Towards CNN Representations for Small Mass Spectrometry Data Classification

Khawla Seddiki, Arnaud Droit
Univ. Laval, Québec, Canada

Frédéric Precioso
Univ. Côte d'Azur, France

Isabelle Fournier, Michel Salzet
Univ. Lille, France

ABSTRACT

An important step to develop rapid and accurate automated clinical diagnosis from Mass Spectrometry (MS) data is building effective classification models. Various Machine Learning (ML) approaches have been tested so far, but most of these require time-consuming processing to remove data artifacts. Even though Convolutional Neural Networks (CNNs) perform well under such circumstances, their effectiveness decreases when the number of available samples is small, which is common in medical applications. Thus, we investigate transfer learning by CNNs to classify small 1D-MS data and then develop a new cumulative learning method when transfer learning is not powerful enough. We proposed to train the same model through several classification tasks over various small datasets to accumulate MS knowledge in the resulting representation. We showed that the use of cumulative learning using datasets generated in different biological contexts, on different organisms, and acquired by different instruments can have a classification accuracy exceeding 98%. Our approach represents a promising strategy to improve classification when the number of samples is small.

CCS CONCEPTS

• **Applied computing → Life and medical sciences**.

KEYWORDS

CNNs, Cumulative Learning, Classification, Small Datasets

1 INTRODUCTION

CNNs are one of the most successful learning architectures. They represent an interesting approach to address rapid classification of cancer and infection data since they can learn representations from raw data [3]. However, efficiency of CNNs trained using a small number of spectra drops rapidly which is unfortunately the case in many real-world applications such as in medicine where a limited number of samples are usually accessible.

2 METHODOLOGY AND RESULTS

We adapted and compared various successful 2D CNNs to fit our 1D tasks [2–4]. We found that a 5-layer architecture was the optimal model depth for feature extraction on the following MS datasets:

1. Canine sarcoma (2228 spectra): including 1 healthy and 11 sarcoma types acquired by a high-resolution Synapt G2S instrument.
2. Human ovary 1 (253 spectra): including healthy and cancerous serums acquired by a low-resolution PBSII-SELDI instrument.
3. Microorganism (117 spectra): including a five bacterial pathogen collection acquired by a high-resolution Synapt G2S instrument.
4. Rat brain (10100 spectra): including rat gray and white brain matter acquired by a high-resolution Rapiflex-MALDI instrument.
5. Beef liver (2637 spectra): including two classes of healthy samples acquired by a high-resolution Synapt G2S instrument.
6. Human ovary 2 (216 spectra): including healthy and cancerous serums acquired by a high-resolution Hybrid Quadruple-SELDI.

CNNs performed poorly for all datasets due to the low number of samples. By using rat brain data as the initial training dataset, sizable gains in accuracy were obtained after transfer learning for all datasets compared to CNNs from scratch, however transfer learning was accurate only for the microorganism dataset (99%) but not powerful enough in cases of data heterogeneity (canine sarcoma: 92%) or low-resolution (human ovary 1: 83%). Hence, we trained the same representation model for several tasks successively across different datasets to converge to an optimal model as shown in Figure 1. By using a cumulative learning approach, classification accuracy exceeded 98% for canine sarcoma and human ovary 1.

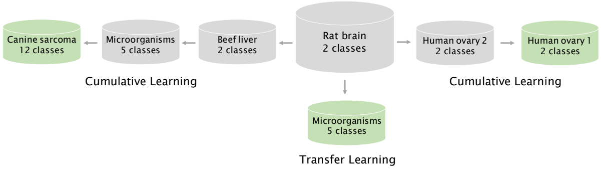

Figure 1: Workflow of transfer and cumulative learning. Source datasets are illustrated in gray and target datasets in green.

Cumulative learning model learned cross-contexts, cross-tasks, cross-organisms, and cross-instruments representation whatever daily, sample or machine variance and from datasets that do not seem to share common features. The learning order was based on the data size, then on the level of resolution. Our approaches outperformed conventional ML algorithms : in the case of canine sarcoma (SVM: 52%, RF: 65%, LDA: 60%), microorganisms (SVM: 54%, RF: 86%, LDA: 67%), and human ovary 1 (SVM: 60%, RF: 88%, LDA: 96%) while not requiring time-consuming processing [1].

3 CONCLUSION

Our methodology goes beyond transferring a representation learnt on one dataset to another since accumulating MS knowledge in the final model without any loss of generalization during the successive phases of training suggests that a "generic" representation of MS classification might exist. In addition, our model was trained to predict 2 classes (rat and gray matter) and was nevertheless efficient for 2, 5 or even 12 classes. The rationale behind this work could be applied to other domains where training samples are scarce. Such end-to-end CNN system efficient on raw data has the potential to contribute to the development of rapid clinical diagnosis workflow.

REFERENCES

[1] Melanie Hilario, Alexandros Kalousis, and Markus Mueller. 2006. Processing and classification of protein mass spectra. *Mass spectrometry reviews* 25 (2006).

[2] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86 (1998).

[3] Jinchao Liu, Margarita Osadchy, and Stuart J Gibson. 2017. Deep convolutional neural networks for Raman spectrum recognition. *Analyst* 142 (2017).

[4] Jie Lu, Vahid Behbood, Peng Hao, Hua Zuo, and Guangquan Zhang. 2015. Transfer learning using computational intelligence. *Knowledge-Based Systems* 80 (2015).