The Learning Performance Vector

One of the main objectives and origins of the Lea's Box project is to contribute novel, theory-driven, and in particular CbKST/FCA-based ideas (cf. www.leas-box.eu) to the existing pool of learning analytics methods and techniques. One of the key challenges for learning analytics, in turn, is to predict learning outcomes and student performance. This is not a trivial challenge.

An educator may want to use such predictions to say if a student will get a question correct or incorrect, or might predict if a student is proficient in a certain skill, task, knowledge component, or competence. Teachers may also build predictive models of which students need intervention to avoid failing a course. These models can then be put back into the systems in which the data was collected. A recent overview of techniques is given for example by Shahiri (2015). Initiatives such as Carnegie Mellon' s DataShop (https://pslcdatashop.web.cmu.edu/) or competitions such as the KDD Cup (http://www.kdd.org/kdd-cup) reflect the state of the art in this field.

A good portion of the existing methods are statistics-based data mining techniques. These perform well on a general, statistical basis however have clear weaknesses when operating on a level of individual learners (a good example is for example described in the use case study of Hughes and Dobbins, 2015), Also, many statistical approaches build upon a set of (at least) debateable statistical assumptions and decision criteria. A well-elaborated overview about the strength and weaknesses of practical applications of such data mining approaches has been published by Papamitsiou and Economides (2014).

In the context of the Lea's Box project we developed an approach for predicting student performance which is based on the theoretical foundations of CbKST and which might offer an interesting, top-down technique to the field of performance prediction. The approach, named Learning Performance Vector (LPV), has been described in deliverable D3.4. In essence, the idea is that the structural information about the learning domain, the atomic units of aptitude (we name them competencies), and the relationships between these competencies provide a pool of important information for predications. In addition, we can add deeper information about the individual competencies which we call "weights". These weights reflect a competency's complexity, difficulty, or importance for a domain. Together with the actual performance data of student's we hypothesized that performance predictions can be improved. In the context of the project the LPV algorithm to predict a student's Learning Horizon (LH) have been developed and implemented in the Lea's Box system.

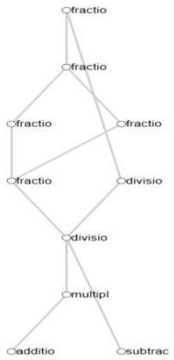
The purpose of this study is to evaluate and perhaps validate the predictive power of the approach in relation to a simple statistical approach. We have chosen a simulation study approach because this enables us to intentionally generate the data basis and observe the algorithmic steps.

Simulating Learner Data

The first step for this evaluation is to simulate realistic performance data of students. To build upon a realistic data set and to be able to make comparisons between simulated and real data, we selected a data set from Carnegie Mellon's DataShop. It is a data set of "Assistments Math 2004-2005", data set

id 92 (accessible at https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=92). This data set cover mathematics (which offers an easy 'playground' because it is a well-defined domain) and includes the data of 912 students. The data set is based on in total 80 competencies (knowledge components). For the simulation study we selected a subset of 11 competencies and established a competence model (prerequisite relation) among them (see the next figure). The weights are based on the inverse solution frequencies of the real data set.

ID	Competency	Weight
1	addition	0,1
2	subtraction	0,15
3	multiplication	0,27
4	division	0,4
5	fraction	0,45
6	division /w decimals	0,55
7	fraction multiplication	0,66
8	fraction division	0,7
9	fraction percents	0,8
10	fraction /w decimals	0,9



The left panel shows the select competencies and the assigned weights; the right panel shows the established prerequisite relation.

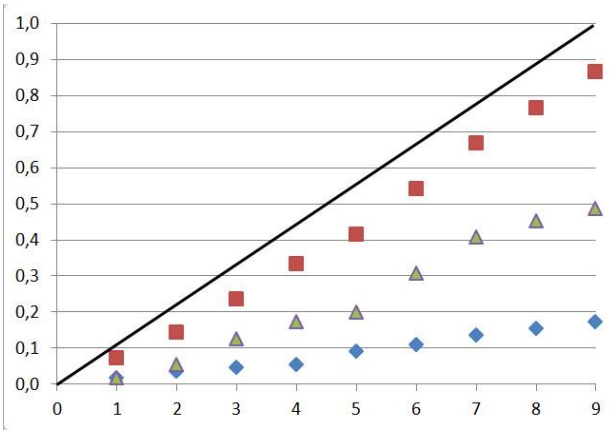
Furthermore, we selected 12 item types and 111 items from the data set. These cover one or more of the selected competencies, partially also other competencies (as shown in the next figure). Of course, also student data is required. In the first step we simulated 15 students with different abilities (or levels of expertise). The ability parameter was defined on a scale from 1 to 10, where 1 means no knowledge in the domain and 10 means having all competencies. The parameters were simulated on the basis of a normal distribution, assuring the medium level abilities are most common and extreme position rather seldom. Finally, because this study is about prediction, we simulated 9 time points with the assumption that in the time intervals learning occurs, depending on the student abilities.

