A MULTIOBJECTIVE MODEL FOR TEST ASSEMBLING OPTIMISATION

Leoneed Kirilov
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences – Sofia, Bulgaria l_kirilov_8@abv.bg,
Delyan Keremedchiev
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences – Sofia, Bulgaria delyan.keremedchiev@gmail.com

Abstract: A test or an examination is an assessment often on paper or computer assissted. It is intended to estimate knowledge, skills and abilities of the students tested or their classification in groups. The estimation which is the objective of the test is called a result and it represents "generalization of the results of the tested student's answers to the questions asked regarding the objective of the estimation". The test results are determined accordng to norms or criteria but most often according to both. The norm can be determined for one tested student or via statistical analysis of a large number of tested students. The basic component of the test is the question or the problem which is stored in a question bank. The increased use of computers in psychological and educational testing leads to the necessity/possibility to use algorithms for optimal test assembling using question banks. There are algorithms and heuristics in scientific literature which gives the test developers the opportunity to generate simultaneously a number of tests which meet the quality limitations such as project contents, test objectives, etc.
This paper presents a new approach for automatic test generation based on a multicriteria model. This model allows test generation based on 5 criteria. The criteria which generate the tests are conflicting. The proposed methodology can be used in a variety of tests. A bank of test units and test questions for the studied subject area is created. Test units perform a number of parameters including test time, difficulty, guessing, etc. The proposed multicriteria model allows assembling of a number of Paretto optimal tests satisfying predefined parameters of the assembled test such as difficulty, accuracy, etc. Pareto optimal tests can consist of a variety of test units. Tests with predefined characteristics are assembled by an expert (teacher in the corresponding subject; Decision Maker) who sets the parameters for the test. The proposed model can be solved analitically with a suitable multicriteria optimisation method . The final choice of Pareto optimal test is performed by the expert..
Keywords: multiobjective optimization, e testing, e-learning

МОДЕЛ ЗА МНОГОКРИТЕРИАЛНА ОПТИМИЗАЦИЯ ПРИ АСЕМБЛИРАНЕ НА ТЕСТОВЕ

Леонид Кирилов
Институт по информационни и комуникационни технологии, Българската Академия на Науките - София, Република България l_kirilov_8@abv.bg,
Делян Керемедчиев
Институт по информационни и комуникационни технологии, Българската Академия на Науките - София, Република България delyan.keremedchiev@gmail.com

Абстракт: Един тест или изпит, представлява оценяване, което често се прилага на хартия или на компютър, и е предназначено за измерване на знания, умения, способности на тестираните (студенти, ученици и др. обучавани), или класификацията им в групи. Измерването, което е целта на тестването, се нарича резултат, и е "обобщение на доказателствата, съдържащи се в отговорите на тестирания за зададените въпроси, свързани с целта или целите на измерването". Резултатите от тестовете се определят чрез норми или критерии, но най-често и чрез двете. Нормата може да се установи самостоятелно, или чрез статистически анализ на голям брой индивиди. Основният компонент на теста е въпроса или задачата, които се съхраняват в банка от въпроси. Навлизането на компютрите в психологическото и образователното измерване е довело до необходимостта/възможността от използване на алгоритми за оптимално сглобяване на тестове от банките. В научната литература са известни алгоритми и евристики, които позволяват на разработчиците на тестове едновременно да генерират множество от тестове, които отговарят на качествените ограничения, като например проектно съдържание, цели на теста и т.н.
В настоящата статия се представя нов подход за автоматично генериране, базиран на многокритериален (МК) модел. Този модел позволява генериране на тестове по 5 показателя/критерия. Критериите по които се генерират тестовете са противоречиви. Предложената методология може да се

