# Generative Modelling and Classification of Students' e-Learning and e-Assessment Results

Monika Simjanoska*, Sasko Ristov*, and Marjan Gusev*
*University Ss. Cyril and Methodius
Faculty of Computer Science and Engineering, Skopje, Macedonia
Email: m.simjanoska@gmail.com, sashko.ristov@finki.ukim.mk, marjan.gushev@finki.ukim.mk

***Abstract*—In this paper we propose an intelligent modelling of the students' knowledge collected from the e-Learning and e-Assessment processes of a particular course. The paper is focused on proposing a methodology for extracting the students' knowledge from the e-Learning activities, which we refer to as Profiling, then modifying it in compliance with their e-Assessment results and eventually, using it to model the probability distributions of the students Profiles that have passed and of those that have failed the course. The probability distributions of the students Profiles are then applied in the Bayes' theorem to perform binary classification analysis, i.e., to classify the students, pass or fail. The purpose of the proposed methodology is to simulate a real teacher, more precisely, to observe the activities of the particular student during the whole course in order to derive a decision of his or hers overall success.**

***Keywords*— *e-Learning, e-Assessment, Machine Learning, Bayes' Theorem;***

## I. Introduction

E-Learning is the process when one uses sophisticated information and communication technologies to obtain knowledge, whereas, e-Assessment is the process when communication technology is used to assess the learner's knowledge. A clear distinction between assessment *of* learning (assessment for the purposes of grading and reporting with its own established procedures) and assessment *for* learning (assessment whose purpose is to enable students, through effective feedback, to fully understand their own learning and the goals they are aiming for) is made by Vonderwell et al. [1]. In our research the assessment for learning is referred to as e-Learning and the assessment of learning is referred to e-Assessment.

Recently, the design of e-Learning and e-Assessment tools has been a challenge for many researchers [2], [3], especially the introduction of intelligence in e-Learning services. A good overview of e-Assessment systems is presented by Gusev and Armenski [4].

The implementation of the e-Assessment systems at universities requires large amount of computational and storage resources. Therefore, as an appropriate solution, the universities have been given a lot of advice to turn their interests towards cloud computing, the paradigm that offers unlimited flexibility and scalability in terms of storage, computation, network access, reduced power consumption and lower costs [5], [6], [7]. Ristov et al. [8] propose a SOA architecture of a cloud hosted e-Assessment system which uses scalability and elasticity in order to achieve sustainable performance with minimal costs since it uses minimum resources utilised only during the e-Assessment.

An e-Learning and e-Assessment system has been implemented at our university since 2002 [9]. The years of experiments provided new directions and proved established strategies for use of the system. Main efforts address creation of knowledge database capable to support the system and this processes required strong commitment of the teachers in building up the course. The knowledge database is organised as a hierarchical structure, precisely, as a tree, in order to enable a good realisation of a navigation algorithm. This kind of organisation is also supported by Baumgartner and Shankararaman [10].

The knowledge database [4] consists of the following items:

1) *Course* is the root of the tree and consists of several lectures;
2) *Lecture* represents areas within the course and consists of several parts;
3) *Part* represents parts of the lectures;
4) *Set* represents a set of basic learning objectives and
5) *Learning objective (LO)* is an essential knowledge item to be learned which consists of a set of questions.

The idea of online learning introduced in [9] consisted of asking the student questions form different LOs after the lecture. The students were providing answers on the specialised e-Testing system. A strategy was chosen to classify the students knowledge about the particular LO after some questions. If the student knowledge was classified as ”pass” the navigation algorithm continued with a new LO.

Several analyses were performed to use different strategies and navigation algorithms. The main idea when developing the navigation algorithm was to simulate a real teacher that asks the student several questions and makes a decision to continue with questions in the same LO, or to move to the next LO. Therefore, from a teaching perspective, a good strategy was built on the following objectives:

- determining the minimal number of questions the system should ask to get relevant assessment of student's knowledge; and
- determining the number of provided correct answers that

should classify the student knowledge as ”pass”.