Item Type	Competencies
1	1
2	1,2
3	1,3
4	2
5	1,2,3
6	1,2,3,4
7	1,2,3,4,5
8	1,2,3,4,5,6
9	1,2,3,4,5,7
10	1,2,3,4,5,7,8,9
11	1,2,3,4,5,7,8,10
12	1,2,3,5

Student	Ability (1-10)
1	2
2	9
3	3
4	3
5	4
6	5
7	5
8	6
9	7
10	7
11	8
12	8
13	9
14	5
15	4

The left panel shows the assignment of competencies to item types; the right panel shows the simulated distribution of student abilities.

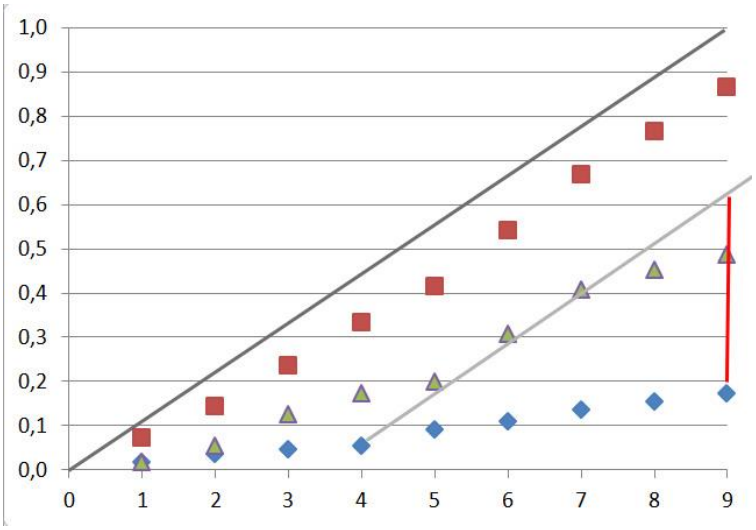
In summary, we simulated the answer patterns of 15 students across 9 time points in 111 fictitious test items, covering 12 competencies. The simulated data set consists of 1665 data points. The following chart shows the prototypical simulated results of an excellent learner (red), a medium learner (green), and a poor learner (blue). The values show the relative increase in correctly solved items over the 9 time intervals. The bold black diagonal indicates the optimal increase, so that with each of the 9 points in time 1/9 of the items is solved correctly – or in other terms, 1/9 of the competencies have been acquired. What the results show is that the increase is determined by the student abilities, due to error rates (lucky guesses and careless errors) we see that the optimal learners is a bit below the ideal diagonal while thee poor learner still shows a slight increase. This is an expected phenomenon we can find in many data sets.



The figure shows the simulated results of three prototypical students as opposed to the ideal learning performance

