Leveraging Machine Learning to Enhance Aerosol Classification using Single-Particle Mass Spectrometry

https://doi.org/10.5194/egusphere-2025-3616

General comments

The work submitted for publication reports an interesting study on ways to improve the accuracy of classifying aerosol particles - which were ionized and analyzed by Single-Particle Mass Spectrometry - by means of Machine Learning. The proposed semi-supervised learning, in which unlabeled data is used for learning, is undoubtedly of great importance in practical applications.

To date, only a few approaches to semi-supervised learning (even beyond SPMS) are known and have been cited in the paper. The efforts undertaken in this study are very welcomed and promising. The chosen approach can be considered largely novel.

Nevertheless, a tailored implementation with convincing results and 'design guidelines' to achieve the best results would be of considerable significance for many applications.

The text is well written and very informative, with only little but disturbing redundancies. E.g. Table 2 and Fig. 5 as well as Table 3 and Fig. 7 bear the exact same information. It is recommended to omit Figs. 5 and 7.

Major issues

(1) The study's aim is to propose sophisticated Machine Learning models capable of bringing the classification performance closer to the optimum of 100 %. The obtained accuracies for the four described algorithms are surprisingly similar to each other (90.0% to 91.1%), with a significant gap to the optimum. This means, looking at the dataset as a whole, almost 10% of the assignments are incorrect. It is worth discussing how these incorrect assignments (false negatives and false positives) would be handled in practical applications.

From the results one might draw the conclusion, that systematic weaknesses common to the different approaches prevent better results from being achieved. The authors speculate on some of the causes (imbalanced dataset, number of classes, similarities between spectral features), but the dependence on these factors is not investigated.

(2) It is suggested to take a closer look to one of the most prominent difficulties for Machine Learning models which is a heterogeneous, limited, imbalanced training dataset.

(a) The dataset chosen by the Authors is very heterogeneous. It contains mass spectra of aerosol particles from very different emission sources, collected in various measurement campaigns. Part of the dataset (it remains unclear, what proportion) was used in a historical reference (Christopoulos et al., 2018).

(b) The dataset is comparatively small (less than 20,000 labeled spectra), nevertheless comprising samples of as much as 20 (!) different classes of aerosol particles. Hence, on average, there are less than 1,000 labeled samples per class in the dataset. The test is performed on 10 % of the dataset, which for the under-represented classes (soot, pollen, agar) leaves less than 20 labeled test samples.

(c) The class sizes vary greatly, from 21% to 0.8% of the total number of spectra. Such strong class imbalance is a well-known obstacle for high-performance ML applications. Methods to balance the class sizes via data augmentation are mentioned and cited in the text, but were not applied. Moreover, the

greatest advantage of semi-supervised learning and probably its core motivation is that the training dataset can be balanced and enlarged with almost no effort by adding unlabeled data to it. To exploit this advantage was apparently not considered by the Authors.

3) In the Introduction, the Authors criticize the common practice of assigning all samples in the dataset to a fixed number of predefined classes, without the option to classify certain samples as 'unknown'. In the presented implementation, however, such class comprising all samples of 'uncertain' or 'unknown' origin is still missing. The authors apparently quietly assume that all unlabeled mass spectra can be assigned to one of the 20 defined classes.

4) To improve the significance and practical applicability of the presented novel promising self-training and autoencoder classifiers, is it recommended to demonstrate their potential by a step-wise approach, starting from a sufficiently large, homogeneous, balanced dataset with only a few classes, to achieve a classification accuracy close to 100%. Then, step-by-step the dataset can be made more 'complicated' in various ways (increasing the share of unlabeled data in the first place), to draw implications for the usability of the sophisticated classifiers for various applications. Certainly, only few applications will need to classify unknown mass spectra into 20 very different classes like feldspar and agar.

Minor issues

1) In Lines 142-145, the dataset is defined as being composed of data collected during a FIN Workshop (reference from 2024) and data already used by (Christopoulos et al., 2018). In Lines 311-314, it is stated that the achieved overall accuracies of 90-91% surpass the 87% accuracy previously reported using the (Christopoulos et al., 2018) dataset. Apparently, the results are only comparable when the same data were used, hence when the data from the FIN Workshop were excluded. Was this the case?

2) Line 209ff and Figure 1: Are the numbers correct? In Line 158 it was given that 14,487 unlabeled mass spectra were included in the dataset. As can be read from Figure 1, for confidence threshold 0.95, about 12,000 labeled spectra were used (why not all?). How does this match to the 25 % of the unlabeled spectra incorporated in the training set? Does it mean that the fraction of unlabeled data in the training set is fixed and roughly 30 % (~4,000/~12,000)? Was the same dataset or the same proportion of unlabeled to labeled samples used for all algorithms?

3) Lines 231ff. How the parameter values (e.g. 96 latent dimensions, 48 features) were found?

4) Line 295: "Reconstruction quality shows no correlation with classification accuracy". That's a bold claim! It might be true only for the specific dataset.

5) Line 411: "Our analysis revealed patterns in how different model architectures approach classification." This sentence is difficult to understand. What are its consequences?

6) The subscript to Fig. 8 is cryptic. It should be explained, that for every of the 4 Feldspar species up to 4 score points can be gained for 4 models.

Conclusion

The work presents a valuable but rare approach to classify a mix of labeled and unlabeled data based on semi-supervised Machine Learning. Four algorithms and their classification results are presented in greater detail. For better understanding, the manuscript needs some clarifications and corrections. Recommendations are given how to re-design the study in order to improve the practical value and significance of the results.