A neural network regression model for predicting student learning success based on prior achievements

Mikhail Dorrer*

Siberian State University of Science and Technology named after M.F. Reshetnev, Krasnoyarsk, Russia

Abstract. The paper describes a project utilizing data analysis tools to predict student performance based on their prior achievements. The task was addressed using historical educational data from over 35,000 students over a span of seven years, containing information on 1.24 million grades. Neural network regression tools were employed to build models that predict future grades, thereby enhancing educational processes. The predictive capability of the model was assessed using the coefficient of determination and the root mean square error (RMSE) through 10-fold cross-validation of the dataset into training and testing sets. More than 70% of the developed grade prediction models demonstrated a coefficient of determination greater than 0.7, with the RMSE of predicted grades from actual values being less than one point on a five-point scale. This indicates a satisfactory solution to the prediction problem.

1 Introduction

Student Performance Prediction (SPP) [1] aims to forecast the grades a student will receive prior to enrollment in a course or taking an exam [2]. This task is crucial for a personalized approach to education and is garnering increasing attention in the fields of artificial intelligence and Educational Data Mining (EDM) [3].

Many organizations in the academic sector recognize the importance of leveraging the potential of digital transformation to enhance the educational process [4]. Consequently, universities must refine their structures, processes, and educational approaches to align with societal needs, thereby educating students in accordance with these demands [5].

It is noteworthy that the current state of digitalization in business process management is characterized by a range of issues, including excessive complexity and labor-intensive procedures [6], as well as a disconnect between analytical phases across various tools.

In response to these challenges, the technology of Digital Twins (DT) is being developed within the framework of the Fourth Industrial Revolution [7]. This issue is also acknowledged at Siberian State University named after M.F. Reshetnev, where, as part of the strategic academic leadership program "Priority-2030", a project titled "Implementation of Data

* Corresponding author: dorrer_mg@sibsau.ru

Analysis and Machine Learning Tools in the Management System of Core and Auxiliary Business Processes of the University" is being executed.

The experimental data utilized consisted of historical performance records from Siberian State University named after M.F. Reshetnev, encompassing approximately 1.24 million grade entries for 35,000 students over the period from 2017 to 2024, representing 13 institutes and faculties within the university.

This paper presents a description of the results obtained during the first phase of this project - addressing the task of predicting student performance across various disciplines of educational programs by constructing a machine learning model that enables forecasting subsequent grades based on prior marks and certain prerequisites of the students.

2 Related works

Educational Data Mining (EDM) is a relatively young field of research that focuses on uncovering hidden patterns within various data related to the educational process. Key areas of interest include the analysis of student knowledge [2], the analysis of student behavior during learning [3], teacher planning of curricula [8], and the organization of course scheduling [9]. All studies in this domain address a common objective: improving student performance [10], while also achieving additional goals, such as reducing the costs associated with the educational process [11].

As a result of this focus over recent decades, there has been a significant number of studies concentrated on predicting student performance (SPP). Among these works is [12]. An alternative approach to addressing this task involves assessing students based on their final academic grades, as presented in [13].

For students, SPP can assist in selecting suitable courses or exercises and in formulating individualized learning plans for academic periods [14]. For educators, SPP can help adjust teaching materials and curricula based on student capabilities and identify at-risk students [15]. For the administration of educational institutions, SPP can aid in evaluating the effectiveness of curricula and optimizing course offerings [8].

A trend in the digitalization of business processes is the development of digital twins for organizations, organizational projects, and processes. For instance, the work [16] is directly dedicated to creating a digital twin of the educational process in engineering training.

Implementing a digital twin of an entity necessitates the development of an adequate mathematical model for that entity. As noted in [17], there is a clear deficiency of research dedicated to quantitative mathematical modeling of business processes. This article builds upon the results of previous works, where the author examines the issues surrounding the creation of digital twins for business processes at various levels of detail - from simulation models of discrete-event systems to models of organizational maturity level dynamics.

