Dual Path Convolutional Neural Network for Student Performance Prediction

Yuling Ma[1,2], Jian Zong[1], Chaoran Cui[3(✉)], Chunyun Zhang[3], Qizheng Yang[1], and Yilong Yin[1(✉)]

[1] School of Software, Shandong University, 250100 Jinan, China
mayuling@mail.sdu.edu.cn, ylyin@sdu.edu.cn
[2] School of Information Engineering, Shandong Yingcai College, 250104 Jinan, China
[3] School of Computer Science and Technology,
Shandong University of Finance and Economics, 250014 Jinan, China
crcui@sdufe.edu.cn

Abstract. Student performance prediction is of great importance to many educational domains, such as academic early warning and personalized teaching, and has drawn numerous research attention in recent decades. Most of the previous studies are based on students' historical course grades, demographical data, in-class study performance, and online activities from e-learning platforms, e.g., Massive Open Online Courses (MOOCs). Thanks to the widely used of campus smartcard, it supplies an opportunity to predict students' academic performance with their off-line behavioral data. In this study, we seek to capture three student behavioral characters, including duration, variation and periodicity, and predict students' performance based on the three types of information. However, it is highly challenging to extract efficient features manually from the huge amount of raw smartcard records. Besides, it is not trivial to construct a good predictive model for some majors with limited student samples. To address the above issues, we develop a novel end-to-end deep learning method and propose Dual Path Convolutional Neural Network (DPCNN) for student performance prediction. Moreover, we introduce multi-task learning to our method and predict the performance of students from different majors in a unified framework. Experimental results demonstrate the superiority of our approach over the state-of-the-art methods.

Keywords: Student performance prediction · Campus behavior · Convolutional Neural Networks (CNN) · Multi-task learning

1 Introduction

As one of the most popular topics in educational data mining, student performance prediction plays a crucial role in many educational domains, e.g., student academic early warning and personalized teaching [1–3]. For example, based on the results of a predictive model, the instructor can provide personalized

R. Cheng et al. (Eds.): WISE 2019, LNCS 11881, pp. 133–146, 2019.
https://doi.org/10.1007/978-3-030-34223-4_9

intervention and guidance to improve student learning, especially for those low-performance students [2]. In recent decades, extensive research effort has been devoted to student performance prediction [4]. Owing to the convenience of collecting data, a large portion of studies focus on e-learning platforms, e.g., MOOCs [5,6], and predict students' performance based on online study activity logs. However, these data concerned with online activities is hardly captured in off-line learning scenarios. With students' historical course grades, demographical data, and their study records on target course (i.e., the course to be predicted), the other series of researches construct predictive models by direct use of part or whole of the aforementioned data [2,3,7]. However, these studies generally suffered from limited efficient features/predictors.

Thanks to the development of information technology, campus smartcards are widely used in colleges, which record about students' campus activities in an unobtrusive way. It supplies an opportunity to predict students' academic performance from the new perspective of campus behaviors. Recent studies illustrate that such real-time digital records generally can reveal some behavioral factors correlated with student academic performance [8,9]. Intuitively, a certain swiping card behavior of a student may reflect an incident that happened to him/her, e.g., swiping smartcards in the library, campus supermarket and dormitory generally means studying, shopping and relaxing, respectively. Each implicit incident may be correlated with students' academic performance, e.g., if a student spends long time in library, there is a very high probability that he/she is a diligent student and will achieve good academic performance.

Motivated by the aforementioned analysis, in this paper, we aim to model three behavioral characters, including duration, variation and periodicity, and construct predictive models for students' performance with the three types of information. Besides, considering inconsistent course settings across majors, we view constructing predictive model for different majors as different tasks. However, it is highly challenging to construct good predictive models owing to the two following issues:

(1) Traditional handcrafted features are highly dependent upon human experts and domain knowledge, and it is thus highly challenging to extract efficient features manually from huge amount of raw smartcard records.
(2) The number of students varies from major to major. For the majors with limited student samples, it is not trivial challenging to train a good predictive model.

