Application of the Dynamic Time Warping Distance for the Student Drop-out Prediction on Time Series Data

Alexander Askinadze
Institute of Computer Science
Heinrich Heine University Düsseldorf
askinadze@cs.uni-duesseldorf.de

Stefan Conrad
Institute of Computer Science
Heinrich Heine University Düsseldorf
conrad@cs.uni-duesseldorf.de

ABSTRACT

It is reported by different universities that over 40% of students do not complete their studies within 6 years. Especially in technical courses, the drop-out rate is already very high at the beginning. Therefore an automatic drop-out prediction is useful for a monitoring system. Since the study progress data can be sorted by time, we show how they can be transformed into a multivariate time series. Then we examine the dynamic time warping (DTW) distance in conjunction with the k-nn classifier and show how DTW can be used as an SVM kernel for drop-out prediction on the time-series data. With this approach, we are able to recognize about 67% of the drop outs from the course of study after the first semester and about 60% after the second semester.

1. INTRODUCTION

The number of drop out is a big problem for many universities. Over 40% of students do not complete their studies within 6 years [1]. Especially in technical courses, the number of drop outs in the first semesters is high. So in [5] it is reported that in the Electrical Engineering course the drop-out rate of beginners is about 40%. Human monitoring is used to solve this problem [5]. With a large number of students, this can lead to a huge manual effort, so that a machine-made pre-selection could facilitate the work of a human decision-maker. Most students fail in the first semesters, which requires an early prediction. The quality of the available data is very important for automatic drop-out prediction. However, due to data protection laws, often little data are available for use. The data is often restricted to only a small amount of private data and the study progress data, so that only examinations, their corresponding grades, and the number of attempts per semester are given. Because of the dearth of data, it is important to obtain as much semantics as possible from the data, such as temporal aspects. The study progress data can be viewed as a multivariate time series. In this paper, we will investigate methods that can perform drop-out predictions on time-series data.

2. RELATED WORK

Many studies have been published on student drop-out prediction like [1], [5], [6]. The data mining methods used include SVM, decision trees, k-nn, and neural networks. Studies were also made in the field of time series analysis. In [6] the authors investigated time series clustering to identify at-risk online students. Several studies, for example [7], have used DTW for time series clustering to identify distinct activity patterns among students. The results of the individual publications are difficult to compare with each other because the data used and the goals are very different. While some seek to prevent drop outs from a study subject, others seek to prevent drop outs from the whole study. We also use DTW, but not for clustering, but as distance for classifiers.

3. METHOD

If only the data of the study progress are available per semester, as much semantic information as possible must be collected from the data. Assuming that the study progress of a student S consists of n semesters $s = \{sem_1, ..., sem_n\}$, a function $\Phi : s \rightarrow T^S_{n,m}$ with $\Phi(s) = \Phi(\{sem_1, ..., sem_n\}) = \{\phi(sem_1)^\top, ..., \phi(sem_n)^\top\} = \{s_{1\leq k\leq n} = [s_{1,k}, ..., s_{m,k}] \in \mathbb{R}^m\} = T^S_{n,m}$ which transforms the ordered set s into a multivariate time series is needed.

In each semester the students have the possibility to take q courses. The results of each course can be expressed by a number p of properties such as the final score or the number of trials. All information of a semester can thus be represented in a vector of size $m = q \times p$. If, for example, the 3 properties were: 1) achieved grade (numeric), 2) passed (binary), and 3) number of attempts (numeric), and in a certain semester, a student had taken the first and last course from the list of all possible courses then the resulting vector for a semester could look like the one shown below.

$$\phi(sem_i) = \left[\ \underbrace{5\quad 1\quad 1}_{course\ 1}\quad \underbrace{0\quad 0\quad 0}_{...}\quad \underbrace{1\quad 0\quad 2}_{course\ q}\ \right]$$

Thus, we can represent a student as a temporal sequence of his completed semesters. To compare these two sequences, we need a distance for multivariate time series. A well-researched distance for time series is the d_{DTW} distance.

Dynamic Time Warping (DTW) [3] is an algorithm from the domain of time series. It is generally defined for univariate time series and can be used to calculate a distance of the two time series $a = (a_1, ..., a_n), a_i \in \mathbb{R}$ and $b = (b_1, ..., b_m), b_j \in \mathbb{R}$ with different length. To extend the DTW distance for multivariate time series, various methods have been proposed in the literature like DTW_D [8]. DTW_D is calculated just as in the one-dimensional case, except that the pairwise distance $d(a_i, b_j)$ is calculated with the Euclidean distance.

The drop-out prediction is a binary problem. One of the most popular binary classifiers is the *support vector machine* (SVM) [4] because it can separate linear separable sets optimally from each other. If the training dataset is not linearly separable, a kernel trick is used to solve the problem. An often used kernel is the Gaussian kernel. In [2], an adaptation of the Gaussian kernel to the Gaussian DTW (GDTW) kernel was made for sequential data. The GDTW kernel K_{GDTW} can be defined by $K_{GDTW}(x, y) = e^{-\gamma d_{DTW}(x,y)}$.

