ISSN 0027-1349, Moscow University Physics Bulletin, 2024, Vol. 79, Suppl. 2, pp. S898–S905. © Allerton Press, Inc., 2024.

Nonlinear Relevance Estimation of Multicollinear Features for Reducing the Input Dimensionality of Optical Spectroscopy Inverse Problem

N. O. Shchurov[1,2*], I. V. Isaev[1,3], S. A. Burikov[1,2], K. A. Laptinskiy[1,2], O. E. Sarmanova[1,2], T. A. Dolenko[1,2], and S. A. Dolenko[1]

[1] *Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Moscow, Russia*
[2] *Physics Department, Lomonosov Moscow State University, Moscow, Russia*
[3] *Kotelnikov Institute of Radio Engineering and Electronics, Russian Academy of Sciences, Moscow, Russia*
Received October 1, 2024; revised October 10, 2024; accepted October 20, 2024

Abstract—In this study we considered an inverse problem of optical spectroscopy. It consists in determining concentrations of the ingredient ions of multicomponent water solutions by their spectra. The problem of describing the spectra of multicomponent solutions is nonlinear and has no adequate mathematical model. Because of this, machine learning methods using experimental data were chosen to solve this problem. At the same time, inverse problems of spectroscopy are characterized by high input dimensionality with a large number of features, more or less relevant. In their turn, some of the relevant features are redundant due to their multicollinearity. This is caused by the fact that the characteristic lines have a width of several spectrum channels. Presence of redundant features leads to a deterioration in the quality of machine learning solution of the problem. Thus, there is a need for a feature selection procedure that takes into account both their relevance and redundancy, as well as their nonlinear relationship with the determined parameters. In this study, we considered a feature selection procedure based on the iterative selection of features with the highest relevance to the target variable and on the elimination of features with a high relationship with each other. In this selection process, we used a trained neural network to analyze weights and determine feature importance in a nonlinear way. We also used the Pearson correlation coefficient to measure how features are related to one another. Finally, we compared the quality of a neural network solution using spectroscopic data of the full set of input features and of its subsets. These subsets were compiled using the selection procedure under consideration. We also used traditional methods for selecting significant input features as baseline methods.

Keywords: dimensionality reduction, feature selection, multicollinearity, entropy, spectroscopy, data aggregation

DOI: 10.3103/S0027134924702357

1. INTRODUCTION

Inverse problems (IPs) of optical spectroscopy (OS) consist in determining the composition of the sample under study based on its spectra. Many problems of this type are characterized by the absence of an analytical solution for both the IP itself and the corresponding direct problem: constructing a spectrum based on the known composition of the sample. Therefore, approximation methods based on the use of experimental data are practically the only way to solve such problems [1]. In addition, IPs of OS are characterized by nonlinearity, mathematical incorrectness, as well as high input and output dimensions, which further complicates their solution. Currently, approximation methods of machine learning are often used to solve such problems [1–5]. These methods are capable of approximating complex nonlinear dependencies and they do not require any information on the type of relationship between the input and output data. In this paper, artificial neural networks (NN), which are universal approximators, were chosen to solve IPs of OS [6].

When using NN to solve a wide range of problems, there is a need to reduce the dimensionality of the input data to obtain a more accurate and stable solution with a reduction in computational complexity. In addition, reducing the dimensionality of the input data allows the model to increase its generalizing

*E-mail: shchurov_no2@mail.ru

ability [7]. When solving the OS problem, in addition to the high dimensionality of the data, multicollinearity of the input features is often observed. This is explained by the fact that spectral lines during spectrum registration can occupy several channels. As a result, close spectrum channels can carry close information. Because of this, it is necessary to use feature selection methods that take into account the relationships between the input features.

According to the currently accepted classification, methods for selecting essential features are divided into filters, built-in methods, and wrappers [7, 8]. Filter-type selection methods are based on calculating some metric based on the data itself and selecting features based on the threshold value of this metric. They have the lowest computational cost, but most methods of this type work poorly with multicollinear features, since they select or reject the entire group of such features at once. Built-in selection methods are based on the analysis of the trained model. In the case of multicollinear features, the model can take information from any of them during the training process, so the results of such selection can differ significantly from run to run of the training procedure. In addition, the presence of multicollinear features can negatively affect the model training procedure itself, leading to unstable results of its operation. Wrapper-type selection methods are based on repeatedly solving the problem on different subsets of features and choosing the subset on which the solution is best. They allow finding optimal sets of features even in the presence of multicollinearity between them, however, this method has the highest computational cost, and with a high dimensionality of the problem being solved, they become practically inapplicable.

