Automatic correction of SVM for drifted data classification

Alexandra Degeest*,**, Benoît Frénay*, Michel Verleysen*

*Machine Learning Group, ICTEAM, Université Catholique de Louvain,
Place du Levant 3, 1348 Louvain-la-Neuve, Belgium
**ISIB, Haute-Ecole Paul-Henri Spaak
Rue Royale 150, 1000 Bruxelles, Belgium
alexandra.degeest@uclouvain.be, degeest@isib.be

Abstract. Concept drift is an important feature of real-world data streams that can make usual machine learning techniques rapidly become unsuitable. This paper addresses the problem of sudden concept drift in classification problems for which standard techniques may fail. To this end, support vector machines (SVMs) are automatically corrected to cope with a new suddenly drifted dataset. Results on real-world datasets with several types of sudden drift indicate that the method is able to correct the SVM in order to better classify the new data after the concept drift, using a correction based on the difference between the initial dataset and the new drifted dataset, even when the new dataset is small.

1 Introduction

Concept drift in real-world data streams is a standard concern. Concept drift happens when the distribution underlying the data changes over time. It can occur in several ways: sudden drift (abrupt modification of the data statistical distribution), gradual drift (period of time when two different distributions are active), incremental drift (progressive modification of the data statistical distribution) or reoccurring contexts (periodic reappearance of different statistical distributions) (Zliobaite, 2010). This paper deals with sudden drift. The reasons why sudden concept drift happens in datasets are numerous and diverse. As an example, in industrial applications, data can suddenly drift because of engine maintenance, because the tool has been relocated or because the new data have been collected by another person. The occurrence of concept drift in data streams often disturbs usual machine learning techniques and, therefore, brings the necessity to develop new learning techniques or to adapt existing ones.

Support vector machines (SVMs) are a powerful tool, commonly used for classification of high-dimensional data. However, SVMs may quickly become useless when data are subject to concept drift. The objective of this paper is to propose a method to automatically correct the SVM model facing sudden drift in a classification problem. The goal is twofold. First, the goal is to allow us using the first dataset (before the drift) for model learning, what is particularly useful when the second dataset (after the drift) is small. The second goal is to correct the decision hyperplane to better classify new datasets, even when their statistical distribution (slightly) differs from the distribution of the first dataset. Two assumptions have been made in this work. First, the labels of the drifted data are unknown. Second, the drifted dataset is small