СРАВНЕНИЕ АЛГОРИТМОВ КОМПЬЮТЕРНОГО ТЕСТИРОВАНИЯ, БАЗИРУЮЩИХСЯ НА РАЗЛИЧНЫХ МЕТОДАХ

А. И. Безруков[1], Л. В. Грахольская[2]

[1]*Саратовский государственный технический университет им. Ю. А. Гагарина, Россия*
[2]*Поволжский институт управления им. П. А. Столыпина – филиал Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации, Саратов, Россия*
E-mail: bezr_alex@mail.ru, graholskayalv@yandex.ru

Рассматриваются результаты имитационного моделирования различных методов компьютерного тестирования. Приводятся сравнительные характеристики алгоритма, использующего метод максимального правдоподобия и байевского алгоритма адаптивного тестирования. Даются рекомендации по выбору алгоритма компьютерного тестирования и его параметров, в зависимости от целей и обстоятельств проведения тестирования.

COMPARISON OF COMPUTER TESTING ALGORITHMS BASED ON DIFFERENT METHODS

A. I. Bezrukov, L. V. Graholskaya

The results of simulation of various methods of computer testing are considered. Comparative characteristics of the algorithm using the maximum likelihood method and the Bayes adaptive testing algorithm are presented. Recommendations are given on the choice of the computer testing algorithm and its parameters, depending on the goals and circumstances of testing.

Компьютерное тестирование является удобной и мало затратной формой оперативной проверки знаний и компетенций студентов. Важность и полезность компьютерного тестирования особенно ярко проявились в условиях удаленного обучения, когда применение других форм контроля затруднено или вообще невозможно.

Однако, применение компьютерного тестирования на практике вызывает множество споров и возражений. Большинство из них связано с несовершенством используемых тестов и методов обработки результатов тестирования, реализованных в популярных системах MOODLE и АСТ.

Активно развивающиеся в настоящее время математические модели, методы и алгоритмы организации компьютерного тестирования [1, 2] позволяют существенно повысить достоверность и, одновременно, снизить трудоемкость проведения тестирования. В данной работе рассматриваются классический алгоритм тестирования с обработкой результатов методом максимального правдоподобия и байевский алгоритм адаптивного тестирования.

Цель данной работы - сравнить «потребительские характеристики» исследуемых алгоритмов и дать практические рекомендации по выбору алгорит-

ма и его параметров с учетом целей и обстоятельств проведения тестирования.

В классическом (неадаптивном) алгоритме компьютерного тестирования студенту предлагается выполнить заранее определенный перечень заданий, уровень его подготовленности θ оценивается по тому, какие задания он успешно выполнил. Предполагается, что вероятность успешного выполнения задания, зависящая от θ и трудности задания δ оценивается трехпараметрической моделью Раша (моделью Бирнбаума) [3]:

$$P(\theta,\delta) = c + (1-c)\cdot\frac{\exp\left(\alpha\cdot(\theta-\delta)\right)}{1-\exp\left(\alpha\cdot(\theta-\delta)\right)},$$

где *c*- вероятность угадывания правильного ответа без выполнения задания а $\alpha \approx 1{,}71$ – чувствительность задания.

С математической точки зрения, алгоритм позволяет оценить латентную характеристику θ, как параметр функции распределения вероятности. Согласно многочисленным публикациям, посвященным проблемам компьютерного тестирования, наилучший результат оценки уровня подготовленности получается при применении метода максимального правдоподобия (Maximum likelihood method, (MLM)) [4].

В отличие от описанного выше классического алгоритма компьютерного тестирования, при устном опросе преподавателю обычно быстро становится понятно, насколько силен опрашиваемый студент. Поэтому, сильному студенту нет смысла задавать простые вопросы, а слабому – сложные. Выбирая следующее задание, преподаватель стремится уточнить уровень подготовленности студента. Для реализации такой стратегии были разработаны алгоритмы адаптивного тестирования [5,6]. Если до начала тестирования преподавателю известны результаты предыдущих тестирований, считают, что задано начальное значение θ для каждого студента, если же таких данных нет, предполагается, что все студенты обладают неким средним уровнем.

В качестве первого вопроса каждому студенту предлагаются задание, выполнение или невыполнение которого даст максимальную информацию об уровне его подготовленности. Выбор каждого следующего задания также направлен на получение максимально возможной информации уровне его подготовленности, но зависит от того, как студент справился с предыдущим вопросом. Наиболее эффективным алгоритмом, реализующим описанную методику, является байесовский алгоритм.

