Математические методы современной теории тестирования

Ахламов А.Г., Белоус Н.В., Пархоменко С.А., Бекетова Е.А.
Харьковский национальный университет радиоэлектроники,
Украинская Академия государственного управления при Президенте Украины,
Одесский региональный институт государственного управления,
Харьков, Украина
E-mail: masterofrings@ukr.net

Abstract

The work considers the problem of knowledge estimation for remote education. This article considers the modern mathematical models, which destined for reliable examination of test results and takes reliable level of knowledge for students.

Дистанционное обучение – одно из наиболее быстро развивающихся направлений системы образования. Это качественно новый прогрессивный вид обучения, базирующийся на современных информационных технологиях. Дистанционное обучение позволяет выбрать удобное место и время для обучения; получить образование лицам, лишенным возможности получить традиционное образование в силу тех или иных причин. Особо важен в дистанционном обучении регулярный, объективный и полный контроль за степенью усвоения материла, который осуществляется с помощью тестирования. Почти каждый педагог пытался разрабатывать тестовые задания по своей дисциплине, но неумение правильно обработать результаты теста и интерпретировать их сводят ценность такой разработки почти к нулю. Напротив, грамотное конструирование теста на основе знания теории тестирования позволит создать инструмент, позволяющий провести объективное измерение обученности по данному курсу с необходимой точностью.

В настоящее время существуют два теоретических подхода к созданию тестов: классическая теория и современная теория IRT (Item Response Theory). Оба подхода базируются на последующей статистической обработке так называемого сырого балла (raw score), то есть балла, набранного в результате тестирования. Только после проведения многократных статистических обработок можно говорить о создании теста с устойчивыми параметрами качества (надежностью и валидностью). Однако классическая теория имеет ряд недостатков, главный из которых - зависимость результатов измерения от инструмента измерения (конкретного теста). Неудовлетворенность такой ситуацией и привела к созданию IRT.

Можно выделить следующие модели:

- однопараметрическая модель G.Rasch;
- двухпараметрическая модель A.Birbaum;
- трехпараметрическая модель A.Birbaum.

Аналитическое задание однопараметрической модели представлено формулой:

$$P_j(\Theta) = 1/\{1+\exp[-1{,}7(\Theta-\beta_j)]\}, \quad (1)$$
$$P_i(\beta) = 1/\{1+\exp[-1{,}7(\Theta i-\beta)]\};$$

где Θ_i – уровень знаний i-го испытуемого; β_j – уровень трудности j-го задания теста; P_j - вероятность правильного выполнения j-го задания теста; P_i - вероятность правильного выполнения i-м испытуемым различных по трудности заданий, Θ и β – независимые переменные для первой и второй функций соответственно.

P_j является возрастающей функцией переменной Θ. Это свойство функции P_j легко интерпретируется и согласуется с практическим опытом педагога. Естественно ожидать, что чем выше уровень знаний испытуемого, тем больше вероятность правильного выполнения им j-го задания теста.

Вероятность правильного выполнения i-м испытуемым различных по трудности заданий P_i является убывающей функцией переменной β. Это означает, что с ростом трудности заданий значения вероятности будут уменьшаться.

Для построения характеристических кривых заданий теста и индивидуальных кривых испытуемых необходимо знать значения параметров Θ и β. Оценка параметров проводится в предположении нормальности распределений эмпирических данных тестирования как по множеству испытуемых, так и по множеству заданий. Считаются нормально распределенными и значения латентных переменных. В процессе разработки теста приходится оценивать оба параметра: Θ и β

Начальная оценка уровня знаний i-го испытуемого в логитах находится по формуле

$$\Theta_i^0 = \ln(p_i/q_i),\ i=1,2,..,N, \quad (2)$$

где N – число испытуемых, p_i – доля правильных ответов i-го испытуемого на все задания теста, q_i – соответственно доля неправильных ответов, причем $p_i=1-q_i$

Аналогично, начальное значение параметра β_j^0 в логитах определяют как

$$\beta_j^0 = \ln(q_j/p_j),\ j=1,2,..,n, \quad (3)$$

где n – число заданий, p_j – доля правильных ответов всех испытуемых группы на j-е задание теста, q_j – доля неправильных ответов, причем $p_j=1-q_j$.

Теоретически значения параметров Θ и β могут меняться в интервале $(-\infty, +\infty)$, т.е. $-\infty<\Theta<+\infty$ и аналогично $-\infty<\beta<+\infty$. Но практически при $\beta<-6$ значения P_i близки к единице ($P_i=0{,}999...$). С этими заданиями в тесте справляются все, и они оказываются просто лишними. В равной мере бесполезны задания при $\beta>6$. С такими заданиями не может справиться ни один испытуемый в группе, и они не несут никакой информации о различиях в знаниях студентов. Однако даже этот интервал рекомендуется еще уменьшить и рассматривать только задания, для которых значения параметра трудности

находятся в интервале (–3; +3). Аналогичные рассуждения можно было бы провести для переменной Θ.

