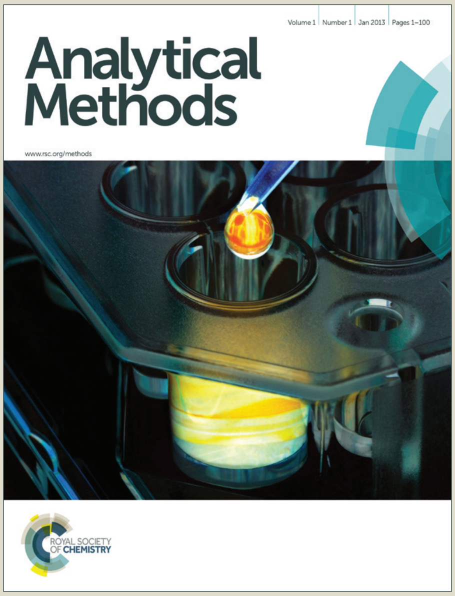
Analytical Methods

Accepted Manuscript



Volume 1 | Number 1 | Jan 2013 | Pages 1–100

Analytical Methods

www.rsc.org/methods

ROYAL SOCIETY OF CHEMISTRY

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

A simple and effective method for picking training samples in neural networks

Fanbiao Kong, Guoqing Chen*, Chun Zhu, Run Li, Yangjun Hu, Yong Zhang, Zhuowei Zhu

School of Science, Jiangnan University, Wuxi 214122, China

Abstract: A simple and effective method for picking training samples in neural networks is proposed. The synchronous fluorescence spectra of 85 standards of Azorubin and New Red mixed with concentrations ranging from 5μg·ml^{-1} to 20μg·ml^{-1} were obtained by synchronous scanning the excitation and emission monochromator maintained at an offset of 70nm. The radial basis function neural networks (RBFNN) were used. The whole analytical properties domain was divided into nine small areas. A sample was placed into every small area. Numbers and distribution of the training samples were decided according to the accuracies of the samples placed. The method was completed in three steps in this work. Finally, the completed RBFNN was fully tested and the results were satisfactory with the root mean square error (RMSE) was 0.4745 and the total mean relative error (MRE) was 0.0338. The testing results show that the method proposed is simple and effective.

Key words: Training samples, synchronous fluorescence, RBFNN, Azorubin, New Red.

1. Introduction

Linear regression and multiple variable linear regression are often used in the quantitative determination. Reference[1] uses high-performance liquid chromatography(HPLC) and linear regression to determine the pyrethroids in porcine tissues. In reference [2], fluorescence spectroscopy coupled with multivariable linear regression is used to the simultaneous determination of multicomponent mixtures in micellar medium. Compared with chromatography and other spectroscopy, the fluorescence spectroscopy has an advantage because of its high sensitivity and selectivity[3,4]. Synchronous fluorescence which can be obtained by simultaneous scanning the excitation and the emission monochromators at a constant offset is one kind of the fluorescence spectroscopies. The synchronous fluorescence is widely used with the advantages of narrowing bandages and avoiding of Rayleigh scattering[5,6].

However, one of the disadvantages of the linear regression and the multivariable nonlinear regression is that their applications are limited to linear situations[7,8,9], in which the relationship between the analytical property and the concentrations of the target analytes must be linear. In some simple nonlinear cases, some mathematical methods may still be used[10]. But in other nonlinear cases, the mathematical method can hardly be applied, especially when the analytical property is more than one. Although any high concentration can be diluted until the relationship becomes linear.

* Corresponding author. Tel.:+8613906176695;
E-mail address: cgq2098@163.com.

Analytical Methods Accepted Manuscript

But some of the reagents used for diluting may affect the analytical results. Thus, a method which can be applied straightforward without any pretreatment is advantageous.

Artificial neural networks are one of the most effective solutions for nonlinear analysis. It has the potential to approximate any continuous function[11]. Although, the concrete form of the function can not be got, the relationship is built on the structure of the artificial neural networks and a series of its parameters[12]. However, one of the difficulties of the applications of artificial neural networks is to decide how many training samples should be used and the distribution of these training samples. Too many training samples are used; the networks may be over trained. Too little, the networks may be insufficiently trained. Both of the situations will lead to the incorrect results and will cause some deviations. Too many training samples will be a waste of time and resources. Some methods are proposed, but the theory is complicated and the procedure is tedious[13].

