Unsupervised Outlier Detection: students profiling in an effort to indicate learning problems in Higher Educational Institutions

Daria Novoseltseva, Nadine Jessel, Florence Sèdes

Unsupervised Outlier Detection: students profiling in an effort to indicate learning problems in Higher Educational Institutions

Daria Novoseltseva, Nadine Jessel, Florence Sedes
Toulouse Institute of Computer Science Research (IRIT), Paul Sabatier University Toulouse III, Toulouse, France

1 Purpose

Nowadays, institutions of higher education aim to create a well-organized structure of educational indicators' monitoring in order to improve the educational process. They seek to provide a qualitative estimation of students' learnability and even to predict students' performance by discovering hidden information in educational systems. Classical systems perfectly work for the majority of students with common and consistent characteristics. However, such systems have drawbacks in terms of ignoring the minority of students, which are not aligned with the majority. This minority of students are called outliers. Outliers can contain significant information for domain experts. Thus, outlier detection becomes an essential tool for extracting this information. Our main goal is to detect outliers as students, which strongly deviate from majority of students in considered dataset in an effort to reveal the nature of outliers as well as the factors, which cause them.

2 Design

The data for analysis was collected during the monitoring of studying process at Tomsk Polytechnic University, Tomsk, Russia. The dataset includes grades that students obtained during the first semester of higher mathematics course. Apart from this, the dataset was enriched by results of enrollment tests and time of the passing attestation tests. Students pass enrollment tests of main subjects (mathematics, physics, and chemistry) in the begging of the semester in order to reflect the students' background before study. The time of passing attestation tests demonstrates the time point of submission for each task. Finally, 8 educational indicators were considered for 1066 students from 6 engineering faculties. The estimated histogram-based outlier score (HBOS) for each student demonstrates the level of their uniqueness. According to obtained scores, outliers were detected as students with HBOS higher than a given threshold.

To investigate the nature of outliers and factors, which cause them, the former ones were clustered using the agglomerative method that merged samples by similar characteristics.

3 Results

Proposed outlier detection approach for educational dataset was considered for 2 cases in order to investigate the HBOS sensitivity to changes in the initial values of indicators. The dissimilarity between cases is caused by different values of HBOS thresholds (t = 10 and t =12), which correspond to 3% and 5% average normalized frequency. We implied that student is an outlier if he/she has a value of normalized frequency for each indicator less than 3% and 5%. Both thresholds divided students into inliers and outliers — students with HBOS less and more than a given *t* respectively. The bigger value of HBOS the more unique students' characteristics. Outliers clustering allowed to determine groups of outliers with common features. The analysis of clusters revealed following outcomes:

1. Students with top marks during the entire term are rare as well as students with problems in learning. Majority of first year students have decline in the grades after enrollment tests, which might be due to the complexity of mathematics courses in Russian Universities.
2. Students with learning problems at the beginning of their study usually fail the course.

4 Implications

The outlier detection model for institutions of higher education enables to extract hidden information from investigated dataset. The implemented approach contains the combination of univariate methods that allow identifying outliers in dataset as objects, which are different from the majority of the data. We investigated outliers and their characteristics in terms of analyzed indicators by the case of students from a real data set. Our findings contain relevant knowledge about educational system and can be further used by domain experts to enhance the educational process.

5 Acknowledgments

Authors would like to acknowledge Dr. A.A. Mikhalchuk, associate professor from Tomsk Polytechnic University, who provided the anonymized educational dataset for this work.

References

1. Ogor, E.N., 2007. Student academic performance monitoring and evaluation using data mining techniques. In: Paper presented at the Electronics, Robotics and Automotive Mechanics Conference, 2007. CERMA 2007, p 354-359;

2. Chandola Varun, Kumar Vipin. (2009). Outlier Detection : A Survey. ACM Computing Surveys. 41, No. 15;

3. Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): a fast unsupervised anomaly detection algorithm. KI-2012: Poster and Demo Track, pages 59–63, 2012;

4. Hanan Aldowah, Hosam Al-Samarraie, Wan Mohamad Fauzy (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis, https://doi.org/10.1016/j.tele.2019.01.007.