The experiments showed that the strategy of providing 3 correct answers in consecutive questions [4] is the best alternative for the given criteria. Finally, the system provided the following features:

- *Corrective measure* - whenever the student gives a wrong answer the system suggests corrections;
- *Adaptive learning path* - better students can move faster towards the more complex LOs; and
- *Clear objectives and goals* - the navigation explains what else has left to be learned.

The developed 3-in-a-row strategy confirms the student's knowledge of a particular LO; however, it cannot predict the student's overall success at the end of the course.

In this paper we aim to use the results from the developed 3-in-a-row strategy for modelling the initial students' knowledge of the given LO during the e-Learning process. Afterwards, considering all the results from the e-Assessment processes during the course, we update and re-evaluate the initial model of the students' knowledge, thus, creating revised students' Profiles. The students' Profiles are then used in a Machine Learning (ML) analysis to model the classes of passing, or failing the course at the end of the semester.

In classification problems, the main idea is to model the existing relationships between a set of multivariate data items and a certain set of outcomes. Although there are many ML techniques that can be used to analyse e-Learning data, we propose an approach for building a classifier based on the different probability distributions of the Profiles that have passed and the Profiles that have failed the course. Once we model the classes probability distributions, we use them in the Bayes' theorem to calculate the posterior probabilities of the new Profiles.

The rest of the paper is organised as follows. Section II presents the recent ML analyses and their application in e-Learning. The new proposed methodology is presented in Section III. The results from the experiments are presented in Section IV and a conclusion is derived in Section V.

## II. Related Work

In this section we give a review of the recent ML applications in e-Learning.

Castro et al. [11] provide an up-to-date snapshot of the current state of research and applications of Data Mining methods in e-Learning. They provide a taxonomy of e-Learning problems to which Data Mining techniques have been applied, including: students' classification based on their learning performance, detection of irregular learning behaviours, e-Learning system navigation and interaction optimisation, clustering according to similar e-Learning system usage and systems' adaptability to students' requirements and capacities.

Hanna and Tang [12], [13] present studies on how Data Mining techniques could successfully be incorporated to e-Learning environments and how they could improve the learning tasks. Sison et al. [14] present an analysis on how ML techniques have been used to automate the construction and induction of student models, as well as the background knowledge necessary for student modelling. Their paper sheds light on the difficulty, suitability and potential of using ML for student modelling processes, and, to a lesser extent, the potential of using student modelling techniques in ML.

Minaei and Punch [15] present an approach for classifying students in order to predict their final grade based on features extracted from logged data in an education web-based system. A combination of multiple classifiers leads to a significant improvement in classification performance.

Fei et al. [16] describe exploring automatic question classification tests, which can be used in an e-Learning system. These tests can take the form of multiple-choice testing, as well as fill-in-the-blank and short-answer tests. They proposed a text categorisation model using an artificial neural network trained by the back-propagation learning algorithm as the text classifier. Their test results showed that the system achieved the performance of nearly 78%.

Chang et al. [17] state that with the growing demand in e-Learning, much research indicated that adaptive learning is a critical requirement for promoting the learning performance of students. The first step for achieving adaptive learning environments is to identify students' learning styles. Therefore, in their paper they propose a learning style classification mechanism to classify and then identify students' learning styles. The proposed mechanism improves k-nearest neighbour classification and combines it with genetic algorithms. To demonstrate the viability of the proposed mechanism, the proposed mechanism is implemented on an open-learning management system. The experimental results indicate that the proposed classification mechanism can effectively classify and identify students' learning styles.

## III. Methods and Methodology

In this section we present the methods that we developed to extract knowledge from the e-Learning processes, to preprocess the data and build a classification model.

### A. Modelling the Students' Knowledge

To apply the results from the e-Learning processes onto a classification procedure, we must extract the raw knowledge results from the database and do some processing to create each student an initial Profile of success.

Considering the knowledge database organisation, we can perceive that for each course there are a lot of LOs, and each LO consists of sets of questions. Each LO is an important learning objective that each student should learn and be evaluated with ”pass”, by such as using the three-in-a-row strategy. After the student is evaluated with ”pass” for a given

LO, the system moves forward to the next LO, by using the appropriate navigation algorithm.

