УДК 681.3: 658.56

В.Е. Бахрушин, С.В. Журавель, М.А. Игнахина

Автоматизация определения результатов тестирования

Рассмотрены алгоритмы автоматизированного определения результатов тестирования, основанные на использовании обучающих выборок. Их применение дает возможность уменьшить субъективизм при построении шкалы оценок и повысить качество оценивания.

Some algorithms of the testing results automated determining based on the use of learning samples are considered. Their application gives an opportunity to decrease the subjectivism at constructing a mark scale and to raise the quality of the estimation.

Розглянуто алгоритми автоматизованого визначення результатів тестування, що базуються на використанні навчаючих вибірок. Їх застосування дає можливість зменшити суб'єктивізм при побудові шкали оцінок та підвищити якість оцінювання.

Введение. Широкое распространение новых, в том числе дистанционных и электронных форм обучения, а также недостатки традиционных форм оценивания результатов обучения (субъективизм, неравенство условий контроля, трудоемкость процедуры и др.) приводят к тому, что сегодня все шире начинают использоваться тестовые формы контроля знаний, теоретические основы которых рассмотрены в [1] и других. Они базируются на предположении, что результаты тестирования подчиняются нормальному закону распределения. Однако, как было показано ранее [2], такое предположение на практике выполняется далеко не всегда. Часто распределение результатов тестирования описывается другими однородными функциями распределения, в частности, логнормальной или может быть представлено в виде смеси нескольких однородных компонент. Вопрос об устойчивости традиционных параметрических методик обработки результатов тестирования к таким отклонениям исследован недостаточно. Однако известно [3], что отклонения данных от нормального закона распределения во многих случаях ведут к необходимости использования непараметрических статистических методов, а также некоторых параметров, применяемых для статистического описания результатов. В частности, может оказаться неправомерным использование таких величин, как среднее арифметическое и стандартное отклонение. Поэтому для обработки результатов тестирования целесообразно использовать методы, устойчивые к отклонениям данных от нормального распределения.

Наряду с несомненными достоинствами тестированию свойственны существенные недостатки. Среди них следует отметить трудность учета разнообразия личных особенностей тестируемых и целей тестирования. Так, по мнению директора Российского центра тестирования В. Хлебникова [4], Единый государственный экзамен (аналог украинского Внешнего независимого оценивания) годится только для средних детей, а высокие и низкие оценки характеризуются недопустимо большими погрешностями. Кроме того, в реальных системах тестирования назначение баллов за правильные ответы на отдельные вопросы обычно бывает субъективным и не всегда достаточно обоснованным. Об этом, в частности, свидетельствуют результаты анализа тестирования студентов во Львовском национальном университете им. И. Франко [5].

В связи с этим, целью данной статьи была разработка алгоритмов, предназначенных для автоматизации определения результатов тестирования и позволяющих повысить их объективность.

Алгоритм, основанный на использовании эмпирической функции распределения результатов тестирования

В основу данного алгоритма положена идея выделения нескольких классов тестируемых, различающихся значением признака «Уровень знаний», измеряемого в порядковой шкале. Типичный пример такого подхода – шкала, применяемая в Европейской кредитно-трансфертной системе *ECTS* [6].

В этом случае перевод первичных баллов в итоговые результаты тестирования выполняется по такому алгоритму.

1. Задаем функцию распределения итоговых результатов $G^*(y)$ в виде таблицы значений ее α-квантилей:

$$G^*(y_1)=\alpha_1;\ G^*(y_2)=\alpha_2;\ \ldots;\ G^*(y_k)=\alpha_k,$$

где y – итоговые баллы, k – максимальная оценка. Пример реализации такого подхода – таблица, предложенная в [1].

