A new learning algorithm for incomplete data sets and multi-layer neural networks

Keiichi Bitou*, Yan Yuan*, Tomoo Aoyama*, and Umpei Nagashima**
* The Faculty of Engineering, Miyazaki University, Miyazaki, Japan
(Fax: +81-0985-58-7411; E-mail: tgb235u@student.miyazaki-u.ac.jp)
** Grid Technology Research Center, National Institute of Advanced Industrial Science and Technology
(Tel : +81-3-5246-6204; Fax: +81-3-5246-6208; E-mail: u.nagashima@aist.go.jp)

Abstract: We discussed a quantitative structure-activity relationships (QSAR) technique on incomplete data set. We proposed a new solver that used 2 kinds of multi-layer neural networks. One is to compensate the defect data, and another is to evaluate the QSAR. The solver can predict the defects in model QSAR data. By using them, we get very high precision QSAR. It is 5-10 times higher than that of a traditional method. However, in case of anti-cancer Carboquone, the prediction is not so complete. It was about *O(3)* wrong than the model calculation. The predicted values would have rather large error. It is caused by noisy observations of Carboquone. However, if we used the uncertain predictions, new data are included in QSAR. If not, they were omitted. The effect would not be little. Therefore, we evaluated the QSAR. The results are contrary to the expectation, are not so wrong. We believe that the wrong effect is suppressed by including information of new data.

Keywords: inverse problem, inverse optimization, QSAR, Carboquone, compensation, interpolation

1. INTRODUCTION

On many pharmaceuticals and chemical industries, multi-layer neural networks are useful computing resources, i.e., classifications and predictions. The reason is that the networks have forecasting functions on the quantitative structure-activity relationships (QSAR). It is an important technique to develop new compounds.

In the QSAR problems, many optimization techniques were published [1]. However, there is an unsolved problem. It is called as inverse (optimization) problem. That is to find functional relations in data including defects, and at the same time, to compensate the defects. The defect parts are distributed on whole data. The neural networks don't fit to solve the problem.

2. DEFINITION OF THE PROBLEM

A multi-layer neural network is vector-vector 'group' transformer, where these elements are different each other. We define the transformer as,

$$\boldsymbol{Tr}=\mathbf{NN}(\boldsymbol{Xr}) \qquad (1)$$

where "***Xr***" is a set of input vectors, and "***Tr***" is a set of output vectors. Using the character, we can do QSAR [2].Where "***Xr***" is observation data for compound's physical and chemical structures, and "***Tr***" is for physiological activities of that compounds. However, there are restrictions;

Restriction 1: Never defect part in learning data

Restriction 2: Must not fit in prohibition,

$$\boldsymbol{Xr} = \boldsymbol{Xs} \;\&\; \boldsymbol{Tr} \;!= \boldsymbol{Ts}.$$

QSAR data include defect parts in all cases; therefore, the restriction 1 limits application fields strictly. Even if the solver were not based on neural networks but statistical ones, the restriction 1 remained. Thus, the development of a new method is very important, which removes the restriction.

We have processed the defect part as followings [3];

Method-1: Remove all data including the defects from QSAR calculations.

Method-2: Replace the defects parts to averages that are calculated from other data set.

Method-3: Using prediction methods, compensate the defects, and calculate QSAR as non-defect case.

Method-4: Assuming a statistical model for the data set, evaluate the definitive parameters based on likelihood around the defects.

The method-1 is a simple approach, but it is not so wrong. It is handy and effective in case of small defect part. We often find the method equipped in commercial statistical program products.

However, if the defect part increases more, the efficiency becomes less; at the limit, the method cannot be applied.

On simple considerations, the method-2 is superior to method-1. However, recent researches show that there are wrong cases by comparison with method-1 [3].

We are also sure that there is no ground for replacement of averages.

The method-3 is practical and has wide application

fields; however, prediction methods are required. Efficiency of this method depends on ability of the prediction methods.

The method-4 would be the most accurate approach; however, the rule equation or distribution for observations must be known in advance. On QSAR, the requirement is a fatal restriction.

Then, we believe that method-1 and 3 are practicable. Our objective is, by using multi-layer neural networks, to show a kind of method-3 for QSAR, and examine it on actual medicines.

