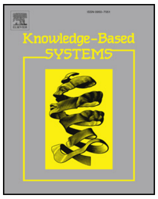
Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Are we meeting a deadline? classification goal achievement in time in the presence of imbalanced data

Martin Hlosta*[,a,b], Zdenek Zdrahal[a,c], Jaroslav Zendulka[b]

[a] Knowledge Media Institute, The Open University, Milton Keynes, UK
[b] Faculty of Information Technology, IT4Innovations Centre of Excellence, Brno University of Technology, Brno, Czech Republic
[c] CIIRC, Czech Technical University, Prague, Czech Republic

ARTICLE INFO

Keywords:
Classification
Imbalanced data
Learning analytics
Educational data mining

ABSTRACT

This paper addresses the problem of a finite set of entities which are required to achieve a goal within a predefined deadline. For example, a group of students is supposed to submit a homework by a specified cutoff. Further, we are interested in predicting which entities will achieve the goal within the deadline. The predictive models are built based only on the data from that population. The predictions are computed at various time instants by taking into account updated data about the entities. The first contribution of the paper is a formal description of the problem. The important characteristic of the proposed method for model building is the use of the properties of entities that have already achieved the goal. We call such an approach "Self-Learning". Since typically only a few entities have achieved the goal at the beginning and their number gradually grows, the problem is inherently imbalanced. To mitigate the curse of imbalance, we improved the Self-Learning method by tackling information loss and by several sampling techniques. The original Self-Learning and the modifications have been evaluated in a case study for predicting submission of the first assessment in distance higher education courses. The results show that the proposed improvements outperform the specified two base-line models and the original Self-Learner, and also that the best results are achieved if domain-driven techniques are utilised to tackle the imbalance problem. We also showed that these improvements are statistically significant using Wilcoxon signed rank test.

1. Introduction

Student retention has been recognised as a common problem both in distance Higher Education institutions and in Massive Open Online Courses (MOOCs) [1,2]. Learning Analytics (LA) and Educational Data Mining (EDM) are research fields that are trying to tackle this issue by examining available student data. They may include both static, e.g. mainly demographic data, and fluid data, e.g. data generated by students when interacting with a Virtual Learning Environment (VLE). These data are available for developing methods for identification of students who are at risk of failing courses. If such students are identified early enough, cost-effective support can be provided. Machine learning (ML) techniques are usually used to build models for predicting at-risk students. Predictions can be either made available directly to students [3] or mediated by tutors [4,5] who may offer additional knowledge not captured by the data, and take into account wider context, such as a student's personal circumstances.

The standard way to train the predictive models is to take advantage of the information from previous runs of the course. These models are applied to data of the current run. This approach is based on the assumption that the same or similar patterns of student behaviour prevail across subsequent years. The existing approaches differ in (1) specification of who are at-risk students; (2) available data for predictions; and (3) the machine learning algorithms used. For example, the "at-risk student" could be defined as one expected to achieve a final grade lower than C in [4]; or less than 60% in [6]; not submitting the next due assessment in [7], or one likely not to submit any other following assessment [8]. In [7], Wolff et al. show that not submitting the first assessment is a strong predictor of future failure.

For new courses, data from the previous courses ("legacy data") are not available and therefore cannot be used to build predictive models. For such cases, we proposed the Self-Learning approach [9].

In this paper, we further develop the Self-Learning philosophy and demonstrate how to predict students likely to fail based on failure to submit the first assessment. In addition, we propose further generalisations and improvements.

* Corresponding author at: Knowledge Media Institute, The Open University, Milton Keynes, UK.
E-mail addresses: martin.hlosta@open.ac.uk (M. Hlosta), zdenek.zdrahal@open.ac.uk (Z. Zdrahal), zendulka@fit.vutbr.cz (J. Zendulka).

https://doi.org/10.1016/j.knosys.2018.07.021
Received 26 January 2018; Received in revised form 27 June 2018; Accepted 3 July 2018

1.1. Self-Learning in the educational domain

To overcome the lack of legacy data, Self-Learning utilises the behaviour of those students who submit assessments in advance. We assume that the relevant patterns can be discovered in the VLE and demographic data. It is expected that learners who are about to submit will follow a similar pattern to those who have already submitted, and that such a pattern will be missing in the VLE data of students who will not submit.

1.1.1. Classification from imbalanced data

At the beginning, only a few students submit an assessment and the problem is inherently imbalanced. Classification from imbalanced data is a well-recognised problem in the ML field [10]. In many real-world supervised learning scenarios, a class exists that has significantly lower number of instances in the data than the other classes. The typical example is the medical domain where many fewer ill individuals exist compared to healthy ones. Another example includes enterprise credit evaluation environment, where at-risk companies are much rarer than normal ones [11,12]. It is not only the large disproportion between the number of instances representing different classes which causes the problem. Intuitively, if the concept that separates the data is not complex and if, for example, one attribute discriminates between the two classes perfectly, the classifier would still be able to provide predictions with high accuracy. However, as the complexity of class characteristics grows, the higher imbalance ratio causes greater errors [13]. In the last decade, the impact of imbalanced data in ML attracted significant attention from the research community and hundreds of papers have been published that discuss the sources of imbalanced data or how to improve performance under imbalanced data. As usual in ML, there is no guaranteed approach to all the problems and datasets, and many of these solutions are domain-dependent. The majority of the research is focused on binary classification but some recent works take the multi-class problem into consideration [14]. The most recent survey that covers many of the issues and also provides a taxonomy of the solutions comes from Branco et al. [15]. In the case of Self-Learner, the dataset evolves in time, more students submit and the imbalance ratio decreases.

Our previous experiments in [9] focus on daily prediction analysis, and compare various ML methods and their ability to deal with imbalanced data. Area Under the Precision-Recall Curve *(PR AUC)* is selected for evaluation because it is a convenient criterion when dealing with imbalanced data [16].

The evaluation shows that the performance is lower the further the prediction is to the deadline. The best performance is achieved by ensemble-based classifiers, XGBoost [17] based on boosting followed by Random Forest based on bagging. Some algorithms, e.g. Support Vector Machines (SVM) or Logistic Regression, offer the ability to compensate for the lower number of instances of the minority classes in the training process. Such algorithms perform better than their original, uncompensated versions.

1.2. Generalisation of the concept

The proposed Self-Learning method is primarily targeted on identifying students at risk of not submitting the first assessment. As suggested in [9, sec. Discussion], the same approach could predict the results of other milestones in the course, i.e. submission of further assessments. Given appropriate data, the application domain does not need to be limited to education. However, two conditions need to be satisfied: (1) the existence of the deadline within which the goal must be satisfied and also (2) the existence of students/entities that achieve this goal before the deadline. Motivated by this, we posed the first research question:

- *RQ1:* How can we formalise the problem of classification whether individuals in a population will satisfy a goal within a specified deadline?

1.3. Time in imbalanced data classification

Temporal changes of the class imbalance ratio have generated considerable research interest. The survey from 2016 by Krawczyk et al. [18] discusses open challenges in ML from imbalanced data, and mentions learning from imbalanced data streams among them. The usual problem of data streams is their dynamic nature: the distribution of the data can change. For example, the imbalance ratio between classes can change, and also a different class can dominate as time progresses. In particular, a topic related to imbalanced data that still needs to be researched further is the problem of *new class emergence* [18], where the number of instances of the minority class is highly under-represented in the beginning and then grows over time.

Wang et al. [19] investigate changes of imbalance ratio depending on different speeds of change. The experiments compare over-sampling and under-sampling bagging methods, with over-sampling bagging being better. The performance, however, drops immediately after the imbalance has changed. The results improve when combining both methods with adaptive weights. Together with synthetic data, the results are examined on two real-world scenarios of fault detection. A similar task is studied by Tan et al. in [20], where they focus on predicting defect introducing changes in the source code from the versioning system of open source projects. The goal is to detect changes of the source code that are later fixed and marked as bugs. Changes of code arrive permanently. The results show improved performance when using sampling methods against baseline and against *updatable classification methods*. Although four types of sampling have been used, the results presented in [20] do not provide sufficient details, e.g. which sampling performs best.

The specificity of the problem with student assessment submissions, or generally goal achievement as introduced above, lies in the presence of the deadline. Although the tasks presented in [19,20] generate imbalanced data by their nature, the absence of the deadline makes their problem different. Compared to their scenario, in our case, we receive new observations about a stable set of entities. Also, rather than an abrupt change, we expect a gradual increase of submissions at the beginning followed by a steep increase closer to the deadline. Consequently, most of the submissions usually occur close to the deadline. This is also confirmed by our previous results in [9] and by other studies [21–23]. This phenomenon can be attributed to the well-known psychological problem of procrastination, i.e. postponing or avoiding of starting, engaging in, or completing a task [24]. Since the models are constructed from the data of the same course that is being predicted, in the beginning, the methods suffer from the imbalanced data, i.e. the lack of information.

A concept similar to the Self-Learning framework is *Self-Training*, which is used in some semi-supervised learning problems [25]. This technique utilises both labelled and unlabelled datasets to improve the performance of the classification. First, the model is trained solely on the labelled examples, and the unlabelled ones are then iteratively added until the performance of the classifier stops improving. In [25], Stanescu and Caragea use the original Self-Training method with several modifications tailored to imbalanced data, achieving the best results when the training set is extended only with the examples predicted as a minority class. The difference between Self-Learning approach and the Self-Training in [25], and semi-supervised methods in general, is the absence of annotated entities of the negative class, *NotAchieve* in our case. In contrast, Self-Learning uses the temporal character of the data to construct the negative class examples from the pool of available entities, e.g. students in our case.

Our previous results [9] compared existing ML methods and methods for dealing with imbalanced data (sampling and algorithm based methods). In the beginning, the lack of information worsens the performance. The improvement due to the use of methods developed to tackle imbalanced data is negligible. This opens the potential for improvement, for instance, using domain knowledge. The dynamic nature