In this paper, we will explore an approach to creating a system for collecting the digital shadow of the educational process and the predictive component of the digital twin of the educational process within higher education as a discrete business process.

3 Applied methods

3.1 Object of analysis

The object of analysis and forecasting in the digital twin of a business process is a discrete process (containing a finite countable number of states - stages of the product/service life cycle) described by typical scenarios executed on a set of operations with a collection of organizational and material resources.

In this article, the educational process of a higher education institution serves as the business process under investigation. One semester of study will be considered as a stage of the process. The scenario for executing this stage of the business process is described by the semester curriculum. Scenarios are applied to unique instances - batches of products, clients, etc. In the case of a university, the unique instance is the student.

The indicators (metrics) of the process outcome are represented by the performance metrics of the product (service) that is the output of each stage of the life cycle. In the context of a university, such metrics are the results of student performance assessments, specifically the grades awarded to students in subjects at the end of the semester. It is important to note that this modeling approach complicates the forecasting of results using time series forecasting methods. This is due to the fact that while the indicators of various stages characterize the same process, they are associated with different outcomes; thus, the set of metrics at different stages may vary, and even their dimensionality does not necessarily have to coincide from stage to stage.

3.2 Forecasting task

For a process belonging to the aforementioned class, it is necessary to predict the values of process performance metrics across its stages. The basis for prediction consists of the indicators of the prerequisites of the process (quality of raw materials, client status); for the student, these will be their prerequisites - academic achievements prior to entering the university, information about passing the Unified State Exam, successes in subject Olympiads, etc. - as well as the indicators from previous stages of the process.

3.3 Input data of the process

As a source of information regarding the state of the process, a measurement database of process indicators is utilized, which represents the results of tracking process indicators over time and includes the following set of fields:

- Measurement date
- Responsible unit for the process (Institute/Faculty)
- Process variant (field of study)
- Process instance (student, identifiable by their academic record number)
- Process stage (in the university—semester of study)
- Process metric
- Unit of measurement
- Metric value
- Unit responsible for the indicator (department)
- Employee responsible for the indicator (subject instructor).

3.4 Forecasting idea

It is necessary to develop a model that ensures the prediction of process indicator values across its stages. To achieve this, a regression model is constructed for each instance of the process, which forecasts the specified metric in given units of measurement at a specified stage of the process based on input data (prerequisites) for the instance and process indicators from previous stages for that instance.

Let us denote the process variant number as $v \in [1..V]$, where V is the total number of process variants. Let us denote the process stage as $t_v \in [0..T_v]$, where T_v is the number of stages in this variant. Stages are numbered as $[1..T_v]$, with the zero index corresponding to the prerequisites of the process.

Let us denote the metric number for the process as $m_{t_v} \in [1..M_{t_v}]$, where M_{t_v} is the number of metrics describing the process variant at its stage.

Thus, the set of process indicators can be characterized by the set X, where an element represents a value $x_{t_v}^m$, describing the m-th indicator of process v at stage t. We also introduce vector $\bar{X}_{t_v}$, which contains all metric values for the process at stage t_v.

The forecasting task then consists of forming models that allow for predicting indicators of the next stage of the process based on indicators from previous stages and prerequisites, that is, identifying function $F_{t_v}^m$, which performs the calculation $\widehat{x_{t_v}^m}$ – the forecasted value of the m-th indicator at the t-th stage of the v-th variant of the process:

$$\widehat{x_{t_v}^m} = F_{t_v}^m(\bar{X}_{t_v-1}, \bar{X}_{0_v}) \quad (1)$$

This task is framed as a machine learning problem, which translates into selecting a regression model and parameters of the regression model that minimize the error functional (or maximize the coefficient of determination) across the well-known set of experimental data pertaining to the process.

3.5 Formulation of the machine learning task

The task is reduced to selecting a regression model and optimizing the model parameters to minimize the error functional or maximize the coefficient of determination on a known set of experimental data related to the process.