The level of student's knowledge about a particular LO can be presented by

- **the number of trials** $NT$ to be evaluated with "pass" for a given LO, and
- **the number of questions** $NQ$ answered before the decision making strategy (such as the three-in-a-row pattern) classifies the student knowledge as a "pass" in the learning trial.

Considering that there are $n$ different LOs in a particular course, the knowledge from the e-Learning process can be presented by a $n$-dimensional vector. Denote by $N$ the number of students. In the following analysis we will use indexes:

- $i$ to present a particular student, where $i = 1, \dots, N$ and
- $j$ to present a particular LO, where $j = 1, \dots, n$.

To express the student's efficiency assigned to each LO, we propose a new metric $IS$ to measure the student's *Initial Success*. The dimensionality of the $IS$ vector is $n$, equal to the number of all LOs in a particular course, calculated by (1) for $i = 1, \dots, N$. In the beginning, the $IS$ indicator for every student enrolled in the course is a null vector.

$$IS_i = (IS_{i1}, IS_{i2}, \dots, IS_{ij}, \dots, IS_{in}) \quad (1)$$

The knowledge level evaluated of the $i$-th student for the $j$-th LO is calculated by (2), for $i = 1, \dots, N$ and $j = 1, \dots, n$. Considering the fact that the students are assigned each week with questions from a new LO, the $IS$ is calculated as a sum of the number of trials ($NT$) necessary for the student to pass the given LO, and the weighted product of the number of questions ($NQ$) with weight 0.01 in the last trial. The minimum value of the variable $NT$ is 1, since the student can pass the given LO in at least one trial. Considering the range of the variable $NQ$, we use the value 0.01 in order to scale the number of questions so that it will always be lower than 1. Therefore, it will vary in range $0.03 \leq NQ < 1$, since the maximum number of answered questions per LO is a double digit number.

Both the variables $NT$ and $NQ$ are determined by the students themselves. Hereupon, the students can perform as much trials and answer as much questions as needed until they achieve three correct answers in a row, a passing manner determined by the three-in-a-row strategy explained in Section I.

The lower the $IS$ score is, the more the student has learned the LOs.

$$IS_{ij} = NT_{ij} + (0.01 * NQ_{ij}), \quad (2)$$

The information that we extracted from the database is raw data depending only on the results from the e-Learning processes. Therefore, it needs some revision prior to modelling the probability distributions of the classes as "pass" and "fail".

To get a real picture of how much the students have learned during the course, we must take into account the results from the e-Assessments processes. For this purpose, we propose another metric, *Revised Success (RS)* to re-evaluate the students' knowledge after the e-Assessments on the particular LOs. Thus, the students' success is revised every time the student has taken an e-Assessment. By introducing this approach, we simulate a real environment where the teacher evaluates the students' activities during the whole course to determine their overall success. The following methodology is used to calculate the *Revised Success*, $RS$, in (3) and (4), where $i = 1, \dots, N$ identifies the currently analysed student and $j = 1, \dots, n$ denotes a particular LO.

$$\text{Punishment}: \quad RS_{ij} = IS_{ij} + 1 \quad (3)$$

$$\text{Reward}: \quad RS_{ij} = IS_{ij} - 1 \quad (4)$$

The $RS$ fully depends on the student's result from answering the questions from the e-Assessment. If the student fails to answer a question of a particular LO, then the student is punished by adding one more trial on the corresponding $IS$. When the student answers correctly a question of a particular LO, the student is rewarded by reducing the number of trials by 1 in the e-Learning history. That is how we use the e-Assessment results to affect the student's impression during the semester.

### *B. Generative Modelling and Classification*

To make a clear distinction between the knowledge of the students that have passed or failed the course, we separated the knowledge data into two classes. The first class $C_1$ contains 151 Profiles of students that have passed the course exams with average score of $> 70$, and the second class $C_2$ contains 128 Profiles of students that have failed the course exams with an average score of $< 30$.

Since the analysis is performed on students from different periods, we realise that majority of the LOs are not commonly elaborated in the two classes. Therefore, to achieve a more realistic knowledge modelling, we analyse only the non-zero values in both classes, and reduce the dimensionality of the vectors from 1361 to 99 LOs.

