Providing a Method to Predict of Students' Academic Status in order to Improve Quality of Educational Process

Fahimeh Hakimi[1], Abdolreza Rasouli Kenari[2], HamidReza Hakimi[3] and Mahboubeh Shamsi[4]
[1] Department of Electronic and Computer Engineering, Qom Ta'ali University, Qom, Iran
FH.hakimi@Gmail.com
[2] Department of Electronic and Computer Engineering, Qom University of Technology, Qom, Iran
Rasouli@Qut.ac.ir
[3] Department of Mechanical Engineering, Foolad Shahr University of Technology, Isfahan, Iran
Hakimi.1743@Gmail.Com
[4]Department of Electronic and Computer Engineering, Qom University of Technology, Qom, Iran
Shamsi@Qut.ac.ir

Received: Sep 2014 Revised: Dec 2014 Accepted: Feb 2015

ABSTRACT
Funding for training human resources in most countries is very important and costly. Hence in training, prediction of students with expelled risky is one of the today's key issues and researches. There are imbalances in the training data that causes reduce prediction accuracy in fail students. In this paper, experiments based on data mining techniques have been tried to improve prediction accuracy of fail students. To do this, data from the UCI site are used that contains 5820 records in Turkish students. First, the training data are clustered to select the most appropriate algorithm, Farthest First, and then by doing experiments, best Features including the fitness level of instructor and students' attendance level and the best Test option, 90% for training data, are selected. Questionnaire is used to Weight features. Finally, the cost-sensitive classification algorithms have been implemented with the proposed cost matrix and the model provided the best results. Results prove that this model can play an important role in promoting science education centers with an accuracy rate of 96.47%, TP rate 99.2% and precision rate of 96%

KEYWORDS: Educational Data Mining (EDM), Prediction student status,Cost sensitive Classification, Select Feature, Fail student

1. Introduction
One of the challenges of research [2] has been detection of the features that affect in Prediction of Students' academic Status (typically fail student) at all levels of education using the large amount of data in the database at education centers. discover useful information and rules in these large databases is a difficult task [3]. One solution to this problem is the use of EDM. Data mining in education is called educational data mining (EDM) [4]. In this area, researchers are trying to provide a method to increase students learning [5]. there are examples of data mining in educational to predict of Students' academic Status and student failure typically [6]. These researchs have shown Goodish results With regard to those economic, sociological, or educational characteristics that may be more relevant in the prediction of Students' academic Status[7]. most of the research on educational data mining have Focused to predict of fail [8] and more specifically to online or distance education [9].

The rest of the paper is structured as follows; in the Section 2, proposed method for predicting of Students' academic Status is presented. Section 3, used data and the information sources are described. Section 4, the data preprocessing is described. Section 5, the different experiments carried out are described. In section 6, the interpretation of results is presented. Finally, section 7 concludes the paper.

2. Method
In this article a method is proposed. In the method is used data mining phases (see Fig1).

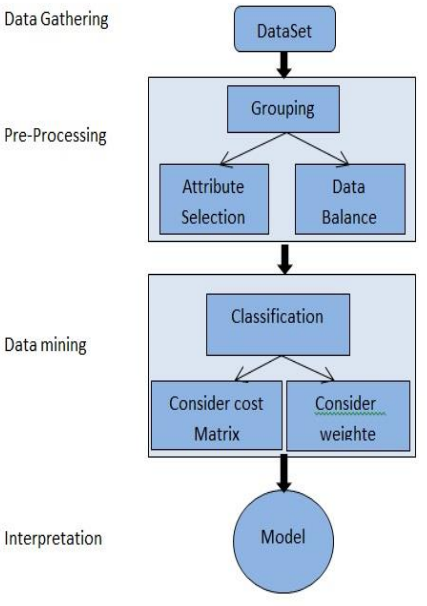

Fig1. The proposed Method

1) Data Collection: in this section, standard data students is used from machine learning site. This data set contains 33 attributes and 5820 records.

2) Pre-processing: This set of data are raw data, so clustering techniques for grouping data is applied. Then selection attributes algorithms are execute to reduce the problems of high dimensionality. There are imbalanced data problem in these dataset so SMOT algorithms is used.

3) Data mining: In this section, in order Predicting students' academic status, data mining algorithms are applied. In this paper, using decision tree and rule-based algorithms have been proposed. Moreover, the cost-sensitive classification method, weighting features, by distributing questionnaires, are used to solve the problem of imbalanced data and identify the most effective attributes.

4) Interpretation: At this section, the obtained models are analyzed to detect student failure. For this purpose, obtained rules and decision tree are analyzed. Factors and relationships that lead students to be expelled or succeeded are interpreted.

3. Data Collection

In this section, standard data set is used from UCI site.This data set contains a total 5820 evaluation scores. There are a total of 28 course specific questions and additional 5 attributes that are shown in table 1. Q1-Q28 are all Likert-type, meaning that the values are taken from {1,2,3,4,5}.

Table1: Features in educational dataset

course specific questions
Q1: The semester course content, teaching method and evaluation system were provided at the start.
Q2: The course aims and objectives were clearly stated at the beginning of the period.
Q3: The course was worth the amount of credit assigned to it.
Q4: The course was taught according to the syllabus announced on the first day of class.
Q5: The class discussions, homework assignments, applications and studies were satisfactory.
Q6: The textbook and other courses resources were sufficient and up to date.
Q7: The course allowed field work, applications, laboratory, discussion and other studies.
Q8: The quizzes, assignments, projects and exams contributed to helping the learning.
Q9: I greatly enjoyed the class and was eager to actively participate during the lectures.
Q10: My initial expectations about the course were met at the end of the period or year.
Q11: The course was relevant and beneficial to my professional development.
Q12: The course helped me look at life and the world with a new perspective.
Q13: The Instructor's knowledge was relevant and up to date.
Q14: The Instructor came prepared for classes.
Q15: The Instructor taught in accordance with the announced lesson plan.
Q16: The Instructor was committed to the course and was understandable.
Q17: The Instructor arrived on time for classes.
Q18: The Instructor has a smooth and easy to follow delivery/speech.
Q19: The Instructor made effective use of class hours.
Q20: The Instructor explained the course and was eager to be helpful to students.
Q21: The Instructor demonstrated a positive approach to students.
Q22: The Instructor was open and respectful of the views of students about the course.
Q23: The Instructor encouraged participation in the course.
Q24: The Instructor gave relevant homework assignments/projects, and helped/guided students.
Q25: The Instructor responded to questions about the course inside and outside of the course.
Q26: The Instructor's evaluation system (midterm and final questions, projects, assignments, etc.) effectively measured the course objectives.
Q27: The Instructor provided solutions to exams and discussed them with students.
Q28: The Instructor treated all students in a right and objective manner.
additional attributes
instr: Instructor's identifier; values taken from {1,2,3}
class: Course code (descriptor); values taken from {1-13}
repeat: Number of times the student is taking this course; values taken from {0,1,2,3,...}
attendance: Code of the level of attendance; values from {0, 1, 2, 3, 4}
difficulty: Level of difficulty of the course as perceived by the student; values taken from {1,2,3,4,5}

4. Pre-processing

Before applying DM algorithms, it is necessary to carry out some pre-processing tasks such as cleaning,

integration, discretization and variable transformation [10].

At this stage, the clustering algorithms applied to grouping data. In previous research, the students were divided into two clusters: Pass and Fail. In this paper, it is suggested that students divide into three clusters: pass, moderate, Fail. By this clustering, prediction accuracy of educatinal status in students is enhanced because for moderate students, a separated group is defined and moderate students are not distributed incorrectly in other class. For clustering, different algorithms are existed. First, 4 algorithms Kmeans, SimpleKmeans, Farthest First, Em that is suitable for data collection were selected.

We need that the similarities between objects of the same cluster are maximized and the similarities between objects in different clusters are minimal. Since a number of distance measures have been introduced. The most commonly used is Euclidean distance and in this study is used the method which is defined by the following equation [11]:

$$Dist(p,q)=\left(\sum_{k=1}^{d}\left(p_k-q_k\right)^2\right)^{1/2} \quad (1)$$

Where p and q are data points and d is a number of dimensions.