Formal Problem Statement:

- Input Data: A set of historical data consisting of vectors $\bar{X}_{t_v-1}, \bar{X}_{0_v}$.
- Output Data: Forecasted values $\widehat{x_{t_v}^m}$.
- Objective Function: Minimization of prediction error

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} (x_{t_v}^m - F_{t_v}^m(\bar{X}_{t_v-1}, \bar{X}_{0_v}))^2 \quad (2)$$

where $x_{t_v}^m$ represents the true values of metrics at stage t_v. The index i refers to the number of a specific instance.

• Additional Quality Metric for Forecasting: Coefficient of Determination (Measure R^2):

$$R^2 = 1 - \frac{SS_{\text{res}}}{S_{\text{tot}}} \quad (3)$$

where $SS_{\text{res}} = \sum_{i=1}^{N} (x_{t_v}^m - \widehat{x_{t_v}^m})^2$ – the sum of squared residuals, and $SS_{\text{tot}} = \sum_{\text{i=1}}^{\text{N}} (x_{t_v}^m - \overline{x_{t_v}^m})^2$ – the total sum of squares, where $\overline{x_{t_v}^m}$ is the mean of the true values of metrics.

Taking into account formulas (2) and (3), the machine learning task is formulated as the minimization of the residual measure, where it is required to find such model parameters that minimize the error functional:

$$min_\theta J(\theta) \quad (4)$$

As an additional condition, the maximization of the coefficient of determination is considered, for which it is required to find such model parameters that maximize the coefficient of determination:

$$max_\theta R^2 \quad (5)$$

Thus, for the developed digital twin of the business process, the machine learning task is reduced to selecting a regression model and parameters that ensure minimization of the error functional and maximization of the coefficient of determination on a known set of experimental data related to the process.

3.6 Neural Network Regression Model

In this work, we utilize a neural network architecture designed to address the regression task. The model is implemented using the Keras library and consists of three fully connected layers.

In the notation of formula (1), $F(\bar{X})$ is the function that represents the output of the neural network for the input vector $\bar{X}$. It can be expressed as a sequence of operations performed on the input data.

1. The first layer takes the input (for our digital twin, this consists of vectors $\bar{X}_{t_v-1}$, $\bar{X}_{0_v}$), applies a weight matrix and a bias, and then uses the hyperbolic tangent activation function $tanh$(x).
2. The second layer performs the same operations, taking the output from the first layer.
3. The third layer simply applies a linear activation.

Formally, this can be expressed as:

$$F(x) = g_3(g_2(g_1(\bar{X}_{t_v-1}, \bar{X}_{0_v}))) \quad (6)$$

where:

$$g_1(\bar{X}_{t_v-1}, \bar{X}_{0_v}) = th(W_1\,(\bar{X}_{t_v-1}, \bar{X}_{0_v}) + \bar{b}_1)$$
$$g_2(\bar{X}_{t_v-1}, \bar{X}_{0_v}) = th(W_2\, g_1(\bar{X}_{t_v-1}, \bar{X}_{0_v}) + \bar{b}_2)$$
$$g_3(\bar{X}_{t_v-1}, \bar{X}_{0_v}) = W_3\, g_2(\bar{X}_{t_v-1}, \bar{X}_{0_v}) + \bar{b}_3$$

here:

W_1, W_2, W_3 are the weight matrices of the corresponding layers,

$\overline{b_1}, \overline{b_2}, \overline{b_3}$ are the bias vectors of the neurons in the corresponding layers.

Based on expression (6), the parameter space for which extrema are sought for expressions (4) and (5) includes the matrices W_1, W_2, W_3 and the vectors $\overline{b_1}, \overline{b_2}, \overline{b_3}$.

To optimize the model parameters, we employ the Adam (Adaptive Moment Estimation) algorithm. The mean squared error is selected as the loss function, which is a standard choice for regression tasks.

