*Avrupa Bilim ve Teknoloji Dergisi*
*Özel Sayı, S. 485-489, Ekim 2019*
*© Telif hakkı EJOSAT'a aittir*
***Araştırma Makalesi***

www.ejosat.com ISSN:2148-2683

*European Journal of Science and Technology*
*Special Issue, pp. 485-489, October 2019*
*Copyright © 2019 EJOSAT*
***Research Article***

# Improving classification performance for an imbalanced educational dataset example using SMOTE

Yavuz Ünal[1*], Ahmet Sağlam[2], Osman Kayhan[3]

[1] Amasya University, Technology Faculty, Department of Computer Engineering, Amasya, Turkey (ORCID: 0000-0000-0000-0000)
[2] Amasya University, Merzifon Vocational High School, Department of Computer Programing, Amasya, Türkiye (ORCID: 0000-0002-2616-8253)
[3] Amasya İlduş Hatun Vocational Technical Anatolian High School, Amasya, Turkey (ORCID: 0000-0000-0000-0000)

*(This publication has been presented orally at HORA congress.)*

(First received 1 August 2019 and in final form 26 October 2019)

(**DOI:** 10.31590/ejosat.638608)

**ATIF/REFERENCE:** Ünal, Y., Sağlam, A. & Kayhan, O. (2019). Improving classification performance for an imbalanced educational dataset example using SMOTE. *European Journal of Science and Technology*, (Special Issue), 485-489.

**Abstract**

With technology, a lot of data is formed in digital environments. One of the areas with intensive data is educational data sets. By analyzing educational data sets, students' situatiokjgjjööÖns can be predicted by foreseeing. In this way, students can be assisted by anticipating situations such as drop-out due to failure. Educational institutions can take measures to prevent such dropouts and reduce student drop-out. Thus, financial losses of students and educational institutions can be prevented. In this study, the data of five separate associate degree students who were enrolled in Amasya University Distance Education Center in 2016-2017 were used. These are associate degree programs in child development, medical documentation and secretarial, electricity, mechatronics, and internet and network technologies. It was estimated whether the students could graduate or not at the end of the IV. Semester with looking at their I. and II. semester course notes. These data were analyzed by k nearest neighbor (K-NN) and KStar algorithms. Some of the data were obtained from the distance education center as imbalanced data due to the low number of students. In Educational Data Mining, researchers usually overlook the balance of the distribution on a dataset. Unbalanced data can seriously affect the success of classification. Synthetic minority oversampling technique (SMOTE) method was applied to these unbalanced data and how it affected the success of classification was examined. First, the raw data were analyzed with K-nearest neighbors classifier and KStar classifier. In this study, the analysis results of these five chapters are given in tables and comparatively. In this study, it has been seen that SMOTE oversampling method increase the classification success. In areas where unstable data such as educational data mining may exist, higher classification accuracy can be achieved with the help of different oversampling methods.

**Keywords:** Oversampling method, Educational datamining, distance educational, resampling, imbalanced class

## 1. Introduction

Interesting and useful information can be obtained by analyzing large amounts of data. Analysis of these data, which can be boring for most of the humans, is the main subject of data mining. Data mining draws the attention of scientists in recent years. Data mining is successfully applied in many fields today, such as finance, marketing, banking, insurance, and healthcare (Ginsberg et al., 2009). One of the other areas which it is successfully used in is education. Lots of data stacks up about students in education institutions. Methods such as classification, clustering and association rules are used in order to identify students' their interests and tendencies, cluster them according to their achievements and interests, present learning contents automatically and reveal misconceptions (Güldal ve Çakıcı, 2017; Peña-Ayala, 2013). Educational data mining contributes to understanding the students' learning styles, and also enables data-driven decision-making to develop existing educational practices and learning materials (Öztürk, 2018). Additionally, the results of data mining analysis will improve the quality of education. There are many studies in literature

[1] Corresponding Author: Amasya University, Department of Computer Engineering, Amasya, Turkey, ORCID: 0000-0000-0000-0000, yavuz.unal@amasya.edu.tr

about education field. Most of these studies were about student behaviour or success (Aydemir, 2019). İnreased education quality will bring well-educated individuals and well-educated individuals will bring a well- developed society.

