Journal of Ceramic Processing Research. Vol. 9, No. 4, pp. 393~397 (2008)

J O U R N A L O F
Ceramic
Processing Research

A virtual-sample technology based artificial-neural-network for a complex data analysis in a glass-ceramic system

Wen Qi-Ye[a,*], Zhang Huai-Wu[a], Yang Qing-Hui[a] and Zhang Pei-Xin[b]

[a]*State key Laboratory of Electronic Thin Films and Integrated Devices, University of Electronic Science and Technology of China, Chengdu, 610054, China*

[b]*Normal College, Shenzhen University, Shenzhen, Guangdong 518060, P.R China*

Artificial neural network has becoming a mainstream technology in the domain of complex materials data analysis. Based on a slag glass-ceramic system we brought forward a virtual sample technology to increase the training samples by fluctuating the content of main compositions in a proper small amplitude. Simulation results proved that a good virtual sample set can not only improve the network's prediction ability considerably, but can also suppress the "overtraining" phenomenon. Therefore a virtual sample improved neural network model can learn the relationship from a small size experimental data set and give an accurate and stable prediction for the test samples. This is more helpful to the material data analysis and can facilitate the design and development for new materials.

Keywords: Artificial neural network; Material Data Analysis; Virtual sample technology; Slag Glass-Ceramics.

Introduction

Artificial neural networks (ANNs) are mathematical constructions that are believed to loosely model the working of the human brain. They are nowadays applied in different domains especially in the area of data analysis. Compared to traditional methods, the ANNs were found to be a more efficient tool for multimensional, complex, and quantitative problems such as material data analysis and property prediction [1]. With much effort concentrating on ANN, some of the bottlenecks encountered when developing NNs for data modeling are, at least in part, overcome. Also neural network has become a mainstream technology in the domain of materials data [2]. However, establishing a good neural network model for a specific problem requires not only the selection of an appropriate neural network type, but also a good problem description. Many of the failures in the application of NN are in fact not caused by the NN technology, but are a consequence of improper problem definition and data selection. At present the acquisition of sufficient and effective experimental data is of the most importance for complex materials-analysis problems [3, 4].

Exploiting of slag glass ceramics has both environmental and commercial value. But the relationships of composition, microstructure and properties of slag glass-ceramics are very intricate. The data from experiments are multidimensional, complex and nonlinear, which is hard to analye with ordinary statistical methods such as regression analysis. This becomes a main obstacle for the exploiting of new materials. In our previous study, an artificial neural network (ANN) model was applied to a glass-ceramic system for data analysis and property prediction. This network model, which uses a new robust back propagation algorithm as a learning law, proved to have a strong learning ability in the slag glass ceramic domain [4, 5]. However, its applications in material data analysis and property prediction are not always satisfactory because there is insufficient and effective data to train the network.

In this paper, based on the basic conception of slag glass-ceramic material, we introduce a virtual sample technology to fabricate abundant data from a small size original data set. The results of our research indicate that the data developed by this technology can not only improve the network learning and prediction ability, but can also suppress the so-called "over-training" phenomenon effectively. Therefore, with this new technology, an ANN model can be applied to a complex material domain more effectively and reliably even when not much experimental data has been provided.

The basic principle of virtual sample technology

Glass ceramics are composite materials consisting of a glass matrix and a crystalline phase, which generally are produced in two steps: 1) glass melting and forming and 2) crystallization. The slag glass-ceramics, which take slag and waste as raw materials, are also environment-friendly materials [7, 8]. The compositions of slag glass-ceramic are nearly the same as ordinary glass or ceramic

*Corresponding author:
Tel : +86-28-83203793
Fax: +86-28-83201810
E-mail: qywen@163.com

except for the "minor components" such as TiO_2, P_2O_5, MnO_2, ZnO, sulfide etc, which are introduced by the slag or waste used. The main components such as SiO_2, Al_2O_3, CaO, and MgO, whose content is usually above 90% of the total, determine the basic structure and crystalline phases that can be produced from the glass matrix in the process of thermal treatment. These "minor components", although minor in content, also have a considerable and complex influence on the microstructure and properties of the glass through the effection on crystallization.

