ENSURING RELIABILITY OF EDUCATIONAL DATA MINING DETECTORS FOR DIVERSE POPULATIONS OF LEARNERS

Baker, Ryan S.J.d., Ocumpaugh, Jaclyn L., Gowda, Sujith M., Gowda, Supreeth M., Heffernan, Neil T.

In recent years, classroom use of educational software for tutoring and assessment has increased considerably. Aside from the benefits to learners, these systems provide extensive data on student learning processes. In recent years, researchers in educational data mining (EDM) have utilized this data in order to develop models of a range of constructs which have been historically difficult to assess at scale, including learner engagement, emotion, knowledge, and preparation for future learning. These models, sometimes termed "automated detectors," can be embedded into adaptive educational software or used to analyze the contexts where learning and engagement are enhanced.

Using these models creates the opportunity to better understand and, in turn, improve learning and engagement, which may particularly benefit underserved populations for whom the current generation of learning interventions are less effective. However, this will only be true if the models are accurate for the populations for who use the software. Unfortunately, insufficient attention has been placed on the reliability of these models when applied to populations – e.g. populations in different countries or among different demographics within-country – that are culturally distinct from those which the model was trained upon.

In this paper, we analyze the population-level reliability of automated detectors that assess four educationally relevant affective states (boredom, engaged concentration, frustration, and confusion) in ASSISTments, online software that supports learning and formative assessment in middle school mathematics. These detectors are developed through a two-step process, where field observations conducted by trained coders are synchronized to log files of student interaction with the ASSISTments system, and data mining algorithms such as Regression Trees, J48, JRip and K* are used to infer the observed emotions from log files patterns. Detectors are evaluated based on two metrics commonly used in EDM research: Kappa, which assesses the degree to which the detector's predictions are better than chance, and A', which assesses the proportion of time the detector can distinguish a specific emotion.

We analyze the goodness of these models when applied across urban (predominantly lower-income and Latino), rural (predominantly lower-income and White), and suburban (predominantly high-income and White and East Asian) populations in the Northeastern United States. We find that models trained only on students from a single population perform better than chance when applied to new students from the same population (average kappa = 0.27, average A' = 0.66), but are inappropriate for different populations, where they perform essentially at chance (average kappa = 0.003, average A' = 0.51).

In order to develop a model that can be confidently used with a broader population, we develop models using data from all three populations. These models perform better than chance when applied to new students from any of the three populations (average kappa = 0.24, average A' = 0.66) – approximately as well as models perform when developed and tested within a single population.

Finally, we discuss how these results may inform future research and development in EDM-based assessment, towards creating standards for how to develop automated detectors and interventions that are fully culturally responsive and appropriate.