3. NOMENCLATURE

We assume that an index "i" corresponds to a chemical compound, and write the physical and chemical properties as;

$$\{X_{i0}, X_{i1}, X_{i2}, \cdots, X_{in}\} \quad (2)$$

Hereafter, we rewrite index "i" as digit series, ($0,1,2, \cdots, m$). We write the physiological activities of "i" compound as followings,

$$\{T_0, T_1, T_2, \ldots, T_3\} \quad (3)$$

The activities are sorted,

$$T_0 < T_1 < T_2 < \cdots < T_n \quad (4)$$

In this paper, we consider a kind of activities. The elements of $\{X_{ij}\}$ and $\{T_j\}$ correspond to each other. Where, the generality is kept.

Next, we introduce a matrix $\{M_{ij}\}$ whose elements are 0 or 1. Similarly, a vector $\{N_j\}$ is defined.

The $\{M_{ij}\}$ and $\{N_j\}$ correspond to $\{X_{ij}\}$ and $\{T_j\}$, respectively. When $M_{ij}=0$ or $N_j=0$, the corresponding data X_{ij} or T_j are lost.

4. TRADITIONAL NEURAL NETWORK APPROACHES

Using neural networks, to solve inverse optimization problems, some trials have been published [4]. When we solve the problems by using neural networks, a difficult task is that there is no means to take account of "non datum calculation". By considering the means simply, we may get that the learning process is to be locked as for the connections related to non-datum part. However, as tracing of the back-propagation algorithm, the idea is equivalent to a calculation in case of "defect-data=0." On the back-propagation algorithm, any means is not defined for non data.

If we leave out of the learning algorithm and adopt the reconstruction learning, the non data is replaced by uniform random numbers in interval [0, 1]. This would be only one way to take account of non-data. However, the replaced input data are random numbers. That is, we cannot expect the convergence of the reconstruction learning. In fact, the learning is never converged and the back-propagation errors are swaying for ever. But, the learning of other data except the non-data is converged. When the learning becomes such situations, the endless learning is forced to stop.

Next, uniform random numbers are inputted in the network. The differences between the network output and teaching data are compared each other. Then, the minimum difference is selected and the input random number is a solution of the inverse problem. We call these schemes "swaying reconstruction-learning method."

The method has some problems, which are too many CPU-times and to require teaching-data. Therefore, restricted and few type of inverse problems is solved.

5. INTER/EXTRAPOLATION THEORY

We discuss a new inter/extrapolation theory in this section. We consider a case, $M_{ik}=0$ and $0 \le k \le n$. Where, "k" represents plural cases. Then, observations, $X_i=\{X_{i0}, X_{i1}, X_{i2}, \cdots, X_{in}\}$, include plural defects, whose locations are pointed by $M_{ik}=0$. Now, we introduce a vector, $Y=\{Y_0, Y_1, Y_2, \cdots, Y_n\}$.The element number is equal to X_i. Except the case $M_{ik}=0$, using Y and Xi as input and teaching data, we can make a multi-layer neural-network learn them. After the learning completed, a relation, $X_{ij}=\mathrm{NN}(Y_j)$, j != k, is organized in the network. We can get Y-vector by sampling of an elementary function. So, Y_k can be evaluated by the function value corresponding to $M_{ik}=0$. If there were plural Y_k, the evaluation would be done. As the elementary function, we adopt a linear function. So, **Y** is arithmetic progression. In the function, prediction of Y_k is easy. If we substitute the Y_k into the neural network, we get $X_{ik}=\mathrm{NN}(Y_k)$. Thus, we can know the defect datum X_{ik}. Using the same processing, we can get $T_k=\mathrm{NN}(Y_k)$.

Where, we can consider an advanced method. That is, the **Y**-vector is replaced by the compensation vector $\{X_{ij}\}$. The method is iterative scheme. This is a kind of mehod-4 in section 2. However, the neural network is a non-linear function converter; therefore, the revising points would be small. As stated above, any defect in input and teaching data is compensated, and the compensation doesn't depend on the number of defect data. The estimated data are complete; therefore, we can analyze them by using another neural network. This is normal neural network QSAR. Hereafter, we call the method CQSAR (Compensated-QSAR).On the CQSAR, all observed data are used, where a lack of observations doesn't affect whole processing, and the wrong effect is suppressed. We wish to evaluate CQSAR numerically.

