Computers in Human Behavior 47 (2015) 168–181

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory

Xing Wanli [1], Guo Rui, Petakovic Eva [2], Goggins Sean *

School of Information Science & Learning Technologies, University of Missouri, Columbia, MO 65211, USA

ARTICLE INFO

Article history:
Available online 24 November 2014

Keywords:
Learning analytics
Educational data mining
Prediction
CSCL
Activity theory
Genetic Programming

ABSTRACT

Building a student performance prediction model that is both practical and understandable for users is a challenging task fraught with confounding factors to collect and measure. Most current prediction models are difficult for teachers to interpret. This poses significant problems for model use (e.g. personalizing education and intervention) as well as model evaluation. In this paper, we synthesize learning analytics approaches, educational data mining (EDM) and HCI theory to explore the development of more usable prediction models and prediction model representations using data from a collaborative geometry problem solving environment: Virtual Math Teams with Geogebra (VMTwG). First, based on theory proposed by Hrastinski (2009) establishing online learning as online participation, we operationalized activity theory to holistically quantify students' participation in the CSCL (Computer-supported Collaborative Learning) course. As a result, 6 variables, *Subject*, *Rules*, *Tools*, *Division of Labor*, *Community*, and *Object*, are constructed. This analysis of variables prior to the application of a model distinguishes our approach from prior approaches (feature selection, Ad-hoc guesswork etc.). The approach described diminishes data dimensionality and systematically contextualizes data in a semantic background. Secondly, an advanced modeling technique, Genetic Programming (GP), underlies the developed prediction model. We demonstrate how connecting the structure of VMTwG trace data to a theoretical framework and processing that data using the GP algorithmic approach outperforms traditional models in prediction rate and interpretability. Theoretical and practical implications are then discussed.

1. Introduction

The ability to predict a student's final performance has gained increased emphasis in education (Baker & Yacef, 2009; Romero & Ventura, 2010). One of the practical applications of student performance prediction is for instructors to monitor students' progress and identify at-risk students in order to provide timely interventions (Bienkowski, Feng & Means, 2012). It is already difficult to detect at-risk students in a regular classroom, not to mention when classes are much larger and learning happens online, as in MOOCs (Gunnarsson & Alterman, 2012). It would be desirable to expand beyond the at-risk students to predict the future performance of all students to allow a feedback process to enhance learning and awareness for a greater number of students during the course (Zafra & Ventura, 2009). As an automated method, student performance prediction has the potential to decrease teachers' duty in assessment.

The objective of performance prediction is to estimate an unknown value – the final performance of the student. In order to accomplish this goal, a training set of previously labeled data instances is used to guide the learning process (Espejo, Ventura, & Herrera, 2010) while another set of correctly labeled instances, named the 'test set', is employed to measure the quality of the prediction model obtained (Márquez-Vera, Cano, Romero, & Ventura, 2013). Previous studies that have documented student performance prediction models have focused on statistical modeling and data mining techniques (Gunnarsson & Alterman, 2012; Thomas & Galambos, 2004; Wolff, Zdrahal, Nikolov, & Pantucek, 2013). These traditional modeling techniques have their own limitations. From the perspective of educational data mining (EDM), which focuses on model and algorithm development to improve predictions of learning outcomes (Siemens & Baker, 2012), existing statistical and data mining methods typically lack an established

* Corresponding author. Tel.: +1 484 683 5037.
E-mail addresses: wxdg5@mail.missouri.edu (W. Xing), evarooka@gmail.com (E. Petakovic), GogginsS@missouri.edu (S. Goggins).
[1] Tel.: +1 281 309 8515.
[2] Tel.: +1 215 948 2729.

http://dx.doi.org/10.1016/j.chb.2014.09.034
0747-5632/

paradigm for optimizing performance prediction. For example, statistical models such as linear regression or logistic regression have requirements related to the distribution of data and a priori regression function structures. Poor estimation and inaccurate inferences would be generated if the basic premises of the regression models are breached (Harrell, 2001); and it is difficult for end users to detect when such breaches occur. In addition, there is a strong tradition in the domain of education of employing linear or quadratic models, limiting exploration of potentially more useful models for predicting student performance.

