Machine Learning Methods Application to Search for Regularities in Chemical Data

© N.N. Kiselyova[1], ©A.V. Stolyarenko[1], ©V.A. Dudarev[1,2]
[1]Institution of Russian Academy of Sciences A.A. Baikov Institute of Metallurgy and Materials Science RAS (IMET RAS), Moscow
[2]National Research University Higher School of Economics (NRU HSE), Moscow, Russia
kis@imet.ac.ru stol-drew@yandex.ru vic@imet.ac.ru

Abstract. The possibility of searching for classification regularities in large arrays of chemical information by means of machine learning methods is discussed. Tasks peculiarities in inorganic chemistry and materials science are considered. The short review of these methods applications to inorganic chemistry and materials science is presented. The system for computer-assisted inorganic compounds design based on machine learning methods has been developed. The developed system usage makes it possible to predict new inorganic compounds and estimate some of their properties without experimental synthesis. The results of this information-analytical system application to inorganic compounds design are promising for new materials search.

Keywords: machine learning, database, inorganic chemistry, design of inorganic compounds.

1 Introduction

Throughout the centuries of its evolution chemistry and materials science accumulated huge information. In common with other experimental sciences chemistry got through several stages: information accumulation, data analysis and development of classification schemes and rules that allow classifying a new object to a particular substances class. The substances division into inorganic and organic ones, Periodic table of elements, compounds classification according to crystal structure type, etc. are examples of such classifications. Essentially, in all cases these classifications are imprecise, and classes intersect partially. For example, organic chemistry is determined as the carbon compounds chemistry but carbides and carbonates belong to inorganic chemistry objects as well as boron hydrides (boranes) or silicon hydrides (silanes) which are closer to hydrocarbons (organic chemistry objects) in many properties. In large measure, it is caused by imperfection in the classification rules which were developed by chemists. One way to get around these problems in inorganic chemistry and materials science is machine learning methods application to information analysis aimed at discovery of complicated classifying regularities that allow considering of substances to particular classes. It is noteworthy that obtained regularities include substance components properties as variables, and for this reason, their use allows us to predict the class for the substances that is not yet synthesized knowing only the well-known parameters values for chemical elements forming this substance.

Half a century ago IMET pioneered in applying such approach to machine learning use to search for classifying regularities that allows a prediction of new inorganic compounds and some of their properties estimation [1, 2]. The machine learning methods application [3, 4] allowed new binary compounds prediction with 90% reliability knowing constituent chemical elements properties only. The success of approach that was put forward in IMET has given an impetus to many investigations which were connected with machine learning application to inorganic chemistry and materials science and carried-out in various countries. The investigations geography in this field is very wide: Europe, America, Asia, Africa (figure 1). The most representative teams work in Russia, the USA, and China. More detailed reviews of these researches are given in the monograph [5] and reviews [6, 7]. It should be noted that in recent years in the developed countries the governmental initiatives aimed at IT application (as well as machine learning methods) to chemistry and materials science were announced: Materials Genome Initiative (the USA) [8], Materials Research by Information Integration Initiative (Japan) [9], and Chinese Materials Genome (China) [10]. It is expected that the theoretic methods use will provide essential progress achievements in chemistry and materials science that will lead to cost reduction during new materials research, development, and production.

2 Problem Statement and Decision Methods

Suppose that every inorganic substance is described by a vector $\mathbf{x} = (x_1^{(1)}, x_2^{(1)},\ldots x_M^{(1)}, x_1^{(2)}, x_2^{(2)},\ldots x_M^{(2)},\ldots, x_1^{(L)}, x_2^{(L)},\ldots x_M^{(L)})$, where L is the number of chemical elements that form a compound and M is the number of chemical elements parameters. Each substance is also characterized by a class membership parameter: $a(x) \in \{1, 2,\ldots, K\}$, where K is the number of classes. The

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10-13, 2017

learning sample consists of N objects: $S = \{x_i, i = 1, \ldots, N\}$. We denote the learning sample objects subset from class a_j, $j = 1, 2, \ldots, K$, by $S_{aj} = \{x: a(x) = a_j\}$. The machine learning aim is to construct a classification rule that distinguishes not only different classes objects of the learning sample but also preserves prognostic ability to generate new combinations of chemical elements that were not used for learning.

