Machine Learning for Science: Classification of Skin Samples Using Mass Spectrometry Analysis

Alice Juanico, Dorian Laforest, Laura Testa

CS-433 Machine Learning
Supervisor: Yingdi Zhu - EPFL SB ISIC LEPA
December 2020

Abstract. In this report, we describe how we used simple machine learning algorithms to classify skin samples into 3 conditions (dehydrated skin, skin mole and normal skin), using mass spectrometry results. By using SVM algorithm, we were able to correctly classify all the skin samples. With statistical analysis, we identified the most relevant mass spectral peaks for the classification task.

Introduction

Mass spectometry is a well-known analytical technique based on the mass-to-charge (m/z) ratio in ions. This tool is widely used in various research fields to analyze chemical or biological components. Thanks to the preciseness of this technique, one can easily identify specific proteins or biomarkers present in an analyzed sample [1]. In the present case, mass spectometry was used to collect information about various skin samples coming from different body regions of several volunteers. The resulting data set is the one we are working with in this report.

During this project, a classification task was implemented to differentiate the skin samples according to three conditions: normal skin, dehydrated skin and skin mole. The classification was based on the mass fingerprints of each skin sample. This method of identification is relevant as healthy skin composition is roughly the same across all individuals [2]. Dehydration and moles are characterized by changes in skin cell metabolites; skin presenting those condition will thus have a different fingerprint than normal skin and can be identified using ML algorithms. A statistical analysis was also conducted to determine the most important mass spectrum peaks for identification of the skin conditions.

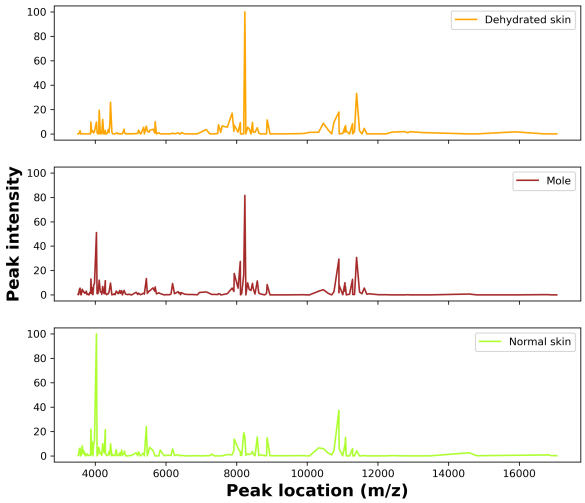

Figure 1: Mean mass spectrum of each skin category.

Models and Methods

Data preprocessing and visualization

The data set is composed of the mass spectra generated from all the investigated skin samples coming from 9 volunteers. There are 66 skin samples, comprising 30 normal healthy skins, 12 dehydrated skins and 24 skin moles. For each sample, the mass fingerprint is described in the form of a peak list, with the normalized peak intensity for each labeled peak location (m/z). For matters of simplicity, all the mass spectra were put together, forming a sparse matrix of 66x179 samples (66 being the number of skin samples and 179 all the peak locations of the mass spectrometry experiment). The skin samples are thus described by the peak intensity for the specific peak location, intensity set to 0 if the sample showed no signal for the particular peak.

From this peak intensity data set, a similar binary sparse matrix was formed to summarize the peak presence for each skin sample, "1" representing peak presence and "0" peak absence. Classification and analysis were performed on both peak intensity data set and peak presence data set. Note that throughout this report peak intensity data set will be referred to as "PI" and peak presence data set as "PP". Moreover, different standardization methods on PI data set were tested: MinMaxScaler (normalization), StandardScaler (standardization) and finally no standardization at all.

Before beginning any further analysis on the data, we visualized the mass spectrum of the skin samples. Therefore we took the mean of the peak intensity across all skin sam-

ples belonging to the same skin category: normal, moles and dehydrated. We then visualized independently the mean mass spectrum of each category and observed the peaks that best characterized each skin category (**Fig.1**).

Finally, we performed the PCA dimension reduction technique on the PI data set projecting it into a 2 dimensional space using 2 components (**Fig.2**).

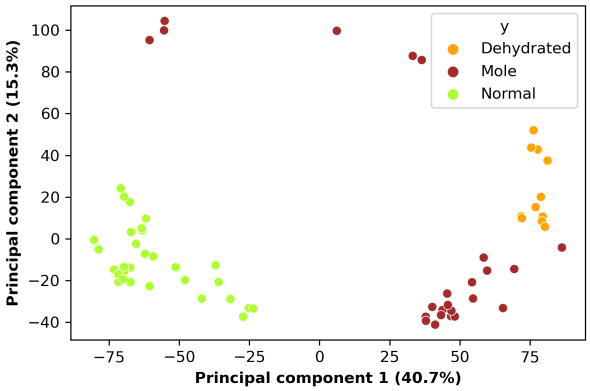

Figure 2: Principal component analysis with 2 components.

