International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 9, No. 2, 2023

Research on the Application of Machine Learning-Based Scientific Argumentation Automatic Scoring in Course Evaluations

Qianqian Lu

School of Education and Physical Education, Yangtze University, Jingzhou, 434023, China

Abstract: In modern education, course evaluation is one of the important means to improve the quality of teaching. However, traditional methods of course evaluation suffer from subjectivity and lack objectivity, making it difficult to accurately reflect the teaching situation. Therefore, the emergence of machine learning-based automatic scoring technology provides a new approach and method for course evaluation by analyzing student assignments, exams, and other data to generate objective and accurate scoring results. This article discusses the application research of machine learning-based automatic scoring technology in course evaluation, including algorithm selection, model training, and analysis of scoring results. It also provides prospects for the future development direction and application prospects of this technology.

Keywords: Machine learning, Automatic scoring, Course evaluation.

1. Preface

In recent years, the development and popularization of machine learning technology has brought new opportunities to the field of education. Among them, automatic scoring technology based on machine learning has become one of the hot research directions. Automatic scoring technology can effectively solve the problems existing in traditional scoring methods, improve scoring efficiency, and also enhance the objectivity and accuracy of scoring results. Furthermore, with the application of artificial intelligence in the field of education, human-computer interaction and collaborative research have received high attention [1]. In recent years, automatic assessment of scientific arguments in text format has shown this trend, where computers can not only automatically score students' learning works in real-time but also provide automatic feedback, achieving the goal of formative evaluation. For example, Lee et al. developed a system called HASbot, which can not only perform real-time scoring of open-ended questions but also provide real-time feedback. Students can receive new real-time scores and feedback even after revising their answers based on the feedback [2]. Currently, this research field has started to focus on the impact of different types of real-time feedback on courses [3-4].

2. The Definition and Disadvantages of Curriculum Evaluation

2.1. Definition of Course Evaluation

When discussing course evaluation, it is necessary to provide a brief definition. Course evaluation involves the scientific examination of the course's objectives, design, and implementation based on certain criteria and systematic information about the course. It aims to determine whether the educational goals have been achieved and to what extent, in order to assess the effectiveness of the course design and make decisions for its improvement. Therefore, the objects of course evaluation include evaluating the course design, evaluating the course implementation, assessing students' academic performance and personal development, evaluating the course system, and evaluating the course evaluation itself. Among these, evaluating students' academic performance and personal development is also one of the main focuses of course and teaching evaluation.

2.2. The drawbacks of traditional course evaluation.

Traditional methods often employ quantitative measures such as exams and tests, which are simple and easy to implement. The evaluation criteria are clear, such as scores or percentages, allowing for direct assessment and comparison of students' performance, hence ensuring a certain level of fairness. Although traditional course evaluation has its advantages, it also has significant issues.

Firstly, the evaluation standards are overly narrow. Traditional course evaluation primarily focuses on students' exam scores, disregarding other aspects that influence students' overall qualities, such as practical skills, communication abilities, and innovation capabilities. Furthermore, the evaluation content lacks comprehensiveness. Traditional course evaluation emphasizes the mastery of knowledge points while neglecting students' depth of understanding and application abilities. This can lead to a "test-oriented education" approach, where students primarily prioritize memorization and mechanical application, rather than being able to flexibly apply their learned knowledge in real-life situations. Thus, it hinders the comprehensive development of students' potential and various abilities.

Secondly, the evaluation methods are limited. Traditional course evaluation usually relies on written exams, even for courses that are not suitable for such an assessment format, such as art or physical education. As a result, it fails to fully assess students' true abilities and performances. Moreover, the reliability of evaluation results is questionable. Traditional course evaluation is susceptible to various factors' interference, such as students' personal circumstances, exam environments, and paper difficulties. Evaluation results often contain errors and lack objectivity, thus not accurately reflecting students' actual abilities and levels.

In conclusion, the problems with traditional course evaluation are evident. It is necessary to reform the evaluation system in terms of methods, content, and approaches to establish a more scientific, comprehensive, and objective course evaluation framework.

3. The Working Principle and Application of Machine Learning-based Automated Scoring Technology

The focus of this article is on the automatic assessment of students' creative and practical abilities, with a specific emphasis on their scientific reasoning skills. The National Assessment of Education Quality (NAEQ) framework for science education in China in 2017 explicitly includes students' ability to reason and argue as an important aspect of their creativity [5]. Therefore, this article aims to explore the working principles and applications of machine learning-based automatic scoring for scientific reasoning, and attempts to integrate real-time feedback systems to improve classroom teaching.

