An Information System for Inorganic Substances Physical Properties Prediction Based on Machine Learning Methods

V.A. Dudarev[1,2][0000-0001-7243-9096], N.N. Kiselyova[1][0000-0002-3583-0704], A.V. Stolyarenko[1], A.A. Dokukin[3], O.V. Senko[3], V.V. Ryazanov[3], E.A. Vashchenko[4], M.A. Vitushko[4] and V.S. Pereverzev-Orlov[4]

[1] A.A. Baikov Institute of Metallurgy and Materials Science of RAS (IMET RAS), Moscow, 119334, Russia
[2] HSE University, Moscow, 109028, Russia
[3] Federal Research Center "Computer Science and Control" of RAS (FRC CSC RAS), Moscow, 119333, Russia
[4] A.A. Kharkevich Institute for Information Transmission Problems of RAS (IITP RAS), Moscow, 127051, Russia
kis@imet.ac.ru

Abstract. ParIS (Parameters of Inorganic Substances) system was developed for predicting inorganic substances physical properties. It is based on the use of machine learning methods to find the relationships between inorganic substances parameters and the properties of chemical elements. The main components of the system are an integrated database system on inorganic substances and materials properties, a subsystem of machine learning and prediction results analysis, a knowledge base and a prediction database. The machine learning subsystem includes programs based on the algorithms developed by the authors of this paper and the algorithms included in the scikit-learn package. The results of the ParIS system application are illustrated by an example of predicting chalcospinels crystal lattice parameter. To get prediction results, only the properties of chemical elements included in the composition of not yet synthesized chalcospinels were used. Moreover, the prediction accuracy was within ± 0.1 Å.

Keywords: Machine Learning, Databases, Prediction of Inorganic Substances Physical Properties.

1 Introduction

Machine learning methods are widely used in chemistry. Object classification and qualitative (categorical) characteristics prediction tasks are among the most successfully solved problems. In inorganic chemistry machine learning methods made it possible to make predictions knowing only the parameters of chemical elements. For example, it's possible to predict compound formation of a certain composition and / or with a given crystal structure type under certain external conditions with an average accuracy of

more than 80% (according to the results of predictions comparison with new experimental data) [1-3]. However, categorical properties prediction is only a small part of the practical problems in chemistry. The vast majority of problems are associated with the quantitative objects characteristics prediction (for example, crystal lattice parameters, melting and boiling points, impact strength, elasticity, electrical conductivity, etc.). Despite the great importance of such tasks of predicting the quantitative (numerical, scalar, vector) objects properties, machine learning methods, which in this case are often connected with regression problems, are used to solve them much less frequently. In applications to inorganic chemistry, this paradox is associated, in particular, with the classical regression analysis limitations: multicollinearity problems (and as a result in poor conditionality of feature description matrices), the approximated dependencies non-smoothness, the large feature description dimension combined with a small number of precedents, presence of erroneous outliers in data etc. The above-mentioned problems are task's peculiarities in inorganic chemistry. The use of regularization (ridge regression methods, LASSO, LARS, elastic networks, regularized neural networks, etc.), various methods of the most important features selecting, filtering of erroneous outliers, in many cases allows to circumvent some of these limitations. However, the task of developing combined methods that would overcome most of the limitations in solving the problems of reconstructing multivariate regression cannot be considered as completely solved. One of the common ways to develop such methods is to consider problem domain (inorganic chemistry) peculiarities. The development of methods and systems for predicting inorganic substances quantitative parameters based on machine learning (ML) methods allows to speed up the search, research and introduction of new materials with specified functional properties. We have developed such an information system for searching for relationships that connect physical and chemical properties of inorganic compounds with the properties of chemical elements that form compound. The developed system allows a solution of various tasks in inorganic chemistry.

2 Selection of Machine Learning Methods for Prediction of Inorganic Compounds Physical Properties

Let's define some terms that we use. In this study, an *object* is a chemical system (inorganic compound, solid solution, heterogeneous mixture, etc.) formed by *components* (chemical elements or simpler inorganic compounds), represented in computer's memory as a set of *attribute* values (component properties) with indication of the value of a given physical or chemical property. A *quantitative property* is an object parameter expressed as a numeric-scaled (scalar, vector) variable.

Let's consider the various most commonly used methods and their limitations in solving the problem, specified by the characteristics of inorganic chemistry.

