Brief Survey on Student Performance Prediction

Yuqing Zhao

***Abstract*—**

I. Introduction

SOME crucial aspects differentiate state of the art for learning performance prediction research. According to the basic structure of student performance prediction task (Fig.1), existing research on Learning Performance Prediction can be divided according to data(input), predicted values(output), and assumption.

II. Input data

Performance prediction has different prediction effects because of the different input data. The input data in the existing research can be divided into direct factors and indirect factors.

A. Direct feature

Direct factors refer to the original data in the education system. These data can be divided according to the type of education system.

In traditional education systems

The commonly used mining data in the traditional education system includes course performance, demographic feature, and learning material.

1) Learning outcome: Existing articles usually use pre-course performance data or current course performance, assuming that the past or the current stage can be used for future state prediction. In the traditional education system, these course performance data can be quiz results, test scores, accuracy rate, GPA for a semester, a school year, graduation, rewards and punishments, and homework completion.

2) Demographic feature: Demographic data is commonly used in traditional education systems, including family, Gender, Age, Level of Schooling, Country of Origin, Primary Language, Employment Status, etc. [1] [2]

In distance education systems

Distance education system refers to a learning system for teaching and learning online. Compared with traditional education systems, distance education systems have more complex and diverse data types. Data records are complete and coherent so that more information can be obtained from them. The unique data types in the distance education system are learning material and learning behaviour data. Learning behaviour data refers to the data that students interact with the system when they participate in online teaching.

3) Learning material: There are many learning material types in the distance education system, such as teaching videos [3], text, and picture courseware.

4) Learning behaviour data: Learning behaviour data refers to the log [4] and engagement data generated by the interaction between students and the system during online learning. Here, learning behaviour data is divided into video related, exercise-related, forum related and platform use according to online teaching.

Video-related: Learning behaviour data contains data generated by student interaction during video learning and video software teaching. For example, log data such as play, pause, skip, adjusting rate, etc. when watching the video, and students answering questions and discussing during the video lecture.

Exercise-related: Exercise-related learning behaviour data [5] refers to related data such as quiz [6], assignment [7], and test. For example, test scores, whether the exercises are CFA (Correct at First Attempt), Distinct Problems Attempted, Submissions, Time Spend on One Problem, Number of Correct Problems, etc.

Forum related: Contains data like Negative/Positive Rating Number, Posts Number, Words Number, discussion [8].

Platform use: Platform use data is platform-related data in addition to other data. Contains such as the number of access, the period of attending courses, access time, and registration time, etc.

B. Indirect feature

Indirect factors refer to factors that require processing to be indirectly extracted.

1) Social network: Some studies use graph structure, namely social network, to simulate the relationship between students, teachers and learning process, and introduce it as an additional feature into the prediction task [6] [8] [9] [10].

Researchers at Princeton University brought social networking into the education system, creating the Social Learning Network (SLN) [6] [8]. SLN contains the dynamic changes of learning behaviour, and the graphs of these dynamic changes represent the relationship between people and the learning process. Factors like path-based features, post-based features and neighbourhood-based features are added to the prediction model as additional features.

2) Personality and Learning style: Some researchers [11] [12] [13] extract learning style from student data as an additional feature to predict student performance. For example, Hoang et al. adopt Feld Silverman Learning Styles Model (FSLSM) that divides learning styles into four dimensions, each of which follows two opposite directions: ac-

tive/reflective, perceptual/intuitive, visual/verbal, and sequential/global.

III. Predicted value

Predicted value refers to the output generated in the process of predicting student performance, that is, the definition of student performance. Other common predicted values are also mentioned below.

A. Student performance

Commonly used outputs to measure student learning performance prediction include GPA, CFA (Correct on First Attempt), Exam / Test Score, Course Grade Range, Program or Module Graduation / Retention / Dropout, Assignment Performance, Number of Passed Course, Certificate Earners, etc.

B. Other predicted values

In addition to students' learning performance, the existing prediction values based on educational data mainly includes Interaction between Learner [8], Concept associations [14] [15], Question difficulties [7], Student behavior [16] [17] [18] [19] [20], etc..

