CZECH TECHNICAL UNIVERSITY IN PRAGUE

THESIS SUPERVISOR'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Classification in data streams with abrupt concept drift in a subset of features**
Author's name:	**Martin Procházka**
Type of thesis :	master
Faculty/Institute:	Faculty of Mathematics and Physics, Charles University
Department:	Department of Algebra
Thesis reviewer:	doc. Mgr. Viliam Lisý, MSc., Ph.D.
Reviewer's department:	Department of Computer Science, FEL, CTU in Prague

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment *(How demanding was the assigned project?)* **Challenging**

The assignment of the thesis is challenging, since it requires that the student understands a new practical problem along with several existing algorithmic and mathematical concepts that I assume are out of the scope of the curriculum of his study program. Since the real world problem has many parameters and does not have an established formalization in the literature, it is challenging to properly formalize it in the right level of detail.

Activity and independence when creating final thesis **Excellent**

The student was very active and hard working during the whole time. He independently researched new literature and suggested solutions to encountered problems. His deep understanding of the necessary background and the amount of time spent working on the thesis is demonstrated in an exceptionally extensive and thorough background section, which is further extended in the appendix. The whole thesis, including the appendices, is 108 pages long.

Technical level **Very good**

The formalization of the problem is rigorous, but not very elegant, which is one of the reasons why Chapters 2-4 are harder to read.

The proposed solution is a novel combination of existing components for detecting concept drift and an existing classifier that accepts a subset of features. The student made some effort to formally derive the parameters and thresholds in the algorithm, but since they depend on the properties of the data, they were eventually largely chosen empirically.

The experimental validation is correct in principle, even though the discussion of the statistical significance of the results could have been more thorough. The experiments section is well structured and understandable. The thesis clearly formulates the hypothesis the experiments test, exhaustively reports the experiment settings and parameters of the algorithms.

Formal level and language level **Very good**

The clarity of the writing itself is good, but could be improved. For example, the data preprocessing in the experiment with real world data is very important for the interpretation of the results and its explanation is not very clear. The summary of the formalization in Section 2.3. and solution outline in Section 3.3 are important, but they were a little confusing and unfocussed. In general, the thesis analyzes the background material more thoroughly and extensively than the novel aspects of the thesis. I appreciate that the student wrote the thesis in English, but I have to note that the formulations are not always natural and I have found multiple typos.

Selection of sources, citation correctness **Excellent**

The thesis cites 58 references, many of which the student studied to great details. All the reused material is properly acknowledged. The list of references is consistent and follows the standards in the field.

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS

The assignment of the thesis is entirely fulfilled. The student formalized the problem, reviewed the related work, suggested a novel solution method and performed thorough evaluation of the solution on both synthetic and real world data. The student has demonstrated his ability to find and understand relevant scientific literature out of the scope of his studies, apply it to a new problem and achieve novel scientific results. On the other hand, the thesis might have been clearer and more focused on the novel contribution rather than the background material. I believe that the submitted thesis fulfills the requirements of a master thesis.

Suggested question for the defense:

In the experiment evaluated in Section 5.4.1., one of the requirements tested is that "the 95% confidence interval of false detections contains zero". What are the possible disadvantages of this test and how could it be improved?

Date: **29.8.2024**

Signature: