Automatic Rating of User-Generated Math Solutions

Turadg Aleahmad, Vincent Aleven, Robert Kraut
{aleahmad, aleven, kraut}@cs.cmu.edu
Human Computer Interaction Institute, Carnegie Mellon University

Abstract. Intelligent tutoring systems adapt to users' cognitive factors, but typically not to affective or conative factors. Crowd-sourcing may be a way to create materials that engage a wide range of users along these differences. We build on our earlier work in crowd-sourcing worked example solutions and offer a data mining method for automatically rating the crowd-sourced examples to determine which are worthy of presenting to students. We find that with 64 examples available, the trained model on average exceeded the agreement of human experts. This suggests the possibility for unvetted worked solutions to be automatically rated and classified for use in a learning context.

1 Introduction

Intelligent tutoring systems have made great progress on adapting instruction to match students' cognitive needs. They match instructional content to the learner's changing state of knowledge, the cognitive variables. There is little work, however, on matching instruction to the learner's interests, motivations and identity, the affective and conative variables [5], which can improve student engagement and test scores [4]. Personalizing to these other factors may require a large set of socially and topically varied problems and worked example solutions. Our previous work demonstrated that crowd-sourcing is a feasible approach to covering the gamut of learners with many worked examples [1]. However, it is important to determine which of the contributions are worthy of presenting to learners, which need more work, and which should be discarded. In the current paper we assess the feasibility of using data mining to automatically classify crowd-sourced worked example solutions by their readiness to present to learners.

Machine rating of quality has been studied in many domains (e.g. Wikipedia articles [2], student essays [3]). The strongest predictors are often simple, and sophisticated features add little. The 1960s Project Essay Grader using just word count, average word length, counts of punctuations and prepositions, etc. achieved *0.78* correlation with teachers' scores, almost as strong as the *0.85* correlation between two or more trained teachers [3].

Our data set comes from a corpus of worked examples to practice the Pythagorean Theorem. Each example consists of problem statement and a series of steps to solve it. The 278 examples were contributed by 10 math teachers, 20 non-math teachers and 110 amateurs. Each solution was coded manually by three experts on a 4 point scale: Useless, Fixable, Worthy, or Excellent. The ratings of the three coders (Cronbach's α=.71) were averaged to create a single *solution quality* measure, for which the models were trained.

2 Results

Model accuracy is evaluated as the correlation of the machine prediction with the *solution quality* measure. Because a subsequent analysis on human correlations requires unseen instances, a test set was made holding out a stratified random sample of 28 instances. The remaining 250 worked examples were used to train a model evaluated by 10-fold cross-validation. The attribute space was restricted to features of the example (e.g. count of the word “solve”) and whether the contributor agreed to be further contacted. The learning algorithm selected was REPTree and was enhanced through a Bagging meta-classifier using Weka.

To see how the performance is affected by the number of instances, we trained on random subsets each of size 8, 16, 32, 64, 128 and 250. Twenty subsets were created with random sampling with replacement. Training on all 250 instances the mean accuracy (correlation with *solution quality* value) over the 20 runs was *r=.67*.

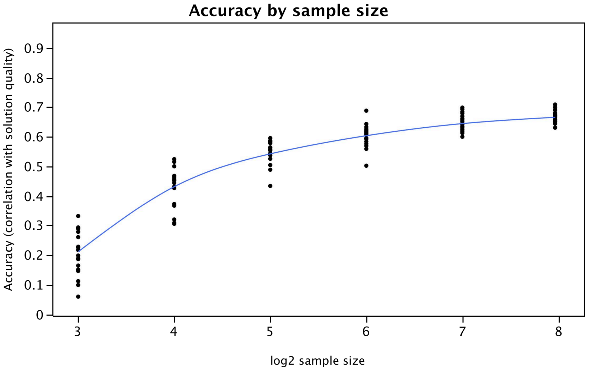

Figure 1 Accuracy by log2 sample size

For the examples in this data set, the machine-based rater correlated better with the human experts than they did with each other (*humans pair-wise average r=.53 and model average r=.67*). We found that 64 training examples were enough to beat the human raters on average and that with 128 examples the machine outperformed consistently. For domains in which these results hold therefore just 64 rated examples are needed to create a model that can automatically rate future contributions. Because it can do this instantly and on-demand, such models could be used to facilitate a peer-produced worked example system. Anyone could contribute a solution, and learners would only be shown those of the highest quality. This work supports the scalability and sustainability of such a system.

References

1. Aleahmad, T., Aleven, V., and Kraut, R. Creating a Corpus of Targeted Learning Resources with a Web-Based Open Authoring Tool. *IEEE Transactions on Learning Technologies 2*, 1 (2009), 3-9.

2. Dalip, D.H., Gonçalves, M.A., Cristo, M., and Calado, P. Automatic Quality Assessment of Content Created Collaboratively by Web Communities: A Case Study of Wikipedia. *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, ACM (2009), 295-304.

3. Hearst, M.A. The debate on automated essay grading. *IEEE Intelligent Systems and Their Applications 15*, 5 (2000), 22–37.

4. Ku, H. and Sullivan, H. Student performance and attitudes using personalized mathematics instruction. *Educational Technology Research and Development 50*, 1 (2002), 21-34.

5. Martinez, M. and Bunderson, C.V. Building interactive World Wide Web (Web) learning environments to match and support individual learning differences. *J. Interact. Learn. Res. 11*, 2 (2000), 163-195.