The clustering algorithm are executed. distance between clusters are Calculated by formule 1. They are shown in table 2.

Table2: The calculation of the distance between clusters

Algorithm	Distance between clusters 0,1	Distance between clusters 0,2	Distance between clusters 1,2
EM	15.78	7.17	8.68
Farthest First	11.9	11.57	21.16
kmeans	8.44	6.88	15.25
Simplekmes	8.47	7.02	15.43

Table3: Details data clustering by Farthest First

	Number of instance	Percentage of instance	Label
Cluster 0	758	65%	Good
Cluster 1	186	16%	Fail
Cluster 2	220	19%	Moderate

According to this table, distance between clusters 1 and 2 in Farthest First algorithm is more than other algorithms. However, by comparing distance between clusters 1 and 0, it is found that the highest distance is associated with EM algorithm which is approximately 4.21 more than Farthest First algorithm. According to these results, Farthest First algorithm is strong in student groupings and it has better results.

The number of clusters obtained from the Farthest First algorithm implementation is three clusters. By examining the nature of samples in each cluster, they are labeled appropriately. Given the nature and number of samples, it can be found that cluster 1 consists of minority class. The results are shown in table 3.

The dataset used in this study contains 33 features. To reduce the dimensions space, feature selection algorithms have been used. Weka provides several feature selection algorithms. To do this, CfsSubSetEval and PrincipalComponet algorithms have been used. attributes subset obtained from this algorithm are shown in Table 4. The mean results of white boxes classification algorithms have been used for evaluating.it means that classification algorithms are executed on 33 features, 20 features and 14 selected features. In this evaluation, the different percentage of learning data is evaluated to determine the best subset of features. The results of this evaluation are shown in Tables 5,6,7,8.

Table4: selected attributes by algorithms

Algorithm	Selected attribute
PrincipalComponet	Repeat, attendance, difficulty, Q1,Q2,Q3,Q4,Q5,Q6,Q7,Q8,Q9,Q10,Q11
CfsSubSetEval	attendance, difficulty,Q2,Q3,Q4,Q5,Q6,Q9,Q10,Q13,Q14,Q15,Q16,Q20,Q21,Q23,Q24,Q25,Q26,Q28

Table 5: Classification results using the all attributes obtained

Test option	Average of TP rate	Average of accuracy rate	Average of precision rate
%60	%72.7	%83.3	%77.12
%70	%69.2	%84.1	%78.92
%80	%72.5	%85.01	%86.26
%90	%73	%85.81	%67.01
Tenfold cross validation	%73.12	%85.9	%77.12

Table 6: Classification results using the attributes obtained from Greedy algorithm

Test option	Average of TP rate	Average of accuracy rate	Average of precision rate
%60	%81.20	%86.39	%84.59
%70	%81.25	%85.40	%87.2
%80	%85.48	%87.88	%87.11
%90	%83.61	%91.12	%89.12
Tenfold cross validation	%86.1	%91.93	%89.15

Table 7: Classification results using the attributes obtained from Ranker algorithm

Test option	Average of TP rate	Average of accuracy rate	Average of precision rate
%60	%72.7	%84.3	%77.18
%70	%70.6	%84.76	%79.92
%80	%73.77	%85.90	%86.26
%90	%74.16	%86.68	%90.1
Tenfold cross-validation	%74.5	%87.1	%66.1

However, existing results table are compared in the previous. According to this comparison, best subset of features and learning data percentage are specified.

As it can be observed in table 5, the maximum average of TP rate and accuracy is respectively 73.12%, 85.9%, and 78.92%. As it can be observed in table 6, the maximum average of TP rate and accuracy is respectively 74.5%, 86.68%, and 90.1%. As it can be observed in table 6, the maximum average of TP rate and accuracy is respectively 86.1%, 91.93%, and 89.15%.