В отличие от классического алгоритма, в байевском алгоритме решается задача классификации. Результатом работы алгоритма на каждом шаге *n* является не оценка уровня подготовленности, а набор вероятностей $\left\{\mathrm{P}_1^{(\mathrm{n})}, \mathrm{P}_2^{(\mathrm{n})}, \ldots \mathrm{P}_\mathrm{M}^{(\mathrm{n})}\right\}$ принадлежности тестируемого студента каждому из заранее определенных классов. Алгоритм останавливается, когда достигается указанный максимум заданий в тесте или максимум различия вероятностей, полученных на предыдущем и последующим шаге, становится меньше заданного значения *t* - параметра остановки алгоритма:

$$\left|P_i^{(n+1)} - P_i^{(n)}\right| < t.$$

Уровень подготовки θ становится дискретной величиной, определенной для каждого класса. Поэтому сравнивать точностные характеристики результатов тестирований, проведенных по разным алгоритмам весьма затруднительно. К тому же вряд ли стоить тестировать одних и тех же реальных студентов по одной теме различными методами. Поэтому, реальные данные, необходимые для сопоставления методов тестирования, найти трудно.

Для сравнения алгоритмов и оценки влияния их параметров на качество тестирования была использована ранее разработанная нами имитационная модель [7]. Транзактами модели являются «студенты» с заданными характеристиками уровня подготовленность и «задания» с заданными уровнями сложности. Результат выполнения данного задания данным студентом разыгрывается с учетом модели Бирнбаума. Многократные прогоны модели с одним и тем же набором данных позволяют выявить статистически значимые зависимости результатов тестирования от характеристик банка тестовых заданий и методов проведения тестирования. При этом, дисперсия характеристик, полученных при различных прогонах модели, может использоваться для оценки точности и достоверности оценок этих характеристик.

Рассмотрим зависимость среднего отклонения результата тестирования от истинного, заданного в модели значения (средней ошибки тестирования) от параметра *t*.

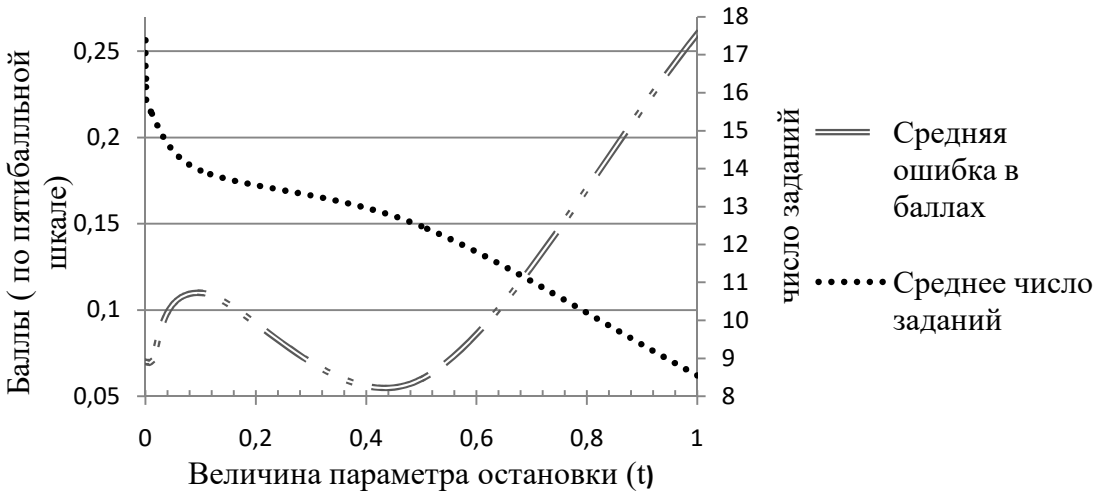

Рис.1. Зависимость средней по прогонам ошибки тестирования и среднего количества заданий в тесте от параметра остановки алгоритма Байеса

На рис. 1 видно, что увеличение *t* приводит к росту ошибки тестирования. Но чем больше параметр *t,* тем быстрее останавливается алгоритм, следовательно, тем меньше число выполненных заданий. Для определения оптимального значения *t* нужно решить, что для нас важнее: сократить ошибку тестирования или уменьшить число выполняемых заданий. На рис. 2 приведены те же графики, каждый из которых нормирован на 100%.

