Modeling learning behaviour and cognitive bias from web logs

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Rashmi Jayathirtha Rao,

Graduate Program in Department of Computer Science

The Ohio State University

2017

Dissertation Committee:

Dr. Christopher Stewart, Advisor

Dr. Arnulfo Perez

Dr. Alan Ritter

Abstract

Students who can link algebraic functions to their corresponding graphs perform well in STEM courses. Increasingly, early algebra curricula teaches these concepts in tandem. However, it is challenging to assess whether students are linking the concepts. Tests, video analyses, interviews and other traditional methods that aim to quantify how students link the concepts taught in school require precious classroom and teacher time. In this paper, we use web logs collected by a smart classroom web server to infer learning. Compared to traditional methods, web logs are widely available, voluminous and amenable to data science. However, web logs are constrained by factors such as data collection practices by system managers and the design of the web interface. Our approach partitions the web interface into components related to data and graph concepts. We collect click and mouse movement data as users interact with these components. We used statistical and data mining techniques like k means clustering, principal component analysis and frequent sequence patterns to model their learning behavior. We compared our models with traditional methods to assess learning behavior for a workshop presented in Summer 2016. Students in the workshop were middle-school math teachers planning to use a new early algebra curriculum in their own classrooms. First, we used our models to assess participation levels, a prerequisite indicator for learning. Our models aligned with ground-truth traditional methods for 17 of 18 students. The results from machine learning models

which do not consider the computational thinking aspect of the web components aligned with ground-truth methods in only 10 of 18 students. Unlike traditional methods, our models were computed in minutes and not days. The results of the models w.r.t the two types of components of the web portal have been used to infer possible data or graph oriented cognitive bias in the students.

Acknowledgments

I would like to thank all my collaborators, without whom this work would not have been possible.

- Dr. Christopher Stewart
- Dr. Arnulfo Perez
- Bailey
- Tony
- Siva Meenakshi Renganathan.

Table of Contents

Page

List of Figures

Chapter 1: Introduction

Computational thinking has been defined as the thought processes involved in formulating problems and their solutions so that the solutions are represented in a form that can effectively be carried out by an information-processing agent. [10]. The Dept. of Teaching and Learning is conducting research on computational thinking in the teachers and students. In this, the subjects of research are introduced to science experiments. The experiments involve computational aspects of linear algebra in mathematics. The users enter and interpret the observations from these experiments on a web application developed on top of the Moodle platform. Moodle is used to manage the user login and provide access to the appropriate view of the web application which we address as the web portal in this paper. Once the input data is entered, the results of the experiments can be observed as a graphical representation. For example, one of the experiments is the Ohm's law. According to Ohm's law, voltage is a linear function of current with resistance as the constant of proportionality. In this experiment, the participants connect a resistor to a voltage source and measure the amount of voltage passed as well as the current flowing through the resistor. The experiment is repeated for different voltage inputs and resistors. The participants then create tables in the 'web portal' and enter the input voltage and observed current values. The portal has an ability to display this data in the form of a graph. The

graph in this case indicates a line created by the linear regression of the data points entered by the user. The slope represents the resistance value. The users perform several other functions on the portal like share their tables, import the tables shared by group mates, move sliders to change the angle of the regression line and even update the length of the axes on the graph.

The subjects of research are video graphed while performing the experiments and this helps to visually observe each person's computational thinking behaviour. Their computational thinking is also tested by means of surveys and quiz. This accounts for the traditional way of inferring participation levels and computational thinking. Research indicates that higher usage of a web portal related to projects correlated to higher motivation levels towards the project and correspondingly higher grades. [5] Based on this analogy, we tried to use the data corresponding to the web portal usage to understand the participation levels of the users in the experimental activities. Block diagram of the assessment methodology can be found in Figure 1.1

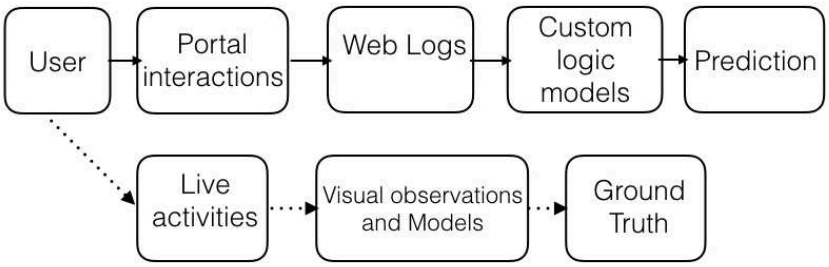

Figure 1.1: Block Diagram

Further, by the analysis of the usage of the two types of components on the portal, we try to identify any preference towards one type of component. In this case, we identify if the person has data or graph oriented cognitive bias. As the analysis is mainly based on the mouse movements and clicks, we assume that the mouse movements provides information about the eye movement. This is supported by the research [9] which indicates that mouse movement has a high correlation with the eye movement. The current research [3] is fast growing and will soon be used by more students across different districts of Columbus,Ohio. It'd be highly informative to the teacher if he/she could be informed about the less participating students in the classroom in real time. Hence there's a need to explore a solution for processing the data quickly and also provided an approximate result before the collection of the entire data. Hence we perform different types and levels of sampling on the dataset to get the quality of the result in each case.