By comparing three previous tables, it is found that the subset of attributes obtained from the greedy algorithm has the highest TP rate and average accuracy rate. However, by comparing the accuracy rate, it is found that the highest accuracy rate belongs to Ranker which is approximately 0.85 more than Greedy algorithm, while it has lower TP and accuracy rates compared to Greedy. Given that in TenFold Cross Validation method, 20 features selected by Greedy algorithm are stronger in predicting student's academic status; therefore, it has a better result because it provides more accuracy and TP rate.

5. Data mining and experimentation

This section describes the experiments and data mining techniques used for obtaining the prediction models of students' academic status semester.

In this article, several experiments are performed in order to obtain the best result. In a first experiment 10 classification algorithms are applied by using the selected attributes selected (20 attributes). In a second experiment, we repeated the executions by using weighting attributes. In the third experiment we considered different costs in the classification. In a final experiment, the combination of previous experiments is done. It means that classification algorithms are executed with the best weighting attributes by attending cost matrix. To do this, popular weka data mining software and 10 universally classification algorithms is used. Result of this experiment shows the rates correct classifications for minority class (TP rate for fail students), the total Accuracy rate (Acc) and the overall precision rate.

In the first experiment, classification algorithms were executed using tenfold cross-validation and with the selected attributes. The results of classification algorithms are shown in Table 8.

All attributes are as not equal as others. Weighting specified the importance of each attribute relative to other attributes. In this study, the average weight for weight attributes have obtained from questionnaires. This questionnaire is designed to work. The questionnaire was distributed among experienced instructor. The questionnaire‘s questions are the same features found in data set. To obtain the desired number of questionnaires, Cochran's formula is used in which 193 questionnaire were distributed. To gain weight attributes, mean scores of each feature were calculated separately for each feature.

There are several criteria for determining the position of attributes in the decision tree. One of these criteria is entropy calculation and information interest. To validate the weights obtained from the questionnaire, entropy and information gain were calculated. The weighted correlation coefficient obtained from the questionnaire and the information gain is calculated. Spss software is used to do this. Given that the correlation coefficient was 0.682, so there is a significant correlation between statistically. So, obtained weights from questionnaire distribution are reliable. In the second experiment, the classification algorithms are executed using the best weighted features and tenfold cross-validation. The results of classification algorithms are shown in Table 9.

cost-sensitive classification can be used to solve the problem of imbalanced data. In this article, The main objective is detection Fail students (the minority class) than other students (the majority class). In the case of three classes, costs can be placed into a 3 × 3 matrix. In the problem have used the values of the matrix [0, 1; 4, 0] as the cost matrix. these values are obtained by Trial and error. This matrix indicates that performing the classification takes into consideration that it is four times more important to correctly classify Fail students than Pass students.

the meta classification algorithm Cost Sensitive Classifier has been used for cost-sensitive classification that are available in weka software.

In the third experiment, we executed the cost-sensitive classification algorithms with The proposed cost matrix and using tenfold cross-validation (with the selected 20 attributes). Table 10 shows the results.

In the fourth experiment, classification algorithms are executed using tenfold cross-validation and consideration the cost matrix and the best of weighted attributes. This test is a combination of previous experiments. In fact, the cost-sensitive classification

with the best of weighted attributes is implemented. The results are shown in Table 11.

Table 8: Classification results using the the selected attributes

Algorithm	ACC	TP Rate	Precision
JRip	94.75	93.5	94.8
OneR	95.01	92.5	95
NNg	90.2	82.3	90.3
Ridor	95.01	93	95
FT	93.9	88.2	93.9
J48	94.58	94.1	94.6
RepTree	95.61	94.1	95.6
Random Forest	96.56	94.1	96.6
LADtree	92.78	87.1	92.8
SimpleCart	93.9	90.3	93.9

Table 9: Classification results using the best weighting attributes

Algorithm	ACC	TP Rate	Precision
JRip	94.5	95.6	94.5
OneR	95.01	97.2	95
NNg	90.2	95.01	90.3
Ridor	95.01	96.06	95
FT	93.9	95.5	93.9
J48	94.58	95.3	94.6
RepTree	95.61	96.7	95.6
Random Forest	96.21	97.9	96.2
LADtree	92.86	94.9	92.9
SimpleCart	93.9	95.1	93.9