The most widely-spread method for predicting the quantitative properties of objects is multivariate regression analysis (multiple regression) [4]. It is designed to analyze

the relationship between several independent variables (also called regressors or predictors – in our case, the properties of components) and the dependent variable (compound property). Limitations (taking into account the problem domain peculiarities):

- it is assumed that the residuals (the dependent variable calculated values minus experimental values) are distributed normally, while the independent variables do not contain errors in values. However, any experimental data are not error-free, and moreover the normality of corresponding distributions is always a moot point;
- the assumption of the absence of property multicollinearity, which leads to poor conditionality of the feature description matrix and the instability of the regression coefficients estimates. It should be noted that the chemical compounds properties description is strongly correlated due to the dependences of chemical elements properties on their atomic number;
- the approximated function smooth character requirement. It must be taken into account, that the dependences of inorganic compounds properties on chemical elements parameters often take the "saw" form with different tooth sizes (due to the Periodic Law).

Support vector regression (SVR) [5, 6] is widely used in chemistry to predict the quantitative properties of substances [7]. In its implementation the regression model parameters are determined by the quadratic programming problem solution, which has a unique solution. The problems of using SVR are connected with the lack of recommendations for choosing the kernel function parameters that are most suitable for solving a specific problem, as well as other effective algorithm parameters (for example, the penalty coefficient). In addition, the algorithm is very sensitive to data outliers while chemical problems, as a rule, contain erroneous and out-of-date experimental values. One of the ways to solve the last problem is outliers detection and filtering, for example, using a system developed and used by us in chemistry [8, 9]. The overfitting problem can be solved by means of regularization.

Artificial neural networks learning can be used both for calculating functions of qualitative and quantitative parameters. In the latter case (for example, for training networks with radial basis functions (RBF) [10] or generalized regression neural network (GRNN) [11]) for the network stability to measurement errors of input vectors, it's required continuity of the conduction functions of edges and functions of neurons activation, and for the network learning using gradient methods, their differentiability is required also. Recent requirements limit the opportunity of using these methods in inorganic chemistry. The overfitting problem can be solved by means of regularization also. It should be noted that neural networks are weak in properties extrapolation. The method disadvantages include the lack of modeling transparency, which does not allow a physical interpretation of the results obtained, the complexity of choosing a network architecture, high requirements for measurement errors, the complexity of choosing a learning algorithm, and high resource consumption of neural networks learning process.

It should be noted that all the above methods, as well as many others, are included in many free distributed and widely used software packages: the scikit-learn package [12], which contains a number of ML-algorithms based on the Python programming language, and packages for statistical data processing in R language [13].

The creation of combined algorithms is one of the promising modern trends in the development of methods for predicting quantitative properties. This approach makes it possible to compensate the shortcomings of some algorithms at the expense of the advantages of others and is aimed at improving the prediction accuracy of quantitative parameters, as one of the main criteria for methods effectiveness. Possible approaches are a combination of classification algorithms, an elastic network, combinations of SVR and multidimensional regression, etc. The following algorithms and programs that implement this approach are included in the system for predicting inorganic compounds physical properties.

2.1 Locally Optimal Convex Combinations (LOCC)

The multilevel method and algorithm for constructing a multidimensional regression model based on convex combinations of predictors has some similarities with deep learning technology. For example, in [14], when solving the problem of predicting the halides melting points at the first level, a family of optimal convex combinations of simple one-dimensional LSM-regressions was generated. To achieve this an approach was used [15], which makes it possible to generate families of locally optimal convex combinations (LOCC) of one-dimensional regressions. Selected regressions were considered as new properties for the initial task. An elastic network was the second level of the proposed learning method. It was shown that the use of a two-level scheme based on weighted collective decisions over near-optimal sets of LOCCs allows one to achieve a higher generalization ability compared to the simple elastic network method. It is also possible to use arbitrary methods for organizing ensembles from other regression algorithms, (for example, combinations of "random forest", LOCC and "elastic network" or "random forest" and "elastic network").

2.2 Gluing Classifications for Regression (GCR)

Another approach to creating combined algorithms was proposed based on the method of gluing classifications for regression (GCR). Unlike the previous approach, the developed method allows to work with substances descriptions that contain various attributes types (quantitative, qualitative, ordinal and more complex). The developed algorithm introduces the degree of objects relevance to each class in the "linear corrector" regression model [16, 17]. To obtain the metric of objects relevance to each class during recognition, algorithms for calculating estimating (ACE) are used.