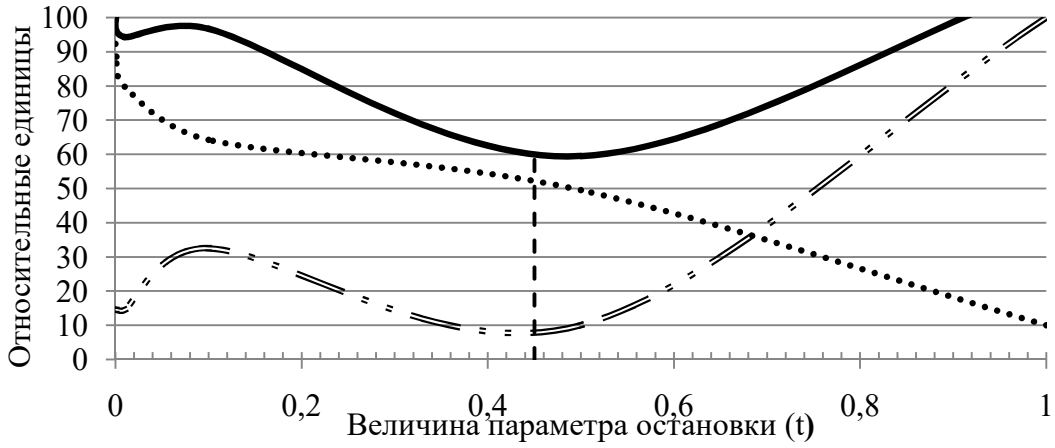

Рис. 2. Определение оптимального значения параметра остановки

Такая нормировка соответствует предположению, что нам одинаково важны оба критерия. Жирная линия – сумма нормированных значений, имеет минимум при $t \approx 0{,}43$. Это и есть оптимальное значение параметра остановки.

Ранее мы полагали, что вероятность угадывания правильного ответа без выполнения задания равна нулю. Однако в реальных тестах она всегда больше нуля. Рассмотрим, как изменяются ошибки тестирования различных методов при увеличении вероятности угадывания. Для сравнения рассмотрим случай, когда вероятность угадывания *c* равна нулю и 0,4.

На рис. 3 представлены зависимости средних значений оценок уровня подготовленности студентов, полученных методом ММП от истинных (заданных) значений. Пунктиром выделены доверительные области ($\pm 3\sigma$).

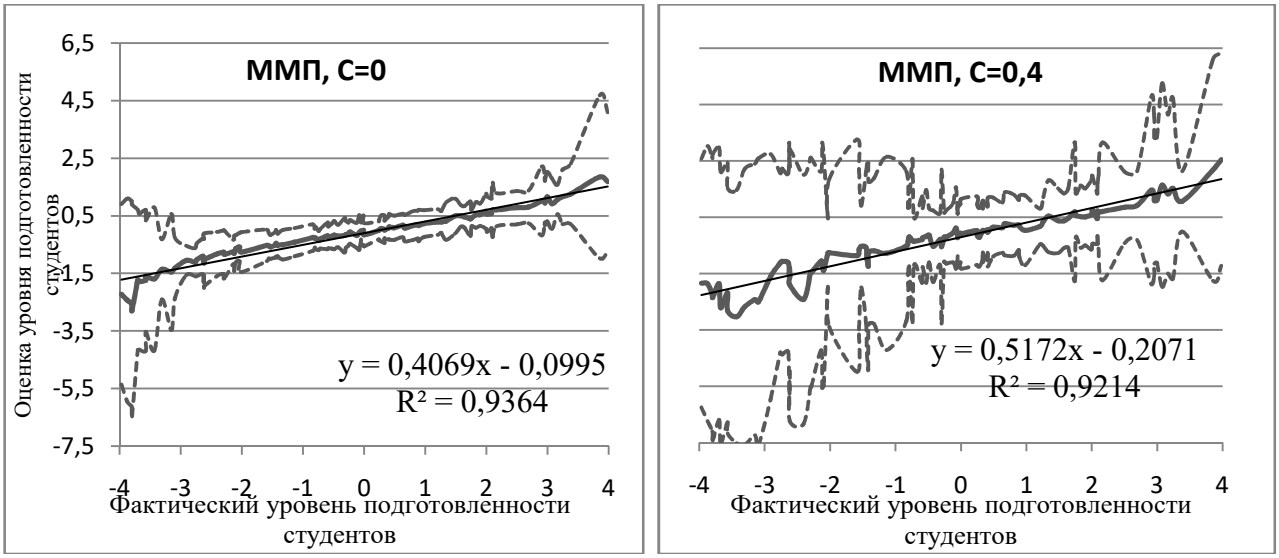

Рис. 3. Связь оценочного и фактического уровней подготовленности (метод максимального правдоподобия)

Из графиков на рис. 3 видно, что с увеличением вероятности угадывания расширяются границы доверительной области, т.е. снижается достоверность

оценки θ. Наибольшая достоверность оценки θ достигается в середине интервала, при приближении к краям интервала существенно снижается.

На рис. 4 представлены те же зависимости для оценок θ полученных алгоритмом Байеса.

