Available online at www.sciencedirect.com
ScienceDirect
Mendeleev Commun., 2024, **34**, 786–787

Mendeleev Communications

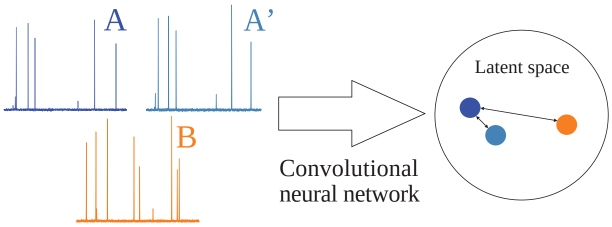
Contrastive representation learning for spectroscopy data analysis

Artem P. Vorozhtsov* and Polina V. Kitina

Department of Fundamental Physical and Chemical Engineering, M. V. Lomonosov Moscow State University, 119991 Moscow, Russian Federation. E-mail: artem2001qaz@gmail.com

DOI: 10.1016/j.mencom.2024.10.006

Metric-based representation learning showed good accuracy in identifying objects from one-dimensional spectroscopy data, robustness to small dataset size and the ability to change the data domain without fine-tuning.



Keywords: spectroscopy, machine learning, representation learning, neural network, metric learning, spectra analysis.

Spectroscopic methods have found a wide range of applications in chemistry and related scientific fields, but processing spectroscopic data often takes a lot of time from researchers. A number of machine learning-based approaches have been proposed that allow automating the processing of spectroscopic data[1,2] and solving the problem of classifying objects based on their spectra.[3,4] The most typical approach to solving the classification problem is the use of neural networks with several convolutional layers capable of efficiently extracting the spatial structure of the input data, and subsequent fully connected layers, the output of which generates the probabilities of an object belonging to a particular class. This model architecture requires a large amount of training data and does not allow applying the model to new classes of objects without changing the number of neurons in the output layer and fine-tuning.[5]

This work focuses on evaluating the benefits of using representation learning, which shows good performance in small-data learning tasks[6] and can be transferred to a new data domain without fine-tuning.[7] Previously, representation learning was used for MS/MS data analysis[8] and fast compound identification from ^{13}C NMR data.[9] A universal synthetic spectroscopic dataset[10] containing one-dimensional spectra of objects of 500 different classes was used as the dataset for training and testing the model (for example data, see Figure S1 in Online Supplementary Materials). Our spectroscopic data identification algorithm consists of two parts. First, a convolutional neural network (for architecture, see Figure S3) encodes the input data into a latent representation space such that objects of the same class are close to each other and distant from objects of other classes (Figure 1). Then, an object from the test dataset is fed to this neural network, and its class is determined from the learned latent representation of the object using the nearest neighbor algorithm.[11] To achieve class separation in the latent representation space, the neural network weights are optimized during the training phase to minimize the Triplet Loss function (see Online Supplementary Materials).

Previously, a neural network with identical convolutional layer architecture was used to solve a classification problem on the same test dataset[10] and achieved an accuracy of 99.02 ± 0.21%. Using the algorithm described above, we were able to obtain an accuracy of 99.03 ± 0.27%. Moreover, training the model on 40% of the training set with the same number of classes yielded an accuracy of 98.95 ± 0.31%, while training on only 10% of the training set yielded an accuracy of 97.64 ± 0.21%. These results show that the use of representation learning can maintain high accuracy even when using significantly fewer training samples than required by the classical multiclass classification algorithm. We also analyzed how effective it is to use a pre-trained neural network to deal with classes not encountered during training. For this purpose, we trained the model on the objects of 400 out of 500 classes in the training set, and then tested it on the remaining 100 classes without fine-tuning. This resulted in an accuracy of 99.667%. In contrast to this, a conventional classification algorithm cannot be applied to a new data domain without fine-tuning and modifying the last fully connected layer.

An important feature of latent representations is the meaningfulness of arithmetic operations performed on them. In particular, it was found that the normalized latent representation of several spectra superimposed on each other (which simulates a mixture of compounds in a real experiment) is close to the

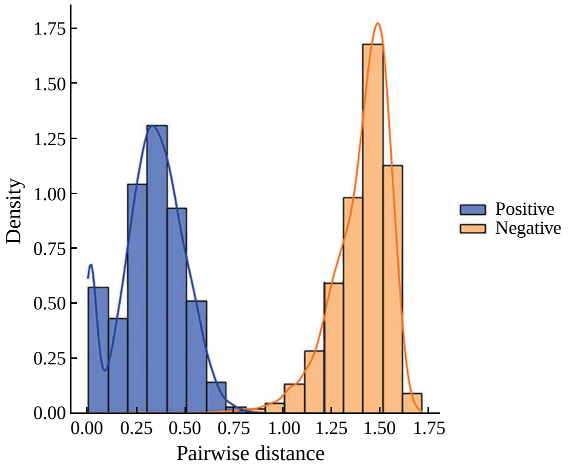


Figure 1 Distribution of pairwise distance between elements of the same classes (positive) and elements of different classes (negative).

normalized sum of the latent representations of each spectrum separately. This property can be useful for analyzing spectra containing signals from several components simultaneously.