The aim of this work was to develop a simple and effective method to decide the numbers of the training samples and the distribution of these samples. In this work, radial basis function neural networks (RBFNN) are used. RBFNN is a class of feed-forward neural networks which is widely used for classification problems[14,15] and function approximation[16], which have some useful advantages, namely, fast convergences, smaller extrapolation errors and higher reliability[17].

2. Experiment

2.1. Reagents

Azorubin and New Red were purchased from Dr.Ehrenstorfer GmbH. Azorubin and New Red are synthetic colorants which are widely used to supplement and enhance natural colors of food destroyed during processing and storage[18,19]. Stock standard solutions (50μg.ml^{-1}) of Azorubin and New Red were prepared by dissolving the pure solid in ultrapure water bought from Huajing Research Laboratory.

2.2. Apparatus

All the synchronous fluorescence spectra were recorded on an Edinburgh FLSP920 spectrofluorimeter produced by Edinburgh instruments. The FLSP920 is a modular and computer controlled spectrofluorimeter for measuring steady state luminescence spectra in the ultraviolet to near infrared spectra range with single photon counting sensitivity. It combines ultimate sensitivity with high spectral resolution and excellent stray light rejection. Four 10mm x 10mm x 45mm cuvettes were used for the synchronous fluorescence spectra acquisition at the right-angle geometry. The cuvettes were washed with ultrapure water and dried with air.

The data obtained were exported in ASCII format and transferred to a Core(TM)2 PC having 4G for RAM(Windows XP operating system) for subsequent manipulation. The RBF neural network programs and the wavelet programs were written in MATLAB (Mathworks, Version2008a). The figures were also plotted in MATLAB.

2.3. Experimental Design

Analytical Methods Accepted Manuscript

An RBF neural network consists of three layers, namely the input layer, the hidden layer and the output layer. Using the MATLAB Neural Network Toolbox functions, we chose the exact RBF neural network. The nodes of the input layer are equal to the dimensionality of the input vector. The hidden layer consists of the same number of computation units as the size of the training samples, by which the centers of the RBF functions are directly determined. In an exact RBF neural network, the training error is zero. The performance of the RBF neural networks is highly dependent on the choice of centers and width of the RBF functions. Details of the introduction of RBF neural networks can be found in several references[20, 21, 22].

85 samples were prepared for the experiment. They were prepared by mixing the appropriate quantities of the two stock standard solutions and the ultrapure water. Before filling into the cuvettes, the samples were shaken vigorously to be as homogenized as possible. 25 samples were picked and divided into two sets as depicted in Fig1. The circles represent the samples for calibration while the stars for validation. The concentrations domain was divided into nine small areas. A sample was placed in every small area. In this article, the neural networks were used for further prediction of the unknown samples. Therefore, the results will be better if the numbers and distributions of the training samples are decided based on the samples placed. The function relationship between the fluorescence intensities and the concentrations of the Azorubin and the New Red in different areas can be exploited according to the accuracies of the results of the samples placed. Then some training samples will be added to improve the accuracies which are not good enough. The more complicated the relationship was in an area, the more samples needed to be added to that area.

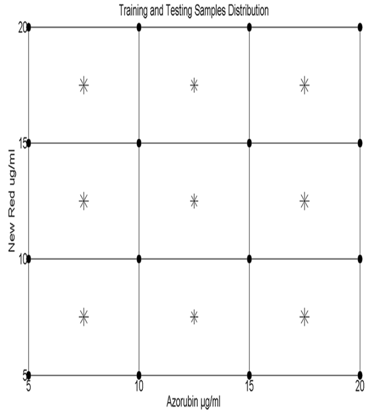


Fig.1 The distribution of the samples in the first step. The circles represent the training samples while the stars the testing samples. The whole area is divided into nine small areas which are labeled with numbers.

Once the training samples were settled, the number of the computation units and the centers of the RBF functions were decided. The whole spectra of the samples would be inputs of the RBF neural networks. The Gaussian function was chosen as the RBF functions in this work. Finally, the width of the Gaussian functions was settled when the minimum of RMSE reached.