Two ACE models were considered in which the proximity functions were: (1) the metric function, (2) the function for arbitrary ordinal features. On an example of solving the melilite crystal lattice parameters estimation problem using the program, we compared two different methods for determining the proximity function in algorithms for calculating estimating as linear corrector classifiers. It was shown that the first model works slightly better than the second one [17].

2.3 Soft Voting Clique-Based Solvers

A peculiarity of the most prediction tasks in inorganic chemistry is the small volume of learning sets in relation to the description space dimension. To solve such problems, versions of soft voting programs of clique-based solvers modified for use in chemistry were used [18]. As a basis for research, two variants of cliques were chosen: «Syndrome Analysis» [19] and «Fragment-Potential» [20], which have a wide range of properties and capabilities for solving such kind of problems.

The first program was based on the version of the syndrome analysis algorithm (SAND). The basic idea of the algorithm: for a classified object a search is made in a learning set in a sense for the closest properties values (one or more) on the basis of which the predicted value is calculated. Proximity is determined using the aggregate set of syndromic rules (syndromes), constructed individually for each object of the learning set on the principle of "one against all the others". The syndrome rule has many symptoms at input, each of which corresponds to one of the properties.

The second program was based on the version of the voting piecewise-linear rules algorithm (FRAGMENT). The basic idea of the algorithm, as in the first program, is for a predicted object in the learning set to search in a sense for the closest values in properties (one or more) based on which the predicted value is built. Proximity is determined by voting of an aggregate set of small piecewise-linear rules, constructed individually for each object of the learning set on the principle of "one against all others". A piecewise-linear rule at the input has many initial features and is a collection of hyperplanes dividing objects into two classes in all a priori given pairs of classes. The results of applying these rules are accumulated in the “voting matrix” of size K * K, where K is the number of classes in the original data classification. Votes are summarized over the rows and columns of this matrix, allowing us to determine the integrated measures of similarity and dissimilarity of the tested object with the classes, and their relations are then used to form the final decisions about the similarity with specific ones from K classes.

The ParIS (Parameters of Inorganic Substances) system, that we developed for predicting inorganic substances physical properties, includes the above-mentioned programs developed by us and the scikit-learn software package [12].

3 System structure for inorganic substances physical properties prediction

The information base for searching the dependences of inorganic substances parameters on the properties of components in the ParIS system is the integrated database system on properties of inorganic substances and materials (DB PISM) that we created [21] (Fig. 1). It virtually unites seven databases developed in Russia and Japan, and contains information on tens of thousands of inorganic substances and materials.

The machine learning subsystem includes three components (Fig. 1):

- a subsystem for searching for dependencies between the substances' properties and components parameters (a machine learning subsystem) based on the programs we developed and the scikit-learn software package;
- prediction subsystem using the found dependencies;
- a subsystem for estimating prediction accuracy, which allows one to estimate the mean absolute and mean square errors (with cross-validation), the R^2 determination coefficient, etc., as well as construct a diagram of deviations of the calculated parameter values from the experimental ones for the substances, information about which was used for machine learning.

The dependencies obtained during machine learning process are entered into the knowledge base. They can be used to predict the parameters of substances not yet obtained in certain composition.

The prediction database contains the prediction results for substances not yet synthesized, the composition of which was set by an expert conducting machine learning. These data are also exported to special tables in "Phases" database on inorganic compounds properties [21]. This allows us to expand the functionality of the DB PISM, allowing the user to query data not only about already experimentally studied compounds, but also predictions.

The managing subsystem orchestrates the work of the information system and controls access to it from the Internet for specialists.

