

Достовірність оцінки знань методом закритого тестування

Володимир Бондарев
кафедра програмної інженерії
Харківський національний університет
радіоелектроніки
Харків, Україна

volodymyr.bondariev@nure.ua

Олексій Галуза
кафедра комп'ютерної математики і аналізу даних
Національний технічний університет «Харківський
політехнічний інститут»
Харків, Україна
alexey.galuz@nure.ua

Reliability of knowledge assessment by multiple-choice tests

Volodymyr Bondariev
dept. of Software Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
volodymyr.bondariev@nure.ua

Alexey Galuz
dept. of Computer Mathematics and Data Analysis
National Technical University "Kharkiv Polytechnic
Institute"
Kharkiv, Ukraine
alexey.galuz@gmail.com

Анотація—Тести закритого типу розглядаються з ймовірнісної точки зору, що дозволяє оцінювати їх достовірність в термінах вибіркового середнього і довірчого інтервалу. Пропонується спосіб оцінки реальної компетентності випробуваного, який враховує можливість випадкового вибору вірних відповідей.

Ключові слова—контроль знань; тест закритого типу; біноміальний розподіл; довірчий інтервал

Abstract—Multiple-choice tests are considered from a probabilistic point of view, which makes it possible to assess their reliability in terms of the sample mean and confidence interval. A method is proposed to assess the subject's real competence, which considers the possibility of randomly choosing the correct answers

Keywords—knowledge assessment; multiple-choice test; binomial distribution; confidence interval

I. ВСТУП

Контроль знань за допомогою тестів завжди був затребуваним в сфері освіти. Особливо часто вдаються до тестів закритого типу [1], оскільки і проведення, і перевірка таких тестів може бути повністю автоматизована.

У зв'язку з цим хотілося б прояснити, що саме і з якою точністю вимірюють подібні тести [2]. Зрозуміло, завжди можна сказати, на яку частку питань випробуваний дав правильну відповідь, але ж мета тесту – виміряти рівень знань, а не полічити кількість вірних відповідей. Інтуїтивно ясно, що точність вимірювання залежить від

розміру тесту, але великі тести важко укласти і утомливо проходити. Тому хотілося б зрозуміти, який компроміс «ціна-якість» можливий при проведенні тестування. Слід зазначити, що всі подальші міркування і висновки засновані на аналізі тільки кількісного боку тестів.

II. МОДЕЛЬ ЗНАНЬ

Почати слід з визначення того, що таке «знання», які ми маємо намір вимірювати. Залишивши в стороні філософський аспект питання, уявімо модель цього поняття у вигляді безлічі питань, на які людина може або не може дати відповідь. Чим більше правильних відповідей здатна дати людина, тим вище рівень її знань. Таким чином, рівень знань – це відношення $P/(P+Q)$, де P – число питань, на які випробуваний знає відповідь, Q – кількість питань, на які випробуваний відповіді не знає. Виміряти рівень знань означає встановити величину цього відношення, вочевидь, його значення перебувають в інтервалі $(0, 1)$.

Зробимо припущення, що, якщо випробуваний знає відповідь на деяке питання, він вірно відповідь на це питання в тесті, а якщо не знає, відповідь невірно. Насправді це не зовсім так, але ми повернемося до цього пізніше.

III. СХЕМА БЕРНУЛЛІ

З таким припущенням проблема оцінки знань зведеться до добре вивченою статистичної задачі. Є генеральна сукупність великого розміру, що складається, скажімо, з чорних і білих куль. Є випадкова вибірка з n куль, в якій

k білих і $n-k$ чорних куль. Необхідно оцінити за допомогою цієї вибірки питому кількість білих куль в генеральній сукупності.

Про всяк випадок уточнимо, що білі кулі – це питання, на які випробовуваний знає відповідь, чорні кулі – питання, на які випробовуваний відповіді не знає. Вибірка це тест. Питома кількість білих куль в генеральній сукупності – це рівень знань випробуваного, який ми хочемо виміряти.