At this stage of the research, the task of structural optimization of function F is not set; however, it could be formulated in principle and provided that sufficient computational resources are available.

4 Implementation

4.1 Implementation tools

In the development of the program for predicting student grades, a combination of popular libraries in Python was utilized, allowing for efficient model building and evaluation of their quality.

- For data processing and analysis – the Pandas library.
- For numerical calculations – the NumPy library.
- For supporting the construction and evaluation of machine learning models, the Scikit-learn library was chosen, specifically:
    - For standardizing feature values – StandardScaler;
    - For assessing the effectiveness of various models – metrics r2score and meansquared_error;
    - For reliable evaluation of the predictive capability of models – the KFold cross-validation method.
- For creating neural network models – the TensorFlow library with the Keras interface, using the Sequential class for layered neural networks.

4.2 Data collection

For the purposes of the project described in this article, the source data is derived from the databases of the corporate information system 1C:University Prof. The following registries are used:

- Certification (since 2017, 1.2 million records) – information about grades.
- Certification MRTO (since 2017, 1.2 million records) – information about intermediate grades of students according to the modular-rating education system
- Assigned calculations – information about assigned scholarships.
- Achievements of applicants (since 2020, 85 thousand records) – information about applicants upon admission, including data on Unified State Exam scores.

The data was extracted from the 1C:University Prof LMS-system and subsequently processed using programs implemented in Python. During data loading, all grades, including pass/fail assessments, were transformed to a five-point scale.

4.3 Data preparation for forecasting

The historical values for analysis are collected within the framework of a single variant of the process. The data is grouped into N_{i_k+1} subsets D_j, where D_{0_k} represents the indicators of prerequisites for the k-th variant of the process, and $D_{1_k}, \ldots, D_{T_k}$ denote the values of indicators for the corresponding stage of the k-th variant of the process. The result of this grouping is a summary table, referred to as "MetricsVariantProcess_Stage" which has instances of the process in rows and the names of process metrics in columns. The cells contain, in cases of multiple metric values at a given stage, the maximum value among them; if there is only one metric value at a given stage, that value itself is recorded; and if there are no values, it displays NaN (not a number).

4.4 Formation of the forecasting model

The following algorithm is employed for the formation of the forecasting model:

1. Selection of Initial Data: Division (institute), variant, and sub-variant of the process (specialty and form of education).
2. For the selected process, all datasets with indicators by stages are selected.
3. The prerequisites of students are identified (gender, age, admission achievements).
4. Machine learning models are formed for each assessment in each semester. This involves:
    a. Normalization of all column values using the StandardScaler normalizer, which adjusts the mean to 0 and the variance to 1, while retaining the normalizer object for reverse transformation of results back to the natural scale.
    b. In a loop through semester numbers from 1 to the maximum semester encountered while reading datasets for the program, a set of machine learning models is constructed for each semester as neural networks of the type specified in formula (6), predicting each column of grades (process metrics) by using as input the grades from the previous semester as well as the prerequisites of all students enrolled in that semester.
5. For the first semester, only prerequisites are utilized.
6. For each model, metrics (R^2 and MSE) are evaluated through averaging results with 10-fold cross-validation.

5 Results and discussion

This section presents the results of constructing and evaluating forecasting models for process indicators (student grades), based on data collected from students of the Institute of Informatics and Telecommunications (IITC) at Siberian State University named after M.F. Reshetnev, who are enrolled in bachelor's programs in the field of 09.03.04 "Software Engineering" in full-time education.

The volume of datasets corresponding to the number of students who have gone through this program from 2017 to 2024 is shown in Table 1.

Table 1. Volume of samples for solving the forecasting task.

Semester	Number of records	Number of columns (grades)
1	593	53
2	583	64
3	451	50
4	405	70
5	297	74
6	274	67
7	170	77
8	169	46

Table 2 presents the results of predicting grades for subjects in the curricula, broken down by semesters (using the first semester as an example) and indicating the quality metrics of the forecasts.