Таблица 1. Пример задания функции распределения итоговых результатов

Итоговый балл (y_i)	Лексико-оценочные эквиваленты	Z_i	$G^*(y_i)$
1	Низшая оценка	< – 2,25	0,01
2	Неудовлетворительно	–2,25 … – 1,75	0,04
3	Малоудовлетворительно	–1,75 … – 1,25	0,11
4	Удовлетворительно	–1,25 … –0,75	0,23
5	Ниже среднего	–0,75 … – 0,25	0,40
6	Средняя оценка	–0,25 … 0,25	0,60
7	Выше среднего	0,25 … 0,75	0,77
8	Хорошо	0,75 … 1,25	0,89
9	Очень хорошо	1,25 … 1,75	0,96
10	Отлично	1,75 … 2,25	0,99
11	Высшая оценка	> 2,25	1

2. Проводим тестирование обучающей выборки объемом n и строим эмпирическую функцию распределения итоговых результатов $F^*(x)$, где x – первичные баллы тестирования. При этом объем обучающей выборки должен быть не менее 50 студентов, поскольку в противном случае погрешность определения значений функции распределения будет слишком высокой [6]. Минимальный объем обучающей выборки увеличивается с ростом количества выделяемых классов.

3. Рассчитываем α-квантили функции $F^*(x)$: $x_1, x_2, \ldots, x_k$.

4. Строим шкалу соответствия, позволяющую осуществить переход от первичных баллов к итоговым результатам тестирования:

$$x < x_1 \ \Rightarrow\ y = y_1;$$
$$x \in [x_1, x_2) \ \Rightarrow\ y = y_2;$$
$$\ldots\ldots\ldots\ldots\ldots;$$
$$x \in [x_{k-1}, x_k] \ \Rightarrow\ y = y_k.$$

5. Периодически (например, по результатам тестирования $2n$, $5n$, $10n$ испытуемых) возвращаемся к п. 2 для корректировки эмпирической функции распределения и уточнения шкалы перевода.

Предложенный алгоритм – достаточно общий и может быть использован за отсутствием специальных требований к процедуре и результатам тестирования. При этом следует отметить, что благодаря использованию эмпирической функции распределения обучающей выборки, отпадает необходимость в использовании предположений о законе распределения исходных данных.

Рассмотренный алгоритм не учитывает процедуру назначения первичных баллов за правильные ответы на отдельные вопросы. Как правило, эту задачу решают эксперты. Однако при необходимости автоматизации соответствующую процедуру также можно формализовать. В простейшем случае, когда задания считаются равноценными по значимости, но различаются по сложности, можно использовать такой алгоритм.

1. Определяем доли тестируемых обучающей выборки (p_j), справившихся с каждым из заданий. Они характеризуют степень сложности заданий. Чем меньше величина p_j, тем сложнее для испытуемых соответствующее задание.

2. Строим шкалу начисления первичных баллов. В зависимости от целей тестирования может быть использована равномерная или неравномерная шкала перевода. Пример равномерной пятибалльной шкалы:

$$p_j < 0,2 \ \Rightarrow\ x_j = 5;$$
$$0,2 \le p_j < 0,4 \ \Rightarrow\ x_j = 4;$$
$$0,4 \le p_j < 0,6 \ \Rightarrow\ x_j = 3;$$
$$0,6 \le p_j < 0,8 \ \Rightarrow\ x_j = 2;$$
$$p_j \ge 0,8 \ \Rightarrow\ x_j = 1.$$

Пример неравномерной пятибалльной шкалы:

$$p_j < \overline{p} - 3\sigma_p \ \Rightarrow\ x_j = 5;$$
$$\overline{p} - 3\sigma_p \le p_j < \overline{p} - \sigma_p \ \Rightarrow\ x_j = 4;$$
$$\overline{p} - \sigma_p \le p_j < \overline{p} + \sigma_p \ \Rightarrow\ x_j = 3;$$
$$\overline{p} + \sigma_p \le p_j < \overline{p} + 3\sigma_p \ \Rightarrow\ x_j = 2;$$
$$p_j \ge \overline{p} + 3\sigma_p \ \Rightarrow\ x_j = 1.$$

Здесь $\overline{p}$ – среднее арифметическое величин p_j, σ_p – их стандартное отклонение.

Использование неравномерной пятибалльной шкалы позволяет более точно выделить группы наиболее сильных и наиболее слабых тестируемых.