However, there are a number of selection methods that are capable of working with multicollinear features [9]. In the case of using filter-type methods, an additional procedure for excluding interrelated features is introduced into the selection algorithm. For example, in the fast correlation-based filter (FCBF) algorithm, it is based on calculating the correlation between features [10], and in the minimal-redundancy-maximal-relevance (mRMR) algorithm, it is based on calculating the mutual information criteria [11]. Built-in methods can also be supplemented with a procedure for excluding interrelated features, for example, also based on calculating the correlation between features [12]. In addition, excluding multicollinear features can be implemented at the training stage of algorithms by introducing penalty terms into the loss function. Examples of such algorithms are ridge regression, LASSO, ElasticNet [13]. Modifications of wrapper-type methods are aimed at reducing their computational complexity in order to ensure their applicability to high-dimensional problems. Here we can note such methods as sequential inclusion/exclusion of features [14]. In this paper, the application of the FCBF method to the inverse problem of Raman spectroscopy and optical absorption spectroscopy was considered. Also, we considered modifications of this method based on the calculation of entropy or nonlinear significance of features.

The paper compares the quality of the neural network solution to the problem of determining the concentration of heavy metal ions in water using Raman spectra and absorption spectra, as well as their combined use (integration), on a full set of input features and on its subsets. These subsets of features are compiled by the considered selection method using traditional methods for selecting significant input features and methods for assessing the relationship of features, such as feature selection based on mutual correlation, cross-entropy and nonlinear significance. This work is a continuation of the research begun by the authors in [15].

2. PURPOSE OF THE STUDY

The purpose of this study is to test the efficiency of the method for selecting essential features taking into account their multicollinearity in the case of solving the inverse problem of spectroscopy. The paper will consider determination of the optimal parameters of the algorithm and compare the results obtained using this method using various metrics (cross-correlation, cross-entropy or neural network weights), as well as when training neural networks on the full feature set. In addition, the effect of combining selected features from two types of spectroscopy was tested.

3. STATEMENT OF THE PROBLEM

Experimentally obtained absorption and Raman spectra of solutions were used as data for training the neural network. The original data set contained 3806 patterns corresponding to solutions of various compositions. The solutions under consideration contained from 1 to 6 salts and from 2 to 5 ions in the concentration range of 0–0.14 M with a step of 0.01 M. The parameters to be determined in this problem were the pH value and the concentrations of five ions:

$$Cu^{2+}, Ni^{2+}, Co^{2+}, SO_4^{2-}, NO_3^-$$

The input features when using neural networks were the intensity values in the corresponding channels of the spectra. Thus, the initial input dimension of the problem was determined by the type(s) of spectra used:

- 2048 features for the Raman spectrum.

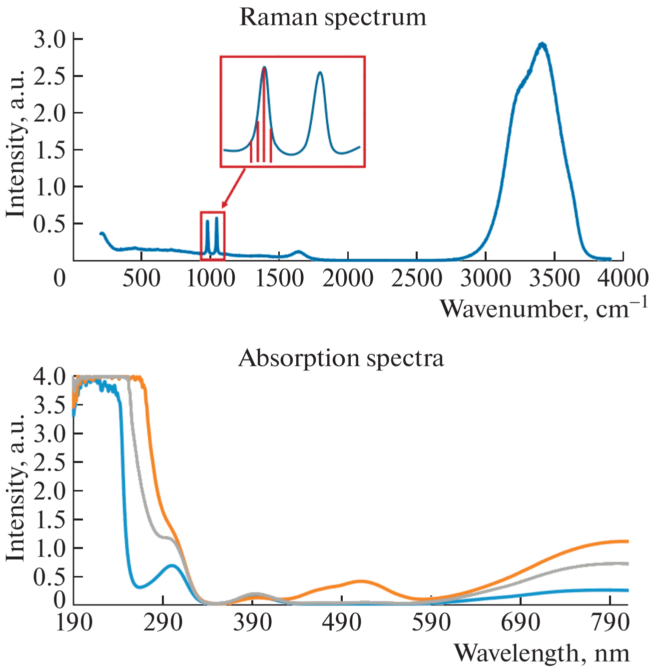

Fig. 1. Examples of Raman spectrum (top) and absorption spectra (bottom).

