Information-Analytical System for Inorganic Compounds Design

Nadezhda KISELYOVA, Andrey STOLYARENKO, Victor DUDAREV

A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, Moscow

1 Introduction

The methods of search for complex regularities in the data known as Knowledge Discovery in Databases (KDD) now advance to the forefront both on tools development and on the applications [1, 2]. The cause to this is development of integrated systems of databases, such as data warehouses, and development of intelligent data analysis methods, which were designated by the term “data mining”. Data mining systems combined with databases (DBs) are a KDD-systems core. Now the latter have included algorithms of applied statistics methods, pattern recognition, artificial intelligence, etc.

In present paper organization principles and the special KDD-system development results - information-analytical system (IAS) - in the field of inorganic chemistry and materials science are considered. This system integrates DBs with the heterogeneous information using different Database Management Systems (DBMSs) and computer platforms, and the subsystem of search for regularities that unites the programs based on various pattern recognition algorithms.

2 Structure of the Information-Analytical System

The information-analytical system for inorganic compounds design includes an integrated subsystem of DBs for the properties of inorganic substances and materials, a subsystem for data analysis, based on computer training, a subsystem of selecting the most important attributes, a visualization subsystem, a knowledge base, a predictions base for various inorganic substances classes, and a managing subsystem (figure 1).

2.1 Integrated Databases Complex for the Properties of Inorganic Substances and Materials

Databases with the large volumes of the data are information basis of KDD-systems. Often the integrated DBs systems are this basis. The information traffic schema in KDD-system substantially depends on a DBs integration method. Principally there are three approaches to database integration: data warehouses **-** ETL (Extract, Transform, Load) [3], live data integration systems - EII **(**Enterprise Information Integration) [4], and systems of applications integration - EAI (Enterprise Application Integration) [5].

Last two approaches are used in a case, when DBs are developed and are maintained by means of different DBMSs and computer platforms, thus they contain the data, for example, measured with different accuracy or/and by different methods. The complete and smooth integration of these systems is impossible because of a different level in data quality stored in various DBs, and problems, which inevitably will arise at change of the procedure of work with the integrated system. Such problems arise at integration DBs on properties of inorganic substances and materials [6].

The information traffic procedure in KDD-system is the simplest and the most logical in case of information extraction for the computer analysis from Data Warehouse. If the DBs system with heterogeneous data, which is integrated on the basis of the EII or EAI approaches, is an information kernel of KDD-system, the information interchange procedure in many aspects will become more complicated. In this case automatic data extraction for the analysis is possible only from one of DBs, included into the integrated system. The expert estimation in a certain subject domain is necessary for decision making about data association from different DBs, distinguished, for example, in accuracy and measurements methods. Hence subsystem ensuring data estimation realization process must be developed.

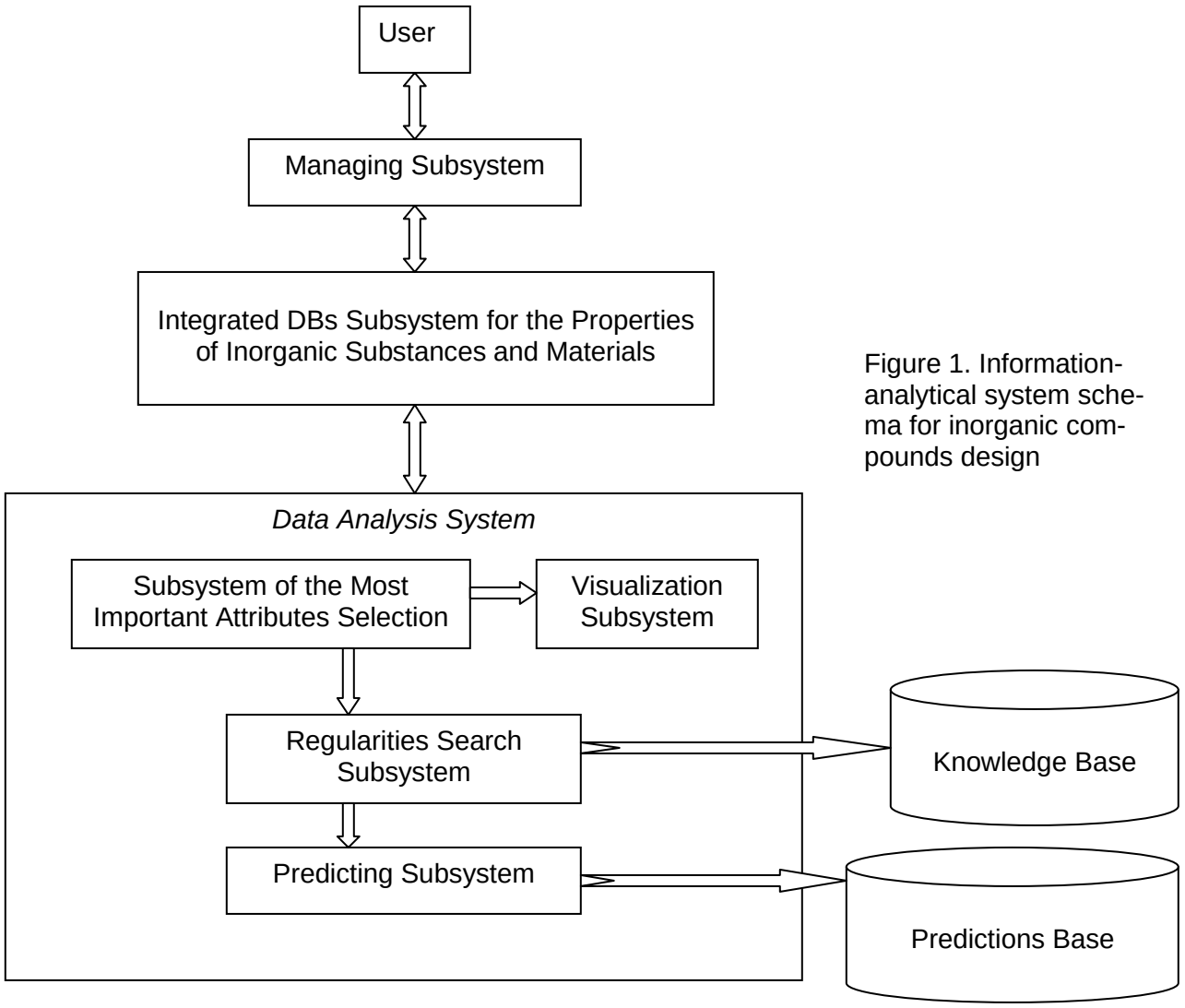

Figure 1. Information-analytical system schema for inorganic compounds design

The KDD-system information kernel, developed by us, is a DBs set for properties of inorganic substances and materials - now integrates five databases:
(1) The DB “Phases” for the properties of inorganic compounds [7, 8] now containing information on more than 49 000 ternary compounds (i.e., compounds made up of three chemical elements) and more than 27 000 quaternary compounds, extracted from more than 28 000 publications.
(2) The DB “Elements” for the elements chemical properties [7] containing data for more than 90 parameters.
(3) The DB “Diagrams” for semiconductor systems phase diagrams [7-9] containing the information on phase diagrams of semiconductor systems and the physical and chemical properties of phases forming in those, collected and evaluated by experts. At present this DB comprises the detailed information on several tens of systems that are the most important in semiconductor electronics.
(4) The DB “Crystal” for the properties of acousto-optical, electro-optical, and nonlinear optical substances [7-9] now containing the information on the parameters of more than 140 materials.
(5) The DB “Bandgap” for the forbidden band width of inorganic substances [7] currently containing the information on more than 3000 substances.

The total DBs size is about 8 GBytes. The integrated DBs subsystem allows specialists to gain aggregate information on the properties of substances and materials from different databases at one time. Authorized users can access the DBs subsystem via Internet (http://imet-db.ru).

The databases were developed with use of different DBMSs (MS SQL Server and Oracle), on different computer platforms (Intel and Sun), with various operational systems (MS Windows and Solaris) and contain the information on different substances properties. The external DBs keys are the substance chemical composition (a set of chemical elements and their quantity).

2.2 Regularities Search Subsystem

Major development task of this subsystem is the data mining methods selection, that is the most suitable for regularities search and taking into account the certain subject domain. At selection of pattern recognition methods intended for the analysis of the chemical information the long-term experience of these methods application to inorganic compounds design was taken into account [8]. As the result the following methods and programs were chosen:

- a wide class of algorithms of system RECOGNITION developed by A.A. Dorodnicyn Computer Center of Russian Academy of Sciences (CCAS) [10]. This multifunctional pattern recognition system includes the well-known methods of *k*–nearest neighbors, Fisher's linear discriminant, linear machine, multi-level perceptron (neural networks), support vector machine, genetic algorithm, and the special algorithms which were developed by CCAS: estimates calculation algorithm, LoReg (logical regularities of recognized objects), deadlock test algorithm, statistical weighted syndromes, etc.;
- a system of concept formation ConFor developed by V.M. Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine. The system is based on special data structure in a computer memory named as growing pyramidal networks [11].

As a rule, beforehand it is impossible to specify, what algorithm is most effective at the decision of a certain task. In this connection a use of recognition methods by collectives of algorithms is very promising. In many cases at synthesis of the collective decision it is possible to compensate recognition errors caused by separate algorithm by the correct predictions of other algorithms. From the above reasoning the developed IAS was supplemented by software subsystem realizing different strategy of the collective decision making based on: the Bayesian method, clustering and selection, decision templates, convex stabilizer, Woods dynamic method, committee methods, etc. [10].

The above-stated pattern recognition algorithms and collective decision making are based on various principles and use various representation forms of required regularities (algebraic functions, logic expressions, neural or growing pyramidal networks, etc.). The service-oriented architecture (SOA) was used for integration of such heterogeneous programs. It has allowed taking into account differences in the data and information structures used in the programs to be integrated, the complicated mechanisms of their interaction and were supplied with an opportunity of quite simple addition of the new data analysis programs into a subsystem of search for regularities. At integration of the applications the special middleware was applied instead of the specialized interfaces between the separate programs. It plays the role of a universal program clue connecting all applications [12, 13]. Advantage of used technology on the basis of integrating middleware is, first of all, simplicity of support and system expansion, developed on its basis.

2.3 Subsystem of the Most Important Attributes Selection

The selection of chemical elements properties, providing the most important information for the substances classification, is of double significance. On the one hand, it enables a drastic reduction of attribute description that includes hundreds of elements' properties for multi-component substances. On the other hand, the selection of the most important elements properties in chemical substances classification affords physical interpretation of the resulting classifying regularities improving the confidence in the obtained predictions and makes it possible to find substantial causal relationships among the parameters of subjects and to develop physical and chemical phenomena models. This subsystem is based on a minimization procedure of generalized error functionals for convex correcting procedures with respect to ensembles of predictors constructed on the individual attributes basis [14, 15].

2.4 Visualization Subsystem

It allows an illustration of attribute selection results. The information about chemical compounds properties is represented in habitual for the chemists and materials scientists form: as projections of the points corresponding to certain type compounds in chemical elements properties space (or more simple compounds – oxides, chalcogenides, halogenides, etc. of elements). The visualization subsystem is intended for information representation about the properties coordinates for components included into compounds. Use of algebraic functions of these properties is possible, for what the special complex attributes formation subsystem was developed.

2.5 Knowledge Base

This subsystem contains the discovered regularities, which can be used for not yet synthesized substances prediction and their properties estimation when there is no information on a certain chemical system in databases. The regularities are stored in the tasks base in the intrinsic format of those software products for the data analysis by whose means they were obtained. Such implementation makes it possible to integrate new software products for the data analysis into the IAS and resolves the problem associated with the fact that the representation forms of the resulting regularities in the computer training methods used are substantially different. By a task, it is meant the training procedure by the selected methods on a particular training sample. Here it is suggested that not the results of training, as such (like logical expressions or the structure of a trained neural network), but so-called labels for the tasks should be stored in the tasks base. The term label is taken to mean the necessary information for the task, which permits distinguishing this task from others. The following information on the task is stored in the IAS: the unique task number; the training sample in standard format; data for the attributes used to form the training sample; the identifier of the software product for the data analysis by whose means the regularities were obtained; the methods list involved in training, with their parameters; information on the quantitative and qualitative compounds composition used in training; the compounds' parameter identifier to be predicted; etc.

2.6 Predicting Subsystem

Only component' properties and found regularities are used in this subsystem for new compounds prediction.

2.7 Predictions Base

The predictions, as well as references to service information from the knowledge base are stored in the predictions base. The use of this base made it possible to improve the functionality of the IMET RAS DB for the properties of inorganic substances and materials "Phases" through providing the user not only with the available information on substances that have already been studied, but also with inorganic compounds predictions which have not yet been obtained and their properties estimations.

2. 8 Managing Subsystem

It organizes the computing process and carries out interaction among all IAS functional subsystems, as well as provides access to IAS via Internet (http://ias.imet-db.ru). Besides, the managing subsystem provides user with software for the preparation for data analysis, produces reports in the form habitual for chemists, and provides other service functions. In particular, a special subsystem has been designed to retrieve the DB information that, after its estimation by chemist, is used to train the computer, and to prepare this information for the subsequent analysis. The subsystem allows the chemist to edit found information and to form an attribute description of compound, which is a complex description made up of few chemical elements parameters included into its composition. A chemist selects the chemical elements properties in a training sample form, and the subsystem for training set preparation retrieves the chosen elements' properties values from the DB "Elements", makes up complex attributes as algebraic functions of the initial elements parameters when needed, and merges the attribute description to produce a table that is there upon passed to the predicting subsystem input. The subsystem for results generation is intended for presenting predictions in the tabular form customary among chemists and material scientists.

3 The Prediction Results of New Inorganic Compounds

Now the IAS is the main tool for predicting new inorganic compounds in our investigations. Some of potentialities of this system are demonstrated below.

3.1 New Inorganic Compounds with Composition ABC_2 (C = S, Se, or Te) Prediction

Chalcogenides with ABC_2 composition (A and B – various elements; C = S, Se, or Te) are a promising semiconductor materials class. Materials for solar batteries were developed on the basis of chalcopyrites $CuInS_2$, $CuInSe_2$, and $CuGaSe_2$ [16]. The compound $AgSbTe_2$ is a promising thermoelectric material for use at average temperatures (300–600°C) [17]. However, nonlinear-optical chalcogenides applications with composition ABC_2 are the most interesting [18-20]. In connection with the prospect of this composition compounds practical use, it was obviously important to design their unstudied analogs [21].

Data on 158 examples of the ABC_2 (C = S, Se, or Te) compounds formation with the α-$NaFeO_2$ (space group R3(–)m) crystal structure under ambient conditions (298 K and 1 atm), 44 compounds with NaCl structure type (space group Fm3m), 47 compounds with chalcopyrite structure (space group I4(–)2d), and 24 compounds with TlSe structure (space group I4/mcm) were used for the data mining.

In computer training the inorganic substances are represented in the computer memory as a set of chemical element properties values. This set included the following parameters: the pseudopotential radius (according to Zunger); the melting and boiling points, Debye temperature (only for A and B elements); the first ionization potential (I_1); the second ionization potential (I_2); the third ionization potential (I_3); the distances to valence and core electrons (according to Schubert); the regular number (according to Mendeleev–Pettifor); the electronegativity (according to Pauling); the Miedema chemical potential (only for A and B); the quantum number; the thermal conductivity; the group number in the periodic system (only for A and B); the molar heat capacity; the enthalpies of melting, vaporization, and atomization; the solid state phase entropy; formal valence (only for A and B); ionic radius (by Bokii and Belov); covalent radius (R_{cov}); functions: $(R_{covA} – R_{covB})/R_{covC}$, $(I_{1A} – I_{1B})/I_{1C}$, $(I_{2A} – I_{2B})/I_{2C}$, $(I_{3A} – I_{3B})/I_{3C}$, $(X_A – X_B)/X_C$ (X is electronegativity by Martynov–Batsanov), etc. The initial information (training set) for computer analysis is a matrix, each line of which corresponded to a set of elements properties values of forming a known chemical system and indicated to what systems class it belongs.

Using collectives of pattern recognition algorithms included into IAS we had found classifying regularities and predicted new compounds with composition ABC_2 (Table 1). The following designations were used: 1 - prediction of the α-$NaFeO_2$ crystal structure type for ABC_2 compound ; 2 - prediction of the NaCl structure type; 3 - prediction of the chalcopyrite structure; 4 - prediction of the TlSe structure type under the ambient conditions (298 K and 1 atm); 5 - prediction of the structure distinct from the ones mentioned above; 6 - prediction of the absence of compound with composition ABC_2; the symbol # is used for known substances for the computer training; empty entries indicate predictions discrepancies using different collective methods or uncertain prediction.

In more than two years 16 our predictions were checked experimentally (they are marked by signs * or @) and have found that the calculated results do not coincide with experimental data in only one case ($RbYTe_2$).

Table 1. Predictions of Compounds with Composition $A^{I}B^{III}C_2$

C	S								Se								Te							
A / B	Li	Na	K	Cu	Rb	Ag	Cs	Tl	Li	Na	K	Cu	Rb	Ag	Cs	Tl	Li	Na	K	Cu	Rb	Ag	Cs	Tl
B	#5	#5	#5	3	#5	3	#5	#5	1	1		#5	1	3	1	4	1	1	4		1		1	4
Al	#5			#3		#3		4	#5	#4	#5	#3		#3		#5		#4	#4	#3		#3	4	#4
P			5	#5		#5	5	#5	2	1	1	#5	1		5	#5	2		4	3	5		5	4
Sc	#1	#1	1	#5	1			#1	1	1	1		1	#5	1		1		1		1		1	
Ti	#5	#1	#1		#5			#4	#5	#1	1	3	1	3	5	#4		4	4		5	3	5	4
V	#5	#1	1		1				#5	#1	1	3	#5	3	5		#5	#5			5		5	4
Cr	#5	#1	#1	#5	#1	#5		#5	2	#1	1	#5	#1	#5	5				5	3	5	#5	5	#5
Mn	5		1		1				#5	#5	1		#5	3	5		#5	#5	#5		#5	#5	#5	
Fe	#5		#5	#3	#5	#3		#5		1	#5	#3	#5	#3		#5			1	#3	1	#3	1	5
Co	5								2	1	1	3	1	3	5	5	2			3	5	5	5	
Ni	5	1		#5						#2	1		1	#5	5				1	3	5	#5	5	
Ga	#5	#5	#5	#3	#5	#3	5	#5	5	4	#5	#3	4	#3		#5	#3	#4	#5	#3		#3		#4
As		#5		#5		#5	5	#5		#5	#5	#5	#5	#2	#5	5	2		4	#6	5	#2	5	
Y	#2	#1	#1	5	1	#5	5	#1	#1	#1	1		1	#5	1	#1	1	1		#5	1@	#5	1	#1
Mo	#5	1	1						2	1	1	3	1	3	5	5			4	3	5		5	
Rh		1	1	3					2		1	3	1	3	5		2		4	3	5	3	5	
In	#5	#1	#5	#3	#5	#3	#5	#4	#5	#1	5	#3	#5	#3		#4	#3	#4	#4	#3		#3	4	#4
Sb		5	#5	#5	#5	2	5*	2*	2	5	#5	#5	#5	#2		5	#2	#2	4	#5	#5	#2	#5	#1
La		#2	#1		#1		#1	#6	#5	#1	1*	#5	#1		1	#6	2		1*		1		1	6
Ce		#2	#1	#5	#1		#1	#6	#5	#1	1	#5	#1	#6	1	#6			1*		1*		1	#6
Pr	#2	#1	#1	#5	#1		#5		1	#1	1	#5	#1	#6	1	#1	2		1*	#5	1		1	#1
Nd	#2	#1	#1	#5	#1		#5			#1	1		#1		1	#1	1	1	1*	#5	1*		1*	#1
Pm		1	1	5	1	5		1	1	1	1		1		1	1	1	1			1		1	
Sm	#2	#1	#1	#5	#1	#5	#5	#1		#1	1	#5	#1	#6	1	1		1	1*		1*		1	#1
Eu	#2	#1	#1	#5	#1	#5	#5	#1		#1	1	#5	1	#6	1	#6		1	1		1		1	
Gd	#2	#1	#1	#5	#1	#5	#5	#1	#1	#1	1	#5	#1	#5	1	#1	1	1	1*		1*	#5	1	#1
Tb	#2	#1	#1	#5	#1	#5	#5	#1	#1	#1	1	#5	#1		1	#1	1	1			1	5*	1	#1
Dy	#2	#1	#1	#5	#1	#5	#5	#1	#1	#1	1		1	#5	1	#1	1	1			1	#5	1	#1
Ho	#1	#1	#1	#5	#1	#5	#5	#1	#1	#1	1		#1	#5	1	#1	1	1			1	#5	1	#1
Er	#1	#1	#1	#5	#1	5	#5	#1	#1	#1	1		#1	#5	1	#1	1	1	#1		1	#5	1	#1
Tm	#1	#1	#1	#5	#1	#5	#5	#1	1	1	1		1	#5	1	#1	1	1			1	#5	1	#1
Yb	#1	#1	#1	#5	#1	#5	#5	#1		#1	#1		1	#5	#5	1	2	1	1	#5	1		1	
Lu	#1	#1	#1	#5	#1	#2	#5	#1	1	1	1		#1	#5	1	#1	1	1			1	#5	1	#1
Tl	2		#5	#3	#5	3	#5			#4	#4	#3	4	#3				4	4	#3	4	#3		4
Bi	#2	#2	#2	#5	#1	#1		#1	#2	#2	#2	#2	#5		#5	#2	#2	#2		#5		1		#1
Th	2	1	1		1		1	1	1	1	1		1		1	1	1	1	4		1		1	4
Pa	2	1	1		1		1	1	1	1	1		1		1	1	1	1			1		1	
U			1		1		5	1			1		1		5						5		5	
Np					1			1	1	1	1		1		1				4		1		1	4
Pu	2		1		1			1	1	1	1		1		1	1			4		1		1	

3.2 Estimation of the Physical Properties of Inorganic Compounds

The prediction of the numerical intrinsic physical properties (for example, compound melting point at atmospheric pressure, critical temperature of transition to superconducting state, etc.) is the most difficult problem at inorganic compounds computer-assisted design using computer training. In this case, only a threshold estimation of the property (more or less than a threshold) is possible. We succeeded in some physical properties estimation of inorganic compounds (the critical temperature of transition to superconducting state [22], the melting point and bandgap [23, 24], etc.). In particular we predicted wide bandgap (ΔE) semiconductors with ABC_2 composition and chalcopyrite crystal structure using IAS [23, 24].

The following chemical elements parameters were used for the chemical compounds description:

- the electronegativity in the Martynov-Batsanov scale of values;
- formal valence Z_A, Z_B, Z_C (for transition metals, the group number of elements is used as Z);
- mean Born exponent $\bar{n} = \frac{n_A + n_B + 2n_C}{4}$;
- the number of Pettifor;
- the proportion: $(I_z / Z)_{AC} = (\frac{I_z}{Z})_A - \left[6 + 0.1(\frac{I_Z}{Z})_C \right]$), where I_z - final ionization potential;
- atomic radius.

Data for computer training has been extracted from database “Bandgap”. The two classes were considered: 1 – chalcopyrites with ΔE>2 eV and 2 – chalcopyrites with ΔE<2 eV. Table 2 contains experimental values and band gap examination prediction results of known chalcopyrites using separate pattern recognition methods.

The following methods were used: **EC** - estimate calculation algorithm, **LDF** - Fisher's linear discriminant, **LM** - linear machine, **MR** - logical regularities algorithm, **NN** - neural networks, **KNN** – *k*–nearest neighbors, **SVM** - support vector machine, **SWS** - statistically weighted syndromes, **TA** – deadlock test algorithm, **LG** – logical regularities of recognized objects, **DT** – binary decisive trees method, **CF**-ConFor, **SVR** - support vector regression. Recognition uses cross-validation procedure (excepting CF and SVR). The best results were achieved using logical regularities algorithms, linear machine and ConFor. The bandgap prediction of new chalcopyrites was carried out using results of logical regularities algorithms and linear machine on the basis of the following collective methods application to making a decision: **BM** - Bayesian method, **C&S -** clustering and selection, **DT -** decision templates, **WDM -** Woods dynamic method, **CS -** convex stabilizer, **CM-MV -** committee method - majority voting, **CM-A -** committee method – average value, and **LC** - logical correction.

Previous control recognition showed that the best collective recognition results could be achieved using Bayesian method and convex stabilizer strategies (prediction error equals 0 %). These algorithms results were used for making a decision at new chalcopyrites bandgap prediction (Table 3). Thus three new compounds ($ZnAlS_2$, $ZnAlSe_2$ and $BeCN_2$) are promising for opto-electronic applications (Table 3). The following designations were used: 1 – chalcopyrites with ΔE>2 eV and 2 – chalcopyrites with ΔE<2 eV.

Table 2. ΔE Prediction of Known Chalcopyrites (Threshold = 2 eV)

Com-pound	Experimental		Prediction												
	class	ΔE, eV	EC	LDF	LM	MR	NN	KNN	SVM	SWS	TA	LG	DT	CF	SVR
$CuAlS_2$	1	3.5	1	1	1	1	1	1	1	1	1	1	1	1	1
$CuGaS_2$	1	2.44	2	1	1	1	1	1	1	2	2	1	2	1	1
$CuInS_2$	2	1.5	2	2	1	?	2	2	2	2	2	1	2	2	2
$CuAlSe_2$	1	2.67	2	1	1	1	1	1	1	2	2	1	2	1	1
$CuGaSe_2$	2	1.63	2	1	1	1	1	1	1	2	2	1	2	2	2
$CuInSe_2$	2	0.95	2	2	2	2	2	2	2	2	2	2	2	2	2
$CuAlTe_2$	1	2.06	2	2	2	2	1	1	1	2	2	2	2	1	1
$CuGaTe_2$	2	1.18	2	2	2	2	1	1	2	2	2	2	2	2	2
$CuInTe_2$	2	0.88	2	2	2	2	2	2	2	2	2	2	2	2	2
$AgAlS_2$	1	3.13	2	1	1	1	1	1	1	1	1	1	1	1	1
$AgGaS_2$	1	2.75	2	1	1	1	1	1	1	2	1	1	2	1	1
$AgAlSe_2$	1	2.55	2	1	1	1	1	1	1	2	1	1	2	1	1
$AgGaSe_2$	2	1.65	2	1	2	2	1	1	1	2	2	2	2	2	2
$AgInSe_2$	2	1.24	2	2	2	2	2	2	2	2	2	2	2	2	2
$AgAlTe_2$	2	1.8	2	2	2	2	1	1	2	2	2	2	2	2	2
$AgGaTe_2$	2	1.1	2	2	2	2	1	2	2	2	2	2	2	2	2
$AgInTe_2$	2	0.96	2	2	2	2	2	2	2	2	?	2	2	2	2
$ZnSiP_2$	1	2.07	2	2	1	1	1	2	2	1	1	1	2	1	1
$ZnSiAs_2$	1	2.1	2	2	1	1	1	2	2	1	1	1	2	1	1
$ZnGeN_2$	1	2.9	2	2	1	1	1	1	2	1	1	1	2	1	1
$ZnGeP_2$	1	2.1	2	2	1	1	1	2	2	1	1	1	2	1	1
$ZnGeAs_2$	2	1.16	2	2	2	2	2	2	2	2	2	2	2	2	2
$ZnSnP_2$	2	1.45	2	2	2	2	2	2	2	2	2	2	2	2	2
$ZnSnAs_2$	2	0.74	2	2	2	2	2	2	2	2	2	2	2	2	2
$ZnSnSb_2$	2	0.4	2	2	2	2	2	2	2	2	2	2	2	2	2
$CdSiP_2$	1	2.2	2	2	1	1	1	2	2	1	1	1	2	1	1
$CdGeP_2$	2	1.8	2	2	2	2	1	2	2	1	1	2	2	2	2
$CdGeAs_2$	2	0.53	2	2	2	2	1	2	2	2	2	2	2	2	2
$CdSnP_2$	2	1.16	2	2	2	2	2	2	2	2	2	2	2	2	2
$CdSnAs_2$	2	0.3	2	2	2	2	2	2	2	2	2	2	2	2	2
$AgInS_2$	2	1.9	2	2	2	2	2	2	2	2	2	2	2	2	2
$CdSiAs_2$	2	1.51	2	2	2	2	1	2	2	1	1	2	2	2	2
$CuFeS_2$	2	0.53	1	2	2	1	1	1	2	1	1	1	2	2	2
$CuFeSe_2$	2	0.16	2	2	2	?	1	2	2	2	2	2	2	2	2
$CuFeTe_2$	2	0.1	2	2	2	2	1	2	2	2	2	2	2	2	2
$LiGaTe_2$	1	2.31	2	2	2	1	2	2	2	2	2	2	2	1	1
$LiInTe_2$	2	1.46	2	2	2	2	2	2	2	2	2	2	2	2	2
$AgFeSe_2$	2	0.23	2	2	2	2	1	2	2	2	2	2	2	2	2
$MgSiP_2$	1	2.35	2	1	1	1	1	2	2	1	?	1	2	1	1
$MnGeP_2$	2	0.24	2	2	2	1	1	2	2	1	1	1	2	2	2
$MnGeAs_2$	2	0.6	2	2	2	2	2	2	2	2	2	2	2	2	2

Table 3. ΔE Prediction of New Chalcopyrites (Calculations Using Collective Methods) (Threshold = 2 eV)

Compound	Prediction of ΔE	BM	C&S	DT	WDM	CS	CM-MV	CM-A	LC
$ZnAlS_2$	1	1	1	1	?	1	1	1	1
$ZnAlSe_2$	1	1	1	1	1	1	1	1	1
$ZnAlTe_2$	2	2	2	2	?	2	2	2	2
$AgFeS_2$	2	2	2	2	2	2	2	2	2
$AgFeTe_2$	2	2	2	2	2	2	2	2	2
$ZnGaTe_2$	2	2	2	2	?	2	2	2	2
$CdGaTe_2$	2	2	2	2	?	2	2	2	2
$HgGaTe_2$	2	2	2	2	?	2	2	2	2
$BeCN_2$	1	1	1	1	1	1	1	1	1

4 Conclusion

The information-analytical system for knowledge discovery in databases for properties of inorganic substance and materials was developed. The IAS data analysis subsystem is based on pattern recognition methods. The IAS application to the inorganic compounds computer-assisted design allows one to find complex classification regularities that make it possible to predict the membership of new chemical systems in one class of substances or another on the knowledge basis of the well-known properties of these systems components — chemical elements or more simple compounds. Using IAS it was possible to carry out the prediction of thousands of new compounds and estimation of some of their properties. Computer-assisted design allows one to substantially reduce the number of complex and expensive experiments in the search for inorganic compounds with predefined properties, replacing them by computation. The experimental results verification of computer-assisted design shows that the average prediction accuracy is higher than 80%.

Acknowledgments

This work was partially supported by the Russian Foundation for Basic Research (project nos. 06-07-89120, 08-01-90427, 08-07-00437, 05-03-39009, 11-07-00715, 09-07-00194, and 12-07-00142).

We are grateful to V.P. Gladun, V.V. Ryazanov, O.V. Sen'ko, A.A. Dokukin, and V.Yu. Velichko for long-term help and collaboration.

References

[1] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. *Advances in Knowledge Discovery and Data Mining*, AAAI Press/The MIT Press, Cambridge, 1996.
[2]. Larose D.T. *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley & Sons, UK, 2004.
[3] Kimball R., Caserta J. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*, John Wiley & Sons, UK, 2004.
[4] Morgenthal J.P. *Enterprise Information Integration: A Pragmatic Approach,* Lulu.com, Morrisville, 2005.
[5] Morgenthal J.P. *Enterprise Applications Integration with XML and Java,* Prentice Hall PTR, New York-Sydney-London, 2000.
[6] Kiselyova N., Iwata S., Dudarev V., et al. Integration principles of Russian and Japanese databases on inorganic materials, *Int.J. "Information Technologies and Knowledge"*, 2008, v.2, №4, pp. 366-372.
[7] Kiselyova N.N., Dudarev V.A., & Zemskov V. S. Computer information resources in inorganic chemistry and materials science, *Russian Chemical Reviews,* 2010, v.79, №2, pp. 145–166.
[8] Kiselyova N.N. *Computer Design of Inorganic Compounds. Application of Databases and Artificial Intelligence,"*Nauka", Moscow, 2005 (Russ.).
[9] Kiselyova N.N., Prokoshev I.V., Dudarev V.A., et al. Internet-accessible electronic materials database system, *Inorganic Materials,* 2004, v.40, №3, pp. 321-325.
[10] Zhuravlev Yu.I., Ryazanov V.V., Sen'ko O.V. *RECOGNITION. Mathematical Methods. Software System. Practical Solutions*, Phasis, Moscow, 2006 (Russ.).

[11] Gladun V.P. *Processes of Formation of New Knowledge,* SD "Pedagog 6", Sofia, 1995 (Russ.).
[12] Kiselyova N., Stolyarenko A., Ryazanov V., Podbelskii V. Information-analytical system for design of new inorganic compounds, *Int.J. "Information Theories & Applications"*, 2008, v.15, № 4, pp. 345-350.
[13] 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,* 2011, v. 21, № 1, pp. 88-94.
[14] Senko O.V. An optimal ensemble of predictors in convex correcting procedures, *Pattern Recognition and Image Analysis,* 2009, v. 19, № 3, pp. 465-468.
[15] Senko O.V., Dokukin A.A. Optimal forecasting based on convex correcting procedures, in *New Trends in Classification and Data Mining*, ITHEA, Sofia, 2010, pp. 62-72.
[16] Siebentritt S. Wide gap chalcopyrites: material properties and solar cells, *The solid films*, 2002, v. 403-404, pp.1-8.
[17] Xu J., Li H., Du B., et al. High thermoelectric figure of merit and nanostructuring in bulk $AgSbTe_2$, *J. Mater. Chem.*, 2010, v.20, №29, pp. 6138–6143.
[18] Zondy J.-J. Experimental investigation of single and twin $AgGaSe_2$ crystals for CW 10.2 μm SHG, *Opt. Commun.,* 1995, v.119, №3-4, pp. 320-326.
[19] Ohmer M.C., Goldstein J.T., Zelmon D.E., et al. Infrared properties of $AgGaTe_2$, a nonlinear optical chalcopyrite semiconductor, *J.Appl.Phys.,* 1999, v.86, №1, pp. 94-99.
[20] Isaenko L., Yelisseyev A., Lobanova S., et al. $LiInS_2$: A new nonlinear crystal for the mid-IR, *Materials Science in Semiconductor Processing*, 2001, №.4, pp. 665-668.
[21] Kiselyova N.N., Podbelskii V.P., Ryazanov V.V., Stolyarenko A.V. Computer-aided design of new inorganic compounds with composition ABX_2 (X = S, Se, Te), *Inorganic Materials: Applied Research*, 2010, v. 1, № 1, pp. 9–16.
[22] Savitskii E M., Gribulya V.B., & Kiselyova N.N. Cybernetic prediction of superconducting compounds. *CALPHAD,* 1979, v. 3, №3, pp. 171-173.
[23] Kiselyova N.N., Stolyarenko A.V., Gu T., & Lu W. Computer-aided design of new wide bandgap semiconductors with chalcopyrite structure, *Advanced Materials (Perspectivnye Materialy), Special Issue*, 2007, pp. 351-355.
[24] Kiselyova N.N., Stolyarenko A.V., Gu T., et al. Computer-aided design of new inorganic compounds promising for search for electronic materials, in *Proc. The Sixth Int.Conf. on Computer-Aided Design of Discrete Devices (CAD DD 07),* UIPI NASB, Minsk, 2007, v.1, pp. 236-242.