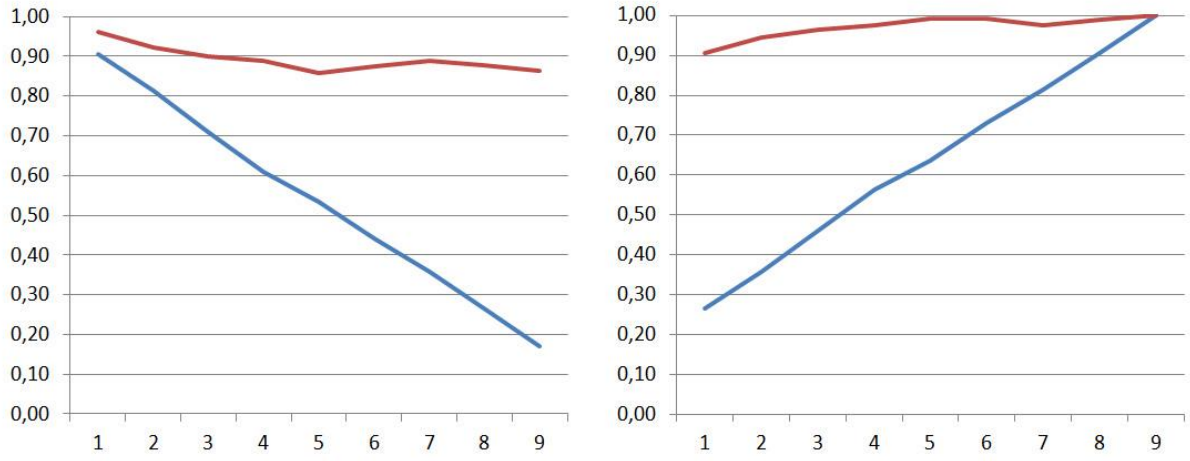
Simple Statistics-based Predictions

In order to evaluate and compare the quality and characteristics of the LPV, we applied a simple statistics-based idea to performance prediction, based on a retrospective view on a particular student's performance. If a student exhibits a certain performance at a certain point in time, for example the poor (blue) student in the following chart reaches a relative solution frequency of 0,05405 at the end of interval 4. One assumption that is inherent to many prediction methods is that a student might perform normally in the future. The grey diagonal that is shifted to the right indicates this idea in the following figure. This, however, is a significant overestimation of a student's abilities. There is a strong discrepancy between the final results of a student and such estimations (red line in the figure).



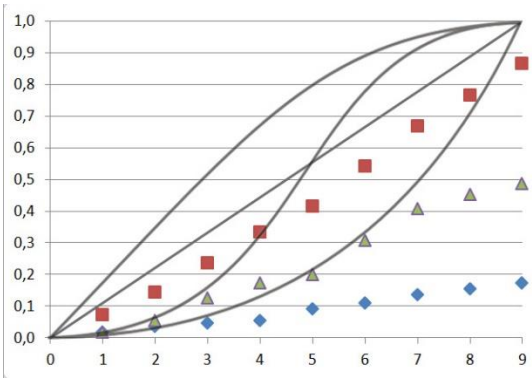
An over-simple predication approach.

The following figure shows the predictive power of this approach over time. The left panel shows the predicted end values over time for the good (red) and the poor (blue) student. The right panel shows the accuracy (difference of simulated end values and predictions) of the approach. It is evident that the method overestimates the achievements by far, even for a nearly optimally performing student. This optimal and average linear increase is a problematic approach, obviously.



The left panel shows the predicted performance over time, the right panel the method's accuracy. The red line displays an optimal student, the blue line a poor student.

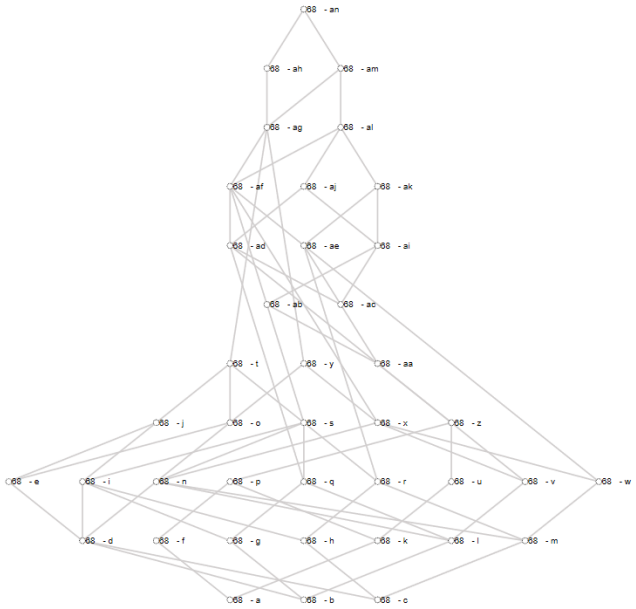
Existing approaches claim other prediction functions, e.g., various non-linear functions (as illustrated in the following figure). But also these are strong assumptions because they must find substrate in existing data. Thus, predication methods add the information of a large number of prior students to the predication model. This, however, includes again the assumption that the model is valid for all types of students (see Kim et al., 2014 for a case study).



Various non-linear predication models.