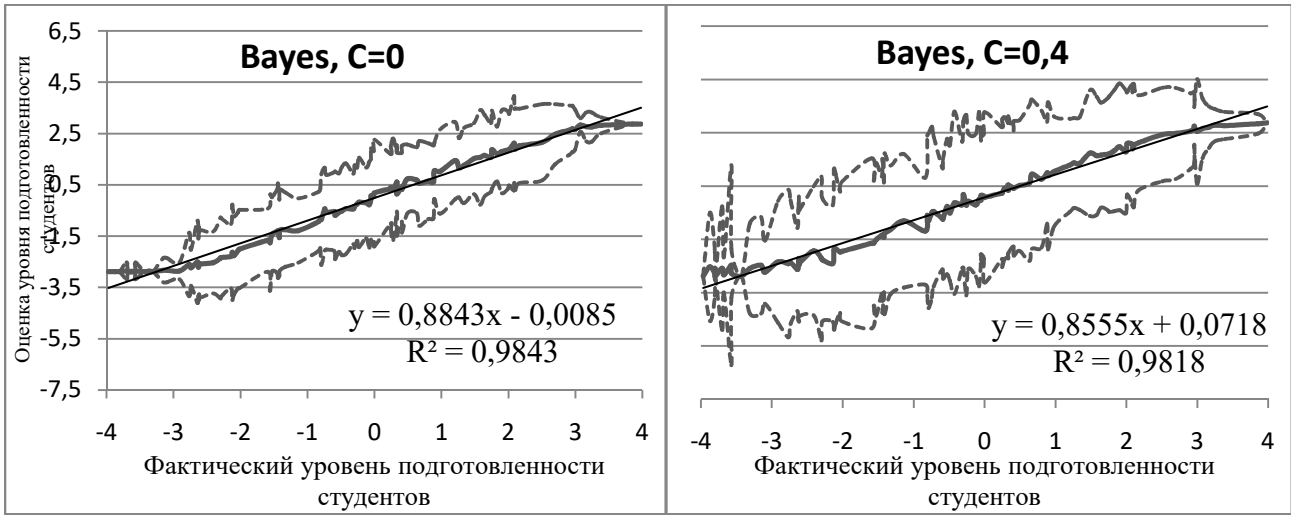

Рис. 4. Связь оценочного и фактического уровней подготовленности (метод Байеса)

Как видно, байесовский алгоритм обеспечивает вчетверо меньшую ошибку, чем классический.

Тестирование следует рассматривать не как процедуру выявления латентного параметра θ для каждого студента, а как классификационную процедуру отнесения студента к одному из заранее выбранных классов. Современные алгоритмы позволяют выразить результат тестирования в виде нечеткого утверждения, что существенно повышает информативность результата в неоднозначных случаях.

Численные эксперименты на имитационной модели продемонстрировали явное преимущество байевского алгоритма. Его применение позволит одновременно сократить количество выполняемых заданий и повысить достоверность результатов тестирования.

Понимание целей проведения тестирования позволяет сформулировать требования к его результатам и выбрать оптимальные параметры алгоритма тестирования для каждого конкретного случая. Будущие системы компьютерного тестирования должны быть снабжены инструментом настройки, позволяющим в диалоговом режиме уточнять цели, обстоятельства и требования к планируемому тестированию и, на основании этого, подбирать алгоритм и его параметры проведения.

Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 20-013-00783 «Развитие методов анализа данных для оценки компетенций, формируемых в процессе обучения».

СПИСОК ЛИТЕРАТУРЫ

1. *Ivailo Partchev* A visual guide to item response theory // Jena : Friedrich-Schiller-Universitat, 2004. 61 p. [Электронный ресурс]. URL: https://docplayer.net/20748000-A-visual-guide-to-item-response-theory.html (дата обращения: 01.10.2021).

2. *Mike Wu, Richard L. Davis, Benjamin W.* Variational Item Response Theory: Fast, Accurate, and Expressive. Domingue, Chris Piech, Noah Goodman // Department of Computer Science, Education, and Psychology. Stanford University / {wumike, rldavis, bdomingu, cpiech, ngoodman}@stanford.edu

3. *Hambleton R., Swaminathan H.* Item response theory : Principles and applications. Kluwer, 1985.

4. *Аванесов В. С.* Композиция тестовых заданий / Учеб. книга. 3 изд.. доп. М. : Центр тестирования, 2002. 240 с.

5. *Летова Л. В.* Точность моделирования латентных переменных с помощью модели Раша (часть 1) // Современные научные исследования и инновации. 2014. Ч. 1. № 6. [Электронный ресурс]. URL: https://web.snauka.ru/issues/2014/06/34399 (дата обращения: 05.10.2021).

6. *Деменчёнок О.* Анализ моделей для адаптивного тестирования // Педагогические измерения. 2011. № 1. С. 3-18.

7. *Безруков А. И., Грахольская Л. В.* Имитационная модель для выбора стратегии адаптивного тестирования // Математическое и компьютерное моделирование в экономике, страховании и управлении рисками. 2020. № 5. С. 145-151.