използва в широк спектър от тестове. Създадена е банка от Тестови Единици (ТЕ) и Тестови Въпроси (ТВ) за Изучавана Предметна Област (ИПО). Тестовите Единици се характеризират с множество от параметри (вкл. времеви ограничения, равнище на трудност, вероятност за налучкване и др.). Предложеният многокритериален модел позволява асемблиране на множество от Парето оптимални тестове, удовлетворяващи предварително зададени параметри на проектирания тест - граници на трудност, (статистическа) точност и др. Парето оптималните тестове могат да съдържат различен брой и различен набор ТЕ. Проектирането на тестовете с предварително зададени свойства се извършва от експерт (преподавател от съответната област; Лице, Вземащо Решения). Той задава желаните параметри на теста. Предложеният модел може да се реши аналитично с подходящ метод за многокритериална оптимизация. Окончателният избор на Парето оптимален тест се извършва от експерта.
Ключови думи: многокритериална оптимизация, електронно тестване, електронно обучение

1. ВЪВЕДЕНИЕ

Образованието е право на всеки човек. Нещо повече, то не е само право, а и задължение. В много национални законодателства, включително и в България обучението е задължително до достигането на определена възраст или до покриването на минимална образователна степен. Един от факторите и измерител за успех при обучението е оценката на обучаемите. Оценяването следва да се разглежда не просто като описващо постиженията, а като мощна движеща сила за промяна в образователната система, водеща до подобряване на качеството и до по-високи стандарти на обучение - (Wolf, Bixby, Glenn & Gardner, 1991).

Оценяването чрез тестове е популярно в много от държавите по света. В България оценяването чрез тестове е част от държавните образователни изисквания и се използва както за задължителните държавни зрелостни изпити, така и за приемен изпит при кандидатстване в някои университети. Един тест или изпит, представлява оценяване, което често се прилага на хартия или на компютър, и е предназначено за измерване на знания, умения, способности на тестираните (студенти, ученици и др. обучавани) или класификацията им в групи. Тестовете са инструмент или техника за измерване, използвани за квантифициране на поведението или подпомагане на неговото разбиране и прогнозиране. Теста може да не измерва пълното разбиране на материала в Изучавана Предметна Област (ИПО). Тестовите резултати не са идеална мярка на поведението или характеристиката, но значимо подпомагат процеса на прогнозиране - (Kaplan, Saccuzzo, 2009). Измерването, което е целта на тестването, се нарича резултат, и е "обобщение на доказателствата, съдържащи се в отговорите на тестирания за зададените въпроси, свързани с целта или целите на измерването". Резултатите от тестовете се определят чрез норми или критерии, но най-често и чрез двете. . Нормата може да се установи самостоятелно, или чрез статистически анализ на голям брой индивиди. Данните, които дават тестовете са експлицитни и подлежат на научно изследване.

2. ПОДХОДИ ЗА РЕШАВАНЕ

Основният компонент на теста е въпроса или задачата, които се съхраняват в банка от въпроси. Навлизането на компютрите в психологическото и образователното измерване е довело до необходимостта/възможността от използване на алгоритми за оптимално сглобяване на тестове от банките. В научната литература са известни алгоритми и евристики, които позволяват на разработчиците на тестове едновременно да генерират множество от тестове, които отговарят на качествените ограничения, като например проектно съдържание, цели на теста. сложност на теста и т.н. Конструирането на теста е комплексен, продължителен и в някои свои фази – итеративен процес с изследователски и приложен характер. Той включва етапите на планиране, разработване и анализ на данните, както и оценяване на показаните резултати на изпитаните лица. Дейностите във всички етапи на създаването и използването на теста се извършват в рамките на определена тестова теория. С развитието на психологическите теории се развиват и теоретичните им основи и се усъвършенстват на съществуващите и се разработват нови психометрични теории и модели. Почти всяка тестова теория съществува под формата на теоретични модели.

В настояще време има няколко основни психометрични теории, които са ориентирани към едни и същи данни, но използват различни подходи за тяхното моделиране.