CbKST-based Prediction: The LPV

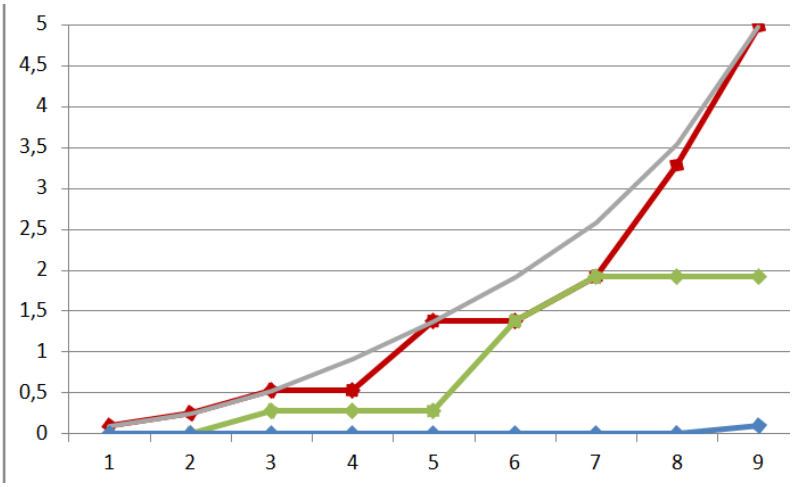
It doesn't come as a surprise that the more information are included in a prediction model, the better and more accurate it will perform. The contribution of CbKST and the Lea's Box project is to use the formal, combinatorics framework of CbKST to add information about the nature of a learning domain to the model. This information includes the number and complexity of competencies as well as the relationships between them. As opposed to one-dimensional models (see the last figure), the concept of multi-dimensional competence spaces allows for a multitude of individual learning paths and learning trajectories. The following image shows the competence space of the CbKST model introduced initially.



Competence space for the learning domain.

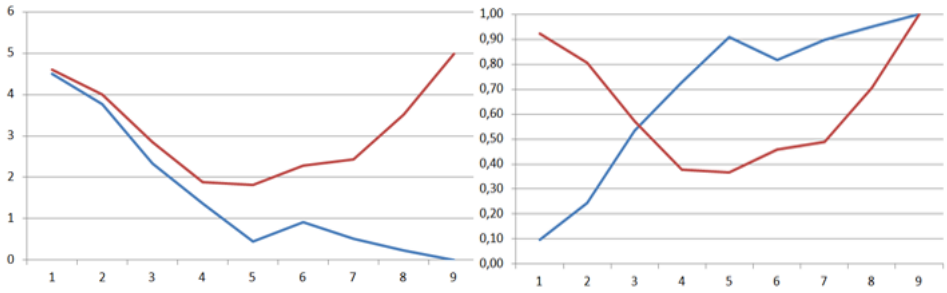
The prediction logic of the LPV is to assume a finite number of learning paths leading from the trivial competence state of having none of the competencies (the empty set) to the trivial state of having all

competencies (the full set). We assume a well-graded space, claiming that in each step in the learning paths only one competency is acquired. The set of learning paths a learner is on can be identified on the basis of the current and past answer patterns (i.e., which item types having been mastered and which not). The various paths can be characterized by their complexity, which is determined by the weight of the individual steps. Thus, the performance of a learner at a pint t is characterized by the weights this student mastered up to the current point in time. This means that the order of learning specific competencies and also the order of the assessment do not distort the prediction. In other words, mastering a lot of low complexity items at an early stage is a weak indicator because major challenges are still ahead for the student. In turn, mastering highly complex items (with high weights and perhaps a larger number of prerequisites) is a very strong indicator because all prerequisite items are assumed to be possessed by the learner. The following figure shows the prediction results for the same simulated data set and the same students.



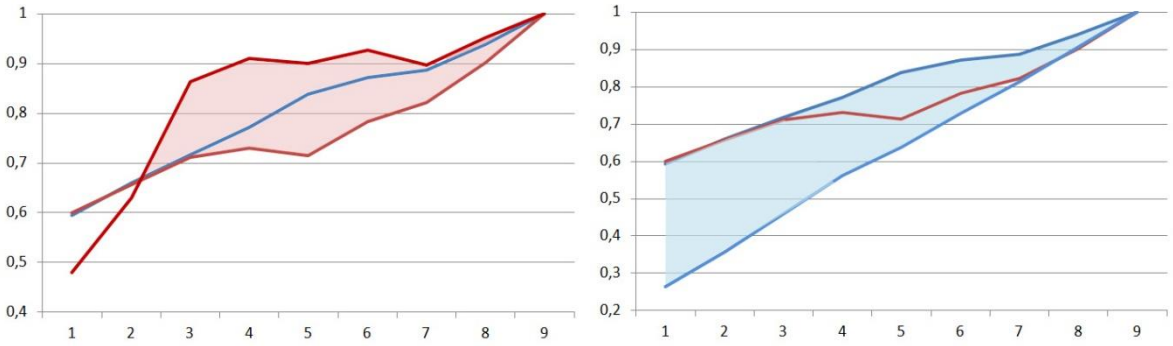
Predication results of the CbKST approach for a good (red), medium (green) and poor (blue) student.