Table 2. Results of grade forecasting

Indicator	R^2	MSE
Algebra and Analytical Geometry	0.67	0.78
Introduction to Discrete Mathematics	0.00	0.00
Introduction to Software Engineering	0.00	0.00
Engineering and Computer Graphics	0.58	0.92
Foreign Language	0.06	0.58
Computer Science	0.00	0.00
Computer Science and Programming	0.58	2.41
Information Technologies in the Digital Economy	0.57	0.59
History	0.00	0.00
Mathematical Analysis	0.61	0.78
Organization and Technology of Document Support for Software Products	0.00	0.00
Fundamentals of Software Engineering	0.37	0.75
Fundamentals of Russian Statehood	0.93	0.20
Professionally Applied Physical Culture	0.00	0.00
Russian Language and Culture of Speech	0.24	0.46
Physical Culture and Sports	0.14	0.59
Programming Languages	0.52	1.02

It can be observed that a significant portion of the models is characterized by an acceptable value of the coefficient of determination – greater than 0.5 (to be analyzed in detail below) and forecasting accuracy – with a mean squared error of prediction of less than one point on a five-point scale.

The averaged results of grade forecasting within a semester are presented in Table 3.

Table 3. Average quality metrics of forecasting.

Semester	Average MSE	Average R^2
1	0.53	0.31
2	0.62	0.40
3	0.58	0.39
4	1.04	0.30
5	2.03	-0.17
6	0.57	0.25
7	0.96	0.32
8	0.86	0.25
Overall Result:	0.95	0.24

The averaged values demonstrate, on average, a rather weak predictive capacity of the constructed models according to the R^2 metric, but also a small magnitude of the mean squared error estimate. The forecast for interim assessments has been presented in a separate table due to the different scales of output values (grades – on a five-point scale, assessments – on a hundred-point scale).

Table 4. Results of Forecasting Assessments.

Indicator	R^2	MSE
Algebra and Analytical Geometry_Assessment1	0,71	22.10
Algebra and Analytical Geometry_Assessment2	0.69	105.72
Algebra and Analytical Geometry_Assessment3	0.70	258.35
Engineering and Computer Graphics_Assessment1	0.67	31.95
Engineering and Computer Graphics_Assessment2	0.65	138.77
Engineering and Computer Graphics_Assessment3	0.67	334.67
Foreign Language_Assessment1	0.72	26.46
Foreign Language_Assessment2	0.70	113.05
Foreign Language_Assessment3	0.64	323.55
Computer Science and Programming_Assessment1	0.66	35.05
Computer Science and Programming_Assessment2	0.63	134.48
Computer Science and Programming_Assessment3	0.59	455.74
Information Technologies in the Digital Economy_Assessment1	0.74	27.28
Information Technologies in the Digital Economy_Assessment2	0.72	111.47
Information Technologies in the Digital Economy_Assessment3	0.70	315.15
Mathematical Analysis_Assessment1	0.73	23.49
Mathematical Analysis_Assessment2	0.72	90.81
Mathematical Analysis_Assessment3	0.71	226.56
Fundamentals of Software Engineering_Assessment1	0.65	42.82
Fundamentals of Software Engineering_Assessment2	0.63	169.28
Fundamentals of Software Engineering_Assessment3	0.66	349.70
Foundations of Russian Statehood_Assessment1	0.90	6.15
Foundations of Russian Statehood_Assessment2	0.88	23.56
Foundations of Russian Statehood_Assessment3	0.91	64.50
Russian Language and Culture of Speech_Assessment1	0.74	14.66
Russian Language and Culture of Speech_Assessment2	0.74	65.84
Russian Language and Culture of Speech_Assessment3	0.65	238.17
Physical Culture and Sports_Assessment1	0.77	18.58
Physical Culture and Sports_Assessment2	0.74	82.19
Physical Culture and Sports_Assessment3	0.66	238.42
Programming Languages_Assessment1	0.69	36.95
Programming Languages_Assessment2	0.66	151.95
Programming Languages_Assessment3	0.62	407.72