According to the peak intensity distribution that can be seen in **Fig.1** and the 3 distinct clusters formed in the 2 components analysis in **Fig.2**, we concluded that a simple multi-classification algorithm should be sufficient and should perform well on the data.

Classification model

Firstly, the project aimed to classify the three skin conditions - moles, dehydrated and normal skin - according to the mass spectra of each skin sample. Therefore, we applied a multi-classification model to our data set. To do so, the peaks intensity data set was used, but the results were then compared with a classification based on the peaks presence only. Both classifications performances were evaluated and compared using the following measurers: accuracy score, f-score, kappa score and AUC.

As previously discussed, basic classification algorithms were implemented for this task. All algorithms used are taken form the scikit library [3]. The main algorithm we used was the standard Support vector machine (SVM) model. This classification was then compared with other popular machine learning algorithms such as Gaussian naive Bayes Classifier, Logistic regression and K-nearest neighbors with the ball tree algorithm using a K value of 2.

Because of its small size, the whole data set was used to train each classification model and accuracy of the model was tested in the leave-one-out cross validation, meaning that 66-fold CV was performed.

Statistical analysis

Secondly, we performed a statistical analysis on both PI and PP data sets to identify the most relevant peaks for the classification. To do so, an univariate feature selection was run on the data. Using chi-squared statistical test, the features (peak locations) with the highest scores were determined, and a new data set was formed containing only these features. The SVM model was then fitted on this newly formed data set. Using the classification accuracy score, we could identify the minimum number features with the highest scores that are needed to correctly classify all the skin samples (accuracy of 1.0). These features are thus the most relevant ones for this classification task. We proceeded in the same way for both PI and PP.

Finally, we searched for the most relevant features for each individual category. To do so, we implemented a binary classification ("one-Vs-all"): we isolated one category by putting the two other categories with a same label (i.e "Dehydrated" labelled -1 Vs. "Normal" and "Mole" labelled 1), and performed univariate features selection on this new PI data set. That way, we obtained for each category the features that best distinguish it from the two others.

Results

Classification

We used a leave-one-out cross validation grid search to find the best hyper-parameter for all tested models. Those were SVM, Naive Bayes Classifier, Logistic regression and K-nearest neighbor. All were tested on both the PI and PP data set. Every classifier yield an accuracy of 1.0 (notebook *project2.ipynb*). For all the following figures, only SVM results are displayed, but the same can be found in notebook *project2_graphs_and_stats.ipynb* for the other models. For SVM the hyper-parameter to find were the regularization parameter C and the preprocessing step (no standardization, standardization, normalization). The mean-accuracy scores obtained after classification of the intensity data set and the presence data set are reported in **Fig.3** and **Fig.4** respectively.

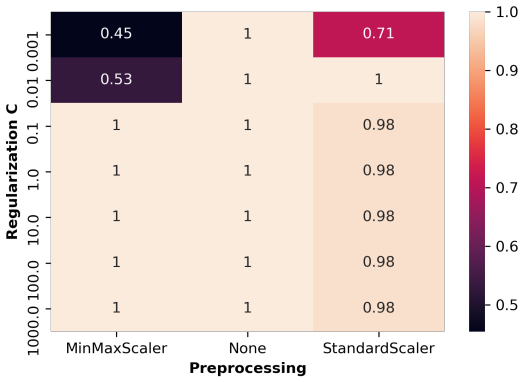

Figure 3: Mean accuracy on peak intensity data set obtained with LOOCV

As we can see, for both data set, we can use no preprocessing and a regularization parameter C of 0.1 to obtain an accuracy of 1.0.

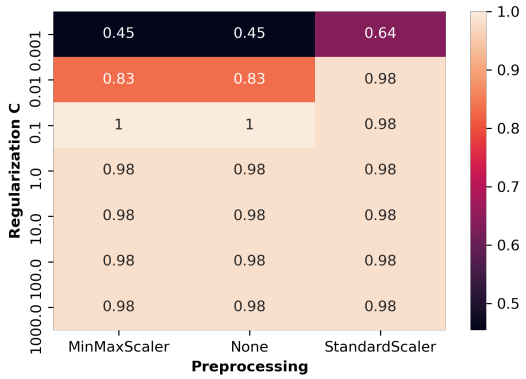

Figure 4: Mean accuracy on peak presence data set obtained with LOOCV

The accuracy score, f-score, kappa score and AUC were all around 1.0 for every classification method implemented. Since this does not provide any useful information about the model selection, the full description of these measures was not included in the report, but it can be found in the notebook *project2_graphs_and_stats.ipynb*.

Statistical analysis

Using univariate selection, features of both PP and PI data set were classified from the highest chi-squared score to the lowest. The accuracy of the SVM model according to the number of best features selected for both data sets is shown in **Fig.5**.