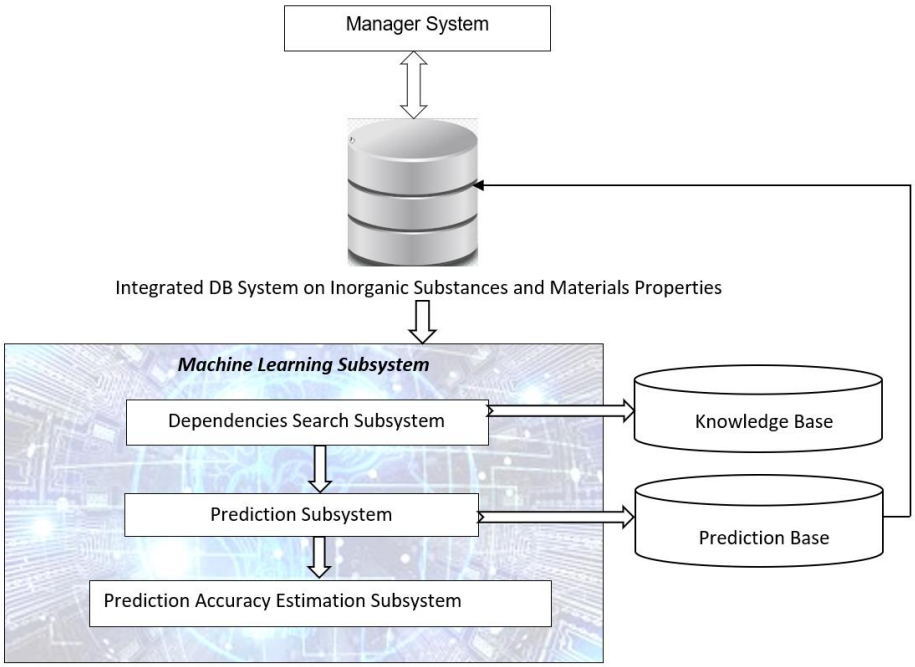


Figure 1. ParIS system structure for inorganic substances physical properties prediction.

The system is designed as an ASP.Net Core 3 Web application in C#. The Web application itself is developed using microservice architecture and it is an extensible shell that consolidates several independent calculation modules available through a single API followed by end-user interface. Calculation modules are responsible for problem solving using different mathematical methods.

Each module is implemented as an independent REST API microservice, that uses common conventions for API methods and custom configuration in JSON document, that consists of a set of mathematical algorithms parameters, suitable for a particular module. As a short example, a couple of config.json files containing default parameters are provided for K Nearest Neighbors and SVM methods (document format is defined individually for a corresponding module with respect to mathematical method's parameters):

```
{
  "int": {
    "k": {
      "type": "int",
      "min": 1,
      "max": 10,
      "default": 5,
      "fullName": "neigh-
bors count"
    }
  },
  "select": {
    "method": {
      "fullName": "K
Neighbors",
      "type": "select",
      "default": "auto",
      "options": [
        "auto",
"ball_tree",
        "kd_tree", "brute"
      ]
    }
  }
}
```

```
{
  "select": {
    "kernel": {
      "default": "rbf",
      "options": [
        "rbf", "linear",
        "poly", "sigmoid"
      ]
    }
  },
  "int": {
    "degree": {
      "default": 3,
      "min": 1,
      "max": 10
    }
  }
}
```

The modules are minimal and stateless to scale well. Every module implements a single algorithm: performs calculations on input data specified by the caller and returns a result.

The architecture allows adding modules easily to extend a set of available methods and gives a lot of flexibility, allowing us to deploy different ParIS modules on various hosts. This implies that modules can be written in any suitable programming language on any software platform. The only requirement is API accessibility via HTTP(S) endpoint. Currently, all modules (written in Python, C++, C#) run on Microsoft IIS within

a single virtual machine, but in future, they could be relocated, if required, e.g. for load balancing or other purposes.

The calculation results obtained from the calculation modules are processed by the Web-application to build a report on solving the problem by various methods or by a high-level collective solution module. Web-application contains a list of active modules in a config file which makes it easy to add module or update its default settings (all URIs are currently relative, illustrating a single host application, but in future, they could be reconfigured to reside on separate hosts referred by absolute URIs):

```
{
  "algorithms": {
    "knn": {
      "type": "classification",
      "uri": "/algorithms/knn",
      "fullName": "K Neighbors"
    },
    "svm": {
      "type": "classification",
      "uri": "/algorithms/svm",
      "fullName": "SVM"
    },
...
  },
  "combinations": {
    "average": {
      "uri": "/combinations/average",
      "fullName": "Average value"
    },
    "majority": {
      "uri": "/combinations/majority",
      "fullName": "Voting by majority"
    },
...
  }
}
```

4 The ParIS System Application for Inorganic Compounds Physical Properties Prediction

The developed system was used to predict the crystal lattice parameter of not yet obtained chalcospinels – promising materials for creating magneto-optical memory elements and sensors [22]. Predicting of crystal lattice parameters of compounds is of great interest for both chemical research and materials science investigations. Machine learning methods are widely used to solve this problem. For example, in [23-26], the crystal

lattice parameters of orthorhombic perovskites with ABO_3 composition not yet obtained were predicted using the methods of neural network training and support vector machine. Using the neural network training and regression on support vectors, it was possible to predict the crystal lattice parameters of cubic and monoclinic perovskites with ABX_3 composition (X is halogen or oxygen) [27]. The same methods and random forest learning were used to predict the crystal lattice parameters of cubic perovskites of $A^{2+}{}_2BCO_6$ composition [28, 29] and apatites [30-32]. The lattice parameters and band gap were predicted for compounds of ABX_2 composition with chalcopyrite structure using neural network training and various statistical methods (discriminate analysis, principal component analysis, etc.) [33].