Chapter 2: Learning Analytics

The analysis on educational and academic data falls into the category of Data Science called Learning Analytics(LA). Learning Analytics is defined as the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs [4]. LA process is often an iterative cycle and is generally carried out in three major steps as shown in Figure 2.1:

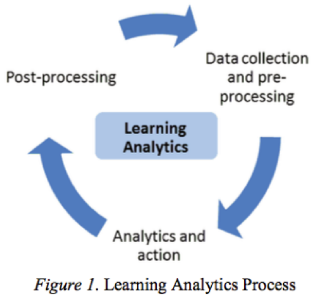

Figure 1. Learning Analytics Process

Figure 2.1: Learning analytics process

- Data collection and pre-processing: Every action performed by the user is collected as a css path by means of the java script and is recorded in the database along with the timestamps as shown in Figure 2.2. Each component in the web

portal is assigned a unique id which helps to identify the component in the css path. Primarily, we use the data collected from the portal usage by a group of 18 students who participated in a 2 week long workshop conducted during Summer 2016. The students in this workshop were math instructors intending to use this curriculum in their respective classrooms. The portal was used for 4 days and around 1-2 hours per day. The total data collected was 15Mb. At a later stage we use the data collected from the student interaction with the portal. The data corresponding to 365 students belonging to 15 class rooms has been collected with a total size of 90Mb till date. This data is then processed to extract features indicating the user behaviour on the portal.

- Analytics and action: We analyse the learning behaviour by modelling the data to predict the participation level and cognitive bias. We also perform clustering to identify students with similar behaviour. A group of five researchers in the Dept. of Teaching and Learning at The Ohio State University have visually observed the video recordings collected during the workshop and assessed the computational thinking levels of the students. Each of the 18 students have been classified as either "high" or "less" participating with the participation levels indicating a computational thinking oriented behaviour. We use this as the ground truth for the classification on our dataset.

- Post-processing : In general post-processing can involve compiling new data from additional data sources, refining the data set, determining new attributes required for the new iteration, identifying new indicators/metrics, modifying the variables of analysis, or choosing a new analytics method. Based on the

performance of the models, we try to improve the prediction by visualizing the data w.r.t the classification and develop better analytic models. The best model is used to infer the learning behaviour in the students in classrooms.

userid	timestamp	estimatedtime	sessionid	event	time	pagex	pagey	clientx	clienty	element	csspath
0	2016-10-03 12:34:36	*NULL*	4mPUTWitMU	mousemove	341797	298	251	298	251	DIV+id:	div.container-fluic
0	2016-10-03 12:34:36	*NULL*	4mPUTWitMU	mousemove	345578	105	384	105	384	SPAN+id:	div.container-fluic > div.container > div.row > di...
0	2016-10-03 12:34:36	*NULL*	4mPUTWitMU	mouseclick	346923	170	358	170	358	INPUT+id:tblAppendGrid_x_1	div.container-fluic > div.container > div.row > di...
0	2016-10-03 12:34:36	*NULL*	4mPUTWitMU	mousemove	347144	170	358	170	358	INPUT+id:tblAppendGrid_x_1	div.container-fluic > div.container > div.row > di...
0	2016-10-03 12:34:36	*NULL*	4mPUTWitMU	mousemove	351879	170	358	170	358	INPUT+id:tblAppendGrid_x_1	div.container-fluic > div.container > div.row > di...

Figure 2.2: Web Logs

Before starting the Learning Analytics process,the challenge is to define the right Objective / Indicator / Metric (OIM) triple [7]. In our case they are:

- Objective is to understand computational thinking/learning behaviour in students in terms of linking concepts of graph and data.
- Indicator is the actions performed by the users on the web components
- Metric is the results from traditional analysis methods

Chapter 3: Related work

The advantages of web based learning is that it not only provides visual interactivity with the resources that encourages students to actively participate in the learning process (Lie and Cano, 2001; Sims, 2000) but also facilitates accessing information about the sophisticated interactive resources. The idea of using web log data to assess the computational thinking is supported by a similar work by Werner and colleagues [6] as a way to measure computational thinking skills in middle school students. They refer to the assessment as the Fairy Assessment. They use a gaming application which focuses on three tasks designed to test comprehension, design, and complex problem solving. They collect the actions performed by the students as log files. Each student is then assessed, hand-graded by two experimenters along a 24 point rubric and identified as high or low performing student. The authors found that assessment scores correlated with students interest in taking a computer science class, confidence with computers and attitude toward computing. Nicholas Diana and team [1] have identified the key features of this dataset namely the human graded rubric scores for each student and the collection of web log data to predict the high and low performing students.

