Better Data Beat Big Data

Michael V. Yudelson, Stephen E. Fancsali, Steven Ritter, Susan R. Berman, Tristan Nixon, Ambarish Joshi
Carnegie Learning, Inc.
437 Grant St.
Pittsburgh, PA 15219
(412) 690-2442
{myudelson, sfancsali, sritter, sberman, tnixon, ajoshi}@carnegielearning.com

ABSTRACT

Generalizability of models of student learning is a highly desirable feature. As new students interact with educational systems, highly predictive models, tuned to increasing amounts of data from previous learners, presumably allow such systems to provide a more individualized, optimal learning path, give better feedback, and provide a more effective learning experience. However, any large student/user population will be heterogeneous and likely consist of discernable sub-populations for which specific models of learning may be appropriate. Student sub-populations may differ with respect to cognitive factors, the level and quality of instruction, and many other environmental and non-cognitive factors.

The era of both "big data" and widely deployed educational software, including Carnegie Learning's Cognitive Tutor (CLCT) intelligent tutoring system, presents opportunities to analyze increasingly large volumes of data collected during learners' interactions with educational systems. These data cover a broad spectrum of learners, allowing researchers to investigate the structure of an increasingly representative student population. In this work, we investigate discovering student sub-populations from "big data." Using a year's worth of data from CLCT, we test the hypothesis that commonly used stratifications of student sub-populations (e.g., school location, socio-demographic factors) offer ways to meaningfully partition learners. We discover that, rather than finding distinct subpopulations that should be treated differently, a particular sub-population of learners provides especially "high quality" data and that models learned from this sub-population outperform all other models even when predicting student learning for the sub-population on which other models were trained. In this way, "better data beat big data."

Keywords

Big data, student modeling, learner sub-populations.

1. INTRODUCTION

Generalizability is an important property of any model of student learning developed by researchers and practitioners in educational data mining, learning analytics, and cognitive modeling. As such, investigators generally aim to iteratively refine models of student learning based on data as it is acquired; experimental iteration informs future versions of computer-based educational systems so that such systems can adapt to (and better serve) larger populations of learners.

Discovering the appropriate grain size (e.g., learning models at the group-, school-, or class-level versus individualized, student-level models) to achieve such generalizability is a topic of recent interest in the literature. The student population (i.e., the user base of an educational system) is likely to be heterogeneous, and important aspects of its structure can potentially be identified. Student sub-populations may have particular characteristics and profiles that can be stratified with respect to demographics, learning capabilities, instructional quality, among other factors. Less clear are ways in which such stratifications can be useful for determining sub-populations over which better models of student learning might be learned.

A body of prior work goes beyond building models of undifferentiated populations, modeling individual student differences [4, 8] and also modeling groups of students (e.g., classes and schools) [5, 7]. Other work builds models of student behavior and compares sub-populations defined by school setting (e.g., urban, suburban, or rural) [1]. Most efforts to model individual student differences or to stratify student sub-populations consider relatively small datasets, with an exception of work by Pardos and Heffernan that uses the largest open access dataset on student learning currently available – the KDD Cup 2010 dataset.[1]

On an industrial scale, adapting at the student- and/or group-level provides an opportunity to deliver an optimized learning experience to a large user base, for example, the hundreds of thousands of users of Carnegie Learning's Cognitive Tutor® (CLCT) intelligent tutoring system (ITS) [6]. Using CLCT data, we focus on the discovery of student sub-populations over which parameters used to track student mastery of knowledge components (KCs) or skills can be learned (i.e., "tuned") to better deliver instructional content to different sub-populations. Little (if any) prior research considers what data to include in an *a priori* school profile that might determine appropriate sub-populations (i.e., groups of schools) for such tuning and similarly for *a posteriori* profiles that include student interaction data after CLCT has been used for a substantive period of time.