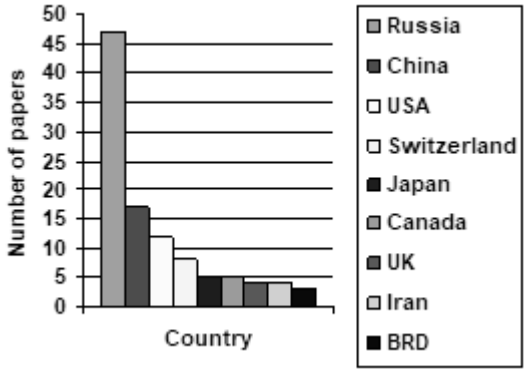

Figure 1. Distribution of publications related to machine learning methods applications to inorganic chemistry and materials science over the countries.

Among the numerous machine learning methods, various of Artificial Neural Network (ANN) learning algorithms modifications and Support Vector Machine algorithms (SVM) are the most popular (figure 2). This is due to appropriate software packages accessibility and seeming exam score accuracy (many investigators do not take into account an influence of overfitting effect on subsequent prediction reliability that is inherent in these methods).

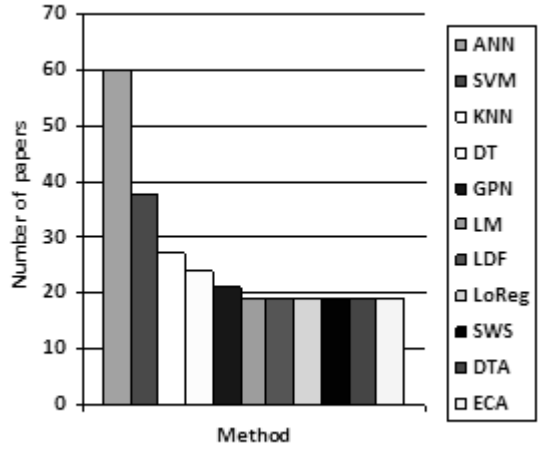

Figure 2. Various machine learning methods popularity in inorganic chemistry.

Notation: ANN – artificial neural network learning; SVM – support vector machine; KNN – k-nearest neighbors method; DT – decision trees learning; GPN – concept formation using growing pyramidal networks; LM – linear machine method; LDF – linear Fisher discriminant; LoReg – voting algorithm where estimations for classes are calculated by means of voting by logical regularities system; SWS – statistical weighted syndromes; DTA – deadlock test algorithm; ECA – estimate calculating algorithm.

A great diversity of chemical and materials science tasks were solved successfully using machine learning methods, e.g.:

theoretic tasks of prediction of:

- inorganic system phase diagram type [5, 11];
- inorganic compounds formation with certain stoichiometric composition [1, 2, 5-7, 12];
- inorganic compounds crystal structure type [5-7, 13, 14];
- some of inorganic compounds properties (melting point [15], critical temperature of superconductivity [16], band gap energy [17], enthalpy of formation [18], etc.);

technologic tasks of prediction of:

- mechanical properties of steels [19];
- acoustic properties of tellurite glasses [20];
- tribological behavior of aluminum–copper based composite [21];
- functional properties of ceramic materials [22], and so on.

3 Experience in machine learning system development for chemical applications

A special information-analytical system (IAS) that allows an automation of task solution procedure in the field of inorganic chemistry using machine learning was developed in IMET [23]. The subject field peculiarities were taken into account at the IAS creation, namely:

1) Attribute description composite structure: chemical elements (inorganic substance components) parameters set is repeated as many times as the number of elements which are included into the compound.