Затем начальные значения логитов уровней знаний и логитов трудностей заданий сводятся в одну интервальную шкалу. В формуле для такого искусственного перевода значений Θ_i^0 в Θi заложена идея уничтожения эффекта влияния трудности заданий на оценки тестируемых студентов. На этом этапе оценки параметра Θ в логитах вычисляются по формуле:

$$\Theta_i = \bar{\beta} + X \ln\frac{p_i}{q_i} \text{ или } \Theta_i = \bar{\beta} + X\Theta_i^0, \; i=1,2,\ldots,N, \qquad (4)$$

где $X = \left(1+\frac{W^2}{2.89}\right)^{1/2}$, $\bar{\beta}$ – среднее значение логитов трудности заданий теста, W – стандартное отклонение распределения начальных значений параметра β, N – число испытуемых.

Эта формула позволяет получить объективную оценку уровня знаний каждого испытуемого, не зависящую от трудности заданий, включенных в тест. На основе таких оценок можно корректно сравнить уровни знаний испытуемых, выполнивших различные по трудности задания теста и даже разные тесты.

Объективное значение переменной βj для j-го задания теста можно найти по формуле

$$\beta_j = \bar{\Theta} + Y \ln\frac{q_j}{p_j} \text{ или } \beta_j = \bar{\Theta} + Y\beta_j^0, \; j=1,2,\ldots,n, \qquad (5)$$

где $Y = \left(1+\frac{V^2}{2.89}\right)^{1/2}$, $\bar{\Theta}$ – среднее значение логитов уровней знаний, V – стандартное отклонение распределения начальных значений параметра Θ, n – число заданий в тесте.

Эта формула позволяет получить устойчивые оценки параметра β, не зависящие от свойств выборки испытуемых.

В случае двухпараметрической модели A.Birnbaum условную вероятность правильного выполнения j-го задания теста испытуемыми с различными значениями Θ можно записать в виде

$$Pj\{x_{ij} = 1|\beta_j\} = 1/\{1+\exp[-1,7a_j(\Theta-\beta_j)]\}, \qquad (6)$$

где кроме прежних обозначений вводится новое a_j для второго параметра j-го задания теста.

При геометрической интерпретации первый параметр βj можно рассматривать как характеристику положения кривой j-го задания относительно оси Θ, второй параметр a_j связан с крутизной кривой задания в точке ее перегиба. А именно, значение a_j прямо пропорционально тангенсу угла наклона касательной к характеристической кривой задания теста в точке $\Theta = \beta_j$. Это означает, что более крутые кривые соответствуют большим значениям a_j, соответственно для пологих кривых $a_j \to 0$. Теоретически значения параметра a_j могут изменяться в интервале (-∞, +∞), но практически далеко не все эти

задания можно помещать в тест. Анализ характеристических кривых заданий одинаковой трудности, но разной крутизны позволяет отобрать лучшие задания и определить разумные границы интервала для значений параметра a_j.

Значения aj близкие к нулю, соответствуют случаю, когда испытуемые с разными уровнями знаний правильно отвечают на j-е задание с приблизительно равной вероятностью, что, естественно, противоречит ожидаемым прогнозам разработчика теста. Эти задания оказываются бесполезными при дифференциации испытуемых группы по оцениваемому параметру, так как они не несут информации об индивидуальных различиях испытуемых.

Еще более бесполезны задания с отрицательными значениями aj, на них отвечают правильно с большой вероятностью испытуемые с низким уровнем знаний, а для знающих студентов с большими значениями Θ вероятность правильного ответа стремится к нулю. Число заданий в тесте должно сокращаться в первую очередь за счет устранения таких неудачных заданий в тестовой форме. Как правило, такое сокращение приводит к повышению надежности и валидности теста.

Отбор заданий с большими значениями aj является одним из важных принципов при подготовке эффективного теста. Для сокращения теста за счет удаления части заданий равной трудности сравнительный анализ крутизны характеристических кривых с одинаковой точкой перегиба позволяет выделить одно, наиболее эффективное задание с наибольшим значением параметра aj. На практике рекомендуется, как правило, оставлять задания со значениями aj, лежащими в интервале (0,5; 3). Значение aj=1 соответствует однопараметрической модели G.Rasch.

Трехпараметрическая модель. Для тестов с заданиями в закрытой форме было отмечено существенное отклонение эмпирических данных от теоретической кривой, предсказывающей вероятность правильного выполнения задания при различных значениях переменной Θ. Этот эффект наиболее характерен для испытуемых с низкими значениями параметра Θ при ответах на наиболее трудные задания теста. Попытки выяснить причины такого отклонения привели ряд создателей современной теории тестов к выводу о влиянии эффекта угадывания правильного ответа на достоверность эмпирических данных.

Возможно, что испытуемые с различным уровнем знаний пользуются различными методами при выборе правильного ответа. Вернее, методом пользуются только те, кто обладает достаточными знаниями для правильного выбора. Другие же, знания которых характеризуются низкими значениями параметра Θ, просто угадывают правильный ответ на задание. И чем труднее задание, тем вероятнее, что ответ получен именно таким путем. Для того чтобы учесть фактор угадывания и была предложена трехпараметрическая логистическая модель.