In this work, we explore the possibility of utilizing information about a particular school (e.g., demographic and socioeconomic indicators) and about its students (e.g., prior performance) to effectively structure a large selection of schools into distinctive groups to determine if and how groups of schools might benefit

[1] KDD Cup 2010 http://pslcdatashop.web.cmu.edu/KDDCup/

from a specific parameter tuning of the CLCT. We set out to discover generalizable sub-populations of schools, but rather we find that a subset of schools provides "high quality" data, models of which effectively generalize to all schools in our sample and outperform (in terms of prediction accuracy on held out data) models learned on other subsets and larger samples of data. In this sense: better data beat big data.

2. CARNEGIE LEARNING COGNITIVE TUTOR

CLCT is an ITS for mathematics that uses cognitive modeling to structure a target domain (e.g., algebra) into knowledge components (KCs). CLCT adapts instruction based on its assessment of which KCs a learner has or has not mastered at any given moment. CLCT provides feedback as to the correctness of their actions on problem-solving steps and also provides context-sensitive hints upon request. Curricula, like algebra, are divided into units of instruction; units are comprised of topical sections, and sections consist of individual problems that are broken up into steps. Problem-solving steps are tagged with one or more KCs.

As students solve problems, CLCT updates its assessment of students' KC mastery using a probabilistic framework called Bayesian Knowledge Tracing (BKT) [3]. BKT is a Hidden Markov Model with two hidden states, representing whether a particular KC is un-mastered or mastered. Observations of student performance on opportunities to practice a KC are binary: a student either solves a problem step correctly or not (due to error or because of a hint request). While students might go through dozens of attempts to get a particular step correct, traditionally, only students' first attempts are considered for updating KC mastery estimates.

BKT uses probabilistic parameters to capture the nature of mastering a skill. These parameters are the probability of knowing the skill *a priori*, the probability of learning the skill at the next practice attempt (i.e., transitioning from the unknown state to the known state), the probability of guessing correctly while in the un-mastered state, and the probability of slipping (i.e., answering incorrectly despite being in the mastered state). In the commercial deployment of CLCT, BKT parameters are set by hand by cognitive scientists and also go through revisions based on data.

3. DATA

We consider a large set of CLCT student usage data, collected in 2010. Although the tutor was used in several thousand schools across the United States, we do not collect detailed interactions for all schools, so our initial data covered 144,080 registered student accounts in 899 schools with close to 473 million records overall, including activity unrelated to problem-solving (e.g., login) as well as solving practice problems. Unfortunately, not all registered students used the tutor or attempted more than one unit of the curriculum. After trimming down the data we arrived at a dataset that included 342 schools, 72,082 active students, and 88.6 million problem-solving actions.

We queried the National Center for Education Statistics (NCES)[2] for school metadata that included: the number of students enrolled (as a proxy of school's relative size), student-teacher ratio, number of students eligible to receive free or reduced price lunch (as a proxy for socioeconomic status), and the school's location (metropolitan area): rural, suburban, or urban. Although some of the school metadata from NCES were from the year 2011, we assume that year-to-year fluctuations are negligible. We matched NCES data and our data and arrived at a set of 232 schools, narrowing our selection to 55,012 students with substantive usage (i.e., attempting more than one unit of instruction) and 67.3 million problem-solving transactions.

In addition to school metadata, we computed school-level student performance statistics from our logs. For each school, we have computed: the average number of distinct units students were attempting, the standard error of the mean number of units attempted, number of distinct units students attempted. We have also retained a binary vector of units attempted by schools' students for grouping schools based on the similarity of attempted units.

To further characterize schools, we ran a mixed effects logistic regression model on the data (see Eq. (1) and Eq. (2)). Here, θ_i represents the ability of student i (a student intercept), and β_j is a problem complexity intercept. For each skill k relevant to problem j, δ_k is general skill easiness (i.e., a skill intercept), and γ_k represents skill k's learning rate; t_{ik} captures student i's number of prior attempts at skill k.

$$m_{ij} = \theta_i + \beta_j + \sum_k \left(\delta_k + t_{ik}\gamma_k\right) \quad \text{Eq. (1)}$$

$$\Pr(Y_{ij} = 1 \mid \theta, \beta, \delta, \gamma) = \frac{1}{1 + e^{-m_{ij}}} \quad \text{Eq. (2)}$$

In this regression model, we treat the student- and problem-intercepts as random factors. From the regression coefficients, we calculated the following values to describe, per school: average student intercept (denotes relative prior preparation of students), average skill intercept (to capture each school's general level of skill difficulties on top of student preparation), and average skill slope (to denote the relative speed of learning for students). Thus, overall we have collected, for each school, four *a priori* metadata descriptive factors and seven *a posteriori* student performance descriptive factors.