First, using the information-analytical system developed by us for the computer-aided inorganic compounds design [34], new chalcospinels with $A^IB^{III}C^{IV}X_4$ compositions (A, B and C – hereinafter, various chemical elements, and X – S or Se) and $A^{II}B^{III}C^{III}S_4$ were predicted. In the first case, the sample for machine learning included information on 20 known chalcospinels of $A^IB^{III}C^{IV}X_4$ composition, 103 compounds with a crystal structure different from spinel under ambient conditions, and 10 A_2X – B_2X_3 – CX_2 systems in which compounds of $ABCX_4$ composition are not formed. For the second composition, information on 13 chalcospinels with $A^{II}B^{III}C^{III}S_4$ composition and on 20 compounds with a crystal structure different from spinel under normal conditions were selected for the learning set. This learning set was extended by examples of 48 spinels with $A^{II}B^{III}{}_2S_4$ composition, 90 compounds of this composition having a crystal structure different from spinel, and 18 AS - B_2S_3 systems in which compounds of AB_2S_4 composition are not formed. When predicting new chalcospinels, only data on the chemical elements' properties were used. According to examination prediction using cross-validation, the accuracy of new chalcospinels prediction was not lower than 80%.

Chalcospinels have a cubic crystal lattice; therefore, the only one parameter was predicted further – "*a*". Because of this parameter value is not known for all obtained chalcospinels, two learning samples for various compositions were prepared. The first one included 19 examples of the "*a*" parameter values for $A^IB^{III}C^{IV}X_4$ (X – S or Se) chalcospinels composition and the second one included 53 examples for $A^{II}B^{III}C^{III}S_4$ composition, including information on $A^{II}B^{III}{}_2S_4$ spinels composition. The feature description included 11 property values for each element that is a part of the chalcospinel, i.e. 44 features values for $A^IB^{III}C^{IV}X_4$ composition, and 33 property values for $A^{II}B^{III}C^{III}S_4$ composition. Prediction accuracy was determined by calculating the mean absolute percentage error (MAPE) and a standard mean squared error (MSE) (in the leave-one-out cross-validation mode).

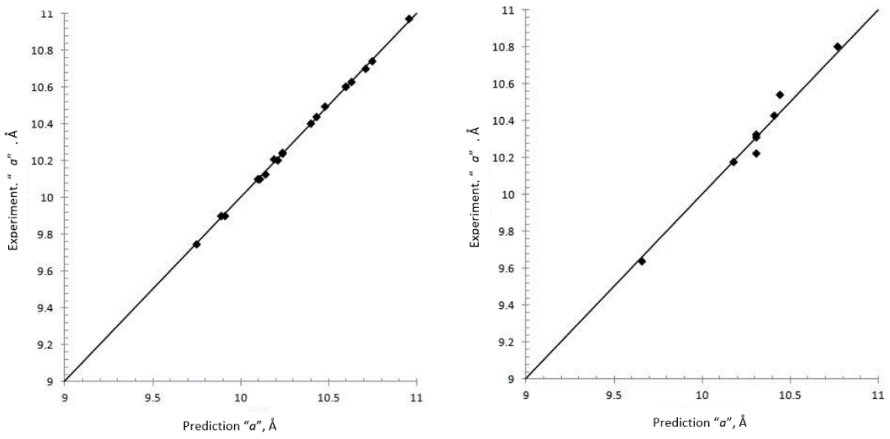
To illustrate the developed information system capabilities we used, the prediction results of the crystal lattice parameter for known chalcospinels are presented in the Table 1 (for a part of the methods with the smallest error rate values). In fig. 2, the prediction results using a multilevel approach, which is a combination of Random Forest and Elastic Net machine learning methods, are presented in graphical form. It should be noted that such a multilevel method provided the smallest prediction errors (see Tables 1 and 2). The prediction results using this method for $A^IB^{III}C^{IV}X_4$ composition is shown in Table 3.