According to the phase diagram of a certain glass system such as SiO_2-Al_2O_3-CaO, the samples with only a small difference in the contents of SiO_2, Al_2O_3 and CaO will have nearly the same possibility for crystal precipitation. In this context the nucleation and crystallization processes of a glass ceramic are mainly affected by the minor components. So if the contents of minor components are fixed, the samples with a small variation in the main components will have the same microstructure and of course identical properties. Based on this simple principle we can develop "virtual samples" from a real one by changing only the contents of the main component over a small amplitude. The number of these virtual samples is usually several times more than these originals. More importantly, they can be regarded as real ones if a proper content range is chosen. More effective "information" about the relationship which is implied in the original data can be provided by them and thus the performance of the trained network will have be considerably improved. The advantages of these virtual samples technology are: 1) large numbers of training samples can be developed from a small size original data set by a numerical method rather than from experiments, thus much money, time and manpower are saved. 2) If the original samples are valid, the virtual samples derived from them with a proper content range amplitude are also effective, then the trouble to eliminate invalid or conflicting data from a large size data set is avoided.

Simulation experiments and discussion

The fabrication of virtual samples

A CaO-Al_2O_3-SiO_2 system glass ceramic was developed from a blast-furnace slag [5]. A series of samples with different contents of compositions were prepared by an ordinary melting and thermal treatment method [8]. The coefficients of thermal expansion (CTE) were measured between 25-300 °C.The experimental data of composition and the CTE of different samples are shown in table 1. We define the first 5 components as "Major components" and last 5 components as "Minor components". Note that we call an components, for example MgO, as "major component" not because it is large in content but because it is common and/or necessary in the glass formation. There are 14 specimens in this data set. The first 11 samples were chosen as the training set and the last three as the test set.

According to the principle of virtual sample technology brought forward above, we altered the contents of main components over a small range, for example ± 0.5%, while keeping the content of minor components fixed. Then 2^K virtual samples were obtained for each original sample, where K is the number of input parameters allowed to alter. In the case where we take the first 5 components in table 1 to alter then 2^5 virtual samples can be developed form the original sample. However, it is unnecessary to present all these 2^5 samples to the network for training. According to the theory of orthographic design, 8 virtual samples developed by a seven factors and two levels orthogonal design table can represent all the 2^5 samples. Thus 11 × 8 samples are obtained from the original training set which contains 11 samples. Table 2 lists the 8 virtual samples developed from the first original sample in table 1 with a content alteration of ± 1%. In the following cases, we fabricate five virtual sample sets by fluctuating the contents of major components with amplitude of 0.5%, 1%, 2%, 3%, and 5% respectively, while keeping

Table 1. Element compositions and properties of CaO- Al_2O_3-SiO_2 system slag glass-ceramic

No.	Major component (w%)					Minor component (wt%)					CTE (10^{-7}/°C)
	SiO_2	Al_2O_3	CaO	Na_2O	MgO	Mn_2O_3	Fe_2O_3	TiO_2	ZnO	S	
01	74.03	2.51	12.51	5.97	2.75	0.14	0.75	0.21	1.61	0.95	62.1
02	74.02	5.02	10.01	5.97	1.49	0.26	1.04	0.41	1.61	1.08	58.3
03	69.02	5.02	10.01	5.50	1.49	0.26	1.04	0.41	1.61	1.08	58.8
04	64.23	5.14	19.83	5.44	1.47	0.26	1.12	0.40	1.93	1.24	77.8
05	64.23	15.09	9.86	5.44	1.47	0.26	1.12	0.40	1.93	1.24	60.9
06	59.23	5.14	24.81	5.44	1.47	0.26	1.12	0.40	1.93	1.24	80.2
07	54.23	5.14	29.86	4.97	1.47	0.26	1.12	0.40	1.93	1.24	82.8
08	49.43	20.21	19.71	4.97	2.93	0.51	1.70	0.80	1.93	1.50	71.9
09	44.23	5.14	39.82	4.50	1.47	0.26	1.12	0.40	0.13	1.24	105.3
10	44.43	25.25	19.71	4.50	2.93	0.51	1.70	0.80	0.13	1.50	71.6
11	39.43	20.21	29.68	4.50	2.93	0.51	1.70	0.80	1.93	1.50	76.2
12	69.02	5.02	14.99	5.50	1.49	0.26	1.04	0.41	1.61	1.08	67.6
13	64.23	10.12	14.84	5.44	1.47	0.26	1.12	0.40	1.93	1.24	66.1
14	54.43	15.25	19.71	4.97	2.93	0.51	1.70	0.80	1.93	1.50	75.1