4. APPROACH

We seek to determine if, based on one or more descriptive factors described above, it is possible to effectively separate schools in our dataset into groups such that schools within groups are more similar to each other in terms of learning than to schools in other groups. We propose to use the accuracy of student modeling as a measure of similarity. That is, if a student model fit to a particular group of schools predicts performance of students in these schools better than models fit to the data of other groups of schools and this is true for all group models, then the school grouping in question effectively separates schools into distinguishable sub-populations.

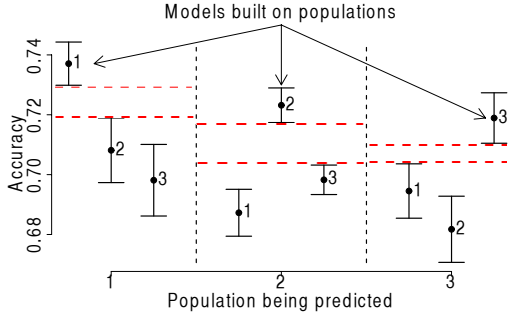

Figure 1. An example criterion of a good split into sub-groups

[2] National Center for Education Statistics http://nces.ed.gov

An illustration of an effective separation of schools into groups is shown in Figure 1. In this graph using idealized data, all schools are split into three groups (or populations). Based on the data from each of the groups we built three models. Each of the three models are used to predict held out data from each of the three groups of schools giving us 3*3=9 predictions. Prediction of held-out data for group of schools #1 is shown in the leftmost column where the accuracy of each of the three models' predictions are shown as dots with serifs denoting standard errors of the mean. Here, we see that model built on group #1 performs better on held out data than models built on the data from groups #2 and #3. Since the range of the serif denoting standard error of the mean for model #1 does not overlap with serif ranges for models #2 and #3, the advantage of model #1 is deemed "significant." Columns 2 and 3 show the same phenomenon: a model built on the data from the respective subgroup outperforms models built on other subgroups.

4.1 Dividing Schools

We have considered all eleven descriptive factors to guide groupings of schools: 1) school locale, 2) percentage of students eligible for free and reduced priced lunch, 3) student-teacher ratio, 4) enrollment, 5) average student units attempted, 6) standard error of student units attempted, 7) number of unique units students attempted, 8) school unit coverage group (based on similarity of binary vectors of distinct units attempted by students in particular school)[3], 9) average student intercept from the logistic regression model (a proxy of average student preparation in the school), 10) average skill intercept for the school from the logistic regression, and 11) average logistic regression skill slope for the school. The factors are grouped into three batches: school metadata factors that are known *a priori*, student usage statistics factors that can be computed from surface logs of student activity, and student model factors that require detailed data to be derived.

Among all factors, school locale and the school unit coverage group are categorical factors. We binned the remaining nine continuous factors into three value ranges – low, medium, and high – so that the number of students in all three is roughly the same. In addition to splitting schools using just one factor, we have computed school splits based on multiple factors. Namely, all factors from all groups[4], only school metadata factors, only student usage factors, only student model factors, and all *a posteriori* student factors (student usage and model factors). The multi-factor groupings were produced with the help of `R` package `cluster` using Goward distances metric and Ward's hierarchical clustering algorithm via function `hclust` with the number of clusters set to 3 for simplicity.

4.2 Cross-Validating School Groups

Since the number of the schools varied across single-factor and multi-factor splits, we sampled 30 schools from each group where 20 schools were used for training a group model and 10 schools were set aside as held out test data. Rather than relying on single-point estimations of model accuracy, we repeated sampling 20 times and obtained the means and the standard errors of prediction accuracies. Thus, for each grouping we selected 20 (samples)*3(groups)*2(fit and test)=120 data sets; within each of the 20 samples fit and test data for a particular group of schools did not overlap, while across samples they could.

