# Анализ изображений в образовательном тестировании с помощью машинного обучения (на примере инструмента измерения креативности)

***Тарасов С.В.***
Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0003-4151-115X, e-mail: svtarasov@hse.ru

***Гельвер Е.С.***
Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0001-9365-801X, e-mail: egelver@hse.ru

***Грачева Д.А.***
Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-4646-7349, e-mail: dgracheva@hse.ru

***Угланова И.Л.***
Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0001-9117-5997, e-mail: iuglanova@hse.ru

***Вырва Е.Е.***
Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0001-8174-421X, e-mail: evyrva@hse.ru

***Ключевые слова:*** машинное обучение, анализ изображений, образовательное оценивание, креативность

На сегодняшний день оценка сложных, метапредметных компетенций является актуальной задачей системы образования. Примером такой компетенции является креативность. Креативность включена в большинство рамок ключевых грамотностей и навыков современного мира: доклад Всемирного Экономического Форума 2018, исследовательские работы [1] и российские образовательные стандарты (ФГОС НОО, 2009, ФГОС ООО, 2010). Для оценивания сложных компетенций требуется прибегать не только к традиционным форматам тестирования (задания с выбором ответа или эссе), но и к более современным форматам.

Часто инновационные форматы предлагают среду тестирования, в которой требуется графическое решение тестового задания. Это осо-

бенно актуально для заданий на креативность, где тестируемых просят создать некоторый продукт. Оценивание продуктов деятельности тестируемых часто проводится с привлечением экспертов. При этом экспертное оценивание обладает рядом существенных недостатков (ресурсозатратность на обучение экспертов, субъективные искажения, отложенная обратная связь и др.). Кроме того, для валидной оценки заданий с графическим решением недостаточно учитывать каждое конкретное действие тестируемого, а требуется учет целостности продукта, созданного в ходе тестирования. Для повышения качества оценивания таких заданий все более востребованными становятся использование машинного обучения.

Технологии машинного обучения в образовательном тестировании широко применяются при автоматическом анализе текстов [2—4]. Однако методология автоматического анализа изображений для сбора доказательств валидности результатов измерения еще не разработана в той мере, чтобы ее можно было использовать при разработке инструментов оценивания компетенций. Таким образом, анализ изображений представляет собой новый этап при доказательстве валидности результатов тестирования сложных компетенций, который учитывает целостность продукта, созданного в ходе тестирования, и нивелирует недостатки экспертного оценивания.

Целью нашего исследования было представить результаты анализа изображений с применением машинного обучения для образовательного оценивания. В частности, представлены результаты автоматического анализа изображений как способа измерения креативности учеников в конце начальной школы.

Для достижения цели мы рассматриваем инструмент для оценки навыков 21го века «4К» среди учеников 4 класса, разработанный Центром психометрики и измерений в образовании (Лаборатория измерения новых конструктов и дизайна тестов) Института образования НИУ ВШЭ в рамках договора о научно-исследовательской работе по разработке инструмента оценки навыков «4К» с фондом «Вклад в будущее». В данный инструмент входит шесть заданий сценарного типа, оценивающих четыре навыка: критическое мышление, креативность, коммуникацию и кооперацию.

Инструмент предъявляются тестируемому в компьютерной форме. Экран заданий интерактивен: тестируемый нажимает на выбранную им область, после чего видит заранее подготовленную специфическую реакцию системы тестирования. Такой формат позволяет проявить сложные компетенции, а также поддерживает мотивацию тестируемых и снижает тестовую тревожность. В то же врем, имита-

ция реальной среды позволяет более точно зафиксировать наблюдаемое поведение, то есть свидетельство того, что тестируемый обладает конкретным навыком.

В целом, в инструменте 4К существует тенденция, что одно задание позволяет измерить несколько навыков. В рамках данной работы рассматривалось одно задание под названием «Монстр», которое в большей степени направлено на оценку навыка креативности. Для создания инструмента измерения разработчиками была определена теоретическая рамка инструмента, которая основывается на классических и современных теориях креативности [5—6]. Креативность в данном инструменте понимается как способность разработать и представить принципиально новые, уникальные, необычные идеи/продукты, полезные для решения стоящей перед субъектом проблемы и состоит из двух субконструктов, реализуемых в инструменте измерения:

— Оригинальность. Этот поведенческий индикатор является характеристикой разрабатываемого тестируемым решения поставленной проблемы и означает необычность, нетипичность, непохожесть этого решения на другие решения.

— Детальность, также как оригинальность, относится к свойствам решения задачи, которое предлагает тестируемый и выражается в умении тщательно продумать свое решение проблемы и детально проработать его, создав целостный образ.

В задании «Монстр» каждый тестируемый, используя элементы конструктора, создает трех монстров, которые, по его мнению, «удивительны, необычны и отличаются от остальных монстров в городе». Элементы конструктора разложены в несколько категорий (туловище, рот, глаза, разное), каждый элемент можно использовать несколько раз. Остальные монстры в городе были представлены на рабочем экране и считались референсами. Для каждого тестируемого сохраняются значения индикаторов по трем монстрам.