- 811 features for the absorption spectrum.
- 2048 + 811 features when combined.

In addition to the high input dimension, the problem is also characterized by multicollinearity of input features, since spectral lines or bands occupy at least several channels, and the intensity values in adjacent channels carry similar information (Fig. 1).

4. USING NEURAL NETWORKS

To reduce the output dimension of the problem, autonomous parameter determination was used, i.e., six separate networks with a single output each were trained—one for each parameter being determined. To prevent overtraining, the early stopping method was used. For this, the original data set was divided into subsets in the following ratio:

- Training 70% 2656 patterns.
- Validation 20% 750 patterns.
- Test 10% 400 patterns.

All neural networks used to solve the inverse problem had the following parameters:

- Early stopping: after 800 epochs without error reduction on the validation set.
- Architecture: multilayer perceptron with one hidden layer.
- Number of neurons in the hidden layer: 32.
- Optimizer: Adam.
- Activation function:
  - Hidden layer—logistic
  - Output layer—linear.

5. DESCRIPTION OF THE FEATURE SELECTION ALGORITHM

This paper considers an algorithm based on the iterative selection of features with the greatest relevance to the target variable and the exclusion of redundant features with high statistical dependence between them. The need to use such a feature selection method arises due to high values of multicollinearity between the input features (Fig. 2).

The algorithm is a two-stage iterative procedure (iterative feature selection, IFS) (Fig. 3). At the first stage, the feature with the highest relevance to the target variable is selected. At the second stage, features are excluded, the similarity of each of which to the feature selected at the previous stage is higher than a certain threshold value. Then the procedure is repeated for the remaining set of features. The process is repeated until the stopping criterion is reached.

Thus, the method under consideration has the following adjustable parameters:

1. The maximum permissible value of similarity of the feature under consideration with other selected features (hereinafter referred to as the threshold value T_{XX}).
2. The stopping criterion, which can be a specified number of selected features N or the minimum permissible value of relevance of the feature under consideration to the target variable (hereinafter referred to as the threshold value T_{XY}).

6. APPLICATION OF THE FEATURE SELECTION ALGORITHM

The following algorithm parameters were considered in this study:

- The Pearson correlation coefficient and cross-entropy normalized to the range from 0 to 1 were considered as relevance and similarity metrics.