Collecting educational data may not always be easy. It can sometimes take a long time. Lack of sufficient data due to lack of students in the department or having a few data left after clearing the outliers may force us to work with imbalanced data sets., working with unstable data sets often adversely affects the performance of many machine learning methods, classification in particular (Pristyanto et al., 2018).

In this study, data of associate's degree students of five different departments who enrolled Amasya University Distance Learning Center in 2016-2017, were used. Some of these departments having a few students resulted with imbalanced data sets. In order to decrease this imbalance, Synthetic Minority Oversampling Technique (SMOTE) was used and the change in classification reliability was analyzed.

# 2. Material and Method

## 2.1. Data Collection and Preprocessing

Data used in this study consist of the information gathered from the sudents of five different associate's degree departments of Amasya University. The data of the students who enrolled in 2016-2017 and revieved distance education in associate degreedepartments of Child Development, Medical Documentation and Secretariat, Electrics, Mechatronics, and Internet and Network Technology were used. The data set attributes are given in Table 1.

*Table 1. Dataset attributes*

| **Attributes** | **Attribute Characteristics** |
|---|---|
| Age | Numeric |
| *Gender* | Numeric (0 – female, 1-Male) |
| *Lesson 1* | Numeric (0-100) |
| *Lesson 2* | Numeric (0-100) |
| *Lesson 3* | Numeric (0-100) |
| *Lesson 4* | Numeric (0-100) |
| *Lesson 5* | Numeric (0-100) |
| *Lesson 6* | Numeric (0-100) |
| *Lesson 7* | Numeric (0-100) |
| *Lesson 8* | Numeric (0-100) |
| *Lesson 9* | Numeric (0-100) |
| *Lesson 10* | Numeric (0-100) |
| *Lesson 11* | Numeric (0-100) |
| *Lesson 12* | Numeric (0-100) |
| *Graduation status* | Class (0-No,1-Yes) |

Age, gender, and marks of first and second semester courses are included as attributes in the dataset. The number of courses may vary according to departments. Moreover, information of whether or not associate degree students have graduated in 2 years (on time) is also included in the data set.

Missing data in the data set were excluded as data preprocessing. Some students were exempt from some courses and so they did not have any marks for them. Some courses were excluded from the dataset for this reason. Removing the unnecessary and missing data had led to a decrease in the number of samples and thus an imbalanced database.

## 2.2. SMOTE algorithm

Synthetic Minority over-sampling Tecnique (SMOTE) is one of the oversampling methods and was recommended by Chawla et al. It is used to deal with imbalaced problems in machine learning (Ge ve ark, 2017; Tallo et al., 2018).

The basis of SMOTE alghoritm is to synthesise minority samples and decrease the imbalance in data set. The SMOTE algorithm performs linear interpolation, especially in minority class samples that are close to each other. The SMOTE algorithm searches for the nearest neighboring samples for each sample in the minority class (Zeng ve ark. 2016). New samples are produced for each of the original minority samples. Then, interpolation is carried out between the original minority class samples and neighboring samples yapmaktır (Han et al., 2005; Bunkhumpornpat et al., 2009).

## 2.3. Kstar algorithm

K-star alghoritm is one of the sample based classifications. This alghoritm is used for determining the distance or similarity between two dots. Unlike the other sample based learning methods, entropic based distance function is used (Sultana et al., 2016; Kalıpsız ve Cihan, 2015). The K-star algorithm takes Kolmogorov distance as the shortest distance between two features (Kalıpsız ve Cihan, 2015; Çölkesen ve Kavzoğlu, 2011).

