Realtime-Robust Decision-Making via Concept Drift Learning

Kun Wang, Jie Lu, Guangquan Zhang, Anjin Liu
University of Technology Sydney, Faculty of Engineering and Information Technology (FEIT),
Australia Artificial Intelligence Institute (AAII)

Introduction

Concept drift describes changes in the data distribution of streaming data that indicate the current model cannot sufficiently maintain accuracy and efficiency. For example, in the prediction of customer churn, due to the customer's own choice and the external environment will change randomly, the prediction result will be biased, as shown in Fig. 1. In order to improve the robustness of the data stream learning model, it is necessary to focus on the time segmentation optimization of training data to meet testing needs in the short-term and the model learning process optimization in the long-term, then think about how to leverage them to avoid model underfitting or overfitting. To address these challenges, this research will develop a method of balancing time-dependent over-fitting and under-fitting to obtain robustness on streaming data learning.

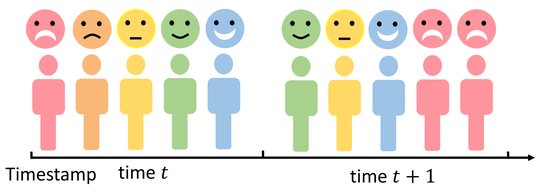

Fig. 1 Customer churn change.

Materials

Under concept drift, the global minimum of a learning model might drift to a new location so that the model no longer adequately performs the task it was designed for, as shown in Fig. 2.

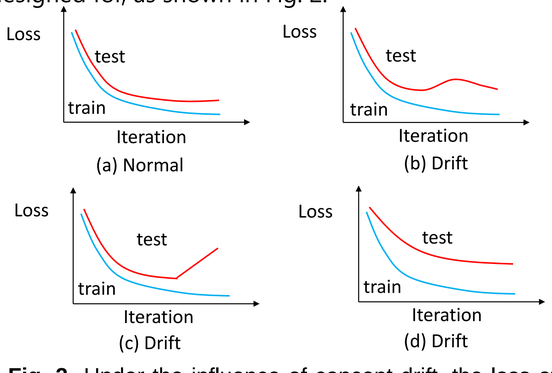

Fig. 2 Under the influence of concept drift, the loss of model will difficult to get to the global minimum.

Methodology

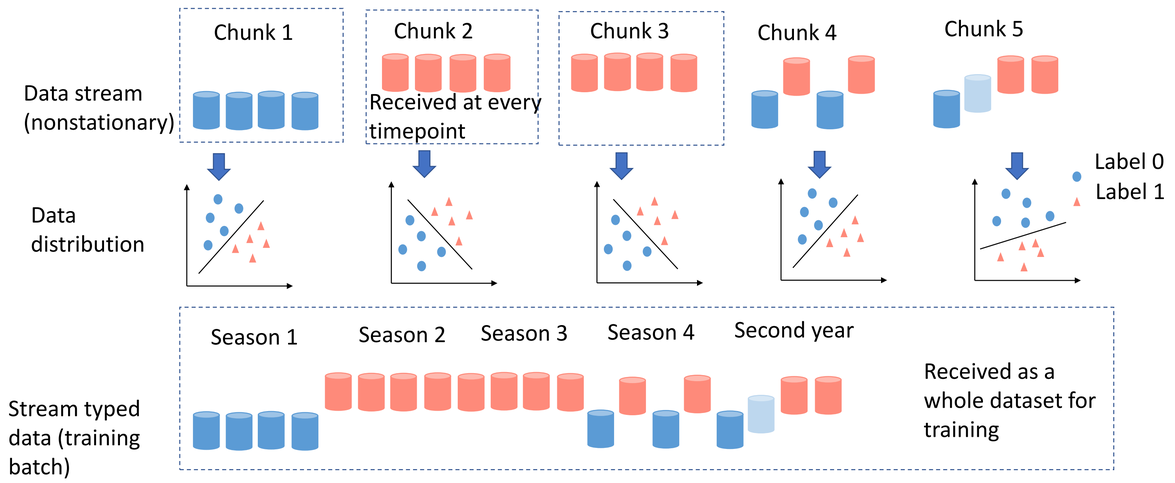

Fig. 3 Schematic diagram of concept drift in data stream and stream typed data.

Data stream has the time sequence, as shown in Fig. 3. Our aim is to make reasonable time segmentation on streaming data for vary concept drift recognition and measurement. Then, we want to balance the learning system between lifelong optimization from the long-term perspective and time-segmentation optimization from the short-term perspective.