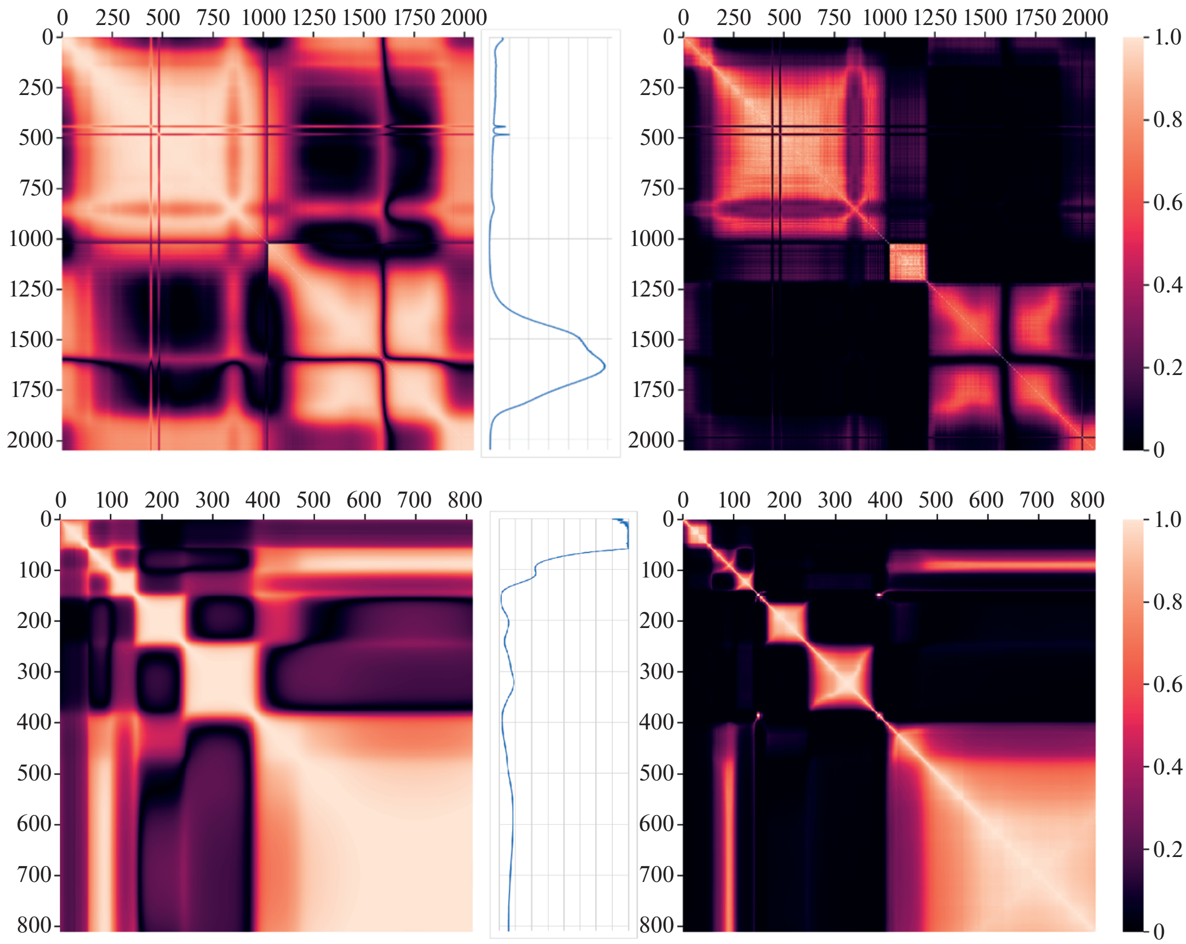

Fig. 2. Heat maps of cross-correlation (left) and cross-entropy (right) values for Raman (top) and absorption (bottom) spectra.

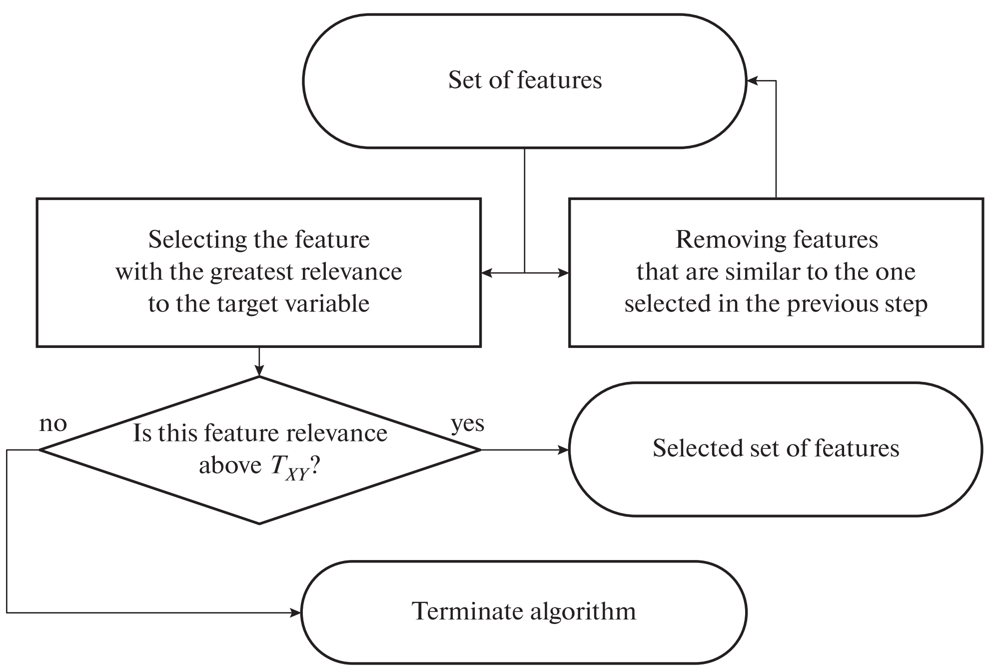

Fig. 3. Scheme of the iterative feature selection algorithm.