Table 1. Crystal lattice parameter examination prediction results for $A^{I}B^{III}C^{IV}X_4$ composition chalcospinels.

Composition	**MAPE**	**0.10**	**0.10**	**0.11**	**0.09**
	MSE	**0.02**	**0.02**	**0.02**	**0.01**
	Method	**Ridge Regression**	**Bayesian Ridge Regression**	**ARD Regression**	**Random Forest + Elastic Net**
	***a*, Å (experimental)**	***a*, Å (prediction)**			
$LiInSnS_4$	10.629	10.63	10.63	10.61	10.63
$CuVTiS_4$	9.902	9.91	9.91	9.91	9.91
$CuCrSnS_4$	10.2	10.17	10.17	10.17	10.21
$CuCrTiS_4$	9.9	9.90	9.90	9.90	9.89
$CuCoTiS_4$	9.744	9.75	9.75	9.75	9.750
$CuTiZrS_4$	10.236	10.22	10.22	10.21	10.24
$CuTiSnS_4$	10.244	10.25	10.25	10.24	10.24
$CuVZrS_4$	10.209	10.15	10.15	10.15	10.19
$CuVSnS_4$	10.124	10.19	10.19	10.19	10.14
$CuCrZrS_4$	10.1	10.13	10.13	10.14	10.11
$CuCrHfS_4$	10.1	10.10	10.10	10.10	10.10
$CuInSnS_4$	10.4938	10.48	10.48	10.49	10.48
$CuCrSnSe_4$	10.7	10.67	10.67	10.68	10.71
$CuCrTiSe_4$	10.4	10.40	10.40	10.40	10.40
$CuCrZrSe_4$	10.6	10.64	10.64	10.64	10.60
$CuCrHfSe_4$	10.6	10.60	10.60	10.60	10.60
$AgCrSnS_4$	10.44	10.44	10.44	10.44	10.43
$AgInSnS_4$	10.74	10.75	10.75	10.76	10.75
$AgCrSnSe_4$	10.97	10.95	10.95	10.95	10.96

Table 2. Crystal lattice parameter examination prediction results for $A^{II}B^{III}C^{III}S_4$ composition chalcospinels.

Composition	**MAPE**	**0.18**	**0.17**	**0.18**	**0.10**
	MSE	**0.05**	**0.04**	**0.05**	**0.02**
	Method	**Ridge Regression**	**Bayesian Ridge Regression**	**ARD Regression**	**Random Forest + Elastic Net**
	***a*, Å (experimental)**	***a*, Å (prediction)**			
$MnCrInS_4$	10.4297	10.42	10.42	10.42	10.41
$FeCrInS_4$	10.323	10.30	10.30	10.30	10.31
$CoCrInS_4$	10.31	10.26	10.25	10.29	10.31
$NiCrInS_4$	10.22	10.15	10.16	10.20	10.31
$CdCrGaS_4$	10.1784	10.22	10.24	10.24	10.18
$CuCoRhS_4$	9.64	9.65	9.66	9.67	9.66
$CdSbInS_4$	10.8	10.78	10.77	10.783	10.77
$CdCrInS_4$	10.54	10.51	10.51	10.49	10.44



A b

Fig. 2. Comparison of the values predicted using multilevel predicting (random forest + elastic network) of the chalcospinel crystal lattice parameter with experimental values for $A^{I}B^{III}C^{IV}X_4$ (a) and $A^{II}B^{III}C^{III}S_4$ (b) compositions.

Table 3. Crystal lattice parameter prediction results for $A^{I}B^{III}C^{IV}X_4$ composition chalcospinels.

Composition	*a*, Å	Composition	*a*, Å	Composition	*a*, Å
$CuInTiS_4$	10.10	$AgCoZrS_4$	10.19	$CuCoZrSe_4$	10.46
$CuCoZrS_4$	9.99	$AgInZrS_4$	10.62	$CuVSnSe_4$	10.48
$CuInZrS_4$	10.41	$AgVSnS_4$	10.40	$CuCoSnSe_4$	10.43
$CuCoSnS_4$	9.95	$AgCoSnS_4$	10.22	$CuVHfSe_4$	10.52
$CuTiHfS_4$	10.23	$AgVHfS_4$	10.37	$CuCoHfSe_4$	10.46
$CuVHfS_4$	10.17	$AgCrHfS_4$	10.25	$AgCoZrSe_4$	10.71
$CuCoHfS_4$	9.98	$AgCoHfS_4$	10.17	$AgCrHfSe_4$	10.79
$CuInHfS_4$	10.39	$AgInHfS_4$	10.59	$AgCoHfSe_4$	10.70
$AgInTiS_4$	10.35	$CuCoTiSe_4$	10.26		
$AgVZrS_4$	10.40	$CuVZrSe_4$	10.53		

5 Conclusion

The ParIS system was developed for inorganic substances physical properties prediction. It allows a search for the relationships between inorganic compounds physical properties and chemical elements parameters by means of machine learning analysis of information contained in databases on inorganic substances properties. The main components of the system are an integrated system of databases on inorganic substances and materials properties developed in Russia and abroad, a machine learning-based data analysis subsystem for making predictions and a knowledge base for prediction results. The ML-subsystem includes programs based on the original algorithms developed by the authors of this paper together with methods implemented in the scikit-learn package. Using the developed system, "*a*" crystal lattice parameter values have been successfully predicted for not yet obtained chalcospinels with $ABCX_4$ composition (A, B and C are various chemical elements, and X is S or Se). During prediction chemical elements properties values were used only. Moreover, the prediction accuracy was ± 0.1 Å. Thus, it is shown that the original multilevel method developed by the authors provided the smallest predicting errors.

This work was supported in part by the Russian Foundation for Basic Research, project nos. 18-07-00080 and 20-01-00609. The study was carried out as part of the state assignment (project no. 075-00947-20-00).

References

1. Zhuravlev, Yu.I., Kiselyova, N.N., Ryazanov, V.V., Sen'ko, O.V., Dokukin, A.A.: Computer-assisted design of inorganic compounds using precedent-based pattern recognition methods. Pattern Recognition and Image Analysis, 20(4), 94–102 (2010).
2. Burkhanov, G.S., Kiselyova, N.N.: Prediction of intermetallic compounds. Russ. Chem. Rev. 78 (6), 569–587 (2009).

3. Kiselyova, N.N.: Komp'yuternoe konstruirovanie neorganicheskikh soedinenii. Ispol'zovanie baz dannykh i metodov iskusstvennogo intellekta (Computer Design of Inorganic Compounds: Use of Databases and Artificial Intelligence Methods). Nauka, Moscow (2005).
4. Draper, N., Smith H.: Applied Regression Analysis, Third Edition. John Wiley & Sons, Inc. Print ISBN: 9780471170822. Online ISBN: 9781118625590. DOI:10.1002/9781118625590 (1998).
5. Vapnik V.N.: Estimation of Dependences Based on Empirical Data.-N.Y., Springer-Verlag. Online ISBN: 978-0-387-34239-9. DOI: 10.1007/0-387-34239-7 (1982).
6. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and Computing. 14(3) (2004).
7. Chen, N. Y., Lu, W. C., Yang, J., Li, G. Z.: Support vector machine in chemistry. World Scientific Publishing Co. Pte. Ltd., Singapore (2004).
8. Ozhereliev, I.S., Senko, O.V., Kiseleva, N.N.: Methods for searching outliers objects using parameters of learning instability. Systems and Means of Informatics. 29(2), 122-134 (2019). (in russian)
9. Kiseleva, N.N., Stolyarenko, A.V., Ryazanov, V.V., et al.: Prediction of New $A^{3+}B^{3+}C^{2+}O_4$ Compounds. Russ. J. Inorg. Chem. 62(8), 1058-1066) (2017).
10. Powell, M.J.D.: Radial basis functions for multivariable approximations: A review. In: Proc. IMA Conf. on Algorithms for the Approx. of Functions and Data. Oxford University Press, Oxford. 143-167 (1985).
11. Specht, D.F.: A general regression neural network. IEEE Transactions on Neural Networks. 2(6), 568-576 (1991).
12. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in Python. J. Machine Learning Research 12(Oct.), 2825-2830 (2011).
13. The R Project for Statistical Computing Homepage: URL: https://www.r-project.org/, last accessed 30/05/2020.
14. Senko, O. V., Dokukin, A. A., Kiselyova, N. N., Khomutov, N. Yu.: Two-Stage Method for Constructing Linear Regressions Using Optimal Convex Combinations. Doklady Mathematics. 97(2), 113-114 (2018).
15. Senko, O.V., Dokukin, A.A.: Optimal convex correcting procedures in problems of high dimension. Comput. Math. Math. Phys. 51(9), 1644-1652 (2011).
16. Tkachev, Yu.I.: Methods for solving the dependence estimation problem using groups of recognition algorithms. PhD thesis. Dorodnicyn Computing Centre of RAS, Moscow (2013). (in russian)
17. Lukanin, A.A., Ryazanov, V.V., Kiselyova, N.N.: Prediction Based on the Solution of the Set of Classification Problems with Supervised Learning and Degrees of Membership. Pattern Recognition and Image Analysis. 30(1), 63-69 (2020).
18. Vaschenko, E., Vitushko, M., Dudarev, V., et al.: On the possibility of predicting the parameter values of multicomponent inorganic compounds. Information processes. 19(4), 415-432 (2019). (in russian)
19. Vitushko, M., Gurov, N., Pereverzev-Orlov, V.: A Syndrome as a Tool for Presenting Concepts. Pattern Recognition and Image Analysis. 12(2), 194-202 (2002).
20. Vaschenko, E., Vitushko, M., Pereverzev-Orlov, V.: Potentials of Learning on the Basis of Partner System. Pattern Recognition and Image Analysis. 14(1), 84-91 (2004).
21. Kiselyova, N.N., Dudarev,V.A., Stolyarenko, A.V.: Integrated system of databases on the properties of inorganic substances and materials. High Temperature, 54(2), 215-222 (2016).
22. Kiselyova, N.N., Dudarev, V.A., Ryazanov, V.V., et al.: $ABCX_4$ (X – S or Se) Chalcospinel Prediction. Perspectivnye Materialy. 7, 5-18 (2020). (in russian)

