Processing of spectrophotometric array signals using an artificial intelligence method

Ling Gao and Shouxin Ren
Department of Chemistry
Inner Mongolia University
Huhhot, China
lingyuxi@hotmail.com
cersx@mail.imu.edu.cn

***Abstract*—This paper addresses processing of spectrophotometric array signals based on genetic algorithms (GA) least square support vector machines (LS-SVM) regression to provide a powerful model for machine learning and data mining. The key to complete LS-SVM regression is to choose its optimal parameters. Due to their outstanding ability in solving global optimization problems in complex multidimensional search space, GA are used in this study to obtain the optimal parameter combination of the LS-SVM model. Experimental results showed the GA-LS-SVM method to be successful for simultaneous multicomponent determination even where severe overlap of spectra was present.**

Keywords- Least squares support vector machines; Genetic algorithms; Spectrophotometric array signals; Overlapping spectra; Artificial intelligence

Nowadays, with the application of photometric diode array detector and computers, rapid scanning commercial spectrophotometers are capable of quickly generating huge data consisting of hundreds and even thousands of absorbance values per spectrum. The array data named full-spectrum contain sufficient information to be able to determine the contents of various compounds. The main drawback of ultraviolet-visible (UV-VIS) is its poor selectivity because in many cases UV-VIS spectra display strong overlaps, especially some less specific and selective chromagenic reagents often give rise to strongly overlapped spectra in many cases. The combination of artificial intelligence methods with the computer-controlled spectrophotometers was proven to be effective in overcoming this difficulty [1-3]. Artificial neural network (ANN) is a form of artificial intelligence that mathematically simulates biological nervous system [4, 5]. However, ANN often has slow convergence, is prone to the existence of many local minima during training, and has a tendency of overfitting. Recently, a promising technology called support vector machines (SVM) has been used for classification and regression problems. SVM pioneered by Vapnik is a kind of machine learning method based on modern statistical learning theory and has notable properties including absence of local minima and high generalization ability [6, 7]. Suykens and his coworkers [8] introduced a modified version of SVM called least square SVM (LS-SVM) , which requires solving a set of linear equations instead of a quadratic programming problem and is much easier and computationally simpler than SVM. SVM and LS-SVM represent relatively recent artificial intelligence method and have found some applications in image analysis, classification and disease diagnosis etc. [9, 10]. It is worth mentioning that the success of LS-SVM model is highly dependent on the optimum choice of two parameters, the relative weight of regression error γ and the kernel width σ of radial basis function (RBF). Genetic algorithms (GA) [11, 12] introduced by John Holland are probabilistic optimization techniques based on natural evolution and genetics and Darwin's theory of survival of the best. With their efficient and robust global search ability, GA are used to search two optimal parameters for the LS-SVM model simultaneously and automatically. The LS-SVM model then performs the regression task using these optimal parameters.

I. THERY

A. Support Vector Machine Regression

The original theory of SVM introduced by Vapnik was a valuable tool for solving pattern recognition and classification problems [6, 10]. The basic idea of SVM is to map the data set X into a higher dimensional feature space F via non-linear mapping Φ and then perform linear regression in the hyperspace. Vapnik expanded the concept of SVM and developed support vector machine regression (SVMR) by introducing an alternative cost function. In general, SVMR involves a solution of a quadratic programming problem. With the help of the Lagrange multiplier method and a quadratic programming algorithm, the constrained optimization problem is solved. For the details of SVM and SVMR algorithm please refer to the reference of this paper [6, 7, 10].

B. Least squares support vector machines

Considering a regression data set x and a dependent data set y, the LS-SVM method always fits a linear relationship shown in Eq. (1):

$$y = wx + b \tag{1}$$

The optimization problem is to minimize the cost function (J)

$$J_{LS\text{-}SVM} = \frac{1}{2} \| w \|^2 + \gamma \sum_{i=1}^{n} e_i^2 \quad (2)$$

subject to:

$$y_i - wx_i - b = e_i \quad (3)$$

The first part of the cost function is weight decay, which is used to regularize weight sizes and penalize quadratically large weights to make them converge to smaller values in order to avoid deteriorating the generalization ability of the SVM. The second part of the cost function is the regression error (e_i) for all the n training objects. The parameter γ is the regularization parameter, which indicates the relative weight of the error term as compared to the first part, and must be optimized by the user. Analyzing Eq. (7) and its restriction given by Eq. (8), a typical problem of convex optimization is formulated. Thus, the Lagrange function is used to solve this optimization problem.