Table 10: Classification results using the best attributes and cost Matrix

Algorithm	ACC	TP Rate	Precision
JRip	95.44	97.8	95.5
OneR	94.75	95.2	94.8
NNg	88.23	86	88.2
Ridor	93.72	93	93.8
FT	92.35	96.2	92.7
J48	91.40	94.1	91.5
RepTree	94.15	93.5	94.3
Random Forest	96.04	96.8	96.1
LADtree	92.43	97.3	92.8
SimpleCart	92.43	96.2	92.8

Table 11: Classification results using the best weighting attributes and cost Matrix

Algorithm	ACC	TP Rate	Precision
JRip	95.88	97.1	95.2
OneR	93.12	97.9	93.3
NNg	89.60	95.9	89.7
Ridor	95.18	97.8	95.2
FT	91.92	97.2	92.1
J48	92.43	97.1	92.5
RepTree	94.67	97	94.7
Random Forest	96.47	99.2	96.5
LADtree	95.44	97.9	95.5
SimpleCart	93.29	96	93.3

6. Interpretation of results

In this section, some rules obtained by some of the algorithms are shown in order to compare their interpretability and usefulness for early identification of students with risk of failing and for making decisions about how to help this student. These rules show us the relevant factors and relationships that result a student to pass or middle or fail class.

Results in Table 8 shows the results of classification algorithms on the best features. In this table, the highest TP rate is equal to 94.1% that is achieved from LADTree,J48 and Random Forest. Algorithm J48 is obtained the highest accuracy (95.61%) and Precision (95.6%). Table 9 is obtained by performing classification algorithms with best weighed attributes. The best results presented in this table are as follows: algorithm Random Forest with TP rate 97.9% and accuracy rates 96.21% and total precision 96.2%. as presented in table 10 has better results related to the other algorithm with algorithm Random Forest with TP rate 96.8% and accuracy rates 96.4% and total precision 96.1%. The results in Table 11, the execution of cost-sensitive classification algorithms using the proposed cost matrix and the weighted features is obtained. In this method, an TP rate 99.2% and an overall accuracy rate 96.47% and total precision 96.5% is obtained. The obtained results of the latest test and other tests have the highest TP rate, precision and accuracy. Therefore, this method becomes more powerful in predicting student academic status.

In the model shown in Table 12 it is observed that the algorithm JRip discovers few rules. With respect to the attributes that are associated to Fail, we found that Level of preparedness instructor and level of attendance student have the most impact.

Furthermore, the decision tree obtained from the algorithm SimpleCart is shown in Table 13. In the Table, according the features of the fail students are correlated, it becomes clear that the fitness instructor level and the student attendance level have the greatest impact on the school. So the decision tree is used to emphasize that the important feature is the presence of a teacher.

Table12: rules that obtained from JRip using the selected attributes and cost classification on balance data

(Q14 <= 0.05) => Academic Status = Fail
(Q25 <= 0.1) and (attendance <= 0.3) and (Q4 <= 0.09) and (difficulty <= 0.01) => Academic Status = Fail
(Q25 <= 0.1) and (Q2 <= 0.02) and (Q24 <= 0.04) => Academic Status = Fail
(Q10 >= 0.15) and (Q28 >= 0.175) => Academic Status = Moderate
(Q4 >= 0.18) and (attendance <= 0) and (difficulty <= 0.01) and (Q23 >= 0.14) => Academic Status = Moderate
(Q24 >= 0.2) and (Q10 >= 0.12) and (Q13 >= 0.25) => Academic

```
Status = Moderate
(Q25 >= 0.25) and (Q5 >= 0.12) and (Q16 >= 0.2) and (Q4 >=
0.225) => Academic Status = Moderate
 (Q24 >= 0.2) and (Q9 >= 0.16) and (Q5 >= 0.12) => Academic
Status = Moderate
 => => Academic Status =Good
```