Observing the probability distribution in Figure 1, where the x-axis presents the values of the revised success, and the y-axis present the number of Profiles that have the particular value, we can see that the histograms for the both classes, are equal. This is not an unexpected phenomena, since we derived the histograms from the initial knowledge $IS$ and we know that initially all students should have more or less equal knowledge due to the 3-in-a-row strategy discussed above. Considering the Profiles that appear to have $IS$ values less than 1 in Figure 1, this phenomena is due to the large amount of Profiles whose initial success values are 1.03 and it is impossible to represent all of them in a single point.

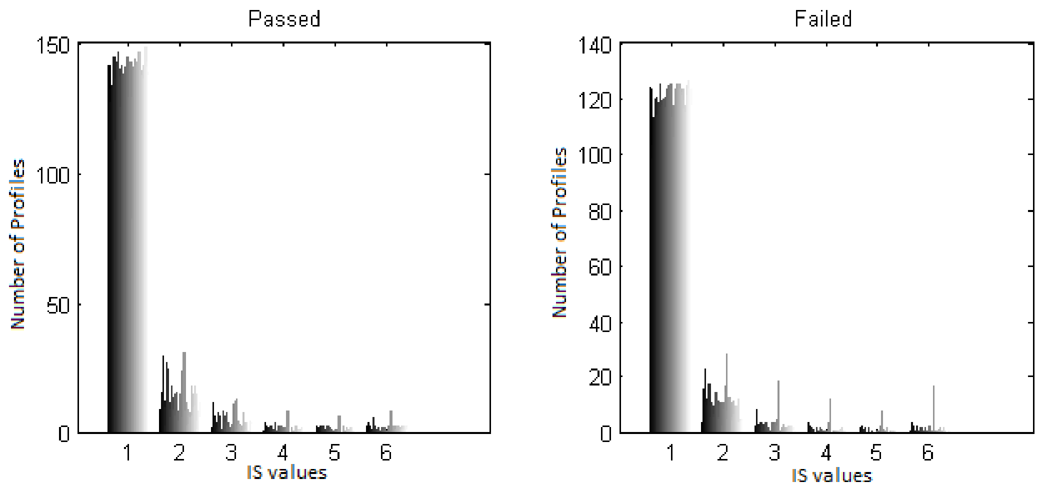


Fig. 1. Distributions of IS at $C_1$ and at $C_2$

On the other hand, observing the distributions of the classes in terms of the revised student knowledge $RS$ discussed in Section III-A, as presented in Figure 2, we can conclude that the two classes have different distributions. Therefore, we will proceed to perform further normalisation of the revised students' success in order to obtain data applicative for Bayesian modelling.

In order to normalise the $RS$ data, we did some data preprocessing by calculating the z-score (5) to normalise the LOs $RS$ values across the Profiles in both the classes $C_1$ and $C_2$.

$$NRS_{ij} = \frac{RS_{ij} - \mu(RS_j)}{\sigma(RS_j)} \tag{5}$$

Once we obtained the z-scores, we wanted to extend the vectors and simulate the students have elaborated more LOs in a same manner. Therefore, in order to achieve more accurate distributions, we performed bootstrapping using the $mode$ values of the vectors and produced results of 200 LOs. Then we smoothed the data and obtained the new, obviously different, distributions of the classes presented on the left sides of the 3 and 4.

In order to confirm our assumption of difference, we tested the distributions using few hypothesis for few different types of distribution. However, most of the Profiles in $C_1$ showed to have $Lognormal$ distributions, whereas the Profiles in $C_2$ showed to be $Normally$ distributed. The distributions fitting is performed by using the Chi-square goodness-of-fit test and the results are presented in figures 3 and 4, respectively.

Once we determined the prior distributions of the two classes, i.e., the class-conditional densities $p(\vec{x}|C_i)$ for $i = 1, 2$, we can use them in the Bayes' theorem (6) to calculate the posterior probability $P(C_i|\vec{x})$.

$$p(C_i|\vec{x}) = \frac{p(\vec{x}|C_i) * P(C_i)}{\sum_{1}^{2} p(\vec{x}|C_i) * P(C_i)} \tag{6}$$

However, since the prior probabilities, $P(C_i)$, for belonging to one of the classes are equal, and the total probability, $\sum_{1}^{2} p(\vec{x}|C_i) * P(C_i)$, is used to scale the posterior probability in range [0,1], the classification is deduced to the calculation only of the class-conditional densities, $p(\vec{x}|C_i)$. The class of the new vector is determined by the maximisation of the $p(\vec{x}|C_i)$.