All the measurements were carried out with the consistent instrument parameters. The slit widths of excitation and emission monochromators were set at 5nm. The integration time was kept 0.1s. Room temperature and humidity were controlled. For the 85 samples, Δλ=70nm was chosen as the offset between the excitation monochromator and the emission

Analytical Methods Accepted Manuscript

monochromator and the synchronous spectra of the samples were acquired with excitation wavelengths scanning from 200nm ~ 650nm in the interval of 1nm.

2.4. Data processing

The wavelet transform (WT) has been extensively used for the digital image and signal processing. A discrete wavelet transform (DWT) up to the third level was performed to denoise the spectra data using the Haar wavelet[23]. Owing to the fact that the details associated with noise and the approximation signal mostly contained the important part of the original signal[24], the high frequency coefficients of the third level were replaced by zeros while the other coefficients were kept. Then the inverse wavelet transform was applied to reconstruct the spectra data.

After that the spectra data were normalized in the range 0.1~0.9. The normalization was done using Eq(1)

$$y = 0.1 + 0.8 * \left(\frac{x - x_{\min}}{x_{\max} - x_{\min}} \right) \qquad (1)$$

Where $x_{\max}$ and $x_{\min}$ were the maximum and minimum values of the parameters, respectively. x was the actual value and y was the normalized value of x.

3. Results and Discussion

3.1 Synchronous Fluorescence Spectra

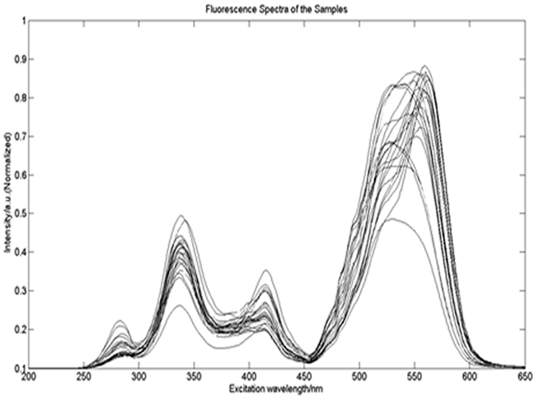


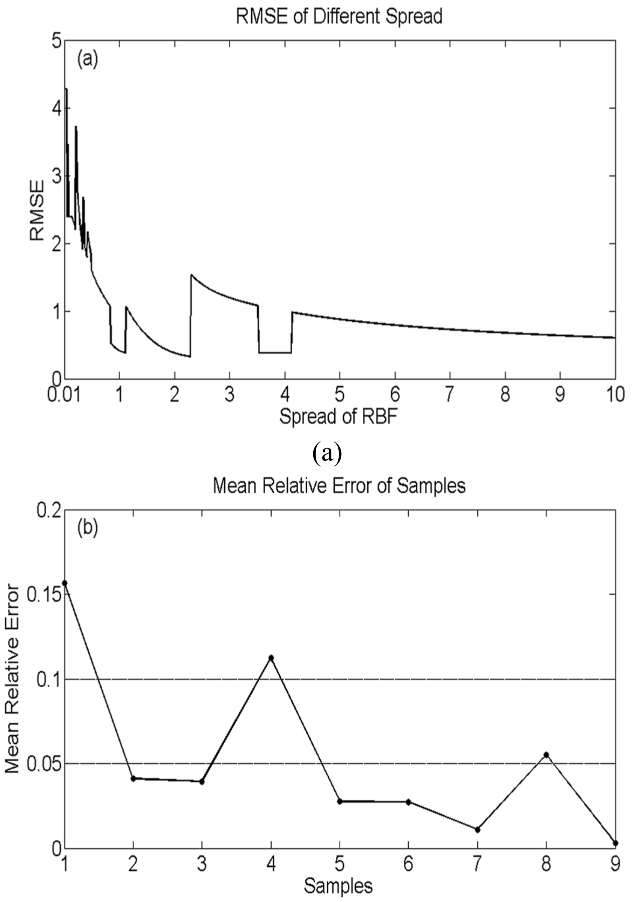
Fig.2 The synchronous fluorescence spectra of the samples used in this work.

The synchronous fluorescence spectra of the samples used in this work are shown in Fig.2. Every spectra have four main peaks. The fluorescence intensities of the peaks change with the concentrations of the Azorubin and the New Red.

