ELBA: Exceptional Learning Behavior Analysis

Xin Du
Eindhoven University of Technology
x.du@tue.nl

Wouter Duivesteijn
Eindhoven University of Technology
w.duivesteijn@tue.nl

Martijn Klabbers
Eindhoven University of Technology
m.d.klabbers@tue.nl

Mykola Pechenizkiy
Eindhoven University of Technology
m.pechenizkiy@tue.nl

ABSTRACT
Behavioral records collected through course assessments, peer assignments, and programming assignments in Massive Open Online Courses (MOOCs) provide multiple views about a student's study style. Study behavior is correlated with whether or not the student can get a certificate or drop out from a course. It is of predominant importance to identify the particular behavioral patterns and establish an accurate predictive model for the learning results, so that tutors can give well-focused assistance and guidance on specific students. However, the behavioral records of individuals are usually very sparse; behavioral records between individuals are inconsistent in time and skewed in contents. These remain big challenges for the state-of-the-art methods. In this paper, we engage the concept of **subgroup** as a trade-off to overcome the sparsity of individual behavioral records and inconsistency between individuals. We employ the framework of **Exceptional Model Mining (EMM)** to discover exceptional student behavior. Various model classes of EMM are applied on dropout rate analysis, correlation analysis between length of learning behavior sequence and course grades, and passing state prediction analysis. Qualitative and quantitative experimental results on real MOOCs datasets show that our method can discover significantly interesting learning behavioral patterns of students.

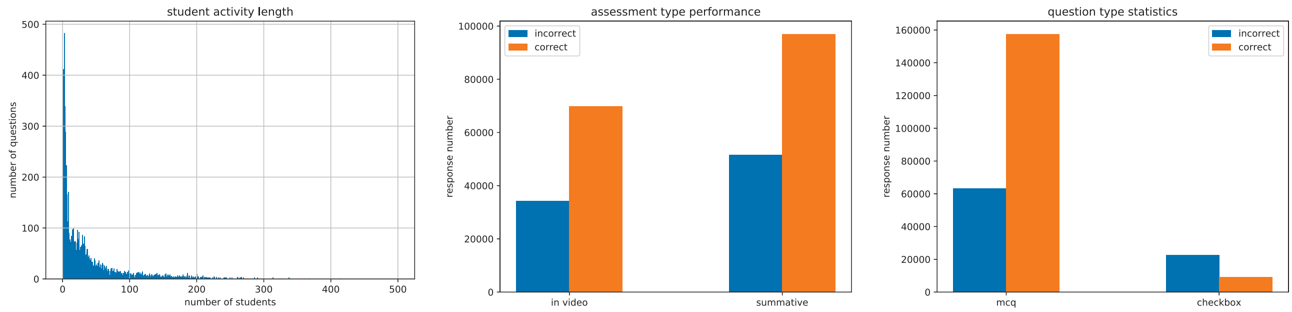
Keywords
Exceptional Model Mining, MOOCs, Learning Analytics

1. INTRODUCTION
Massive Open Online Courses (MOOCs) make it possible for educators to analyze learning behavior of students in multiple views. In contrast to traditional classes, which only have limited learning behavioral records, MOOC platforms such as Coursera, edX and Udacity provide huge amounts of learning behavioral records. These platforms collect very detailed course information and students' learning behavior such as course assessments, peer assignments, programming assignments, forum discussions and feedback [19], which can reflect the knowledge and skill achievements and the study performance of students. Modeling students' learning behavior and trying to discover interesting behavioral patterns are non-trivial. Most recent research is focused on how to predict the learning results based on the learning behavior model. It can help the tutors to design the courses and give specific guidance and assistance to specific students. However, due to the complexity of the behavioral records, there are still several challenges to be overcome:

Individual sparsity. Even when many students are enrolled in a course, the duration of their involvement varies substantially. Figure 1a displays a histogram of assessment question frequencies, which shows an obvious Power-Law distribution [2]. Only a few students participate in hundreds of assessment questions. Most of the students have activity length less than 20 records, which is very sparse. This makes evolutionary activity sequence based user modeling methods [16, 17] ineffective.