Table 2. Virtual samples developed from the first real sample in table 1 with an altering in composition of 1%

No.	SiO_2	Al_2O_3	CaO	Na_2O	MgO	Mn_2O_3	Fe_2O_3	TiO_2	ZnO	S	CTE
1	74.77	2.54	12.63	6.03	2.84	0.14	0.75	0.21	1.61	0.95	62.1
2	74.77	2.54	12.63	5.91	2.72	0.14	0.75	0.21	1.61	0.95	62.1
3	74.77	2.48	12.38	6.03	2.84	0.14	0.75	0.21	1.61	0.95	62.1
4	74.77	2.48	12.38	5.91	2.72	0.14	0.75	0.21	1.61	0.95	62.1
5	73.29	2.54	12.38	6.03	2.72	0.14	0.75	0.21	1.61	0.95	62.1
6	73.29	2.54	12.38	5.91	2.84	0.14	0.75	0.21	1.61	0.95	62.1
7	73.29	2.48	12.63	6.03	2.72	0.14	0.75	0.21	1.61	0.95	62.1
8	73.29	2.48	12.63	5.91	2.84	0.14	0.75	0.21	1.61	0.95	62.1

the content of minor components fixed. We denote these five virtual sample sets as *I*, *II*, *III*, *IV*, and *V* respectively.

The prediction performance of the networks trained by virtual sample sets.

The original sample set as well as the five virtual sample sets, after being preprocessed with a statistical method, were presented to train the 3-layer feedforward neural networks (the fabrication details of the network are presented in ref. 6). The study error was set to 0.001. After that the test samples were presented to each trained network for predictions of thermo-expansion. The relative error (ε_r) between the prediction and experimental value was calculated, and for each network the prediction error on the test set ($\varepsilon_{ra} = \frac{1}{m}\sum_{i=1}^{m}\xi_{ri}$, where m is the number of test samples) were also calculated. This prediction error, ε_{ra}, represents the prediction ability of the trained network, which in turn reflects the effectiveness of the data set presented to train the network.. We noticed that if different networks converge to about the same final error and then give close predictions, it is probably the best result that can be obtained with the data set used. We call this the networks' "stability". If only one or two networks could be trained to a very low error, or the predictions are saliently different with different networks, we believe the real relationship was not disclosed and the predictions are "unstable". Therefore, for each training set, nine networks with different hidden nodes are trained. These networks have 10 input nodes and 1 output nodes, and the number of hidden nodes, N, varies from $2M$-3 to $2M$+ 5, where M= 10 is the number of input nodes.

Figure 1 shows the curves of prediction error, ε_{ra}, versus the number of hidden nodes (N_h) for networks trained by the original set and the five virtual sample sets. For the nine networks trained by each sample set, the average value and standard deviation of prediction errors were calculated and listed in table 3. We can see from Fig. 1, the prediction errors of networks trained by original data are as big as 3-5%, and oscillate distinctly with different N_h. The virtual sample set *I* has only an improvement of the prediction performance, which is possibly due to an "error dilute" effect; that is, errors inevitably exist in the content of compositions, which arise either in the weighing of raw materials or in the material fabrication

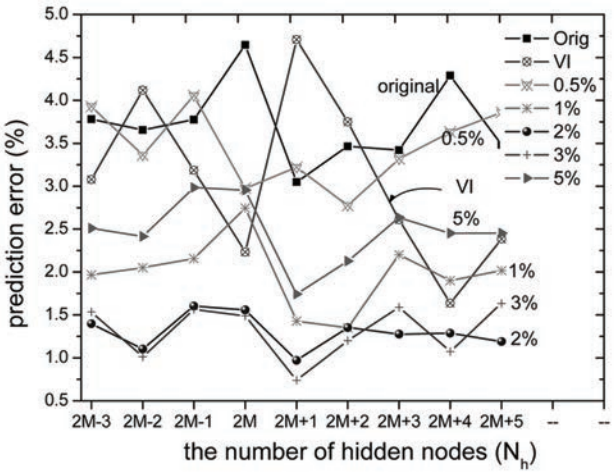

Fig. 1. Curves of prediction error versus the number of hidden nodes for networks trained by different data sets.

Table 3. The average value and standard deviation of prediction error for networks trained by different data sets

Bathe of network	Average value (%)	Standard deviation
Original	3.729	0.481
I(0.5%)	3.458	0.441
II (1%)	1.981	0.415
III (2%)	1.304	0.204
IV (3%)	1.315	0.318
V (5%)	2.474	0.384
VI	3.078	0.980

process. The data we get are in fact approximate values that are around the real ones. Therefore, a fluctuation of the content of composition will cover the real value and thus the error is "diluted". In this case, because the fluctuation amplitude is so small (0.5%) that the "information" included in the virtual samples is nearly identical to the original data, thus the improvement is insignificant.

When the fluctuation amplitude is up to 1% in virtual sample II, the prediction error shows a considerable decrease by nearly a factor of two as compared to that of the networks trained by the original data. This is because, as we have pointed out in above, the virtual samples developed in a proper fluctuating amplitude are essentially

real samples, and more "information" about the relationship is provided than that in the original data. Thus the increase of the prediction ability is not surprising. As we can see from Fig. 1 or table 3, when the fluctuation amplitude is up to 2% the performance of the prediction has been further improved in both the prediction value and the stability, as the average prediction error drops to 1.30% and the standard deviation to 0.204. With an increase in the fluctuation amplitude to 3%, the prediction error is still small but the stability of the prediction is a little worse. This gives a hint that 2% is the best amplitude to develop the virtual samples for the slag glass-ceramic system under study. With a further increase in the amplitude to 5%, we can discover that the virtual samples developed from No. 2 and No. 3, or those developed from No. 6 and No. 7 original samples in table 1, overlap in the content of major compositions. As a result conflicting data were introduced in this virtual sample set and consequently the prediction performance of the networks trained deteriorates.

Another mode is also used to fabricate virtual samples. In this mode, all the 10 components in table 1 were used to develop virtual samples with a fluctuation amplitude of 3%. An eleven factors and two levels orthogonal design table was used to produce virtual samples, thus from each original sample 12 virtual samples were developed. By this mode a data set consisting of 132 virtual samples were obtained from the original sample set and denoted as data set VI. The prediction performance of the networks trained by virtual sample VI, as shown in Fig. 1, is very poor especially in the prediction stability. This is because the microstructure and properties are very sensitive to the contents of minor components, virtual samples developed by this mode distort the real relationship between composition and property thus the predictions are poor. This simulation hints that strictly controlling the the contents of minor component, especially the crystal nucleation agents, is very important in the preparation of slag glass ceramic materials.

An interesting phenomenon appears in Fig. 1 is, as we have pointed out in our previous study, that when a network has $2M+1$ hidden nodes, the prediction error ε_{ra} is relatively small. Therefore in the following study networks with a topology of 10-21-1 were adopted.

The influence of virtual samples on the "overtraining" phenomenon

The backpropagation network, which is probably used in 95% of the existing applications [9], suffers from a phenomenon called "overtraining" shown in Fig. 2. In this case, the error in the training set will monotonically decrease and finally reach a minimum value as training continues. However, the performance on the test set increases for a while but then gets worse again when it reaches to the overtraining point. The best results for the specified neural network architecture are obtained when the training process is stopped at the moment the

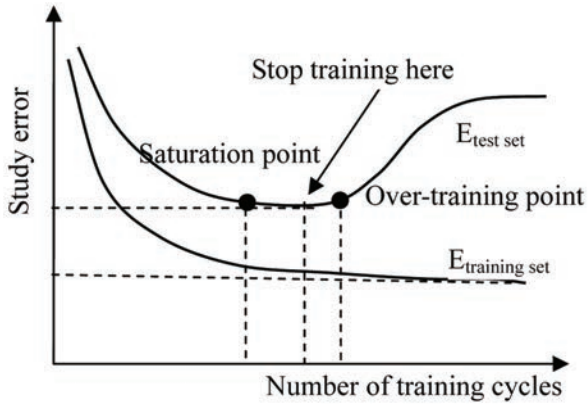

Fig. 2. The overtraining phenomenon.

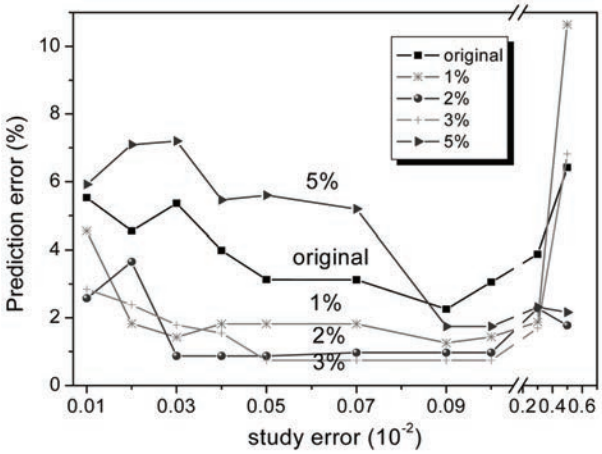

Fig. 3. Curves of prediction error versus study error for networks trained by different data sets.

error in the test set reaches its minimum [2]. A small positive numerical value, η, which is called the "study error", is preset to control whenever the training process terminates. In the study present here, for each training sample l if $|t_l-y_l| < \eta$ then the training process terminates, where t_l is the network's output value and y_l the experimental value. If an improper small η is chosen the neural networks will encounter the "overtraining" problem, while an improper large η will make the training stop before the network learns the exact relationships. This subtle choice of study error is troublesome to most developer for, as our knowledge, only a "trial and error" method can be used to decide a proper study error.

Networks with a topological structure of 10-21-1 were trained by the original data set as well as virtual data sets II, III, IV, and V, respectively, with study errors varying from 0.006 to 0.0001. The curves of prediction error as a function of study error are shown in Fig. 3. With a decrease of the study error, all curves have a rapid descent first and then reach to a minimum. A further decrease of η the performance of prediction will retain the saturation status for a while before becoming worse again. This is the typical "overtraining" phenomenon. We denote the range from the saturation point to the

Table 4. The saturation value and range for networks trained by different data sets

Batch	Saturation Value(%)	Saturation Range
Original	3.2	0.001-0.0005
II(1%)	1.8	0.001-0.0002
III(2%)	0.9	0.001-0.0003
IV(3%)	0.74	0.001-0.0005
V(5%)	1.7	0.001-0.0009

overtraining point as the saturation range (Fig. 2), and the average error in this range as the "saturation value". The saturation value represents the network's prediction ability and the saturation range denotes the stability of the prediction performance with regard to the study error. The latter will decide the space for a network developer to choose the study error. The saturation value and range for each curve are listed in table 4.

As shown in Fig. 3 and table 4, nearly all curves reach the saturation status at $\eta = 0.001$. As to the network trained by the original data its prediction performance gets worse at $\eta = 0.0005$. But for the networks trained by data set II, Their overtraining points do not appear until $\eta = 0.0002$. Thus the saturation range expands by a factor of two. The best performance of prediction is obtained from the networks trained by data set III which have a range of content of 2% in the major components. The saturation range is from 0.0002 to 0.001, and the prediction error in this range is as small as 0.9%. The network trained by data set IV has the smallest saturation value, 0.74%, but the saturation range is relatively small compared to that of the networks trained by data sets II and III. The prediction performance of the network trained by data V gets worse quickly after reaching its saturation value and thus has a very short saturation range. This simulation result agrees well with the quality of the virtual sample set we have analyzed in section 3.2. From these simulation results we can conclude that a good virtual sample set can not only improve the network's prediction ability, but can also suppress the "overtraining" phenomenon effectively, the reason for which is obvious, that is, plenty of effective training samples are used to train the network.

Conclusion

Based on the complex slag glass-ceramic system, we introduced a virtual sample technology to enlarge the training data for artificial neural networks through altering the content of composition over a proper small range. Research shows that a good virtual sample set can not only considerably improve the network's learning ability and prediction performance, but can also suppress the "over-training" phenomenon. We ascribed this to the extra data which are used to train the network and the extra information about the relationship to be disclosed provided by these virtual samples. It is worthwhile noticing that not all parameters are suitable to fabricate a virtual sample, for example, the parameters that the relationship are very sensitive to or those which only have discrete values. For most material systems such as ceramic, concrete, glass-ceramic, etc. this virtual sample technology can help the ANN model learn the intricate relationships from a small size data set, and give accurate predictions that won't vary with the change of network's topology of the hidden layer and the "study error" over a considerable range. So this new technology is possibly a powerful and reliable tool for data analysis with ANNs and can facilitate the design and development of complex material system.

Acknowledgements

This work was supported by the National Basic Research Program of China (973) under Grant No. 2007CB31407, Foundation for Innovative Research Groups of the NSFC under Grant No. 60721001, and the International S&T Cooperation Program of China under Grant No. 2006DFA53410.

References

1. Z.N. Xia, S.G. Lai, Z.B. Hu and Y.W. Lu, In: C.P. Sturrock and E.F.Begley (Eds.), Computerization and Networking of Materials Databases: Fourth Volume, ASTM STP 1257, American society for testing and materials, (Philadelphia, 1995) pp. 224-234.
2. H.M.G Smets and W.F. L.Bogaerts, In: C.P.Sturrock and E.F. Begley (Eds.), Computerization and Networking of Materials Databases: Fourth Volume, ASTM STP 1257, American society for testing and materials, (Philadelphia, 1995) p. 211.
3. J.J. Zhou, Q. Xie, J. Feng, S.M. Li, Z.H. Xu, L.J. Chen and Z.L. Gui, In: C.P. Sturrock and E.F. Begley (Eds.), Computerization and Networking of Materials Databases: Fourth Volume, ASTM STP 1257, American society for testing and materials, (Philadelphia, 1995) p. 235.
4. J. Kasperkiewicz, J. Mater. Proc. Technol. 106 (2000) 74-81.
5. Q.Y. Wen, Expert System for Slag Glass-Ceramics Based on Artificial neural networks, [MS thesis], Guangxi University, Nanning, 2001. (in Chinese)
6. Q.Y. Wen, P.X. Zhang and H.W. Zhang, Journal of inorganic materials, 18 (2003) 561-568. (in Chinese)
7. S.Q. Pan. New Glass, (Press of Tongji University, Shanghai, 1992) p. 122. (in Chinese)
8. E. Hisashi, N. Yoshikazu and K. Suzuki, Wat. Sci. Tech. 36 (1997) 235-241.
9. P. Harmon, Intelligent software strategies, 8 (1992) 1-16.