23. Aleksovska, S., Dimitrovska, S., Kuzmanovski, I.: Crystal structure prediction in orthorhombic ABO_3 perovskites by multiple linear regression and artificial neural networks. Acta Chim. Sloven. 54(3), 574–582 (2007).
24. Javed, S.G., Khan, A., Majid, A., et al.: Lattice constant prediction of orthorhombic ABO_3 perovskites using support vector machines. Comput. Mater. Sci. 39(3), 627–634 (2007).
25. Khan, A., Javed, S.G.: Predicting regularities in lattice constants of $GdFeO_3$-type perovskites. Acta Crystallogr. B64(1), 120-122 (2008).
26. Li, C., Thing, Y., Zeng, Y., et al.: Prediction of lattice constant in perovskites of $GdFeO_3$ structure. J. Phys. Chem. Solids. 64(11), 2147-2156 (2003).
27. Majid, A., Khan, A., Javed, G., Mirza, A.M.: Lattice constant prediction of cubic and monoclinic perovskites using neural networks and support vector regression. Comp. Mater. Sci. 50(2), 363-372 (2010).
28. Dimitrovska, S., Aleksovska. S., Kuzmanovski, I.: Prediction of the unit cell edge length of cubic $A^{2+}{}_2BB'O_6$ perovskites by multiple linear regression and artificial neural networks. Central Eur. J. Chem. 3(1), 198-215 (2005).
29. Majid, A., Khan, A., Choi, T.-S.: Predicting lattice constant of complex cubic perovskites using computational intelligence. Comp. Mater. Sci. 50(6), 1879-1888 (2011).
30. Kockan, U., Evis, Z.: Prediction of hexagonal lattice parameters of various apatites by artificial neural network. J. Appl. Cryst. 43(4), 769-779 (2010).
31. Legrain, F., Carrete, J., van Roekeghem, A., et al.: Materials screening for the discovery of new half-Heuslers: Machine learning versus Ab initio methods. J. Phys. Chem. 122(2), 625-632 (2018).
32. Oliynyk, A.O., Adutwum, L.A., Rudyk, B.W., et al.: Disentangling structural confusion through machine learning: Structure prediction and polymorphism of equiatomic ternary phases ABC. J. Amer. Chem. Soc. 139(49), 17870-17881 (2017).
33. Zeng, Y., Chua, S.J., Wu, P.: On the prediction of ternary semiconductor properties by artificial intelligence methods. Chem. Mater. 14(7), 2989-2998 (2002).
34. Kiselyova, N.N., Stolyarenko, A.V., Ryazanov, V.V., et al.: A system for computer-assisted design of inorganic compounds based on computer training. Pattern Recognition and Image Analysis. 21(1), 88-94 (2011).