2) Strong correlation within set of these attributes for each component due to their dependence on common parameter - chemical elements atomic number (it follows from the Periodic Law).

3) Individual chemical elements properties give small informative gain therefore more informative parameters of single compounds (for example, single oxides, halogenides, chalcogenides, etc.) and component properties algebraic functions are widely used additionally.

4) Blanks of attributes' values that are filled by various methods including interpolation taking into account a periodicity in chemical elements properties variation with their atomic numbers.

5) Large asymmetry of learning set sizes for different classes (at that often the least of representative – as a rule newly obtained classes of substances – are the most interesting for chemists).

6) Errors and discrepancies in inorganic compounds experimental classification of learning set decreases the prediction accuracy drastically.

Machine learning procedure involves several stages:

1) objects selection for machine learning,

2) attribute description formation (including the most informative attributes selection and filling attribute values blanks also),

3) machine learning algorithms selection,

4) machine learning including application of algorithms ensembles and collective solution synthesis in a case of several algorithms usage,

5) machine learning quality estimation,

6) new objects status prediction and results interpretation.

3.1 Objects selection for machine learning

Representative and reliable set formation for machine learning preconditions subsequent prediction accuracy in a great measure. Objects selection (known inorganic substances examples) for machine learning is performed by experts in subject domain by means of information stored in data bases (DBs) on inorganic substances and materials properties including DBs that were developed in IMET [17, 23-25]. The latest include data on tens of thousands of substances and are Internet-accessible [25]. Data on substances were extracted from thousands of publications. In common with other intellectual fields papers can involve errors and inaccuracies. The experimental errors in object classification contribute significantly to prediction accuracy decrease. However, classification reliability estimation of tens of thousands of substances is massively expensive and practically impossible task. Partial automation of procedure of search for data outliers using machine learning is proposed by us. This can be best done in detecting of errors which were caused by incorrect and incomplete experimental knowledge of the class to which the substance belongs (for example, crystal structure type) as well as by erroneous property values of components which form the substance description. In the latter case errors can be incorrect experimental property value measurement result or they can be associated with incorrect interpolation in the case of filling attribute values blanks as well. The machine learning results analysis allows detection of substances which fall within another class and provision for chemist with information on substance expert assessment and making a decision for its status. The problem solution principal possibility is specified by the subject domain specific that is connected with inorganic compounds properties variation periodicity depending on atomic number of elements – the chemical system components.

3.2 Attribute description formation

Attribute description formation problem is complicated and hard-to-solve task of modern machine learning theory. There are a large number of approaches which have proved their effectiveness at various task types solution. However, it is impossible to evolve a surely optimal universal method of attributes selection. In this regard few alternative methods with subsequent collective decision synthesis are used by us for attribute selection. 2D-projections visualization tools are applied additionally for points corresponding to certain type compounds in chemical elements properties space. The parameters set includes not only initial attributes but also the algebraic functions of these attributes which are selected by user.

3.3 Machine learning algorithms selection

The IAS includes a set of machine learning algorithms which are the most popular among chemists (figure 1). At present time IAS involves the following software: programs based on well-known linear machines methods, Fisher linear discriminant, k-nearest neighbors, support vector machine, neural-network algorithms, and also algorithms which were developed by the Computing Centre, Russian Academy of Sciences and based on estimates calculation, deadlock tests voting algorithms, logical regularities voting algorithms, weighted statistical voting algorithms, etc. [26]. IAS includes also the ConFor system for machine learning according to procedure for concept formation, developed by the Institute of Cybernetics, National Academy of Sciences of Ukraine [27]. This system is built upon computer memory data arrangement in the form of growing pyramidal networks. At solution of each task at hand a selection of the most exact machine learning algorithms is carried out for subsequent use in decision making and prediction procedures.