$$L(w, b, e, \alpha) = \frac{1}{2} \| w \|^2 + \gamma \sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_i (wx_i - b - e_i - y_i) \quad (4)$$

In Eq. (4), the first two parts are the cost functions as defined earlier. The third part is the Lagrange term, which is multiplied by the so-called Lagrange multipliers (α_i). Each Lagrange multiplier corresponds to a certain training point. To obtain the final LS-SVM solution, the partial first derivatives of this Lagrangian function are obtained and are set to zero.

$$\frac{\partial L}{\partial W} = w - \sum_{i=1}^{n} \alpha_i x_i = 0 \quad \therefore w = \sum_{i=1}^{n} \alpha_i x_i \quad (5)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i = 0 \quad (6)$$

$$\frac{\partial L}{\partial e} = \sum_{i=1}^{n} \gamma e - \alpha = 0 \quad \therefore \alpha = \gamma e \quad (7)$$

$$\frac{\partial L}{\partial \alpha} = wx + b + e_i - y_i = 0 \quad (8)$$

The weight coefficients w can be written as a linear combination of the Lagrange multipliers with the corresponding training objects (x_i):

$$w = \sum_{i=1}^{n} = \alpha_i x_i = \sum_{i=1}^{n} \gamma e_i x_i \quad (9)$$

From the equation, the Lagrange multipliers α are calculated. By substituting α and w into the original regression equation Eq.1, the following results are obtained:

$$y = \sum_{i=1}^{n} \alpha_i x_i^T x + b = \sum_{i=1}^{n} \alpha_i \langle x_i, x \rangle + b \quad (10)$$

The inner product $\langle x_i, x \rangle$ can be changed by a kernel function $k = \Phi(x_i)\Phi(x)$ to simplify the use of a mapping operation that is necessary to transform the non-linear input space to a high dimensional feature space where linear regression is possible. Thus, the mapped data are analyzed using conventional linear statistical analysis in the feature space, which is equivalent to nonlinear analysis in the original space. According to Mercer's theorem, the resulting LS-SVM model can be expressed as:

$$y_i = \sum_{i=1}^{n} \alpha_i k(x, x_i) + b \quad (11)$$

The LS-SVM method can perform both linear and nonlinear multivariate calibrations rather rapidly. During the training of the LS-SVM model, it is only necessary to solve a set of linear equation， so the computational complexity is reduced.

C. Genetic algorithms

GA are a group of robust search and optimization techniques, which are based on principles of genetic and natural selection derived from the theories of biological evolution. GA mimic the evolution principle of survival of the best in natural genetic to seek solutions from vast search space at reasonable computation costs. In GA, the collection of variables whose values are to be optimized is termed a chromosome, and the individual variables are called genes. A set of chromosomes is named a population. The initial population of number of possible candidate solutions is generated randomly across the search space. In the population each chromosome has an associated fitness, the chromosomes are evaluated according to the fitness. A new population is formed using three genetic operators: selection, crossover and mutation to generate the offspring of the existing population with the best fitness. The main aim of GA is that the new population will be better than the old one. The optimal solution can be obtained after a series of iterative computations, which is known as generation. The iterative evolution process terminates after a fixed number of generations or when a user-specified stop condition is achieved. Overview of the basic GA can be summarized briefly below:

1. Population initialization

An initial population is generated randomly across the search space. Each chromosome of the population is made by a set of genes and represents a candidate solution to the optimization problem.

2. Evaluating fitness

Once the population is initialized, the fitness value of each chromosome is evaluated by use of fitness function.

3. Genetic operations: selection, crossover and mutation

The GAs use selection, crossover and mutation operators to generate the offspring of the existing population.

4. Stop standard

A fixed number of generations is used as stop criterion or a user-specified -fitness value, where the best fitness value has not been improved.

