Wydział Elektroniki
Katedra Systemów i Sieci Komputerowych
Politechnika Wrocławska

Classifier selection for imbalanced data stream classification

Selekcja klasyfikatorów w zadaniu klasyfikacji niezbalansowanych strumieni danych
(streszczenie)

Paweł Zyblewski

Słowa kluczowe: *Rozpoznawanie wzorców; uczenie indukcyjne; klasyfikacja; zespół klasyfikatorów; selekcja klasyfikatorów; przetwarzanie wstępne danych; dane trudne; dane niezbalansowane; strumienie danych; dryf koncepcji; uczenie aktywne.*

Wrocław 04.2021

Abstract

The thesis focuses on the use of the *Dynamic Ensemble Selection* algorithms in conjunction with data preprocessing techniques in the tasks of the stream and imbalanced data classification. The aim was to present the natural ability of classifier selection algorithms to deal with data imbalance and to propose new, effective solutions to the rarely discussed problem of highly imbalanced data stream classification. Based on these assumptions, the following hypothesis was formulated

> *There exist such methods employing data preprocessing and classifier selection that can outperform state-of-the-art classifiers for difficult data classification tasks.*

The hypothesis was substantiated by achieving the following goals:
***Goal 1* – Developing an ensemble selection algorithm for imbalanced data classification, as well as designing a dedicated combination rule.**
This goal was met by developing three algorithms based on the clustering of models in a one-dimensional space of classifier diversity. To construct this clustering space, the H measure, informing about about the impact of individual classifiers on the ensemble diversity, was proposed.

The *Diversity Ensemble Pruning* (DEP) prunes the ensemble by selecting, from each cluster, only the model with the highest BAC value. The *Two-step majority voting organization* (TSMV) algorithm classifies imbalanced data using the two-step voting structure. The *Random Sampling Multistage Organization* (RSMO) algorithm, additionally uses sampling with replacement to reduce the number of similar models involved in the decision-making process.

***Goal 2* – Proposing a novel distance-based *Dynamic Ensemble Selection* method for imbalanced data classification.**
This goal was met by proposing novel *Dynamic Classifier Selection* algorithms for the imbalanced data classification problem. Two methods were proposed, namely *Dynamic Ensemble Selection using Euclidean distance* (DESE) and *Dynamic Ensemble Selection using Imbalance Ratio and Euclidean distance* (DESIRE), which use the Euclidean distance and *Imbalance Ratio* in the training set to select the most appropriate model for

the classification of each new sample. DESE performs the selection based on local competencies and distance to classified neighbors, while DESIRE additionally scales the obtained weights by *Imbalance Ratio* of the problem.

Goal 3 **– Developing a chunk-based ensemble algorithm, aimed specifically for the task of highly imbalanced data stream classification.**
This goal was achieved by proposing the *Minority Driven Ensemble* (MDE) algorithm. This algorithm classifies highly imbalanced data streams using a decision rule exploiting local data characteristics to prefer the minority class instances.

Goal 4 **– Designing a novel framework combining *Dynamic Ensemble Selection* and preprocessing techniques for imbalanced data stream classification.**
This goal was achieved by proposing two batch-based approaches, combining *Dynamic Classifier Selection* algorithms and preprocessing techniques for the task of highly imbalanced data stream classification. The *Dynamic Ensemble Selection for Imbalanced Stream Classification* (DESISC) method generates a single model on each data chunk, while the *Dynamic Ensemble Selection for Imbalanced Stream Classification approach using Stratified Bagging* (DESISC-SB) employs a stratified version of *Bagging* for the base classifier generation.

Goal 5 **– Proposing a strategy for learning from drifting data stream under limited access to labels scenario.**
This goal was achieved by the introduction of the *Budget Active Labeling Strategy* (BALS) algorithm. The proposed approach, in addition to the pool of objects selected for labeling based on their distance to the decision boundary, also received a small number of randomly selected objects.

Goal 6 **– Evaluating the behavior of the previously proposed data stream classification framework, taking into account the limitation in the label access.**
This goal was achieved by combining the proposed DESISC-SB framework with the active learning method based on selecting patterns located at a certain distance from the decision boundary.

Goal 7 **– Conducting an experimental evaluation of the proposed methods in comparison to *state-of-the-art* approaches.**
Goal 8 **– Developing a *Python* Machine Learning library for difficult data stream analysis.**
Goals 7 and 8 were achieved by designing an experimental environment for imbalanced data classification, as well as by creating the *stream-learn*[2] package for difficult data stream analysis, which was used to conduct all experiments related to data stream classification.

[2]Ksieniewicz, P. and Zyblewski, P., 2020. stream-learn–open-source Python library for difficult data stream batch analysis. arXiv preprint arXiv:2001.11077.

Keywords

Pattern recognition; inductive learning; classification; classifier ensemble; classifier selection; difficult data; imbalanced data; data stream; data preprocessing; concept drift; active learning.