To address the above issues, in this study, we exploit a novel end-to-end deep learning approach to predict student academic performance. Recent results indicate that the implicit features extracted from the Convolutional Neural Networks (CNN) are very efficient [10], and it has been empirically illustrated that CNN has powerful ability to hierarchically capture the spatial structural information [11]. Benefiting from these findings, we employ CNN to learn features from the raw smartcard records, and propose a Dual Path CNN method, called DPCNN, to model the aforementioned three behavioral characters. Specifically,

we represent students as tensors by direct use of raw records, each dimension of which denote time, location and date of swiping card behaviors, respectively. Then two types of filters are designed according to the size of student tensor, and utilized in the dual path network to model duration and variation, respectively. By taking the date axis as the depth of convolutions, periodicity can be modeled. Besides, given limited student samples in some majors, we introduce multi-task learning [12] to our method. Through the shared convolutional layers followed by task-specific fully-connected networks, predictive models for different majors can be trained in a unified framework. In this way, the problem of data scarcity can be alleviated. Our contributions are four-fold:

- Instead of extracting features manually, we exploit end-to-end learning style to predict students' performance. To the best of our knowledge, such a deep learning approach for student performance prediction in traditional teaching scenarios has not been previously reported.
- We propose a dual path CNN method, which is comprised of dual path convolutions followed by three-layer fully-connected networks, and three aforementioned behavioral characters can be modeled based on well-designed filters.
- We construct predictive models for different majors simultaneously following the idea of multi-task learning. Benefiting from relatedness between majors, the problem of data scarcity can be effectively alleviated.
- Experimental results demonstrate the superiority of our approach over the state-of-the-art methods.

In the following, we will briefly review related works, then the proposed method DPCNN is detailed in Sect. 3. We report experimental results and analysis in Sect. 4, followed by the conclusion and future work in Sect. 5.

2 Related Work

As one of the most important research branches of educational data mining, there has been a large body of work on student performance prediction in recent decades. Owing to the convenience of collecting data, many efforts have been devoted to predicting performance based on online activity logs from e-learning platforms, including MOOCs [5,6,13–17], Intelligent Tutoring Systems (ITS) [18–20], Learning Management Systems (LMSs) [1,21–24], Hellenic Open University (HOU) [25,26], and other platforms [27–31]. For example, Ren et al. predicted grades using data from MOOC server logs, such as the average number of daily study sessions, total video viewing time, number of videos a student watches, and number of quizzes [6]. Macfadyen et al. developed predictive models of student final grades based on LMS tracking data, including the number of discussion messages posted, number of mail messages sent, and number of assessments completed [23]. Zafra et al. predicted students' performance (i.e., pass or fail) with the information about quizzes, assignments and forums stored in Moodle, which is a free learning management system [24]. As can be seen, the above studies for e-learning platforms have mainly relied on the data about

students' online activities, which is hardly accessed in off-line study scenarios. Another line of studies utilized students' demographical data, in-class study performance, and their past course grades to construct predictive models for student performance [2,3,7,32–36]. To name a few, Huang et al. predicted students' final grades for a course based on scores in three mid-term exams and grades in four pre-requisite courses [2]. Meier et al. predicted students' final grades based on the performance assessments on homework assignments, mid-term exam, course project, and final exam [3]. Ma et al. predicted students' performance prior to a course's commencement with their historical course grades as well as course description [7]. Marbouti et al. utilized the in-class performance factors, including grades for attendance, quizzes, and weekly homework, to predict at-risk students [32]. However, these researches generally suffer from limited efficient features.

Recently, there is a growing trend to predict students' performance based on their behavioral data, which is instantly recorded in campus smartcards [8,9]. In [8], the authors extracted two high-level behavioral characters, including orderliness and diligence, to predict students' GPA ranking. In [9], besides orderliness and diligence, two more factors, i.e., sleeping pattern and friend factors, were extracted to construct predictive models. However, these features are extracted in a manual way, which are highly dependent upon human ingenuity and prior knowledge.

3 Framework

In this section, we first represent student samples as tensors based on their smartcard records, then the DPCNN framework is proposed, followed by multi-task learning and the implementation Details.