6. NUMERICAL CALCULATION FOR MODEL QSAR DATA

6.1 Generation of model data

We examine CQSAR for model functions. Because, they don't include noise, and it is convenience to check the true character. We selected 5 kinds of elementary functions;i.e,

$$\sqrt{x}\ ,\ x\ ,\ x^2\ ,\ 2(x-0.25)^2\ ,\ 10(0.1+(x-0.1)(x-0.5)(x-0.9))\ .$$

These functions simulate physical and chemical characters of substitute of various compounds. The definition interval is [0, 1]. Sampling them at 37 points regular intervals, and the sampled vectors are scaled within [0, 1]. The 5 vectors were used as input data for neural networks. Next, we sample one function, $1-(x-0.7)^2$. The function emulates physiological characters of the compound. We believe that such data should have a maximum point. Using same process, we got a teaching vector. We plotted them in Fig 1.

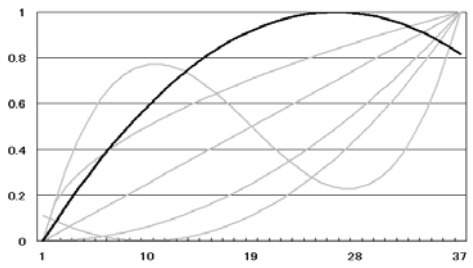

Fig 1: Plotting of QSAR model data.

Horizontal axis indicates compounds' numbers. Vertical axis is function values scaled in interval [0, 1]. A bold curve is plotting of teaching data, and pale curves are 5-kinds input data. The Fig 1 shows that there is a simple relation in learning data.

Those 6 kinds of vectors are all complete. We introduce following ***M***-matrix and ***N***-vector, and under them, we process the 6 vectors as incomplete ones.
We considered 5 M-matrices and 5 N-vector as followings.

Case 1: $M_{ii}=0\ ,\ 0\le i\le 4\ ,\ N_5=0$
Case 2: $M_{ii}=0\ ,\ 18\le i\le 22\ ,\ N_{23}=0$
Case 3: $M_{ii}=0\ ,\ 32\le i\le 36\ ,\ N_{37}=0$
Case 4: $M_{ii}=0\ ,\ 0\le i\le 4,6\le i\le 10,\cdots,N_5=N_{11}=\cdots=0$
Case 5: $M_{ij}=0\ ,\ \begin{matrix} i=even,\ j=\{0,2,4\} \\ i=odd,\ j=\{1,3\} \end{matrix}\ ,\ N_{(odd)}=0$

An object of the selection is;

Case 1: Defects are found continuously in the first part of observation.
Case 2: They are found in central part.
Case 3: They are found in terminal part.
Case 4: They distribute over whole data.
Case 5: They distribute as checkered pattern.

Since the defect ratio is small in cases 1-3 (=16.2%), by using method-1, they can be processed. However, in case 4, the ratio is 100%, and moreover in case 5, whole data is 50% lost. (In case 4, whole data is =16.7% lost.) On such cases, the method-1 cannot be applied.

When such a large percentage data are lost, normally usual methods would not process them. On other hand, CQSAR can be done for all cases; and it can take account of the unused data in column including defect datum.

6.2 Definition of neural networks

The defect parts are predicted by using 3-layer neural networks whose structure is;

(1) Input, hidden, output-layer's neurons are 2, 4, and 1, respectively.
(2) The neuron's emulation functions are a sigmoid-function for hidden and output-layers.

The non-linear function fitting ability of neural networks is caused by the sigmoid-function and the number. Therefore, it should be limited in a small number. The 4-sigmoid functions can simulate any function having two peaks. The fitting ability is not so much. We would prevent the neural networks from excessive learning. The design is important at first predictions for defect data.

