Course-Specific Model for Prediction of At-Risk Students Based on Case-Based Reasoning

Haris Supic and Dzenana Donko
Faculty of Electrical Engineering, University of Sarajevo, hsupic, ddonko@etf.unsa.ba

Abstract - **Identifying at-risk students is a crucial step in different learning settings. Predictive modeling technique can be used to create an early warning system which predicts students' success in courses and informs both the teacher and the student of their performance. In this paper we describe a course-specific model for prediction of at-risk students. The proposed model uses the case-based reasoning (CBR) methodology to predict at-risk students at three specific points in time during the first half of the semester. In general, CBR is an approach of solving new problems based on solutions of similar previously experienced problem situation encoded in the form of cases. The proposed model classifies students as at-risk based on the most similar past cases retrieved from the casebase by using the *k*-NN algorithm. According to the experimental evaluation of the model accuracy, CBR model that is being developed for a specific course showed potential for an early prediction of at-risk students. Although the presented CBR model has been applied for one specific course, the key elements of predictive model can be easily reused by other courses.**

Index Terms – at-risk students, case-based reasoning, prediction, student performance.

INTRODUCTION AND RELATED WORK

Whether we are teachers, teaching assistants, or administrators, we want all of our students to experience academic achievements and desired learning outcomes. But the reality is that not all students will reach their academic potential and will be seen as at-risk students. Here the term at-risk refers to students who are considered to have a higher probability of failing or dropping some specific course. The successful implementation of integrated quality management systems in educational settings relies, among other things, on the ability to address early identification of at-risk students. One of the primary reasons for attempting to identify at-risk students is that it allows lecturers to respond appropriately to students' learning needs. Lecturers can then provide at-risk students personalized and adaptive learning paths for improving their performance in the course. Personalized and adaptive learning environments can provide support in learning activities based on students' characteristics [1]. One of the main problems in personalized and adaptive learning is the learning path recommendation. In personalized and adaptive learning environments learning path gives step-by-step guidance and is continually adapted to students' characteristics in order to provide students to achieve their objectives in the shortest possible time [2], [3]. This guidance should be based on student's individual knowledge state.

To help students' who may fail a course, it is crucial to identify them as early in the semester as possible. Instead of relying only on experience, another way to help at-risk students is to analyze students' academic performance data. Over the last decade, the use of learning management systems (LMS) and student information systems (SIS) in different educational institutions has increased. A wide variety of data can be collected from these systems, including the students' individual characteristics, results of tests and other assessment scores, behavior data, students' activity records, students' attendance records, etc. As a result, large datasets that store many aspects of students' performance including previous academic achievements are available. Finding an appropriate predictive model in this information can help educational institutions improve learning processes [4]. Through the use of an appropriate predictive model, it is possible to forecast at-risk students early in the semester. The majority of research on predictive models focused on predicting students' success in a course using academic performances available before the start of the semester. A few studies have utilized the academic performance data available during the semester [4], [5]. On the other hand, it is natural to expect that relevant data available during the semester have the highest prediction power. Unlike mid-term exams, many learning activities, such as for example quizzes, homework and pre-lab assignments, start earlier in the semester and using them as predictors may result in accurate prediction early in the semester. As the academic performance data becomes available during the semester, students' achievements should be predicted more accurately. The student success prediction model reported in [6] predicts sudents' performance based on their online behavior. Several other studies have been conducted to investigate the effectiveness of different models for prediction of at-risk students [7], [8].

One of the problems with early warning systems is that they usually employ a general model that cannot adequately

capture the specificities of all courses [4], [5]. For example, the course components can vary considerably from one course to another. One of the common problems with general prediction models is that they cannot address the complexity of all courses [4], [5]. Creating predictive models at the specific course level increases the accuracy of the model [4].

In this paper, we describe a case-based reasoning model that allows students' success prediction in a course at three specific points in time during the first half of the semester.

The main two research questions (RQs) of this paper can be stated as follows:

1. RQ_1: Can the case-based reasoning methodology provide a students' success predictive models.
2. RQ_2: To what extent does accuracy of the proposed case-based prediction model change during the first half of the semester?

The rest of the paper is organized as follows. Section II provides a short description of a general case-based reasoning problem solving methodology. Section III describes the case structure and representation. Section IV gives the description of the case-based reasoning cycle for prediction of at-risk students. Experimental results are given in Section V. Section VI outlines conclusions and some future work.

CASE-BASED REASONING METHODOLOGY

Case-based reasoning is a problem solving methodology that may use different techniques. The choice of concrete techniques used in a particular CBR system highly depends on the problem domain. In order to facilitate the understanding of the proposed model, the case-based reasoning as a problem solving methodology need to be briefly described. The problem solving methodology using case-based reasoning is depicted in Figure I.

In general, CBR is a process of solving new problems based on the solution of similar past problems. There are four steps in case-based reasoning [9]-[11]:

1. **RETRIEVAL.** This step matches a new target case with past cases stored in casebase and retrieves one or more the most similar cases.
2. **REUSE.** In the reuse step, the solution component of one or more retrieved cases is used to build a solution for the new problem.
3. **REVISE.** The proposed solution to the new problem is tried out and evaluated to check whether it actually solves the new problem.
4. **RETAIN.** After the proposed solution has been applied and the new problem is solved, the new experienced problem solving episode is stored in the casebase for possible future use.

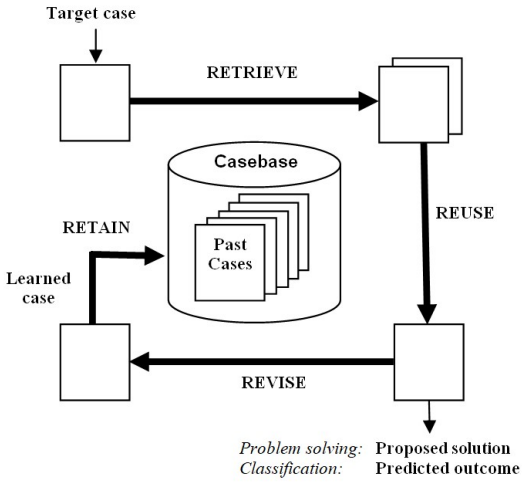

FIGURE I
CASE-BASED REASONING CYCLE (ADAPTED FROM [3], [10])

CASE STRUCTURE AND REPRESENTATION

An adequate case representation scheme is an essential element for successffful application of the case-based reasoning methodology in a concrete problem solving situation. Definition of an expressive case representation scheme involves, among other things, an appropriate selection of information content that should be encoded in the form of a case.

In general, a typical case c in a CBR system consists of three components [10], [11]:

$$c = (p, s, q),$$

where

- p is a *problem description* component of a case. This component describes the new problem.
- s is a *solution* component of a case. This component describes a solution for the problem described by p.
- q is an *outcome* component of a case. This component represents the resulting state of the world after the case occured, i.e. the solution s was applied.

If CBR is used for prediction, the solution component s is not a part of the case structure [12]. In other words, in this paper a case consists of two components:

$$c = (p, q)$$

where

- p describes an academic performance state of the student before the prediction. This component consists of values of all attributes for that case. A full description of the case attributes is given in Table I.
- q is an outcome component. The model proposed in this paper uses only two possible values: 0 and 1. The value 0 means that a student was *not successful*, and the value 1 means that a student was *successful*.

The cases we used in this paper are extracted from the student information system used at the Faculty of Electrical Engineering, University of Sarajevo. This system provides tracking of student attendance, documenting results of pre-lab and homework assignments, as well as documenting other assessment scores. The selection of cases was based on the following:

- Only cases of which all the attributes were available were included. In other words, we excluded missing data cases.
- Only cases representing students who enrolled in the course for the first time were included.

The casebase *CB* is modeled as a collection of cases:

$$CB = \{c_1,.....,c_i,\ldots,c_{|\mathrm{CB}|}\},\ 1 \le i \le\ |\mathrm{CB}|.$$

II. Case Attributes

The basic approach of this paper is to use all the attributes that are shared by all the cases. Formally, a description component *p* of a case is represented by attribute-value pairs:

$$p = \{A_1{:}a_1, A_2{:}a_2, \ldots, A_n{:}a_n\},$$

where A_i, $1 \le i \le n$, denotes an attribute of a case, and a_i denotes the value of the attribute A_i. In this paper, we used the following three types of attributes:

- Numerical attributes,
- Categorical attributes (values with no order in rank), and
- Ordinal attributes (values with order in rank).