3.2 Results of the RBFNN and Discussion

3.2.1 Results and discussion of the first step

Analytical Methods Accepted Manuscript



(a)

(b)

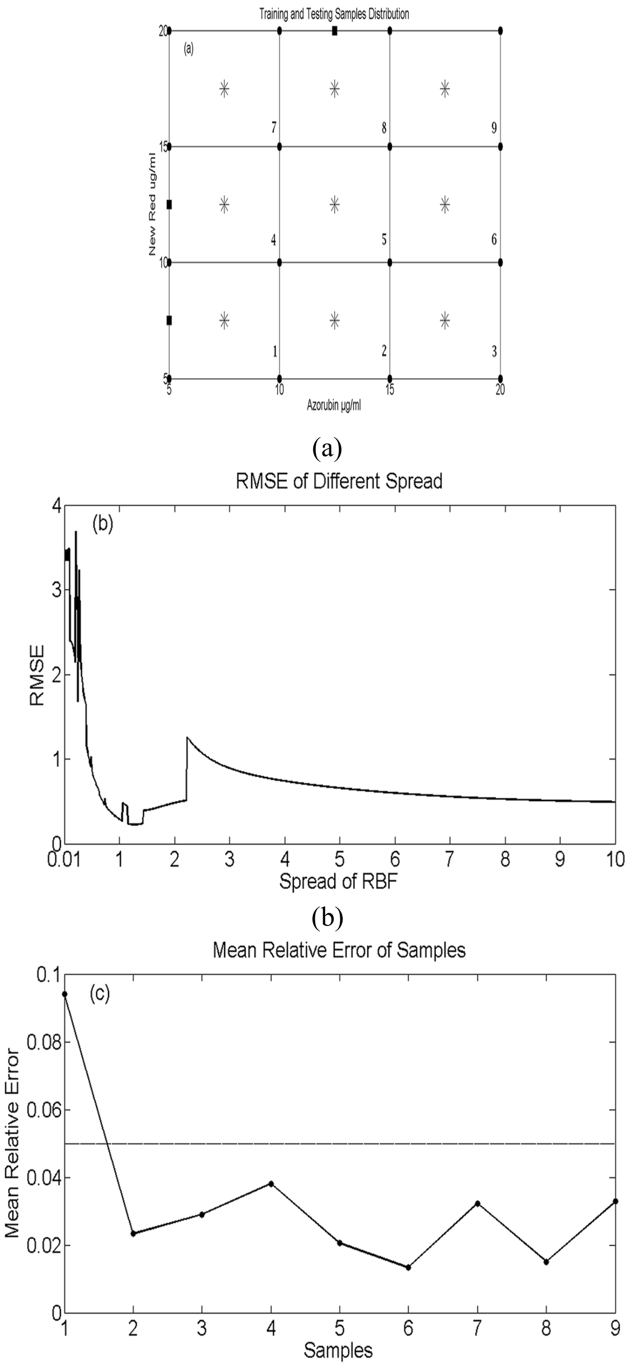
Fig.3 Results of the first step. (a) the RMSE with the spreads of RBF. (b) the mean relative error of the testing samples in the first design.

The results of the experiment designed above are shown in Fig.3. The RMSE changes with different values of the width (spread) of the Gauss functions. The minimum of the RMSE is reached when the spread is 2.29. With the optimized parameter, the MREs of the testing samples are depicted in Fig.3 (b). The MREs can be divided into three ranges, the range above 0.1, the range between 0.05 and 0.1 and the range below 0.05. The MREs of area 1 and area 4 are in the first range, the MRE of area 8 was in the second range and the MREs of the other areas are in the third range. The larger the deviation of the predicting result of the sample placed in every area, the more complicated the relationship is in that area. Therefore, the complication rankings were areas 1, 4> area 8 > other areas according to the accuracy of the sample placed in every area. The more complicated the ranking is, the more training samples should be added to the area.

Thus areas 1, 4, 8 should be added more training samples in that their accuracies were not good enough.

3.2.2 Results and discussion of the second step

Analytical Methods Accepted Manuscript



(a)

(b)

(c)

Fig.4 The design and results of the second step. (a) the squares represent the training samples added. (a) the relationship of the RMSE and the spread of RBF. (c) the mean relative error of the testing samples.