## IV. Experiments and Results

This section presents the results obtained by experiments following the approach discussed in Section III. To test the ability of our approach to classify the critical Profiles whose average score is in interval [30,70], we defined the following test cases:

- Test Case 1 (Training set):
  $C_1 = Profiles(score > 70) = 151;$
  $C_2 = Profiles(score < 30) = 128.$
- Test Case 2:
  $C_1 = Profiles(score > 60) = 369;$
  $C_2 = Profiles(score < 40) = 312.$
- Test Case 3:
  $C_1 = Profiles(score > 50) = 644;$
  $C_2 = Profiles(score \leq 50) = 629.$

In Table I, we present the performance of the classifier in terms of true positive rate, $TP$, and true negative rate, $TN$. $TP$ denotes the percentage of correctly classified Profiles that have passed the exams, whereas the $TN$ is the percentage of correctly classified Profiles that have failed the exams.

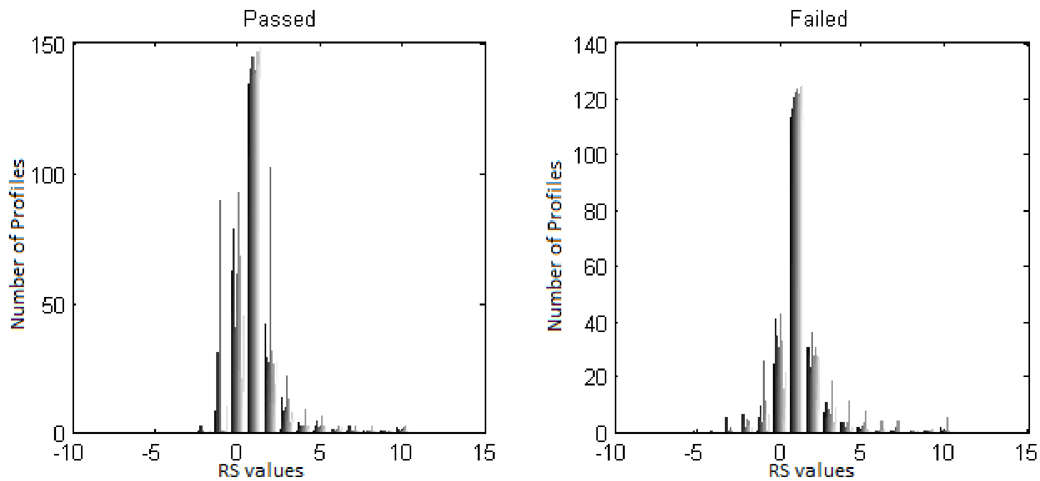


Fig. 2. Distributions of RS at $C_1$ and at $C_2$

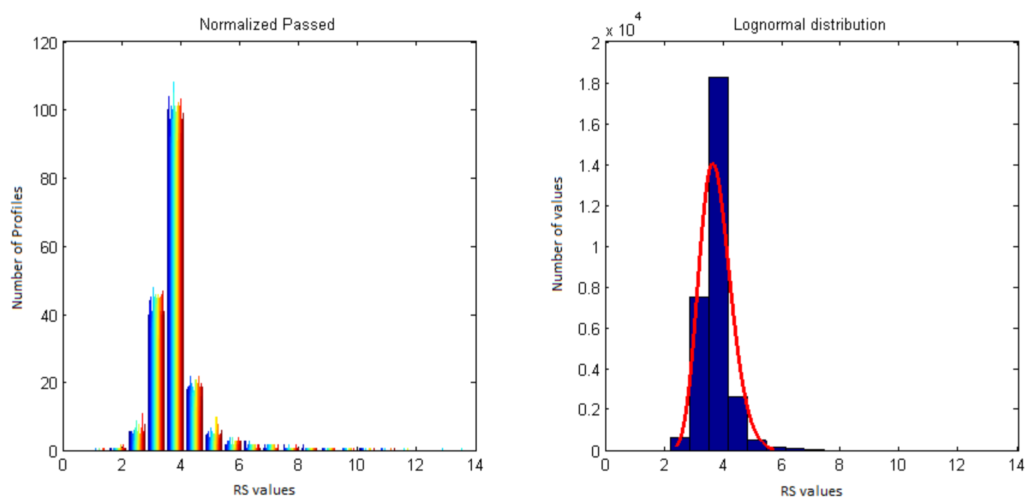


Fig. 3. Normalized Passed Profiles and their fitting to Lognormal distribution

TABLE I
CLASSIFICATION RESULTS

| Test Cases | TP | TN |
|---|---|---|
| Case 1 | 96.68 % | 96.09 % |
| Case 2 | 95.93 % | 100 % |
| Case 3 | 95.80 % | 97.93 % |

TABLE II
POLYNOMIAL KERNEL CLASSIFICATION

| Performance | Polynomial kernel |
|---|---|
| TP | 90 % |
| TN | 75 % |

According to the classification results in Table I we can conclude that our classifier is able to correctly classify even the Profiles that we defined as critical.

