Predicting Academic Performance: A Bootstrapping Approach for Learning Dynamic Bayesian Networks

Mashael Al-Luhaybi(✉), Leila Yousefi, Stephen Swift, Steve Counsell, and Allan Tucker

Intelligent Data Analysis Laboratory, Brunel University London, London, UK
{Mashael.Al-luhaybi,Leila.yousefi,stephen.swift,steve.counsell,
Allan.Tucker}@brunel.ac.uk

Abstract. Predicting academic performance requires utilization of student related data and the accurate identification of the key issues regarding such data can enhance the prediction process. In this paper, we proposed a bootstrapped resampling approach for predicting the academic performance of university students using probabilistic modeling taking into consideration the bias issue of educational datasets. We include in this investigation students' data at admission level, Year 1 and Year 2, respectively. For the purpose of modeling academic performance, we first address the imbalanced time series of educational datasets with a resampling method using bootstrap aggregating (bagging). We then ascertain the Bayesian network structure from the resampled dataset to compare the efficiency of our proposed approach with the original data approach. Hence, one interesting outcome was that of learning and testing the Bayesian model from the bootstrapped time series data. The prediction results were improved dramatically, especially for the minority class, which was for identifying the high risk of failing students.

Keywords: Performance prediction · Bayesian networks · Resampling · Bootstrapping · EDM

1 Introduction

According to the Higher Education Statistics Agency (HESA) in the UK [1], the drop-out rate among undergraduate students has increased in the last three years. The statistics published by the HESA reveal that a total of 26,000 students in England in 2015 dropped out from their enrolled academic programmes after their first year. Also, the statistics show that the higher education (HE) qualifications obtained by students for all levels, including undergraduate and postgraduate levels, decreased from 788,355 in 2012/13 to 757,300 in 2016/17.

Supported by the Intelligent Data Analysis Research Laboratory, Brunel University London, United Kingdom.

S. Isotani et al. (Eds.): AIED 2019, LNAI 11625, pp. 26–36, 2019.
https://doi.org/10.1007/978-3-030-23204-7_3

The growing availability of such reports provides new opportunities to educational researchers. A large body of research has investigated the issues associated with students' learning and academic performance. For instance, educational data mining (EDM) researchers have attempted to analyse and evaluate student data to enhance their education and provide solutions to failure issues using the state-of-the-art Artificial Intelligence (AI) methods.

Predicting student performance is a major area of interest within the field of EDM in terms of ascertaining accurately what, as yet, unknown knowledge regarding this performance, such as final grades [2], will transpire to be. However, it is a very difficult task as it is influenced by social, environmental and behavioral factors [3,4]. Thus, machine learning algorithms are increasingly being used to discover the relationships between these factors and the academic performance of students. There are different educational predictive models for student assistance aimed at helping them to achieve an improvement in their studies. What is interesting about these models is that they model students' achievement by using Bayesian networks (BNs) to handle uncertainty as well as representing the student's knowledge. For instance, [5] used Dynamic Bayesian networks (DBNs) to analyze students' cognitive structure over time. BNs [6] involve a classification approach based on probability theory [7] and are considered the best predictors. Such probability predicts the membership of all student-related factors and the class factor by assuming that the independency of the latter is based on the associated values with the other attributes in the prediction model [8].

However, from a practical point of view, a common issue with classifying students is that the educational datasets usually contain imbalanced data, especially for high risk or failed students compared to the excellent or medium performance ones. Because of this, we exploited a resampling method on students' obtained grades and other students related attributes, namely bootstrapping, to ensure that more states of student overall performance are obtained than using the original time series datasets.

This paper provides, first, a novel DBN approach for predicting university students' academic performance from time series educational data using a probabilistic modeling approach. Secondly, it explores the use of a bootstrapping method to resample the educational datasets in order to improve the learning of the BN structure, whilst also enhancing the detection of the students of the minority class, who are those at high risk, as early as possible.

2 Related Work

A considerable number of studies have been conducted to predict the performance of the students based on the Bayesian method. For instance, [8] compared the Bayesian approach with other classification approaches to identify useful patterns that could be extracted from students' personal and pre-university data in order to predict students' performance at the university. Similarly, authors of [9] used the same approach for a comparative study of classification algorithms but their aim was to classify the students and identify the most influencing attributes

on students' failure. Though some researchers were attempting to use the BNs for characterizing performance and exploring the correlation between the performance and the attributes, others were interested in detecting and modeling students' learning styles [10,11]. For example, [12] conducted a study to analyze demographic, social and assessment data to predict the slow learning students in order to improve their performance and reduce failure rate prior to the exam.