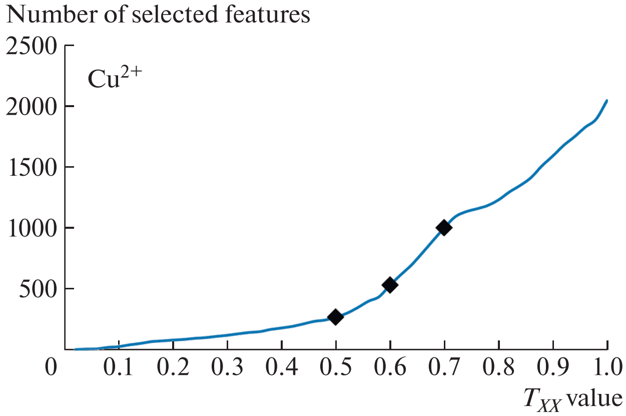

Fig. 4. Dependence of the number of selected features on the value of the T_{XX} threshold (the maximum permissible value of similarity of a feature with other input features) using Pearson correlation coefficient as the similarity metric.

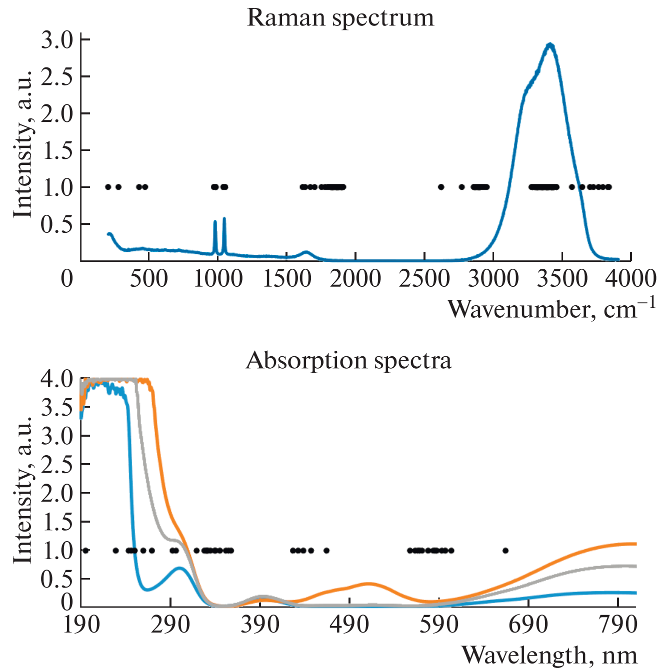

Fig. 5. An example of features selected by the studied algorithm (black dots) for Raman spectroscopy (top) and for absorption spectroscopy (bottom) using Pearson correlation coefficient as the similarity metric.

- The nonlinear estimation of relevance by neural network weight analysis (NNWA) [17] was considered as an additional relevance metric.
- The maximum permissible threshold value of similarity of the considered feature with other input features T_{XX} was selected during the computational experiment.
- To ensure the possibility of comparing the efficiency of the algorithm using different parameters, the stopping criterion was the selection of a given number of features N.

The nonlinear estimation of relevance was based on weight analysis of perceptron neural networks with the following parameters: single hidden layer with 16 neurons, stochastic gradient descent, learning rate 0.01, moment 0.5, early stopping at 100 epochs after minimum of the error on the training set, logistic activation function in the hidden layer and linear in the output layer. Each network was trained 5 times with various initial weights, and the obtained relevance values were averaged over the 5 runs.

In the first experiment, the threshold value T_{XY} remained unchanged and equal to zero, i.e., no restrictions were imposed on the minimum acceptable value of the relevance of the feature under consideration with the target variable. To select the threshold T_{XX}, graphs of the dependence of the number of selected features on the threshold value T_{XX} were constructed (Fig. 4) using Pearson correlation coefficient as the similarity metric. For further consideration, three threshold values were taken from the inflection section of the graph: $T_{XX} = 0.5, 0.6, 0.7$.

An example of the set of features selected by this algorithm using Pearson correlation coefficient as the similarity metric is shown by black dots on the Raman and absorption spectrum graphs (Fig. 5). The absorption spectra of three different solutions are displayed on the graph in different colors. The graph shows that for the Raman spectrum, the algorithm selected features corresponding to the characteristic bands of ions (900–1100 cm^{-1}) and stretching vibrations of water (3200–3500 cm^{-1}). This choice corresponds to existing physical concepts of spectra formation.