Коли укладач додає питання до тесту, йому не відомо, знає або не знає випробуваний відповідь на це питання, тобто колір кулі, що обирається, прихований від упорядника, і тому тест можна вважати випадковою вибіркою. Якщо так, то ймовірність появи будь-якої білої кулі в вибірці дорівнює p , будь-якої чорної – q .

$$p = P/(P+Q), \quad q = Q/(P+Q), \quad p + q = 1$$

Кількість білих куль у вибірці – це випадкова величина, що має біноміальний розподіл [3]. Згідно зі схемою Бернуллі, найбільш імовірним значенням p є частка k/n , але наскільки достовірною буде така оцінка? Відповідь можна дати у формі довірчого інтервалу, який визначається, виходячи з бажаного рівня значущості. Імовірність того, що у вибірці розміру n буде рівно k білих куль, розраховується за формулою:

$$P_n(k) = \binom{n}{k} p^k q^{n-k}$$

Біноміальний розподіл унімодальний, тому довірчий інтервал d можна визначити із співвідношення (1).

$$S(k, d) = \sum_{i=k-d_1}^{i=k+d_2} \binom{n}{i} p^i q^{n-i} \leq P, \quad (1)$$

де d_1 і d_2 – відхилення вліво і вправо від найбільш ймовірного значення k/n ; $S(k, d)$ – сума ймовірностей всіх вибірок з числом чорних куль від $k-d_1$ до $k+d_2$; P – обраний рівень значущості.

На графіку функції біноміального розподілу площа $S(k, d)$ зафарбована (рис.1).

При тих значеннях n , які характерні для тестування, визначити довірчий інтервал можна прямим підрахунком сум (1), прийнявши, для параметру p вибіркоче значення k/n . Підрахунок можна виконати за допомогою комп'ютерної програми, як це і робиться в системі автоматичного тестування [4].

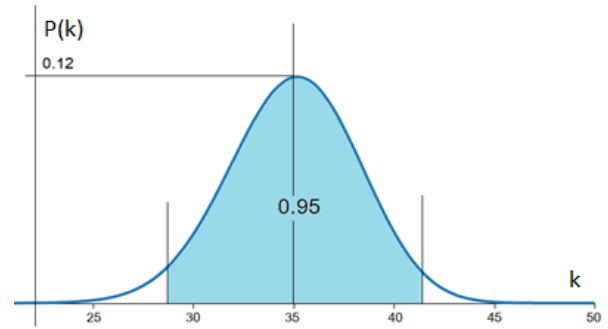


Рис. 1. Графік функції біноміального розподілу

З обчислень видно, що найменш надійні оцінки тих випробовуваних, які показують середній рівень знань; при тому ж рівні значущості їх довірчий інтервал найширший. Видно також, що тести з числом питань більш 50 не дають помітного виграшу в точності оцінок, а тести з числом питань менше 30 недостатньо надійні.

IV. ВГАДУВАННЯ ВІДПОВІДІ

На початку було зроблено припущення, що якщо випробуваний знає відповідь, він вірно відповідає на питання тесту, якщо не знає, відповідає невірно. Ми не обійдемося без припущень в наших міркуваннях, але нехай вони будуть природними, тобто справедливими по відношенню до більшості людей. В даному випадку, природно припустити, що коли випробуваний знає відповідь, він відповідає правильно, а коли не знає, то відповідає навмання або намагається вгадати правильну відповідь.

Таким чином, якщо випробовуваний правильно відповів на k питань, то це число складається з k_1 питань, на які він дійсно знав відповідь, і k_2 питань, на які він відповідь вгадав, $k = k_1 + k_2$.

Імовірність випадково дати правильну відповідь залежить від кількості варіантів відповіді на дане питання, а саме,

$$p_i = 1/a_i \quad \text{- для одноразового вибору}$$

$$p_i = 2^{-a_i} \quad \text{- для множинного вибору}$$

де a_i – число можливих варіантів відповіді на i -те питання тесту.

Припустимо, що випробуваний не знає нічого і на кожне запитання відповідає навмання. Уявімо, що тест він проходить багато разів (N разів). Тоді число вірних відповідей на i -те питання дорівнює $p_i N$, а середня кількість вірно вгаданих відповідей при одноразовому проходженні тесту дорівнює

$$\frac{1}{N} \sum_{i=0}^n p_i N = \sum_{i=0}^n p_i$$

Така сума є важливою характеристикою тесту, але нам буде зручніше мати справу з її відносним вираженням – середньою ймовірністю вгадування.

$$z = \frac{1}{n} \sum_{i=0}^n P_i$$

Якщо випробовуваний відповідає навмання не на всі, а тільки на t питань тесту, то в середньому він вгадає правильні відповіді на z питань.

Повернемося до того, що з k питань випробовуваний знав відповіді лише на $k1$, а решту $k2$ вгадав. Отже, випробовуваний не знав відповіді на $n - k1$ питань тесту і відповідав на них випадково, тобто $k2 = z(n - k1)$.