3.1. The scientific rationale behind the orientation of evidence in automated scoring

Currently, there is a lack of consensus on the understanding of the value orientation of scientific argumentation, which can be categorized into two main perspectives. First, scientific argumentation is seen as a practical activity for problem-solving and knowledge growth. In this view, argumentation in science is considered a practice of establishing and validating knowledge. Individuals strive to understand the natural world by proposing, supporting, questioning, and refining ideas [6-7]. Second, argumentation is viewed as non-formal reasoning. Non-formal reasoning refers to reasoning beyond formal logic and mathematics, including reasoning about the causes and consequences of specific claims or decision-making processes, providing reasons for or against a particular decision, and analyzing the pros and cons of specific claims or decision-making processes. This type of reasoning is based on certain attitudes and perspectives. It is often associated with inductive reasoning rather than deductive reasoning [9]. This article focuses on the automatic assessment of scientific argumentation using the first value orientation, which regards scientific argumentation as a scientific practice for establishing and validating knowledge.

3.2. The statement of the characteristics and scoring rules of automatic test questions for scientific argumentation

Scholars have developed various evaluation frameworks suitable for scientific reasoning. These evaluation frameworks focus on different dimensions of arguments. For instance, Er-duran et al. [10] focus on the structure or complexity of arguments, which refers to the constituent parts of an argument. Clark et al. [11], on the other hand, emphasize the quality of argument content, evaluating the accuracy or sufficiency of individual components from a scientific perspective. Currently, automatic evaluation of scientific reasoning primarily adopts Toulmin's argument pattern (TAP) as the basic framework, combining subjective and objective questions. The scoring criteria are formulated based on research achievements in the field of scientific argument evaluation frameworks. This paper uses the scientific reasoning assessment questions and scoring criteria employed by Mao et al. [15] as an illustrative example.

3.2.1. Based on the TAP evaluation framework

The TAP framework, developed by Mao et al., is a widely used and highly accepted model for evaluating students' argumentation skills based on scientific evidence. The framework consists of two components: the basic pattern and the extended pattern. The basic pattern includes a claim, data, and warrants, while the extended pattern incorporates additional elements such as backing, rebuttal, and qualifiers. Mao et al. selected the basic pattern of TAP as the scientific argumentation evaluation framework because of its simplicity, universality, and wide application, which meet the requirements of automated assessment.

3.2.2. A test structure combing subjective and objective questions

The scientific assessment test developed by Mao consists of four questions, each examining different aspects of scientific reasoning ability. These are: 1) presenting scientific claims (multiple-choice); 2) explaining scientific claims based on theoretical evidence (subjective question); 3) expressing the level of uncertainty in the scientific claims (Likert five-point scale); and 4) describing the sources of uncertainty.

4. Application of the Automatic Scoring Tool c-rater-ML

In 2016, building upon the early scientific measurement automated scoring tool c-rater, ETS developed a new generation of scientific assessment automated scoring tool called c-rater-ML (c indicating constructed-responses, ML indicating machine learning). The aim was to utilize machine learning techniques to establish a model between student responses and scores [18]. Machine learning techniques have gained significant recognition in scientific assessment research in recent years [19]. Compared to c-rater, c-rater-ML not only reduces the consumption of manpower but also enhances scoring accuracy. This article introduces c-rater-ML from three aspects: development principles and technical roadmap, accuracy validation, and advantages and disadvantages.

4.1. The principle and technical roadmap of c-rater-ML

The c-rater-ML automatic scoring tool utilizes support vector regression (SVR) to establish a relationship model between student responses and scores, and to identify the mapping from text answers to scores. SVR is a statistical analysis method based on machine learning technology specifically designed to handle large datasets with numerous correlated predictor variables. In terms of the principle of establishing the relationship model between predictor variables and dependent variables, SVR is very similar to multiple regression.

c-rater-ML employs the open-source software package SciKit-Learn Laboratory to execute SVR. The generated model can be used to automatically predict student responses. The c-rater-ML automatic scoring consists of three steps [18].