Compared to the results from our previous research [18] where we followed a discriminative approach for modelling the classifier, Table II presents the best results from the classification with Polynomial kernel (PK) which clearly shows decreased accuracy when discriminating between the Profiles that have passed from those that have failed the exams.

## V. Conclusion and Future Work

In this paper we focus on intelligent modelling of the students' results obtained from e-Learning and e-Assessment processes of a particular course. The purpose of the proposed research is to simulate a real teacher, who will be able to observe the activities of a particular student during the whole course in order to derive a decision of the overall success.

We propose a methodology for building students' Profiles based on their e-Learning activities, then we modify the Profiles in compliance with their e-Assessment results and

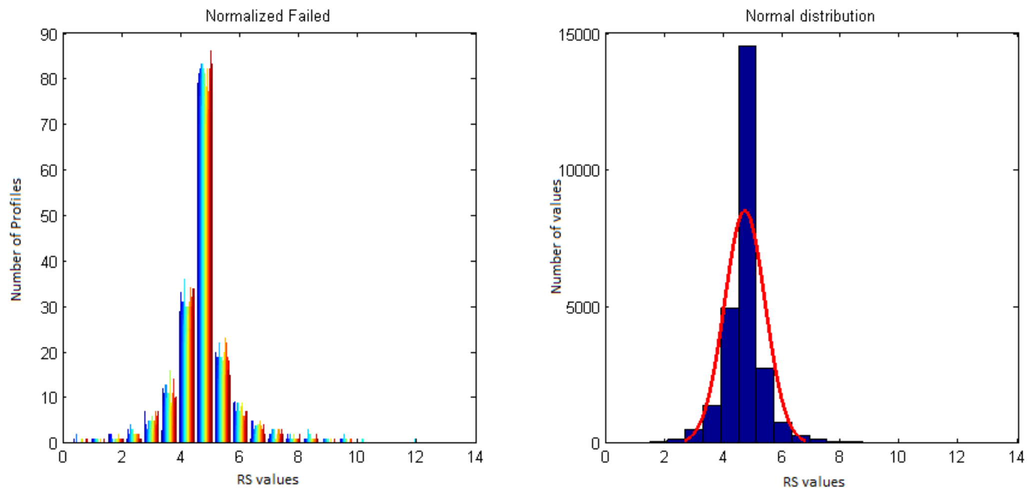


Fig. 4. Normalized Failed Profiles and their fitting to Normal distribution

eventually, we model the probability distributions of the students Profiles that have passed, and of those that have failed the course. The model is then applied in the Bayes' theorem to perform a binary classification analysis, i.e., to classify the students into one of the classes, pass or fail.

The classification analysis showed that our generative model is capable of recognising the passing and the failing Profiles and when compared to our previous discriminative model for classification, shows high percentage of increased accuracy.

Since the generative approach showed to be better option for this kind of problem, in our future work we will go deeper into the problem to apply our approach for modelling the students' final grades.