(3) Initial guess of connection weights among neurons is uniform random numbers in [-0.5, 0.5].
(4) Back propagation learning, learning coefficients are 0.2 and 0.15 for hidden and output-layers.
(5) The learning iterations are 20K.

If the back propagation error is not converged to 0, the learning is stopped by force.

It is also a treatment to stop excessive learning. In case of model data, the errors were under $O(-5)$.

(6) Whole calculations are done by using 64-bits IEEE floating point format.
(7) The compiler is digital FORTRAN version 6 (produced by Dec).

After the predictions, normal QSAR was calculated by using secondary 3-layer neural network.

The network structure is below.

(1) Input, hidden, output-layer's neurons are 6, 8, and 1,

respectively.

The second network has many kinds of input data; therefore, we used large number of hidden-layer's neurons.

(2) Other conditions are same as the first networks.

6.3 Examination of CQSAR method

We examined the effects of CQSAR based on standard deviations.

Table 1: Standard deviations for defect parts

	CQSAR	Method-1
Case-1	0.0152	0.1301
Case-2	0.0046	0.0063
Case-3	0.0343	0.1107
Case-4	0.0132	impossible
Case-5	0.0195	impossible

The table shows prediction ability of two methods. The ability is represented by the standard deviations; therefore, "0" signifies complete predictions.
"0.1" means that error of 10% order is found.

Then, we can know the ideal standard deviation when complete data are given. It was 0.0062, which signifies calculation-precision of second network. Even if a lower value might be got, it was calculated by an accident. We believe the value under 0.0062 is non-significant.

CQSAR method is developed in this paper, and method-1 is traditional, which is column or row including a defect is left out of the calculations. On whole cases, CQSAR is superior to traditional.

On cases-3,4,5, CQSAR's predictions are 8.61, 1.37, 3.23 times accurate. Especially, on the cases-4 and 5, CQSAR gave reasonable standard deviations. These results are to be noticed.

Cases-1 and -3 show the network's precisions near two terminal points. They are less precisions compared with that of the case-2, which indicates the defects in central parts. It is well known that extrapolation ability of neural networks is less than interpolation. The simulations indicated the same results.

Comparing CQSAR and method-1, CQSAR gives high precision standard deviations at both terminals of observations. This shows that CQSAR method has extrapolation, which is a useful character for QSAR.

We checked the compensation ability of CQSAR in case-4 and 5. To check it, we plotted the differences between the outputs and real values in Fig 2.

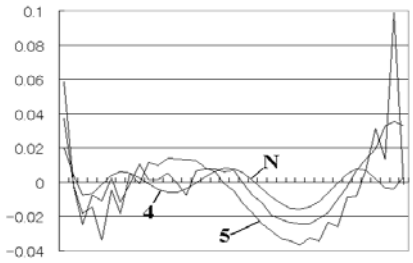

Fig 2: Differences between outputs and teaching data

Horizontal axis indicates compounds' numbers. Vertical axis is amplitude of the differences. Curve N is the case of non-defect data, which shows fitting-ability of the second neural network. Curve "4" is case-4, and "5" is case-5. They show fitting-abilities of this CQSAR method.

Thus, we believe that CQSAR is effective for a model QSAR calculation. We evaluated other model calculations that include many kinds of bend-functions and step-functions, which are not differential and continuous. Even if there are such un-fit points, where the influences are limited at local regions, CQSAR gave reasonable results.

7. NUMERICAL CALCULATION FOR ANTI-CANCER MEDICINES, CARBOQUONE

7.1 Characters of Carboquone observations

The effects of CQSAR should be examined for real medicinal data. Carboquone are anti-cancer medicines, the physical, chemical structure data, and physiological activities are published [5]. The activities are categorized as two cases that are continuous/one and mean-effective/optimum dose. We write them as MED/OD. These data have defect parts; however, the ratio is very small. So, we added $\mathbf{M}_{ij}$-matrix and $\mathbf{N}_j$-vector in section 5 to the published data, and examined the CQSAR effects. The observations (Fig 3) are quite different compared with the model data (Fig 1) in section 6.

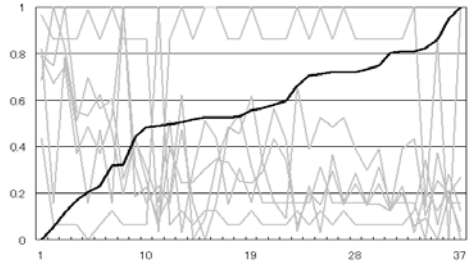

Fig 3: Plotting of Carboquone in case of continuous injection and MED.

MED is “mean effective dose.” Horizontal axis indicates Carboquone - derivatives’ numbers. Vertical axis is observations scaled in interval [0, 1]. A bold curve is plotting of teaching data, i.e., physiological anti-cancer activities, which are sorted. Pale curves are plotting of 6-kinds input data, which are physical and chemical observations of the derivatives. The figure shows that there are complex relations in learning data.

7.2 Neural network structures

The defect parts are predicted by using 3-layer neural networks whose structures are same in section 6. The back-propagation error converged to $O(-1.5)$ for compensations of input data defects, and $O(-3)$ for teaching data. The values are not sufficient. But, considering the characters of learning data, they should be accepted. If we used more neurons on a hidden-layer, the error might be less. However, the direction doesn’t equal to increase the prediction ability of the neural network. By using many parameters on neural networks, it is impossible to get accurate predictions, as well as statistical methods. Here, it is true that less information gives uncertain predictions. In this section, the objective is to research influences of the uncertain compensations to QSAR of Carboquone.

The structure of second neural network is followings;

(1) Input, hidden, output-layer’s neurons are 7, 8, and 1, respectively.
(2) Other structures are same as section 6.

When the structures were used, the back- propagation- error of 20K iterations was 0.0047 and 0.0046 for CQSAR and method-1. They are $O(-2.5)$ and are same level to that of model data in section 6.

Some one may feel it is not sufficient. We believe that forced convergence means excess-learning. So, we stopped the learning on this level. To check content back-propagation-error for full set of learning data, we calculated the non-defect case on 20K iterations. The error value was 0.0059; that is $O(-2.5)$. So, we believe that the CQSAR extracts whole information in Carboquone observations.

7.3 Calculated standard deviations

We show that the prediction values of CQSAR are not so accurate. If there were the wrong effects, standard deviations of Carboquone-QSAR were larger than that of traditional method-1. Because, in the method-1, there is no uncertain prediction. We test them on the continuous mean-effective dose and the optimal dose. The results are listed in table 2.

Table 2: Standard deviations for Carboquone continuous MED/OD.

	continuous MED		continuous OD	
	CQSAR	Method-1	CQSAR	Method-1
Case-1	0.044	0.377	0.043	0.371
Case-2	0.042	0.087	0.042	0.067
Case-3	0.046	0.141	0.030	0.159
Case-4	0.052	impossible	0.054	impossible
Case-5	0.040	impossible	0.079	impossible

The table-2 shows prediction ability of two methods. The ability is represented by the standard deviations.
The base data are Carboquone physical, chemical observations, and physiological activities.
These data would be much noises. The standard deviations are calculated by artificial masking data for defects. The masking data are categorized by 5-cases. When the masks do not exist, we can evaluate ideal standard deviations. They were 0.039 and 0.041.

On whole cases, CQSAR is superior to traditional method-1.On cases-3, 4, 5, CQSAR’s predictions as MED were 8.6, 2.1, 3.1 times accurate. As OD, they were 8.6, 1.6, and 5.3.

Here, extrapolation of CQSAR method was found, same as the case of model data. On the cases-4 and 5, CQSAR gave reasonable standard deviations. We interpret it as followings.

Even if we could make uncertain predictions, the wrong effects might be small. The predictions saved some observations that were not used in traditional methods; and the effects would exceed the demerits. The extrapolation function was also found.

There is another observation, 1-injections, for Carboquone. We calculated them, and checked effects of CQSAR. That is in table 3.

Table 3: Standard deviations for Carboquone 1- Injection MED/OD.

	1-injection MED		1-injection OD	
	CQSAR	Method-1	CQSAR	Method-1
Case-1	0.043	0.281	0.026	0.063
Case-2	0.071	0.103	0.040	0.137
Case-3	0.049	0.110	0.045	0.047
Case-4	0.045	impossible	0.057	impossible
Case-5	0.067	impossible	0.083	impossible

The table-3 shows prediction ability of two methods, CQSAR and traditional method-1.