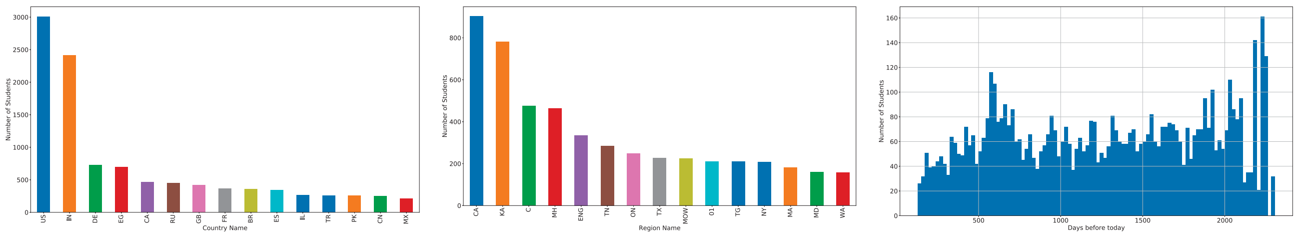
Activity inconsistency. Beyond the distribution in activity length of assessment questions, students' learning behavior in forum discussion, click stream and peer review are also shown to follow a Power-Law distribution. In Table 4, we can see that among the 18 courses on Coursera, enrolled students, grades and students who passed the course are highly diverse. This inconsistency makes the data very imbalanced, which results in difficulties for Matrix factorization based modeling methods [24]. These methods might merge infrequent behavior with common behavior.

Content heterogeneity. Behavior diversity is not only shown in activity length and course status, but also shown in informative contents. There are 7 types of assessments and 12 types of questions in the courses, such as video, summative, checkbox and multiple checkbox. Proportions of these assessments and questions are skewed in different courses. On the other hand, students also have varying participation records on these contents. In Figure 2, it is shown that distributions of students are obviously different in specific demographic categories. It is a big challenge for modeling methods to handle these heterogeneous contents for tasks like dropout prediction or passing state prediction.



(a) Histogram of the number of assessment questions in which students participate. (b) Performance per assessment type. (c) Performance per question type. 'mcq' represents multiple checkbox questions.

Figure 1: Heterogeneity and inconsistency of student behavior.



(a) Number of students by country. (b) Number of students by regions. (c) Number of students by join date.

Figure 2: Student distributions across various demographic categories.

To overcome these challenges, we propose to employ Exceptional Model Mining (EMM) [4] for exceptional learning behavior analysis. Instead of looking for anomalies or outliers of individuals, we look for exceptional behavior on the subgroup level [7], which can provide interpretable descriptions such as 'Students: Country = US, Region = Manhattan, Join dates > 365 (days)' having exceptional learning behaviors that are predominantly different from those in the whole dataset. We employ EMM to discover interesting learning behavioral patterns in subgroups. We establish various model classes for specific learning behaviors, such as discovering correlation between length of behavior sequence and course grades, finding out subgroups with exceptional dropout ratio, and looking for specific subsets where the classifier does not perform well. Experimental results on a real dataset illustrate the type of meaningful learning behavioral patterns EMM can discover in MOOCs. This can help us build an improved behavior model in the future research. In summary, our main contributions are:

1. We employ Exceptional Model Mining (EMM) to learning behavior analysis in MOOCs, which can help us to overcome the sparsity, inconsistency and heterogeneity in the behavioral records.
2. We employ several EMM model classes for different tasks to discover exceptional learning behaviors on the subgroup level. Our results show very interesting learning behavioral patterns, which can help the tutors conduct specific guidance and assistance to the students.

2. RELATED WORK

Local Pattern Mining (LPM) [6, 14] is a subfield of data mining, concerned with discovering subsets of the dataset at hand where something interesting is going on. Typically, a restriction is imposed on what kind of subsets we are interested in: only those subsets that can be formulated within a predefined *description language* are allowed. A canonical choice for this language is conjunctions of conditions on attributes of the dataset. Hence, if the records in our dataset concern people, then LPM finds results of the form:

$$\text{Age} \geq 45 \wedge \text{Smoker} = \text{yes} \quad \rightsquigarrow \quad \text{interesting}$$

This ensures that the results we find with an LPM method are relatively easy to interpret for a domain expert: the subsets will be expressed in terms of quantities with which the expert is familiar. We call a subset that can be expressed in such a way a *subgroup*.

