно быстро относительно возможно затраченного времени на расчет.

VI. ЗАКЛЮЧЕНИЕ

В работе рассмотрена разработка моделей машинного обучения с использованием таких методов оптимизации, как GridSearchCV и Optuna, необходимые для подбора гиперпараметров. С помощью представленных методов были получены более точные результаты психологического теста «Большая пятерка». Optuna показала наилучший результат оптимизации, так как данная библиотека использует методы оптимизации на основе байесовского подхода и дает возможность использовать более гибкие и сложные пространства гиперпараметров в то время, как GridSearchCV выполняет полный перебор всех гиперпараметров и использует только их фиксированные значения.

Результаты данного исследования могут быть использованы для разработки более сложной модели, учитывающей большее количество пользовательских данных с категориальными признаками.

СПИСОК ЛИТЕРАТУРЫ

- [1] Цапкова Е.А. «Большая пятерка» личностных качеств // Международная научно-практическая конференция «Современная наука: вызовы, проблемы, решения — взгляды молодёжи» (Биробиджан, 2023). С.660–665.
- [2] Можейкина Л.Б. Личностные характеристики активных пользователей социальных сетей // Научно-практический журнал «Медицина Кыргызстана». 2015. № 2. С. 93–95.
- [3] Maharani W., Effendy V. Big Five Personality Prediction from Social Media Data using Machine Learning Techniques // International Journal of Engineering and Advanced Technology. 2020. С.2412–2417.
- [4] Oktafiqurahman A., Kusriani K., Nasiri A. Personality Prediction Based on Facebook Media Social Status Using the Method Naïve Bayes and KNN // International Journal of Social Science and Human Research. 2023. P. 4257–4262.
- [5] Ледовая Я.А., Тихонов Р.В., Боголюбова О.Н. Социальные сети как новая среда для междисциплинарных исследований поведения человека // Вестник Санкт-Петербургского университета. Психология и педагогика. 2017. Т. 7. №. 3. С. 93–95.
- [6] Рязанцев Г.Е., Олисеенко В.Д., Абрамов М.В. Предсказание результатов теста Р.Кеттелла на основе подписок пользователя в социальной сети // Международная конференция по мягким вычислениям и измерениям. 2023. С. 42–44.
- [7] Кисельникова Н.В., Станкевич М.А., Данина М.М., Куминская Е.А., Лаврова Е.В. Выявление информативных параметров поведения пользователей социальной сети ВКонтакте как признаков депрессии // Психология. Журнал Высшей школы экономики. Т. 17. № 1. 2020. С.73–88.
- [8] Браницкий А.А., Дойникова Е.В., Котенко И.В. Методика классификации пользователей социальной сети по психологическим шкалам теста Аммона на основе искусственных нейронных сетей // Информационные технологии в управлении: 13-я Мультиконференция по проблемам управления. Санкт-Петербург, 2020. С. 300–302.
- [9] Станкевич М.А., Смирнов И.В., Игнатьев Н.А., Кисельникова Н.В., Данина М.М. Анализ личностных черт пользователей социальных сетей на основе автоматической обработки их профилей // Системы высокой доступности. 2018. Т.4. № 4. С. 15–19. Doi: 10.18127/j20729472-201804-04.
- [10] Горелова Г.Г., Иноземцев Д.В. Анализ психологического благополучия на основании параметров поведенческой активности пользователей социальных сетей // Научный журнал «Ученые записки университета им. П. Ф. Лесгафта». 2020. № 2. С. 489–494.
- [11] Дойникова Е.В., Браницкий А.А., Котенко И.В. Использование нейросетей для прогнозирования подверженности пользователей социальных сетей деструктивным воздействиям // Информационно-управляющие системы. 2020. № 1. С. 24–33. <https://doi.org/10.31799/1684-8853-2020-1-24-33>
- [12] Ихсанов И.Р., Шахова И.С. Применение методов машинного обучения для выявления взаимосвязи академической успеваемости и данных профиля социальной сети // Электронные библиотеки. 2019. Т. 22, № 2. С. 95–118.
- [13] Сербина Г.Н. Мацута В.В., Гойко В.Л. Анализ связи психологических характеристик пользователей социальной сети «ВКонтакте» с подписками на сообщества с девиантным контентом // Вестник Томского государственного университета. 2021. № 467. С. 164–169. DOI: 10.17223/15617793/467/20 14.
- [14] Puślecki T., Walkowiak K. Hyperparameters Optimization Using GridSearchCV Method for TinyML Models // Lecture Notes in Networks and Systems .2023. Vol. 766. P.63–69. DOI:10.1007/978-3-031-41630-9_7.
- [15] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019. P. 2623–2631.
- [16] Maziar Yazdani, Fariborz Jolai. Lion Optimization Algorithm (LOA): A nature-inspired metaheuristic algorithm. // Journal of Computational Design and Engineering. 2016. Vol. 3. Iss. 1. P. 24–36.