3.1 Student Representation

In our dataset, smartcard records cover the period from September 01, 2013 to August 31, 2015 (i.e., 730 days totally), and the time of swiping card in each day varies from 6am to 12 pm (i.e., 18 h totally), which may occur at 12 campus places, e.g., the library, canteen and dormitory. As aforementioned, instead of extracting features manually, we seek to learn representations for student samples with deep learning methods. Therefore, we denote a student sample as a tensor $X \in R^{t \times l \times d}$ by using the raw records directly. Here, t denotes the number of time intervals that a day is split into, l denotes the number of campus places where swiping card behaviors may occur, and d denotes the number of days during the period covered by smartcard records. If a student X has a record of swiping smartcard at the j^{th} campus place in the i^{th} time interval of the k^{th} day, X_{ijk} equals 1, otherwise it equals 0. As can be seen easily, in this study, the value of l and d is 12 and 730, respectively. Additionally, we divide the swiping card period in a day into 18 time bins, each of which spans 1 h. Thus, t is equal to 18. With tensors, we can analysis students' behavioral data

from multiple views, i.e., temporal dimension, spatial dimension and periodic dimension.

3.2 Dual Path CNN

As aforementioned, CNN has powerful ability to hierarchically capture the spatial structural information. We thus choose CNN as the backbone to construct our framework. In this part, we first analyze how to adopt convolutional operations to model the behavioral characters, and then details the proposed method.

Given a student tensor $X \in R^{t \times l \times d}$, we attempt to employ filters of different size to model different behavioral characters. They are as follows: (1) filters of size $\alpha \times l$ with taking the date axis as the depth of convolutions. Here, l equals to the width of the tensor, which denotes the number of campus places (i.e., 12 in this work), and $\alpha \leq t$ is a hyperparameter, which means how many time intervals can be observed per convolution. Trough convolutional operations upon the tensor X along the axis of time intervals, swiping smartcard behaviors at different time in a day can be observed, and the changing patterns of these behaviors in temporal domain can be captured. In this way, the behavioral character *duration* can be modeled; (2) filters of size $t \times 1$ with taking the date axis as the depth of convolutions. Here, t is consistent with that in $X \in R^{t \times l \times d}$, both of which equal the number of time intervals a day split into (i.e., 18 in this study), and the width of such filters is set to be one. The reason is that we consider only one campus place per convolutional operation. Similarly, with the proceeding of convolutions along the axis of location, swiping card behaviors at different places can be observed, and the changing patterns of these behaviors in spatial domain can be captured. In this way, the behavioral character *variation* can be modeled. Additionally, the depth of both the above-mentioned convolutions are the date axis, and thus periodicity of campus behaviors can be modeled.

Based on the above analysis, we propose a novel Dual Path Convolutional Neural Network, called DPCNN, to model the aforementioned high-level behavioral characters. Figure 1 presents the architecture of DPCNN, which is comprised of dual path convolutional layers followed by a three-layer fully-connected neural network. The aforementioned two types of filters are exploited as a start in dual path convolutions, i.e., filters of size $\alpha \times l$ for the top path (i.e., Path1) to model duration, and filters of size $t \times 1$ for the bottom path (i.e., Path2) to model variation. Here, l and t is equal to 12 and 18, respectively, as aforementioned. The hyperparameter α is empirically set to be 3 based on the data used in our study. In order to fetch more information, more filters of size $\alpha \times l$ / $t \times 1$ can be adopted in the Path1/Path2 as the first-level convolutional layer. Besides, it was empirically observed that layerwise stacking of convolutions often yielded better representations [37]. Thus more levels of convolutional layers are employed in the dual path structure. Due to limited storage, the first level of convolutions in dual path both exploit 64 filters, which is followed by three more levels of convolutions with 128, 256, 512 filters, respectively. Finally, the output features generated from path1 and features from path2 are both rearranged to be feature