Firstly, experts manually score all student responses according to the scoring criteria. Initially, two raters independently score 50 student responses for each question,

which should encompass all possible scores. The two raters compare their own scores, identify any discrepancies, and engage in discussions until reaching a consensus. If the two raters cannot reach an agreement, a third rater is involved in the discussion until a consensus is reached. This scoring process iterates until the average consistency of manual scoring reaches 90%. Once a consensus is reached, the raters clarify the scoring criteria, and the remaining un-scored student responses are independently scored by the raters after distributing the workload evenly among them. Raters intermittently select 20 student responses for cross-evaluation to ensure consistency in scoring results. In Mao et al.'s study, quadratic-weighted kappa (QWK), Pearson correlation coefficient (r), and standardized mean difference (SMD) were used to describe inter-rater consistency.

Secondly, a computerized automatic scoring model based on student responses is established. Specifically, the student responses for each question are randomly divided into two groups. The first group comprises approximately 2/3 of the data and is used for model training and preliminary evaluation. The second group consists of the remaining approximately 1/3 of the data and is used for testing or validating the accuracy of the trained predictive model.

Lastly, scoring analysis is conducted for questions with significant discrepancies between human and automated scoring.

4.2. The accuracy evaluation of c-rater-ML

Evaluate the consistency of c-rater-ML scores with human ratings, referred to as "human-machine consistency," using QWK and Pearson correlation. QWK and Pearson correlation are the most commonly used methods to assess the accuracy of automated scoring. QWK measures the proportionality of the squared differences between two ratings, making it easier to detect score discrepancies. When judging the QWK value, Landis and Koch's kappa coefficient consistency strength standards are employed: poor (QWK < 0.00), slight ($0.00 \leq$ QWK ≤ 0.20), fair ($0.21 \leq$ QWK ≤ 0.40), moderate ($0.41 \leq$ QWK ≤ 0.60), substantial ($0.61 \leq$ QWK ≤ 0.80), and almost perfect ($0.81 \leq$ QWK ≤ 1) [20]. Pearson correlation is commonly used as a general criterion for evaluating consistency between human and machine scores. The Pearson correlation coefficient reflects the correlation between human and machine ratings, following Cohen's consistency strength standards: none ($0 \leq r \leq 0.09$), small ($0.10 \leq r \leq 0.30$), moderate ($0.31 \leq r \leq 0.50$), and large ($0.51 \leq r \leq 1.00$) [20]. Additionally, Mao et al., based on Williamson et al.'s work [21], proposed precise accuracy testing criteria for automated scoring models, stating that a reliable automated scoring model should meet three conditions: 1) QWK and Pearson correlation coefficients between c-rater-ML scores and human ratings should not be lower than 0.70; 2) the standardized mean difference (SMD) between c-rater-ML and human ratings should not exceed 0.15; 3) based on the QWK and Pearson correlation coefficients between human ratings, the difference between the QWK and Pearson correlation coefficients between c-rater-ML scores and human ratings should not exceed 0.10 [15].

In addition to the aforementioned analysis, a differential analysis can be conducted on the ratings between subgroups, such as grouping by gender, language, and whether homework is done using a computer. The differences between each group will be estimated using t-tests, and the effect size in terms of standardized mean difference (SMD) will be provided for each comparison [20].

5. Research Prospect

With the rapid development of artificial intelligence technology, machine learning-based automated assessment has become an indispensable part of modern education. This study aims to explore the application of this technology in course evaluation and delve into its feasibility and practical effects. By introducing the multi-criteria rating machine learning (ML) technique, it is evident that automated assessment based on machine learning offers high accuracy and efficiency. It also significantly reduces labor costs and enhances the objectivity and scientificity of evaluation results.For instance, in language courses, the use of automated assessment technology can better evaluate students' abilities in grammar, spelling, pronunciation, and other aspects. Additionally, it can provide customized evaluations for individual students according to their specific needs, while offering teachers more detailed assessment reports to aid them in guiding students effectively. However, machine learning-based automated assessment still faces certain limitations and challenges. For example, due to the inability of machines to perceive students' emotional changes and thought processes, this technology cannot completely replace the role of teachers. Nevertheless, overall, the application of machine learning-based automated assessment in course evaluation has achieved some positive outcomes. The evaluation results are more accurate and require less time and effort.

In the future, we can further strengthen research and development efforts to improve this technology and better meet the demand for scientific, objective, and efficient evaluation in the field of education. Simultaneously, in practical applications, it is crucial to combine the technology with human care to ensure it truly becomes a tool serving students and teachers, thereby promoting comprehensive student development.