However, there is not much-related work in the educational system that handles the issues with the imbalanced educational data through exploiting the bootstrap approach [13,14]. Feng and co-authors [15] utilized and validated their statistical results by using bootstrapping with logistic regression to evaluate students' learning based on different educational interventions. Similarly, a study has been conducted by [16] to evaluate students' understanding of statistical inference with a bootstrapping approach while did not consider time. To the best of our knowledge, there is no previous work in this field that applied DBNs on the bootstrapped "time series" dataset of students progression data. Hence, this work is a first attempt to use them, with the aim of achieving an improvement in student performance overall.

3 Method

As with many prediction issues, educational datasets usually include imbalanced data, because the number of students in Low, Medium and High risk classes is not equally balanced (see Fig. 5(A)). We focus here on first predicting students performance at university based on the original educational time series records. Then, we investigate our resampling approach in order to have some insights into the current problems with the imbalanced educational time series records. Hence, the issue of predicting students' performance based on imbalanced data can be determined using our resampling approach. For this purpose, a DBN model was learned from student's temporal data, taking into consideration the imbalance issue of the predictive classes (see Fig. 1).

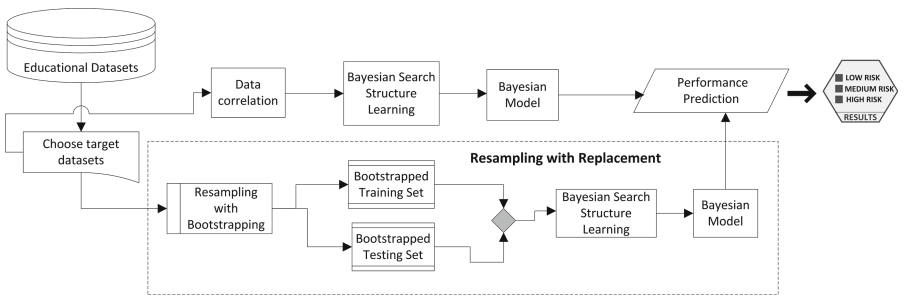

Fig. 1. This diagram presents Bayesian structure learning and the resampling strategy used for learning the DBN model.

3.1 Datasets

The datasets used in this paper were collected from Brunel University admissions and computer science databases. This consists of 377 records for students' progression and other student-related data in 2014, 2015 and 2016, respectively. The datasets contained the following data categories:

- **Admissions dataset:** includes students' application data when entering the university. such as: nationality, ethnicity, country of birth, disability, been in care, socioeconomic class ...etc.;
- **Progression dataset:** this includes student final grades for all Year 1 and Year 2 modules at the university for measuring students' overall academic performance;
- **Engagement attitude:** this includes students' attitude towards turning up to classes and labs in Year 1 and Year 2 at the university;
- **Online temporal assessment profiles:** this includes a student online assessment profile based on their online time-series assessment trajectories. These profiles were obtained using dynamic time warping (DTW) and hierarchical clustering algorithms.

3.2 Resampling with Bootstrapping

Resampling strategies are fundamental approaches in the pre-processing phase, which are used to change the distribution of data in a dataset [17]. After we had discretised the students' overall grade bands from (A, B, C, D, E and F) to qualitative states of low, medium and high risk students, we still encountered an imbalance issue especially for the high risk students (see the confusion matrix in Fig. 5(A)). As aforementioned, imbalance data is a very common issue in educational datasets, which affects learning the predictive models as well as making difficulties in identifying the cases of the minority classes. The minority class in this work is the high risk class, which is the class assigned for those students who obtained low grades (D, E and F) in most of the modules. We exploited here a resampling approach on the datasets to obtain reliable accuracy of the prediction results using bootstrapping.

We exploited the bootstrapping approach using bootstrap aggregation (bagging) [18] to sample the original data with replacement. This approach was also applied to estimate the accuracy of the BN in predicting more student records of all classes and to avoid overfitting. To implement the bootstrap approach, we used the REPTree algorithm with the classification and regression tree algorithm (CART) in the WEKA [19] mining tool. We decided to resample the data using the decision tree algorithm as it is widely used for the low bias and high variance models. To this end, we spilt the dataset into 60% training set and 40% testing set. We tested the model on the training data through 10 iterations for bagging using the same size as the original dataset. It is important to mention that, we could have resampled any size from the dataset, but we decided to obtain the same number of students' records as in the original dataset for better comparisons to the imbalanced data and decision making using our proposed approach.

