Prediction of Thermodynamic Properties by Artificial Intelligence Techniques

CALIN I. ANGHEL* , **MIRCEA V. CRISTEA**
Faculty of Chemistry and Chemical Engineering,, Department of Chemical Engineering and Materials Science, 11 Arany Janos Str., 400028, Cluj-Napoca, Romania

Thermodynamic models are solid foundations for many theoretical investigations of multicomponent systems. Beside of cumbersome theoretical approaches necessary to develop new insights in this context of potential support is the integration of artificial intelligence innovative ideas. The paper implements a novel procedure of the artificial intelligence based on support vector machine in a minimax approach. The main goal of the paper is to compare the performance of the novel procedure with artificial neural networks or other theoretical approaches and to promote it as an effective technique for thermodynamic analyses. Comparative results demonstrate the capability of the proposed procedure to be used by engineers dealing with multicomponent systems analysis but also its potential application in other engineering fields.

Keywords: thermodynamic model, multicomponent solid solution, artificial neural networks, support vector machine, minimax approach.

The development and processing of new materials are dynamic and complex fields. Although scientific investigations have produced basic knowledge of the underlying phenomena many problems still remain unsolved, where quantitative deterministic characterization or theoretical approaches are dismally lacking or are cumbersome.

The knowledge of thermodynamic properties for multicomponent solid solutions systems are still of great interest to understand the physical chemistry of the underlying phenomena [1-8]. Despite of the recent important advances theoretical methods cannot always fulfil in a satisfactory manner all the demands. Within this context, over the past decades intense effort has been constantly devoted to bring ideas and methodologies from the artificial intelligence filed into solving engineering or scientific problems. It is well known that artificial intelligence procedures are capable of replicating a lot of varieties of non-linear relationships of considerable complexity and avoid uncertainty in input parameters. It is commonly recognised the ability of these methods to model parameters, predict properties or trends of interest or to investigate cases of new phenomena where the information cannot be easily described theoretically or without the physics, chemistry or biology being explicitly provided. Among the well known artificial intelligence techniques developed during the last decades the following approaches may be pointed out: artificial neural networks, genetic algorithms and support vector machines. The application of artificial neural networks (*ANNs*) in many domains has been a rapidly growing field. The most important limitations of *ANN*s procedures are: (i) sensitivity to the dimension of available data, over-fitting and local minima, (ii) accuracy, not directly linked by the ANN architecture (number of layers and number of neurones in the hidden layers), (iii) inability to adequately identify unnecessary weights in the network, (iv) their intrinsic nature as a black-box thinking. To overcome these drawbacks of the artificial neural networks the present work is proposing the support vector machine (*SVM*) in minimax approach. Important incentives of the support vector classifier (that is primarily a two-class classifier) over the *ANN*s are: (i) being a non-parametric classifier makes no prior assumptions concerning the data distribution, (ii) provides better accuracy even with a small number of training samples and is fast and simple in implementation, (iii) avoids the specific *ANN*s problems such as over-fitting and local minima, (iv) it has a relative explicitly nature. A major drawback of the support vector machine consists in simple assessment of the same covariance for each class and thus the margin should be defined in a local way. The main advantages of the support vector machine in minimax approach are: (i) avoids the drawback of SVM consisting in the simple assessment of the same covariance for each class and defining the margin in a local way, (ii) unlike SVM, for which the closest points to the decision boundary are most important, the minimax approach looks at the margin between the means of both classes, (iii) provides an explicit direct upper bound on the probability of misclassification of new data, without making any specific distribution assumptions and (iv) obtains explicit decision boundaries based on a global information of available data.

The support vector machine and minimax approach, named minimax probability machine classification, has become an active research topic. There is a rich literature on this subject [9-17]. To provide predictive power for the minimax probability machine classification a new regression model was built for maximising the minimum probability of future predictions, such as to keep them within some bound of the true regression function. The authors [15] referred to this regression framework as the minimax probability machine regression. Based on these procedures the present work develops and implements in the MATLAB object oriented language a minimax decision procedure. The procedure casts both classification (class labels as outputs) and regression algorithms (numerical values as outputs) into a global unified technique. The aim is to introduce and compare the performance of the proposed procedure (minimax decision procedure) with theoretical approaches or traditional procedures such as the artificial neural networks. Another significant target is

* email: canghel@chem.ubbcluj.ro

to demonstrate the feasibility of the minimax decision procedure in such kinds of scientific applications. Fundamentals and basic principles were presented elsewhere [18-20]. According to these previous developments, probabilistic certainty of the predictions may be also obtained.