Идея использования квантильных оценок для построения шкалы перевода первичных баллов в итоговые оценки не нова. Она используется, например, в украинской системе внешнего независимого оценивания (ВНО). Однако следует заметить, что в отличие от рассматриваемого алгоритма ВНО предполагает получение итоговых оценок в виде большого по объему набора значений числовых величин. Это некорректно статистически и требует серьезного обоснования условий, при которых такая операция не приводит к существенным ошибкам. Кроме того, рассмотренные варианты предлагаемого алгоритма дают возможность более точно выделять группы «сильных» и «слабых» студентов.

Алгоритм, основанный на предварительном описании классов эквивалентности

Во многих случаях возможно предварительное разбиение обучающей выборки на некоторые классы эквивалентности, соответствующие тем или иным качественно различным группам испытуемых. Например, могут быть сформированы классы «хороших», «средних» и «плохих» студентов; классы школьников, склонных к различным видам деятельности и т.п. Следует отметить, что при таком подходе возможно учитывать различный смысл понятий «хороший», «средний», «плохой» студент. В частности «хорошим» можно назвать студента, умеющего быстро выполнять большое количество относительно простых заданий, а можно того, кто способен выполнить сложное задание.

Можно ожидать, что при тестировании, например, по математике будут проявляться существенные различия между абитуриентами, «хорошими» с точки зрения успешности их дальнейшего обучения по разным направлениям подготовки (математика, физика, инженерные специальности, экономика). Это обусловлено как различием используемого разными приложениями математического аппарата, так и различиями психологических портретов «оптимальных» кандидатов.

В таких ситуациях для перевода первичных баллов в итоговые результаты тестирования желательно использовать алгоритмы, которые позволили бы определить степень принадлежности испытуемых к заданным классам. Примером может служить алгоритм, рассматриваемый далее.

1. Формируем обучающие выборки, соответствующие заданным классам эквивалентности. Объем каждой выборки должен быть не менее 50 человек.

2. Проводим тестирование на обучающих выборках.

3. Определяем характеристические параметры классов эквивалентности. Для нормально распределенных классов это могут быть средние арифметические и стандартные отклонения оценок за отдельные задания теста для каждого класса. В других случаях можно использовать такие параметры, как медиана, центр сгиба, размах и т.п.

4. Решаем задачу о принадлежности результатов испытуемого к каждому из классов эквивалентности на основе результатов сравнения результатов теста с характеристиками классов. В зависимости от характера имеющихся данных, в качестве меры сходства можно использовать коэффициент корреляции Пирсона, коэффициент детерминации, коэффициенты ранговой корреляции Спирмена и Кендалла, точечно-бисериальный коэффициент корреляции и др. [6].

5. Периодически модифицируем обучающие выборки и уточняем параметры классов.

Заключение. Предложенные в статье алгоритмы автоматизированного перевода первичных результатов тестирования в итоговые оценки, основанные на применении обучающих выборок позволяют повысить качество оценивания и учитывать различие целей тестирования при определении его результатов.

Окончание на стр. 21

1. *Аванесов В.С.* Научные основы тестового контроля знаний. – М.: Исследовательский центр, 1994. – 135 с.
2. *Бахрушин В.Е.*, *Игнахина М.А.*, *Шумада Р.Я.* Эмпирические функции распределения результатов тестирования // Зб. пр. III Міжнар. конф. «Нові інформаційні технології в освіті для всіх: система електронної освіти». – К.: МННЦ ІТС, 2008. – С. 79–84.
3. *Орлов А.И.* Прикладная статистика. – М.: Экзамен, 2006. – 671 с.
4. *Лемуткина М.* Единый государственный обмен. – http://www.gazeta.ru/education/2006/01/30_a_528596.shtml
5. *Петрів В.Ф.*, *Рикалюк Р.Є.* Візуалізація критеріїв якості тестових завдань // Зб. наук. пр. XVI Всеукр. наук. конф. «Сучасні проблеми прикладної математики та інформатики».– Львів: ЛНУ, 2009. – С. 166 – 169.
6. *Бахрушин В.Є.* Аналіз даних. – Запоріжжя: ГУ «ЗІДМУ», 2006. – 128 с.

E-mail: *Vladimir.Bakhrushin@zhu.edu.ua*