To validate the bootstrapped data, we computed the mean μ of the distribution to give an 95% bootstrap confidence interval. The mean was $x = 0.67$ for students overall performance, which we used as an estimated value of the mean for the underlying distribution. To calculate the confidence interval we needed to measure the difference between the distribution of x around the mean μ, as follows:

$$\delta = \overline{x} - \mu \tag{1}$$

To get this distribution, we found the standard deviation for the entire student records $\delta.1$ and $\delta.9$, the 0.1 and 0.9, which are critical values of δ to achieve a 95% confidence interval of $[\overline{x} - \delta.1, \overline{x} - \delta.9]$. The StdDev for the full data was obtained from the following equation:

$$P(\delta.9 \leq \overline{x} - \mu \leq \delta.1|\mu) = 0.95 \quad \Longleftrightarrow \quad P(\overline{x} - \delta.9 \geq \mu \geq \overline{x} - \delta.1|\mu) = 0.95 \tag{2}$$

However, the bootstrap offers a direct approach to obtain the distribution of δ, which can be measured by the distribution of:

$$\delta^* = \overline{x}^* - \overline{x} \tag{3}$$

where, $\overline{x}^*$ indicates the mean of the bootstrap data. We generated one bootstrapped sample of size of 377, which was the size of the original data.

3.3 Bayesian Structure Learning

The learning of the BN structure was performed using GeNIe [20], software implemented for learning and modelling Bayesian Networks (BNs) and Dynamic Bayesian networks (DBNs). For learning the BN structure, we used a Bayesian search on two datasets: the original and bootstrapped datasets. We trained the BNs on these two datasets as we wanted to obtain a very accurate and reliable predictive model. The BNs with temporal links inferred from the admissions and students historical grades, are represented in three time slots (t, t+1 and t+2) (see Fig. 2). In our discrete time BNs, three-time slots are observed to identify the correlation between students' overall performance and other related attributes, such as module grades, online temporal assessment profiles and students' engagement attitude. For example, Fig. 3 shows that the disability attribute at time (t) affects the states of some grades at Year 1 and Year 2.

3.4 Bayesian Parameter Learning

The principle goal of learning a DBN is to find the posterior distribution that is adapted to students' progression data, which allows for identifying the states of all students' attributes as well as overall performance. The parameters of the Dynamic Bayesian model were learned using the expectation maximization (EM) algorithm [21] with the bootstrapped data. We implemented this algorithm to estimate the posterior distribution of students' attributes in time slots t, t+1 and t+2. We used the EM algorithm as it performs the maximum likelihood for temporal data, which supports learning from time series data.

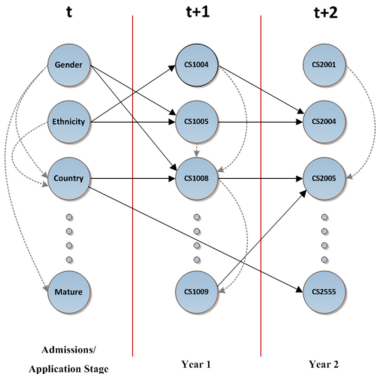

Fig. 2. Dynamic Bayesian network proposed approach for three time slots (t, t+1 and t+2).

4 Experimental Results

The results provide an evaluation of our proposed DBN approach in predicting third year university students' performance. Two key experiments were undertaken: learning from the original data and learning from the bootstrapped data. We set up these two experiments as we wanted to show improvement in predicting the performance, especially for high risk students, who belonged to the minority class in these experiments. In the learning process, we learned the BN structure from students' data in three different time slots. These were students' admissions attributes (time slot t), their obtained grades at Year 1 (time slot t+1) and Year 2 (time slot t+2). In Fig. 3, we present the discovered correlations between students' admission and progression data (grades). It is interesting to note that some of the admission nodes influence students' achievement in some of the Year 1 and Year 2 modules. In addition, students' performance in Year 2 was mainly influenced by their grades in CS1811 (Fundamental Programming Assessment), CS2003 (Usability Engineering) and CS2005 (Networks and Operating Systems), which are compulsory modules for computer science students.

To examine the bootstrapping on improving prediction of the high risk students and the other classes, we provide a comparison between the two approaches used, as shown in Fig. 4. The accuracy results were obtained using the 10 fold cross validation for predicting the academic performance in time slot t+2. Figure 4 shows a significant improvement in identifying the low, medium and high risk students for the bootstrapped data. For example, the accuracy obtained for the high risk class using the bootstrapped data was 0.94, whilst when using the original data it was only 0.63.