For each of the 20 samples we fit three group models. Each of the three models is used thrice to predict three held-out data sets for each of the groups (9 predictions overall). Fitting models and producing prediction accuracies was done with the help of a BKT utility built for use with large datasets [8].

We stipulate that, in order for a grouping of schools to be considered producing distinct groups, for every group, the in-group prediction should be significantly better than out-group prediction (cf. Figure 1).

5. RESULTS

First, we consider several school metadata factors, knowable *a priori* (prior to any student usage of CLCT). Figure 2 is a group split graph for school enrollment. As we can see, models built on groups of low and middle ranges of enrollment are not discernable from each other across all prediction tasks. The model built on high enrollment schools is visibly worse even when predicting held out data of high enrollment schools.

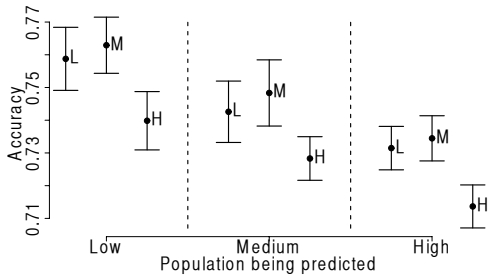

Figure 2. Group separation by school enrollment

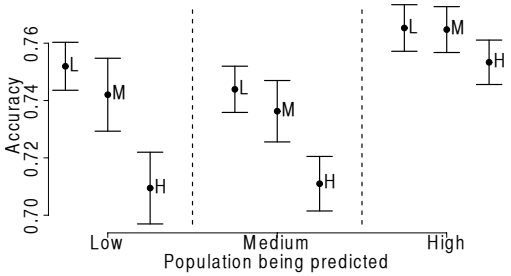

Figure 3. Group separation by the ratio of students eligible for free and reduced price lunch

Figure 3 is a group separation graph for the ratio of students eligible for free and reduced price lunch. Again, we see that this factor is not separating schools into reliably discernable groups. Models built on schools with a high proportion of students eligible for free and reduced price lunch are visibly worse across all populations, while models of low and medium groups are not discernable, again across all populations.

Neither school metadata factors separately nor a grouping based on a clustering solution of these metadata factors produce a desirable split. Instead, we see model accuracies lined up in identical fashion: one particular model is a slightly better predictor universally; a second model is slightly worse, and the remaining model is worse than the second.

However, for 3 out of 7 remaining individual factors and one multi-factor case (all factors but metro area), models built on one group of schools are consistently and significantly better than other models in at least 2 prediction tasks. See, for example,

[3] The grouping was done with the help of `R` package `cluster` using Euclidean distances and Ward's hierarchical clustering algorithm via function `hclust` with k=3.

[4] School locale factor was excluded since using it defaulted the clustering to be identical to the metro area factor itself.

Figure 4. Here, schools where students finish a high number of units on average (more than 9.2 units) produce a model that outperforms another model in two out of three comparisons and ties in third.

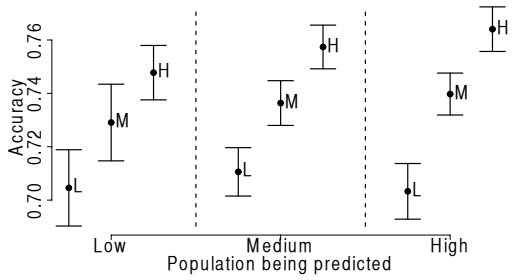

Figure 4. Group separation by average student units attempted

We find a similar pattern for average student intercept (a proxy of average student preparation), where the model built on a group of better-prepared students wins in two comparisons and ties in one. The third factor with one-model-trumps-all is the average skill slope (a proxy of speed of learning), where the winning model actually is built on the group of schools where the average skill slope is in the medium range. When cross-correlated, only the correlation of average units attempted and average student intercept is relatively high and significant ($r=0.56$, $p<0.001$).

