May 06, 2022

A Survival Analysis based Volatility and Sparsity Modeling Network for Student Dropout Prediction

Feng Pan[1], Bingyao Huang [1], Chunhong Zhang[1], Xinning Zhu[1], Zhenyu Wu[1], Moyu Zhang[1], Yang Ji[1], Zhanfei Ma[2], Zhengchen Li[3]

[1]Beijing University of Posts and Telecommunications; [2]Baotou Teachers' College; [3]Shenyang Polytechnic College

dx.doi.org/10.17504/protocols.io.b4duqs6w

P Feng Pan
Beijing University of Posts and Telecommunications

Student Dropout Prediction (SDP) is of pivotal significance in mitigating withdrawals in Massive Open Online Courses. Research in these areas are usually carried out using deep learning to detect complex nonlinear patterns in students' learning sequences. However, the volatility and sparsity of data always weaken the performance of deep neural networks. Prevailing approaches always required an additional data smoothing or interpolation step independent of the prediction model, which may lose valuable information or introduce inauthentic data. Besides, when modeling the SDP problem as a binary classification task, previous works often required to specify an observation window, which may lead to inconsistent prediction results with different settings. To address these issues in an end-to-end learning framework, we propose a Survival Analysis based Volatility and Sparsity Modeling Network (SAVSNet). Particularly, SAVSNet smooths the volatile time series by convolution network while preserving the original data information using Long-Short Term Memory Network (LSTM). Furthermore, we propose a Time-Missing-Aware LSTM unit to mitigate the impact of data sparsity by integrating informative missingness patterns into the model. To achieve consistent predictions along time, a survival analysis loss fuction is adopted for parameter estimation and the model outputs monotonically decreasing survival probabilities. In the experiments, we compare the proposed method with state-of-the-art methods in two real-world MOOC datasets and the experiment results show the effectiveness of our proposed model.

DOI

dx.doi.org/10.17504/protocols.io.b4duqs6w

https://doi.org/10.1371/journal.pone.0267138

Citation: Feng Pan, Bingyao Huang , Chunhong Zhang, Xinning Zhu, Zhenyu Wu, Moyu Zhang, Yang Ji, Zhanfei Ma, Zhengchen Li A Survival Analysis based Volatility and Sparsity Modeling Network for Student Dropout Prediction
https://dx.doi.org/10.17504/protocols.io.b4duqs6w

Feng Pan, Bingyao Huang , Chunhong Zhang, Xinning Zhu, Zhenyu Wu, Moyu Zhang, Yang Ji, Zhanfei Ma, Zhengchen Li 2022. A Survival Analysis based Volatility and Sparsity Modeling Network for Student Dropout Prediction. **protocols.io** https://dx.doi.org/10.17504/protocols.io.b4duqs6w

protocol

Pan F, Huang B, Zhang C, Zhu X, Wu Z, Zhang M, Ji Y, Ma Z, Li Z (2022) A survival analysis based volatility and sparsity modeling network for student dropout prediction. PLoS ONE 17(5): e0267138. doi: 10.1371/journal.pone.0267138

Student Dropout Prediction, Survival Analysis, Volatility, Sparsity

protocol ,

Jan 27, 2022

May 06, 2022

57492

1 Prepare the dataset:
We conduct experiments on two benchmark datasets to evaluate our proposed model. Both of them are drawn from the largest MOOC platform in China, XuetangX (see https://www.xuetangx.com/). The first dataset KDDCup 2015 is available at https://www.biendata.xyz/competition/kddcup2015/data/, which have been widely used for various MOOC dropout prediction studies. The dataset contains information about 39 courses and 72,395 enrolled students. Each course takes 30 days as the history time window and the prediction period is set to 10 days. 7 different event types of student learning activity (i.e., features) are provided in KDDCup 2015. We use KDDCup 2015 to compare our proposed method with existing methods. The second dataset XuetangX is available at http://moocdata.cn/data/user-activity, which is much larger than KDDCup 2015. The original XuetangX dataset contains 246 courses, 202,000 students and 22 event types. But to facilitate model training, we adopt the data processing method used in "Prenkaj B, Velardi P, Distante D, et al. A reproducibility study of deep and surface machine learning methods for human-related trajectory prediction[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 2169-2172.", i.e., pruning all e-courses with less than 350 student trajectories, which leaves us with 19 courses and 23839 students. For the XuetangX dataset, the history period is 35 days and the prediction period is still 10 days. We use XuetangX to test the robustness and generalization of the proposed method.

2 Data processing:
For each dataset, we perform a daily grouping of events concerning each student to help in deriving the temporal sequences. Therefore, we represent each student's trajectory with a time-matrix T_u∈ $R^{(}l, n)$, where l is the length of the adopted time-window in days and n is the number of different events types available in the dataset. Moreover, we add the elapsed time Δt and the missing indicator u_t to each time step. Therefore, the final input matrix is a tensor of shape (N,l,n+2).

3 Build model:
Our proposed model, SAVSNet, is designed as three stages deep learning networks: a volatility modeling network, a sparsity modeling network and a survival analysis network. (1) In the volatility modeling network, we leverage a dual-branch structure, which is consisted of an 1D convolutaional network to capture the spatial dependencies of multivariate time series and an LSTM network to learn the temporal dependency, respectively. Then we ultilize an update gate to adaptively fuse the two hidden representations. (2) In the sparsity modeling network, we use a customized Time-Missing-Aware LSTM (TM-LSTM) unit to take the elapsed time and missing indicator between the consecutive timestamps of a sequence into consideration to adjust the long-term and short-term memory content of the LSTM unit. (3) In the survival analysis network, we apply a softplus function to transform the output of TM-LSTM into hazard rate that represents the instantaneous hazard of the student given no dropout event occurred before. Then we employ a survival analysis loss function to estimate the hazard rates. Accordingly, the survival probability at each time step is obtained by the formula ${S_t} = {e^{ - \sum\nolimits_{k = 1}^t {{\lambda _k}} }}$. By comparing the survival probability at the last timestamp with a threshold τ, we can predict whether a user would dropout of the course.We use the Tensorflow1.x platform to build this model and the source code will be published at: https://github.com/leondepf/SAVSNet.

4 Model training and testing:
The data is trained in the built model, and this process is carried out on the GPU. At the end of this process, we get some metrics that evaluate our model's perfromance.

Citation: Feng Pan, Bingyao Huang , Chunhong Zhang, Xinning Zhu, Zhenyu Wu, Moyu Zhang, Yang Ji, Zhanfei Ma, Zhengchen Li A Survival Analysis based Volatility and Sparsity Modeling Network for Student Dropout Prediction
https://dx.doi.org/10.17504/protocols.io.b4duqs6w