IV. Assumption

In the process of modelling, we need to make some assumptions. A more discriminating assumption is whether the educational system or learning process by students is seen as sequential.

A. Sequential

Assume that the system is in a sequential state. In the student modelling section, some studies are based on the assumption that there is an interdependency on the sequence of the defined stages. Therefore, the performance in the future or the present stage can be predicted through this interdependency. These stages can be defined by periods of equal interval such as semesters, quarters, and years, as well as by "steps the student takes to complete a problem", past and future stages, and new and old educational systems.

For example, Pardos, Zachary A et al. divide the one response from a student on a problem step into different stages, and use the Hidden Markov property between the different steps to predict whether the student can be correct the first attempt (CFA) [21]. Some researchers [22] [23] use Three-mode Tensor Factorization (on student/problem/time) to introduce a time tensor into the prediction model. Given a three-mode tensor Z of size U × I× T, where the first mode describes U students, the second mode describes I tasks (problems), and the third mode describes the time. A study [4] published on AAAI in 2017 introduced the Quarter-wise Based Model, which divides education data into different quarters in order. The forecast results of the previous quarter can be used for forecasting this quarter. Studies at Stanford University [15] and USTC [5] also take into account the sequential nature of students' learning. They used the Recurrent Neural Network (RNN) for knowledge tracing and sequential modelling.

Prior performance Some research work [15] [22] [24] [25] uses students' previous performance to predict current or future grades, which basically model the education system as two stages, and assume that the present or future stage has dependencies on the previous phases.

B. Non-sequential

Some studies do not consider the sequential effects of learning, but rather make predictions as a single event. Without considering the temporal/sequential factor, they emphasize some relational features or adopt more efficient models. For example, Thai-Nghe et al. used multi-relational matrix factorization (MF) to take into account the relationship between students and multiple tasks and skills. Many other machine learning techniques, such as decision trees[26], artificial neural networks [26], collaborative filtering [27] [28] and probabilistic graphical models [29] [8] have been used to develop student performance prediction algorithms.

C. Cold start problem

Due to the problem of insufficient data for freshmen, we often encounter the problem of "cold start" in the process of predicting their grades.

Several existing studies have attempted to solve the "cold start" problem in different ways. Yu Su et al. proposed an Exercise-Enhanced Recurrent Neuro Network (EERNN), which solves the "cold start" problem by associating students' performance with their corresponding exercise patterns when predicting the performance of new students and new exercises. [5]. Here, Cold Start is defined by limited data.

Coleman et al. used an ensemble-based algorithm to study dropout rates in data-rich school districts and predict dropout rates in data-poor districts based on the assumption that students in these districts share the same learning patterns as students in data-rich districts. [30].The cold start problem here refers to predicting performance in districts suffering from high data missingness.

The Item Response Theory (IRT) is used in a study [31] to extract the learner's initial ability and corresponding performance in various difficulty levels from a set of learners' feature(such as age, relevant courses, IQ, and predicted grades) and their known responses to predict the proficiency of new learners.

Mack Sweeney et al. demonstrated that the use of course and student bias terms is sufficient to mitigate the weakness of predicting for unseen student performance and provide reasonable solutions using the factorization machine [32] and hybrid FM-RF method [33]. In these papers, the cold start problem refers to predicting next-term student grade, in which either or both students and courses occurred in a semester and did not occur in any previous semester.

In 2010, Thai-Nghe et al. used recommendation system technology to carry out educational data mining, significantly to predict students' performance. The Future Work section stated that they would use a sounder approach to the cold

start problem using matrix factorization [34]. In another paper published in 2011, they used multi-relational Factorization Models to predict student performance. This paper provides the global average score for new students or new tasks, as "in the educational environment, the cold-start problem is not as harmful as in the e-commerce environment" [35]. Nevertheless, with MOOCs' emergence, the "cold start" problem has gained some traction [36]. The new students and tasks here are defined by Thai-Nghe et al. as data in the test set but not in the training set.