In the present work the thermodynamic significance of the results is not considered as the main task and therefore was not thoroughly investigated. Similar approaches based on support vector machine in minimax approach applied in the filed of thermodynamics of multicomponent solid solution system have not yet been reported. In a comparative manner, a numerical experiment based on data reported in related studies [5] is used to show incentives of the proposed approach. Based on expression of the excess Gibbs energy of the mixture and partition function, the related work [5] predicts the thermodynamic properties in a multicomponent solid solution system using only binary activities parameters or infinite dilute activity coefficients. Our experiment replicates [5] and compares the performance of the proposed procedure with the well-known artificial neural networks approach and theoretical achievements involving the molecular interaction volume model. The experiment deals with the prediction of the activity coefficients to describe the thermodynamic properties for the solid solution of C-Fe-Co-Ni system. Good results point out the efficiency of the proposed minimax decision procedure based on support vector machine in a minimax approach for multicomponent systems analysis. It is worthwhile to mention that, by proper developments, this minimax decision procedure has a large potential to be applied in many other engineering fields.

Procedures implementation

Implementation of the minimax decision procedure

The main targets of this minimax decision procedure consist in getting predicted values of interest without directly using the explicit relational information, together with obtaining classification and probabilistic certainties. Every time it is possible, the procedure is conducted in a manner of a data mining application. In a way specific to common data mining procedures, the error may be estimated by testing rather than by calculation. To carry out the most basic testing method (simple testing) a random percentage of the database (10-30%) is set aside and not used in any way in the model building and estimation (training). This set, named testing data, will be used for the final test of the procedure. The remaining set of data is used during the training step (learning and validation) in order to build the model. Figure 1 presents a general framework of the global minimax decision procedure. The implementation was developed as a user-friendly computer application in MATLAB software environment and works in a multiple step and cyclic approach. Computer programs were coded in a convenient way to find the best results or the best case over a number of "k" cyclic experiments (simulations). The best case over these experiments and the corresponding output values emerged from the minimax probability machine regression, equations [19,20], was defined as the sample model. Formally, it represents the best procedure and outputs for some particular task. To ensure accuracy and stability of the procedure the sample model is generated based on data randomly divided into a number of distinct training (learning-validation) and testing subsets. Subsequent to its establishment, the model is used for further simulations and for predictions on the testing data (the subset randomly extracted from the total database). This final step stands for a global testing step of the procedure. Long random trials (k > 100) do not get improved accuracy or more reliable predictions. Taking this aspect into account, it was considered appropriate to obey recent statements [16,17] and to work with a reduced learning set, i.e. to limit the trials to k = 50…100. The performance of the procedure was investigated in detail based on the following criteria:

- empirical correlation factors ERC (1)
- relative errors

$$\mathrm{RE} = 100 \times \left| Y_{predicted} - Y_{test} \right| \Big/ \left| Y_{predicted} \right| \ [\%] \quad (2)$$

- simple equivalent linear dependency between predicted and corresponding testing values, shown as a linear regression equation described by:

$$Y_{predicted} = a \cdot Y_{test} + b \quad (3)$$

Consequently, better predictions means *a* index close to unity and *b* index close to the zero value. The performance criteria are evaluated with all values brought back into the original R^d space. The probabilistic certainty of predicted values was obtained according to a minimax development with convergence towards the mean [18,19]. Basically, the probabilistic certainty was achieved in a binary classification manner. The data for binary classification are given based on the limit state function:

$$LSF = \frac{\mathbf{Y}_{\mathrm{Predicted}}}{\mathbf{Y}_{\mathrm{Test}}} = 1 \pm \Delta\varepsilon \,. \quad (4)$$

A decision regarding the probabilistic certainty of predicted values should be done by separating the values of the limit state function reported at unity, with a confidence value of $\pm\Delta\varepsilon$. A confidence value of $\pm\Delta\varepsilon$ = 0.05 corresponds to a confidence interval around 95%. As it was previously stated [13,14], the performance of the presented procedure might be completed with the test set accuracy (percentage of well-classified test data) and the lower bound *a* on correct classification of future data. In all experiments the lower bound on the correct classification of the future data must be smaller than the test set accuracy. As a result, the lower bound is not violated and the linear approach is robust.