6. DISCUSSION

We set out to discover subsets of schools for which models of practice could be built for sub-populations to optimize the CLCT learning experience for students in that sub-population. Instead, we find that particular sub-populations of schools can be used to learn parameters that perform best over the *entire population*. In essence, we have identified a set of schools for which particular aspects of their interaction with the CLCT provide high-quality (e.g., less "noisy") data for such model building.

While this substantial subset may still count as "big" data, we disregard a large number of students to arrive at this generalizable model, and the characteristics along which the group of schools from which these students are drawn are not obvious *a priori*. While much focus is placed on the revolutionary potential of big data applications in education, careful consideration and attention must be paid to the quality of such data for particular purposes and application contexts.

We find that the sub-populations that yield a universally better model tend to contain students who are better prepared and students who attempt more CLCT units.. However, with respect to average skill learning rates, the best model contains many students in the "middle" group. At this point we hypothesize that students that should be considered for inclusion in learning a generalizable model are not just better students but those that yield a substantial data footprint in terms of curriculum coverage. Students who should likely be excluded are those who only cover a fragment of units, insufficient to provide for a "good" model.

Several caveats could hinder how strongly the phenomenon of "better data" vs. "big data" manifests itself. One is that CLCT allows instructors to deploy "custom" curricula; different schools sometimes use different content units and, as a result, practice different skills. Consequently, when validating the model on the held out data where a particular unit was not practiced, we used default modeling parameters that could potentially lead to lower accuracy. Together with known issues with fitting BKT models (e.g., local maxima and non-identifiability [2]), this might have led to the inter-group differences being underestimated and the effect of "one group model takes all" – lessened.

Second, we cannot judge, for example, whether our 2010 dataset constitutes a representative sample of all US schools with respect to the school metadata variables we considered. However, we estimated whether our selected subset of 232 schools maintains the same distribution of the school locale (i.e., whether schools are rural, urban, and suburban) as that over 729 school of our original 899 schools for which we have appropriate data to make the comparison. The split between rural, suburban, and urban schools in the larger sample of 729 schools are 29%, 33%, and 38%, respectively. Our smaller sample of 232 schools breakdown as 29%, 25%, and 46%, respectively. While the percentage of the rural schools is the same, the percentage of urban schools significantly grew, and the ratio of suburban schools declined. While this may introduce bias, it is unclear whether such bias, given the relatively large sample overall, would have a substantive impact on the generalizability of our results.

7. REFERENCES

[1] Baker, R. S. J. de & Gowda, S. M. (2010). An Analysis of the Differences in the Frequency of Students' Disengagement in Urban, Rural, and Suburban High Schools. In 3rd International Conference on Educational Data Mining (EDM 2010), Pittsburgh, PA, USA, 2010 (pp. 11-20).

[2] Beck, J. E. & Chang, K.-m. (2007). Identifiability: A Fundamental Problem of Student Modeling. In User Modeling, Corfu, Greece, 2007 (pp. 137-146). Springer.

[3] Corbett, A. T. & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4), 253-278.

[4] Lee, J. I. & Brunskill, E. (2012). The Impact on Individualizing Student Models on Necessary Practice Opportunities. In Proceedings of the 5th International Conference on Educational Data Mining, Chania, Greece, June 19-21, 2012, Chania, Greece, 2012 (pp. 118-125).

[5] Pardos, Z. A. & Heffernan, N. T. (2010). Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. In 18th International Conference on User Modeling, Adaptation, and Personalization (UMAP 2010), Big Island, HI, USA, 2010 (pp. 255-266). Springer.

[6] Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.T. (2007). Cognitive Tutor: applied research in mathematics education. Psychon Bull Rev, 14:249-255.

[7] Wang, Y. & Beck, J. (2013). Class vs. Student in a Bayesian Network Student Model. In 16th International Conference on Artificial Intelligence in Education (AIED 2013), Memphis, TN, USA, 2013 (pp. 151-160). Springer.

[8] Yudelson, M., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian Knowledge Tracing Models. In 16th International Conference on Artificial Intelligence in Education (AIED 2013), Memphis, TN, USA, 2013 (pp. 171-180). Springer.