The confusion matrices in Fig. 5 indicate the predicted low, medium and high risk students using the original and the proposed bootstrapped data approach. It also reveals the percentage of classification accuracy for each predicted class using the original dataset (A) and bootstrapped dataset (B).

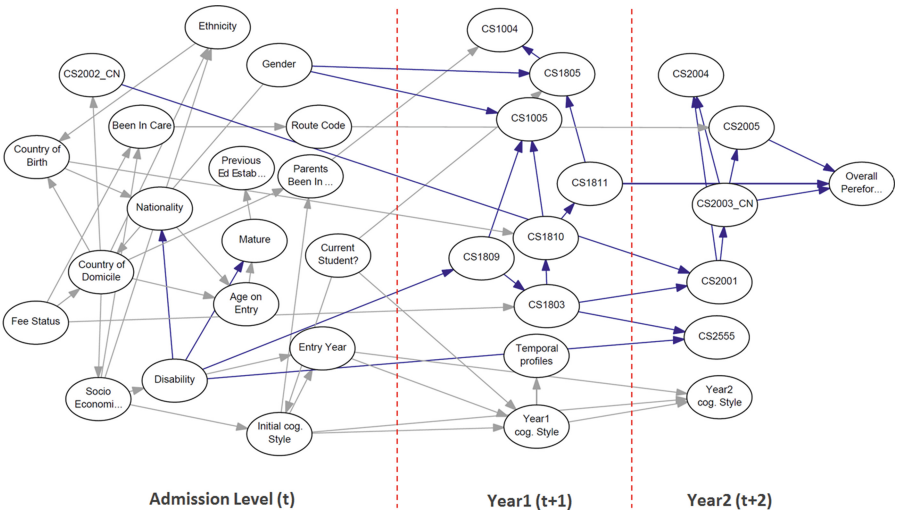

Fig. 3. Dynamic Bayesian structure learned from the bootstrapped temporal educational data. The strong relationships between students' attributes were coloured in blue. (Color figure online)

For evaluating the performance of the BN model, we performed sensitivity and specificity analysis on the cohort of students who were predicted to be at low, medium and high risk. To this end, we visualized the Receiver Operating Characteristics curve (ROC) and the Area under the Curve (AUS), as shown in Fig. 6. We used these two performance measurements as we had a multi class predictive model. It can be seen from the ROC curves in Fig. 6 that for the low risk prediction (A) and high risk prediction (C) are very close to 100% sensitivity and 100% specificity, which means a perfect discrimination of the overall prediction accuracy based on the bootstrapped data.

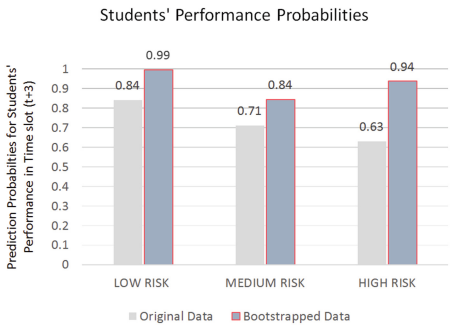

Fig. 4. Prediction probabilities for students' overall performance using two approaches (original and bootstrapped data). It represents the accuracy results for the three classes.

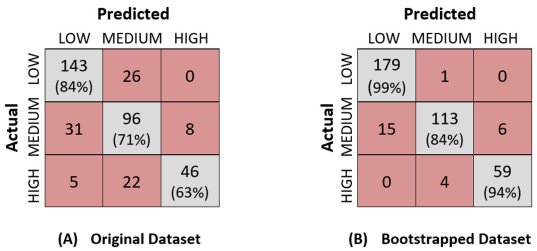

Fig. 5. Academic performance confusion matrices comparing prediction results for the class attribute using the original dataset (A) and bootstrapped dataset (B).

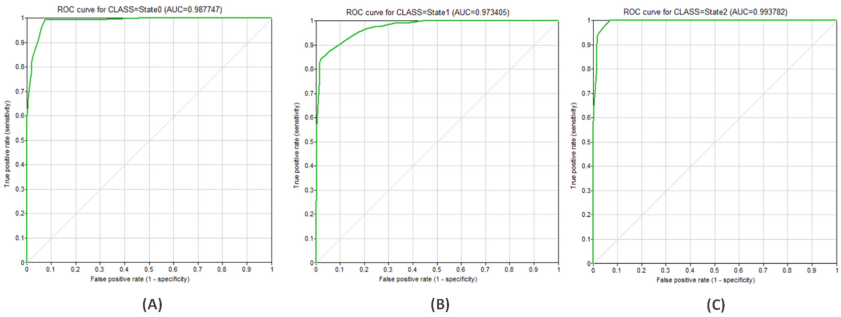

Fig. 6. ROC curves of students' overall performance for the three states, these being: (A) state 0 for the low risk, (B) state 1 for the medium risk and (C) state 2 for the high risk students.