3.4 Machine learning

Our experience in inorganic chemistry prediction tasks solution shows [6, 7, 12, 17, 23, 24] that algorithms ensembles application allows a considerable increase of accuracy in inorganic compounds prediction. In decision making process the most accurate machine learning algorithms are used that were selected on the previous stage. The IAS includes the following programs realizing various collective decisions strategies, which are based on Bayes method, clustering and selection methods, decision templates, logical correction, convex stabilizer method, Woods dynamic method, committee methods, etc. [26].

3.5 Machine learning quality estimation

The cross-validation on learning set objects is the most widely used universal and reliable tool for machine learning quality estimation. IAS contains special software for this procedure realization that is used in the best machine learning algorithms selection. However, an attempt of cross-validation application to machine learning accuracy estimation at use of algorithms ensembles as optimizable criterion results the loss of estimate unbiasedness. In this case, there is a certain overfitting risk. In this regard, the traditional approach to collective algorithms accuracy evaluation using examination recognition of N examples chosen randomly from learning samples and unused in learning (at the final prediction stage, reference examples are returned to the learning set) is applied. The corresponding program was included to IAS.

The learning set sizes asymmetry for different classes is an important problem at machine learning

accuracy estimation. Naturally in this case the generalized examination recognition accuracy does not represent the prediction error for small classes, therefore the ROC curves application is appropriate to different algorithms prediction quality analysis. ROC curves allow recognition accuracy comparison for the targeted and alternative classes at variation of cut-offs which identifies belonging to different classes.

It should be pointed out that machine learning quality estimation procedure belongs to yet hardly unsolved machine learning task. Some algorithms (SVM, ANN, etc.) characterized by overfitting effect, show high examination recognition accuracy often but this fact does not always provide high predicting reliability for new objects.

3.6 Prediction of new inorganic compounds formation and some of their properties estimation

To increase predicting accuracy in the case of learning sets with K classes (K > 2) the following method is used. Firstly, multi-class learning and prediction are carried out. Next, K dichotomies are calculated: the targeted class and all the alternative classes, followed by subsequent K predictions. The results of multi-class prediction and dichotomies series are intercompared, and if the predictions are not contradictory the decision on the object status is made. The special tools for collective decision formation based on comparison of multi-class prediction results and dichotomies series were developed. The efficiency of such approach that allows to increase prediction accuracy was approved during numerous tasks solution [5-7, 12, 17, 23, 24].

4 IAS application illustration to regularities search in chemical information

The machine learning application allowed a search for inorganic compounds formation regularities, a prediction of thousands not yet synthesized substances and some their properties estimation using obtained regularities. This approach efficiency to inorganic compounds design can be illustrated by comparison of the predictions results with newer experimental data obtained after publication of our predictions [12].

The table contains AB_3X_3 compounds formation possibility predictions in the A_2X_3–B_2X systems (A and B are various elements, and X = S, Se, or Te) under normal conditions, which could be promising for search for new semiconductor, nonlinear optical, electro-optical, and acousto-optical materials. Experimental information on 117 examples of AB_3X_3 compounds formed and 58 examples when no such composition compounds were formed in the A_2X_3–B_2X systems under normal conditions was used for computer analysis. To describe the compounds in computer memory we selected A, B, and X elements properties (the melting and boiling points; covalent, ionic (by Bokii and Belov), and pseudopotential (by Zunger) radii; the first three ionization potentials; electronegativity (by Pauling); the standard enthalpies of atomization and evaporation; thermal conductivity; molar heat capacities, etc.), simple A_2X_3 and B_2X chalcogenides properties (standard entropy and enthalpy), and some algebraic functions of these properties (for example, the ratio of the covalent radius to the metal radius for elements A, B, and X). The table 1 presents predictions examples for the AB_3X_3 compounds and their experimental verification results. The following notation is used: 1, the prediction of AB_3X_3 formation under normal conditions; 2, the prediction of AB_3X_3 absence under normal conditions; #, examples, the information on which is used for machine learning; empty cells, uncertain prediction; ©, the prediction of AB_3X_3 formation matches new experimental data; and Θ, the prediction of compound absence matches experimental data. All 27 tested predictions coincided with the experimental data.

Table 1. AB_3X_3 compounds formation possibility prediction

A B	Fe	Ga	In	Sn	Sb	La	Ce	Pr	Nd	Sm	Eu	Gd	Tb	Dy	Bi
X = S															
K	©	#2	1	1	#1	1		1	1	1	1	1	1	1	#2
Rb	©		#1	1	©	1			1	1	1	1	1	1	#1
Tl	1	Θ	#1	#1	#1	#2	2	#2	Θ	2	2				
X = Se															
K	#1	#1	1	©	#1		1			1	1	1	1	1	#1
Rb	1	1	1	1	©	1	1			1	1	1	1	1	#1
Ag	2	#2	Θ		#2	Θ	Θ	Θ	2	Θ	Θ		Θ	2	Θ
Cs	1	#1	1	1	©	1				1	1	1	1	1	#1
Tl	1	Θ	#2	#1	#1	2	2	2	2						#2
X = Te															
Rb	1	1	1	©	1	1	1			1	1	1	1	1	1
Ag	2	#2	#2	Θ	#2	2	2	2	2	2	2	#2	Θ	Θ	Θ
Cs	1	1	1	©	1	1	1			1	1	1	1	1	1
Tl	©	Θ	Θ	#1	#2	#2	2	2	Θ						#2

Conclusions

During half of the century the predictions of thousands of inorganic compounds in binary, ternary and more complicated chemical systems were obtained and some their properties (melting point, critical temperature of superconductivity, band gap energy, etc.) were estimated in IMET [1, 2, 5-7, 12, 16, 17, 23, 24]. The obtained predictions usage allows an essential progress provision in a search for new magnetic, semiconductor, superconductor, nonlinear optical, electro-optical, acousto-optical and other materials. Hundreds of predicted compounds were synthesized and our results experimental verification shows that the average prediction accuracy is higher than 80% [2, 5-7]. Machine learning methods application to search for regularities in big chemical data gives an opportunity for theoretic design of new inorganic compounds that allows substantially reduce the costs for search for new

materials with predefined properties, replacing them by computations. It is important to note that only information on components properties (chemical elements or more simple compounds) is used in prediction process.

This work was partially supported by the Russian Foundation for Basic Research (project nos. 16-07-01028, 17-07-01362, and 15-07-00980). We are grateful to V.V. Ryazanov, O.V. Sen'ko and A.A. Dokukin for long-term help and collaboration.

References

[1] E. M. Savitskii, Yu. V. Devingtal', and V. B. Gribulya. Prediction of metallic compounds with composition A3B using computer. Dokl. Akad. Nauk SSSR (English translation - Doklady Physical Chemistry), 183(5), p.1110-1112, 1968

[2] E. M. Savitskii and V. B. Gribulya. Application of computer techniques in the prediction of inorganic compounds. New Delhi-Calcutta: Oxonian Press Pvt., Ltd. 1985.

[3] Yu. V. Devingtal'. About optimal coding of objects at their classification using pattern recognition methods. Izvestiya Akademii Nauk SSSR. Tekhnicheskaya Kibernetika, 1, p.162-169, 1968.

[4] Yu. V. Devingtal'. Coding of objects at application of separating hyper-plane for their classification. Izvestiya Akademii Nauk SSSR. Tekhnicheskaya Kibernetika, 3, p.139-147, 1971.

[5] N.N. Kiselyova. 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). Moscow: Nauka, 2005.

[6] G.S. Burkhanov and N.N. Kiselyova. Prediction of intermetallic compounds, Russ. Chem. Rev., 78(6), p. 569-587, 2009.

[7] N. Kiselyova, A. Stolyarenko, V. Ryazanov, et al. Application of Machine Training Methods to Design of New Inorganic Compounds. In Diagnostic Test Approaches to Machine Learning and Commonsense Reasoning Systems. Ed. By X.A. Naidenova & D.I. Ignatov. Hershey: IGI Global, p. 197-220, 2012.

[8] Site of Materials Genome Initiative: https://www.mgi.gov/ .

[9] Site of Center for Materials Research by Information Integration: http://www.nims.go.jp/eng/research/MII-I/index.html .

[10] X.-G. Lu. Remarks on the recent progress of Materials Genome Initiative, Sci. Bull., 60(22), p.1966–1968, 2015.

[11] P. Villars, K. Brandenburg, M. Berndt, et al. Binary, ternary and quaternary compound former/nonformer prediction via Mendeleev number, J. Alloys and Compounds, 317-318, p.26-38, 2001.

[12] N.N. Kiselyova. Prediction of Formation of AB3X3 (X = S, Se, Te), Inorg. Mater., 45(10), p.1077-1080, 2009.

[13] A.O. Oliynyk, E. Antono, T.D. Sparks et al. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds, Chem. Mater., 28(20), p.7324−7331, 2016.

[14] G. Pilania, P.V. Balachandran, J.E. Gubernatis, and T. Lookman. Classification of ABO3 perovskite solids: a machine learning study, Acta Crystallogr., B71(5), p.507-513, 2015.

[15] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka. Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids, Phys. Rev., B89(5), p.054303/1-9, 2014.

[16] E.M. Savitskii, V.B. Gribulya, and N.N. Kiselyova. Cybernetic prediction of superconducting compounds, CALPHAD, 3(3), p.171-173, 1979.

[17] N.N. Kiselyova, V.A. Dudarev, M.A. Korzhuyev. Database on the Bandgap of Inorganic Substances and Materials, Inorganic Materials: Applied Research, 7(1), p. 34-39, 2016.

[18] S.P. Sun, D.Q. Yi, Y. Jiang, et al. Prediction of formation enthalpies for Al2X-type intermetallics using back-propagation neural network, Mater. Chem. and Phys., 126(3), p. 632–641, 2011.

[19] A. Bahrami, A. S. H. Mousavi, and A. Ekrami. Prediction of mechanical properties of DP steels using neural network model, J. Alloys and Compounds, 392(1-2), p.177-182, 2005.

[20] M.S. Gaafar, M.A.M. Abdeen, and S.Y. Marzouk. Structural investigation and simulation of acoustic properties of some tellurite glasses using artificial intelligence technique, J. Alloys and Compounds, 509, p. 3566-3575, 2011.

[21] M. Hayajneh, A.M. Hassan, A. Alrashdan, and A.T. Mayyas. Prediction of tribological behavior of aluminum–copper based composite using artificial neural network, J. Alloys and Compounds, 470, p. 584-588, 2009.

[22] D.J. Scott, P.V. Coveney, J.A. Kilner, et al. Prediction of the functional properties of ceramic materials from composition using artificial neural networks, J. Eur. Ceram. Soc., 27(16), p. 4425–4435, 2007.

[23] N.N. Kiselyova, A.V. Stolyarenko, V.V. Ryazanov, et al. A system for computer-assisted design of inorganic compounds based on computer training, Pattern Recognition and Image Analysis, 21(1), p. 88-94, 2011.

[24] N.N. Kiselyova, V.A. Dudarev, and V.S.Zemskov. Computer information resources in inorganic

chemistry and materials science, Russ. Chem. Rev., 79(2), p. 145-166, 2010.

[25] Site of IMET RAS DBs: http://imet-db.ru .

[26] Yu. I. Zhuravlev, V. V. Ryazanov, and O. V. Sen'ko. RECOGNITION. Mathematical methods. Software system. Practical solutions. Moscow: Phasis. 2006.

[27] V. P. Gladun. Processes of formation of new knowledge. Sofia: SD "Pedagog 6". 1995.