www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 12 December 2014, Page No. 9395-9398

Using Educational Data Mining (EDM) to Prediction and Classify Students

Samira Talebi[1],Ali Asghar Sayficar[2]

[1]Islamic Azad University Garmsar Branch, Department ofInformation Technology,
University Square, Student Street, Iran
samiratlb86@gmail.com

[2]Islamic Azad University Garmsar Branch, Department ofInformation Technology,
University Square, Student Street, Iran
a_sayficar@yahoo.com

Abstract:***The aim of this paper is to predict the students' academic performance. It is useful for identifying weak students at an earlier stage. In this study, we used WEKA open source data mining tool to analyze attributes for predicting students' academic performance. The data set comprised of 180 student records and 21attributes of students registered between year 2010 and 2013. We chosethem from FERDOWSIUniversity of Mashhad .We applied the data set to four classifiers (Naive Bayes, LBR,NBTree, Best -First Decision Tree) and obtained the accuracy of predicting the students' performance into either successful or unsuccessful class. The student's academic performance can be predicted by using past experience knowledge discovered from the existing database. A cross-validation with 10 folds was used to evaluate the prediction accuracy. The result showed that Naive Bayes classifier scored the higher percentage of prediction F-Measure of 83.9%.***

Keywords: Data Mining, Prediction, Average, Attributes for predicting students, Educational Data Mining (EDM)

1.Introduction

Classification and prediction are of highimportance in data mining techniques and usedin many fields. Recently, researchers haveutilized machine learning in order to makewise career decisions. It is useful for both thestudents and the instructors getting better intheir performances. We got our dataset fromthe Information system of the biggest virtualuniversity of Iran. We decided to extract theattributes that have significant contribution tothe prediction of academic performance. Theprediction can be done by using data miningtools such asWeka software.

2. Methodology

Many studies were undertaken in order to explain the academic performance or to predict the success or the failure (Kotsiantis *et al*., 2003; Chamillard,2006;Minaei-Bidgoli *et al*., 2003;Merceron and Yacef, 2005; Romero *etal*., 2008;Superby *et al*.,2006;Vandamme *et al*., 2007;Ardila, 2001; Gallagher, 1996; King,2000;Minnaert and Janssen, 1999;Parmentier, 1994.)they highlighted a series of explanatory factors associated to the student.
We first considered a set of attributes to betaken into account based on a model used byParmentier (1994). Secondly, we created aquestionnaire allowing us to collect a largeamount of interesting information on a certainnumber of students. We distributed thisquestionnaire by paper to students in theFERDOWSI University of Mashhad.

We used WEKA open source data mining. Itsupports many machine learning algorithmsand data processing tools. In the datapreprocessing step, we collected 205 records of students admitted from year 2010 to 2013 atthe FERDOWSIUniversity of Mashhad.According to the total average, thestudents were classified into four classes:
Grade A (Total Average>=17),
Grade B (15=<Total Average< 17),
Grade C (13=<Total Average< 15),

Grade D (Total Average<13)

We splitedthe data for training(119 records ˜ 66%)and testing (61 records˜ 34%). We used the Naïve Bayes, LBR, NBTree and Best-First Decision Tree classifiers for prediction. Table 1 shows the attributes and their valid values we considered for predicting student's academic performance.

Table 1: The attributes used for classification

Attribute	*Value*
Sex	Female / male
Marital status	Single / married
Job status	Employed/ unemployed
City	Mashhad / others
Right handed or left handed	Right hand/ left hand
The method study	Solo/with the group
How to study	During the semester/the night before the exam
Do my projects	Alone, use the preparation projects
The source of the study	Booklet, reference
Diploma average	A/B/C/D
The First university semester average	A/B/C/D
The amount of interest in the field of	Very high/high/medium /low/very low
Internet accessibility	Very high/high/medium /low/very low
Break between high school and university	Very high/high/medium /low/very low
Mother’s level of education	Very high/high/medium /low/very low
Type of high school in the pre-university course	Very high/high/medium /low/very low
The number of terms has fallen	Very high/high/medium /low/very low
The number of children of the family	Very high/high/medium /low/very low
The rate of attendance in class	Very high/high/medium /low/very low
English language level	Very high/high/medium /low/very low
Total Average	A/B/C/D

2.1.Confusion Matrix

A confusion matrix (Kohavi and Provost,1998) contains information about actual andpredicted classifications done by aclassification system. Performance of such

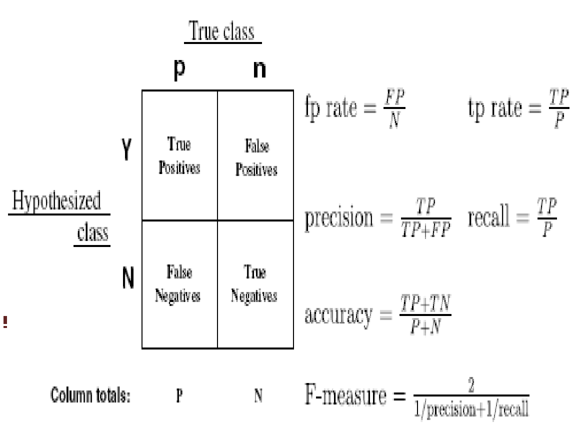

Fig.1. Confusion matrix and common performancemetrics calculated from it.