В этом случае вероятность правильного ответа испытуемых на j-е задание теста находят по формуле (7):

$$P_{ij}=c_j+(1-c_j)/\{1+\exp[-1,7a_j(\Theta_i-\beta_j)]\} \qquad (7)$$

где кроме прежних обозначений введен третий параметр cj, характеризующий вероятность правильного ответа испытуемых на j-е задание теста при полном отсутствии знаний у тестируемых студентов ($\Theta \to -\infty$).

Однако следует заметить, что введение третьего параметра cj не только существенно снижает точность оценок параметров Θ и β, но и ухудшает сходимость итерационных методов, используемых при оценивании объективных значений латентных переменных Θ и β.

На основе вышеизложенного материала была разработана программа "TestStat", которая, предназначена для обработки результатов тестирования и определения качества теста, его надежности, а также выдачи рекомендаций для коррекции заданий теста с целью его улучшения с точки зрения качества. В основе данного программного продукта заложена однопараметрическая модель G. Rasch.

Программа "TestStat" может быть использоваться как самостоятельно так и в качестве одной из составных частей какого-нибудь программного комплекса, предназначенного для автоматизированного создания тестовых заданий и тестирования уровня знаний испытуемого. Для этого программа реализует механизм OLE-автоматизации и предоставляет интерфейс при помощи которого главная программа может взаимодействовать с функциональными возможностями программы "TestStat", такими как:

- загрузка матрицы результатов тестирования (результаты тестирования должны быть представлены в дихотомическая шкале: 1 - знает, 0 - не знает);
- расчет основных статистических параметров;
- получение оценки испытуемого по выбранной шкале оценок (шкала от 5 до 255 баллов);
- рекомендации о несостоятельных заданиях, а также заданий плохого качества;
- определение надежности теста;
- прогнозирование надежности теста;
- латентный анализ полученных результатов.
-

Анализ реализованной модели показал, что однопараметрическая модель G. Rasch полностью не решает поставленной задачи. Это связано с определенными ограничениями, накладываемыми на крутизну характеристических кривых заданий в рамках данной модели. В частности, она считается одинаковой у всех кривых, что, конечно, обеспечивает определенную простоту в практических приложениях модели G.Rasch, но вместе с тем является и недостатком. Этот недостаток особенно заметен, когда нужно отдать предпочтение одному из заданий равной трудности. В этом случае можно легко прийти к неверному решению и существенно снизить надежность и валидность теста, удалив задания с более крутыми характеристическими кривыми, а оставив с более пологой. Для того чтобы избежать таких случаев необходимо использовать двух либо трехпараметрическую модель A.Birnbaum.

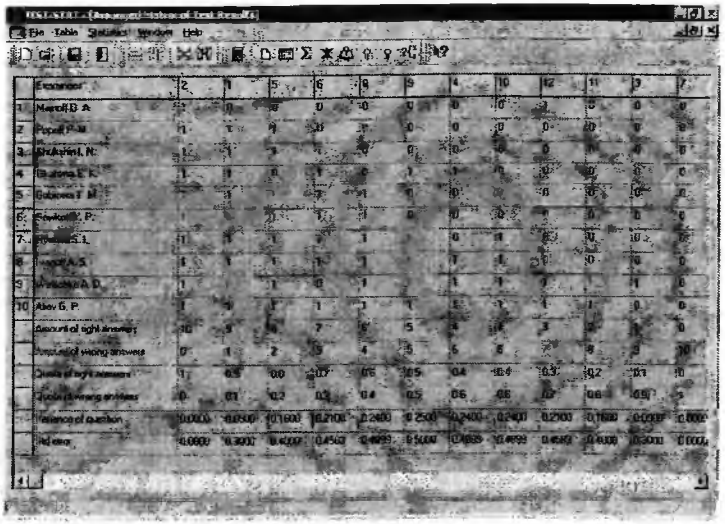

Рис. 1. Внешний вид программы "TestStat"

Литература

1. Rasch, G. Probabilistic Models for Some Intelligence and Afteword by B.D.Wright. The Univ. of Chicago Press. Chicago & London, 1980.
2. Lord F.M., Novick M. Statistical Theories of Mental Test Scores. Addison-Wesley Publ. Co. Reading, Mass., 1968.
3. From p-values and raw score statistics to logits. Stenner AJ, Wright BD, Linacre JM. 1994.

— ◙ —

Система дистанционного контроля успеваемости студентов

Азаренков В.И., Менделис В.В., Былым Т.Ю.
Харьковский национальный университет радиоэлектроники,
Харьков, Украина
E-mail: tvicg@knure.kharkov.ua

Abstract

The questions of creation of tool environment for testing of student's knowledge are considered. The software system for testing of distant student's knowledge was developed. It includes client and server parts. The content of the material is represented with laboratory works, theoretical materials and the tests. It is provided the ability of changing the labs, tests and theoretical materials.