Different LPM methods give a different answer to the question what exactly constitutes "where something interesting is going on". The most famous form of LPM is *Frequent Itemset Mining* (FIM) [1], where interestingness is equivalent to occurring unusually frequently: things that happen often are interesting. Hence, FIM finds results of the form:

$$\text{Age} \geq 45 \wedge \text{Smoker} = \text{yes} \quad \rightsquigarrow \quad \text{(high frequency)}$$

The methods we are mainly concerned with in this paper, however, seek a more complex concept on the right-hand side of this arrow. The task of *Subgroup Discovery* (SD) [9, 23, 7] typically singles out one binary attribute of the dataset as the *target*: subgroups are deemed interesting if this one target has an unusual distribution, as compared to its distribution on the entire dataset. In our example, if the target column describes whether the person develops lung cancer or not, SD finds results of the form:

$$\text{Smoker} = \text{yes} \rightsquigarrow \text{lung cancer} = \text{yes}$$
$$\text{Age} \leq 25 \rightsquigarrow \text{lung cancer} = \text{no}$$

These subgroups make intuitive sense in terms of our knowledge of the domain. Smokers have a higher-than-usual incidence of lung cancer. People under the age of 25 often have not yet had the chance to develop lung cancer, so the incidence in this group will be lower. When the connection between subgroup and unusual target distribution is not immediately intuitively clear, the result of SD is a new hypothesis to be investigated by the domain experts.

2.1 Exceptional Model Mining

Exceptional Model Mining (EMM) [12, 4] can be seen as an extension of SD: instead of a single target, EMM typically selects multiple target columns. A specific kind of *interaction* between these targets is captured by the definition of a *model class*. EMM finds a subgroup to be interesting when this interaction is exceptional, as captured by the definition of a *quality measure*. For instance, when two numerical columns are selected as the targets, we can consider Pearson's correlation ρ as the model class. Quality measures for this model class could be ρ itself (to find subgroups on which the target correlation is unusually high), $-\rho$ (to find subgroups with unusually strongly negative target correlation), $|\rho|$ (to find subgroup with unusually strong positive or negative target correlation), or $-|\rho|$ (to find subgroups with unusually weak target correlation). Hence, the model class fixes the type of target interaction in which we are interested, and the quality measure fixes what, within this type of interaction, we find interesting. Several model classes have been defined and explored; for instance, Bayesian networks [5], and regression [3]. Popular quality measure for SD/EMM include WRAcc [10], z-score [13], and KL divergence [11].

2.2 Learning Behavior Modeling

Learning behavior modeling for students in MOOCs is generally aimed at predictive analytics such as dropout prediction, passing state prediction, and grades prediction. For instance, latent factors and state machines are employed to model the hidden study state of students for a predictive task [18, 16, 21]. Khajah et al. [8] integrate Latent factor and knowledge tracing with a hierarchical Bayesian model, which can consider the study skill for prediction tasks. Recurrent neural network and LSTM have been used to model study trajectories for the learning results prediction [15, 22]. Most of these existing methods focus on modeling individual behavior but do not consider the sparsity, inconsistency and heterogeneity of learning behavior data. Our methods focus on discovering exceptional learning behaviors on the subgroup level, which provide interpretable information about where the predictive model does not perform well. This allows us to establish an improved model for prediction tasks for both normal and exceptional behavioral patterns.

3. PRELIMINARIES

We assume a dataset Ω: a bag of N records $r \in \Omega$ of the form $r = (a_1, \ldots a_k, l_1, \ldots, l_m)$, where k and m are positive integers. We call $a_1, \ldots, a_k$ the *descriptive attributes* or *descriptors* of r, and $l_1, \ldots, l_m$ the *target attributes* or *targets* of r. The descriptive attributes are taken from an unrestricted domain $\mathcal{A}$. Mathematically, we define descriptions as functions $D : \mathcal{A} \to \{0, 1\}$. A description D covers a record r^i if and only if $D(a_1^i, \cdots, a_k^i) = 1$.