Для оценки оригинальности использовались шесть индикаторов, которые оценивали отличие созданного монстра от референса (расположение конечностей, глаз, рта и др.)

Для оценки детальности использовались два индикатора, которые оценивали изменение цвета монстра и вращение элементов монстра, и семь индикаторов на общее количество элементов и количество элементов из разных категорий.

В исследовании принимали участие 1780 учащихся четвертых классов из трех городов России — Москвы, Калуги и Ярославля. Среднее время выполнения задания «Монстр» — 10 минут.

Так как каждый тестируемый строил по три монстра, то исходная база содержала 5340 изображения. После удаления профилей с пропущенными значениями по всем индикаторам количество монстров в базе сократилось до 5286 (1762 профиля учащихся).

В рамках данного исследования используется направление машинного обучения, объединяющее алгоритмы и методы построения моделей нейронной сети на основе размеченных данных (методика обучения с учителем, supervised learning). Для задачи классификации используются данные вида «элемент — метка класса». Обучение модели происходит за счет минимизации ошибки между предсказанным и фактическим значением. Следовательно, для достижения максимальной точности модели, процесс разметки данных является ключевым. Источником информации для разметки выступают результаты психометрического анализа данных в методологии Mixture Modeling. Возможность не обращаться к экспертам на этом этапе также позволяет сохранять все преимущества оценивания без привлечения экспертов.

Методология Mixture Modeling позволяет выявить латентные классы (группы) среди продуктов деятельности тестируемых (созданных монстров). Интерпретация результатов позволяет говорить о существовании групп тестируемых с различным уровнем выраженности субконструкта креативности — оригинальности и детальности.

Проведенный анализ будет описан подробно на примере субконструкта «оригинальность», аналогичный анализ был проведен и для субконструкта «детальность».

На первом этапе был проведен разведывательный (эксплораторный) анализ для выявления оптимального количества латентных классов. Были построены модели с двумя, тремя и четырьмя классами на основе шести индикаторов оригинальности. Качество разбиения определялось по показателю энтропии, где значения больше 0.8 считаются приемлемыми. Сравнение моделей с разным количеством классов происходило по информационному критерию Акаике (AIC) и байесовскому информационному критерию (BIC) в пользу моделей, где значения критериев минимальны. Дополнительно, тест Лу-Менделя-Рубина (Lo—Mendell—Rubin Adjusted Likelihood Ratio Test, LRT) использовался для сравнения моделей с количеством классов, отличающихся на один. Значимость критерия свидетельствует о том, что модель с меньшим количество классов подходит данным хуже. Согласно рекомендации [7], отдавалось предпочтение моделям, где минимальная пропорция тестируемых в одном классе составляла не менее 10%. Результаты сравнения моделей LCA с разным количеством классов представлены в табл. 1.

Таблица 1

**Результаты сравнения качества моделей**

| | **Энтропия** | **Минимальная пропорция тестируемых в классе, %** | **AIC** | **BIC** | **Статистика LMR** |
|---|---|---|---|---|---|
| C=2 | 0.833 | 31 | 26075.438 | 26160.884 | 1509.007* |
| C=3 | 0.825 | 6 | 25911.146 | 26042.602 | 175.369* |
| C=4 | 0.573 | 6 | 25865.959 | 26043.425 | 58.217* |

* — $p<0.01$

Предпочтение было отдано двух-классовому решению с наивысшим показателем энтропии модели (0.833). Трех-классовое решение было отклонено по критерию минимального количества тестируемых в классе, четырех-классовое решение отличалось низким качеством разбиения монстров на группы.

Обратимся к характеристикам групп в двух-классовом решении. Монстры из первой группы (69% от всех монстров) отличаются низкой выраженностью оригинальности, согласно средним значениям индикаторов. Монстры из второй группы (31% от всех монстров) характеризуются отличием от референса, то есть нарушением симметрии конечностей, расположением глаз в нижней части монстра и пр. По всем индикаторам вероятность построить оригинального монстра во второй группе выше, чем в первой. Полученные результаты позволяют говорить о том, что монстры в первом классе «скорее не оригинальны», а монстры во втором классе «скорее оригинальны».

Аналогичный анализ был проведен для выявления латентных классов с различным уровнем детальности как субкомпетенции креативности. В результате анализа было выявлено два класса монстров. Первый класс характеризовался высоким уровнем детальности по всем индикаторам (скорее детальны), а второй класс низким уровнем детальности (скорее не детальны).

Полученные разбиения монстров на классы являются начальной разметкой для обучения нейронной сети и применения машинного обучения — второго этапа анализа.

Качество работы нейронной сети для идентификации оригинальности изображений было проверено на тестовом наборе данных, состоящим из 749 изображений (345 — оригинальных, 404 — неоригинальных), при помощи матрицы ошибок (Confusion Matrix). Матрица ошибок — это таблица, которая позволяет визуализировать эффективность алгоритма классификации путем сравнения прогнозируемого значения целевой переменной с ее фактическим значением.