Table I summarizes the relevant information associated with these attributes. The first column contains the names of all attributes, and the type of each attribute is specified in the second column. Because not every attribute is equally important, we assigned a weight to each attribute according to its importance. The different weights have a value in the interval from 0 to 1. These weights are determined by a course teacher in terms of the course specific domain knowledge. Fourth column gives a short description of the attributes.

CBR CYCLE FOR AT-RISK STUDENT PREDICTION

In this section, we describe a case-based reasoning cycle to predict at-risk students at three different points in time during the first half of the semester.

I. Retrieval of similar cases (RETRIEVE)

The *k*-nearest neighbors (*k*-NN) method is one of the most widely used techniques in CBR systems for the retrieving of similar cases. This method uses an exhaustive search algorithm that involves the assessment of similarity between the target case representing a new case to be classified and stored cases representing previously classified cases. The result of *k*-NN is a fixed number of retrieved cases that have the highest similarity scores with the target case. In this paper, we used seven retrieved cases (k=7) with the highest value of similarity to identify at-risk students.

The retrieve step in CBR cycle relies heavily on similarity measure function. This function assigns a number between 0 and 1 as a measure of similarity between two cases, where 0 means no match and 1 means an exact match. The overall similarity *SIM* (*T*, *S*) between a target case *T* and a stored case *S* is computed as follows:

$$SIM(T,S) = \frac{\sum_{i=1}^{n} w_i \cdot sim(t_i, s_i)}{\sum_{i=1}^{n} w_i},$$

where

- T, S are target and stored case, respectively;
- t_i, s_i are values of target and stored case on the attribute A_i, respectively;
- $sim(t_i,s_i)$ is a function that defines the similarity measure between T and S on the attribute A_i;
- w_i is the weight of the attribute A_i;
- n is a number of case attributes.

TABLE I
ATTRIBUTES USED IN CASE-BASED REASONING FOR PREDICTION OF AT-RISK STUDENTS

Attribute (Abbrev.)	Type	Weight	Description
Pre-lecture preparation (PR)	Categorical	0.05	Indicates whether students have completed recommended reading before a class lecture
Week workload (WW)	Ordinal	0.20	Estimated student workload per week using questionnaire
Prior GPA (PriorGPA)	Numerical	0.10	Grade point average of all prior successfully completed courses
Prerequisite GPA (PrereqGPA)	Numerical	0.20	Grade point average of prerequisite courses
Course Completion Rate (CCR)	Numerical	0.05	Course completion rate (Earned credits/Attempted credits)
Attendance Rate (AR)	Numerical	0.05	Attendance rate at lectures and tutorials
Homework Grade Average (HGA)	Numerical	0.20	Average homework grade
Average Pre-lab Assignment Grade (APLAG)	Numerical	0.10	Average pre-laboratory assignment grade
In-class Activity (IA)	Ordinal	0.05	Estimated student engagement in the classroom

We use specific similarity functions *sim,* one function for each type of attributes. For numerical attributes, we compute the similarity score as follows:

$$sim(t_i, s_i) = 1 - \frac{|t_i - s_i|}{range(A_i)},$$

where $range(A_i)$ denotes a range of valid values for attribute A_i. Furthermore, for categorical attributes, the similarity score is computed as follows:

$$sim(t_i, s_i) = \begin{cases} 1 & if \quad t_i = s_i \\ 0 & if \quad t_i \neq s_i \end{cases}$$

Finally, for the ordinal attributes, the similarity is computed by using the formula:

$$sim(t_i, s_i) = 1 - \frac{|Ord(t_i) - Ord(s_i)|}{Card(A_i)},$$

where $Ord(t_i)$ and $Ord(s_i)$ denote ordinal numbers of the attribute values t_i and s_i in the range set, and the $Card(A_i)$ is the cardinality of the attribute A_i.

II. Predicting the outcome of the target case (REUSE)

There are two possible outcome values: *successful* and *not successful*. For predicting the outcome of the target case, we used standard *k*-NN rule. In other words, we base our prediction on the most frequent outcome value of the retrieved cases as our predicted value.

III. The last two steps of the CBR cycle (REVISE and RETAIN)

After a student has completed the course, we have the actual outcome whether the student was successful or not successful, based on the final exam. In the REVISE step, we compare the predicted outcome with the actual outcome. We finish this step by inserting the actual outcome as the value of the outcome component in the target case. In the RETAIN step, the target case is added to the casebase.