Approximate computing:

In software-driven approximation bound by quality limits, speeding up the processing of data and reducing delay has been explored in several works such as:

- ApproxHadoop [11]: In this approach, the processing delay is reduced by dropping map tasks to speed up map-reduce computations, lowering the quality of final answers

- Ubora [2]: This work proposes an approach to measure the effect of slow running components on the quality of answers, memorize and improve the computation.

- Sprinting [8]: This approach explores increasing processing rates by exceeding budgets for short bursts before reverting back to safe processing rates.

Chapter 4: The Models

4.1 Feature Engineering

Is the process of using domain knowledge of the data to create features.(Wikipedia) The raw web log data has the information about set of actions performed by the users at different instants of time. But we need cumulative features to represent these actions w.r.t each student. Research suggests that key indicators of the learning behaviour in web based learning are [5]:

- Web access rate: How frequently the web page resources were accessed.
- Web Resources: Which resources on the web were accessed
- Time spent with the resources
- Usage pattern of the resources

The main resources on our portal are mainly components used to enter, assemble data and the components used to visualize the data graphically. Hence we've divided the each of the web components into either data or graph component as shown in Figure 4.1

Features extracted for each user:

Label	Component Name	Associated Behaviour
D1	My Tables - Table Name	Data
D2	My Tables - Share	Data
D3	My Tables - Delete	Data
D4	Tables Shared with me- Table name	Data
D5	Tables Shared with me- Refresh	Data
D6	Input Table	Data
D7	Save Table	Data
G1	Graph Button	Graphical
G2	Graph - SVG	Graphical
G3	Slider- Primary Table	Graphical
G4	Update Axes	Graphical
G5	Slider- Secondary Table	Graphical

Figure 4.1

- Inter data interactions - Total number of interactions from graph to the data components
- Inter graph interactions - Total number of interactions from data to the graph components
- Intra data interactions - Total number of interactions within data components
- Intra graph interactions -Total number of interactions within the graph components
- Intra data time spent - Total time spent within the data components
- Intra graph time spent - Total time spent within the graph components
- Total number of interactions on the page

- Total time spent on the page
- Data patterns - The total number of data components in frequently accessed patterns
- Graph patterns - The total number of data components in frequently accessed patterns
- Participation level-used for prediction

4.2 Data pre-processing

We tried to understand the relation between the different features by plotting their correlation.

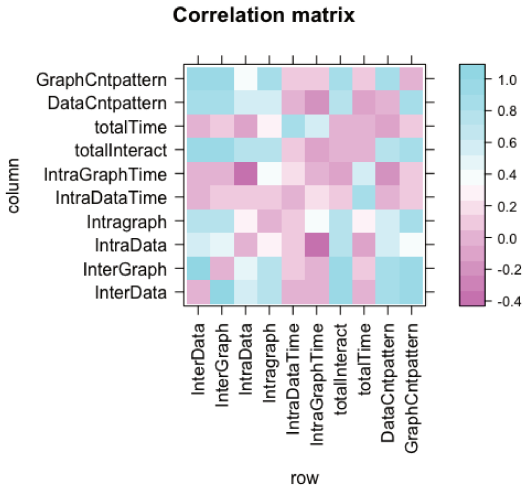

Figure 4.2

There does exist a high correlation between certain features like Inter Data, Inter Graph, Data patterns, total number of interactions. We normalized the data by converting it to zero mean and unit variance.

4.3 Basic Machine Learning models

We applied machine learning models like decision tree and random forest to predict the participation level. We considered 17 out of the 18 data points as the training set and the use the built model to predict the participation level of the 18th student. This was repeated for each of the students and the results are as shown in Figure 4.3 The average accuracy in both the models was around 55%. Hence we could observe

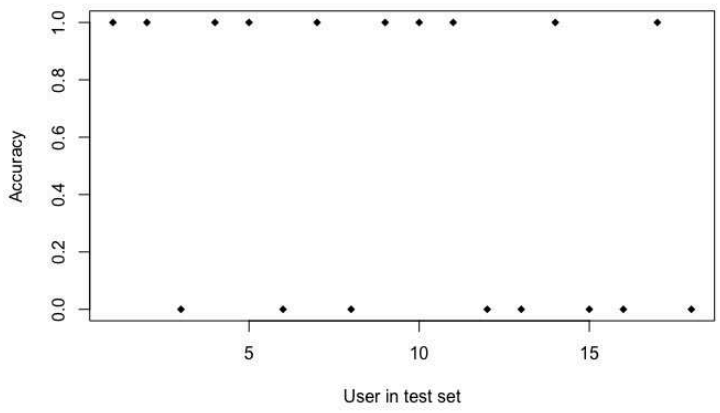

Figure 4.3: Results for Decision Tree and Random Forest

that the accuracy is low. But the key objective of the analytics to detect any cognitive bias is hard to infer from such models. This is mainly because the features providing information regarding the graph and data have not been considered individually.