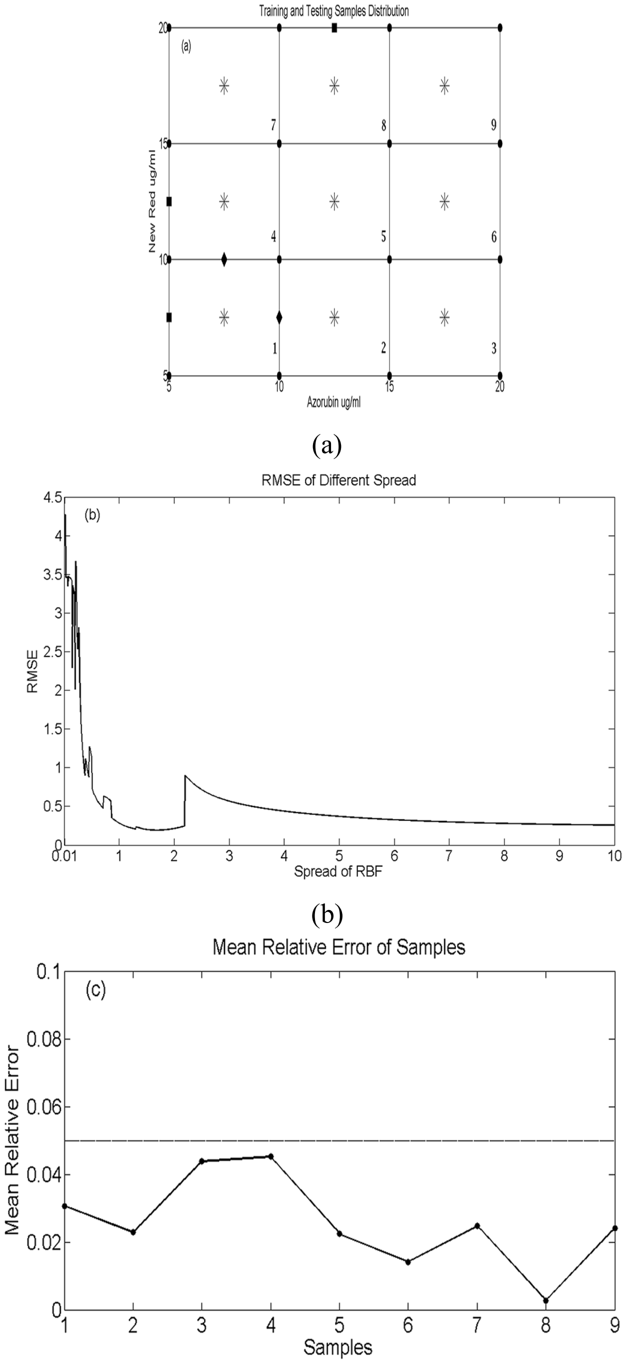
As depicted in Fig.4 (a), the squares represent the samples added to optimize the areas 1,4 and 8. The best spread and the minimum of RMSE which is 1.27 and 0.2237, respectively, can be obtained from Fig.4 (b). It can be seen from Fig.4(c) that the RMEs in area 1,4 and 8 are optimized. The RMEs of number 4 sample and number 8 sample are now in the third

Analytical Methods Accepted Manuscript

range. The RME of number 1 sample drops to the second range. Thus, the complication rankings are area1 > area 4 > area 8 > other areas.

However, the RME of the number 1 sample is still above 0.05 and it should be optimized. More training samples should be added to the area 1.

3.2.3 Results of the third experiment and discussion



(a)

(b)