EXPERIMENTAL RESULTS

Our experiments are based on a casebase created from the undergradute course in algorithms and data structures taught at the Faculty of Electrical Engineering, University of Sarajevo, over the past four years. According to the model previously described, the stored cases are extracted from available student academic performance data related to the course.

In order to evaluate the described CBR model and to determine to what extent does accuracy of the students' success prediction model changes across the weeks during the first half of the semester, we have created a casebase composed of 448 cases for each of the three specific points in time: end of week 3, end of week 5, and end of week 7. For each week specific casebase, we used the student academic performance information available in corresponding weeks.

The evaluation procedure that we followed includes the following steps:

1. Creation of casebase composed of 448 cases that are relevant to the corresponding week.
2. Select one case as the target case and remove this case from the casebase.
3. Retrieve the most similar cases for the selected target cases and predict the outcome.
4. Check whether the predicted outcome is correct and record the results.
5. Repeat steps 2-4 until all cases from the casebase are used as the target case.
6. Evaluate the accuracy of the CBR prediction model for the corresponding week.

The overall accuracy of the CBR model is defined as follows:

$$Accuracy = \frac{TP + TN}{Total\ number\ of\ cases} = \frac{TP + TN}{TP + TN + FP + FN},$$

where

- TP is the number of students who failed the course and were identified as at-risk.
- TN is the number of students who passed the course and were not identified as at-risk.
- FP is the number of students who failed the course but were not identified as at-risk students
- FN is the number of students who passed the course but were identified as at-risk students.

As the goal of this paper is to identify at-risk students, it is important to reach high predictive accuracy for the students who failed the course. In addition to the overall accuracy of the CBR model, we also calculated the accuracy of the model for students who failed and passed the course:

$$Accuracy(Pass) = \frac{TN}{Number\ of\ passed\ students} = \frac{TN}{TN + FP},$$

$$Accuracy(Fail) = \frac{TP}{Number\ of\ failed\ students} = \frac{TP}{TP + FN}.$$

The results of testing CBR prediction model for week 3, week 5 and week 7 are reported in Table II, Table III, and Table IV, respectively. Figure II summarizes the accuracy of the predictions for the three weeks.

TABLE II
PREDICTION RESULTS FOR WEEK 3

	Actual outcome		
Prediction	Failed	Passed	Total
Failed	140	84	244
Passed	51	173	204
Total	191	257	448

TABLE III
PREDICTION RESULTS FOR WEEK 5

	Actual outcome		
Prediction	Failed	Passed	Total
Failed	154	69	223
Passed	37	188	225
Total	191	257	448

TABLE IV
PREDICTION RESULTS FOR WEEK 7

	Actual outcome		
Prediction	Failed	Passed	Total
Failed	160	60	227
Passed	31	197	221
Total	191	257	448

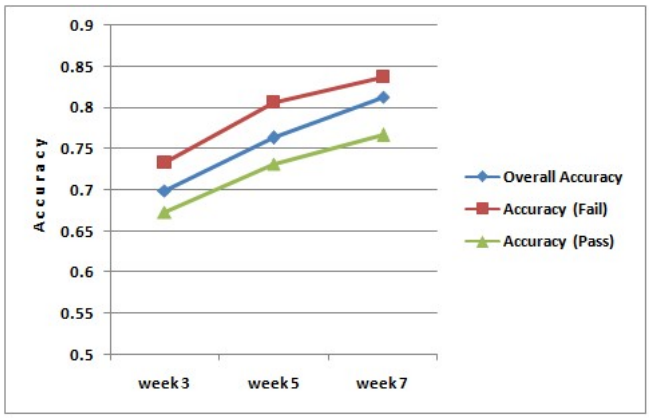

FIGURE II
ACCURACY OF THE MODEL FOR THE THREE WEEKS

The overall accuracy of the predictions made after three weeks was 69.87%. On the other side, the accuracy for the students who passed was 67.31%, and the accuracy for the students who failed was 73.29%. Prediction accuracy of 73.29% at the end of week 3 for the students who failed indicates that these students, despite all teachers' efforts during the semester, never improved their educational achievements in the course. Identification of these students early in the semester and providing them personalized and adapted learning paths might be one way to help these students.

