Modeling Recurring Concepts in Imbalanced Data Streams

Classifying an imbalanced stream of non-stationary data with recurring drift is a challenging task. Recurring concept drifts occur when the same concept, observed in the past, reappears in the data stream. In the real world, rush-hour in traffic flows and seasons are examples of repeated patterns. A common approach to dealing with recurring drift is to retrain a classifier whenever a change is detected in the underlying distribution of data. However, if a recurring concept occurs, the model then needs to relearn the recurring concept. If we can recognise recurring concepts, we may perform faster and better by reverting to the classifiers previously trained on those concepts. Most of the existing approaches such as GraphPool [1], RCD [2], Diversity Pool [3], and ECPF [4], dealing with recurring concepts, do not have any mechanism for handling the class imbalance problem, where the distribution of instances with respect to the classes is not equal, and often evolves in the streaming data. Having no mechanism for handling the class imbalance results in models that have poor predictive performance, particularly for the minority class which is often of higher importance. Therefore, a better approach taking into account the class imbalance is needed.

In this Master's thesis, we aim to model the recurring concepts in data streams whose distribution is skewed. For this purpose, the following subtasks must be done: 1) implementing a graphical model which models the concepts in an imbalanced data stream as states/nodes; 2) incorporating one oversampling method (random oversampling, localized oversampling, SMOTE, ...) to increase the number of minority instances before the classifier gets assigned to one state/node, and thus tackle the class imbalance; 3) dynamic visualization of graphical model using the Directed Graph.

Requirements:

- prerequisite course: Data Mining 2

- solid programming skills in Python and its libraries (Numpy, scikit-learn, scikit-multiflow, ...)

If you are interested, send an email to Amir Abolfazli (abolfazli[at]l3s.de) with your up-to-date CV and transcript of records attached.

References:

1. Ahmadi, Zahra, and Stefan Kramer. "Modeling recurring concepts in data streams: a graph-based framework." Knowledge and Information Systems 55.1 (2018): 15-44.

2. GonçAlves Jr, Paulo Mauricio, and Roberto Souto Maior De Barros. "RCD: A recurring concept drift framework." Pattern Recognition Letters 34.9 (2013): 1018-1025.

3. Chiu, Chun Wai, and Leandro L. Minku. "Diversity-based pool of models for dealing with recurring concepts." 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018.

4. Anderson, Robert, et al. "Recurring concept meta-learning for evolving data streams." Expert Systems with Applications 138 (2019): 112832.