Table 4 provides the results of forecasting interim assessments for subjects in the first semester of the analyzed training direction. It can be observed that the models for forecasting assessment results for the first semester possess a sufficiently high predictive capacity - all models exhibit an average value of the coefficient of determination above 0.6. The accuracy of the forecasts, as assessed by the mean squared error, varies significantly - from 18.58 to 455.74, which corresponds to a range of approximately 4.31 to 21.35 on a natural hundred-point scale. The averaged results of forecasting interim assessments by discipline within the semester are presented in Table 5.

Table 5. Average Quality Metrics of Forecasting Interim Assessments.

Semester	Average R^2	Average MSE
1	0.70	141.97
2	0.70	146.26
3	0.52	175.25
4	0.45	276.12
5	0.13	330.39
6	0.40	227.99
7	0.19	288.61
8	-0.09	435.51
Overall Result:	0.38	250.51

It can be observed that the average coefficients of determination for the models across three out of eight semesters correspond to an acceptable predictive capacity (greater than 0.5), while the situation is less favorable for the remaining semesters, which will be discussed further.

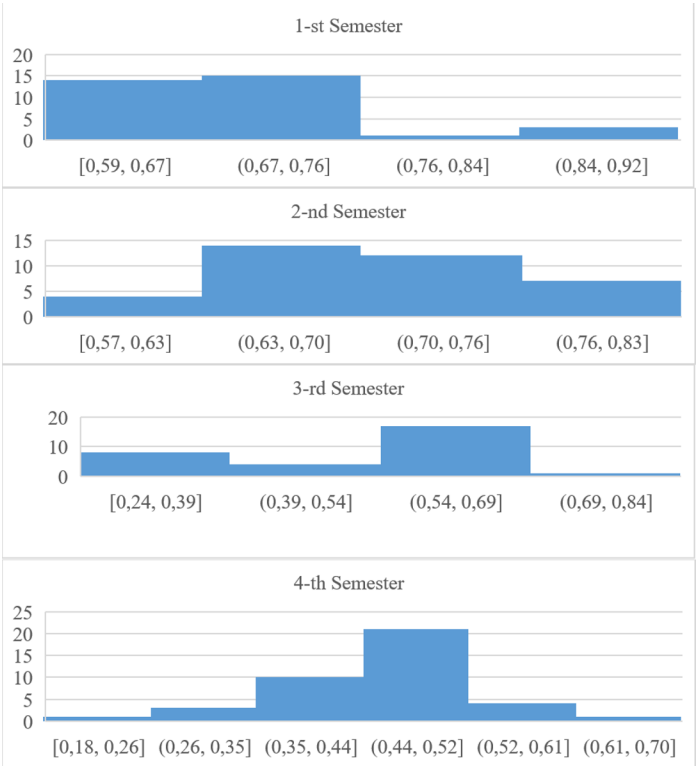

Fig. 1. Histograms of the distribution of frequencies for the coefficients of determination for models predicting grades by semester.

Figure 1 presents histograms of frequencies based on the magnitude of the coefficient of determination (R^2 measure) for the models predicting the results of interim assessments, broken down by semester.

It can be seen that even for semesters with an unfavorable picture regarding the average value of the coefficient of determination, the proportion of models with acceptable values of this measure exceeds 50%. The low average value is attributed to models that do not explain the variance of the target variable at all, having a coefficient of determination equal to zero or even negative.

6 Conclusion

The calculations conducted show that the selected machine learning model for the available initial data (students' prerequisites and their academic performance) demonstrated acceptable predictive capacity when forecasting a significant portion (from 50 to 100%) of academic grades within the chosen higher education program.

