Partial Credit Grading of DFAs: Automation vs Human Graders

Nathan Smearsoll
Rochester Institute of Technology

Ivona Bezáková
Rochester Institute of Technology

ABSTRACT

We examined the efficacy of automatic partial credit approaches for assignments asking students to construct a Deterministic Finite Automaton (DFA) for a given language. We chose two DFA problems, and generated a representative sample of 10 benchmark submissions for each. Next, in order to get an accurate baseline of the results of human graders, we asked professors at our university to submit their grader guides to us. We found that the grader guides, at least within our institution, were very consistent but also quite problem-specific and reliant on human understanding, hence unlikely to lead to an automated process applicable to all DFA problems. We generated a "consensus grader guide" and graded each benchmark submission, obtaining a baseline human partial credit score. Then, we assessed the submissions using three techniques proposed by Alur et al.: The Solution Syntactic Difference (SSD) technique's score corresponds to the number of changes that must be made to the DFA. The Problem Syntactic Difference (PSyD) score is based on converting each DFA into Monadic Second Order (MSO) Logic and examining the number of necessary changes. For Problem Semantic Difference (PSeD), the score is the limit of the ratio of incorrect strings to correct strings. The final score is the maximum of these three scores. In general, the results closely matched the consensus grades, but there were some peculiarities generated by PSeD. Additionally, for each problem, one submission included two separate types of mistakes. These submissions had automatic grades much lower than the consensus grades.

1 BACKGROUND AND RELATED WORK

Automata are key to a strong theoretical background of Computer Science. DFAs, as the most basic of automata, are useful theoretical tools and students are frequently assigned problems that require them to construct DFAs for a given language. Traditionally, these problems are graded by hand by a human who understands how to assess whether a DFA generates the requested language, and if not how many points to deduct. Since figuring out whether a DFA generates a given language is an easily decidable process, grading these problems with either a 0 or a 100% is quite straightforward to automate. However, these problems are usually graded with the possibility of receiving partial credit for a solution that, while incorrect, is close to the correct solution or shows the correct methodology. Alur et al. [1] proposed several techniques for automated computation of partial credit for DFAs. (This is the only work on DFA automated partial credit we are aware of.) We examined these proposed approaches and compared the computed scores with scores awarded by human graders, in order to assess the viability of such automation, and whether automated partial credit DFA grading is generally efficient and effective.

2 METHODS

We chose two DFA problems, and generated a representative sample of 10 benchmark submissions for each, ranging from perfectly correct to very incorrect. For all of these submissions, we first assessed what score we, as the problem creators, felt they deserved. Next, we received grader guides from four instructors at our university and generated a "consensus grader guide" which we used to obtain a baseline human partial credit score for each submission. Then, we assessed our example submissions using the three techniques from [1]: For SSD, we computed how many changes to the DFA must be made in order to convert it to the correct minimal DFA for the problem. For PSyD, we converted each DFA into an MSO Logic formula, and examined how many changes would need to be made to match an MSO definition of the problem. For PSeD, we examined the limit of the ratio of incorrect strings to correct strings, as the length of the strings increases. We also looked at one submission for each problem that included two separate types of mistakes, which we knew was both a realistic example of student submissions and especially difficult to automatically grade. We graded the submissions with each of the above methods: general assessment, consensus grader guide, SSD, PSyD, PSeD, and max of the last three. We also assessed the possibility of adversarial submissions, and whether they would be likely to successfully defeat, or in some way break, an automated grader using the above methods.

3 FINDINGS

In general, the automatic grader does map relatively well to the results from the consensus grader guide for errors that conform well to one of the types. One notable exception is when a student submits Σ^* for a language containing almost all strings, which leads to bloated score due to the string ratio calculation. It is not clear how cases like this can be detected (and scores adjusted) in an automated way. Furthermore, there are many realistic student errors that include multiple of the core categories defined. Fair automated evaluation of such submissions appears elusive.

4 ACKNOWLEDGMENTS

Supported by a Research Experiences for Undergraduates (REU) supplement to NSF award 1819546. We thank the RIT instructors, and the anonymous reviewers for excellent poster suggestions.

REFERENCES

[1] Rajeev Alur, Loris D'Antoni, Sumit Gulwani, Dileep Kini, and Mahesh Viswanathan. 2013. Automated Grading of DFA Constructions. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI/AAAI, 1976–1982.

SIGCSE 2023, March 15–18, 2023, Toronto, ON, Canada

ACM ISBN 978-1-4503-9433-8/23/03.
https://doi.org/10.1145/3545947.3576328