Definition 1. *A subgroup corresponding to a description D is the bag of records $G_D \subseteq \Omega$ that D covers, i.e.:*

$$G_D = \left\{ r^i \in \Omega \middle| D(a_1^i, \ldots, a_k^i) = 1 \right\}$$

This merely formalizes the standard LPM conditions: we seek subgroups that are defined in terms of conditions on the descriptors, hence our results are interpretable. Those conditions select a subset of the records of the dataset: those records that satisfy all conditions. These subgroups must be evaluated, which is done by the quality measure:

Definition 2. *A quality measure is a function $\varphi : \mathcal{D} \to \mathbb{R}$ that assigns a numeric value to a description D. Occasionally, we use $\varphi(G)$ to refer to the quality of the induced subgroup: $\varphi(G_D) = \varphi(D)$.*

Typically, a quality measure assesses the subgroup at hand based on some interaction on the target columns. Hence, a description and a quality measure interact through different partitions of the dataset columns; the former focuses on the descriptors, the latter focuses on the targets, and they are linked through the subgroup.

Since subgroups select subsets of the dataset at hand, and many such subsets exist, we need to employ a search strategy to ensure that we find good results in a reasonable amount of time. To do so, we employ the *beam search* algorithm as outlined in [4, Algorithm 1]. This algorithm holds the middle ground between a pure greedy search algorithm, which is likely to quickly end up in a local optimum, and an exhaustive search, which is likely to require too much time for providing the global optimum. Beam search builds up candidate subgroups in a level-wise manner, by imposing a single condition on a single attribute at each step of the search. In subsequent steps, promising candidates are *refined*, by conjoining to these candidates all possible additional single conditions on a single attribute, and evaluating the results. A purely greedy approach would, at each step, refine the single most promising candidate. By contrast, beam search refines a fixed number w (the *beam width*) of most promising candidates at each step. The larger the choice of w, the more likely we are to escape local optima, and the longer the algorithm will take. An additional parameter of beam search is the number d (the *search depth*), which sets an upper limit to the number of steps in the search process. Hence, by design, any subgroup resulting from a beam search procedure must be defined as a conjunction of at most d conditions on single attributes. The larger the choice of d, the more expressive the results are; the smaller the choice of d, the easier the results are to interpret.

4. EXCEPTIONAL LEARNING BEHAVIOR ANALYSIS

Our dataset originates from the learners involved in the EIT Digital MOOCs at Coursera. EIT Digital, as part of the

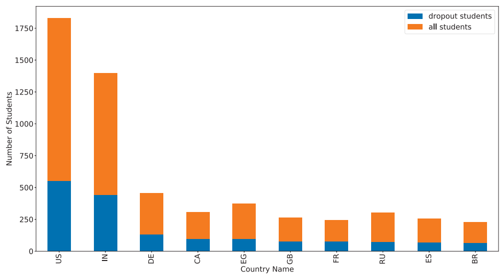


Figure 3: Dropout ratio of students by country.

Table 1: Exceptional dropout rate in subgroups. Results show subgroups with highly exceptional dropout rate. The overall dropout rate is 0.4286.

D	φ_{WRAcc}	dropout	$\lvert G_D \rvert$
Country = OM, Was Group Sponsored != True, Was Finaid Grant != True	0.0338	0.0	42
Region = MOW, Gender != male, Join Date <= 1011, Join Date > 389	0.0336	0.0	57
Country = KR, Gender != female, Profile language != ko	0.0330	0.7812	32
Country = KR, Educational status != MASTERS DEGREE, Gender != female, Was Group Sponsored != True	0.0313	0.7742	34
Country = KR, Was Group Sponsored != True	0.0304	0.7222	36

European Institute for Innovation and Technology, aims to drive Europe's digital transformation, also for education. The EIT Digital academy is focused on mobility and entrepreneurship and is at the forefront of integrating education, research, and business. The MOOCs in the online programme, have been developed by the partner universities involved in the EIT Digital Master School in Embedded Systems, in a best of breeds approach.

Together, the MOOCs form the EIT Digital online programme "Internet of Things through Embedded Systems". The online programme aims to build the reputation of EIT Digital, the partner universities, and the involved teachers. It also helps to renew pedagogy through scalable education technologies and data driven education. Learning analytics are at the core of this feedback mechanism. The online programme is comparable to an edX's micromaster and similarly offers an online equivalent of a 25 ECTS first semester; the online programme offers learners to study at their own pace, any time, any place. Moreover, they first can have a try before they commit themselves to the whole master programme. Once selected and admitted on campus, the learners can finish the double degree master programme of EIT Digital Master School in Embedded Systems.

Figure 2 displays the distributions of students across various demographic categories. In order to catch the inherent imbalance, we use demographic columns as the left hand attributes, to formulate subgroup descriptions. In the data preprocessing process, we convert the join dates, which represents how long a student has registered in Coursera, from the format of 'Datetime' to the integer days. The following three sections illustrate what kind of discoveries can be made by wielding various tools from the EMM toolbox.