4.4 Visual data observation

We considered each of the data and graph related features to observe the distribution of the data w.r.t the classification. The results for Inter Graph vs Inter Data, Intra graph vs Intra data and Intra graph time vs Intra data time has been shown in Figure 4.4, Figure 4.5 and Figure 4.6 respectively.

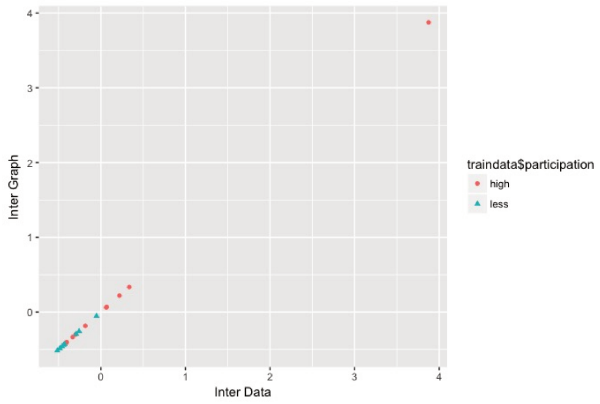

Figure 4.4: Inter Graph vs Inter Data

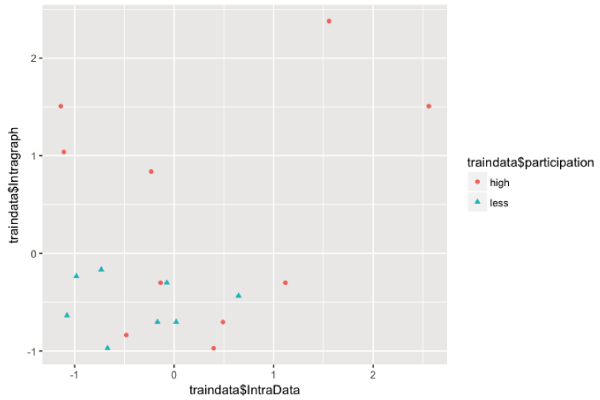

Figure 4.5: Intra graph vs Intra data

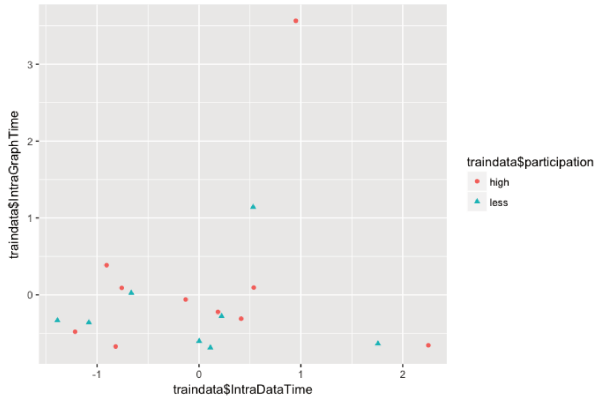

Figure 4.6: Intra graph time vs Intra data time

We can observe that they are not linearly separable. Combination of 2 features and 3 features together can be visualized as shown in the Figure 4.7 and Figure 4.8

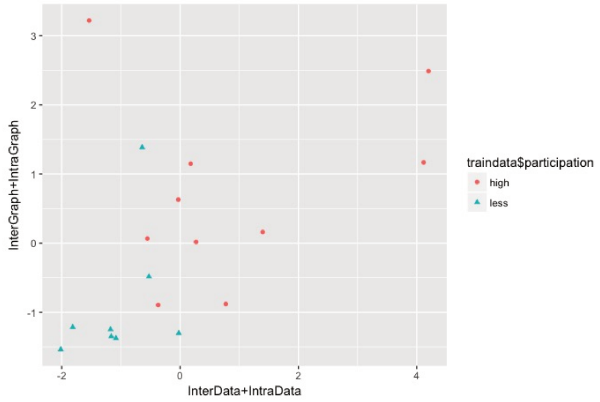

Figure 4.7: Intra Graph+ Intra Graph vs Inter Data+ Intra Data

We can observed that the data is more linearly separable.