7. RESULTS OF SOLVING THE INVERSE PROBLEM OF SPECTROSCOPY

To train neural networks, features were selected by the studied algorithm using different threshold values selected as described above ($T_{XX} = 0.5, 0.6, 0.7$). The quality of the neural network solution increases in some cases with an increase in the threshold value T_{XX}, which can be caused by an increase in the total number of selected input features. However, based on the quality of the obtained models (Fig. 6), the threshold value $T_{XX} = 0.6$ was selected as optimal for further consideration. In Fig. 6, the coefficient of determination R^2 and the mean absolte error of the

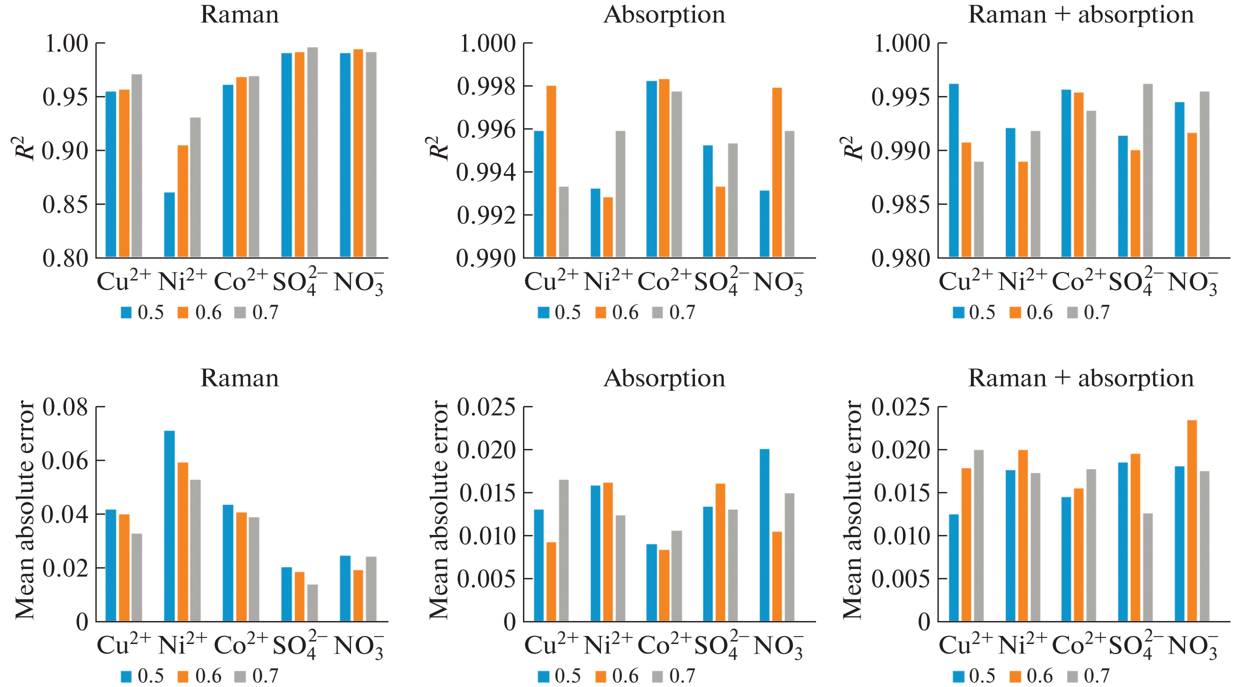

Fig. 6. Quality of the solution (top—R^2, bottom—mean absolute error in M) of the inverse problem for different values of the T_{XX} threshold and different input data.

solution of the inverse problem in M were used as metrics for evaluating the obtained models.

When training neural networks on the full feature set, combining the spectra leads to a decrease in the quality of the decision in respect to the individual use of absorption spectroscopy data (Fig. 7). The reason for this is the too large number of not so relevant features [7]. In [16], absorption and Raman spectra were also combined, and neural networks showed worse results on the combined data set compared to using only the absorption spectra.

When determining the pH of a solution, feature selection methods show an advantage on the combined data set, but when using only the Raman spectrum or absorption, the quality of networks trained on the full feature set is higher (Fig. 7). The lower accuracy of neural networks trained on data sets using feature selection methods may be due to a stronger influence of noise in the data on the result, since in this case the total number of features used in the training process is smaller.

Feature selection using the method under consideration allows improving the quality of the solution to the Raman spectrum problem only when using nonlinear significance as a metric.

