(2016). The 2010 KDD Cup Competition dataset: Engaging the machine learning community in predictive learning analytics. *Journal of Learning Analytics*, *3*(2), 312–316. http://dx.doi.org/10.18608/jla.2016.32.16

The 2010 KDD Cup Competition Dataset: Engaging the Machine Learning Community in Predictive Learning Analytics

John Stamper
Human Computer Interaction Institute, Carnegie Mellon University, USA
john@stamper.org

Zachary A. Pardos
Graduate School of Education/School of Information, UC Berkeley, USA
zp@berkeley.edu

ABSTRACT: In the spring of 2010, the Association for Computing Machinery (ACM) Special Interest Group on Knowledge Discovery and Data-mining (KDD) selected a dataset from an educational technology for its annual competition. The competition, titled "Educational Data Mining Challenge," tasked participants with predicting the correctness of student answers to questions within an Intelligent Tutoring System (ITS) from The Cognitive Tutors suite. This challenge was hosted by the PSLC DataShop, and included data provided by Carnegie Learning Inc., producers of The Cognitive Tutors. Consisting of over 9GB of student data, this was the largest KDD Cup dataset to that time. The competition brought in 655 competitors submitting 3,400 solutions. Five years later, the competition dataset has been the most often cited from an educational technology platform.

Keywords: KDD Cup, Cognitive Tutor, DataShop, competition, Algebra, Intelligent Tutoring System

1 INTRODUCTION

The 2010 KDD Cup challenge engaged the extended machine learning community in a data mining challenge to forecast the correctness of student responses to questions within an Intelligent Tutoring System (ITS) for Algebra. Previous competitions included classification of cancer in mammogram images and classification of individual consumers as current or prospective customers based on a large array of customer analytics signals. The 2010 competition was the first to utilize a dataset from education and the chosen dataset, from Carnegie Learning's Cognitive Tutors, was the largest seen in this competition, underscoring educational technologies as a serious emerging source of big data.

The two datasets used in the primary competition were based on 8th graders' (13–14 year olds) use of the "Algebra" and "Bridge to Algebra" tutoring products during the 2007–2008 school year. In most cases, the tutoring system was purchased at the school district level as part of a full math curriculum sold by Carnegie Learning Inc. (www.carnegielearning.com). Three additional datasets of smaller size,

(2016). The 2010 KDD Cup Competition dataset: Engaging the machine learning community in predictive learning analytics. *Journal of Learning Analytics*, *3*(2), 312–316. http://dx.doi.org/10.18608/jla.2016.32.16

called the development dataset, were released before the official start to the competition to allow participants to become familiar with the data format as well as allow the curators of the datasets an opportunity to respond to data quality issues and rectify them in the official competition sets. The development datasets consisted of the previous two years of data (2005–2007) of the "Algebra" tutor and one previous year's worth of data (2006–2007) for "Bridge to Algebra."

A typical student uses the Cognitive Tutor software to practice a specific skill until mastery has been achieved, after which the student may progress to the next section of the material. As a true adaptive intelligent tutor, students are given feedback on their responses as well as receiving solution hints. The focal learning strategy in an ITS is based on problem solving rather than video lectures featured in many Massive Open Online Courses, including Khan Academy. The specification of skills and assessment of a student's state of mastery are key components of this and many other tutors that allow for adapting the amount of practice based on each individual student's pace. The skills associated with a question in the tutor was a key piece of meta information included in the dataset that allowed competitors to make fine-grained discriminations between questions that would be predicted correct or incorrect. Multiple skill associations, more commonly referred to as knowledge components (KCs) in the ITS literature (Koedinger, et al. 2010), were included in the dataset. The more coarse-grained set came from cognitive task analysis of the content, while the finest-grained set represented the actual rules firing in the tutoring system to produce and check the correctness of the provided mathematical answer.

The top performing solutions incorporated a variety of data mining techniques and machine learned classifiers. Hand feature engineering played a role in several solutions (Pardos & Heffernan, 2010; Yu et al., 2010). An example of a successful hand-engineered feature included Z-score of time spent by a student on questions of a particular skill, among other features found in the educational data mining literature (Koedinger, D'Mello, McLaughlin, Pardos, & Rosé, 2015). Automatic feature generation was a key to the success of Yu et al. (2010), using clustering techniques to create millions of features and then applying a custom logistic regression (Yu, Hsieh, Chang, & Lin, 2012) modified to train a model on such a massive corpus of features using common PC/laptop hardware. Both Pardos & Heffernan (2010) and Yu et al. (2010) used Random Forests (Breiman, 2001), a mainstay in data mining competitions, as a central classifier. Random Forests combines, through bagging, the predictions of many individual decision trees that differ in their predictions by training on both a random re-sampling of the training data and a random sampling of the features.