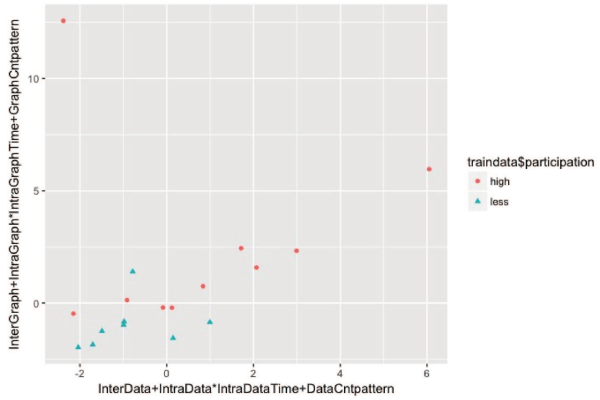

Figure 4.8: Intra Graph+ Intra Graph*Intra Time+Graph Patterns vs Inter Data+ Intra Data*Intra Time + Data Patterns

4.5 Custom models

We came up with a model to combine the data and graph related features. We define a metric called score to quantify the data and graph orientation.

$$(GS, DS) = \left(\hat{G}, \hat{D}\right) \qquad D_1 = \sum_{d_i \leftarrow g_i} weight(g_i, d_i)$$
$$\hat{G} = f\left(G_1, G_2 ...\right) \qquad G_1 = \sum_{g_i \leftarrow d_i} weight(d_i, g_i)$$
$$\hat{D} = f\left(D_1, D_2 ...\right)$$

$$D_2 = \frac{\sum_{j \in DataCluster} I_j T_j}{\sum_i T_i}. \qquad D_3 = \frac{\sum_{seq \in freqSequences} Nd}{}.$$
$$G_2 = \frac{\sum_{j \in GraphCluster} I_j T_j}{\sum_i T_i}. \qquad G_3 = \frac{\sum_{seq \in freqSequences} Ng}{}.$$

GS and DS are Graph and Data Scores respectively and G1,G2.etc are the graph scores obtained from model 1, 2 and so on. Similarly, D1, D2.etc are the data scores obtained from different models.

- Model 1: Considering each of the web components on the portal as a node in the graph with edges representing the pattern of access, weight of the edge from node i to another node j is the total number times node j was accessed after node i. These sub scores indicates the orientation and relative usage of both graph and data components by the participants.

- Model 2: These sub scores considers the participants interaction within the same components as well as the time spent on each component.

- Model 3: Weighs more on the which component was more accessed in a set of patterns.

Each of these models provides a sub score for data as well as graph orientation. These sub scores need to be combined to a single score. This dimensionality reduction can be done by performing principle component analysis and considering the primary component that captures the maximum variance in the data. The graph and data score after performing PCA(Model1, Model2 , Model3) is as shwon in Figure 4.9 A simple linear model of graphscore+data score>=-0.8 is considered as "high" and graphscore+data score<-0.8 is considered as "less" participation level.

By this modelling we're able to classify 17 out of 18 users classified accurately.

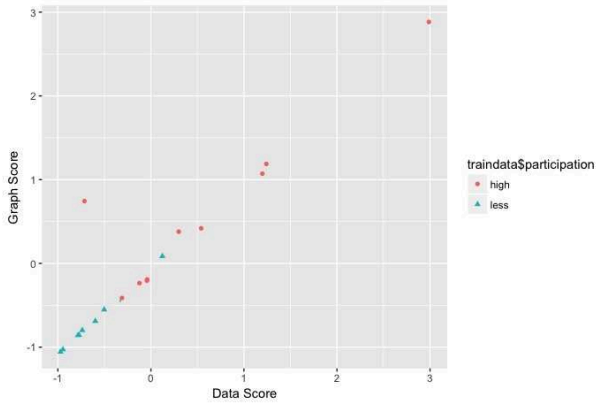

Figure 4.9: Graph Score vs Data Score

4.6 Cognitive Bias

Research [9] indicates there's more than 66% correlation between eye movement and mouse movement in linear web page layouts. Hence we hypothesise that a possible cognitive bias towards either graph or data can be indicated by their respective features. We infer cognitive bias by observing the high value of graph or data score that has contributed to the user being classified as "high" or "less". From the Figure 4.9 we can observe that the point on the top left has been classified as high. But its due to high graph score in spite of low data score. We define a cut off for data and graph scores as -0.5. If a student is classified as high but has a data or graph score less than -0.5, its possible that he/she has a cognitive bias towards the other type of orientation. This bias could indicate higher interest or a struggle to understand the graphical representation.

4.7 Outlier Detection

Its a hard task to group people of similar behaviour and detect any outliers. As shown in Figure 4.9, the student on the top right is an outlier. To perform the outlier detection, we use the un-supervised method of K means clustering. To get an idea of number of clusters we perform a hierarchical clustering and infer that K=5 can be used to determine the outliers as shown in Figure 4.10. Student number 4 is the outlier.