Based on the described property, we proposed a simple algorithm for processing the spectra of such 'mixtures' by means of representation learning. At the first step of this algorithm, latent representations of each class of objects are formed. Then, to form latent representations of mixtures, the algorithm calculates the normalized sums of latent representations of classes for all their unique combinations. After that, a test 'mixture' is fed to the neural network, and the nearest neighbor is determined for the obtained latent representation of the 'mixture'. The class of the mixture (*i.e.*, the list of classes whose objects are included in the mixture) is determined by the class of the found nearest neighbor. The disadvantage of this algorithm is the significant memory costs for processing mixtures of a large number of elements, but its high accuracy (Table 1) makes it promising for analyzing spectra.

To identify the limitations of the model, it was also tested on a specially prepared dataset[10] containing objects of 27 classes divided into three groups: (1) classes that differ from each other by low-intensity peaks; (2) classes where the peak positions differ slightly from each other; (3) classes with the same peak positions but different intensities. On this dataset, the accuracy of our algorithm was 53.08%, while the benchmark accuracy was 52.47%. A more detailed analysis (Figure S6) showed that the algorithm is able to accurately predict the classes of the second group, but not the first and third. Apparently, the limited classification accuracy on this dataset is due to the low ability of the used convolutional network to handle spectra with close peak intensities and does not depend on the way it is trained.

Table 1 Accuracy of prediction algorithms.

Number of correct predictions[a]	Accuracy (%)			
	Quaternary mixture	Ternary mixture	Binary mixture	Identification of individual compound
1	99.3	99.6	99.2	99.0
2	96.0	96.6	92.0	–
3	84.8	78.8	–	–
4	59.4	–	–	–

[a] The number of mixture components whose classes were predicted correctly.

The results show that representation learning is a promising tool for analyzing spectroscopic data due to the low training data quantity requirements and the ability to be used on new data domains without fine-tuning.

This work was supported by the Foundation for the Development of Science and Education 'Intellect'.

Online Supplementary Materials

Supplementary data associated with this article can be found in the online version at doi: 10.1016/j.mencom.2024.10.006.

References

1 R. Luo, J. Popp and T. Bocklitz, *Analytica*, 2022, **3**, 287; https://doi.org/10.3390/analytica3030020.

2 K. Wu, J. Luo, Q. Zeng, X. Dong, J. Chen, C. Zhan, Z. Chen and Y. Lin, *Anal. Chem.*, 2021, **93**, 1377; https://doi.org/10.1021/acs.analchem.0c03087.

3 L. Xu, J. Xie, F. Cai and J. Wu, *Electronics*, 2021, **10**, 1892; https://doi.org/10.3390/electronics10161892.

4 C. Zhang, Y. Idelbayev, N. Roberts, Y. Tao, Y. Nannapaneni, B. M. Duggan, J. Min, E. C. Lin, E. C. Gerwick, G. W. Cottrell and W. H. Gerwick, *Sci. Rep.*, 2017, **7**, 14243; https://doi.org/10.1038/s41598-017-13923-x.

5 X. Liu, H. An, W. Cai and X. Shao, *TrAC, Trends Anal. Chem.*, 2024, **172**, 117612; https://doi.org/10.1016/j.trac.2024.117612.

6 B. Li, M. N. Schmidt and T. S. Alstrøm, *Analyst*, 2022, **147**, 2238; https://doi.org/10.1039/D2AN00403H.

7 Y. Xian, C. H. Lampert, B. Schiele and Z. Akata, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, **41**, 2251; https://doi.org/10.1109/TPAMI.2018.2857768.

8 H. Guo, K. Xue, H. Sun, W. Jiang and S. Pu, *Anal. Chem.*, 2023, **95**, 7888; https://doi.org/10.1021/acs.analchem.3c00260.

9 Z. Yang, J. Song, M. Yang, L. Yao, J. Zhang, H. Shi, X. Ji, Y. Deng and X. Wang, *Anal. Chem.*, 2021, **93**, 16947; https://doi.org/10.1021/acs.analchem.1c04307.

10 J. Schuetzke, N. J. Szymanski and M. Reischl, *npj Comput. Mater.*, 2023, **9**, 100; https://doi.org/10.1038/s41524-023-01055-y.

11 J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon and S. J. Gibson, *Analyst*, 2017, **142**, 4067; https://doi.org/10.1039/C7AN01371J.

Received: 31st May 2024; Com. 24/7516