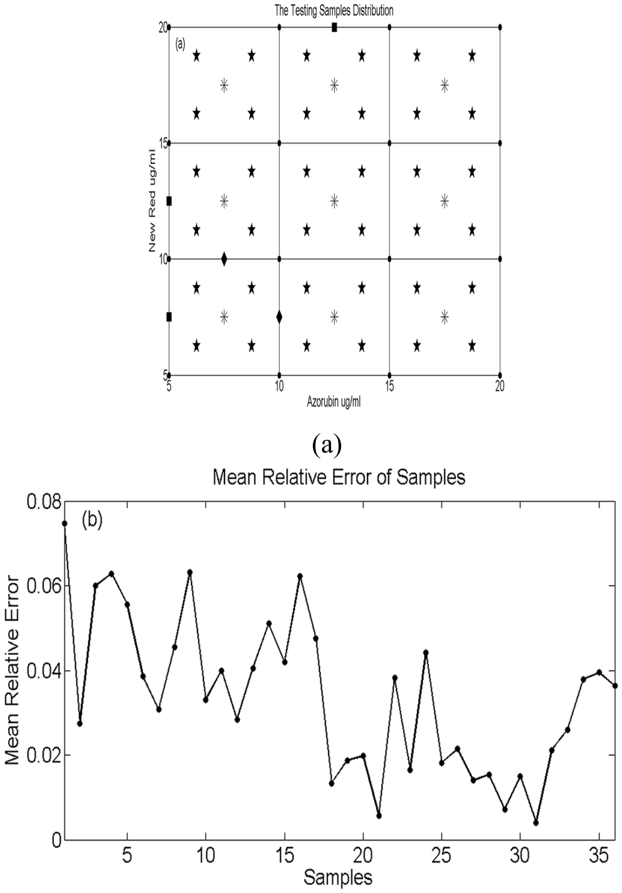
(c)

Fig.5 (a) the diamonds represent the training samples added. (a) the relationship of RMSE and the spread of RBF. (c) the mean relative error of the testing samples in the third design.

Analytical Methods Accepted Manuscript

The diamonds in Fig.5 (a) represents the training sample added to area 1 in the third step. 0.1922 is the minimum of RMSE and 1.68 is the best spread, which can be obtained from Fig.5 (b). Fig.5(c) shows the RME of number 1 sample is optimized into the third range. Now, all the RME are below 0.05 and the training results are acceptable in the whole area.

3.3 Results and discussion of the testing samples



(a)

(b)

Fig.6 Results of the testing samples. (a) The pentagrams represent the samples used for testing.

As shown in Fig.6 (a), the pentagrams represent the samples used for the testing of the well-trained RBFNN. In order to fully test the RBFNN, 36 samples were used with four testing samples placed in every small area. The distribution of the testing samples covers the whole concentrations domain. Fig.6 (b) shows the mean relative error of the testing samples. The RMSE of the results is 0.4745and the total MRE is 0.0338. Results show that the method is of satisfactory accuracy.

The structure of the networks was completed in three steps. The method can be applied to other analytes and in more extensive concentrations. And the accuracy can also be set in light of requirements. The number of areas, training samples and testing samples in every small area can be changed according to the need and more steps may be needed.

Analytical Methods Accepted Manuscript

4. Conclusions

In this work, a method for picking training samples of RBFNN is proposed. Meanwhile, a method for simultaneous determination of the Azorubin and the New red by synchronous fluorescence spectra coupled with RBFNN is demonstrated. Results show that the methods in this work are of satisfactory accuracy. The procedure is simple and effective. It is possible that both of the methods can be extended in other situations and applications.

Acknowledgments

The authors appreciate the financial supporting from the Natural Science Foundation of China (NO.61178032 and NO.61378037) and the Science and the Technology Support Program of Jiangsu Province (BE2011828).

Reference

1. J.H.Cheng , M.Liu , Y.Yu , X.P. Wang, H.Q Zhang , L.Ding , H.Y. Jin, Meat Science, 2009,**82** 407–412.

2. G.D. Armas, M.Miró , J.M.Estela , V.Cerdà, Analytica Chimica Acta, 2002, **471,** 173–186.

3. A.B. Moreira, I.L.T. Dias, G.O.Neto, E.A.G. Zagatto, L.T. Kubota, Analytica Chimica Acta, 2004, **523** 49–52.

4. G.Q.Hui , Y.C.Zhang , F.L.Cui , G.R.Qu, Measurement, 2013, **46,** 1507–1511.

5. D.Patra, A.K. Mishra, trends in analytical chemistry, 2002, vol. 21, no. 12, .

6. F.L.Cui, Q.Z. Zhang, X.J.Yao, H.X.Luo, Y.Yang, L.X.Qin , G.R.Qu , Y.Lu,Pesticide Biochemistry and Physiology, 2008, **90,** 126–134.