- Класическа тестова теория (Classical test theory), фокусирана върху осигуряване на надеждността на резултатите от измерването на равнище цялостен тест.
- Теория на генерализацията (Generalizability theory, Gtheory), представляваща развитие на Класическата теория, предназначена за осигуряване на надеждността и валидността на наблюденията чрез едновременно оценяване на множество източници на грешки в измерването (Cronbach et al., 1972; Steyer, 2001).

- Теория на латентните състояния и черти (Latent statetrait theory, LST theory), представляваща развитие на Теорията на генерализацията, въвежда формални дефиниции на понятията „състояние" и „черта", както и методи за тяхното разграничаване, отчита влиянието на факторите на ситуацията върху резултатите от измерването и разпростира този подход до анализ на отделни въпроси, базиран на нормалната огива (Steyer, Majcen, Schwenkmezger, Buchner, 1989; Courvoisier, Eid, & Nussbeck, 2007)
- Теория за отговор на тестов въпрос (Item response theory, IRT), фокусирана върху анализа на резултатите на ниво тестов въпрос.
- Теория за отговор на група от въпроси (Testlet response theory, TRT), фокусирана върху изследването на малки групи от еднородни въпроси (testlets, contentdependent item sets), които се разглеждат като основна структурна единица на теста. Теорията се базира изцяло на Бейсовския подход за оценка на вероятностите, а параметрите се оценяват чрез използване на методите Монте Карло за верига на Марков (Rosenbaum, 1988; Wang et al., 2006).

Един съвременен обзор на методите за автоматично асемблиране на тестове може да се намери в Belov (2016). Многокритериални подходи за автоматично генериране на тестове на практика има много малко. Във Veldkamp and Matteucci (1999) и Veldkamp (1999) се разглеждат възможности за прилагане на някои известни в литературата многокритериални подходи в автоматичното генериране на тестове. Във van Groen et al. (2014) са изследвани възможностите на някои многокритериални методи за класификация на изпитваните студенти.

За целите на модела, който представяме в разработката ще използваме Класическата тестова теория, която въпреки някои свои несъвършенства дава висока степен на надеждност на измерваните резултати. В Класическата теория действителният бал на индивида е относително устойчива, стабилна величина, която не се променя при многократно измерване с един и същи тест или с различни форми на теста. На индивидуално равнище този компонент е константа, параметър с фиксирана, но неизвестна стойност. Макар че действителният тестов бал не може да бъде наблюдаван пряко, точно той стои във фокуса на Класическата теория, а и на изследователския интерес при провеждане на измервания за научни или практически цели. В Класическата теория наблюдаваният бал служи за оценка на неизвестния действителен бал (Embretson & Reise, 2000)

3. МНОГОКРИТЕРИАЛЕН МОДЕЛ

Предполагаме съществуване на изходна банка от въпроси (тестови единици) с многовариантен избор. Всяка от тестовите единици е решавана електронно от контролна група изпитвани:

DBTU (database of test units) = $\{tu_i, i= 1, \ldots, N\}$

Всяка тестова единица се характеризира с три параметъра – виж. Kaplan and Saccuzo (2009), Totkov et al. (2014)):

$tu_i = tu_i (c_i, p_i, t_i)$, $i= 1, \ldots, N$

където

c_i – сложност/трудност (difficulty) на тестовата единица

p_i – вероятност за отгатване/налучкване (guessing)

t_i – време за решаване при избрана мерна времева единица.

Сложността c_i означава какъв процент от изпитваните са в състояние да решат тази тестова единица:

c_i = (студенти с верен отговор) / (общ брой студенти) * 100

Вероятността за отгатване на верен отговор p_i се пресмята като

$p_i = 1/x$,

където x е броя на отговорите в тестовата единица

Времето за решаване t_i се пресмята като

t_i = средноаритметично време от времето за решаване на тестовата единица от контролната група

Очевидно могат да бъдат генерирани различни тестове Tj, j = 1, …, m на база изходната банка от тестови единици. Те ще имат различна дължина, респ. различен брой неповтарящи се компоненти (тестови единици). Както е известно общият брой на всички възможни тестове с дължина к тестови единици е $\binom{N}{k}$, което е комбинация от N елемента k-ти клас без повторения. Общият брой на всички възможни тестове с всички възможни дължини е доста по-голям. Именно сумата от всички възможни комбинации без повторения, където k = 1, …, N. Да означим това множество от всички възможни допустими тестове с Ω.

Задачата, която си поставяме е да изберем тест (или набор от тестове) с предварително зададени свойства.

Очевидно тази задача е комбинаторна и пълното изброяване/генериране на всички тестове с цел последващ избор на тест(ове) по някакъв критерий(и) не е най-добрият път за нейното решаване.
И така стигнахме до въпроса по какви показатели или критерии да се селектират тестовете.
Ние предлагаме един тест да се описва със следните критерии:

- Точност/прецизност θ
- Дължина на теста TL (test length)
- TC – сложност/трудност (difficulty) на теста (Test Difficulty)
- TG – вероятност за отгатване/налучкване (guessing) на теста (Test Guessing).
- TT – време за решаване при избрана мерна времева единица (Test Time).

По такъв начин един тест T_j се описва като набор от пет параметъра (критерия):

$$T_j = T_j(\theta_j, TL_j, TC_j, TG_j, TT_j)$$

където например теста $T_j = (tu_{j1}, tu_{j2}, ..., tu_{jn})$ е с дължина jn, а индексите на тестовите единици показват тяхното глобално място в банката от данни DBTU.
Първите два критерия – точността и дължината на теста, са известни в литературата и се пресмятат както следва:

1. Точност/прецизност на теста $\theta = \left(\frac{N}{N-1}\right)\left(1 - \frac{\sum_{i=1}^{N} \sigma_{yi}^2}{\sigma_x^2}\right)$,

   където N е броят тестови единици в теста, σ_{yi}^2 е вариацията на i-та тестова единица, σ_x^2 е вариацията на наблюдавания резултат.

2. Дължина на теста TL (test length) = цяло число N, брой тестови единици в теста.

Последните три критерия произлизат по естествен начин от характеристиките на тестовата единица, която е съставна компонента на теста.
Ще дефинираме тяхното аналитично пресмятане както следва.

3. Сложност на теста $TC_j = \sum_{l=1}^{n} Ctu_{jl}$ като сума от сложностите на съставляващите тестови единици.
4. Вероятност за отгатване $TG_j = \left(\sum_{l=1}^{n} P\,t_j u_l\right) /$ като средноаритметично на отделните вероятности на съставляващите тестови единици.
5. Време за решаване $TT_j = \left(\sum_{l=1}^{n} Ttu_{jl}\right)$ като сума от времето за решаване на съставляващите тестови единици.

Вече можем да формулираме многокритериална задача за намиране на оптимален тест по така формулираните пет критерия както следва:

$$\underset{T_j \in \Omega}{Opt}(\theta_j, TL_j, TC_j, TG_j, TT_j)$$

където оптимизацията е по избраните пет критерия, а тестовете T_j описват множеството от всички възможни тестове Ω. Символът „Opt“ означава едновременна оптимизация на критериите, която може да бъде минимизация или максимизация според конкретния критерий, или постигане на желана стойност.
Горната многокритериална задача е коректно дефинирана, тъй като два от критериите са противоречиви. Именно точността на теста и дължината на теста. Максимална точност на теста се получава при „голяма“ дължина на теста. Ние обаче се интересуваме да постигнем максимална точност при минимална дължина на теста. Горният многокритериален модел може да бъде решен с подходящ метод за многокритериално вземане на решения.

