

# Adaptive Concept Drift Detection

Anton Dries <sup>a</sup>

Ulrich Rückert <sup>b</sup>

<sup>a</sup> *Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium*

<sup>b</sup> *International Computer Science Institute, Berkeley, USA*

This is an extended abstract. The full version of this paper was presented at SDM 2009 [1].

## 1 Introduction

Concept drift is an important problem in the context of machine learning and data mining. It can be described as a change in the fundamental concepts underlying the data, or, in its most basic form, as a significant change in the distribution of the data. From a learning theoretic point of view, one can say that concept drift is a violation of the i.i.d. assumption, which states that each example in a dataset is drawn independently from an identical distribution. When concept drift occurs, the second part of this assumption no longer holds. This has important consequences because most of the learning theoretic performance guarantees used in machine learning are based on this assumption. This means that the performance of most learning algorithms becomes unreliable when concept drift occurs. Detecting when this happens is therefore of vital importance for many applications working in dynamic environments (e.g. data streams [4, 5]).

This *concept drift detection* problem is often addressed by statistical methods. More formally, the problem can be framed as follows: Given a sequence of training examples, are the last  $n_1$  examples sampled from a different distribution than the  $n_2$  preceding ones? Statistical decision theory has come up with a broad range of established methods that can be used for this purpose [2, 3]. These methods typically compute a statistic that catches the similarity between the two example sets. The value of the statistic is then compared to the expected value under the null hypothesis that both sets are sampled from the same distribution. The resulting *p-value* can be seen as a measure of to what extent concept drift has happened.

It must be noted, though, that it is impossible to come up with a universally best test statistic. This is because for every test statistic one can construct a pair of distributions, which differ from each other to some degree, but lead to the same distribution of the test statistic. The question on whether or not a particular test works well in a particular setting depends on the match of the applied test statistic with the underlying distribution. In the following we propose and evaluate three new methods, which adjust the test statistic depending on the actual data. This ensures that the test statistic captures the most important properties of the underlying distributions and adjusts itself well in a broad range of settings.

## 2 Adaptive approach

Let us frame the problem of concept drift detection and analysis more formally. We are given a continuous stream of examples  $x_1, x_2, \dots$ . Each example is an  $m$ -dimensional vector in some pre-defined vector space  $\mathcal{X} = \mathbb{R}^m$ . At every time point  $p$  we split the examples in a set  $\underline{X}$  of  $n$  recent examples and a set  $\overline{X}$  containing the  $\bar{n}$  examples that appeared prior to those in  $\underline{X}$ . We would now like to know whether or not the examples in  $\overline{X}$  were generated by the same distribution as the ones in  $\underline{X}$ . The traditional statistical approach would be to apply a statistic directly to the two samples, but this approach does not take into account specific properties of the data at hand and it requires a (computationally expensive) multi-variate statistic. In our approach we first apply a well-chosen transformation function to the two samples. This function serves two purposes: (1) reduce the dimensionality of the data to allow us to use (fast) univariate statistics, and (2) maximize the difference between the two samples if they are from different distributions. However, in choosing this function, we must take care of some limitations on the information we can use. For example, most statistics

for the two-sample problem only work under the assumption that the two samples are independent. Using a transformation function based on the samples themselves would clearly violate this assumption. In this paper we use two approaches to overcome this limitation. In the first approach we use an independent training set to determine a good transformation, allowing us to apply a standard univariate statistic on the transformed samples. In the second approach we use results from statistical learning theory to develop a test statistic that does allow us to use the two samples directly by restricting the class of transformation functions.

Based on these two approaches we develop three methods. The first one uses a binary CNF rule learner as a density estimator for the original concept. This algorithm learns a set of rules represented in Conjunctive Normal Form, which can be done incrementally based on examples from a single class (or, in this case, concept). By using an incremental algorithm we can apply it directly to the data stream and maximize the size of our training set.

The second method is based on results from statistical learning theory applied to linear support vector machines. In this approach we reformulate the problem of finding a transformation function that maximizes the difference between two samples as a problem of maximizing the margin of a one-norm support vector machine. Concretely, we assign class labels to the two samples and try to find a linear separation that maximizes the margin between the two samples. The intuition behind this is that it will be hard to distinguish the two samples if they come from the same distribution. By using statistical learning theory we can formalize this idea and formulate a bound on this margin, which can be used as a statistic for the two-sample problem.

A similar approach is followed in the third method, where we use the error of a regular (two-norm) SVM as basis of the transformation function. Again, we define bounds on the zero-one and sigmoid loss error, and use them as statistical tests for concept drift detection.

### 3 Results and Conclusions

To evaluate our methods we applied them to 27 datasets from the UCI repository in which we introduced concept drift by reordering the examples according to class label. We compared our methods with the standard statistical Wald-Wolfowitz test. These experiments show that, even for the relatively small sample size of (2x50) data points, these methods are able to detect concept drifts and are not too sensitive to noise in most cases. All of them are faster than the Wald-Wolfowitz test and remain applicable if the concept drift is more gradual in nature.

### References

- [1] Anton Dries and Ulrich Rückert. Adaptive concept drift detection. In *SDM*, pages 233–244. SIAM, 2009.
- [2] Jerome H. Friedman and Lawrence C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Annals of Statistics*, 7(4):697–717, 1979.
- [3] Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel method for the two-sample-problem. In *NIPS*, pages 513–520. MIT Press, 2006.
- [4] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *VLDB*, pages 180–191. Morgan Kaufmann, 2004.
- [5] Matthijs van Leeuwen and Arno Siebes. StreamKrimp: Detecting change in data streams. In *ECML/PKDD*, pages 672–687. Springer, 2008.