D. Genetic algorithms least square support vector machines (GA-LS-SVM)

A hybrid method named genetic algorithms least square support vector machines (GA-LS-SVM), which combines

the advantages of genetic algorithms and least square support vector machines, is developed in this case. Two parameters, the relative weight of regression error γ and the kernel width σ of radial basis function (RBF) of LS-SVM model are optimized simultaneously by the real-valued GA (RGA) technique. When using RGA, unlike traditional binary GA (BGA), the two parameters of LS-SVM model are directly coded to form chromosome X, which is symbolized as X= {P1, P_2}, where P_1 and P_2 stand for γ and σ, respectively. The chromosomes are evaluated according to the fitness function. Only the fitness function is problem-dependent and has to be carefully designed by user according to practical condition. In this case, the relative standard errors of prediction (RSEP) were used as the fitness function for evaluating the performances for each chromosome. The RSEP is given by Eq.12.

$$RSEP = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\{C_{ij} - \hat{C}_{ij}\}^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}C_{ij}^2}} \qquad (12)$$

where C_{ij} and $\hat{C}_{ij}$ are the actual and estimated concentrations, respectively, for the i-th component in the j-th mixture, m is the number of mixtures, and n is the number of components. Based on the evaluation of fitness, a new population is formed using selection, crossover and mutation operators. In this case, roulette wheel algorithm is performed to selection operation. Scattered and Gaussian algorithms are used as crossover and mutation operations, respectively. The three evolution operations are iterative evolutionary process until a specified stop criteria is achieved.

A program PGALSSVR was designed to perform the simultaneous spectrofluorimetric multicomponent determination.

II. EXPERIMENTAL AND METHODS

A Shimadzu UV-240 spectrophotometers furnished with OPI-2 function was used for all experiments; a Lenovo Pentium IV microcomputer was used for all the calculations; pH measurements were made with a pH-3B digital pH meter with a glass-saturated calomel dual electrode. Spectra were measured between 330 and 550 nm at 2 nm intervals after 15 min, giving values at 108 wavelengths for each standard solution. An absorption matrix for calibration (D) was developed from these data. According to the same procedures an absorption matrix for prediction (Du) was also developed.

III. RESULTS AND DISCUSSION

A. Genetic algorithm least square support vector machine

In order to obtain the optimal LS-SVM model, two crucial problems are required to be solved: proper kernel function and optimal LS-SVM parameters. Currently, there are no systematic methods for the selection of kernel function. Several kernel functions, such as polynomial function, radial basis function and two-layer perceptions, have been proposed in literature. The radial basis function kernel that is a Guassian curve is commonly used. Only two parameters, the relative weight of regression error γ and the kernel width σ of RBF, need to be selected. It is not known beforehand what values of the two parameters are suitable. However, there is no effective guideline for selecting the optimal values of γ and σ^2, so they must be set only by trial and error, depending on the problem and specific data. These usually used methods for searching the two parameters are time consuming and can not converge at the global optimization. Meanwhile, the two parameters cannot be optimized separately due to the strong interaction they exhibit. Thus, it is difficult for traditional methods to solve such problems. Because of its powerful global searching ability, GA is used to optimize automatically the two parameters in the proposed GA-LS-SVM method. Minimum and maximum values of the parameters are selected by the user; here, we select γ values in the range of 1-15000 and σ^2 in the range of`1-20000. According to the data, the initial population ranges are determined. The success of the GA optimization is affected by the configuration of proposed GA parameters. Since no general strategy exists to select GA parameters, they must be chosen by trial and error. After trial, the configuration of the proposed GA is selected. It is well-known that the quality of LS-SVM regression is dependent upon the parameters γ and σ^2 to be selected correctly. The GA-LS-SVM model then performs the prediction using these selected optimal values. After implementing GA process, the optimal LS-SVM parameters are dynamically optimized. Finally the two parameters (γ = 14474 σ^2 = 3172) were selected by GA as optimal parameters simultaneously.

A training set of 16 samples formed by the mixture of Fe (III), Co (II) and Cu (II) was designed according to four-level orthogonal array design with the L_{16} (4^5) matrix. Spectra were measured between 330 and 550 nm at 2 nm interval, giving values at 108 wavelengths for 16 standard samples. The experimental data obtained from the training set were arranged in matrix D, where each column corresponded to the absorbance of different mixtures at a given wavelength and each row represented the spectrum obtained at a given mixture. A set of 9 synthetic unknown samples was measured in the same way. Using LS-SVM method, the concentrations of Fe (III)，Co (II) and Cu (II) for the test set were calculated. The experimental results showed that the RSEP for total elements were 6.2 %.