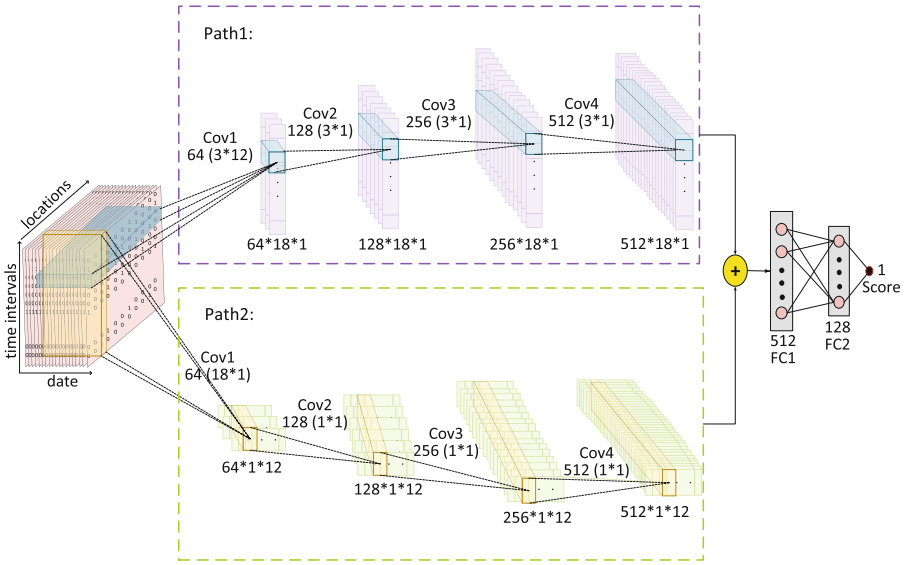

Fig. 1. DPCNN architecture. Cov: Convolution; $\oplus$: Concatenation operation; FC: Fully-connected

vectors, which are concatenated and then fed into a three-layer fully-connected network to make predictions.

In our study, the labels are the ranking of students' accumulated Grade Point Average (GPA), and thus we follow the idea of learning-to-rank to train our model. Formally, we denote each pairwise comparison by a triple $(X_i; X_j; y)$, where $X_i, X_j \in R^{t \times l \times d}$ are two student samples, and $y \in \{+1, -1\}$ is a label. $y = +1$ denotes that the former student (i.e., X_i) is ahead of the latter X_j's ranking, and $y = -1$ means the reverse. We denote the dataset consisted of n pairwise comparisons as $D = \left\{\left(X_i^k, X_j^k, y^k\right)\right\}_{k=1}^{n}$. The goal of our task is to learn a mapping function $f(X) \rightarrow R$ that can give the predictive real value for each student sample. The desired mapping function is obtained by minimizing the hinge loss function as follows:

$$L = \sum_{(X_i, X_j, y) \in D} max(0,\ y(f(X_j) - f(X_i)) + 1) \tag{1}$$

3.3 Multi-task Learning

The data used in this study are behavioral records of 8199 undergraduate students from 19 majors, and the label of a student is his or her ranking based on students' GPA in his or her major. Due to inconsistent course settings across majors, it may be irrational to construct a common predictive model with mixing data of different majors brutally. Therefore, we view constructing models

for different majors as different tasks. However, the number of students varies across majors from 100 to 600. For the majors with limited students, it is highly challenging to train a good predictive model. Multi-task learning is an empirically good solution, which can train classifiers for multiple related tasks simultaneously [12]. Though we view constructing classifiers for different majors as different tasks, these tasks may be correlated owing to the similarity of courses and teaching styles from similar majors, such as computer science and electronic engineering [9]. We thus introduce multi-task learning to our framework. Hard parameter sharing is the most commonly used approach to multi-task learning in neural networks, which generally applied by sharing the hidden layers between all tasks, while keeping several task-specific output layers [38]. Motivated by this, we let all the tasks share representation learning layers (i.e., the dual path convolutional layers of DPCNN) while remaining the final three-layer fully-connected networks task-specific. Extension experiments illustrate the appealing effectiveness of multi-task learning, which is reported later.

3.4 Implementation Details

The python libraries, including "torch" and "torchvision", are used to build our network. As aforementioned, in our study, student samples are denoted by tensors of size $18 \times 12 \times 730$. Thus, filters of size $\alpha \times 12$ and 18×1 are utilized to model duration and variation, respectively. The hyperparameter α is empirically set to be 3. Additionally, we take the axis of date as the depth of convolutions to model periodicity. With the limited storage, we utilize 64 filters to fetch more information in the first-level convolutions, and three more levels of convolutions are exploited to yield a better representation, with 128, 256 and 512 filters, respectively. More specifically, in DPCNN, the top path (i.e., Path1) starts with a 64 filters of size 3×12 convolutional layer, followed by three levels of convolutional layers with 128, 256 and 512 filters of size 3×1, respectively. Similarly, the Path2 starts with a 64 filters of size 18×1 convolutional layer, followed by three levels of convolutional layers with 128, 256 and 512 filters of size 1×1, respectively. In our framework, the output of the fully connected layer is a single real value corresponding to the predictive score of a student. The stride of convolutions is 1. The network can be effectively optimized with the Adam method.