3. Results

In this paper we used the Naïve Bayes, LBR, NBTree and Best-First Decision Treeto predict student's academic performance. A crossvalidation with 10 folds was used to evaluate the prediction accuracy.

3.1Best-First Decision Tree

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.633     0.09      0.745     0.633     0.685      0.772    H1
                 0.915     0.082     0.768     0.915     0.835      0.916    H2
                 0.814     0.116     0.738     0.814     0.774      0.849    H3
                 0.821     0.006     0.97      0.821     0.889      0.907    H4
Weighted Avg.    0.785     0.08      0.791     0.785     0.784      0.853

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 38  9 12  1 |  a = H1
  3 43  1  0 |  b = H2
  9  2 48  0 |  c = H3
  1  2  4 32 |  d = H4
```

Fig.2. Summary of the results of Best-First Decision Tree

As shown in fig 2, the proportion of correct predictions for class H2 are good: 91.5 of thestudents of class H2 were correctly classified bymeans of the Naïve Bayes classifier; but theproportion of correct predictions for class H1 arebad, only 63.3% of the students of class H1 were actually classified into class H1.The weighted average of F-Measure is 78.4% andthis is not such a good result.

3.2NBTree

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.65      0.076     0.78      0.65      0.709      0.916    H1
                 0.915     0.057     0.827     0.915     0.869      0.974    H2
                 0.831     0.103     0.766     0.831     0.797      0.941    H3
                 0.949     0.012     0.949     0.949     0.949      0.997    H4
Weighted Avg.    0.82      0.067     0.819     0.82      0.816      0.952

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 39  6 13  2 |  a = H1
  2 43  2  0 |  b = H2
  7  3 49  0 |  c = H3
  2  0  0 37 |  d = H4
```

Fig.3. Summary of the results of NBTree

As shown in fig 3, the proportion of correctpredictions are better than Best-First Decision Tree, 65% of the students of class H1were correctly classified by means of the NBTreeclassifier; and 94.9% of the students of class H4were actually classified into class H4.The weighted average of F-Measure is 81.6% andthis is a good result.

3.3LBR

Fig 4 shows a summary of the results of LBRclassifier.

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.883     0.069     0.841     0.883     0.862      0.972    H1
                 0.851     0.044     0.851     0.851     0.851      0.939    H2
                 0.814     0.103     0.762     0.814     0.787      0.939    H3
                 0.769     0.012     0.938     0.769     0.845      0.964    H4
Weighted Avg.    0.834     0.062     0.839     0.834     0.835      0.953

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 53  3  4  0 |  a = H1
  2 40  5  0 |  b = H2
  6  3 48  2 |  c = H3
  2  1  6 30 |  d = H4
```

Fig.4. Summary of the results of LBR classifier

As shown in fig 4, the proportion of correctpredictions for class H1 are better than Best-First and LBR classifier: 88.3% of the students of class H1were correctly classified bymeans of MLP classifier; and the proportion ofcorrect predictions for class H4 are better than
Best-Firstbut is equal to NBTree classifier:
76.9% of the students of class H4 were actuallyclassified into class H4.Theweighted average of
F-Measure is 83.5% and this is a good result.

3.4Naive Bayes

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.9       0.103     0.783     0.9       0.837      0.918    H1
                 0.83      0.032     0.886     0.83      0.857      0.961    H2
                 0.78      0.068     0.821     0.78      0.8        0.908    H3
                 0.846     0.018     0.917     0.846     0.88       0.968    H4
Weighted Avg.    0.839     0.061     0.843     0.839     0.839      0.935

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 54  2  3  1 |  a = H1
  4 39  3  1 |  b = H2
 10  2 46  1 |  c = H3
  1  1  4 33 |  d = H4
```

Fig.5. Summary of the results of Naïve Bayes classifier

As you see in fig 5, the proportion of correctpredictions are the best of all: 90% of thestudents of class H1 were correctly classified bymeans of Naïve Bayes classifier; and 78% of the students of class H3 were actually classified into class H3.the weighted average of F-Measure is 83.9% and this is a very good result.

4. Conclusion

Identifying the classifiers that contribute the
Mostsignificant to predict student's academic performance can help to improve theintervention strategies and support services forstudents who perform poorly in their studies,at an earlier stage. The objective of this studywas to introduce and compare some techniquesused to predict the student performance at aAzad university of Mashhad. This is importantas it provides groundwork for furtherevaluation of the program. The findings of thisstudy showed that Naïve Bayes classifierscored the higher percentage of prediction FMeasure of 83.9%. Moreover, the ROC area ofLBR classifier is better than otherClassifiers.

References

[1] B.K. Bharadwaj and S. Pal. " Mining Educational Data to Analyze Students' Performance", *International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp*. 63-69, 2011.

[2] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", *International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp.* 136-140, 2011.

[3] U.K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", *(IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN: 0975-9646*, 2011.

[4] U. K. Pandey, and S. Pal, "A Data mining view on class room teaching language", (IJCSI) *International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN: 1694-0814*, 2011.

[5] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, "Data mining model for higher education system", *Europen Journal of Scientific Research, Vol.43, No.1, pp*.24-29, 2010

[6] Data Mining: A Prediction for Performance Improvement of Engineering Students using Classificatio *World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012*

[7] Kumar, Varun, and Anupama Chadha. Mining Association Rules in Student's Assessment Data.*InternationalJournalofComputerScienceIssues9. 5:211-216*,2012.