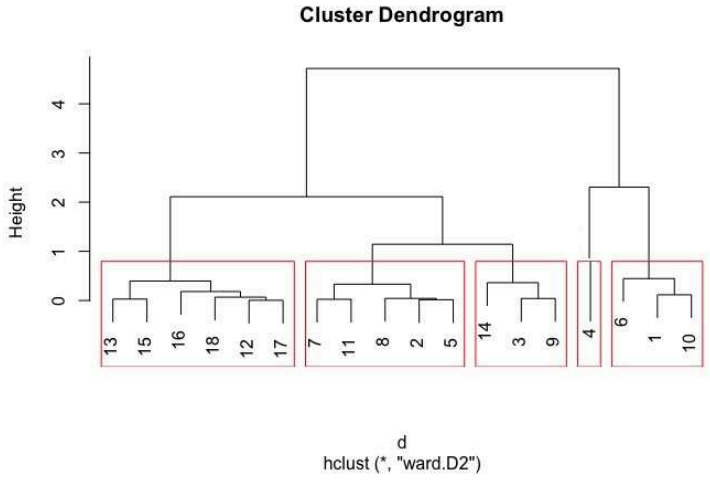

Figure 4.10: Dendogram

Chapter 5: Classroom Data

We applied the same model to the students data. We don't have the ground truth about the classification yet, hence we hypothesise that our model predicts the participation level in students with 95% accuracy. This analysis is highly applicable and useful in the research by the Dept.of Teaching and Learning to identify and understand those classrooms which have many students with high computational thinking and similarly those with very less. Also the it could help to identify students with detected cognitive bias and understand their actual learning behaviour in the classrooms. Some level results for the classroom data is shown in Figure high_percentage indicates the fraction of the students classified in that classroom as "high".low_percentage indicates the fraction of the students classified in that classroom as "high". num_highs indicates the actual number of the students classified in that classroom as "high". num_lows indicates the actual number of the students classified in that classroom as "less".

classid	high_percentage	low_percentage	num_highs	num_lows
10	0.25806451612903225	0.7419354838709677	8	23
8	0.17391304347826086	0.8260869565217391	4	19
26	0.8620689655172413	0.13793103448275862	25	4
7	0.29411764705882354	0.7058823529411765	5	12
28	0.7419354838709677	0.25806451612903225	23	8
13	0.7037037037037037	0.2962962962962963	19	8
12	0.6666666666666666	0.3333333333333333	18	9
24	0.08333333333333333	0.9166666666666666	2	22
27	0.6176470588235294	0.38235294117647056	21	13
11	0.32142857142857145	0.6785714285714286	9	19
23	0.18181818181818182	0.8181818181818182	2	9
15	0.5714285714285714	0.42857142857142855	16	12
6	0.3125	0.6875	5	11
9	0.2903225806451613	0.7096774193548387	9	22
25	0.7727272727272727	0.22727272727272727	17	5

Figure 5.1: Results of student data

Chapter 6: Approximate Computing

The classroom data set we considered was for 15 classes and 365 students accounting for a total of 90Mb of data. The application will be expected to be introduced in more than 100 hundred class rooms over the coming years. Also it'd be highly effective for a teacher to assist the students who're struggling if the lack of participation is detected in the classroom itself. Hence there's a need to explore the computation aspect and efficient ways to reduce the time taken to get the results without compromising much on the quality. The lesser the data, lesser is the time for computation. Hence we explore the effects of sampling the data on the quality of the results. We performed sampling in different manners and calculated the error percentage for each type and amount of sampling. We performed sampling for 10,20,30...90% of data and ran the models for 50 trials for each amount of sampling. Error is calculated as:(Number of mis-predictions with sampled data when compared with predictions from complete data)/total number of students. The results for different ways of sampling is as below:

- Sampling without preserving the order: Performed random sampling with shuffling the data. This mainly destroys the time stamp integrity i.e, the records for a user that follows a increasing time stamp is only considered in data preprocessing. The results are as shown in Figure 6.1 and Figure 6.2 Intuitively,the

error percentage decreases with the increase in the sampling percentage. The maximum and minimum errors observed are around 24 and 13% respectively.

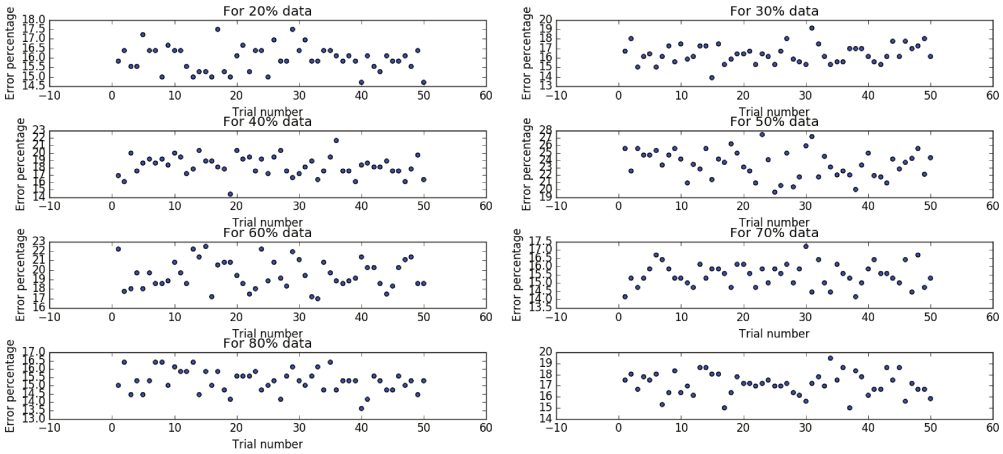

Figure 6.1: Unordered Sampling

- Sampling with preserving the order: Performed random sampling without destroying the sequential order of the data. In this way the time stamp integrity is preserved. The results are as shown in Figure 6.3 and Figure 6.4 Intuitively,the error percentage decreases with the increase in the sampling percentage. The maximum and minimum errors observed are around 26 and 18% respectively.

- Stratified sampling with order: We considered only the 1st 10,20,30..90% of the data and applied our models. The error percentage decreases with the increase in the number of samples considered. The results are as shown in Figure 6.5