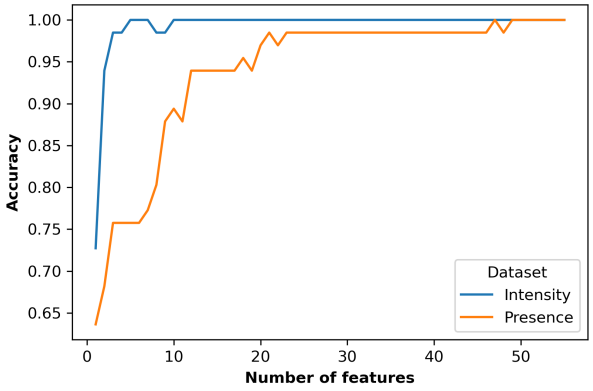

Figure 5: Accuracy on both data set depending of the number of best features selected

On the PI data set we only need 5 features to achieve an accuracy of 1.0. Those features can be seen in listed in **Tab.1**, with their respective chi-squared score. This accuracy was obtained with SVM, a C parameter of 1 and normalization of the data as a preprocessing step.

Concerning PP data set, 49 features were needed to successfully classify all the skin samples with an accuracy of 1.0. Only the 5 features with the highest scores are listed in **Tab.2**. Considering only those 5 features, we obtained an accuracy of 0.758 when using SVM on it, with a C parameter of 0.1 and standardization of the data as a preprocessing step.

Peak label	Peak location (m/z)	Score
PEP32X	8235.289	1600.88
PEP147X	4035.3176	1186.67
PEP121X	11397.069	646.04
PEP30X	8106.5234	613.80
PEP6X	7875.197	407.34

Table 1: Most important peak location based on peaks intensity, classified according to the highest chi-squared score.

Peak label	Peak location (m/z)	Score
PEP9X	7592.109	48.13
PEP8X	7489.842	36.25
PEP113X	5697.2974	27.32
PEP19X	12751.924	22.50
PEP24X	4472.014	22.50

Table 2: 5 first most important peaks location based on peaks presence, classified according to highest chi-squared score.

Results of the "one-vs-all" feature selection are shown in **Tab.3**. The most relevant features to distinguish the category from the two others are listed with their chi-squared score.

Category	Best peaks selection		
	Label	Location	Score
Dehydrated Vs. Mole + Normal	PEP147X	4035.3176 m/z	701.85
	PEP32X	8235.289 m/z	542.89
	PEP6X	7875.197 m/z	353.13
Mole Vs. Dehydrated + Normal	PEP30X	8106.5234 m/z	589.46
	PEP32X	8235.289 m/z	487.91
	PEP139X	10898.25 m/z	295.26
Normal Vs. Dehydrated + Mole	PEP32X	8235.289 m/z	1551.39
	PEP147X	4035.3176 m/z	978.19
	PEP121X	11397.069 m/z	643.42

Table 3: 3 first most important peaks location with One-vs-all category feature selection on peak intensity data set.

Discussion

Classification

According to the results, the previously stated hypothesis that a very good accuracy could be obtained using a simple model is clearly verified. Even with the PP data set - which loses a lot of information since a peak with a very small intensity is considered the same as a peak with a high intensity - we obtained the same accuracy as for PI. As seen in the mass spectra of each category and with PCA, the three classes are easily differentiable. Furthermore, even without data preprocessing, the accuracy of the model remains very high (1.0 for PI, 0.98 to 1.0 for PP, see **Fig.3** and **Fig.4**). Some preprocessing methods even lower the score (i.e Standard Scaler). This is however not surprising

as we do not analyze here raw data directly extracted from the mass spectroscopy recording: the data that we were provided with was already preprocessed and normalized.

SVM was our first choice of classifier, as the data set contains very few data: according to the documentation of the library used to conduct this project, SVM is the optimal method for classification on small data set [4]. To validate this choice, comparison was made with other algorithms. As all the measured statistics were 1.0 for all methods (c.f *project2_graphs_and_stats.ipynb*), we kept SVM as our final model.

The fact that we have a very small data set can also be a double-edged factor: on one hand the task is rapidly and easily done, but on the other hand the risk of overfitting the data set is really high. In order to improve the test accuracy of our model we could either use additional data or reduce the feature space with feature selection task. An idea was to try to find similar data sets published by other researchers, and to use them to test our accuracy. However mass spectometry analysis of skin samples is not very common, and we found no concluding results. We therefore chose to go on with the latter option, feature selection. This process was combined with the statistical analysis of the data, as it basically consists in selection of the most important features. For the following part, only SVM models were tested.