References

[1] Dong Yan, Li Xinyi, Zheng Yafeng, et al. Bidirectional Feedback in Intelligent Educational Applications: Mechanisms, Models, and Implementation Principles [J]. Open Education Research, 2021, 27(2): 26-33.

[2] LEE H-S, PALLANT A, PRYPUTNIEWICZ S, et al. Automated text scoring and real-time adjustable feed- back: supporting revision of scientific arguments involv- ing uncertainty[J]. Science Education, 2019, 103(3): 590-622.

[3] ZHU M, LIU O L, LEE H-S. The effect of automated feedback on revision behavior and learning gains in for- mative assessment of scientific argument writing[J]. Computers & Education, 2020, 143(1): 41-48.

[4] LEE H-S, GWEON G-H, LORD T, et al. Machine learning-enabled automated feedback: supporting stu- dents'revision of scientific arguments based on data drawn fromsimulation[J]. Journal of Science Education and Technology, 2021, 30(2): 168-192

[5] Hu Weiping. Evaluation of academic quality in science based on core literacy [J]. Chinese Examination, 2016(8): 23-25.

[6] DRIVER R, NEWTON P, OSBORNE J. Establishing the norms of scientific argumentation in classrooms[J]. Science Education, 2000, 84(3): 287-312.

[7] SAMPSON V, CLARK D B. The impact of collabora- tion on the outcomes of scientific argumentation[J]. Sci- ence Education, 2009, 93(3): 448-484.

[8] SADLER T. Informal reasoning regarding socioscientif- ic issues: a critical review of the research[J]. Journal of Research in Science Teaching, 2004, 41(5): 513-536.

[9] ZOHAR A, NEMET F. Fostering students'knowledge and argumentation skills through dilemmas in human genetics[J]. Journal of Research in Science Teaching,2002, 39(1): 35-62.

[10] ERDURAN S, SIMON S, OSBORNE J. Tapping into ar- gumentation: developments in the application of Toul- min's argument pattern for studying science discourse [J]. Science Education, 2004, 88(6): 915-933.

[11] CLARK D, SAMPSON V. Assessing dialogic argumen- tation in online environments to relate structure, grounds, and conceptual quality[J]. Journal of Research in Science Teaching, 2008, 45(3): 293-321.

[12] VAN EEMEREN F, GROOTENDORST R, HENKE- MANS A F. Argumentation: analysis, evaluation, pre- sentation[M]. Mahwah, NJ: Erlbaum, 2002: 12.

[13] CLARK D, SAMPSON V, WEINBERGER A, et al. An- alytic frameworks for assessing dialogic argumentation in online learning environments[J]. Educational Psy- chology Review, 2007, 19(3): 343-374.

[14] DUSCHL R. Quality argumentation and epistemic crite- ria[M]//ERDURAN S, JIMENEZ-ALEIXANDRE M. Argumentation in science education: recent develop- ments and future directions. Berlin: Springer, 2007: 159-175.

[15] LEE H-S, LIU O L, PALLANT A, et al. Assessment of uncertainty-infused scientific argumentation[J]. Jour- nal of Research in Science Teaching, 2014, 51(5): 581-605.

[16] LAZAROU D, ERDURAN S, SUTHERLAND R. Argu- mentation in science education as an evolving concept: following the object of activity[J]. Learning, Culture and Social Interaction, 2017, 14(9): 51-66.

[17] MAO L, LIU O L, ROOHR K, et al. Validation of auto- mated scoring for a formative assessment that employs scientific argumentation[J]. Educational Assessment, 2018, 23(2): 121-138.

[18] Ren Hongyan, Li Guangzhou. Research progress on Turmin's argumentation model in science education [J]. Foreign Primary and Secondary Education, 2012(9): 28-34.

[19] Liu Taorong, Xiao Hua, Zhang Junpeng. Analysis perspectives and evaluation models on the scientific reasoning ability of foreign students [J]. Shanghai Educational Research, 2019(2): 53-57.

[20] LIU O L, RIOS J A, HEILMAN M, et al. Validation of automated scoring of science assessments [J]. Journal of Research in Science Teaching, 2016, 53(2): 215-233.

[21] ZHAI X, KRAJCIK J, PELLEGRINO J W. On the va- lidity of machine learning-based next generation sci- ence assessments: a validity inferential network[J]. Journal of Science Education and Technology, 2021, 30(2): 298-312.

[22] COHEN J. Statistical power analysis for the behavioral sciences[M]. 2nd ed. Hillsdale, NJ: Erlbaum Associ- ates, 1988: 3.