www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 12 December 2014, Page No. 9606-9609

Using data mining in prediction of educational status

Samira Talebi[1],Ali AsgharSayficar[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 AZADUniversity 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 88.7%.*

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 be taken into account based on a model used by Parmentier (1994). Secondly, we created a questionnaire allowing us to collect a large amount of interesting information on a certain number of students. We distributed this questionnaire by paper to students in the "IAUM" Islamic Azad University of Mashhad.

We used WEKA open source data mining. It supports many machine learning algorithms and data processing tools. In the data preprocessing step, we collected 180 records of students admitted from year 2010 to 2013 at the "IAUM".

According to the total semesters average, the students were classified into four classes:
class [ma] (Average>= 17),
class [mb] (15=< Average< 17),
class [mc] (13=< Average< 15),
class[md] (Average <13).

We splited the data for training (119 records ˜ 66%) and testing (61 records˜ 34%). We used the Naïve Bayes, LBR, NB Tree 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 (Kohaviand Provost,1998) contains information about actual andpredicted classifications done by aclassification system. Performance of such

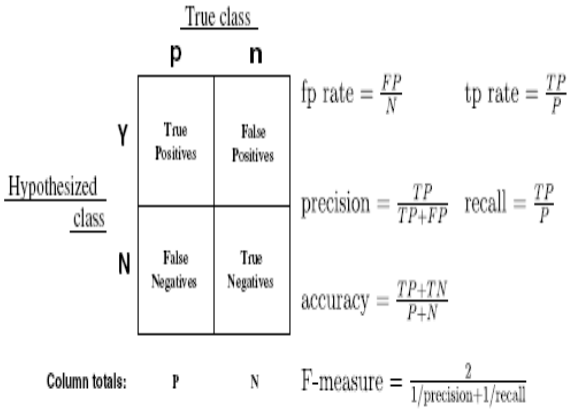


Fig.1. Confusion matrix and common performancemetrics calculated from it.

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.

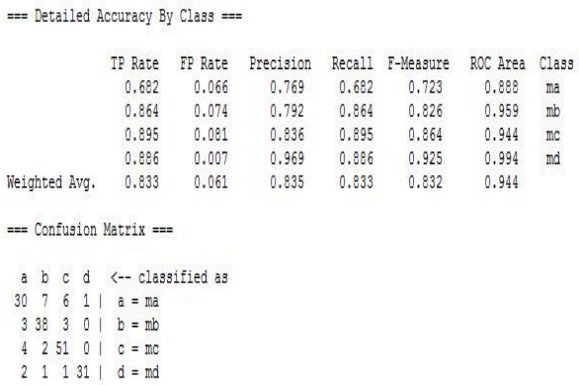
3.1Best-First Decision Tree

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

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.591     0.066      0.743     0.591     0.658      0.856    ma
                 0.818     0.11       0.706     0.818     0.758      0.891    mb
                 0.807     0.154      0.708     0.807     0.754      0.898    mc
                 0.686     0.034      0.828     0.686     0.75       0.952    md
Weighted Avg.    0.733     0.099      0.739     0.733     0.731      0.897

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 26 10  7  1 |  a = ma
  2 36  6  0 |  b = mb
  4  3 46  4 |  c = mc
  3  2  6 24 |  d = md
```

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

As shown in fig 2, the proportion of correct predictions for class [mb] is good: 81.8% of the students of class [mb] were correctly classified by means of the Naïve Bayes classifier; but the proportion of correct predictions for class [ma] is bad, only 59.1% of the students of class ma were actually classified into class [ma]. The weighted average of F-Measure is 73.1% and this is not such a good result.

3.2NBTree



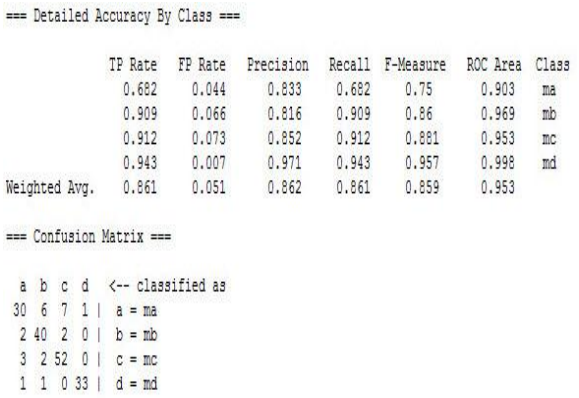
```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.682     0.066      0.769     0.682     0.723      0.888    ma
                 0.864     0.074      0.792     0.864     0.826      0.959    mb
                 0.895     0.081      0.836     0.895     0.864      0.944    mc
                 0.886     0.007      0.969     0.886     0.925      0.994    md
Weighted Avg.    0.833     0.061      0.835     0.833     0.832      0.944