Learning analytics designed to support performance prediction are the type of actionable intelligence teachers and students require to improve learning, and inherently involves the interpretation and contextualization of data (Agudo-Peregrina, Iglesias-Pradas, Conde-González, & Hernández-García, 2013). Model interpretability in performance prediction is important for two primary reasons (Henery, 1994): first, the constructed model is usually assumed to support decisions made by human users – in our context, to facilitate teachers to provide individualized suggestions to students. If the discovered model is a black-box, which renders predictions without explanation or justification, people or teachers may not have confidence in it. Second, if the model is not understandable, users may not be able to validate it. This hinders the interactive aspect of knowledge validation and refinement. Unfortunately, traditional prediction models (e.g. support vector machines, neural networks) require a sophisticated understanding of computation that most teachers do not possess (Romero & Ventura, 2010; Siemens & Baker, 2012). If teachers cannot interpret analytics, they cannot provide meaningful feedback to students. For instance, Campbell, DeBlois, and Oblinger (2007) employed logistic regression, neural networks and other models to search for students that are at-risk of failing and alert instructors to potential issues. While automatic alert messages enable teachers to quickly identify struggling students, the generation of a risk signal is unable to convey enough information to enable personalized interventions for students (Essa & Ayad, 2012). From an application perspective, typical data mining algorithms usually work as black boxes, and as a result, it is difficult to identify the relationship between student performance and the various factors affecting performance. In turn, these models demand far more time and computing resources. Moreover, most previous studies stopped at the level of predicting failure and success of a student in a course or a program (e.g. Hämäläinen & Vinni, 2006; Romero, López, Luna, & Ventura, 2013; Zafra & Ventura, 2009), while few went further to predict student performance at more granular levels. With focus put solely on low performing students interventions have the risk of becoming a tool only for punitive interventions (Mintrop & Sunderman, 2009).

Moreover, previous research in forecasting students' performance has concentrated on methodology and the exploration of algorithms in ways tending to overlook educational contexts, theories, and phenomena (Baker & Yacef, 2009; Romero & Ventura, 2010). Many times, computational model results are at least difficult, if not impossible, for teachers to use and explain (Ferguson, 2012). To gain a deeper understanding of the factors influencing students' learning and to build an interpretable student performance prediction model, researchers must contextualize those data factors using educational theories and corresponding semantics. The number of factors (variables) affecting students' performance makes this a difficult challenge. A large set of selected variables can dramatically diminish both statistical and data mining prediction power (Deegalla & Bostrom, 2006; Vanneschi & Poli, 2012). Data dimensionality can be reduced using feature selection, but in educational situations in which human judgment is key (Siemens & Baker, 2012), it is more suitable to accomplish dimensionality reduction by constructing variables according to human theories (Fancsali, 2011). The automatic processing of data generated by these environments without the additional lens of theory provides a kind of "blunt computational instrument". Feature selection algorithms, statistical models and data mining grounded in mathematical theories lack connection to theories of human behavior that are most relevant in a learning analytics system. In practice, approaches to variable selection and construction are usually based on ad-hoc guesswork or significantly detailed experience in the educational field (Cetintas, Si, Xin, & Hord, 2009; Nasiri & Minaei, 2012; Tair & El-Halees, 2012). A principled, theory-based method for synthesizing factors from raw data will connect the input to computational prediction models more coherently than previous approaches.

1.1. Our framework for exploring more understandable prediction

This paper illustrates the potential for the integration of prediction models focused on automating analytics around humans working in computational systems to increase the understandability and utility of learning analytics. We selected the prediction model (Genetic Programming) that represents what we see in our results as the most optimal tradeoff between model understandability and the predication accuracy. To explore this aim, we synthesize prior work in learning analytics, EDM and activity theory to approach student performance prediction model construction. We draw on a theory proposed by Hrastinski (2009), which emphasizes participation in online learning as a central factor affecting performance. We then contextualize participation-related data factors on a semantic background using an operationalization of activity theory. Integrating activity theory directly into our operationalization of participation indicators allows for a systematic construction of variables and reduces data dimensionality in a CSCL environment to only six aspects.

We then use activity theory derived participation indicators as inputs to a Genetic Programming (GP) model to develop our student performance prediction model. The GP model can build a prediction model without assuming any a priori structure of functions and relies on theoretically grounded factorization of data. Moreover, the proposed GP model is more easily understood by users when compared with traditional statistical and data mining algorithms, providing teachers actionable information to offer individualized suggestions to students in any performance state (at-risk, just survive, average or good) as well as increasing students' awareness provided that prediction results are also presented to them. As a final product, this model defines tangible relationships between student performance and its related variables. Therefore, in terms of practical application, the resulting prediction model may be easily implemented in a real life context.

This study provides a practical and interpretable student performance prediction model that enables teachers to discern differences in performance among students in a classroom full of small group geometry learners who are working in groups of three to five in a synchronous CSCL environment, Virtual Math Teams with Geogebra (VMTwG). The paper is organized as follows: Section 2 discusses related work and background information. Section 3 introduces the theoretical framework behind this study. Section 4 shows the context of the study and data format. Section 5 describes methodology. Section 6 presents experimental results and analysis. Section 7 discusses results. Section 8 summarizes this study, pointing out limitations and future research directions.

2. Literature review

The development of student performance prediction models is one of the oldest and most popular practices in education