An empirical but heuristic principle was applied for setting the type of the kernel function. The kernel type that yields to the best performance, assessed by equations (1-3), was put aside and considered for the final testing. The proper size and selection of the training set (divided into learning and validation subsets) is very important to produce optimal results and to increase the performance of the algorithm. So far, there are no uniquely agreed and generalised approaches to choose the suitable dimension and the selection of the training set. However, it is a commonly agreed statement that the training set must be sufficiently large compared with the number of features. In the present study the procedure was conducted in a basic manner, without features and outliers selection or reduction.

Artificial neural networks implementation

Founded on an idealised model of the biological neuron, the calculation paradigm of *ANN*s is able to represent information on complex systems. The main characteristics

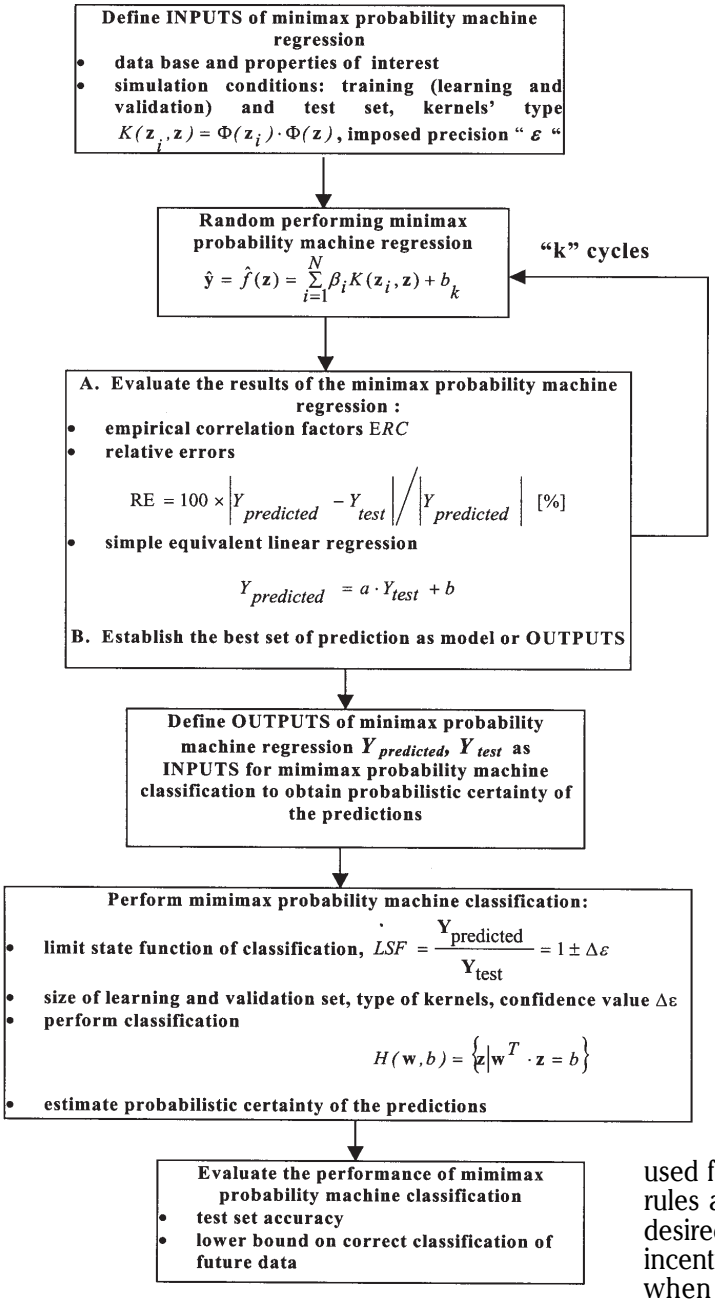


Fig. 1. Basic framework of the minimax decision procedure

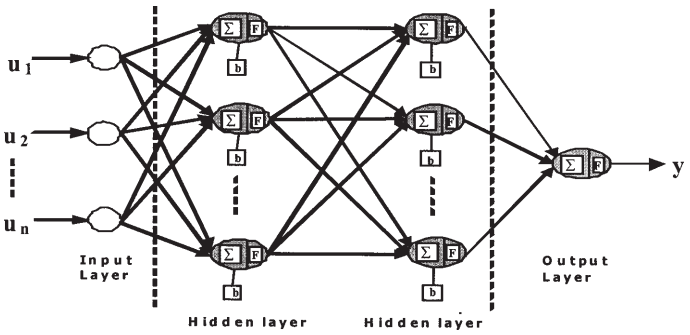


Fig. 2. Basic triple-layer ANN structure

of the *ANN*s model are the inputting of information (signals) from exterior or other units of the network, feeding it to the given unit (neuron) that processes it and then sending it, as output, to other units or output of the network. For the generic case, the weighted connection paths link every two neurons to each other, the weighting structure providing the total network performance. The main benefits of the *ANN*s approach consist in its remarkable ability of learning, generalisation and robust behaviour in the presence of noise [21-23]. As a consequence, the *ANN*s may be successfully used for modelling systems in which detailed governing rules are unknown or are difficult to formalise, but the desired input-output set is known [24]. They also present incentives for cases when input-output data are noisy and when high processing speed is required. The *ANN*s prediction capability of generating new output values for given inputs is greatly appreciated. The artificial neural networks approach was implemented by means of the Neural Network Toolbox from *MATLAB* 7.2.0 software package. Based on the feed-forward *ANN* architecture various designs of *ANN*s were experimented. They had different number of neurons, architecture with single or double hidden layers and different transfer functions. Consequently, a triple-layer feed-forward *ANN* with *tansig*, *tansig* and *purelin* activation functions (for the hidden and output layers) and the back-propagation training algorithm, has been employed for computing the network biases and weights, (fig. 2). For the output layer the linear transfer function was applied. The hyperbolic tangent sigmoid transfer function was implemented in the hidden layers:

$$F(x)=\frac{1-e^{-x}}{1+e^{-x}} \quad \text{and} \quad x=\sum_{1}^{n}\left(u_i\cdot w_i+b\right) \tag{5}$$