Таблица 2

**Матрица ошибок на тестовом наборе данных**

| | | Предсказанные значения | |
|---|---|---|---|
| | | 1 | 0 |
| Истинные значения | 1 | 323 | 40 |
| | 0 | 22 | 364 |

Из матрицы ошибок видно, что всего лишь 62 изображения из 749 были неправильно классифицированы. Так как классы достаточно сбалансированы, то точность (Accuracy) может быть использована, как метрика оценки классификатора:

$$\text{Accuracy} = \frac{323 + 364}{323 + 364 + 22 + 40} = 0{,}92$$

Также, была построена и обучена нейронная сеть для идентификации детальности изображений. Модель была апробирована на выборке, состоящей из 793 изображений (370 — детальных, 423 — не детальных). Точность на данной выборке составила 85%.

Таким образом, в данной работе мы продемонстрировали использование автоматического анализа изображений с применением методов машинного обучения для получения качественных, объективных и масштабируемых результатов тестирования.

Необходимо отметить, что для успешного воспроизведения экспертной деятельности автоматизированными процессами машинного обучения необходимо обеспечить надлежащее качество данных и предоставить качественную изначальную разметку для обучения сети. Оба этих фактора обладают определенной спецификой в образовательном тестировании.

Качество данных в образовательном тестировании зависит от качества самого инструмента измерения. Предыдущие исследования анализа качества инструмента измерения креативности 4К доказал его хорошие психометрические характеристики — было доказано, что данные образуют два фактора (оригинальность и детальность), каждый из которых в достаточной мере объясняет наблюдаемое поведение [8].

Использование методологии латентного классового анализа позволило получить качественную разметку для обучения нейронной сети. Модель LCA продемонстрировала, что оптимальным является разбиение на два класса по уровню выраженности оригинальности и детальности соответственно.

Среди ограничений работы отметим, во-первых, небольшой размеры выборки для обучения нейронной сети, а, во-вторых, принадлежность

трех монстров к одному ученику. Таким образом, мы можем ожидать, что изображения находятся в большем согласии между собой, чем если бы каждое изображение принадлежало независимому создателю. Построение трех изображений одним учеником — необходимое решение для получения точных оценок креативности, поэтому в дальнейших исследованиях, направленных не только на анализ отдельных изображений, но и на предоставление индивидуальной обратной связи тестируемым, мы считаем важным учитывать такой общий источник дисперсии. Отметим, что валидизация и обоснование теоретической рамки креативности не входило в фокус данного исследования.

Работа вносит вклад в развитие междисциплинарных исследований — науке о данных и науке об образовании. Использование методов машинного обучения в образовательном тестировании — совсем новая, но перспективная область для решения исследовательских и прикладных задач [9]. Как показывает наша работа, методы машинного обучения имеют масштабные перспективны для оценки сложных образовательных и психологических характеристик, которые на сегодняшний день являются частью образовательных стандартов (ФГОС, 2008).

**Литература**

1. *Griffin P.,* & *Care E.* (Eds.). (2014). Assessment and teaching of 21st century skills: Methods and approach. Springer.
2. *Shao Z., Li Y., Wang X., Zhao X., & Guo Y.* (2020). Research on a new automatic generation algorithm of concept map based on text analysis and association rules mining. Journal of Ambient Intelligence and Humanized Computing, 11 (2), 539—551.
3. *Liu O.L., Rios J.A., Heilman M., Gerard L.,* & *Lin, M.C.* (2016). Validation of automated scoring of science assessments. Journal of Research in Science Teaching, 53 (2), 215—233.
4. *Leacock C., Chodorow M.* (2003). C-rater: Automated Scoring of Short-Answer Questions. Computers and the Humanities 37, 389—405.
5. *Lubke G.H.,* & *Muthén B.* (2005). Investigating population heterogeneity with factor mixture models. Psychological methods, 10 (1), 21.
6. *Богоявленская Д.Б.* Психология творческих способностей / Д.Б. Богоявленская. М.: Федоров, 2009. С. 342.
7. *Kaufman J.C., Beghetto R.A.* (2009) Beyond big and little: The four c model of creativity. Review of general psychology. Т. 13. № 1. P. 1—12.
8. *Uglanova I., Vasin G., Brun I.* Developing a conceptual framework of creativity and critical thinking: evidence from a validity study, 19th annual AEA-Europe conference 2018 (Arnhem — Nijmegen).
9. *Polyak S.T., von Davier A.A.,* & *Peterschmidt K.* (2017). Computational psychometrics for the measurement of collaborative problem solving skills. Frontiers in Psychology, 8.

***Сведения об авторах***

*Тарасов Сергей Владимирович*, стажер-исследователь Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0003-4151-115X, e-mail: svtarasov@hse.ru

*Гельвер Евгений Сергеевич*, аналитик Института образования, Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0001-9365-801X, e-mail: egelver@hse.ru

*Грачева Дарья Александровна*, стажер-исследователь Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-4646-7349, e-mail: dgracheva@hse.ru

*Угланова Ирина Львовна*, стажер-исследователь Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0001-9117-5997, e-mail: iuglanova@hse.ru

*Вырва Елена Евгеньевна*, стажер-исследователь Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0001-8174-421X, e-mail: evyrva@hse.ru