Table 2: Exceptional correlation analysis between length of behavior sequence and course grades. The overall correlation coefficient ρ is 0.7406.

D	φ_{scd}	ρ	$\lvert G_D \rvert$
Country = LT, Join Date > 701, Browser language != et-EE	0.9999	0.9782	11
Region = 6	0.9994	-0.1272	10
Region = QUE	0.9992	-0.0788	11
Country = NP	0.9985	0.9630	11
Browser language = es-MX	0.9973	0.1203	7

Table 3: Exceptional classifier behavior for course passing state prediction. Results indicate that the classifier cannot work well on these exceptional subgroups.

D	φ_{f1}	$\lvert G_D \rvert$
Country = OM, Profile language = en-US, Browser language != en-US, Educational status != BACHELOR DEGREE	0.5051	32
Country = OM, Profile language != en-US	0.4058	22
Region = MA, Gender = female, Educational status=COLLEGE NO DEGREE	0.3489	24
Country = OM, Met Payment Condition != True	0.3464	31
Join Date <= 390, Region != MA	0.3193	28

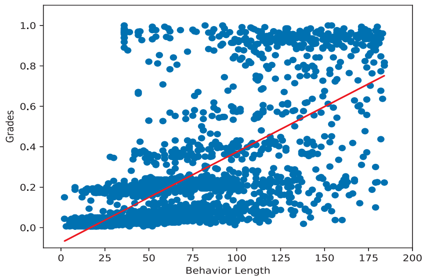
4.1 Exceptional Dropout Rate Analysis

In this section, our task is to find out the subgroups which have significantly different dropout rate compared with the whole dataset. For the purposes of this paper, we define a dropout student to be a student who has participated in at least one assessment question, but has not obtained an overall course grade. In Figure 3, we present the highest-frequency countries, and the dropout rate of students in those countries. From the figure we can see that both frequency and dropout rate vary a lot. The high dropout rate is usually seen as a defect of MOOCs. If we were to discover what kinds of students have exceptional dropout rates, then that would allow us to direct specific guidance to those students that most require it. Traditional partition and clustering methods are not qualified for this task, because they cannot provide interpretable results about the subsets of students and quantitative information about how different the subsets of students are from the whole dataset. To address this problem, we propose to engage subgroups as a partition for the whole dataset, and look for subgroups that have most exceptional dropout rate comparing with the whole dataset, employing *Weighted Relative Accuracy* (WRAcc) [20]:

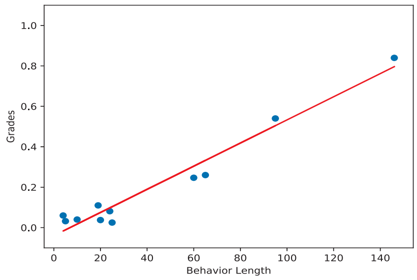
$$\varphi_{\text{WRAcc}} = \frac{|G_D|}{N}\left(\frac{S_D}{|G_D|} - \frac{S_\Omega}{N}\right)$$

Here, $|G_D|$ represents the number of records covered by subgroup description D, S_D represents the number of dropout students in subgroup G_D, S_Ω represents the total number of dropout students in the whole dataset, and N represents the number of students who join this course and participated in at least one assessment question.

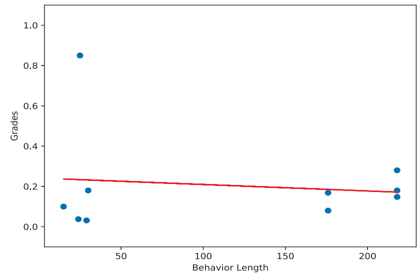
The beam search algorithm as described in [4, Algorithm 1] is parameterized with beam width 20 and search depth 4. The overall dropout rate is 0.4286. In Table 1, we presents the top-5 subgroups with most exceptional dropout rate. The subgroup with description "D: Region = MOW, Gender != male, Join Date between 389 and 1011" has a dropout rate



