

Dr. hab. Mirosław Chorazewski
Director of Institute of Chemistry
University of Silesia

Dr. Agnieszka Smolinska
Department of Pharmacology and
Toxicology
NUTRIM School for Nutrition,
Toxicology and Metabolism
Faculty of Health Medicine and Life
Sciences
Maastricht University
PO Box 616
6200 MD Maastricht
The Netherlands

e-mail:
a.smolinska@maastrichtuniversity.nl

Maastricht 15.09.2022

Dear Dr. hab. Chorazewski

Please find enclosed the evaluation of the thesis entitled Optimization, validation and applicability of one-class classification methods by Zuzanna Malyjurek covers a very relevant topics within chemometrics, data science and machine learning field. The standard methods based on classification via partial least square discriminant analysis or discriminant analysis are powerful, however in certain situation their applicability is very limited and thus built suboptimal models.

One-class modeling based on Soft Independent Modeling of Class Analogy (SIMCA) is a popular method in the field of food authenticity. Therefore, SIMCA is very often used for that types of problems within but also outside chemometrics field. This is clearly indicated by the candidate in the introduction to the thesis. Many products, such as olive oil, wine or cheese are characterized by their distinctiveness and this is essentially linked to their geographical origin. For that purpose SIMCA model is used to find so called "real" and possibly fake products. As presented by the candidate within SIMCA building a unique mathematical model using just the training set data from real samples is the goal of class modeling. In the next step the test set is used to validate the created class-model and comprises of authentic and, if available, non-authentic samples. In order to determine if additional samples of unknown origin belong to the legitimate class under investigation, the

final model is utilized. A sample that has not been classified to an official class may be of poor quality, a fake product sample, or a sample from an unidentified class. All those aspects are well described in the thesis of the candidate and several examples are given. Although, it might sound very straightforward and one-class modeling, such as SIMCA can be easily implemented, there are several aspects related to proper optimization of the built model and selection of the approaches, which can directly influence the outcomes of the results and of course the conclusions.

The current thesis fits perfectly in the scope of the needs within one class-modelling. The class modeling techniques are tools that, although more flexible, than discriminant approaches (used so often in metabolomics field), and they are more suited to deal with disproportionate, asymmetric or imbalanced problems, they are often overlooked or as indicated above, used in suboptimal way. Since there are no proper optimization or even validation approaches presented in various publications and in many cases SIMCA and others one-class modeling techniques are used as a tool without emphasis on the technique and description of the methodology. Therefore, the current thesis, focusing on class-modeling tools, has the value of comprehensively investigating the impact of various aspects of class-modeling on classification performance. This includes and is not limited to choice of the acceptance threshold and whether to include samples from other categories or not during the model selection phase, which strategy to use with respect to type of the data (i.e. presence of one authentic class or representativeness of the classes), on the final classification performances. The candidate gives in each of the published paper a comprehensive procedure, description and several examples of the used data. Since such comprehensive evaluations of one-class modeling has never been properly investigated and no specific recommendations have not been given thus far, in my opinion the thesis is novel and contributes significantly to the field. The thesis provides better understanding of the one-class modeling and thus makes it easier to extend the applicability of the proposed methodology to other fields (e.g. omics, personalized medicine, early diseases-detection etc.). The candidate published five research papers, all as first author in high impact


journals. The first three published papers are combination of various theoretical problems for one-class modeling and last two papers demonstrate great application examples within food field. The thesis is full of rich results and each chapter contains an extensive experimental and computational support, allowing the reader to dive into the methodology and applicability of the presented methods beyond the presented examples.

Although, the content of the thesis is remarkable and the performed scientific work by the candidate is very impressive (several published papers, many international talks and posters which were valued by receiving several awards) I have some questions that I would like to have them clarified during the discussion with the candidate.

1. One aspect that I would like the candidate to elaborate on is the aspects of the heterogeneity and class distribution. More specifically, if we talk about group(s) that are heterogeneous, subgroups are present (e.g. different phenotype of the same diseases) or when there is mislabeling (e.g. in the medical data). In the chapter 3 (paper 2) the candidate indicate that it is not expected that that OCPLS or SIMCA to lead to high classification results. However, if I look at the results shown, I see good performance of the SIMCA. How can the candidate explain that?
2. The candidate compared the proposed methodology to various class-modeling approaches (e.g. SVDD, OCPLS and PFM) but tree-based technique, called isolation forest was never considered. Can the candidate explain the reasons for selecting the abovementioned methods?
3. In the paper „Different strategies of class model optimization” for the analysis of results obtained in the comparative study, the first type ANOVA model was applied. The ANOVA model consisted of with two fixed factors and one fixed blocking factor. Why the blocking factor (datasets) is fixed?
4. From the user point of view, which of methods that the candidate worked with is the easiest or the most convenient to use as a starting point for the class-modelling of data with multimodal structure and why?

5. In the paper „class-modelling of overlapping classes”. A two-step authentication approach” the candidate mentioned that the difference between Cyclopia species studied are subtle and local. Why for authentication the candidate used entire spectra instead of using only the features that differentiate the species studied? It looks like the candidate include a lot of irrelevant information.
6. Why % of variance are missing on PCA score plots in the last two chapters? Are they not informative?
7. The thesis ends with a short conclusion, which is perfectly fine but what I miss is the future steps in the field of one class-modeling. Can the candidate speculate here about that aspect?
8. Can the candidate please explain the RF model used in the chapters (papers) 4 and 5 and its optimization? There are no clear visualization of the model, but only in the graphical abstract I can see a figure which I assume, resembles PCoA. What was used here to generate the figure?

Yours sincerely,

A handwritten signature in black ink, appearing to read 'Agnieszka Smolinska'.

Dr. Agnieszka Smolinska