4 Experiments

In this section, we first details the data used in our study, and then introduce two performance measures. Finally, the proposed framework DPCNN are compared with the state-of-the-art methods, and the effects of dual path structure and multi-task learning are studied.

4.1 Data Description

The data used in this paper has been publicly accessed on a national undergraduate big data competition platform. It consists of 13,700,000 smartcard records

of 8199 undergraduate students from 19 majors, which cover the period from September 01, 2013 to August 31, 2015. These data records a large volume of students' campus behaviors, such as paying for meals, entering/exiting the dormitories, and entering the library. These behaviors may occur at 12 different campus places, including cafeterias, campus supermarket, library, dormitory, laundry room, campus bathroom, boiled water room, printing center, academic affairs office, school bus, campus hospital, and card center. Besides, academic performance data is also supplied, which denotes a student's GPA ranking in his or her major. Following the idea of learn-to-rank, the input samples are pairs of student. In the implements, due to limited computational resources, we randomly select approximately 200,000 pairs of student. 70% and the remained 30% are utilized to train and test our model, respectively.

4.2 Evaluation Metrics

Since we follow the idea of learning-to-rank to train our model, we exploit Spearman's rank correlation coefficient [39], which is one of most important ranking-based methods, to evaluate the performance of the proposed method. Spearman coefficient can measure the correlation between the predicted rank and the actual rank, which can be defined as

$$\rho = 1 - \frac{6\sum_{i=1}^{m}(\hat{r}(X_i) - r(X_i))^2}{m(m^2-1)} \tag{2}$$

where m is the number of students under consideration, $\hat{r}(X_i)$ and $r(X_i)$ are the predicted rank and the actual rank of the student X_i, respectively. The higher the Spearman coefficient, the better the prediction performance.

Besides Spearman coefficient, we also care about accuracy of our model. Given a data set $D = \left\{\left(X_i^k, X_j^k, y^k\right)\right\}_{k=1}^{n}$, the accuracy is defined as below.

$$acc = \frac{1}{n}\sum_{k=1}^{n}\mathbb{I}(\hat{y}^k = y^k) \tag{3}$$

Here for predicate π, $\mathbb{I}(\pi)$ equals 1 if π holds and 0 otherwise, $\hat{y}^k$ and y^k are the output label and the actual label, respectively.

4.3 Performance Comparison with State-of-the-Art Methods

In this part, to demonstrate the effectiveness of the proposed network, we compare the DPCNN model against the following two state-of-the-art methods for student performance prediction.

- **RankNet**, a well-known supervised learning to rank algorithm, was utilized to predict the ranking of students with two high-level behavioral characters including orderliness and diligence [8]. The two features were extracted from smartcard records in a manual way. Specifically, the authors calculated

orderliness based on two behaviours: taking showers in dormitories and having meals in cafeterias, and roughly estimated diligence based on two other behaviours: entering/exiting the library and fetching water in teaching buildings. More details can be found in [8].

- **MTLTR-APP** is a multi-task predictive framework based on a learning-to-rank algorithm proposed in [9]. This method took both the difference of majors and the difference of semesters into account, and considered constructing predictive models for students' performance in different semesters as different tasks, even if the students came from the same major. MTLTR-APP can thus capture inter-semester correlation, inter-major correlation with constraints upon model parameters. Three handcrafted behavior features (i.e., orderliness, diligence, and sleep pattern) as well as student similarity, were employed to predict student performance.

Table 1. Comparison of DPCNN with the state-of-the-art methods

Methods	*acc* (Accuracy)	ρ (Spearman coefficient)
RankNet	0.5980	0.2800
MTLTR-APP	0.6012	0.2905
DPCNN	**0.7658**	**0.6964**

We implement the two above methods on our dataset, to demonstrate the effectiveness of the proposed method. It is necessary to mention that in our dataset, the label information is the ranking based on students' accumulated GPA rather than GPA in each semester. Inter-semester correlation is thus discarded when we implement the MTLTR-APP method. Table 1 shows the performance of DPCNN and the two above methods. It can be observed that the proposed DPCNN has the highest accuracy as well as Spearman coefficient. Compared with the RankNet method, the proposed DPCNN further improves the accuracy and Spearman coefficient by an absolute value 16.78% and 41.64%, respectively. Likewise, DPCNN also makes large improvements against MTLTR-APP, i.e. 16.46% on accuracy and 40.59% on Spearman coefficient. The better results demonstrate the proposed dual path network is capable of learning better representation from huge amount of raw smartcard records, compared with handcrafted features used in the RankNet and MTLTR-APP approaches.

4.4 Effect of Dual Path Structure

In order to demonstrate the appealing effectiveness of the dual path architecture, we intentionally design two kinds of networks, which merely owns the top path (i.e., the bottom path is discarded) and the bottom path (i.e., the top path is discarded), respectively. For convenience, we denote them as Single-Path1 network and Single-Path2 network, respectively. The results are reported in Table 2.

Table 2. Comparison of DPCNN with single path networks

Methods	*acc* (Accuracy)	ρ (Spearman coefficient)
Single-Path1 network	0.7540	0.6353
Single-Path2 network	0.7326	0.5990
DPCNN	**0.7658**	**0.6964**

As can be seen from Tabel 2, the DPCNN is obviously superior to both Single-Path1 network and Single-Path2 network on the two evaluation metrics. In particularly, the DPCNN obtains the Spearman coefficient of 0.6964%, which makes large improvements, i.e. 6.11% compared with Single-Path1 network and 9.74% compared with Single-Path2 network. The reason may be that Single-Path1 network merely utilize filters of size 3×12, and it thus only can capture the two behavioral characters including duration and periodicity, i.e., variation is missing. Likewise, Single-Path2 network merely utilize filters of size 18×1, and it thus can capture variation and periodicity, but lose duration. Benefitting from the dual path structure, DPCNN can capture all of the three behavioral characters, and thus make a better prediction.

4.5 Contribution of Multi-task Learning

As aforementioned, we take student performance prediction for different majors as different tasks due to the consistent course settings across majors. To alleviate the issue of data scarcity, we follow the idea of multi-task learning and construct predictive models for multiple majors in a unified framework, i.e., a novel deep network of sharing the convolutional layers (i.e., representation learning layers) while keeping the final three-layer fully-connected networks (i.e., output layers) task-specific. In this part, we care about the benefits from multi-task learning. To this end, we predict student performance with the two following single-task methods. The results are reported in Table 3.

- **Single-task network** constructs predictive models for each major separately without considering the relatedness between tasks. Specifically, we divide the data set into 19 subsets corresponding to 19 majors. Each task owns the whole network DPCNN (i.e., without sharing convolutions), and predictive model for each major is trained one by one. In this method, the average performance is reported.
- **Mixed-data method** views constructing predictive models for all majors as a whole task. In other words, the whole DPCNN framework are shared, and it trains a common model for all majors through mixing data of different majors brutally, i.e., the whole DPCNN framework are shared, without task-specific layers.

As can be seen from Table 3, the DPCNN achieves obviously better performance compared with Single-task network as well as the Mixed-data method.

Table 3. Comparison of DPCNN with single-task methods

Methods	*acc* (Accuracy)	ρ (Spearman coefficient)
Single-task network	0.7507	0.5889
Mixed-data method	0.7259	0.6100
DPCNN	**0.7658**	**0.6964**

Specifically, first, Mixed-data method obtains the worst performance on accuracy, and thus illustrate that it is irrational to train a common model for different majors. Second, the performance of DPCNN makes a great improvement compared with that of Single-task network, i.e., 1.51% for accuracy and 10.75% for Spearman coefficient. The reason may be that DPCNN constructs different predictive models for different majors in a unified framework, and the relatedness between tasks can be implicitly exploited. Third, intuitively, higher accuracy is generally accompanied by a higher Spearman coefficient. Surprisingly, when we compare Single-task network with Mixed-data method, we find that the Single-task network is superior to the Mixed-method on accuracy, while inferior to Mixed-method on the Spearman coefficient. The result may be that in the implements, we sample some pairs of student from each major rather than utilizing all the pairs to construct predictive models.

5 Conclusion and Future Work

In this paper, we predict academic performance based on a large-scale students' behavioral data. Instead of using handcrafted features, we exploit end-to-end deep learning method. To model the three behavioral characters including duration, variation and periodicity, we propose dual path convolutional neural networks. Through dual path convolutions upon student samples, which are represented as tensors, duration and variation can be modeled, respectively. Besides, by taking the date dimension of tensors as the depth of convolutional operations, periodicity can be modeled. Then we introduce multi-tasking learning into our framework, and let multiple tasks share the common convolutional layers while remaining the final three-layer fully-connected networks task-specific. By comparing with two baselines, we show the effectiveness of our proposed DPCNN for predicting academic performance. Moreover, extension experiments illustrate the effectiveness of both dual path structure and multi-task learning.