Table13: Tree obtain from SimpleCart using the selected attributes and cost classification on balance data

```
Q14 < 0.08 : fail
Q14 >= 0.08
|  Q4 < 0.2
|  |  Q10 < 0.11
|  |  |  Q28 < 0.09
|  |  |  |  Q2 < 0.05
|  |  |  |  |  attendance < 0.15
|  |  |  |  |  |  Q2 < 0.03 : fail
|  |  |  |  |  |  Q2 >= 0.03
|  |  |  |  |  |  |  Q15 < 0.11
|  |  |  |  |  |  |  |  Q25 < 0.13 : fail
|  |  |  |  |  |  |  |  Q25 >= 0.13 : Good
|  |  |  |  |  |  |  Q15 >= 0.11 : Good
|  |  |  |  |  attendance >= 0.15
|  |  |  |  |  |  Q21 < 0.05 : fail
|  |  |  |  |  |  Q21 >= 0.05 : Good
|  |  |  |  Q2 >= 0.05 : Good
|  |  |  Q28 >= 0.09 : Good
|  |  Q10 >= 0.11
|  |  |  difficulty < 0.02
|  |  |  |  attendance < 0.45
|  |  |  |  |  Q6 < 0.12 : Good
|  |  |  |  |  Q6 >= 0.12 : Moderate
|  |  |  |  attendance >= 0.45 : Good
|  |  |  difficulty >= 0.02
|  |  |  |  Q24 < 0.18 : Good
|  |  |  |  Q24 >= 0.18
```

7. Conclusions

Predicting students' academic status is not only a difficult task but is an issue of several factors (personal factors, family, society, economy). There are imbalances in the educational data and reducing forecast accuracy. This study is an attempt to implement a predictive data mining model to predict fail student academic status. The consultation can be prevented by timely expulsion of fail students. The proposed method with an accuracy of 97% can be used as decision support tools in order to improve quality of Educational Processes.

REFERENCES

[1] C. Romero and S. Ventura, (Eds**), "Data Mining int- Learning"**, 2006, pp.261-278

[2] C. Romero, S. Ventura, P. G. Espejo and C. Hervás, **"DataMining Algorithms to Classify Students",** The 1st International Conference on Educational Data Mining Proceedings, Montreal, Quebec , Canada, June 20-21, 2008.

[3] E. Ayers, **"Rebecca Nugent and Nema Dean, Skill Set Profile Clustering Based on Weighted Student Responses",** The 1st International Conference on Educational Data Mining Proceedings, Montreal, Quebec, Canada, June 20-21, 2008.

[4] N. Delavari, M. R. Beikzadeh, 2004, **"A New Model for Using Data Mining in Higher Educational System",** 5th International Conference on Information Technology based Higher Education and Training: ITEHT '04, Istanbul, Turkey, 31st May-2nd Jun 2004.

[5] P. Varapron, et al. 2003**, "Using Rough Set theory for Automatic Data Analysis".** 29th Congress on Science and Technology of Thailand.

[6] K. Mierle, K. Laven, S. Roweis, G. Wilson, 2005, "**Mining Student CVS Repositories for Performance Indicators".**

[7] N. Delavari, M. R. Beikzadeh, S. Amnuaisuk, 2005**, "Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System"** 6th Annual International Conference: ITEHT, July 7-9, 2005, Dolio, Dominican Republic.

[8] C. Márquez-Vera, C.R Morales, S. Ventura Soto, **Predicting School Failure and Dropout by Using Data Mining Techniques, IEEE** Journal of Latin-American learning technologies, vol. 8, no. 1, 2013.

[9] H. Bydovska, L. Popelinsky, **Predicting Student on Database Performance in Higher Education 24th International Workshop and Expert Systems Applications (DEXA),** 2013,pp.141 – 145.

[10] E. Espíndola and A. Leon, **"La desercion escolar en america latina: Un Tema prioritario parala agenda regional",** Revista Iberoamer. Educ., vol. 1, no. 30, pp. 39–62, 2002.

[11] P. Lasek, M. Sc., **Efficient Density-Based Clustering,** Warsaw university of technology, 2011