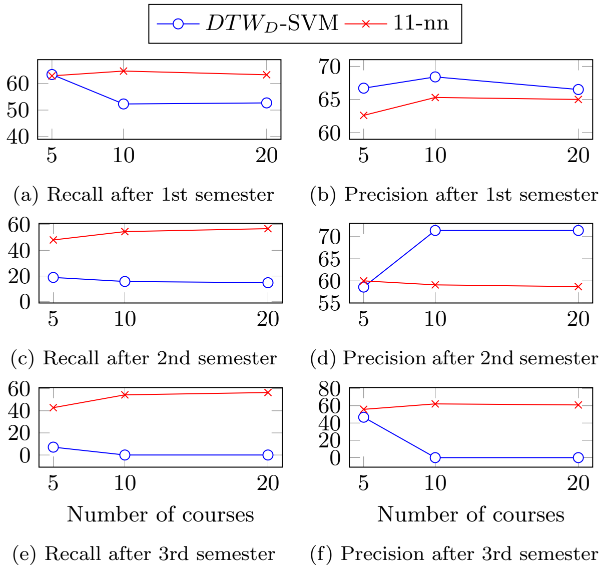
4. EVALUATION

We have a data set with 704 students of which 310 did not successfully complete their studies within 10 semesters. For each student the following information per semester is available: idCourse, number of attempts, examination status (passed, failed), recognized exam (true, false), reached grade, and semester. We use recall and precision as evaluation measures. The evaluation is performed 3 times for all parameters with a 10-fold cross-validation for the two approaches DTW_D-SVM and DTW_D-k-nn (ordinary k-nn classifier that uses the DTW_D distance). It is examined per semester how good the prediction is at the end of the semester. For example, the students who have studied at least 2 semesters are considered for the training and prediction of the drop out after the second semester. The length of the resulting multivariate time series vectors depends strongly on the number of courses used. Therefore, we will examine the influence of the number of courses used to create the vectors. In the dataset there are more than 100 courses. Because most students of our dataset drop out after a few examinations, we sort all courses according to the number of students who have enrolled in them. Then the 5, 10 and 20 courses with the highest enrollment will be used for further study. After the first investigation, we have found that the k-nn parameter $k = 11$ is comparatively well suited and is therefore used for the evaluation.

We first consider the prediction after the first semester. The recall and precision results are shown in Figure 1. In the second semester, 609 students are still active, of whom 215 will be leaving in the future. After the last examination of the second semester, almost 60% of these students can be recognized with 11-nn. The precision of 11-nn is also about 60%. DTW_D-SVM achieves a 10% higher precision when using more than 10 courses to create the multivariate time series vectors. However, the recall value of DTW_D-SVM is significantly smaller. At the end of the 3rd semester, the limits of DTW_D-SVM are recognizable. Both the recall and the precision values are smaller for 5 courses, and decrease to 0 for more used courses. 11-nn remains stable and provides similar results as after the second semester. In the third semester, 542 students are still active, of whom 143 will be leaving. 11-nn can recognize about 84 of these 143 students.

5. CONCLUSION

We have shown how a study progress can be transformed into a multivariate time series. Then we demonstrated that the DTW_D distance can be used within an SVM kernel to make an SVM usable for student time series data. We compared the DTW_D-SVM with the 11-nn classifier, which also uses the DTW_D distance on a dataset with 704 students and found that the k-nn classifier is better suited to achieve higher recall values in the drop-out prediction. The DTW_D-SVM is only suitable until the second semester and provides better precision results. In the later semesters, the values become worse due to most of the students in the first semester fail because of a few specific courses. For the students from the technical courses, it is usually the first mathematics courses. In the later semesters, the reasons cannot be stated so easily. Generally this approach is only for the prediction and not to determine the reasons. In future work we want additionally determine the reasons for drop outs.



(a) Recall after 1st semester

(b) Precision after 1st semester

(c) Recall after 2nd semester

(d) Precision after 2nd semester

(e) Recall after 3rd semester

(f) Precision after 3rd semester

Figure 1: Recall and Precision results

6. REFERENCES

[1] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West. Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*, 2016.

[2] C. Bahlmann, B. Haasdonk, and H. Burkhardt. Online handwriting recognition with support vector machines-a kernel approach. In *Frontiers in handwriting recognition, 2002. proceedings. eighth international workshop on*, pages 49–54. IEEE, 2002.

[3] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[4] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[5] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers. Predicting students drop out: A case study. In *Educational Data Mining 2009*, 2009.

[6] J.-L. Hung, M. Wang, S. Wang, M. Abdelrasoul, W. He, et al. Identifying at-risk students for early interventions? a time-series clustering approach. *IEEE Transactions on Emerging Topics in Computing*, 2015.

[7] E. Młynarska, D. Greene, and P. Cunningham. Time series clustering of moodle activity data. In *24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16), University College Dublin, Dublin, Ireland, 20-21 September 2016*, 2016.

[8] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh. Generalizing dtw to the multi-dimensional case requires an adaptive approach. *Data Mining and Knowledge Discovery*, 31(1):1–31, 2017.