The second place finishers (Toscher & Jahrer, 2010) employed matrix factorization as their primary classifier. This was the same team and method used as part of the million dollar Netflix prize (Bennett & Lanning, 2007). Matrix factorization is based on searching for values for factors of questions and factors of students that when multiplied together result in predictions of correct and incorrect answers. The values for the factors are arrived at based on the accuracy with which they predict the already observed correct and incorrect responses. The predictions to unseen questions in the test set are used as the

(2016). The 2010 KDD Cup Competition dataset: Engaging the machine learning community in predictive learning analytics. *Journal of Learning Analytics*, *3*(2), 312–316. http://dx.doi.org/10.18608/jla.2016.32.16

forecasts. The algorithm, internal to the Cognitive Tutor ITS, which infers if a student learned or mastered a skill, is called Bayesian Knowledge Tracing (Corbett & Anderson, 1995). A heavily extended version of this algorithm that allows for individualized parameters was designed and applied by Pardos & Heffernan (2010).

All top solutions made use of multiple classifiers and combined them using a form of "ensembling," a method by which multiple predictors are combined to produce a result that is more accurate than either predictor alone (Pardos, Gowda, Baker, & Heffernan, 2011). The matrix factorization group used neural networks as their ensemble method to combine multiple matrix factorization models. The first place finishers used SVMs and Random Forests to craft their composition of methods, while the fourth place finisher used Ensemble Selection, a form of simple linear weighting utilized by the previous year's winning team (Caruana, Niculescu-Mizil, Crew, & Ksikes, 2004).

2 THE DATASET

The initial datasets were provided to DataShop in traditional format (transaction level) from Carnegie Learning, originated from two distinct cognitive tutoring systems. The datasets come from multiple schools over multiple school years. The challenge data sets had not been made available to researchers prior to the 2010 KDD Cup. The data was imported into DataShop using the tutor message format (https://pslcdatashop.web.cmu.edu/dtd/). As part of the import process, DataShop automatically created a "step rollup" file of every transaction dataset, which created a single row for all the transactions relating to one student attempt at a step.

Once the step rollup file was provided to the 2010 KDD Cup chairs, the file was split into a training set and a test set, with each data set broken in two. A third submission file was provided for the results. The submission file contains a subset of the columns in the test file. Each data set was split as seen in Figure 1, where each horizontal line represents a student-step (a record of a student working on a step). The data set is broken down by student, unit (a classification of a portion of the math curriculum hierarchy, e.g., "Linear Inequality Graphing"), section (a portion of the curriculum that falls within a unit, e.g., "Section 1 of 3"), and problem. Test rows were determined by a program that randomly selected one problem for each student within a unit, and placed all student-step rows for that student and problem in the test file. Based on time, all preceding student-step rows for the unit were placed in a training file, while all following student-step rows for that unit were discarded. The goal at testing time was to predict whether the student got the step right on the first attempt for each step in that problem. Each prediction was created by the competition participants in the form of a value between 0 and 1 for the column "Correct First Attempt" and entered with the submission file.

(2016). The 2010 KDD Cup Competition dataset: Engaging the machine learning community in predictive learning analytics. *Journal of Learning Analytics*, *3*(2), 312–316. http://dx.doi.org/10.18608/jla.2016.32.16