The greatest advantage of using feature selection is observed when combining the Raman and absorption spectra into one set (Fig. 7). In this case, the best results are shown by the CC IFS and NNWA IFS feature selection methods. Using CE IFS gave worse results in determining ion concentrations. The data set we used is not very large, and there is noise, so the accuracy of calculating the cross-entropy for each of the features is low and may be insufficient for using cross-entropy as a metric for feature selection. An argument in favor of using cross-correlation rather than cross-entropy is that removing distant but correlated channels has a positive effect on the accuracy of the solution, since linear relationships with distant channels (Fig. 2) can be related to experimental factors that affect the spectrum as a whole. Examples of such factors can be laser power fluctuations that are not eliminated by normalization and Rayleigh illumination for small Raman shifts in the Raman spectra.

8. CONCLUSIONS

The features extracted using the method under consideration correspond to existing physical notions and allow significant reduction in the dimensionality of the input data. Also, when using feature selection based on Pearson cross-correlation and nonlinear significance based on NNWA, an improvement in the quality of the neural network solution is observed when combining Raman spectroscopy and absorption spectroscopy data, which allows using this method to extract the most significant features from different types of data sets.

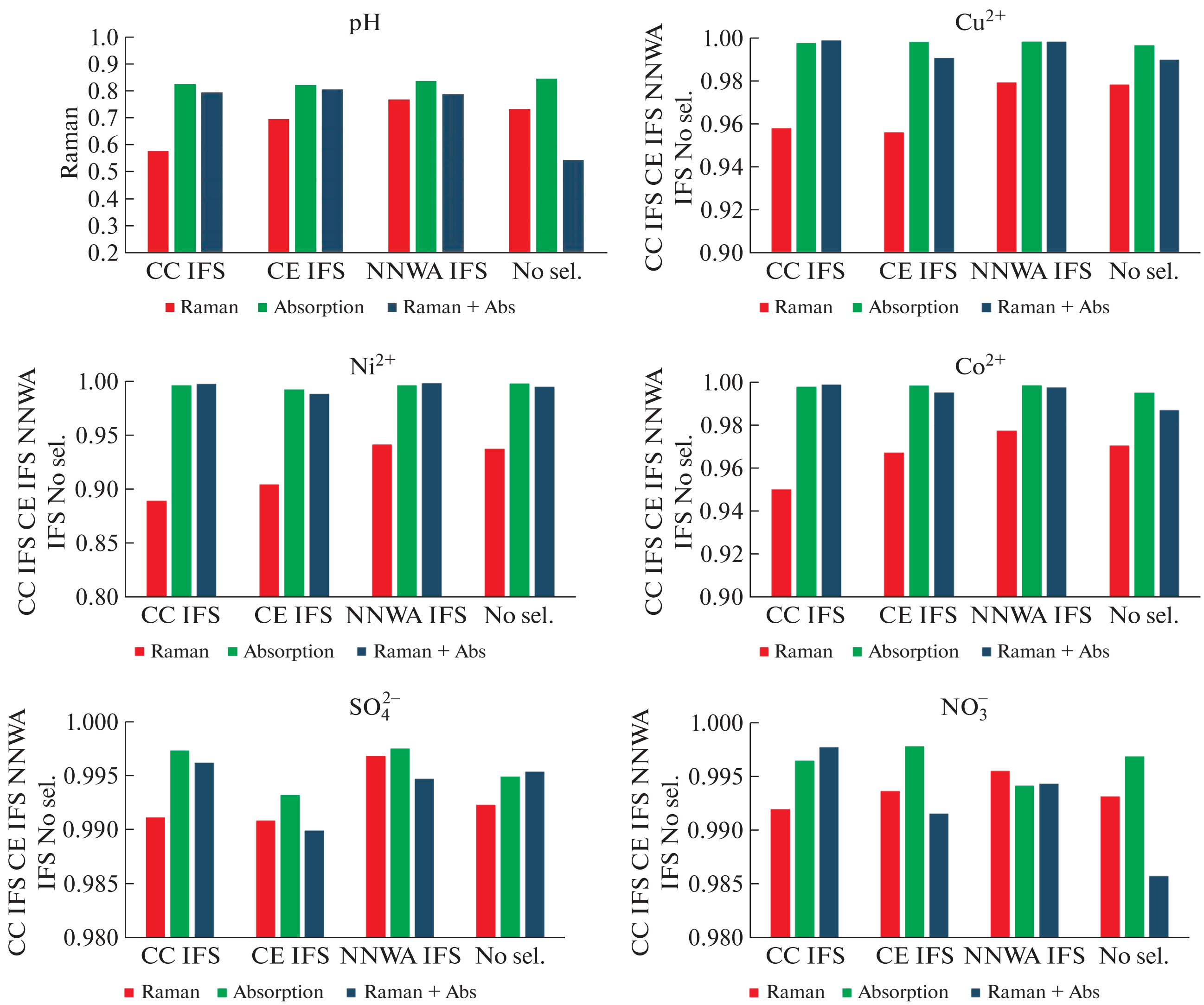

Fig. 7. The R^2-coefficient values for the inverse problems of determination of concentration for the five types of ions, solved by the features selected by the IFS algorithm based on cross-correlation, cross-entropy, neural network weight analysis (nonlinear significance) and on the full set of features for different combinations of spectroscopic data.

Using cross-entropy for feature selection did not yield better results, which may be due to the insufficient size of the training dataset for calculation of entropy with sufficient precision.

FUNDING

This study was supported by the Russian Science Foundation, grant no. 24-11-00266, https://rscf.ru/en/project/24-11-00266/.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. H.-P. Wang, P. Chen, J.-W. Dai, D. Liu, J.-Y. Li, Y.-P. Xu, and X.-L. Chu, TrAC, Trends Anal. Chem. **153**, 116648 (2022). https://doi.org/10.1016/j.trac.2022.116648

