TERM PROJECT

Konstantinos Mexis, Xenios Stefanos

Paper full citation:
Single-Cell Classification Using Mass Spectrometry through Interpretable Machine Learning, Yuxuan Richard Xie, Daniel C. Castro, Sara E. Bell, Stanislav S. Rubakhin, and Jonathan V. Sweedler, *Analytical Chemistry* **2020** *92* (13), 9338-9347, DOI: 10.1021/acs.analchem.0c01660

Link to text: https://pubs.acs.org/doi/10.1021/acs.analchem.0c01660
Link to data: https://drive.google.com/drive/folders/1a__5nrhcNX8ePqD31NZlsYlWKyIZaT8W

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
School of Chemical Engineering

1. **What is the problem studied in the paper you selected?**

The problem addressed in the selected paper is understanding the relationships between chemical heterogeneity, cell function and phenotype in the brain. The authors propose a machine learning workflow that classifies single cells based on their mass spectra, specifically distinguishing between different cell groups of interest such as neurons and astrocytes. In this work mass spectra data from three different mass spectrometer platforms were utilized to classify cell groups or different areas of cell origin. The trained models achieved over 80% classification accuracy, and the authors used a recently developed instance-based model interpretation framework called SHapley Additive exPlanations (SHAP) to assign feature importance for each single-cell spectrum. The challenge of the problem was to build high accuracy classifiers for different cell types based on mass spectra data and extract information using explainability techniques.

2. **Why it is important? What is the state of the art in the topic?**

Single-cell chemical analysis has emerged as a powerful tool for identifying unique cellular markers and detecting rare biological events that contribute to cellular abnormalities. In particular, single-cell RNA sequencing (scRNA-seq) has become the preferred method for studying cellular heterogeneity and its impact on the structure and function of complex multicellular structures like the nervous system. Recent advancements in matrix-assisted laser desorption/ionization (MALDI) mass spectrometry (MS) have enabled the high-throughput analysis of single cells and the detection of small metabolites like neurotransmitters. This technology can be coupled with other techniques, such as immunocytochemistry (ICC), to screen thousands of cells labeled by their canonical cell types, producing comprehensive 'omics' data on the peptidome, lipidome, and metabolome. While not yet commonly used, there are many exploratory computational tools that are available for scRNA-seq, which can be adapted to MALDI MS measurements to determine the chemical differences between single cells and cell types.

A systematic approach of classifying areas of cell origins or general cell type is very important in technologies such as mass spectroscopy which produce plenty and noisy data. Also, different mass spectrometer platforms use different means of operation and may lead to different magnitudes of results.

The state of the art that is proposed in this paper is to build classification algorithms on two different datasets for two different labels (cell type/ area of cell origin) and extract information using interpretable machine learning. Lastly a different dataset is used on which the extracted information in the form of rules is applied to label the unseen data.

3. **What are the data analysis methods and tools used and the main results claimed by the authors?**

Single-cell classification is a supervised machine learning task that learns a function and maps the input data X (single-cell spectra) to the categorical output variable Y (cell GOI). The authors state that since the single-cell mass spectrometry (MS) classification task is driven by the proper interpretations of the classification results in order to find potential biological implications, nonlinear models such as DNN and kernel SVM were not used because of their limited interpretability. For that reason, ensemble tree classifiers (specifically XGBoost) were trained to predict cell GOI. The authors stated that since Gini importance may provide inconsistent and biased feature importance measures, SHAP framework, based on the additive feature attribution methods, which locally assigns consistent feature importance for each individual prediction of a single cell, is the state-of-the-art technique for measuring feature importance. Computing the mean absolute SHAP values allowed the researchers to select features that were important to the classification tasks, enabling global interpretations. Furthermore, PCA was applied on the SHAP values to demonstrate that the models learned group-specific knowledge and captured heterogeneous feature contributions among the single cells within a specific GOI. The main result of this work is the usage of SHAP values in order to better understand the relationships between chemical heterogeneity, cell function, and phenotype.

4. **What is your motivation for your additional analysis?**

This paper offers good, preprocessed datasets on which different classification techniques could be applied to evaluate their efficiency. The accuracy on the dataset for cell type classification is deemed to be quite low (77%) and is seen as an opportunity to increase the accuracy using more advanced pipelines and techniques. In addition, different rule extraction algorithms could be tested, and the results can be compared to the original paper that uses SHAP values. Our working hypothesis is that we could build a better pipeline for the classification task, using more sophisticated machine learning models and techniques. It is very important to have an accurate classifier in order to extract feature importances using SHAP values. First, we will try to reproduce the results from the paper to have a baseline to build new approaches and compare our results. The first problem we will try to tackle is the low accuracy score (77%) on the dataset for cell type classification. We will try to implement different classification algorithms and pipelines and try to evaluate their performances. The goal will be to achieve higher accuracy on the specific dataset. Additionally, we will try to extend the work of this paper by implementing different interpretability techniques and we will compare them with the results from the paper.

5. **Your technical approach: What ML methods you plan to use and why? What programming language(s) and libraries /packages you plan to use (should be R and/or Python).**

The machine learning methods that will be used in this project are;

- *Artificial Neural Networks*: Experiment with ANN classifiers and compare their accuracy with machine learning methods, such as boosting trees (used in the original paper).
- *Partial Dependence Plots (PDPs):* PDPs are a visualization technique that shows the effect of a single input feature on the output variable, while holding all other features constant. PDPs are highly interpretable and can be used to identify non-linear relationships between input features and the output variable.
- *Accumulated Local Effects (ALEs):* ALEs is a visualization technique that is used to interpret the impact of a single input variable on a predictive model's output. It is often used in the field of machine learning, especially for understanding complex models like random forests, gradient boosting, etc.
- *LIME*: LIME (Local Interpretable Model-Agnostic Explanations) is a technique for explaining the predictions of any black-box model by approximating the model locally around the prediction using a simpler interpretable model.
- *Generalized Additive Models (GAMs):* GAMs are a class of models that represent a non-linear relationship between the input features and the output variable using a sum of smooth functions. These models are highly interpretable and can be visualized easily.

- *Dimensionality reduction methods* for data visualization: Experiment with more dimensionality reduction methods, such as t-SNE, UMAP and Autoencoders.

We are planning to implement the above methods mainly using Python libraries: *Pandas, NumPy, Keras and/or PyTorch, scikit-learn, Matplotlib and/or Seaborn, SHAP, LIME, ELI5, InterpretML, OmniXAI*

6. **Your implementation plan: Outline the analysis tasks and the team member roles in each task. Include a mitigation plan if things go wrong.**

This project will be divided into three different tasks, one mandatory and two optional.

- The mandatory task will be to reproduce the results of the paper and offer different data visualization techniques. The workload will be divided as follows:
- Mexis Kostas will be responsible for building the pipeline for the classification algorithm.
- Stefanos Xenios will be responsible for calculating the SHAP values.
- Both students will help with the data visualization part
- The first optional task is assigned to Kostas Mexis who will try to implement different classification algorithms/pipelines to increase the accuracy in the second dataset used in the paper. Useful charts and graphs will be provided.
- The second option task is assigned to Stefanos Xenios, who will try to extract rules and information from the classification algorithms using different interpretable machine learning techniques. Useful charts and graphs will be provided.
- The above split of the project is also the mitigation plan for this project. More specifically, if either optional task proves to be either difficult or fruitless that member's attention will be turned to the other optional task. If both optional tasks are difficult then the only deliverable for this project will be the reproduction of the paper's results.