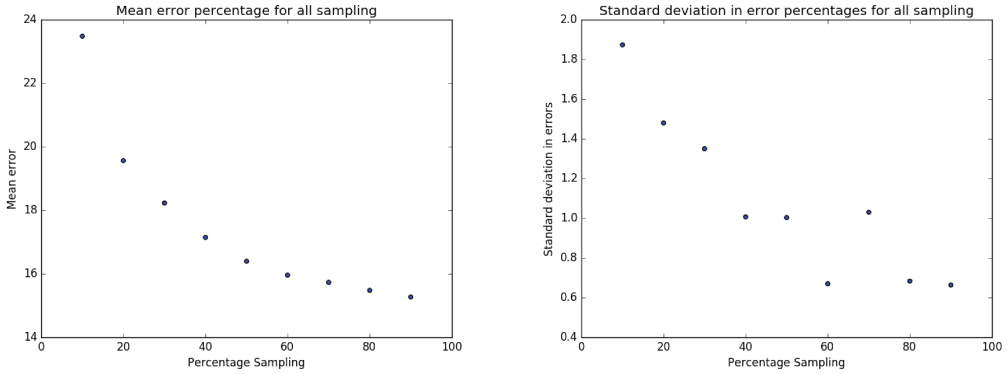

Figure 6.2

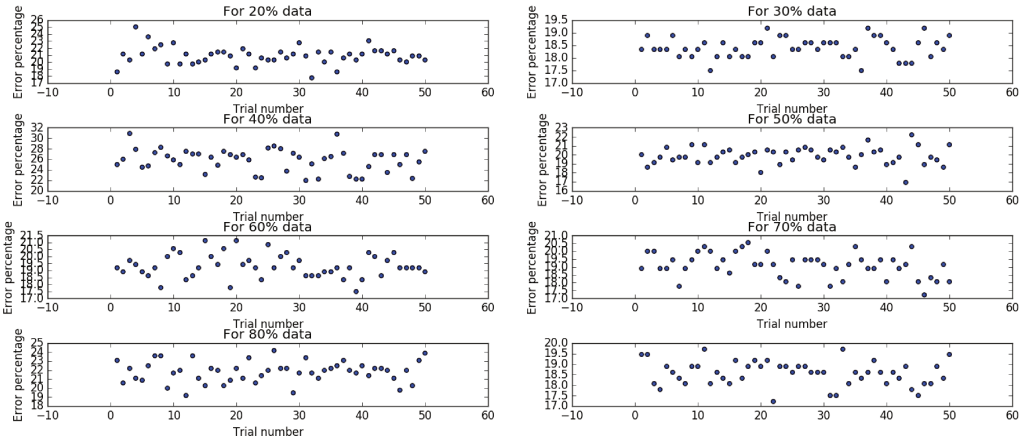

Figure 6.3: Unordered Sampling

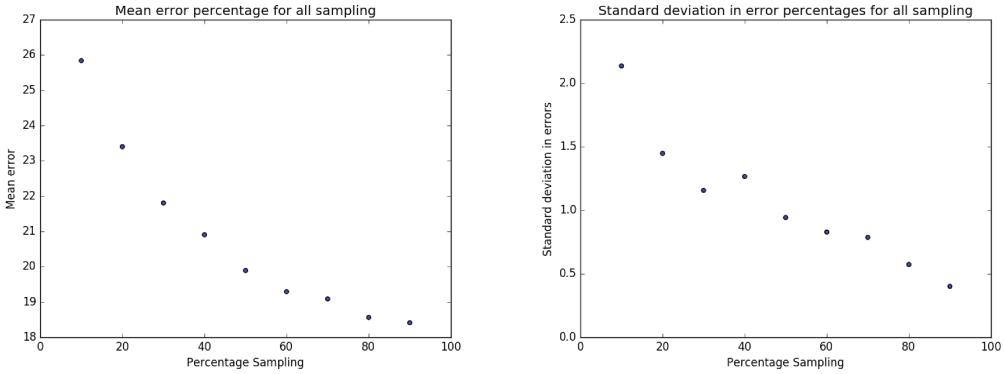

Figure 6.4

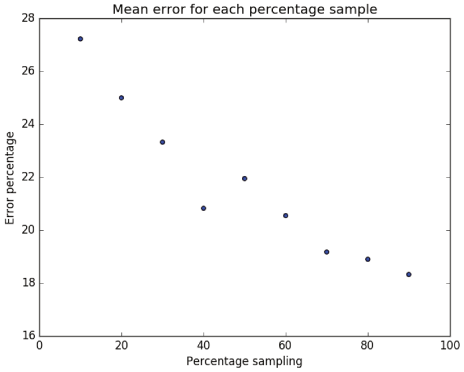

Figure 6.5

Chapter 7: Conclusion and future work

7.1 Conclusion

We were able to devised a model which accurately predicted the relative computational thinking levels of 17 out of the 18 students in the workshop using the data collected from the e-learning portal. From this, we were able to identify a possible cognitive bias towards either graph or data oriented learning. We performed outlier detection by unsupervised learning methods to group students of similar behaviour together and detect anomaly in the behaviours. We further applied the models on the student data to gain insights in the classroom dataset. We performed sampling on the data to observe the accuracy levels for different amounts and types of sampling.

7.2 Future work

Better models could be developed to achieve a 100% accuracy. There's a scope to extract more features that might provide further insights in the computational thinking behaviour of the teachers and students. A possible way to cross verify the prediction of participation levels on student data is by performing NLP and assigning a metric for participation level based on on the field notes collected by the researchers in every class room.

Bibliography

[1] Nicholas Diana. Michael Eagle. John Stamper. Shuchi Grover. Marie Bienkowski. Satabdi Basu. "An Instructor Dashboard for Real-Time Analytics in Interactive Programming Assignments". *Learning Analytics and Knowledge*, 2017.

[2] Jaimie Kelley. Christopher Stewart. Nathaniel Morris. Devesh Tiwari. Yuxiong He. Sameh Elnikety. "Measuring and Managing Answer Quality for Online Data-Intensive Services". *ICAC*, 2015.

[3] Arnulfo Perez. Kathy Malone. Siva Meenakshi Renganathan. Kimberly Groshong. "Computer Modeling and Programming in Algebra". *8th International Conference on Computer Supported Education*, 2016.

[4] gsiemens. Learning analytics and knowledge. *1st International Conference on Learning Analytics and Knowledge*, 2011.

[5] JUDY SHEARD. JASON CEDDIA. JOHN HURST. *"Inferring Student Learning Behaviour from Website Interactions: A Usage Analysis"*. Kluwer Academic Publishers, 2003.

[6] ShannonCampe LindaWerner, JillDenner. "The Fairy Performance Assessment: Measuring Computational Thinking in Middle School". *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education*, 2012.

[7] U. Schroeder H. Ths M.A. Chatti, A.L. Dyckhoff. "A Reference Model for Learning Analytics". *International Journal of Technology Enhanced Learning*, 2016.

[8] Nathaniel Morris. Christopher Stewart. Siva Meenakshi Renganathan. "Sprint Ability: How Well Does Your Software Exploit Bursts in Processing Capacity?". *International Conference on Autonomic Computing XIII*, 2016.

[9] Vidhya Navalpakkam. LaDawn Jentzsch. SRory Sayres. Sujith Ravi. Amr Ahmed. Alex Smola. "Measurement and Modeling of Eye-mouse Behavior in the Presence of Nonlinear Page Layouts". *ACM 978-1-4503-2035-1/13/05.*, 2013.

[10] Jeannette M. Wing. Computational thinking: What and why? *CACM*, 2010.

[11] Thu D. Nguyen In igo Goiri.Ricardo Bianchinil, Santosh Nagarakattee. "ApproxHadoop: Bringing Approximations to MapReduce Frameworks". *ASPLOS*, 2015.