However, a considerable number of grades are predicted unsatisfactorily by models of the same type, reaching even negative values for the coefficient of determination. This situation necessitates analysis in several directions:

1. A substantive analysis of the grading process for the subject and the values of interim assessments. The subjective factor in grading can render predictions impossible.
2. An analysis of the process of forming machine learning models and their architecture. Here, the evaluation and interpretation of the significance of input parameters for each model are of primary interest. Excluding redundant inputs may enhance the predictive capability of the models, especially those that have certain but insufficient predictive power. Furthermore, assessing the impact of factors can be utilized to formulate recommendations for improving student performance.

The proposed solution in this work has several important avenues for development:

- Development of a model predicting events in a business process (using the educational process as an example), such as student expulsions before the completion of their studies.
- Increasing the accuracy of models through optimization of the set of input parameters.
- Implementation of an active part of the digital twin – functionalities for generating recommendations that influence the business process, ensuring the required changes in target metrics as desired by the process owner.

References

1. Y. Zhang, Y. Yun, R. An, J. Cui, H. Dai, и X. Shang, Front. Psychol. **12**, 1 (2021)
2. C.-K. Yeung и D.-Y. Yeung, in *Proc. Fifth Annu. ACM Conf. Learn. Scale* (ACM, New York, NY, USA, 2018), pp. 1–10
3. L. Juhaňák, J. Zounek, L. Rohlíková, Comput. Human Behav. **92**, 496-506 (2019)
4. Á. López-Gracia, T. González-Ramírez, и J. De Pablos-Pons, Profesorado, Rev. Currículum y Form. del Profr. **26**, 75-90 (2022)
5. M. Romero, T. Romeu, M. Guitert, P. Baztán, RIED-Revista Iberoam. Educ. a Distancia **26**, 163-182 (2022)
6. M. Barnett, in *BP Trends Newsletter, White Pap. Tech. Briefs* (2003), pp. 1–10
7. H. van der Valk, H. Haße, F. Möller, M. Arbter, J. L. Henning, и B. Otto, в *26th Am. Conf. Inf. Syst. AMCIS 2020* (2020), pp. 1–10

8. B. Reeves, *Development of rubrics to support teacher judgement of student proficiency in ethical Decision-Making* (Melbourne Graduate School of Education, 2018)
9. H. Zhang, T. Huang, Z. Lv, S. Liu, Z. Zhou, Multimed. Tools Appl. **77**, 7051-7075 (2018)
10. Q. Liu, R. Wu, E. Chen, G. Xu, Y. Su, Z. Chen, G. Hu, ACM Trans. Intell. Syst. Technol. **9**, 1 (2018)
11. T. J. Gronberg, D. W. Jansen, L. L. Taylor, K. Booker, Texas A\&M Univ. Coll. Station. TX. http//www. Sch. info/states/tx/march4\% 20cost\% 20study. pdf (2004)
12. S. Morsy, G. Karypis, in *Proc. 2017 SIAM Int. Conf. Data Min.* (2017), pp. 552–560
13. M. A. Al-Barrak, M. Al-Razgan, Int. J. Inf. Educ. Technol. **6**, 528-533 (2016)
14. Z. Ibrahim, D. Rusli, in *21st Annu. SAS Malaysia Forum, 5th Sept.* (Kuala Lumpur, 2007), pp. 1–6
15. M. Kloft, F. Stiehler, Z. Zheng, N. Pinkwart, in *Proc. EMNLP 2014 Work. Anal. large scale Soc. Interact. MOOCs* (2014), pp. 60–65
16. S. N. Masaev, A. N. Minkin, E. Yu Troyak, A. L. Khrulkevich, J. Phys. Conf. Ser. **1889**, 022045 (2021)
17. K. Vergidis, A. Tiwari, B. Majeed, IEEE Trans. Syst. Man, Cybern. Part C Applications Rev. **38**, 69-84 (2008)