Camacho et al. did a review of the research on alleviating the cold start problem with social network data and scored articles based on whether social network information was used to alleviate the cold start problem of Collaborative-Filtering-based recommender systems [37].

Vie et al. proposed an algorithm based on an approach recently used in machine learning: a row-point process that samples an initial set of different problems for new learners. Based on Item Response Theory, Knowledge Tracing and Cognitive Diagnosis, this model inferred potential variables from the test results, to make better predictions for new learners [36].

Huang et al. analyzed educational "check-in" data using WiFi access logs collected by Purdue University. They proposed a heterogeneous graph-based approach to encode correlations between users, places and activities, and then learn vertex embedding together. They use the same scoring function for both visited and unvisited places(cold-start places) to perform user recommendations [38].

REFERENCES

[1] S. M. M. Rubiano and J. A. D. Garcia, "Formulation of a predictive model for academic performance based on students' academic and demographic data," in *2015 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2015, pp. 1–7.

[2] B. Trstenjak and D. onko, "Determining the impact of demographic features in predicting student success in croatia," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2014, pp. 1222–1227.

[3] X. Wang, J.-F. Hu, J.-H. Lai, J. Zhang, and W.-S. Zheng, "Progressive teacher-student learning for early action prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3556–3565.

[4] J. Xu, Y. Han, D. Marcu, and M. Van Der Schaar, "Progressive prediction of student performance in college programs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[5] Y. Su, Q. Liu, Q. Liu, Z. Huang, Y. Yin, E. Chen, C. Ding, S. Wei, and G. Hu, "Exercise-enhanced sequential modeling for student performance prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[6] C. G. Brinton and M. Chiang, "Mooc performance prediction via clickstream data and social learning networks," in *2015 IEEE conference on computer communications (INFOCOM)*. IEEE, 2015, pp. 2299–2307.

[7] A. S. Lan, C. Studer, and R. G. Baraniuk, "Time-varying learning and content analytics via sparse factor analysis," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 452–461.

[8] T.-Y. Yang, C. G. Brinton, and C. Joe-Wong, "Predicting learner interactions in social learning networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1322–1330.

[9] C. G. Brinton, S. Buccapatnam, F. M. F. Wong, M. Chiang, and H. V. Poor, "Social learning networks: Efficiency optimization for mooc forums," in *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 2016, pp. 1–9.

[10] C. Reffay and T. Chanier, "How social network analysis can help to measure cohesion in collaborative distance-learning," in *Designing for change in networked learning environments*. Springer, 2003, pp. 343–352.

[11] L. Zeng, D. Chen, K. Xiong, A. Pang, J. Huang, and L. Zeng, "Medical university students' personality and learning performance: Learning burnout as a mediator," in *2015 7th international conference on information technology in medicine and education (ITME)*. IEEE, 2015, pp. 492–495.

[12] A. Raza and L. Capretz, "Do personality profiles differ in the pakistani software industry and academia—a study," *International Journal of Software Engineering (3: 4)*, pp. 60–66, 2012.

[13] R. M. Carro and V. Sanchez-Horreo, "The effect of personality and learning styles on individual and collaborative learning: Obtaining criteria for adaptation," in *2017 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2017, pp. 1585–1590.

[14] L. Pan, X. Wang, C. Li, J. Li, and J. Tang, "Course concept extraction in moocs via embedding-based graph propagation," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, pp. 875–884.

[15] C. Piech, J. Spencer, J. Huang, S. Ganguli, M. Sahami, L. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," *arXiv preprint arXiv:1506.05908*, 2015.

[16] Y. Cao, J. Gao, D. Lian, Z. Rong, J. Shi, Q. Wang, Y. Wu, H. Yao, and T. Zhou, "Orderliness predicts academic performance: behavioural analysis on campus lifestyle," *Journal of The Royal Society Interface*, vol. 15, no. 146, p. 20180210, 2018.

