Int. J. Advance Soft Compu. Appl, Vol. 11, No. 2, July 2019
ISSN 2074-8523

Dimensionality Reduction for Predicting Student Performance in Unbalanced Data Sets

Theng-Wai Lim[1], Kok-Chin Khor[2], and Keng-Hoong Ng[1]

[1]Faculty of Computing and Informatics, Multimedia University, Jalan Multimedia, 63100, Cyberjaya, Malaysia
[2]Lee Kong Chian Faculty of Engineering Science, Universiti Tunku Abdul Rahman, Bandar Sg. Long, Kajang 43000, Malaysia
e-mail: limthengwai@gmail.com, khorkc@gmail.com, khng@mmu.edu.my

Abstract

In this study, we evaluated two data sets from two Portuguese schools for predicting student performance. These data sets contain not only the previous grades of the students, but also the demographic, social and school related features. Both data sets are unbalanced in class distribution and contained some irrelevant features. Such characteristics may cause unsatisfactory True Positive (TP) rates for the Fail grade. This grade is important in prediction but it has a low representation as compared with the Pass grade. To improve prediction, dimensionality reduction was performed on both data sets to generate subsets that contained: (i) features selected by a wrapper approach, and (ii) only previous grade(s). The results showed that dimensionality reduction helped to improve the TP rates for the Fail grade. In addition, good classification accuracies were attained. We also noticed that even though the subsets contain only one previous grade, comparable accuracies can also be achieved.

Keywords: *Student Performance Prediction, Data Mining, Dimensionality Reduction.*

1 Introduction

Education systems are the key to provide human capital for achieving a sustainable economic development of a nation. Whether an education system is indeed a success or not is depending on the student performance. To help students to perform, educators pour their efforts to plan better education systems. Since the student performance is a concern, much research has been carried out to help educators. One of the research is about predicting the student performance. There are reasons why such prediction is important.

The research of predicting student performance has set forth in the early years and still ongoing till today. As early as 1963, the University of Aberdeen formed a commission to look into the matter of student failures [1]. Through the analyses of the committee, an "early warning system" was developed to identify students that would possibly fail in a course.

Recent decades, data mining approaches are common in predicting student performance [2-5]. For instance, an early work of Cortez and Silva (2008) predicted the performance of students from two Portuguese secondary schools using various classification algorithms [2]. The intention of the research was to ameliorate the quality of education with a better school resource management. This was achieved by identified the students that would perhaps fail in a subject ahead of time. The authors concentrated on two core subjects in the schools, namely, Portuguese Language and Mathematics. In the research, they modelled the class attribute (the final grade) into three supervised approaches: binary classification, 5-level classification and regression. The main features involved were the previous grades of the students. In addition, the demographic, social and school related features were also added to form the data sets.

As the research of predicting student performance progresses, the data mining techniques were also applied in Learning Management Systems (LMS) and Massive Open Online Courses (MOOC) [6-9]. Elbadrawy et al. (2016) utilised personalized analytics, such as multi-regression and matrix factorization approaches, to accurately forecast student grades and also in-class assessments [6]. The analytics also used transcript data in addition to LMS data and MOOC data.

Depending on subjects, the representation of students who fail could be low or even rare, causing the unbalanced class distribution in a data set. As a result, classification algorithms may not able to learn the detail of failed students effectively and may result the unsatisfactory classification rate problem.

The problem was well reported by Weiss (2004) [10] and various approaches had been proposed to handle the unbalanced class distribution. The study by Longadge et al. (2013) suggested to apply pre-processing techniques such as sampling to attain a better classification result for the problem [11]. For instance, Marquez-Vera et al. (2013) used Synthetic Minority Over-sampling Technique (SMOTE) to rebalance their data set and to consider different classification costs [12]. The techniques used were able to provide good classification results.

The problem could be worsened if the unbalanced data sets involved contain irrelevant features. Longadge et al. (2013) suggested that dimensionality reduction is important and may help to mitigate the unsatisfactory classification results of such unbalanced data sets [11]. In the work by Ramaswami and Bhaskaran (2009), the number of features in the student data were reduced using six feature selection algorithms for a better prediction in student performance [13]. A research by Acharya and Devadatta (2014) also showed that selecting relevant

features in the student data were able to maximize the classifier performance in predicting student performance [14]. There were also some recent research that attempted dimensionality reduction on student data to improve the prediction performance [15, 16].

In this study, we evaluated the effectiveness of dimensionality reduction in predicting the student performance. The evaluations were conducted on two unbalanced full data sets as well as (1) the data subsets that contained features selected by a wrapper approach, and (2) the data subsets that contained only previous grade(s). We then examined how dimensionality reduction may help to improve the overall prediction as well as TP rates for the Fail grade. The Fail grade has a smaller representation than the Pass grade in the data.

2 Data Set Overview

In this study, we used the data sets of two subjects provided by Cortez and Silva (2008) [2]. The data sets contain the student detail from two Portuguese secondary schools and they are available on UCI machine learning repository website. As shown in Table 1, there are 32 features and one class (G3, the final grade) in the data sets. The features include the student grades, demographic, social and school related features. In their previous work, G3 was modelled into three supervised approaches: binary classification, 5-level classification and regression. However, in this study, we only focused on binary classification. The binary classification categorised G3 into binary values which are "Pass" and "Fail". The values of G3 are ranged (0, 20); its value must be at least 10 to categorise a student in the "Pass" class.

Table 1. The features of the data sets.

No.	Feature	Description	Data Type
1	school	School of the student	Binary
2	sex	Gender of the student	Binary
3	age	Age of the student	Numeric
4	address	Home address	Binary
5	famsize	Family size	Binary
6	Pstatus	Cohabitation of the parents	Binary
7	Medu	The education level of mother	Numeric
8	Fedu	The education level of father	Numeric
9	Mjob	Father's job	Nominal
10	Fjob	Mother's job	Nominal
11	Reason	Reason for choosing the school	Nominal

12	Guardian	Guardian of the student	Nominal
13	traveltime	Travel time from home to school	Numeric
14	studytime	Weekly study time	Numeric
15	failures	Number of previous class failures	Numeric
16	schoolsup	Extra educational school support	Binary
17	famsup	Family educational support	Binary
18	paid	Extra paid classes within the course subject	Binary
19	activities	Extra-curricular activities	Binary
20	nursery	Attended nursery school	Binary
21	higher	Wants to take higher education	Binary
22	internet	Internet access at home	Binary
23	romantic	Romantic relationship	Binary
24	famrel	Quality of family relationship	Numeric
25	freetime	Free time after school	Numeric
26	goout	Going out with friends	Numeric
27	Dalc	Workday alcohol consumption	Numeric
28	Walc	Weekend alcohol consumption	Numeric
29	health	Current health status	Numeric
30	absences	No. of school absences	Numeric
31	G1	First period grade	Numeric
32	G2	Second period grade	Numeric
33	G3	Final grade	Numeric

Table 2. The class distribution of the data sets. The data sets are unbalanced as the representation of the Fail grade (numbers with bold font) of each data set is lower than the Pass grade.

Grade	Portuguese Language	Distribution	Mathematics	Distribution
Pass	549	84.6%	265	67.1%
Fail	100	**15.4%**	130	**32.9%**
Total	649	100.0%	395	100.0%

The class distributions of both data sets are as shown in Table 2. They share a common characteristic where the number of records for the Fail grade is less as compared with the Pass grade. Even though the accuracies obtained may be satisfactory, such unequal class distribution may cause unsatisfactory TP rates for the Fail grade because classification algorithms may not be able to learn the Fail grade effectively; irrelevant features may also contribute to the unsatisfactory rates.

In their previous work [2], the previous grades (G1 and G2) were evaluated to determine their importance in predicting the final grade of the students. Along with full data sets (setting A), additional two data subsets were produced. The first two subsets (setting B) contained all the features except the second previous grade, G2. The second two subsets (setting C) had no previous grades at all.

Table 3. The results of the previous work [2] using different data set settings. The numbers in bold show the highest classification accuracies for Portuguese and Mathematics data sets.

Setting	Portuguese Language	Algorithm	Mathematics	Algorithm
A	**93.0%**	Decision Tree	**91.9%**	Naïve Predictor
B	90.1%	Random Forest	83.8%	Naïve Predictor
C	85.0%	Random Forest	70.6%	SVM

Table 3 shows part of their results with the best performing classification algorithms on As, Bs, and Cs, respectively. However, there is no TP rate specifically for the Pass and Fail grades in the data sets. Therefore, we are unable to look into the prediction performance for each grade. But generally, As (the full data sets) gave the best accuracy regardless of classification algorithms followed by Bs and then Cs. Such results indicated that with the presence of the other features, both previous grades helped in predicting student performance effectively.

In the following sections, we attempted to figure out whether or not dimensionality reduction on the full data may improve the prediction on student performance. Evaluations on the dimension reduction were then followed and discussions were provided for the evaluation results.

3 Evaluations

3.1 Finding the Overall Best Performing Algorithm for the Full Data Sets

Table 4. The preliminary study of the two data sets using various classification algorithms. The numbers in bold indicated the highest accuracies for the data sets.

Data set	C4.5	NB	NBT	LibSVM
Portuguese Language	**92.9%** (c:0.2, m:12)	88.6%	87.2%	**92.9%** (cs:10, gm:0.001)
Mathematics	**91.4%** (c:0.6, m:16)	86.3%	81.8%	90.6% (cs:100, gm:0.001)

* c- the confidence factor used for pruning the decision tree
m- the minimum number of instances per leaf
cs - cost, gm - gamma

An evaluation was conducted on these two full data sets using popular classification algorithms: (1) C4.5 Decision Tree, (2) Naïve Bayes (NB), (3) Naïve Byes Tree (NBT), and (4) Library for Support Vector Machine (LibSVM). The reason of conducting this study was to find the overall best performing algorithm for the subsequent evaluations. At this stage, we were looking only at the overall performance of the algorithms. Therefore, only accuracy was considered as the evaluation metric. The classification algorithms were optimized for performance and the parameters involved are as shown in Table 4.

Most of the classification algorithms gave average and above prediction performance (refer to Table 4). For the Portuguese language data set, both C4.5 and LibSVM scored the highest accuracy, which was 92.9%. For the Mathematics data set, C4.5 gave the highest accuracy, which was 91.4%. Looking at the overall good performance by C4.5 in both data sets, the algorithm shall then be utilised in the subsequent evaluations.

3.2 Dimensionality Reduction for the Overall Prediction

3.2.1 Using the Wrapper Approach

We then reduced the dimension of the full data sets. A wrapper approach [17] was utilised in this study to find a useful feature subset. Using the approach, a feature subset was produced and evaluated in the context of the classification algorithm, C4.5.

Table 5. The accuracy attained using the feature subsets produced by the wrapper approach. C4.5 was used for predicting student performance.

Data set	Backward Search	Accuracy
Portuguese Language	Traveltime, studytime, famrel, G2 (4)	94.8% (c:0.3 m:2)
Mathematics	address, famsize, Fedu, Fjob, failures, paid, romantic, goout, Walc, health, absences, G1, G2 (13)	92.2% (c:0.5 m:6)

* c- the confidence factor used for pruning the decision tree
m- the minimum number of instances per leaf

Since the classification algorithm was involved in the evaluation, the feature subset produced was able to yield good classification accuracies. In searching the feature space for the subset, we used greedy hill-climbing with backtracking capability and the backward search direction. The data subsets produced by the wrapper approach are as shown in Table 5.

C4.5 selected in the previous evaluation was then utilised to evaluate the data subsets. The accuracies attained were: (i) 94.8% for the Portuguese Language data subset, and (ii) 92.2% for the Mathematics data subset (refer to Table 5).

3.2.2 Using Only Previous Grades

Table 6. Evaluation using C4.5 on the data subsets that contained only features G1, G2 or both. The numbers in bold indicated the highest accuracies for the data subsets.

Data set	G1 only	G2 only	G1&G2
Portuguese Language	92.0% (c:0.1, m:1)	**93.7% (c:0.1, m:1)**	93.4% (c:0.1, m:1)
Mathematics	82.0% (c:0.1, m:1)	**91.9% (c:0.1, m:1)**	91.4% (c:0.1, m:1)

* c- the confidence factor used for pruning the decision tree
m- the minimum number of instances per leaf

The evaluation was then followed by reducing even more features from the data sets. As oppose to the previous work [2], this evaluation was conducted using only the previous grades without considering the rest of the features. Three data subsets were produced. The first two subsets contained only feature G1 and G2, respectively. The third subset contained both G1 and G2.

As shown in Table 6, utilising the previous grade G2 alone yielded the best accuracies. On the other hand, comparative performances were also attained using both previous grades. The results showed that even without considering the other

features, using only previous grades yielded satisfactory results. The result of using G2 shall be used for comparison purpose in Table 7.

Table 7. The accuracies for the data subsets produced by the wrapper approach, the data subsets that contained only G2, and the full datasets. The numbers in bold indicated the highest accuracies achieved.

Data set	Wrapper	G2 only	The previous work [2]
Portuguese Language	**94.8%** (c:0.3 m:2)	93.7% (c:0.1, m:1)	93.0 (DT)
Mathematics	**92.2%** (c:0.5 m:6)	91.9% (c:0.1, m:1)	91.9% (NV)

* c- the confidence factor used for pruning the decision tree
m- the minimum number of instances per leaf

As shown in Table 7, using feature subsets produced by the wrapper approach is the choice for the overall prediction. The result was not much different from the results of the previous evaluations and also the results of the previous work [2]. However, fewer features were involved using data subsets produced in this study. Therefore, it helped reduce the computational cost of building the prediction model.

3.3 Dimensionality Reduction for Predicting the Fail Grade

Table 8. The TP rate, resulted from using C4.5, for each grade of the full data sets and data subsets.

Data set		Full data set	Subset_$_{wrapper}$	Subset_$_{G2}$
Portuguese Language	Fail	70.0%	76.0%	66.0%
	Pass	97.1%	98.5%	98.7%
Mathematics	Fail	86.2%	88.5%	93.8%
	Pass	94.0%	94.0%	90.9%

The previous evaluations used only accuracies to look into the overall prediction performance. We then further examined whether the dimensionality reduction for both unequal data sets helped improving the prediction performance for the important Fail grade. The examination looked into the TP rates yielded from the previous evaluation, but were not shown or discussed in the previous section.

Using both full data sets, the TP rates for the Pass grade were generally high (refer to Table 8). Learning Pass grade was effective using the classification algorithm because of their high representation in the data sets. On the contrary,

the TP rates for Fail grade were average for the Portuguese data set (70.0%) and above average for the Mathematics data set (86.2%). These average rates could be caused by the low representation of the Fail grade and also the irrelevant features in the data sets.

Reducing the number of features in both data sets improved the prediction performance for the Fail grade. The Portuguese data subset produced by the wrapper approach yielded a better TP rate than the full Portuguese data set (76.0% vs 70.0%). A slightly poor TP rate was obtained if only the feature G2 was used (66.0%).

The Mathematics data subset that contained the features selected by the wrapper approach, yielded a slightly improved TP rate for Fail grade as compared with the full data set (88.5% vs 86.2%). The highest TP rate was attained using the data subset that contained only the feature G2 (93.8%).

In short, competitive results can be attained for the fail grade even though fewer features were involved (particularly the wrapper approach) in predicting the student performance.

4 Conclusion

In this study, we looked for the possibility of gaining competitive prediction performance for both unbalanced Portuguese language and Mathematics data sets via dimensionality reduction. Removing irrelevant features helps producing competitive prediction results. In addition, less computation cost is involved with dimensionality reduction.

The evaluation results showed that the overall student performance, measured using accuracies, can generally be predicted well by relying on the data subsets generated by the wrapper approach. Note that comparable overall predictions can also be achieved by using only one previous grade, G2, without involving the other features. This shows the importance of the previous grade in predicting the student performance.

We had also attempted to improve the TP rates attained for the Fail grade that has a small representation in both data sets. Predicting the Fail grade effectively is important for teachers so that resources can be well managed ahead for the potential Fail students. As compared with full data sets, the evaluation results showed that the wrapper approach gave relatively better TP rates for the Fail grade by reducing the number of features in the full data sets.

References

[1] Nisbet, J., & Welsh, J. (1966). Predicting student performance. *Higher Education Quarterly*, 20(4), 468-480.

[2] Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. *In the Proceedings of 5th Future Business Technology Conference* (pp. 5-12). Eurosis.

[3] Burgos, C., Campanario, M. L., de la Pena, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, *66*, 541-556.

[4] Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177-194.

[5] Devasia, T., Vinushree, T. P., & Hegde, V. (2016, March). Prediction of students performance using Educational Data Mining. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)* (pp. 91-95). IEEE.

[6] Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H. (2016). Predicting student performance using personalized analytics. *Computer*, 49(4), 61-69.

[7] Conijn, R., Van den Beemt, A., & Cuijpers, P. (2018). Predicting student performance in a blended MOOC. *Journal of Computer Assisted Learning*, *34*(5), 615-628.

[8] Hughes, G., & Dobbins, C. (2015). The utilization of data analysis techniques in predicting student performance in massive open online courses (MOOCs). *Research and practice in technology enhanced learning*, *10*(1), 10.

[9] Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in human behavior*, *58*, 119-129.

[10] Weiss, G. M. (2004). Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, 6(1), 7-19.

[11] Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. arXiv preprint arXiv:1305.1707.

[12] Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence*, 38(3), 315-330.

[13] Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. arXiv preprint arXiv:0912.3924.

[14] Acharya, A., & Sinha, D. (2014). Application of feature selection methods in educational data mining. *International Journal of Computer Applications*, 103(2).

[15] Xu, M., Liang, Y., & Wu, W. (2017). Predicting Honors Student Performance Using RBFNN and PCA Method. In *International Conference on Database Systems for Advanced Applications* (pp. 364-375). Springer, Cham.

[16] Punlumjeak, W., Rachburee, N., & Arunrerk, J. (2017). Big data analytics: student performance prediction using feature selection and machine learning on microsoft azure platform. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 9(1-4), 113-117.

[17] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2), 273-324.