Firstly, we propose an elastic Gradient Boosting Descent Tree (eGBDT) algorithm, which can delete redundant trees without increasing the runtime complexity, as shown in Fig. 4. We choose NOAA weather datasets for experiment, as shown in Fig. 6.

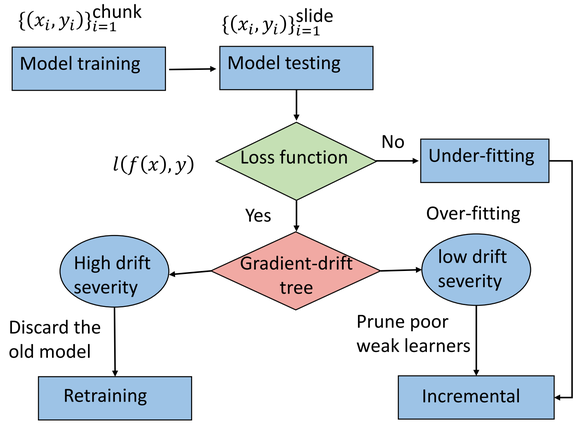

Fig. 4 The eGBDT learning method.

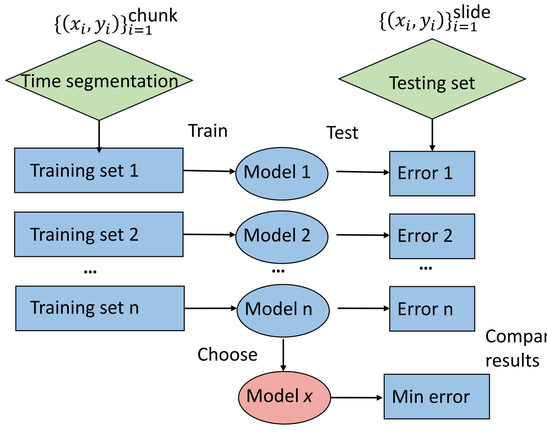

Fig. 5 Time segmentation method.

Secondly, we will propose a strategy of time segmentation on data to make sure model learn the knowledge needed for testing, as shown in Fig. 5.

- Step 1: Consider segment data stream to maintain knowledge.
- Step 2: Separately training model on segmented data and compare the model performance.
- Step 3: Organize these training models and targeted use.

Results

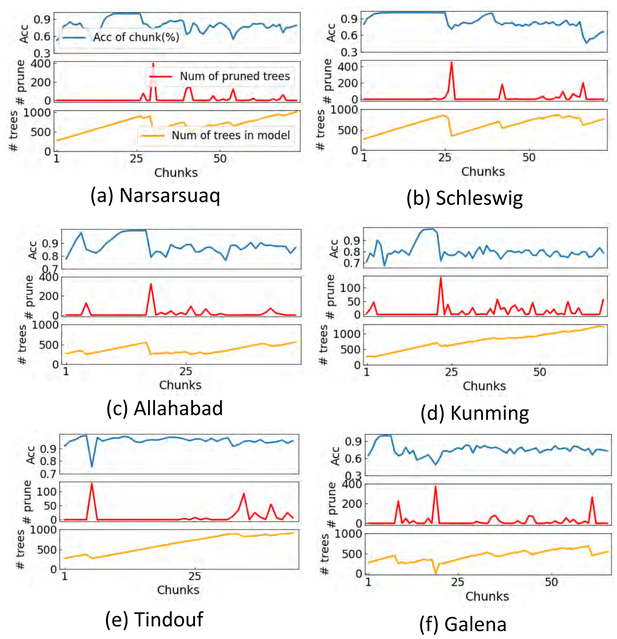

Fig. 6 A plot of the chunk accuracy, the number of pruned trees, and the number of trees in eGBDT during drift learning. We can clearly see that the accuracy dropped, the number of pruned trees increased.

Conclusion

- The proposed eGBDT integrates incremental learning and tree pruning to dynamically adjust the number of trees for different stream situations. The incremental learning of GBDT can help the model to improve performance. The process of tree pruning is a simple but efficient way to increase model stability and accuracy.
- The time segmentation method can be used in real-world stream typed data to help choose the appropriate training samples needed for the test to reduce the model test error.
- The research of balance the lifelong optimization from the long-term perspective and time-segmentation optimization from the short-term perspective is still undertaking. We will also experiment on synthetic and real-world datasets.