B. A comparison of PLS, WT-PLS and GA-LS-SVM

In order to evaluate the GA-LS-SVM method, three methods were tested in this study with a set of synthetic unknown samples. The SEP and RSEP for the three methods are displayed in Table 1. The RSEP for total elements with GA-LS-SVM, WT-PLS and PLS were 6.2%, 7.9% and 10.3%,

respectively. These results indicate that GA-LS-SVM has the best performance among the three methods. The reason may be that the raw data have some nonlinear problem deviating from the Beer-Lambert law, and the LS-SVM method is capable of dealing with both linear and nonlinear problems. Therefore, in this case, the method based on LS-SVM is better than the method based on PLS. The results demonstrated that the GA-LS-SVM method performed well and is a promising technique.

IV. CONCLUSION

The genetic algorithm least square support vector machines (GA-LS-SVM) regression can provide an intelligent predictive model for multicomponent spectrophotometric determination. The GA-LS-SVM method is proven to be successful even when severe spectral overlap was present, and in this case performed better than the WTPLS and PLS methods.

ACKNOWLEDGEMENT

The authors would like to thank the National Natural Science Foundation of China (21067006 and 60762003) for financial support of this project.

REFERENCES

[1] Q. J. Han, H. L.Wu, C. B. Cai, L. Xu, and R. Q. Yu, "An ensemble of monte carlo uninformative variable elimination for wavelength selection," Anal. Chim. Acta Vol. 612, pp. 121-125, 2008.

[2] S. X. Ren, and L. Gao, "Wavelet packet transform and artificial neural network applied to simultaneous kinetic multicomponent determination," Anal. Bioanal. Chem. Vol. 378(5), pp. 1392-1398, 2004.

[3] S. X. Ren, and L. Gao, "Simultaneous quantitative analysis of overlapping spectrophotometric signals using wavelet multiresolution analysis and partial least squares, " Talanta，Vol. 50 (6), pp 1163-1173, 2000.

[4] F. Marini, A. L. Magri, R. Bucci, and A. D. Magri, "Use of different artificial neural networks to resolve binary blends of monocultivar Italian olive oils, " Anal. Chim. Acta, Vol. 599 (2), pp 232-240, 2007.

[5] S. X. Ren, and L. Gao, "Resolve of overlapping voltammetric signals in using a wavelet packet transform based Elman recurrent neural network," J. Electroanal. Chem. Vol. 586 (1), pp. 23-30, 2006.

[6] V. N. Vapnik, "The nature of statistical learning theory," Springer-Verlag, New York, 1995.

[7] V. N. Vapnik, "An overview of statistical learning theory，" IEEE Trans. Neural Networks，Vol. 10，pp. 988-999, 1999.

[8] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, "Least-squares support Machines, " World Scientific, Singapore, 2002.

[9] K. Polat, and S. Gunes, "Brest cancer diagnosis using least squares support vector machine," Digital Signal Processing, Vol. 17, pp. 694-701, 2007.

[10] H. D. Li, Y. Z Liang, and Q. S. Xu, "Support vector machines and its application in chemistry," Chemometr. Intell. Lab. Syst. Vol. 95, pp.188-198, 2009.

[11] O. Devos, and L. Duponchel, "Parallel genetic algorithm co-optimization of spectra pre-processing and wavelength selection for PLS regression," Chemometr. Intell. Lab. Syst., Vol. 107, pp. 50-58, 2011.

[12] D. Ballabio, M. Vasighi, V. Consonni, and M. Komopany-Zareh, "Genetic algorithms for architecture optimization of counter-propagation artificial neural networks," Chemometr. Intell. Lab. Syst., Vol. 105, pp. 56-64, 2011.

TABLE I. SEP AND RSEP VALUES FOR FE(III), CO(II) AND CU(II) SYSTEM BY THE THREE METHODS

Method	SEP (10^{-5} mol l^{-1})				RSEP (%)			
	Fe(III)	Co(II)	Cu(II)	Total elements	Fe(III)	Co(II)	Cu(II)	Total elements
GA-LS-SVM	0.13	0.46	0.08	0.28	3.1	9.1	1.9	6.2
WT-PLS	0.16	0.58	0.22	0.37	3.5	11.4	5.1	7.9
PLS	0.14	0.78	0.25	0.48	3.2	15.4	5.6	10.3