7. D.M. Reynolds, Water Research, 2003, **37**, 3055–3060.

8. C.L.Tong , X.J.Zhuo,Y.Guo,Y.H.Fang, Journal of Luminescence, 2010, **130**, 2100–2105.

9. A.C. Olivieri, G.M. Escandar, A. M.Pena, Trends in Analytical Chemistry, 2011, Vol. 30, No. 4.

10. N.Morsy, D.W.Sun, Meat Science , 2013, **93,**292–302.

11. K.Baddari, T.Aıfa, N.Djarfour, J.Ferahti, Computers & Geosciences, 2009, **35**, 2338–2344.

12. S.Haykin, Neural Networks and Learning Machines, third ed., China Machine Press, 2009, pp. 21-31.

13. F.Tong , X.l.LIU, 2005, **10(2)** , 233-239.

14. G.A.Montazer, H.Khoshniat, V.Fathi, Applied Soft Computing, 2013, **13**, 3831–3838.

15. R.X.Zhang, G.B.Huang, N. Sundararajan, P. Saratchandran. ,Neurocomputing, 2007, **70** 3011–3018.

16. I.Yilmaz, O.Kaynar, Expert Systems with Applications, 2011, **38**, 5958–5966.

17. H.Moradkhani, K.L.Hsu, H.V.Gupta, Journal of Hydrology, 2004, **295**, 246–262.

18. K.S. Minioti, C.F. Sakellariou, N.S.Thomaidis, Analytica Chimica Acta, 2007, **583**, 103–110.

19. Y.M.Huo，J.Wang，H.Zhang，J.H.Zhu，Y.M.Liu，Y.L.Wang, Journal of Instrumental Analysis, 2011, **30(6)**, 670-673.

20. S.Haykin, Neural Networks and Learning Machines, third ed., China Machine Press, 2009, pp. 230-267.

21. A.SZhang, L.Zhang, Computers and Structures, 2004, **82**, 2333–2339.

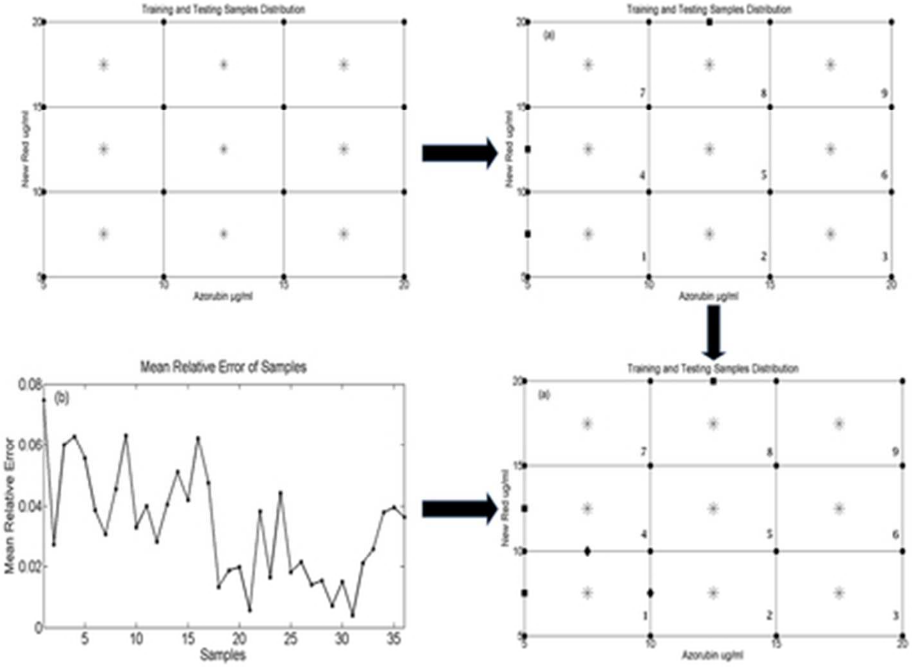
22. M.A. Behrang, E. Assareh, A. Ghanbarzadeh, A.R.Noghrehabadi, Solar Energy, 2010, **84** 1468–1480.

23. D.Giaouris, J.W.Finch, Electric Power Systems Research, 2008, **78**, 559–565.

Analytical Methods Accepted Manuscript

24. L.K.MSc, J.Salau, I.Traulsen, Journal of Equine Veterinary Science, 2012, **32**, 696-703.

Analytical Methods Accepted Manuscript



A simple and effective method for picking training samples in neural networks
39x28mm (300 x 300 DPI)

Analytical Methods Accepted Manuscript