[17] J. Park, K. Denaro, F. Rodriguez, P. Smyth, and M. Warschauer, "Detecting changes in student behavior from clickstream data," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 2017, pp. 21–30.

[18] 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*, 2016, pp. 93–102.

[19] S. Tomkins, A. Ramesh, and L. Getoor, "Predicting post-test performance from online student behavior: A high school mooc case study." *International Educational Data Mining Society*, 2016.

[20] T.-Y. Yang, C. G. Brinton, C. Joe-Wong, and M. Chiang, "Behavior-based grade prediction for moocs via time series neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 716–728, 2017.

[21] Z. A. Pardos and N. T. Heffernan, "Using hmms and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset," *Journal of Machine Learning Research W & CP*, vol. 40, 2010.

[22] Y. Meier, J. Xu, O. Atan, and M. Van der Schaar, "Predicting grades," *IEEE Transactions on Signal Processing*, vol. 64, no. 4, pp. 959–972, 2015.

[23] N. Thai-Nghe, T. Horváth, and L. Schmidt-Thieme, "Factorization models for forecasting student performance," in *Educational Data Mining 2011*, 2010.

[24] C. Romero, M.-I. López, J.-M. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," *Computers & Education*, vol. 68, pp. 458–472, 2013.

[25] Y. Bergner, S. Droschler, G. Kortemeyer, S. Rayyan, D. Seaton, and D. E. Pritchard, "Model-based collaborative filtering analysis of student response data: Machine-learning item response theory." *International Educational Data Mining Society*, 2012.

[26] Y.-h. Wang and H.-C. Liao, "Data mining for adaptive learning in a tesl-based e-learning system," *Expert Systems with Applications*, vol. 38, no. 6, pp. 6480–6485, 2011.

[27] A. Toscher and M. Jahrer, "Collaborative filtering applied to educational data mining," *KDD cup*, 2010.

[28] D. Lemire, H. Boley, S. McGrath, and M. Ball, "Collaborative filtering and inference rules for context-aware learning object recommendation," *Interactive Technology and Smart Education*, 2005.

[29] R. Bekele and W. Menzel, "A bayesian approach to predict performance of a student (bapps): A case with ethiopian students," *algorithms*, vol. 22, no. 23, p. 24, 2005.

[30] C. Coleman, R. S. Baker, and S. Stephenson, "A better cold-start for early prediction of student at-risk status in new school districts." *International Educational Data Mining Society*, 2019.

[31] K. Pliakos, S.-H. Joo, J. Y. Park, F. Cornillie, C. Vens, and W. Van den Noortgate, "Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems," *Computers & Education*, vol. 137, pp. 91–103, 2019.

[32] M. Sweeney, J. Lester, and H. Rangwala, "Next-term student grade prediction," in *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 2015, pp. 970–975.

[33] M. Sweeney, H. Rangwala, J. Lester, and A. Johri, "Next-term student performance prediction: A recommender systems approach," *arXiv preprint arXiv:1604.01840*, 2016.

[34] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, "Recommender system for predicting student performance," *Procedia Computer Science*, vol. 1, no. 2, pp. 2811–2819, 2010.

[35] N. Thai-Nghe, L. Drumond, T. Horváth, L. Schmidt-Thieme *et al.*, "Multi-relational factorization models for predicting student performance," in *KDD Workshop on Knowledge Discovery in Educational Data (KDDinED)*. Citeseer, 2011, pp. 27–40.

[36] J.-J. Vie, F. Popineau, É. Bruillard, and Y. Bourda, "Automated test assembly for handling learner cold-start in large-scale assessments," *International Journal of Artificial Intelligence in Education*, vol. 28, no. 4, pp. 616–631, 2018.

[37] L. A. G. Camacho and S. N. Alves-Souza, "Social network data to alleviate cold-start in recommender system: A systematic review," *Information Processing & Management*, vol. 54, no. 4, pp. 529–544, 2018.

[38] M. Hang, I. Pytlarz, and J. Neville, "Exploring student check-in behavior for improved point-of-interest prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 321–330.