In this example, the sum of weights assigned to the competence structure is 4.98 (the grey curve in the figure indicates the average prediction model of this approach, contrasting the linear approach we described above). The curves show the progress - in terms of mastered weights – of the three prototypical students across the 9 time intervals. The following figure shows the predicted end values at each point in time. We see that at first the predications are too high for the poor student, however, they rapidly decrease until the 5th point in time. Form here, the algorithm quickly approaches the final simulated values. The right panel of this figure shows the prediction accuracy of the final value over the 9 points in time. The result shows that the prediction with the CbKST approach is weaker for an optimally performing student as opposed to a weakly performing student. In other words, the linear approach loses predictive power with an increasing deviation from the ideal linear, diagonal increase.



The left panel shows the predicted performance over time, the right panel the method's accuracy. The red line displays an optimal student, the blue line a poor student.

When simulating different answer patterns in the 12 item types, we found that the order of item types strongly influences the power of the LPV approach. If more difficult items are presented already at an early stage, which means that the additional structural information influence the predictions, highly accurate predictions occur fast, in our example already after time interval 3 (left panel of next figure). In general, the LPV offers effective and accurate predictions. The linear approach, as said, is strongly dependent on the deviation of learning patters from the diagonal. For weak learners, for example, the predication is quit inaccurate (right panel of next figure).



Comparison of predictive power (LPV vs. linear)

Discussion

The aim of this simulation study was to find systematic evidence whether and to what extent the LPV is a suitable method to predict students' performance. Also, simulation studies allow us to explore the characteristics and dependencies of the method in its application to various data characteristics.

By the benchmark of the linear method we found stable and promising results. Specifically since the linear increase was the conceptual basis of the simulation algorithm as well.

Another critical aspect is the weighting process for competencies and subsequently assessment items. This has a strong influence on the predictions and can be based on several approaches. The simplest

is a manual assignment of weights by teachers. This, however, bears the peril of an arbitrary and unfounded weighting. The strength of this approach could be that the weights a grounded on the very concrete and practical experiences of a teacher. A second and more data driven approach is to refer to the solution frequencies of items in large data sets. This is the method we used in this study. If items are solved with a high frequency, we can assume a low difficult of the competencies covered by the item and also a low predictive power in terms of CbKST-type prerequisites between the competencies. A third method we must explore in future steps is the so-called Component Attribute Approach (Albert & Held, 1999). This theoretical approach describes test items (problems) by components and their attributes. Components are major characteristics, for example which algebraic operations are included in a math item. The attributes describe the individual components, for example stating which types of numbers (positive integers, decimals, …) are part of the problem. It could be shown that the prerequisite relation of the problem space could be derived by the set inclusion principle. In our context decomposing and analysing the components and their attributes can support the weighting process. Finally a forth method is to analyse the items on the basis of their cognitive depth. This refers back to the famous taxonomy of Benjamin Bloom, revised by Anderson & Krathwohl (cf. Anderson, 2013). In the so-called Concept – Action Verb approach (Heller, Steiner, Hockemeyer, & Albert, 2006), a competency is defined by a concept, ideally in form of a proposition (e.g., “house – has – window”) and a specific cognitive depth that ranges from mere “knowledge” to the level of “creation”. Bloom proposed 6 such levels. An example would be “understand that a house has windows and apply this understanding in a new situation”. The taxonomy also separates the knowledge dimensions factual, conceptual, procedural, and metacognitive knowledge, which in the end established a 2-dimensional hierarchy. In our context this taxonomy provides a scaffolding to analyse the items, to identify the covered competencies, and to rank the competencies according the taxonomy – which I the end specifies the weight.

Future steps will continue to explore the characteristics of the LPV under various context conditions. Also more in-depth studies using the real data sets from the “Assistments Math 2004-2005” data set.

Main Conclusions

- The LPV is a suitable method to predict learning performance.
- The LPV has particular strength for rather medium and poor students and it works best when at each assessment point in time a broad spectrum of item difficulties are presented.

References

Hughes, G., & Dobbins, C. (2015). The utilization of data analysis techniques in predicting student performance in massive open online courses (MOOCs). Research and Practice in Technology Enhanced Learning, 2015 10:10.

Papamitsiou, Z., & Economides, A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. Educational Technology & Society, 17 (4), 49–64.

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science, 72, 414-422.