Review of Master Thesis

Title: Classification in data streams with abrupt concept drift in a subset of features
Student: Martin Procházka
Reviewer: doc. Mgr. Branislav Bošanský, Ph.D.

Summary:
Master thesis focuses on the acute problem of the concept drift that appears in many practical applications of machine learning models. The concept drift occurs when a single model is being used on data incoming in discrete time steps, and in which the joint probability distribution of input features and labels changes with the time. This can happen not only in mentioned use-case of malware detection but is apparent in many other use-cases where the model is used in an open world (e.g., changes in image preprocessing due to camera improvements, etc.). Understanding this phenomenon and developing methods that are robust to the concept drift is therefore one of the key challenges in machine learning.

The thesis starts with a detailed mathematical introduction to the problem of modeling the concept drift, and introduces the main concepts used later in the thesis (Chapter 1). Next chapter formally defines the problem and specifies which types of concept drift would be the main focus—sudden and severe concept drift—that abruptly causes a change in certain features of input data. Chapter 3 discusses the related work, Chapter 4 introduces the proposed solution methods the author compares using experimental evaluation using various datasets in Chapter 5. The solution methods are twofold – either based on (1) measuring the distance between two distributions or (2) comparing data distribution w.r.t. to the currently analyzed sample. These methods are then incorporated into the Dynamic Weighted Majority framework to improve the online prediction on data with concept drift.

Evaluation:
The topic of the thesis is quite challenging – while there are many works that focus on this problem, due to the inconsistent naming convention (in the literature, authors also use terms like ‘concept shift’, ‘dataset shift’, ‘covariance shift’, ‘out-of-distribution detection’, etc.) it can be challenging to find all relevant related work. Second challenge stems from a lack of good practical datasets that can be used for the experimental evaluation. This is partially substituted with synthetic datasets that confirm the expected results and where the proposed methods provide significant improvement compared to the baseline. However, the experimental evaluation on real-world datasets is less convincing (the improvement is almost non-existent) due to many common issues working with real-world data (the assumptions on the type of the concept drift do not need to be met, there can be noise in data, data are disbalanced, etc.)

Technically, the work is correct, the author provided many mathematical details that are related to the problem of online prediction on data with concept drift. The proposed methods are sound and the experimental evaluation seems correct.

The thesis is written in English. Occasionally, there are some typos and a few incorrect grammatical formulations. My main concern regarding the text is the style and the length of the thesis. The main part of the thesis has 83 pages and there are additional 20 pages in the appendices. The text itself contains many details that are, from my perspective, not necessary. Chapter 1 should have been, in my opinion, significantly reduced and focus only on the methods that are necessary. I do not think that deriving all the mathematical properties is a necessary part of the thesis. Next, the student uses summarizations of the main topic/point of almost all sections, subsections, and even paragraphs. While in general summarizing the main points is good, the way the thesis is written now, makes it very difficult to read sequentially since it contains many repetitive statements.

The student cites relevant works, I appreciate that he uses the references in a very detailed manner (referring to specific sections / parts of the cited works). While there are other works on the topic that were not mentioned, it can be difficult to find all mainly due to inconsistencies in namings among the machine learning community.

Overall evaluation:
In his work, the student demonstrated the ability to learn, understand, formally describe, and solve a complex technical problem. He proposed two methods for tackling the problem of concept drift and experimentally evaluated the impact of these methods compared to the baseline. Hence, he demonstrated the ability of a research work and I recommend this work to be accepted as the master thesis. Due to my objections to the structure and the formal side of the submitted text, I evaluate the thesis with grade **2.**

In Prague, 30. 8. 2024 Branislav Bošanský