## References

[1] S. Vonderwell, X. Liang, and K. Alderman, "Asynchronous discussions and assessment in online learning," *Journal of Research on Technology in Education*, vol. 39, no. 3, p. 309, 2007.

[2] A. Rashad, A. A. Youssif, R. Abdel-Ghafar, and A. E. Labib, "E-assessment tool: A course assessment tool integrated into knowledge assessment," in *Innovative Techniques in Instruction Technology, E-learning, E-assessment, and Education*. Springer, 2008, pp. 7–12.

[3] Z. Harley and E. Harley, "E-learning and e-assessment for a computer programming course," *EDULEARN11 Proceedings*, pp. 2074–2080, 2011.

[4] M. Gusev and G. Armenski, "E-Assessment Systems and Online Learning with Adaptive Testing," in *e-Learning Paradigms and Applications - Agent-based Approach*, M. Ivanovic and L. Jain, Eds. Springer Verlag, 2013, pp. 229–249.

[5] A. C. Caminero, A. Robles-Gómez, S. Ros, R. Hernández, R. Pastor, N. Oliva, and M. Castro, "Harnessing clouds for e-learning: New directions followed by uned," in *Global Engineering Education Conference (EDUCON), IEEE*. IEEE, 2011, pp. 412–416.

[6] M. Mircea and A. I. Andreescu, "Using cloud computing in higher education: A strategy to improve agility in the current financial crisis," *Communications of the IBIMA*, vol. 2011, pp. 1–15, 2011.

[7] D. G. Chandra and M. D. Borah, "Cost benefit analysis of cloud computing in education," in *Computing, Communication and Applications (ICCCA), 2012 International Conference on*. IEEE, 2012, pp. 1–6.

[8] S. Ristov, M. Gusev, G. Armenski, K. Bozinoski, and G. Velkoski, "Architecture and organization of e-assessment cloud solution," *Global Engineering Education Conference (EDUCON), IEEE*, pp. 736–743, 2013.

[9] M. Gusev and G. Armenski, "On-line learning and etesting," in *Information Technology Interfaces, 2002. ITI 2002. Proceedings of the 24th International Conference on*. IEEE, 2002, pp. 147–152.

[10] I. Baumgartner and V. Shankararaman, "Actively linking learning outcomes and competencies to course design and delivery: Experiences from an undergraduate information systems program in singapore," in *Global Engineering Education Conference (EDUCON), 2013 IEEE*, 2013, pp. 238–246.

[11] F. Castro, A. Vellido, À. Nebot, and F. Mugica, "Applying data mining techniques to e-learning problems," in *Evolution of teaching and learning paradigms in intelligent environment*. Springer, 2007, pp. 183–221.

[12] M. Hanna, "Data mining in the e-learning domain," *Campus-wide information systems*, vol. 21, no. 1, pp. 29–34, 2004.

[13] T. Tang and G. McCalla, "Smart recommendation for an evolving e-learning system: Architecture and experiment," *International Journal on E-learning*, vol. 4, no. 1, pp. 105–129, 2005.

[14] R. Sison, M. Shimura *et al.*, "Student modeling and machine learning," *International Journal of Artificial Intelligence in Education (IJAIED)*, vol. 9, pp. 128–158, 1998.

[15] B. Minaei-Bidgoli and W. F. Punch, "Using genetic algorithms for data mining optimization in an educational web-based system," in *Genetic and Evolutionary ComputationGECCO 2003*. Springer, 2003, pp. 2252–2263.

[16] T. Fei, W. J. Heng, K. C. Toh, and T. Qi, "Question classification for e-learning by artificial neural network," in *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, vol. 3. IEEE, 2003, pp. 1757–1761.

[17] Y.-C. Chang, W.-Y. Kao, C.-P. Chu, and C.-H. Chiu, "A learning style classification mechanism for e-learning," *Computers & Education*, vol. 53, no. 2, pp. 273–285, 2009.

[18] M. Simjanoska, M. Gusev, S. Ristov, and A. M. Bogdanova, "Intelligent student profiling for predicting e-assessment outcomes," in *Global Engineering Education Conference (EDUCON), 2014 IEEE*, 2014, pp. 616–622.