## 2.4. K-nearest neighbors (KNN) algorithm

This method is used for determining the class of a new observation, which will be added to the sample, using a sample cluster observation value whose class is specified. These algorithms use a range of field specific distance functions to find a single sample that most closely resembles the training data. Found sample is used to classify a new sample. The nearest neighbor algorithm is based on calculating the distances of each pixel in the training data to one pixel in the test data whose attribute value is unknown, and selecting the k number of observations with the closest distance. Euclidean distance is used. The formula of euclidean distance is given below (Çölkesen and Kavzoğlu, 2011):

$$d(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \tag{1}$$

## 2.5. Performance measurment

After applying classification alghoritm to a new data, the result is placed in one of the four groups given in Table 2, according to the values of predicted class and true class. They are:

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negavtive

*Table 2. Confusion Matrix for Two-Class problem*

| | **Predicted Positive** | **Predicted Negative** |
|---|---|---|
| Positive | TP | FN |
| *Negative* | FP | TN |

The equation of accuracy according to the table is given below. Accuracy is used in order to evaluate the results of the alghoritms used in this study.

$$Accuracy = \frac{(TP+TN)}{TP+FN+FP+TN} \tag{2}$$

# 3. Results and Discussion

Analysis of this study is carried out using weka software. Initially SMOTE was applied and analysis were done to 5 different data sets, using Kstar and K-NN alghoritms. It was predicted whether the sudents will graduate on time or not. 10-fold cross validation was applied. SMOTE oversampling was used for imbalanced data in the raw data sets and after that K star and K-NN alghoritms were used again for analysis. Data set characteristics for each associate degree programme are given in Table 3.

*Table 3. Data set characteristics used in this work*

| **Dataset Name** | **Total Instance** | **Number of Minority** | **Number of Majority** | **Number of Attributes** |
|---|---|---|---|---|
| Electrics | 24 | 8 | 16 | 13 |
| *Mechatronics* | 49 | 19 | 30 | 13 |
| *Internet and Network Technologies* | 20 | 5 | 15 | 13 |

| | | | | |
|---|---|---|---|---|
| *Child Development* | 110 | 27 | 83 | 17 |
| *Medical Documentation and Secreteriat* | 96 | 35 | 61 | 19 |

Study consists data of 5 different associate degree programmes. Child development had the highest student count with 110 samples. It was followed by Medical Documentation and Secreteriat programme with 96 students. Electrics, Mechatronics, and Internet and Network Technologies programmes had fewer students, thus imbalanced data sets. Departments had different counts of courses, so different counts of attributes. The results of Kstar and K-NN alghoritms applied before SMOTE oversampling is given in Table 4.

*Table 4. The Performance of 2 classifiers without SMOTE*

| **Dataset Name** | **Kstar (Accuracy)** | **K-NN (Accuracy)** |
|---|---|---|
| Electrics | 87.5 | 87.5 |
| *Mechatronics* | 77.55 | 77.55 |
| *Internet and Network Technologies* | 60 | 70 |
| *Child Development* | 81.81 | 86.36 |
| *Medical Documentation and Secreteriat* | 77.08 | 78.12 |

The results given in Table 4 shows us that K-NN alghoritm worked well with imbalanced data sets in terms of classification accuracy. The results afterapplication of MOTE oversampling to the imbalanced data sets are given in Table 5.

*Table 5. The performance of 2 classifiers with SMOTE*

| **Dataset Name** | **Kstar (Accuracy)** | **K-NN (Accuracy)** |
|---|---|---|
| Electrics | 90.24 | 93.25 |
| *Mechatronics* | 83.82 | 92.64 |
| *Internet and Network Technologies* | 64 | 72 |
| *Child Development* | 89.78 | 91.24 |
| *Medical Documentation and Secreteriat* | 83.20 | 90.07 |

It is clear from the Table 5 that application of SMOTE oversampling method to data sets increased classification accuracy in all of the associate degree programmes. For Electrics programme: accuracy values increased from 87.5% to 90.24% for Kstar alghoritm and from 87.5% to 93.25% for K-NN alghoritm. For Mechatronics programme: accuracy values increased from 77.55% to 83.82% for Kstar alghoritm and from 77.55% to 93.25% for K-NN alghoritm. For Internet and Network Technologies programme: accuracy values increased from 60% to 64% for Kstar alghoritm and from 70% to 72% for K-NN alghoritm. For Child Development programme: accuracy values increased from 81.81% to 89.78% for Kstar alghoritm and from 86.36% to 91.24% for K-NN alghoritm. For Medical Documentation and Secreteriat programme: accuracy values increased from 77.08% to 83.20% for Kstar alghoritm and from 78.12% to 90.07% for K-NN alghoritm. Moreover, It should be indicated that K-NN alghoritm resulted with higher accuracy values than Kstar alghoritm, in this study.

## 4. Conclusions and Recommendations

It is important to warn students and their parents about the probability of not graduating on time and give them academic counselling. Student drop outs cause losses for both the institutions and the individuals, morally and financially.

In this paper, it was predicted whether the students who attended distance education programmes could graduate on time or not. Difficulties such as low student count and having limited data led us to work with imbalanced data sets. In order to prevent imbalanced data distribution, SMOTE method which is one of the oversampling methods, was used and prediction accuracies were increased this way.

Student failiures or drop-outs can be prevented with on time counselling for students and/or their parents, with the help of applications like this which were derived from educational data. And losses of individuals and/or institutions can be lowered, this way.

Classification accuracy can be inreased using oversampling methods like SMOTE, while working in fields which have missing or insufficient data.

# References

Aydemir, E. (2019). Ders Geçme Notlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi. *Avrupa Bilim ve Teknoloji Dergisi*, (15), 70-76.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). *Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem*. Pacific-Asia conference on knowledge discovery and data mining, Berlin, Germany.

Çölkesen, İ., & Kavzoğlu, T. (2011).*Örnek tabanlı k-star algoritması ile uzaktan algılanmış görüntülerin sınıflandırılması*. UFUAB VI.Teknik Sempozyumu, Belek, Antalya.

Ge, Y., Yue, D., & Chen, L. (2017). *Prediction of wind turbine blades icing based on MBK-SMOTE and random forest in imbalanced data set.* IEEE Conference on Energy Internet and Energy System Integration (EI2), Changsha, China.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012-1014.

Güldal H., Çakıcı, Y. (2017). *Eğitsel Veri Madenciliği*. 12th International Balkan Education and Science Congress, Nessebar, Bulgaria.

Han, H., Wang, W. Y., & Mao, B. H. (2005). *Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning.* International conference on intelligent computing, Berlin, Germany.

Kalıpsız, O., & Cihan, P. (2015). Öğrenci Proje Anketlerini Sınıflandırmada En İyi Algoritmanın Belirlenmesi. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 8(1), 41-49.

Öztürk, A. (2018). Açık ve uzaktan öğrenme ortamlarında eğitsel veri madenciliği. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, 4(2), 10-13.

Peña-Ayala, A. (Ed.). (2013). *Educational data mining: applications and trends* (Vol. 524). Springer.

Pristyanto, Y., Pratama, I., & Nugraha, A. F. (2018). *Data level approach for imbalanced class handling on educational data mining multiclass classification*. International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia

Sultana, M., Haider, A., & Uddin, M. S. (2016). *Analysis of data mining techniques for heart disease prediction.* 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dakka, Bangladeş.

Tallo, T. E., & Musdholifah, A. (2018). *The Implementation of Genetic Algorithm in Smote (Synthetic Minority Oversampling Technique) for Handling Imbalanced Dataset Problem*. 4th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia.

Zeng, M., Zou, B., Wei, F., Liu, X., & Wang, L. (2016). *Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data*. May, 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), Chongqing, China.