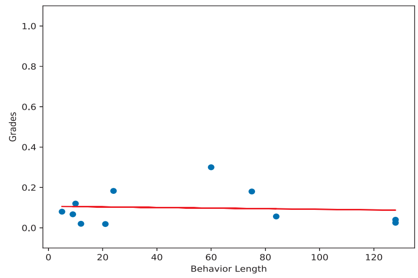
(a) The whole dataset. ρ = 0.7406



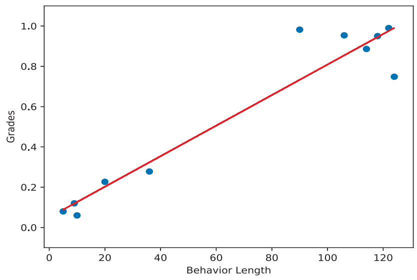
(b) Country = LT, Join Date > 701, Browser language != et-EE. ρ = 0.9782



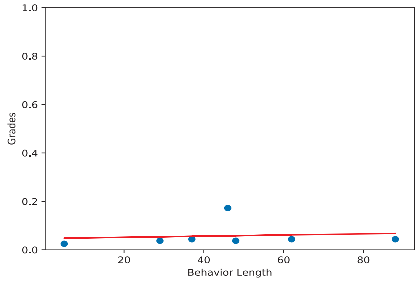
(c) Region = 6. ρ = -0.1272



(d) Region = QUE. ρ = -0.0788



(e) Country = NP. ρ = 0.9630



(f) Browser language = es-MX. ρ = 0.1203

Figure 4: Exceptional correlations in subgroups.

of zero: all students in that subgroup complete the course. On the other hand, the subgroup with description "D: Country = KR, Gender != female and Profile language != ko", has an elevated dropout rate of 0.7812: most of these students drop out. Based on these results, we can conclude that Korean males who have set their profile language to something other than Korean, are in need of more attention. This may be a group of students who are foreigners in Korea, or Koreans who are studying in a language which is non-native to them. By identifying such at-risk groups, educators can more effectively channel their remedial activities.

4.2 Exceptional Correlation Analysis

Generally, more active students can be expected to obtain higher grades. To investigate this phenomenon, we look into the relation between the activity length (denoted by q) of students and the overall grades (denoted by g) in a course. We engage the correlation model class for EMM to realize this task. In this model class, we can estimate the correlation coefficient by calculating the sample correlation as follows:

$$\widehat{r} = \frac{\sum (q^i - \bar{q})(g^i - \bar{g})}{\sqrt{\sum (q^i - \bar{q})^2 \sum (g^i - \bar{g})^2}}$$

$$z' = \frac{1}{2} \ln \left(\frac{1 + \widehat{r}}{1 - \widehat{r}} \right)$$

$$z^* = \frac{z' - z^C}{\sqrt{\frac{1}{|G_D| - 3} + \frac{1}{|G_D^C| - 3}}} \qquad (1)$$

Here, $\widehat{r}$ represents the sample correlation, q^i, g^i represent the activity length and course grade of each student, and $\bar{q}, \bar{g}$ represent their average values over the dataset. Equation (1) is the Fisher z transformation, z' in the lower equation represents the z' computation on the subgroup and z^C on its complement, and $|G_D|$ represents the number of records covered by subgroup with description D. Under the null hypothesis that the correlation between q and g is the same inside and outside of the subgroup, z^* follows a standard normal distribution. Hence, the value for z^* implies a p-value under this null hypothesis. Leman et al. [12] propose to use one minus this p-value as quality measure φ_{scd}: the higher this value is, the more certain we are that the null hypothesis is false and hence exceptional correlations are observed.

Using this quality measure, we conduct the experiment with beam width 20 and search depth 3. In Table 2 and Figure 4, we list the top-5 subgroups with exceptional quality score, coefficients, and coverage. We can see that some students gain extremely high grades with longer behavior sequence (cf. Figure 4b, 4e); some students have longer behavior sequence length but lower grades (cf. Figure 4c, 4d); and for some subgroups, the length of behavior sequences has no obvious correlation with the grades (cf. Figure 4f). We can deduce that the efforts that some students spend in the study are not directly correlated with their learning results.

4.3 Exceptional classifier behavior analysis

Students' behavioral records in MOOCs are sparse, inconsistent and heterogeneous. Learning behavior could be very different between different students. This imbalance increases the difficulty of training a classifier that can perform well on each part of the dataset. This makes it difficult to train a model that is qualified for tasks like dropout prediction and course passing state prediction.