We then examined the validation of our DBN approach in predicting the performance in Year 3 using supplied test sets, as shown in Fig. 7. Firstly, we predicated students' performance based on admissions data only at time slot t for the two datasets, which were the original and the bootstrapped data. Secondly, we added more data, which were students' progressions and final grades at Year 1, to see how better we can predict using the temporal approach. After that, we predicted performance using all students' attributes. It is apparent from Fig. 7 that, the prediction was improved in time slot t+1 when we added Year 1 grades. This improvement was due to the direct relation between students' achieved grades in Year 1 with their overall performance.

5 Confidence Interval Results

This section presents the influence of using bootstrapping to improve learning a BN model. The plotted chart in Fig. 8(A–C) compares the accuracy, the precision and the sensitivity results for predicting the performance of students, among 377 students' time series records for two different datasets (original, bootstrapped). We also show the error bars with a 95% confidence interval, which

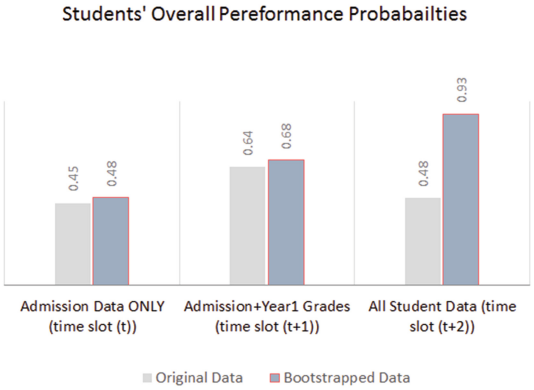

Fig. 7. Validation probabilities for students' performance based on the original and bootstrapped data. It represents the prediction results in time slots (t, t+1 and t+2).

helps in observing the difference between error bars where they overlap or not. It is apparent from Fig. 8(A) the error bars are quite small due to the corresponding confidence interval results. Whenever the confidence interval error bars do not overlap, as clearly illustrated in Fig. 8(B and C), for the precision and the sensitivity, then, this means that the two datasets are statistically significant.

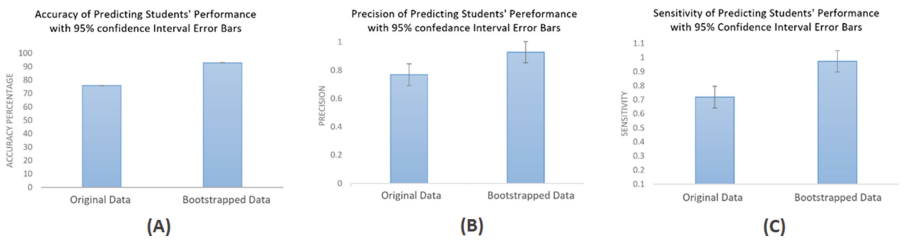

Fig. 8. Confidence interval (CI) error bars charts for the accuracy (A), the precision (B) and the sensitivity (C) for predicting the performance of the students based on the original and the bootstrapped data approaches.

6 Conclusion and Future Work

The prediction of the performance of students has been increasingly emerging in the educational field as it is now possible to transform huge amounts of data into useful knowledge. However, this is a very difficult task because of the issues associated with data. Usually, educational datasets have missing, inaccurate, imbalanced data and so forth, which are also very common issues in all the other research fields. Learning from imbalanced data requires approaches and techniques to transform such data into useful knowledge [22]. To this end, a

resampling approach was explored in this paper for learning DBNs using bootstrap aggregating (bagging). This approach was adopted to tackle the imbalance issue with educational datasets.

The objective of this paper was to model a DBN for predicting the performance of students and the early detection of students at risk of failing or dropping a module based on their time series data. For this purpose, we used students' admission, Year 1 and Year 2 grades in conjunction with other attributes to predict the performance at Year 3, taking into consideration the imbalanced issue of the educational data. A set of two BNs were learned from the educational time series data. The first was learned from the original data, whereas the second model was learned from the resampled data (via bootstrapping). We evaluated the obtained BN models in terms of predicting more states of students' overall performance from temporal educational data using the two different approaches.