Furthermore, the overall accuracy of predictions, as well as accuracy of failed and successful predictions, as it was expected, increases at the next two specific points in time during the first half of the semester: the end of week 5 and the end of the week 7. Prediction model in general has sufficient predictive power. Having in mind that it is crucial to identify at-risk students as early as possible, and taking into account that prediction accuracy of failures made at the end of week 7 is around 10% higher than prediction accuracy of failures made at the end of week 3, we can also conclude that week 3 predictions can provide an early warning indicator to identify at-risk students. Teachers can then use a variety of strategies to help these students and provide them with personalized and adapted learning paths for increasing their success in a course.

CONCLUSIONS AND FUTURE WORK

In this paper, we presented a course specific case-based reasoning model to identification of at-risk students at the three following specific points in time during the first half of the semester: end of week 3, end of week 5, and end of week 7. According to the model, new students' success prediction problems are solved by using case-based classification. The model uses the *k*-NN algorithm to retrieve the most similar cases that represent previously experienced at-risk student identification problems. The crucial element of the proposed model is the formal framework for representation structure for case encoding. Furthermore, this paper reports preliminary experimental evaluation of the proposed model. In general, preliminary results of evaluations showed that the presented prediction model based on case-based classification can provide significant support for prediction of at-risk students, but certainly more research is needed to show whether the proposed model is adequate for predictions in courses characterized by different course-specific learning objectives, different types of learning activities and learning resources, etc. Because the experimental results are to a certain extent dependent on the applied similarity measures and on the distribution of attribute weights, further research should be also devoted to include more different similarity measures and more different attribute weighting techniques.

ACKNOWLEDGMENT

This work is supported by the Ministry for Education, Science and Youth, Canton Sarajevo, Bosnia and Herzegovina, under Agreement No. 11/05-14-277724-1/19.

REFERENCES

[1] Yang, F., Li, Li, F.W.B., and Lau, R. W. H. 2014. "A Fine-Grained Outcome-Based Learning Path Model." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44(2): 235-245

[2] Supic, H. 2009. "An Adaptive Multimedia System for Teaching Fundamentals of Finite Element Method Using the Case-based Content Sequencing." *Proceedings* of 2009 International Conference of Computer Science and Engineering, 2009, July 2009, London, UK, Vol. I, pp. 174-178

[3] Supic, H. 2016. "A model of a case-based approach to context-aware content sequencing in Mobile learning environments." *Proceedings* of 2016 International Conference on Information and Digital Technologies (IDT), Rzeszow, Poland.

[4] Marbouti, F., Diefes-Dux, H.A and s Strobel, J. 2015. "Building course-specific regression-based models to identify at-risk students." *Proceedings* of 122nd ASEE Annual Conference and Exposition

[5] Marbouti, F. Diefes-Dux, H.A., and Madhavan, K. 2016. "Models for early prediction of at-risk students in a course using standard-based grading." *Computers and Education*, Elsevier

[6] Shayan, P. and Zaanen, M. 2019. "Predicting student performance from their behavior in learning management systems." *International Journal of Information and Education Technology*, Vol. 9, No. 5, Emerald Publishing.

[7] Chen Y., Zheng, Q., Ji, S., Tian, et. al. 2019. "Identifying at-risk students based on the phased prediction model." *Knowledge and Information Systems*, Springer-Verlag-London.

[8] Barata, G., Gama, S., Jorge, J., Goncalves, D. 2018. “Early Prediction of Student Profiles based on Performance and Gaming Preferences.” *IEEE Transactions on Learning Technologies*, Vol. 9, No. 3, IEEE Computer Society, pp. 272-284
[9] L. Mantaras, et al. “Retrieval, reuse, revision, and retention in CBR.” *Knowledge Engineering Review*, 20(3), 2005, pp. 215-240.
[10] Watson, I. and Marir, F., 1994. “Case-based reasoning: A review.” *The knowledge Engineering Review*. Vol 9:4, pp. 327-354
[11] Kolodner, J. 1993. *Case–based reasoning*. Morgan Kaufmann Publishers, Inc., San Mateo, CA.
[12] Janssen, R., Spronck, P., and Arntz, A. 2014. “Case-based reasoning for predicting the success of teraphy”, *Expert Systems*, Wiley Publishing.

AUTHOR INFORMATION

Haris Supic, Professor, Department of Computing and Informatics, Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina.

Dzenana Donko, Professor, Department of Computing and Informatics, Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina.