So, H. J. et al. (Eds.) (2020). Proceedings of the 28th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Improving Classification in Imbalanced Educational Datasets using Over-sampling

Zeynab MOHSENI[a*], Rafael M. MARTINS[a], Marcelo MILRAD[a] & Italo MASIELLO[a]
[a] Department of Computer Science and Media Technology, Linnaeus University, Sweden
*zeynab.mohseni@lnu.se

Abstract: Learning Analytics (LA) involves a growing range of methods for understanding and optimizing learning and the environments in which it occurs. Different Machine Learning (ML) algorithms or learning classifiers can be used to implement LA, with the goal of predicting learning outcomes and classifying the data into predetermined categories. Many educational datasets are imbalanced, where the number of samples in one category is significantly larger than in other categories. Ordinarily, it is ML's performance on the minority categories that is the most important. Since most ML classification algorithms ignore the minority categories, and in turn have poor performance, so learning from imbalanced datasets is really challenging. In order to address this challenge and also to improve the performance of different classifiers, Synthetic Minority Over-sampling Technique (SMOTE) is used to oversample the minority categories. In this paper, the accuracy of seven well-known classifiers considering 5 and 10-fold cross-validation and the F_1-score are compared. The imbalanced dataset collected based on self-regulated learning activities contains the learning behaviour of 6,423 medical students who used a web-based study platform—Hypocampus—with different educational topics for one year. Also, two diagnostic tools including Area Under the Receiver Operating Characteristics (AUC-ROC) curves and Precision-Recall (PR) curves are applied to predict probabilities of an observation belonging to each category in a classification problem. Using these diagnostic tools may help LA researchers on how to deal with imbalanced educational datasets. The outcomes of our experimental results show that Neural Network with 92.77% in 5-fold cross-validation, 93.20% in 10-fold cross-validation and 0.95 in F_1-score has the highest accuracy and performance compared to other classifiers when we applied the SMOTE technique. Also, the probability of detection in different classifiers using SMOTE has shown a significant improvement.

Keywords: Learning Analytics, Imbalanced Dataset, Machine Learning, SMOTE, ROC, PR.

1. Introduction

The data generated in educational environments such as courses that use learning management systems or digital learning materials are frequently large, complex and heterogeneous (Martins et al., 2019). Learning Analytics (LA) (Khosravi & Cooper, 2017) can be defined as the measurement, collection, analysis and reporting of data about learners and their contexts and contains a wide range of methods for understanding and optimizing learning and the digital environments in which it occurs. LA is commonly implemented with the use of different ML algorithms (Chung & Lee, 2019), which are strong and flexible methods to produce solutions for problems that are not being handled with traditional statistical approaches. ML algorithms make it possible to predict students' performance and risk of under- or over-performing, for example, as it helps to classify the data into predetermined categories (e.g., high performance, medium performance, low performance) (Akcapınar et al., 2019; Khosravi & Cooper, 2017).

Large-scale datasets usually contain several categories, but when the distribution of such categories is not uniform, i.e. some of them (the *minority*) are heavily under-represented when compared to others (the *majority*), that leads to category imbalance. This *'imbalanced'* or *'skewed'* distribution of category instances results in learning classifiers being biased (Kuncheva et al., 2019; Napierala & Stefanowski, 2016). One possible solution to this problem is over-sampling, where randomly-selected samples from the minority categories are duplicated. SMOTE is one of the most

popular algorithms for over-sampling, relying on the concept of nearest neighbours to create its synthetic data (Fernandez et al., 2018).

In this paper, we describe the improvement of the performance results of different ML classification algorithms using the SMOTE resampling techniques, applied to a learning analytics problem and dataset. Our research question is formulated as follows: *How does the combination of SMOTE and under-sampling perform, in comparison to traditional ML classification algorithms, when handling learning analytics datasets?* In our data, the activity of each student reflects the number of questions she/he answered related to a certain subject or topic. We analyse and compare seven different types of ML algorithms using two threshold metrics, accuracy and F_1-score, together with two rank metrics, AUC-ROC curves and PR curves (Jeni, Chon & De La Torre, 2013).

The rest of this paper is organized as follows. Section 2 describes briefly related work in this field. Sections 3 and 4, respectively, present the case study and the different methods used in our work. An exploratory analysis and the experimental results are described in sections 5 and 6, respectively. Section 7 discusses the conclusions and presents possible lines of future work.

2. Related Work

Imbalanced classification is a problem in data sets with skewed distributions of data points (Chawla et al., 2002; Napierala & Stefanowski, 2016). Data-level approaches are used to address the category imbalance problems. The focus of these methods is on re-sizing the training datasets to balance different labels and make the dataset suitable for a standard learning algorithm. In order to fix category imbalance, resampling methods such as under-sampling and over-sampling methods are used. In under-sampling techniques, samples from the majority category are discarded, while in over-sampling methods, new minority category samples are generated (Mathew et al., 2015). SMOTE, an approach proposed by Chawla et al., (2002), is one of the most well-known over-sampling algorithms. It generates new minority data instances by identifying nearest neighbours in input space and applying a linear interpolation between them. This way, the new data instances populate areas near other points, and should properly resemble real data. Batista et al., (2004) have performed a systematic experimental study with 15 real-world datasets and different pre-processing methods such as SMOTE. The results indicate that the over-sampling methods provided better AUC than the under-sampling ones. Mathew et al., (2015), on the other hand, proposed a kernel-based SMOTE algorithm that generates synthetic data points of minority categories directly in the feature space of Support Vector Machine (SVM) classifier. In Beyan & Fisher (2015), a new hierarchical decomposition method for imbalanced data sets is proposed. The proposed method is based on clustering and outlier detection. These works are focused on general classification problems for imbalanced datasets with the goal of improving the performance of different algorithms but are not applied to the specific challenges and opportunities of imbalanced learning analytics applications. Hasnine et al., (2018) use SMOTE as one of the pre-processing steps in their pipeline, among others such as features selection, in order to apply ML for the prediction of student performance. However, their paper does not focus on the analysis of SMOTE itself, so it is not clear what exactly the advantages and disadvantages of the technique are when applied to an imbalanced learning analytics dataset. In our paper, we perform an experimental comparison that is specifically suited to compare performances before and after SMOTE, isolating its effects and producing objective insights on the impact of its application. By analysing the PR and ROC-AUC curves obtained from different classification algorithms, more in-depth understanding of the behaviour of balancing methods is provided.

3. Dataset Description

We used an imbalanced dataset which includes the learning behaviours of 6,423 medical students (data points) who used an online study platform (Hypocampus[1]) during one year. The students chose their own study path through the material, which is arranged in subjects, e.g., orthopaedics or neurology, and

[1] https://www.hypocampus.se/.

topics (or chapters), e.g., cerebrovascular disease or diabetes. During their studies, they are frequently faced with questions to test what they have learned. We aggregated all students' answers per topic in order to generate the features, resulting (after data cleaning and pre-processing) in 1,445 features which reflect how many questions she/he answered on each of the 1,445 topics (Martins et al., 2019). In summary, each of the 6,423 data points indicate the learning behaviour of one student, quantified by the amount of questions she/he answered in each of 1,445 available topics. Additionally, each student may have a *University ID*, which indicates where that student comes from (or "*Other*" for all the students who do not have a university ID).

4. Methods

In order to identify how well a classifier preformed, a cross-validation procedure was used. In k-fold cross-validation, a partition of the dataset is formed by splitting it into *k* non-overlapping subsets, including *k-1* training sets and one test set. Then, we can train and test the model *k* times, each time using different train and test sets (Goodfellow et al., 2016; Geron, 2017). We used 5-fold and 10-fold cross-validations; 5-fold means for each classifier we choose the mentioned input features that perform best on average when we train on 80% of the data and test on the remaining 20%, and 10-fold when we train on 90% of the data and test on the remaining 10%. To compare cross-validation results from different classifiers, one of the measures used is the average accuracy (or average error), shown in Eq. 1 (Chawla et al., 2002). In this equation *TP* or True Positives is the number of positive examples correctly classified; *TN* or True Negatives is the number of negative examples correctly classified; *FP* or False Positives is the number of negative examples incorrectly classified as positive; and *FN* or False Negatives is the number of positive examples incorrectly classified as negative.

Error rate, that is, *1-Accuracy*, is more appropriate to use in balanced datasets, while other measures such as ROC and PR curves are more suitable to be used when there are unequal error costs. Using PR curves is more suitable for highly-skewed domains where ROC curves may provide an excessively optimistic view of the performance (Chawla et al., 2002). In this paper, two ROC-AUC and PR curves were used to compare different classification algorithms, summarized with the average precision, micro-average and macro-average. A macro-average computes the metric independently for each category, and then takes the average, whereas a micro-average aggregates the contributions of all categories to compute the average metric. In ROC-AUC and ROC curves, the *True Positive* Rate (Eq. 2) is a fraction calculated as the total number of true positive predictions divided by the sum of the true positives and the false negatives, while the *False Positive Rate* (Eq. 3) is calculated as the total number of false positive predictions divided by the sum of the false positives and true negatives. In PR curves, *Recall* (Eq. 4) is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made. Recall is calculated as the number of true positives divided by the total number of true positives and false negatives. In this curve, *Precision* (Eq. 5) is a metric that quantifies the number of correct positive predictions made, and it is calculated as the number of true positives divided by the total number of true positives and false positives (Chawla et al., 2002). Finally, to determine a weighted average of the precision and recall values, F_1*-score* is used (Eq. 6) (Jeni, Chon & De La Torre, 2013). F_1-score range is between 0 and 1, where the maximum shows the perfect classification.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad \text{(Eq.1)}$$

$$TP = TP / (TP + FN) \quad \text{(Eq.2)}$$

$$FP = FP / (TN + FP) \quad \text{(Eq.3)}$$

$$\text{Recall} = TP / (TP + FN) \quad \text{(Eq.4)}$$

$$\text{Precision} = TP / (TP + FP) \quad \text{(Eq.5)}$$

$$F_1 = 2 \times TP / (2 \times TP + FP + FN) = 2 \times ((\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})) \quad \text{(Eq.6)}$$

Additionally, in order to compare and check the statistical significance of the difference in results before and after the use of SMOTE, we ran a set of one-tailed Mann-Whitney U tests, one for each performance measure, with the significance level set to 0.01. These are non-parametric statistical tests suitable for small samples, as is our case.

4.1 Classification Algorithms

We compare seven different types of ML classification algorithms: Linear (Geron, 2017), k-Nearest Neighbors (kNN) (Goodfellow et al., 2016), Decision Tree (Goodfellow et al., 2016), Neural Network (Geron, 2017), Support Vector Machines (SVM) (Geron, 2017), Random Forest (Breiman, 2001) and XGBoost (Geron, 2017). Due to space limitations, and the fact that these are well-known in the ML field, we refrain from discussing them in details here; please see the references for more information. Each of these seven well-known algorithms can be configured with different hyperparameters that will affect their performance; therefore, we report only on the best configurations found after performing a systematic hyperparameter search in each case.

5. Experimental Results

Table 1 contains the results of all the considered metrics for the seven ML algorithms, before applying the SMOTE method. The mean performance of all algorithms stays around 58%, and while micro-average ROC is relatively high (0.85), the rest of the measures including average precision (0.56), macro-average ROC (0.71) and F1 scores (0.57) reflect the expected low-quality results of an imbalanced dataset. The Random Forest classifier had the best results overall, with 61.70% accuracy in 5-fold cross-validation, 61.79% in 10-fold cross-validation, 0.63 average area in PR, 0.88 micro-average area, 0.76 macro-average area in ROC-AUC, and F1 score of 0.62. It was the strongest classifier before the application of SMOTE.

Table 2 shows the results for the same algorithms after applying SMOTE to correct the imbalance of the data. The overall results improved drastically, with mean accuracy now around 78% in all cases, and all the rest of the measures above 0.80 (mean). All algorithms showed increased performance overall, with the Neural Network and Random Forest classifiers staying at the top with 92.77% and 92.40% accuracy in 5-fold cross-validation, and with 93.20% and 92.81% in 10-fold cross-validation, respectively. Random Forest also remains the best classifier according to the rest of the performance measures, with 0.96 average area in PR, 0.98 micro-average area, and 0.98 macro-average area in ROC-AUC. Neural Network, with 0.95, has the highest F1 score, followed closely by Random Forest with 0.94. It is interesting to notice that Linear and XGBoost did not improve much after SMOTE, reflecting the highly non-linear and complex nature of the data (to which these algorithms are not suitable).

The last row of Table 2 provides the U-values resulting from the Mann-Whitney *U* tests for each measure (as described in Section 4). In all the considered cases, the critical value of *U* at $p < .01$ is 6, so a U-value of less than 6 means statistically-significant results. Most of the measures show a statistically-significant increase in the overall performance (highlighted in green), except Micro-average ROC (highlighted in red).

To further illustrate and discuss these results, we focus on the Random Forest algorithm (which performed best overall) and show, in Figure 1, three detailed visual comparisons of its performance: the confusion matrix, as a heatmap (Pryke, Mostaghim & Nazemi, 2007); the ROC curves; and the PR curves. In the confusion matrix (Figures 1a, d), darker cells mean correct class predictions. The main problem with imbalanced datasets is immediately apparent in the matrix before SMOTE (Figure 1a): the classifier assigned the label “other” to most points (since most darker cells are in the last column to the right), resulting in low performance. After SMOTE, however, the problem is mostly solved, as can be seen from the darker cells along the diagonal of the matrix. In the ROC curves, the ideal results are curves that bend along the top-left of the graph, maintaining a large proportion of TP vs. FP, as is the case after SMOTE (Figure 1e). Figure 1b shows that, before SMOTE, the lines are close to the diagonal instead. On the other hand, for the PR curve graph, the ideal results are curves that are close to the top-right corner, as is again the case after SMOTE (Figure 1f). In Figure 1c we can see that, before SMOTE, the lines were near random and scattered among the whole graph.

Table 1. *Performance of ML Algorithms* ***Before*** *SMOTE*

Classifiers	Training (%)	5-fold (%)	10-fold (%)	Avg. Precision	Micro-avg. ROC	Macro-avg. ROC	F1
Linear	54.47	54.29	54.99	0.47	0.80	0.69	0.48
Decision Tree	57.74	57.64	58.23	0.55	0.85	0.70	0.56
Neural Network	53.07	54.85	54.65	0.52	0.84	0.73	0.52
SVM	59.77	59.74	59.91	0.58	0.85	0.69	0.59
kNN	60.16	60.02	59.94	0.59	0.85	0.72	0.59
Random Forest	61.95	61.70	61.79	0.63	0.88	0.76	0.62
XGBoost	62.02	61.30	61.31	0.61	0.86	0.74	0.61
Mean	58.45	58.50	58.69	0.56	0.85	0.71	0.57

Table 2. *Performance of ML Algorithms* ***After*** *SMOTE*

Classifiers	Training (%)	5-fold (%)	10-fold (%)	Avg. Precision	Micro-avg. ROC	Macro-avg. ROC	F1
Linear	64.72	64.15	64.56	0.58	0.82	0.82	0.59
Decision Tree	84.09	84.24	85.11	0.71	0.75	0.74	0.71
Neural Network	92.75	92.77	93.20	0.93	0.88	0.88	0.95
SVM	60.31	61.04	61.64	0.85	0.92	0.92	0.86
kNN	83.69	84.24	84.85	0.89	0.97	0.97	0.9
Random Forest	92.07	92.40	92.81	0.96	0.98	0.98	0.94
XGBoost	69.61	68.29	68.60	0.75	0.89	0.89	0.75
Mean	78.17	78.16	78.68	0.81	0.89	0.88	0.81
***U*-values**	**2 (<6)**	**2 (<6)**	**1 (<6)**	**3.5 (<6)**	**13.5 (>6)**	**1.5 (<6)**	**3 (<6)**

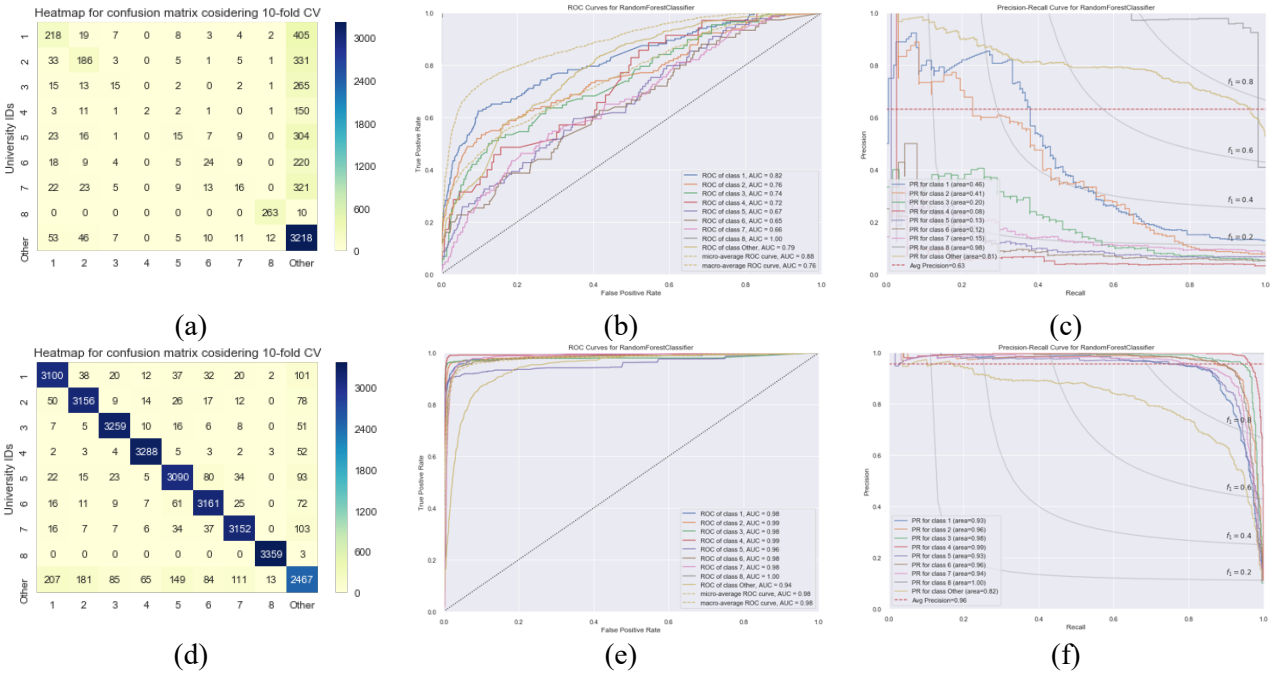

Figure 1. Confusion matrices, ROC curves, and PR curves of Random Forest classifier **before (a, b, c)** and **after (d, e, f)** the use of SMOTE.

6. Conclusion

In this article we have compared the performance of seven machine learning algorithms applied to an imbalanced LA problem, to answer our research question on how the combination of SMOTE and under-sampling performs compared to traditional ML classification algorithms. The students' dataset was collected from a Web-Based Learning Environment during one year and it consists of students (data points) described by multidimensional numerical vectors (features). The approach described here should be generalizable to any other scenario similar to this. According to the results, the performance using SMOTE has widely increased, and Neural Network and Random Forest are the most accurate and high-performance classifiers among the tested ML classification algorithms. Thus, we determine that the combination of SMOTE and under-sampling performs better than traditional ML classification algorithms in an LA context, which reflects other previous and more general results outside of LA. One limitation of this approach is that SMOTE is based on linear interpolation between nearest neighbours, which limits its application for datasets that are too large and contain highly non-linear relationships between its features. The results of a high-performance classification algorithm on educational datasets can have practical implications for teachers, that is, given the right visualization technique this sort of analysis promise to guide teachers in identifying learning issues and possibly, in the future, predicting students' outcomes.

References

Akcapınar, G., Hasnine, M. N., Majumdar, R., Flanagan, B., Ogata, H. (2019). Developing an early-warning system for spotting at-risk students by using eBook interaction logs. *Smart Learning Environments, Springer*.

Batista, G. E., Prati, R. C., Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD explorations newsletter 6* (1).

Beyan, C., Fisher, R. (2015). Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition 48* (5), 1653-1672.

Breiman, L. (2001). Random Forests. *Journal of Machine Learning, Springer*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research 16*, 321-357.

Chung, J. Y., Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Journal of Children and Youth Services Review 96*, 346-353.

Fernandez, A., Garcia, S., Herrera, F., Chawla, N. V. (2018a). SMOTE for Learning from Imbalanced Data: Progress and Challenges. *Journal of Artificial Intelligence Research 61*, 863-905.

Geron, A. (2017). Hands-On Machine Learning with Scikit-Learn and Tensorflow. *O'Reilly Media, Location*.

Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. *MIT Press, Location* (2016).

Hasnine, M. N., Akcapinar, G., Flanagan, B., Majumdar, R., Mouri, K., Ogata, H. (2018). Towards Final Scores Prediction over Clickstream Using Machine Learning Methods. *26th International Conference on Computers in Education—Workshop Proceedings*, 399-404.

Jeni, L. A., Chon, J. F., De La Torre, F. (2013). Facing Imbalanced Data Recommendations for the Use of Performance Metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*.

Khosravi, H., Cooper, K. M. L. (2017). Using Learning Analytics to Investigate Patterns of Performance and Engagement in Large Classes. *Journal of ACM Special Interest Group on Computer Science Education*, 309-314.

Kuncheva, L. I., Arnaiz-González, A., Díez-Pastor, J., Gunn, I. A. D. (2019). Instance selection improves geometric mean accuracy: a study on imbalanced data classification. *Journal of Progress in Artificial Intelligence 8*, 215-228.

Martins, R. M., Berge, E., Milrad, M., Masiello, I. (2019). Visual Learning Analytics of Multidimensional Student Behavior in Self-regulated Learning. *Journal of Transforming Learning with Meaningful Technologies 11722*, 737-741.

Mathew, J., Luo, M., Pang, C. K., Leng, H. (2015). Kernel-based SMOTE for SVM classification of imbalanced datasets. *IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society*, 1127-1132.

Napierala, K., Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems 46*, 563-597.

Pryke, A., Mostaghim, S., Nazemi, A. (2007). Heatmap Visualization of Population Based Multi Objective Algorithms. *International Conference on Evolutionary Multi-Criterion Optimization (EMO2007)*, 361-375.