В той самий час $k2 = k - k1$, і звідси знаходимо $k1$.

$$k1 = (k - zn)/(1 - z)$$

Або, розділивши на n ліву і праву частини

$$k1/n = (k/n - z)/(1 - z)$$

Але $k1/n = r$ – це відношення числа питань, на які випробовуваний знає відповідь, до загальної кількості питань тесту, тобто це вибіркова оцінка реальних знань випробованого.

Частка $k/n = v$ – відношення числа вірних відповідей до числа питань тесту – оцінка спостережуваного рівня знань. Якщо нас цікавить реальний рівень знань, а не спостережуваний, то потрібен перерахунок за формулою (2).

$$r = (v - z)/(1 - z) \quad \square \square \quad (2)$$

Наприклад, в тесті на кожне питання пропонується 5 варіантів відповіді, тобто середня ймовірність вгадування, $z = 1/5$. І коли випробовуваний дає 20% вірних відповідей ($v = 0.2$), ми можемо припустити, що його знання дорівнюють нулю, $r = (0.2 - 0.2)/(1 - 0.2) = 0$.

Якщо ж випробовуваний дає 100% вірних відповідей, то можна сподіватися, що його компетентність максимальна і дорівнює $r = (1 - 0.2)/(1 - 0.2) = 1$.

На малюнку 2 показано співвідношення між спостережуваним (вісь Ox) та реальним (вісь Oy) рівнем знань.

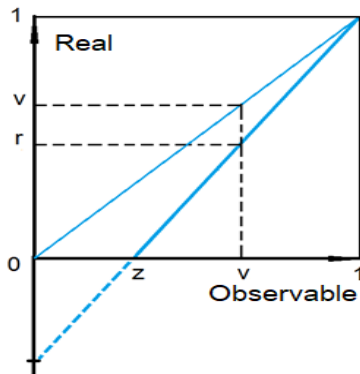


Рис. 2. Спостережуваний та реальний рівень знань

Межі довірчого інтервалу також повинні бути піддані перетворенню (2).

Зауважимо, що знайдуться такі випробовувані, чия реальна компетентність виявиться негативною. Це можливо, тому що, намагаючись вгадати, люди приймають правдоподібні відповіді за правильні. Навіть, якщо результат r невід'ємний, він буде занижений за рахунок невдалих вгадувань, оскільки укладачі тесту зазвичай роблять неправильні відповіді правдоподібними.

Виходить, що, якщо відповіді не знаєш, краще відповідати не замислюючись. А ще краще передбачити в системі тестування кнопку «не знаю», щоб виключити питання з такою відповіддю з перерахунку. Не втомлюючи читача викладками, наведемо формулу перерахунку в тому випадку, коли t питань тесту отримали відповідь «не знаю». Щоб виражатися у відносних величинах, позначимо t/n через u .

$$r = \frac{v - z(1 - u)}{1 - z} \quad (3)$$

При $u = 0$, вираз (3) вироджується в (2).

Якщо ж $u = 1 - v$, тобто випробовуваний завжди чесно визнавав своє незнання, він нічого не втрачає при перерахунку і $r = v$.

V. ВИСНОВКИ

Оцінка знань за допомогою тестів закритого типу носить статистичний характер, тому повинна бути не точковою, а інтервальною. Щоб тест був і інформативним, і не дуже втомлював, він повинен містити від 30 до 50 питань.

Компетентність, яку демонструє випробовуваний в тесті, відрізняється від реальної. Якщо нам цікава реальна компетентність, необхідний перерахунок за формулою (3). До того ж реальна шкала дозволяє порівнювати не тільки результати одного тесту у різних випробовуваних, але і результати різних тестів у одного випробованого, наприклад, при оцінці прогресу студента на певному відрізку часу.

У системах тестування бажано мати кнопку «Не знаю». Натискати її у відповідній ситуації – в інтересах самого випробованого, але це треба йому пояснити.

ЛІТЕРАТУРА REFERENCES

- [1] В.С.Аванесов, Композиция тестовых заданий, 3rd ed М.: Центр тестирования., 2002, с 16-18.
- [2] М.Ф.Бондаренко и др."Оценивание тестовых заданий разных типов и определение их уровня сложности." в журналі НАН України (Відділення інформатики) "Штучний інтелект" 4'2009, сс.322-329.
- [3] Е.В.Гмурман Теория вероятностей и математическая статистика, М.: Высшее образование. 2005, сс.117-123.
- [4] A.J.Almghawish etc., "Software Support For Programming Language Tutorials," in World of Computer Science and Information Technology Journal, Vol. 3, No. 9, 2013, pp.144-149.