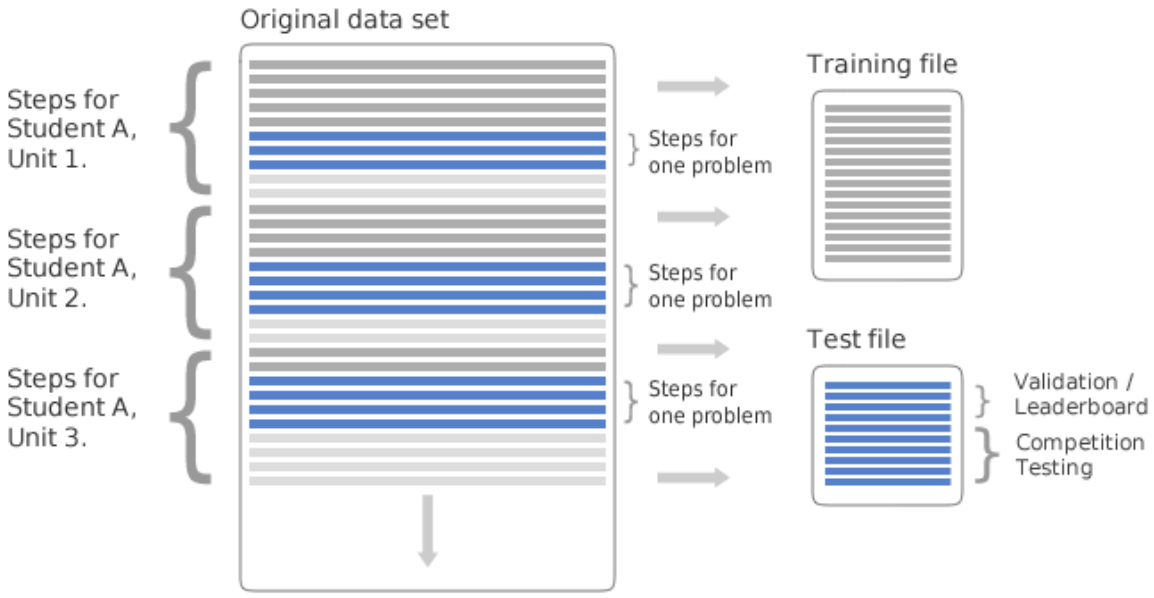

Figure 1: Example of the splitting of the original Step Rollup file into Training and Test sets.

3 ACCESS

The website hosting the competition continues to operate and is the main location to access the datasets (https://pslcdatashop.web.cmu.edu/KDDCup/). The scoring is still active and the competition website continues to be used to test methods and as a learning resource for machine learning and data mining courses in education, as well as more broadly. The dataset is freely re-distributable so long as the original data usage agreement is retained.

4 LIMITATIONS

The procedure for creating the training and testing set, described in section 2, means that the training set does not contain the complete log of student responses. A complete data dump of the 2006–2007 year is being considered. The trade-off would be that the test set for the KDD Cup leaderboard, still being used to date, would now be public, enabling continuing participants to over-fit to the test set.

The second limitation is that this dataset does not contain information on the content of the questions being answered aside from the skill associated with it and the internal name of the problem and its unit.

REFERENCES

Bennett, J., & Lanning, S. (2007). The Netflix prize. *Proceedings of KDD cup and workshop*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.8094&rep=rep1&type=pdf

(2016). The 2010 KDD Cup Competition dataset: Engaging the machine learning community in predictive learning analytics. *Journal of Learning Analytics*, *3*(2), 312–316. http://dx.doi.org/10.18608/jla.2016.32.16

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. *Proceedings of the 21st International Conference on Machine Learning* (ICML-04), 345-353. http://dx.doi.org/10.1145/1015330.1015432

Corbett, A., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction, 4*, 253–278. http://dx.doi.org/10.1007/BF01099821

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rosé, C. P. (2015). Data mining and education. *WIREs Cognitive Science*, *6*, 333–353. http://dx.doi.org/10.1002/wcs.1350

Koedinger, K. R., Baker, R. S. J. d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J., (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. d. Baker (Eds.), *Handbook of educational data mining*, (pp.43–55*)*. Boca Raton, FL: CRC Press.

Pardos, Z. A., & Heffernan, N. T. (2010). Using HMMs and bagged decision trees to leverage rich features of user and skill. Paper presented at the KDD Cup 2010 Workshop as a part of 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC. Retrieved from http://pslcdatashop.org/KDDCup/workshop/

Pardos, Z. A., Gowda, S. M., Baker, R. S. J. d., & Heffernan, N. T. (2011). The sum is greater than the parts: Ensembling models of student knowledge in educational software. *ACM SIGKDD Explorations, 13*(2), 37–44. http://dx.doi.org/10.1145/2207243.2207249

Toscher, A., & Jahrer, M. (2010). Collaborative filtering applied to educational data mining. Paper presented at the KDD Cup 2010 Workshop as a part of 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC. Retrieved from http://pslcdatashop.org/KDDCup/workshop/

Yu, H. F., Hsieh, C. J., Chang, K. W., & Lin, C. J. (2012). Large linear classification when data cannot fit in memory. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *5*(4), Article 23. http://dx.doi.org/10.1145/2086737.2086743

Yu, H. F., Lo, H. Y., Hsieh, H. P., Lou, J. K., McKenzie, T. G., Chou, J. W., et al., (2010). Feature engineering and classifier ensemble. Paper presented at the KDD Cup 2010 Workshop as a part of 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC. Retrieved from http://pslcdatashop.org/KDDCup/workshop/