2. C. A. Meza Ramirez, M. Greenop, L. Ashton, and I. U. Rehman, Appl. Spectrosc. Rev. **56**, 733 (2021). https://doi.org/10.1080/05704928.2020.1859525

3. W. Zhang, L. Ch. Kasun, Q. J. Wang, Yu. Zheng, and Zh. Lin, Sensors **22**, 9764 (2022). https://doi.org/10.3390/s22249764

4. Sh. Guo, J. Popp, and T. Bocklitz, Nat. Protoc. **16**, 5426 (2021). https://doi.org/10.1038/s41596-021-00620-3

5. L. Pan, P. Zhang, Ch. Daengngam, Sh. Peng, and M. Chongcheawchamnan, J. Raman Spectrosc. **53**, 6 (2022). https://doi.org/10.1002/jrs.6225
6. G. Cybenko, Math. Control, Signals Syst. **2**, 303 (1989). https://doi.org/10.1007/bf02551274
7. I. Guyon and A. Elisseeff, J. Mach. Learn. Res. **3**, 1157 (2003). https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf?trk=public_post_comment-text.
8. J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, ACM Comput. Surv. **50**, 94 (2017). https://doi.org/10.1145/3136625
9. J. Y.-L. Chan, S. M. H. Leow, Kh. T. Bea, W. Kh. Cheng, S. W. Phoong, Z.-W. Hong, and Y.-L. Chen, Mathematics **10**, 1283 (2022). https://doi.org/10.3390/math10081283
10. L. Yu and H. Liu, J. Mach. Learn. Res. **5**, 1205 (2004). https://www.jmlr.org/papers/volume5/yu04a/yu04a.pdf.
11. H. Peng, F. Long, and C. Ding, IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1226 (2005). https://doi.org/10.1109/TPAMI.2005.159
12. E. Gasca, J. S. Sánchez, and R. Alonso, Pattern Recognition **39**, 313 (2006). https://doi.org/10.1016/j.patcog.2005.09.002
13. Sh. Altelbany, J. Appl. Econ. Bus. Stud. **5**, 131 (2021). https://doi.org/10.34260/jaebs.517
14. G. Borboudakis and I. Tsamardinos, J. Mach. Learn. Res. **20** (8), 1 (2019). https://www.jmlr.org/papers/v20/17-334.html.
15. N. Shchurov, I. Isaev, S. Burikov, T. Dolenko, K. Laptinskiy, and S. Dolenko, in *Proceedings of The 6th International Workshop on Deep Learning in Computational Physics—PoS(DLCP2022)* (Sissa Medialab, 2022), Vol. 429, p. 26. https://doi.org/10.22323/1.429.0026
16. I. Isaev, N. Trifonov, O. Sarmanova, S. Burikov, T. Dolenko, K. Laptinskiy, and S. Dolenko, Proc. SPIE **11458**, 114580R (2020). https://doi.org/10.1117/12.2564398
17. M. Gevrey, I. Dimopoulos, and S. Lek, Ecol. Modell. **160**, 249 (2003). https://doi.org/10.1016/s0304-3800(02)00257-0