Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Nikita Klyuchnikov

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Multi-fidelity classification and active search

Supervisor: Associate Professor Evgeny Burnaev

Name of the Reviewer: Andrzej Cichocki

I confirm the absence of any conflict of interest	**Signature:**  **Date: 07-01-2021**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications

The summary of issues to be addressed before/during the thesis defense

The thesis is devoted to the important Machine Learning, Surrogate Modeling, and Bayesian Optimization problems related to Mutli-fidelity Classification and Multi-fidelity Active Search, when available multiple data sources have various noise in labels. The author of the thesis Nikita Klyuchnikov investigates two important sub-problems: Classification and active search, in real-life practical special cases when available data sources have both high and low fidelity, i.e., one source is high-fidelity (gives precise class labels), while another source is low-fidelity (some labels are imprecise and even can be wrong). In Multi-fidelity Active Search the objective was to retrieve as much relevant objects as possible within a limited budget on evaluations (assuming that we do not know in advance whether an object is relevant or not, until we retrieve it).

In order to find optimized solutions for these problems Nikita considers models based on Gaussian Processes and a co-kriging scheme for fusing sources. Such models have many potential applications such as law cases, patents search, and medical records retrieval. Moreover, several real-life industrial applications were also investigated that were connected to industrial projects related to design of a muon shield for the SHiP experiment at CERN and optimization of directional drilling of oil-wells, where proposed methods helped to improve quality of datasets annotation.

The proposed approach have some constraints and limitations: Particularly, inference with Gaussian Processes is computationally quite intensive and complex, so they are only applicable when sample size is not larger than, say, several thousand points. Furthermore, models based on Gaussian Processes also suffer from the curse of dimensionality, although in recent years, these challenges have been partially addressed in several research papers.

For the multi-fidelity classification problem, the author considered a specific data fusion scheme that draws a parallel to a widely adopted method of co-kriging regression used in the engineering design. However, in case of classification, the inference is not analytically tractable, whereas a standard computational technique based on MCMC is not effective. This motivate author to develop a new algorithm for efficiently approximate inference for the model. In addition, the author the thesis demonstrated that proposed model is often more robust to noise for low-fidelity data compared to existing alternatives.
Moreover, He has demonstrated experimentally, that in case of high correlation between data sources, the proposed model outperformed state of the art (SOTA) approaches.
Algorithms developed by Nikita for both problems were additionally tested for sensitivity to various hyper-parameters and properties of data.

The developed algorithms were verified and validated against baselines and compared with SOTA methods for a number of numeric experiments for real-life and synthetic datasets. For example: Penn Machine Learning benchmarks and popular crowdsourcing annotation datasets for music genre and sentiments classification were used for testing multi-fidelity classification models; Yahoo music, Yelp challenge and ACL knowledge graphs were used to test multi-fidelity active search algorithm. Moreover, their usefulness was demonstrated in real industrial applications described in Chapter 4.

Source codes of developed algorithms implemented in Python (multi-fidelity classification and active search) were published on github.

In my opinion, the most original new results are related the latent co-kriging Gaussian Process schema along with specially tailored multi-fidelity Laplace approximation for this problem (Chapter 2) and a multi-fidelity active search framework presented in Chapter 3.

All the results of the thesis were already published in scientific journals or conferences. Nikita is co-author of 7 publications, where 5 of them were published in high impact factor journals and two of them in international conferences while he is leading in 3 of them.

Summarizing the most significant contribution of the PhD thesis are as follows:

- The main results are presented in Chapter 2 and Chapter 3. In Chapter 2 was presented a new model with a computationally efficient method of Bayesian inference based on Laplace approximation method, while in Chapter 3, was investigated a new method for active search with modeling the relevance of objects based on data of varying fidelity using Gaussian processes.

- In both chapters, the developed algorithms are analyzed for sensitivity to parameters and characteristics of data in a number of computational experiments, as well as their performance and quality were compared with baselines and SOTA algorithms

Chapter 4 demonstrates the practical applications of methodology developed in the thesis to real-life industrial problems, especially the methods improved speed of manual dataset annotation that allowed substantially reducing noise in manually annotated labels.

Provisional Recommendation

☒ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report*

☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*