For predictions presented in the following numerical example, the *ANN* input layer has 4 neurons, in the hidden layers 7 and 2 neurons have been used and the output layer consists in one neuron. The number of nodes in the

Table 1
INPUT VARIABLES FOR THE PREDICTION OF THE ACTIVITY COEFFICIENTS [5]

Variable	*Mean*	*Standard deviation*	*Minimum value*	*Maximum value*
X_C	0.013111	0.006997	0.006	0.037
X_{Fe}	0.33853	0.19793	0.094	0.794
X_{Co}	0.31958	0.19918	0.085	0.796
X_{Ni}	0.32856	0.20289	0.083	0.791
$\ln\gamma_C$	4.0767	0.4879	2.865	4.844

Table 2
THE MAIN PARAMETERS OF SIMULATION AND RESULTS OF THE MINIMAX DECISION PROCEDURE

Parameters	***Example [5]***
Original data base (ODB)	Variables (columns) = 5 Samples (rows) = 36
Training (learning-validation) set size	70% of ODB
Cycles of simulation	k = 50
Test set size	30% of ODB
Kernel functions for *minimax* regression procedure	$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d$
Performance of the procedure reported for the final testing	
Empirical correlation factor	0.933...0.999
Relative errors [%] (eq. 2)	2.844...0.192
Coefficients of equivalent linear dependency eq. (3)	a = 0. 982 b = 0.071
Probabilistic certainty of predicted values	0.833
Confidence value $\pm\Delta\varepsilon$ (eq. 4)	0.050
Test set accuracy [%]	86.73
Lower bound on correct classification of future data [%]	83.82

hidden layer has been set on the basis of a trial and error process. The quasi-Newton Levenberg-Marquardt algorithm was employed for training the *ANN* [23]. Over fitting has been avoided by early stopping. During the repeated presentations of the training data set in the training steps, random initial conditions have been used for the weights and biases in order to prevent the convergence to undesired local minima. With the aim of improving the training procedure all input-output training data have been normalized using the maximum and minimum values of the input and output sets of data. In a specific way of data mining procedures, the entire set of available data (input-output pairs) has been first divided in two sets: one set used for training the *ANN* and a second set used for the testing and performance assessment of the trained *ANN*. The set of data used for training has been further randomly divided in two subsets: one subset (input-output pairs) directly used for the learning procedure and a second validation subset used for preventing over fitting by the early stopping method. Once the training step was completed, the trained *ANN* model was established and premises for using it in the predicting applications was accomplished. The best-trained *ANN*, according to the results obtained for the training-validation set, has been further used for making the predictions on the not yet seen testing set. The ANN performance was investigated based on the same criteria and the same set of testing data as those previously presented at the minimax decision procedure.

Numerical applications and results

The numerical example follows a study presented by Tao [5] and predicts the activity coefficients of carbon in the quaternary solid solution of C-Fe-Co-Ni system. In the original paper the activity coefficients ($\ln\gamma_C$) of carbon (in the quaternary solid solution C–Fe–Co–Ni, at 1273K) were predicted based on cumbersome materials science theory. They were determined as binary activities or infinite dilute activity coefficients. The predicted output was obtained using a model having the generic form of a function with input variables, the molar fractions of components C, Fe, Co and Ni in the solvent alloy C- Fe–Co–Ni. Predictions have been further compared with the experimental data. The database is relatively small and consists of only 36 samples of activity coefficients, considered as outputs, and 36 samples of 4-dimension vectors (molar fractions of components), considered as inputs. Despite the small size of the database a simple testing based on the randomly set aside testing set of data it is still possible. Table 1 shows the details of the data used in the simulations.

Whenever it was possible the two comparative procedures, artificial neural networks and minimax decision procedure, were conducted based on the same training and testing data sets. It was convenient to extract a 30% fraction from the database and to set aside. This randomly selected data set was intended to the final testing-performance of both the minimax decision procedure and the artificial neural networks processing, in a comparative assessment. The main conditions of the simulations and results are presented in tables 2 and 3. In the minimax decision procedure implementation a kernel type having the form of a polynomial with unit offset (table 2) was proved to work very well for performing the predictions.

The performance of both procedures was established based on: values of relative errors (eq. 2), simple equivalent linear dependency between predicted and corresponding

testing values (eq. 3) and comparative dependency between predicted and testing values. The performance was evaluated on the testing set of data (the set extracted from the database at the beginning of the random and cyclic procedure). The performance was reported alike only to the values of carbon activity coefficients in the quaternary solid solution [5].

Results are comparatively presented. Figure 3a presents the values of the relative errors (eq. 2), reported to the theoretic values obtained where molecular interaction volume model was used [5]. The others, figure 3b and figure 3c present the values of the relative errors (eq. 2), reported to the predicted values obtained from the minimax decision procedure and the artificial neural networks approach. Figure 4 presents the comparative dependencies between the experimental results and the results reported for the predicted values of carbon activity coefficients in the quaternary solid solution (fig. 4a) and respectively dependencies of the activity coefficients obtained based on predicted values from minimax decision procedure and the artificial neural networks approach (fig. 4b-4c) [5]. At first glance, the minimax decision procedure results and the theoretical emerged results are in a reasonable good agreement, but the predictions obtained by the minimax decision procedure seem to be superior to those obtained by the theoretical approaches [5].

Based on the high values of correlation coefficients, reduced relative errors and good linear dependency between predicted and testing values (table 2-3) it may be concluded, in a basic evaluation analysis, that performance of the minimax decision procedure is reasonable. Taking into account the robust mathematical nature of the proposed procedure and its way of development, i.e. being achieved without the need of direct or explicit relational or theoretical information, it may be concluded to be appreciated for its predictive power. When analysing and comparing the results between the ANN and the minimax decision procedure a good agreement may be observed. As the results and figures show, relatively reduced differences are noticed. Generally, all these differences are within 5-10 %. These results demonstrate the real capacity and good accuracy of the proposed procedure and confirm its important predictive capacity.

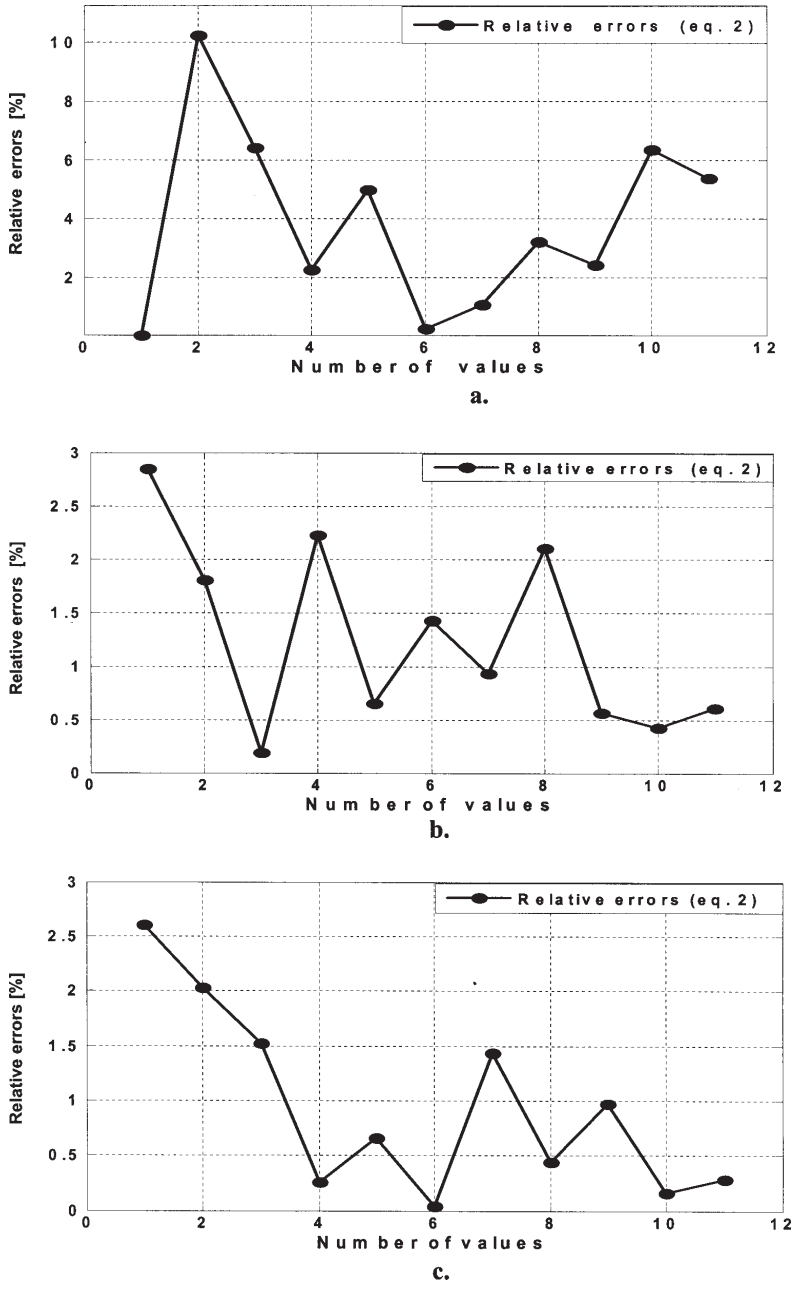
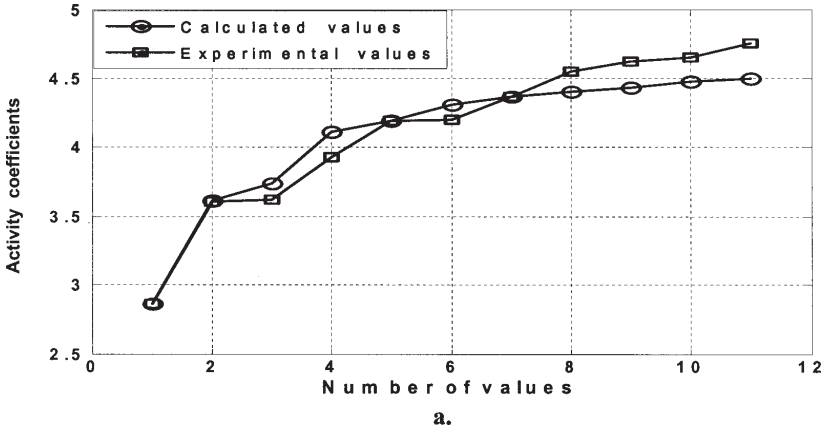
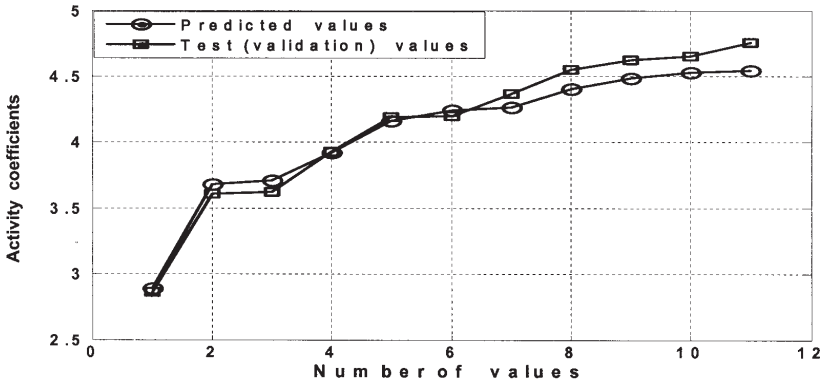


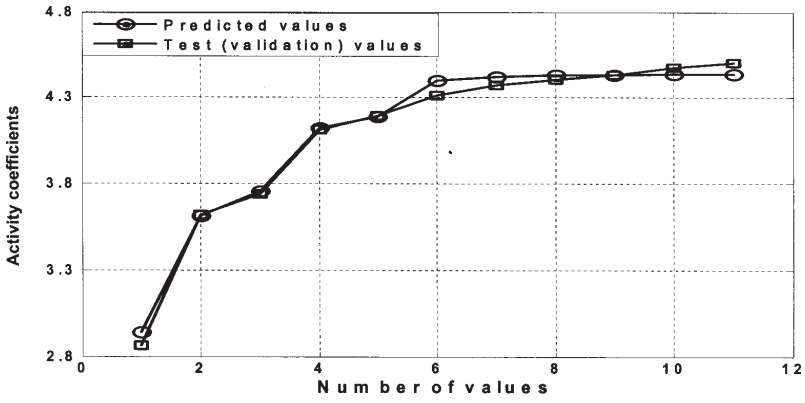
Fig. 3. Comparison between relative errors of the activity coefficients
a. Relative errors computed based on theoretical values presented by Tao [5];
b. relative errors computed based on values of the minimax decision procedure; c. relative errors computed based on values of the artificial neural networks approach



a.



b.



c.

Fig. 4. Comparison between dependencies of the activity coefficients
a. Dependency between theoretical and experimental values [5]; b. dependency between predicted and experimental values in minimax decision procedure; c. dependency between predicted and experimental values in artificial neural networks approach;

Parameters	***Example [5]***
Original data base (ODB)	Variables (columns) = 5 Samples (rows) = 36
Training (learning-validation) set size	70% of ODB
Test set size	30% of ODB
Transfer function for hidden layers	$F(x)=\frac{1-e^{-x}}{1+e^{-x}}$ and $x=\sum_{1}^{n}(u_i \cdot w_i + b)$
Performance of ANN reported for the final testing	
Empirical correlation factor	0.995
Relative errors [%] (eq. 2)	2.604…0.037
Coefficients of equivalent linear dependency eq. (3)	a = 0.960 b = 0.180

Table 3
THE MAIN PARAMETERS OF SIMULATION AND RESULTS FOR THE ARTIFICIAL NEURAL NETWORKS APPROACH

Conclusions

The presented example and the results emphasise the ability of the proposed minimax decision procedure to make good predictions in the thermodynamic multicomponent systems analysis. All the predicted data were situated within a 95 % confidence interval close to the desired values. As it was foreseen and then proved, the capacity of the minimax decision procedure is good over the entire set of investigated cases. For the best-obtained outputs, denoted as the sample model, the capacity of the

procedure is reasonable, and its performance is at least as good or even superior to the *ANN* approach. The depicted example points out the real potential of the proposed approach for being successfully implemented in thermodynamic analyses of multicomponent systems. Despite the facts of minimax decision procedure: (i) seems to be less known like *ANN* approach, (ii) on this relatively small database produces results not significantly superior to *ANN* approach, it proved its potential in thermodynamic analysis. Main advantages of the procedure are: (i) may work properly with a reduced learning set, (ii) needs relative few science-phenomena knowledge and avoids cumbersome theoretical approaches, (iii) alike *ANN* approach is able to investigate new phenomena where the information cannot be easily accessed theoretically or directly by explicitly relational descriptions, (iv) alike *ANN* approach makes possible property or trend predictions.

References

1. KATTNER, U.R., Thermodynamic ModeLling of Multicomponent Phase Equilibria, JOM, 49, 12, 1997, p.14
2. DANIELEVSKI, M., FILIPEK, R., The generalized solutions in non-equilibrium thermodynamics, JSCTA Netsu Sokutei, 24, 4, 1997, p.165
3. ALIKHANIAN, A.S., GUSKOV, V.N., NATOROVSKII, A.M., KOVALENKO, V.V., Thermodynamic Properties of ZnTe–CdTe Solid Solutions, Inorganic Materials, 39, no.3, 2003, p.234
4. KANG, Y.-B., JUNG, In-Ho, DECTEROV, S., PELTON, A.D., LEE, H.Geon, Phase Equilibria and Thermodynamic Properties of the CaO-MnO-Al_2O_3-SiO_2 System by Critical Evaluation. Modelling and Experiment, ISIJ International, 44,no. 6, 2004, p.975
5. TAO, D.P., Mater. Sci. Eng. A 390 (2005) 70
6. MORARIU, L., VLAD, M., Rev. Chim.(Bucuresti), **58**, no. 2, 2007, p.129
7. MOLDOVAN P., POPESCU, G., BUTU, M., CUHUTENCU, M., BUTU, L., Rev. Chim.(Bucuresti), **58**, no. 6, 2007, p.537
8. DINC, E., BALEANU, D., TAS, A., Rev. Chim.(Bucuresti), **57**, no. 6, 2006, p.627
9. VAPNIK, V.N., The nature of statistical learning theory, 2nd edition, Springer, New York, 2000
10 JAIN, A.K., DUIN, R.P.W., MAO, J., Statistical Pattern Recognition: A Review, IEEE T Pattern Anal. 22 (2000) 4-37
11. BURGES, C.J.C., A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Disc. 2 (1998) 121
12. SCHOLKOPF, B., BURGES, C., SMOLA, A. (editors), Advances in Kernel Methods: Support Vector Machines, Cambridge MA MIT Press, 1998
13. LANCKRIET, G.R.G., GHAOUI, E.L., BHATTACHARYYA, C., JORDAN, M.I., Minimax Probability Machine, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14.*** Cambridge MA. MIT Press, 2002a, p. 801. (http://robotics.eecs.berkeley.edu/~gert/)
15. LANCKRIET, G.R.G., GHAOUI, E.L., BHATTACHARYYA, C., JORDAN, M.I., A Robust Minimax Approach to Classification, Journal of Machine Learning Research. 3 (2002) 555 (http://robotics.eecs.berkeley.edu/~gert/).
16. STROHMANN, T.R., GRUDIC, G.Z., A formulation for minimax probability machine regression, in: S. Thrun S. Becker, K. Obermayer (Eds.), Advances in Neural Information Processing Systems 15. Cambridge, MA. MIT Press, 2003 pp. 769
17. BORDES, A., ERTEKIN, S., WESTON, J., BOTTOU, L., Fast Kernel Classifiers with Online and Active Learning, Journal of Machine Learning Research. 6 (2005) 1579
18. LEE, YUH-JYE, MANGASARIAN, O.L., RSVM: Reduced Support Vector Machines, Computer Sciences Department, University of Wisconsin, Madison, WI 53706.(2000) olvi@cs.wisc.edu.
19. ANGHEL, C.I., OZUNU AL, A new insight for the risk of gaseous release assessment based on a Minimax approach, Chem. Pap., 59 (6b), 2005, p.469
20. ANGHEL, C.I.,Rev. Chim.(Bucuresti), **57**, no. 7, 2006, p.773
21. ANGHEL, C.I., OZUNU AL, Prediction of Pollution Level Based on Artificial Intelligence Methods, Chem. Pap. 60(6), 2006, p.410.
22. HAGAN, M.T., DEMUTH, H.B., BEALE, M.H., Neural Networks Design, MA: PWS Publishing, Boston, 1996.
23. HAYKIN, S., Neural Networks A Comprehensive Foundation, Macmillan Publishing Company, Englewood Cliffs, NJ, 1994
24. HAGAN, M.T., MENHAJ, M.H., Training feedforward networks with the Marquardt algorithm, IEEE Transaction on Neural Networks, 5(6) (1994) 989
25. CRISTEA, V.M., BATIU, I., Vapor-liquid equilibrium predictions using neural networks in ternary system (+) fenchone+metyl chavicole+trans anethole, Rev. Roum. Chim.,, 50, 11-12, 2005, p.1009

Manuscript received: 21.05.2009