In this section, we investigate whether learning behavior can predict whether or not a student can pass the course. At

Table 4: Course statistics.

course_name	course_level	complete_number	avg_grades	course_enroll_num	max_grades	min_grades	pass_number
Marketing	I	1141	0.105	4609	1	0.006	52
Design Thinking	I	369	0.167	3483	0.972	0.01	22
IoT	A	8	0.098	241	0.1	0.087	0
System Validation (2)	I	63	0.412	1010	1	0.05	12
Smart IoT	B	905	0.216	6035	1	0.004	100
Computer Architecture	I	913	0.510	7652	1	0.025	299
System Validation (4)	A	17	0.597	985	1	0.071	9
Quantitative Model (1)	I	429	0.395	1807	1	0.007	49
System Validation (3)	A	45	0.418	764	1	0.057	11
Quantitative Model (2)	A	979	0.339	4975	1	0.016	52
System Validation	I	601	0.376	2605	1	0.04	124
Technology	I	258	0.232	3930	1	0.002	34
Embedded Systems	I	549	0.291	3737	1	0.02	67
Software Architecture	A	2710	0.299	10487	1	0.012	331
Real-Time Systems	I	3615	0.203	15123	1	0.006	389
IoT Devices	I	430	0.318	6609	1	0.008	85
Embedded Hardware	I	3943	0.160	19592	1	0.02	128
Open Innovation	I	480	0.137	3150	0.969	0.008	24

the same time, we investigate in which parts of the dataset the classifier does not work well. In Section 4.1 and 4.2, we have presented that EMM can effectively discover exceptional learning behavioral patterns in MOOCs. We will continue using the EMM framework to find where our predictive model does not work well in the dataset. Considering the activities of students in assessments, forum discussions and peer assignments, we formulate the passing state prediction problem as follows:

$$f : \mathcal{X}^i \to Y^i$$

Our aim is to train a classifier f that can automatically map $\mathcal{X}^i$ to Y^i, where $\mathcal{X}^i$ is a 8-tuple $(s^i, m^i, o^i, c^i, b^i, e^i, h^i, p^i)$ feature vector representing the length of assessment and question sequence (s^i), number of assessment types (m^i), number of question types (o^i), number of correctly answered questions (c), number of asked, answered and liked questions in the forum (b^i, e^i, h^i), and peer review score (p^i), and where Y is the label of passing state: $\{0, 1\}$. We normalize the features into 0 to 1 as the input values.

At first, the classifier is trained on the whole dataset. This model will classify some students correctly and some students wrongly; in any case we find a value of predicted labels $\hat{Y}$. These two binary values Y and $\hat{Y}$ will agree and disagree on some students, and that interaction can be used to capture the quality of the classifier predictions in a single number. We use the f1 score to capture this:

$$\varphi_{\text{f1}} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

However, we can perform the exact same computation for a subset of the vectors Y and $\hat{Y}$, for instance the subset induced by a subgroup. Thus, we employ φ_{f1} as a quality measure for EMM.

We conduct the experiment by setting the search depth to 4 and beam width to 10. We engage an SVM classifier as the predictive model[1], which has 0.85 as f1 score on the whole dataset. In Table 3 we list the top-5 subgroups with exceptional behavior. We can see that even though the classifier performs well on the whole dataset, in some subgroups it does not. Particularly for the students described by descriptions like “D: Region = MA, Gender = female, Educational status=COLLEGE NO DEGREE”, the classifier performs poorly on the prediction task at hand: the support vector machine has trouble predicting the study success of Massachusets women without a college degree. Hence, this group requires a more sophisticated classifier.

[1]one may plug in one’s preferred classifier; SVM selection is merely meant as an illustration, not an endorsement.

5. CONCLUSIONS

In this paper, we employ Exceptional Model Mining (EMM) for exceptional learning behavior analysis in MOOCs. Rather than predicting the success of individual students, which is difficult due to the inherent sparsity, inconsistency, and heterogeneity of the data, EMM specializes in identifying coherent groups that behave differently from the norm. Since the subgroups resulting from EMM come with an easily interpretable definition, Exceptional Model Mining allows educators to more effectively channel their remedial activities.