Important analytically relevant findings were found when comparing the two approaches used for learning the DBNs. The results show that more states of student's overall performance were achieved when learning from the bootstrapped data, especially for the minority class which was for detecting the high-risk students. We have also demonstrated how the bootstrapped resampling approach enhances the overall prediction of student performance using DBNs. These findings have significant implications for developing education and enhancing students' learning using artificial intelligence. We intend to use these findings to differentiate between the different cohorts of students who perform with similar dynamics and therefore, simplify them to obtain a better understanding of students' performance.

Further experimental works are needed to explore the extension of these Bayesian models with the investigation of latent attributes, with the aim of capture hidden factors that may influence the dynamics of students' academic performance. In our future works, we attempt to compare the proposed methodology especially DBNs with other classification approaches. Moreover, we try to compare other balancing methods with the Bootstrap approach and being more precise in bootstrapping time series grades.

Acknowledgment. This work was partially funded through an internal Brunel Student Assessment and Retention grant (STARS Project).

References

1. The Higher Education Statistics Agency (HESA) Website, HE student enrolments by level of study. https://www.hesa.ac.uk/data-and-analysis/sfr247/figure-3. Accessed 30 Jan 2019
2. Romero, C., López, M.I., Luna, J.M., Ventura, S.: Predicting students final performance from participation in on-line discussion forums. Comput. Educ. **68**, 458–472 (2013)
3. Bhardwaj, B.K. and Pal, S.: Data mining: a prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418 (2012)
4. Araque, F., Roldán, C., Salguero, A.: Factors influencing university drop out rates. Comput. Educ. **53**(3), 563–574 (2009)

5. Seffrin, H.M., Rubi, G.L. Jaques, P.A.: A dynamic bayesian network for inference of learners' algebraic knowledge. In: Proceedings of the 29th Annual ACM Symposium on Applied Computing, pp. 235–240. ACM (2014)
6. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Elsevier (2014)
7. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington (2016)
8. Kabakchieva, D.: Predicting student performance by using data mining methods for classification. Cybern. Inf. Technol. **13**(1), 61–72 (2013)
9. Kaur, G., Singh, W.: Prediction of student performance using weka tool. Int. J. Eng. Sci. **17**, 8–16 (2016)
10. GarcÍa, P., Amandi, A., Schiaffino, S., Campo, M.: Evaluating Bayesian networks' precision for detecting students' learning styles. Comput, Educ. **49**(3), 794–808 (2007)
11. Carmona, C., Castillo, G., Millán, E.: Designing a dynamic bayesian network for modeling students' learning styles. In: 2008 Eighth IEEE International Conference on Advanced Learning Technologies, pp. 346–350 (2008)
12. Kaur, P., Singh, M., Josan, G.S.: Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Comput. Sci. **57**, 500–508 (2015)
13. Beal, C., Cohen, P.: Comparing apples and oranges: computational methods for evaluating student and group learning histories in intelligent tutoring systems. In: Proceedings of the 12th International Conference on Artificial Intelligence in Education, pp. 555–562 (2005)
14. McLaren, B.M., Koedinger, K.R., Schneider, M., Harrer, A., Bollen, L.: Bootstrapping novice data: semi-automated tutor authoring using student log files. In: Proceedings of Workshop on Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes, Proceedings of the 7th International Conference on ITS-2004: Intelligent Tutoring Systems (2004)
15. Feng, M., Beck, J.E., Heffernan, N.T.: Using Learning Decomposition and Bootstrapping with Randomization to Compare the Impact of Different Educational Interventions on Learning. International Working Group on Educational Data Mining (2009)
16. Pfannkuch, M., Forbes, S., Harraway, J., Budgett, S., Wild, C.: Bootstrapping students' understanding of statistical inference. Summary research report for the Teaching and Learning Research Initiative (2013)
17. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
18. Moniz, N., Branco, P. and Torgo, L.: Resampling strategies for imbalanced time series. In: 2016 IEEE International Conference Data Science and Advanced Analytics (DSAA), pp. 282–291. IEEE (2016)
19. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newslett. **11**(1), 10–18 (2009)
20. Druzdzel, M.J.: GeNIe: a development environment for graphical decision-theoretic models (1999)
21. Moon, T.K.: The expectation-maximization algorithm. IEEE SIgnal Process. Mag. **13**(6), 47–60 (1996)
22. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **9**, 1263–1284 (2008)