The ability is represented by the standard deviations.
The base data are Carboquone physical, chemical observations, and physiological activities. These data would be many noises. The standard deviations are calculated by artificial masking data for defects. The masking data are categorized by 5-cases. When the masks do not exist, we can evaluate ideal standard deviations. They were 0.047 and 0.047.

As well as continuous MED/OD, in whole cases, CQSAR is superior to traditional method-1.
On cases-3, 4, 5, CQSAR's predictions as MED were 6.5, 1.5, 2.2 times accurate, and as OD, they were 2.4, 3.4, 1.0. On the cases-4 and 5, CQSAR gave reasonable standard deviations, 0.045-0.083.

Thus, we are sure that processing's introduced in CQSAR method are reasonable and should be acceptable.

8. CONCLUSIONS

We discussed a QSAR-technique on incomplete data set. We proposed a new solver that used 2 kinds of multi-layer neural networks. One is to compensate the defect data, and another is to evaluate the QSAR.

1) It can solve problems that a traditional method cannot process.
2) It has prediction-ability to compensate 50% defects of observations.
3) It revises standard deviations of QSAR about 2-5 times.

The solver can completely predict the defects in model QSAR data. By using them, we get very high precision QSAR. It gives 5-10 times accurate standard deviations in comparison with a traditional method. We tried to other 5 models, and got same effective results.

However, in case of anti-cancer Carboquone, the prediction was not so complete. It was about $O(3)$ wrong than the model calculation. The predicted values would have rather large error. It is caused by noisy observations of Carboquone. However, if we used the uncertain predictions, new data were included in QSAR, which were omitted in the traditional method. The effect would not be little. Therefore, we evaluated the QSAR. The results are contrary to the expectation, are not so wrong. We believe that the wrong effect is suppressed by including information of new data. Thus, we got 1-9 times accurate standard deviations of Carboquone QSAR.

Same results are got in two cases of anti-anxiety effect of Benzodiazepinooxazolidine and tranquilizer Diazepam. The CQSAR-method is also operated accurately on actual medicines. We believe that the proposed method has practical usability. We would like to publish the Fortran program-codes and their data on Internet; the URL is,

http://www.miyazaki-u.ac.jp/aoyama_t/index.html.

REFERENCES

[1] H.Zhu, T.Aoyama, I.Yoshihara, S.Lee, W.Kim, "Precision indices of neural networks for medicines: structure-activity correlation relationships", *Proc. of 15th Korea Automatic Control Conference*, CD-ROM (40rd.pdf), 2000.10.20.

T. Aoyama, Q. Wang, U. Nagashima, "Reinforcement of extrapolation of multi-layer neural networks", *Proc. of International Joint Conference on Neural Networks'01*, *CD*-ROM (492.pdf), 2001.7.14-19.

[2] T.Aoyama, Y.Suzuki, H. Ichikawa,"Neural networks applied to structure activity relationships", *J. Med. Chem*. Vol.33, pp.905-908, 1990.3.

T.Aoyama, Y.Suzuki, H.Ichikawa,"Neural networks applied to quantitative structure activity relationship", *J. Med. Chem.*, Vol.33, pp.2583-2590, 1990.9.

[3] M. Watanabe, K. Yamaguchi, "EM Algorithm and the problems for incomplete data set (in Japanese)", *Taga Publishing Co.Ltd.*(2000, Tokyo), ISBN4-8115 -5701-8.

[4] Q.Wang, T. Aoyama, U. Nagashima, E-S. Kang, "Inverse optimization problem solver on use of multi-layer neural networks", *Proc. of International Conference on Control Automation and Systems'01*, CD-ROM (949.pdf), 2001.10.17-21.

[5] T. Fujita and Research group for Structure-Activity Relationships, "Structure -Activity Relationships, Quantitative Approaches; The Significance in Drug Design and Mode-of-Action Studies (in Japanese)", *Nanko-do Publishing Co. Ltd.*, (1979, Tokyo).