We employ three EMM model classes for different tasks of learning behavior analysis. Experimental results on a real Coursera dataset show that for some students, the dropout rate is very different from the whole dataset, the learning efforts are not always correlated with course grades, and a classifier that performs very well on the whole dataset has trouble on some subpopulations of the data. In future work, we will make use of these discovered exceptional behavioral patterns to establish an improved model, which can model both normal and exceptional learning behaviors for the students in MOOCs. We plan to develop a modeling method that can perform well on each part of the dataset, including the exceptional ones.

6. ACKNOWLEDGMENTS

This research was funded by EIT 18008-A1803 project. In addition, Xin Du would like to thank China Scholarship Council (CSC) for the financial support.

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo. Fast Discovery of Association Rules. *Advances in Knowledge Discovery and Data Mining*, pp. 307–328, 1996.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] W. Duivesteijn, A. Feelders, and A. Knobbe. Different slopes for different folks: mining for exceptional regression models with cook's distance. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 868–876, 2012.

[4] W. Duivesteijn, A. J. Feelders, and A. Knobbe. Exceptional model mining. *Data Mining and Knowledge Discovery*, 30(1):47–98, 2016.

[5] W. Duivesteijn, A. Knobbe, A. Feelders, and M. van Leeuwen. Subgroup discovery meets Bayesian networks — an exceptional model mining approach. In *10th International Conference on Data Mining (ICDM)*, pp. 158–167, 2010.

[6] D. Hand, N. Adams, R. Bolton (eds). *Pattern Detection and Discovery.* Springer, New York, 2002.

[7] F. Herrera, C. J. Carmona, P. González, and M. J. Del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, 29(3):495–525, 2011.

[8] M. Khajah, R. Wing, R. Lindsey, and M. Mozer. Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. In *Educational Data Mining*, 2014.

[9] W. Klösgen. Explora: A Multipattern and Multistrategy Discovery Assistant. *Advances in Knowledge Discovery and Data Mining*, pp. 249–271, 1996.

[10] M. van Leeuwen and A. J. Knobbe. Non-redundant subgroup discovery in large and complex data. In *Proceedings of the European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases*, pp. 459–474, 2011.

[11] M. van Leeuwen and A. J. Knobbe. Diverse subgroup set discovery. *Data Mining and Knowledge Discovery*, 25(2):208–242, 2012.

[12] D. Leman, A. Feelders, and A. Knobbe. Exceptional model mining. In *Proceedings of the European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases*, pages 1–16. Springer, 2008.

[13] M. Mampaey, S. Nijssen, A. Feelders, R. Konijn, and A. Knobbe. Efficient algorithms for finding optimal binary features in numeric and nominal labeled data. *Knowledge and Information Systems*, 42(2):465–492, 2015.

[14] K. Morik, J. F. Boulicaut, A. Siebes (eds). *Local Pattern Detection.* Springer, New York, 2005.

[15] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep knowledge tracing. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pp. 505–513, 2015.

[16] J. Qiu, J. Tang, T. X. Liu, J. Gong, C. Zhang, Q. Zhang, and Y. Xue. Modeling and predicting learning behavior in moocs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pp. 93–102, 2016.

[17] M. Qiu, F. Zhu, and J. Jiang. It is not just what we say, but how we say them: Lda-based behavior-topic model. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pp. 794–802, 2013.

[18] A. Ramesh, D. Goldwasser, B. Huang, H. Daume III, and L. Getoor. Learning latent engagement patterns of students in online courses. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[19] D. T. Seaton, Y. Bergner, I. Chuang, P. Mitros, and D. E. Pritchard. Who does what in a massive open online course? *Communications of the ACM*, 57(4):58–65, 2014.

[20] L. Todorovski, P. Flach, and N. Lavrač. Predictive performance of weighted relative accuracy. In *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*, pages 255–264, 2000.

[21] F. Wang and L. Chen. A nonlinear state space model for identifying at-risk students in open online courses. *Proceedings of the 9th International Conference on Educational Data Mining*, pp. 527–532, 2016.

[22] L. Wang, A. Sy, L. Liu, and C. Piech. Learning to represent student knowledge on programming exercises using deep learning. In *Proceedings of the 10th International Conference on Educational Data Mining*, pp. 324–329, 2017.

[23] S. Wrobel. An Algorithm for Multi-relational Discovery of Subgroups. In *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 78–87, 1997.

[24] Z. Zhao, Z. Cheng, L. Hong, and E. H. Chi. Improving user topic interest profiles by behavior factorization. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1406–1416, 2015.