4. ЗАКЛЮЧЕНИЕ

В настоящата статия се предлага многокритериален модел за асемблиране на тестове.
Посредством него могат да се генерират Парето оптимални тестове на базата на оптимизация на пет критерия – точност на теста, дължина на теста, сложност, време за решаване и вероятност за налучкване на верни отговори.
Предложеният многокритериален модел позволява асемблиране на множество от Парето оптимални тестове, удовлетворяващи предварително зададени параметри на проектирания тест - граници на трудност, (статистическа) точност и др. Парето оптималните тестове могат да съдържат различен брой и различен набор ТЕ. Проектирането на тестовете с предварително зададени свойства се извършва от експерт (преподавател от съответната област; Лице, Вземащо Решения). Той задава желаните параметри на теста. Предложеният модел може да се реши аналитично с подходящ метод за многокритериална оптимизация. Окончателният избор на Парето оптимален тест се извършва от експерта.

ЛИТЕРАТУРА

[1] B. Veldkamp, M. Matteucci (1999) Dealing with multiple criteria in test assembly, AMS Acta-AlmaDL - Univ. of Bologna Digital Library, http://amsacta.unibo.it/2674/.
[2] B. Veldkamp (1999) Multiple Objective Test Assembly Problems, Journal of Educational Measurement, Vol. 36, No. 3 (Autumn, 1999), pp. 253-266.
[3] M. van Groen, T. J.H.M. Eggen, B. P. Veldkamp (2014) Item Selection Methods Based on Multiple Objective Approaches for Classifying Respondents Into Multiple Levels, Applied Psychological Measurement, 2014, Vol. 8(3), pp. 187–200.
[4] D. Belov (2016) Review of Modern Methods for Automated Test Assembly and Item Pool Analysis, Law School Admission Council Research Report 16-01 March 2016, LSAC Research Report Series, 23 pages, https://www.lsac.org/docs/default-source/research-(lsac-resources)/rr-16-01.pdf
[5] R. Kaplan, D. Saccuzzo (2009) Psycological Testing, 7-th edition, Wadsworth, Cengage Learning, Belmont (CA), USA, ISBN-13: 978-1133492016.
[6] Wolf, D., Bixby, J., Glenn, J. & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (ed.), Review of Research in Education, 17, pp. 31-125.
[7] Cronbach, L., Gleser, G., Nanda, H. & Rajaratnam, N. (1972). The dependability of behavior-al measurement: Theory of generalizability for scores and profiles. New York: John Wiley & Sons, Inc.
[9] Steyer, R. (2001). Classical (Psychometric) test theory. Friedrich-Schiller-Universitaet Jena, Institut fuer Psychologie, Lehrstuhl fuer Methodenlehre und Evaluationsforschung. http://www.metheval.uni-jena.de/materialien/publikationen/ctt.pdf
[10] Steyer, R., Majcen, A., Schwenkmezger, P. & Buchner, A. (1989). A latent state-trait anxiety model and its application to determine consistency and specificity coefficients. Anxiety Research, 1, pp. 281-299.
[11] Courvoisier, D., Eid, M. & Nussbeck, F. (2007). Mixture Distribution State-Trait-Models: Basic ideas and applications. Psychological Methods, 12, 80-104.
[12] Rosenbaum, P. R. (1988). Item bundles. Psychometrika, Vol. 53, pp. 349-359.
[13] Wang, X., Wainer, H., Brown, L., Bradlow, E. Skorupski, W., Boulet, J., & Mislevy, R. (2006). An application of Testlet response theory in the scoring a complex certification exam. In: D. Williamson, R. Mislevy, & I. Bejar (eds). Automated scoring of complex tasks in Computer based testing. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., pp. 169-199.
[14] Embretson, S. E., Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
[15] Totkov, G., Raykova, M., Kostadinova H. (2014) Testing in e-education, Rakursi Ltd, Plovdiv, Bulgaria, 205 pages, ISBN 978-954-8852-42-5 (in Bulgarian).