Statistical analysis

From the one-Vs-all feature selection, the significant peaks for each category were selected (**Tab.3**). For a complete analysis, one should compare the obtained results to the mean mass spectra of **Fig.1**.
Dehydrated skin differentiates from the other samples mainly through PEP147X value (4035 m/z), with a high chi-squared score of 702. Indeed, looking at **Fig.1**, it is quite clear that this peak intensity is way lower in dehydrated skin than in other samples. The low value of PEP147X could therefore be considered as a main characteristic for dehydrated skin identification. Dehydrated skin also present a very high PEP32X (8235 m/z). However this peak is also present in mole mass spectrum, so this peak alone cannot be considered as characteristic of this skin class.

The same analytic logic can be used to determine normal skin spectrum characteristics: for this category, main identification features could be resumed in high PEP147X value but low PEP32X. As for mole skin, the selected features present lower chi-squared scores than for the other classes. The spectrum for this category indeed contains characteristics of both dehydrated and normal skin: this category cannot be identified by the presence of a single peak only. Considering the previous description of normal and dehydrated skin, we could conclude that mole skin characteristic would be the presence of high intensity peaks for both PEP147X and PEP32X.

From all of this analysis, we can deduce that to build a model with a reduced number of features, a PI-based approach would be more effective than a PP one. Indeed a lot of the peaks are present in the three categories and vary only in intensity. Therefore those kind of features (i.e PEP147X and PEP32X) should be useful for PI classification, but not so much for PP. This statement was confirmed by the results of the overall feature selection: peaks selected for the PI data set are clearly different from the ones selected for PP (**Tab.1**, **Tab.2**). The two classification approaches are thus very different from one another.

By comparing the accuracy of sub-models of PI and PP classification, we can confirm that PI classification is more efficient, as the PI task needs fewer peaks than the PP one to obtain an optimal accuracy (**Fig.5**). An accuracy of 1.0 was obtained with only 5 features for PI while 49 features were needed to obtain the same accuracy for PP. Also note that the chi-squared scores are more than 20x higher for PI feature selection (**Tab.1**, **Tab.2**), further supporting the claim that comparison based on simple Boolean values is not sufficient for an optimal classification.

Taking all of that into account, the final model that we kept for this classification task was the SVM model fitted on the PI data set, using only the 5 best features presented in **Tab.1**.

Summary

In the end, with a simple multi-classification algorithm, we could correctly classify all the skin samples, obtaining a perfect accuracy. After a statistical analysis on the data, the few mass spectral peaks most helpful for the classification task were identified, allowing us to build an equally efficient classification model with a reduced number of features. This confirmed the hypothesis that identification of skin conditions can be easily done using mass spectrometry analysis. Note that if the task was more difficult, we could have used similar techniques, as described in other papers [5] [6].

However this analysis was conducted on a very small data set, therefore more skin samples should be measured and analyzed to fully validate the model. But considering the very good results yield by our algorithm, we are confident that even if the accuracy decreased while running on bigger data sets, it would still perform well.

Acknowledgments

We would like to thank EPFL LEPA [7] and in particular Zhu Yingdi for providing us with the data set and giving us the opportunity to work on this project.

Supplementary Material

The data used are confidential thus we can not give a GitHub repository for our code.

References

[1] G. L. Glish and R. W. Vachet, "The basics of mass spectrometry in the twenty-first century," *Nat. Rev. Drug Discovery*, vol. 2, pp. 140–150, Feb 2003.

[2] S. et al., "Consistency of the Proteome in Primary Human Keratinocytes With Respect to Gender, Age, and Skin Localization," *Mol. Cell. Proteomics*, vol. 12, no. 9, p. 2509, Sep 2013.

[3] "scikit-learn: machine learning in Python — scikit-learn 0.23.2 documentation," Dec 2020, [Online; accessed 16. Dec. 2020]. [Online]. Available: https://scikit-learn.org/stable/index.html

[4] "Choosing the right estimator — scikit-learn 0.23.2 documentation," Dec 2020, [Online; accessed 12. Dec. 2020]. [Online]. Available: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

[5] W. Xu, J. Lin, M. Gao, Y. Chen, J. Cao, J. Pu, L. Huang, J. Zhao, and K. Qian, "Rapid Computer-Aided Diagnosis of Stroke by Serum Metabolic Fingerprint Based Multi-Modal Recognition," *Adv. Sci.*, vol. 7, no. 21, p. 2002021, Nov 2020.

[6] F. M. Nachtigall, A. Pereira, O. S. Trofymchuk, and L. S. Santos, "Detection of SARS-CoV-2 in nasal swabs using MALDI-MS," *Nat. Biotechnol.*, vol. 38, pp. 1168–1173, Oct 2020.

[7] "Laboratory of Physical and Analytical Electrochemistry - EPFL," Dec 2020, [Online; accessed 16. Dec. 2020]. [Online]. Available: https://www.epfl.ch/labs/lepa