Though the proposed approach DPCNN can achieve a better presentation automatically as well as better performance, it fails to show the relationship between campus behaviors and academic performance. Thus more deep learning methods, e.g., attention model, can be explored and exploited for our future study. Besides, it should be noted that there exist many other factors affecting student performance, such as psychological status, in-class study behaviors, and historical course grades. Thus, it is also highly appealing to consider more factors to predict student performance in the future.

Acknowledgements. This work was supported by National Natural Science Foundation of China (Grant 61701281, 61573219, 61703234, and 61876098), Shandong Provincial Natural Science Foundation (Grant ZR2017QF009, Grant ZR2016FM34), Shandong Science and Technology Development Plan (Grant J18KA375), Shandong Province Higher Educational Science and Technology Program (Grant J17KA065), and the Fostering Project of Dominant Discipline and Talent Team of Shandong Province Higher Education Institutions.

References

1. Rianne, C., Chris, S., Ad, K., Uwe, M.: Predicting student performance from LMS data: a comparison of 17 blended courses using Moodle LMS. IEEE Trans. Learn. Technol. **10**(1), 17–29 (2017)
2. Huang, S., Fang, N.: Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models. Comput. Educ. **61**, 133–145 (2013)
3. Meier, Y., Xu, J., Atan, O., Schaar, M.V.D.: Predicting grades. IEEE Trans. Signal Proces. **64**(4), 959–972 (2016)
4. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man. Cybern. C **40**(6), 601–618 (2010)
5. Qiujie, L., Rachel, B.: The different relationships between engagement and outcomes across participant subgroups in massive open online courses. Comput. Educ. **127**, 41–65 (2018)
6. Ren Z., Rangwala H., Johri A.: Predicting performance on MOOC assessments using multi-regression models. arXiv preprint arXiv:1605.02269 (2016)
7. Ma, Y.L., Cui, C.R., Nie, X.S., et al.: Pre-course student performance prediction with multi-instance multi-label learning. Sci. China Inf. Sci. **62**(2), 200–205 (2019)
8. Cao, Y., Gao, J., Lian, D., et al.: Orderliness predicts academic performance: behavioural analysis on campus lifestyle. J. Roy. Soc. Interface **15**(146) (2018)
9. Yao, H., Lian, D., Cao, Y., et al.: Predicting academic performance for college students: a campus behavior perspective. ACM Trans. Intel. Syst. Tec. **10**(3), 1–21 (2019)
10. Razavian, A.S., Azizpour, H., Sullivan, J., et al.: CNN features off-the-shelf: an astounding baseline for recognition. arXiv preprint arXiv:1403.6382 (2014)
11. Zhang J., Zheng Y., Qi D.: Deep spatio-temporal residual networks for citywide crowd flows prediction. In: 31st AAAI Proceedings on Artificial Intelligence, pp. 1655–1661. AAAI, San Francisco (2017)
12. Zhang Y., Yang Q.: A survey on multi-task learning. arXiv preprint arXiv:1707.08114 (2017)
13. Wang, F., Chen, L.: A nonlinear state space model for identifying at-risk students in open online courses. In: 9th International Proceedings on Educational Data Mining, Raleigh, NC, USA, pp. 527–532 (2016)
14. Li, W., Gao, M., Li, H., Xiong, Q.Y., et al.: Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning. In: International Proceedings on Neural Networks, pp. 3130–3137. IEEE, Vancouver (2016)
15. He, J.Z., Bailey, J., Rubinstein, B., Zhang, R.: Identifying at-risk students in massive open online courses. In: 29th AAAI Proceedings on Artificial Intelligence, pp. 1749–1755. AAAI, Austin (2015)

16. Mi, F., Dit-Yan, Y.: Temporal models for predicting student dropout in massive open online courses. In: 2015 IEEE International Proceedings on Data Mining Workshop, pp. 256–263. IEEE, Atlantic City (2015)
17. Kim, B.H., Vizitei, E., Ganapathi, V.: GritNet: student performance prediction with deep learning. In: 11st International Proceedings on Educational Data Mining, Buffalo, NY, USA, pp. 625–629 (2018)
18. Trivedi, S., Pardos, Z.A., Heffernan, N.T.: Clustering students to generate an ensemble to improve standard test score predictions. In: International Conference on Artificial Intelligence in Education, Christchurch, New Zealand, pp. 377–384 (2011)
19. Thai-Nghe, N., Schmidt-Thieme, L.: Multi-relational factorization models for student modeling in intelligent tutoring systems. In: 7th International Conference on Knowledge and Systems Engineering. IEEE, Chongqing (2015)
20. Suleyman, C., Luo, S., Yan, P.X., Ron, T.: Probabilistic latent class models for predicting student performance. In: International Conference on Information and Knowledge Management, pp. 1513–1516. ACM, San Francisco (2013)
21. Er, E.: Identifying at-risk students using machine learning techniques: a case study with is 100. Int. J. Mach. Learn. Comput. **2**(4), 476–480 (2012)
22. Hu, Y.H., Lo, C.L., Shih, S.P.: Developing early warning systems to predict students online learning performance. Comput. Hum. Behav. **36**, 469–478 (2014)
23. Macfadyen, L.P., Dawson, S.: Mining lms data to develop an early warning system for educators: a proof of concept. Comput. Educ. **54**(2), 588–599 (2010)
24. Zafra, A., Romero, C., Ventura, S.: Multiple instance learning for classifying students in learning management systems. Expert Syst. Appl. **38**(12), 15020–15031 (2011)
25. Kotsiantis, S.B., Pierrakeas, C.J., Pintelas, P.E.: Preventing student dropout in distance learning using machine learning techniques. Appl. Artif. Intell. **18**(5), 411–426 (2004)
26. Xenos, M.: Prediction and assessment of student behaviour in open and distance education in computers using bayesian networks. Comput. Educ. **43**(4), 345–359 (2004)
27. Wang, A.Y., Newlin, M.H., Tucker, T.L.: A discourse analysis of online classroom chats: predictors of cyber-student performance. Teach. Psychol. **28**(3), 222–226 (2001)
28. Wang, A.Y., Newlin, M.H.: Predictors of performance in the virtual classroom: identifying and helping at-risk cyber-students. J. High Educ. **29**(10), 21–25 (2002)
29. Essa, A., Ayad, H.: Student success system: risk analytics and data visualization using ensembles of predictive models. In: 2nd Proceedings on Learning Analytics and Knowledge, Vancouver BC, Canada, pp. 158–161 (2012)
30. Lopez, M.I., Luna, J.M., Romero, C., Ventura, S.: Classification via clustering for predicting final marks based on student participation in forums. JEDM **4** (2012)
31. Wu, R.Z., Liu, Q., Liu, Y.P., et al.: Cognitive modelling for predicting examinee performance. In: 24th Proceedings of the International Joint Conference on Artificial Intelligence, pp. 1017–1024. AAAI Press, Buenos Aires (2015)
32. Marbouti, F., Diefes-Dux, H.A., Madhavan, K.: Models for early prediction of at-risk students in a course using standards-based grading. Comput. Educ. **103**, 1–15 (2016)
33. Kimberly, E.A., Matthew, D.P.: Course signals at purdue: using learning analytics to increase student success. In: 2nd Proceedings on Learning Analytics and Knowledge, pp. 267–270. ACM, Vancouver (2012)

34. Ashay, T., Shajith, I., Bikram, S., et al.: Predicting student risks through longitudinal analysis. In: 20th Proceedings of International Conference on Knowledge Discovery and Data Mining, pp. 1544–1552. ACM, New York (2014)
35. Gedeon, T.D., Turner, S.: Explaining student grades predicted by a neural network. In: Proceedings of International Joint Conference on Neural Networks, Nagoya, pp. 609–612 (2002)
36. Acharya, A., Sinha, D.: Early prediction of students performance using machine learning techniques. Int. J. Comput. Appl. **107**(1), 37–43 (2014)
37. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. **35**(8), 1798–1828 (2013)
38. Ruder S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098v1 (2017)
39. Spearman, C.: The proof and measurement of association between two things. Am. J. Psychol. **100**(3/4), 441–471 (1987)