Non-stationary data stream processing with metafeature analysis

Joanna Komorniczak

Abstract

The doctoral dissertation focuses on topics related to non-stationary data stream processing. The machine learning methods proposed in the dissertation allow for solving two key tasks in the considered field: *concept drift detection* and *classification of non-stationary data streams* using *hybrid* approaches. The solutions proposed and examined in the dissertation are based on the analysis of *metafeatures* calculated for disjoint data batches arriving in the form of a stream. The dissertation aimed to verify the following hypothesis:

> *It is possible to propose methods employing metafeature analysis for concept drift detection and classification of the non-stationary data streams that demonstrate significantly better or statistically dependent recognition quality compared to state-of-the-art approaches.*

Based on the conducted research, the hypothesis was substantiated by achieving the following objectives:

Objective 1: Proposal of an implicit concept drift detector analyzing the time variability of the classification task complexity measures calculated for disjoint data chunks. The objective was achieved by proposing the *Complexity-based Drift Detector*, which analyzes metafeatures describing the complexity of the classification task. The detection mechanism operates independently of the classification quality, allowing for *implicit* concept drift detection. The proposed approach uses an ensemble of one-class classifiers trained with data metafeatures to distinguish between representations of batches from different concepts, allowing for effective recognition of changes occurring in the data streams.

Objective 2: Proposal of an implicit concept drift detector analyzing drift magnitude measures integrated using the ensemble learning paradigm. The objective was achieved by proposing the *Statistical Drift Detection Ensemble*, which uses *drift magnitude* and *conditioned marginal covariate drift* measures calculated for disjoint data batches of a data stream. The drift detection method uses an ensemble approach to integrate constituent decisions of the concept change recognition.

The metafeatures used in this method are dedicated to analyzing concept changes in data streams, directly indicating probability distribution shifts. The original measures showed low informativeness in the case of processing high-dimensional data. The ensemble approach proposed in the detector allows for effective drift recognition by analyzing low-dimensional subspaces and the integration of constituent detections within the subspaces using the ensemble learning paradigm.

Objective 3: Proposal of an unsupervised drift detection method analyzing the distribution of activations from the last layer of a deterministic neural network. The objective was achieved by proposing the *Parallel Activations Drift Detector* method, which allows for unsupervised detection of concept drifts based on the outputs of a deterministic neural network. The metafeatures used in the proposed approach are defined as random projections generated using the neural network initialized with random weights, invariant during the processing. The metafeatures defined in such a way implicitly describe the location of samples in multidimensional feature space. Analysis of their variability using replicated paired statistical tests allows for identifying significant changes in the data distribution that indicate concept drifts.

Due to the lack of a reliable criterion for assessing the quality of the concept drift detection task in the literature, which would allow for examining the quality of methods in the case of data streams with incremental and gradual drifts, three *drift detection error* measures were proposed in the dissertation. The proposed evaluation criteria employ the analysis of distance and the cardinality of drifts and detections. Additionally, the quality assessment was supported by a visual analysis of the moments of concept change signaling.

Comparison of all proposed concept drift detection methods with other *state-of-the-art* approaches on an extensive pool of synthetic data streams allowed for demonstrating, depending on the examined criterion, significantly better or statistically dependent detection quality compared to the methods known from the literature.

Objective 4: Proposal of an ensemble method for classification of data streams analyzing the distributions of statistical metafeatures calculated for subsequent data chunks to identify recurring concepts. The objective was achieved by proposing the *Metafeature Concept Selector* method, which uses a set of statistical metafeatures to detect concept drifts and to re-identify concepts that occurred in the past in the case of their recurrence. The proposed algorithm uses a pool of one-class classifiers to distinguish concepts based on the calculated metafeatures and, in parallel, a pool of classifiers responsible for performing the recognition task.

In the experiments examining the method's operation, the quality of concept identification was assessed using the *Rand* metric dedicated to the evaluation of clustering tasks. The metric allowed for an unambiguous and reliable assessment of the concept recognition quality, regardless of the classification mechanism of the baseline classifier. In further experiments, the performance of the proposed solution was compared with independent classifiers, which showed a statistically significant improvement in the classification quality when using the proposed approach.

Objective 5: Proposal of a classification method for compensating the bias of baseline classifiers when processing the data streams with dynamic changes in imbalance ratio, using the prior probability estimated based on the metafeatures of the processed data chunk. The objective was achieved by proposing *Prior Probability Assisted Classifier*, which uses an estimated prior probability to compensate for the classifier's bias towards the majority class. The estimation of prior probability is based on metafeatures calculated for subsequent batches of data streams. Based on the estimated level of imbalance in a given data batch, the proposed method corrects the predictions of the baseline classifier, ensuring a specific number of objects of a given class.

In the experiments, different strategies for estimating the prior probability were compared. The most universal approach was the *Dynamic Statistical Concept Analysis*, an original method estimating the current prior probability based on a pair of regressors analyzing the mean values and the standard deviation of samples within the processed classes. The results showed a statistically significant improvement in the classification quality when using the proposed method.

Objective 6: Proposal of a framework for processing data streams with a time-varying level of difficulty, allowing for the selection of an appropriate neural network architecture based on the analysis of the model's certainty. The objective was achieved by proposing a *Certainty-based Architecture Selection Framework* – a processing scheme that uses the classification model's *certainty* level in the classification task of data streams characterized by time-varying difficulty. Using the *certainty* of the classifier as a metafeature allows for avoiding the computational overhead associated with defining additional indicators describing the data. The *support function* values provided by the neural network were used to calculate the metafeatures and then to determine the predictions for the processed objects.

The performed experiments focused on the classification of semi-synthetic computer vision data streams. The proposed processing scheme showed the ability to dynamically switch architectures of *convolutional neural networks* of different complexities, trained for the given problem. The results of the experiments showed that the proposed solution allows for a significant reduction in processing time complexity and the number of operations of the system with a subtle reduction in recognition quality.

Keywords machine learning, data streams, concept drift, concept drift detection, classification, ensemble classifiers, metafeatures, imbalanced data

17.04.2025
Joanna Komorniczak