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 30  7  6  1 |  a = ma
  3 38  3  0 |  b = mb
  4  2 51  0 |  c = mc
  2  1  1 31 |  d = md
```

Fig.3. Summary of the results of NBTree

As shown in fig 3, the proportion of correct predictions are better than Best-First Decision Tree, 68.2% of the students of class [ma] were correctly classified by means of the NB Tree classifier; and 89.5% of the students of class [ma] were actually classified into class [mc]. The weighted average of F-Measure is 83.2% and this is a good result.

3.3LBR

Fig 4 shows a summary of the results of LBRclassifier.

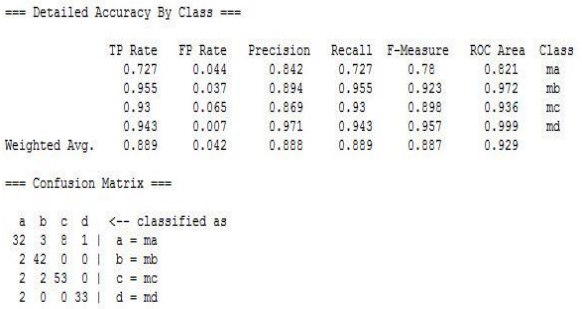


```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.682     0.044      0.833     0.682     0.75       0.903    ma
                 0.909     0.066      0.816     0.909     0.86       0.969    mb
                 0.912     0.073      0.852     0.912     0.881      0.953    mc
                 0.943     0.007      0.971     0.943     0.957      0.998    md
Weighted Avg.    0.861     0.051      0.862     0.861     0.859      0.953

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 30  6  7  1 |  a = ma
  2 40  2  0 |  b = mb
  3  2 52  0 |  c = mc
  1  1  0 33 |  d = md
```

Fig.4. Summary of the results of LBR classifier

As shown in fig 4, the proportion of correct predictions for class 1 are better than Best-First and LBR classifier: 94.3% of the students of class [md] were correctly classified by means of MLP classifier; and the proportion of correct predictions for class [ma] are better than Best-Firstbut is equal to NB Tree classifier: 68.2% of the students of class [ma] were actually classified into class [ma]. The weighted average of F-Measure is 85.9% and this is a good result.

3.4Naive Bayes



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

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.727     0.044      0.842     0.727     0.78       0.821    ma
                 0.955     0.037      0.894     0.955     0.923      0.972    mb
                 0.93      0.065      0.869     0.93      0.898      0.936    mc
                 0.943     0.007      0.971     0.943     0.957      0.999    md
Weighted Avg.    0.889     0.042      0.888     0.889     0.887      0.929

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 32  3  8  1 |  a = ma
  2 42  0  0 |  b = mb
  2  2 53  0 |  c = mc
  2  0  0 33 |  d = md
```

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

As you see in fig 5, the proportion of correct predictions is the best of all: 95.5% of the students of class [mb] were correctly classified by means of Naïve Bayes classifier; and 72.7% of the students of class [ma] were actually classified into class [ma]. The weighted average of F-Measure is 88.7% and this is a very good result.

4. Conclusion

Identifying the classifiers that contribute the most significant to predict student's academic performance can help to improve the intervention strategies and support services for students who perform poorly in their studies, at an earlier stage. The objective of this study was to introduce and compare some techniques used to predict the student performance at a Azad university of Mashhad. This is important as it provides groundwork for further evaluation of the program. The findings of this study showed that Naïve Bayes classifier scored the higher percentage of prediction F Measure of 88.7%. Moreover, the ROC area of LBR classifier is better than other Classifiers.

References

[1] Samira Talebi and Ali AsgharSayficar. "Using Educational Data Mining (EDM) to Prediction and Classify Students",*International Journal of Engineering and Computer Science ,Volume 3 Issue 12,ISSN : 2319-7242, 2014.*

[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.