The new Development in Support Vector Machine Algorithm Theory and Its Application

LIU Taian
College of Environment and Spatial Informatics China University of Mining & Technology, Xuzhou, China
Department of Information and Engineering Shandong University of Science and Technology, Taian, China
Email: LTA999@163.COM

WANG Yunjia
College of Environment and Spatial Informatics China University of Mining & Technology, Xuzhou, China

WANG Yinlei
Department of Information and Engineering Shandong University of Science and Technology, Taian, China
Email: Wylei2003@163.COM

LIU Wentong
Nanyang Technological University, Singapore, Singapore

***Abstract*—As to classification problem, this paper puts forward the combinatorial optimization least squares support vector machine algorithm (COLS-SVM). Based on algorithmic analysis of COLS-SVM and improves on it, the improved COLS-SVM can be used on individual credit evaluation. As to regression problem, appropriate kernel function and parameters were selected based on the analysis of support vector regression (SVR) algorithm. This paper proposes the forecasting model of coal mine ground-water-level based on SVR algorithm and improves on it. In another regression problem, it improves on successive overrelaxation for support vector regression (SORR) algorithm to measure the cholesterol content of a blood sample concerning the three kinds of plasma lipoproteins (VLDL, LDL, HDL) in medical science. The numerical experiment results show that the improved COLS-SVM algorithm and Mine Ground-water-level Forecasting improved Model and improved SORR algorithm are effective.**

***Index Terms*—COLS-SVM, SVR, Individual Credit Evaluation, the Forecasting Model of Coal Mine Ground-water-level, Plasma Lipoprotein Cholesterol Measurement**

I. INTRODUCTION

In the algorithm theory and application of support vector machine, the research of classification and regression problems, has important theoretical significance and application value.

As to classification problem, this paper puts forward the COLS-SVM with algorithm of combinatorial optimization for reference. Based on algorithmic analysis of COLS-SVM and improves on it, the improved COLS-SVM is used on individual credit evaluation. Through comparing with Lagrange Support Vector Machine (LSVM) and K-Nearest Neighbor (KNN), the numerical experiment results show that improved COLS-SVM has effective classified forecast ability. For details in the paper part Ⅱ.

As to regression problem, the SVR is applied to forecast coal mine ground-water-level in this paper. Appropriate kernel function and parameters are selected based on the analysis to SVR algorithm. This paper improves on the forecasting model of coal mine ground-water-level based on the SVR algorithm and determines the forecast of the input factor and the output factor according to the physical geography and the hydrology geology situation of the chosen mining area. The numerical experiment results show that the forecast outcomes have compatibility with the actual measurement results. We verify that the improved forecasting model of coal mine ground-water-level has been effective, and provide a new effective method to the forecasting of coal mine ground-water-level. For details in the paper part Ⅲ.

In another regression problem, as to classification problem of successive overrelaxation for support vector (SOR), it improves on successive overrelaxation for support vector regression (SORR) algorithm, to measure the cholesterol content of a blood sample concerning the three kinds of plasma lipoproteins (VLDL, LDL, HDL)in medical science. The numerical experiment has proved that the improved SORR is effective and has better regression precision compares with the standard SVR algorithm and the speed of study is improved. This paper provides a new method for the cholesterol measurements in clinic. For details in the paper part Ⅳ.

II. THE INDIVIDUAL CREDIT EVALUATION BASED ON IMPROVED COLS-SVM

doi:10.4304/jsw.6.2.273-280

A. Individual Credit Evaluation Background

Individual credit system is an important part of state credit systems and individual credit evaluation is a general reflect of individual capital and credit situation, It has great significance and application value to evaluate individual credit using Support Vector Machine (SVM) and a lot of work has been done in this filed [1][2][3]. We put forward the combinatorial optimization least squares support vector machine algorithm (COLS-SVM) basing on the combinatorial optimization algorithm and by basing on algorithmic analysis of COLS-SVM, and improves on it, we use improved COLS-SVM in individual credit evaluation and have made new development in this filed.

B. The COLS-SVM Algorithm

1) The Combinatorial Optimization Algorithm

LS-SVM doesn't have sparse solutions [4] and we divide the training sample set into two subsets to train LS-SVM respectively and the parameter is the same as full training sample set. This kind of training makes the space complication degree much less than that of the full training sample set concerning LS-SVM, so to provide more feasible condition to establish LS-SVM fast algorithm.

To combine the two subsets used to train LS-SVM respectively together lineally, and then we have $f_1(x)$,$f_2(x)$:

$$\begin{aligned} &y = \beta_1 f_1(x) + \beta_2 f_2(x) \\ &among which: \beta_1 + \beta_2 = 1 \end{aligned} \tag{1}$$

To achieve the minimum derivations of full training samples so to satisfy:

$$\begin{aligned} &\min \ e \cdot e' \\ &\text{s.t.} \ \ \beta_1 + \beta_2 = 1 \end{aligned} \tag{2}$$

Among which: e=y-$(\beta_1 f_1(x) + \beta_2 f_2(x))$

Owing to $\beta_1 + \beta_2 = 1$,

$$\begin{aligned} e &= y - (\beta_1 f_1(x) + \beta_2 f_2(x)) \\ &= (\beta_1 + \beta_2) y - (\beta_1 f_1(x) + \beta_2 f_2(x)) \\ &= \beta_1 (y - f_1(x)) + \beta_2 (y - f_2(x)) \end{aligned}$$

If: $e_1 = y - f_1(x)$, $e_2 = y - f_2(x)$,

Then: $e = \beta_1 e_1 + \beta_2 e_2$ (3)

So the above optimization problem changes into:

$$\begin{aligned} &\min \ (\beta_1 e_1 + \beta_2 e_2) \bullet (\beta_1 e_1 + \beta_2 e_2)' \\ &\text{s.t.} \quad \beta_1 + \beta_2 = 1 \end{aligned} \tag{4}$$

Introduce Lagrange η to get the Lagrange function of the above optimization problem:

$$L(\beta) = (\beta_1 e_1 + \beta_2 e_2)^2 + \eta(\beta_1 + \beta_2 - 1) \tag{5}$$

Owing to the above convex optimization problem according to Karush-Kuhn-Tucke, the minimum sufficient and essential condition of $L(\beta)$:

$$\begin{cases} \dfrac{\partial L(\beta)}{\partial \beta} = 0 \\ \dfrac{\partial L(\beta)}{\partial \eta} = 0 \end{cases} \tag{6}$$

So

$$\begin{cases} 2\beta_1 e_1 \cdot e_1' + 2\beta_2 e_1 \cdot e_2' + \eta_1 = 0 \\ 2\beta_2 e_2 \cdot e_2' + 2\beta_1 e_2 \cdot e_1' + \eta_1 = 0 \\ \beta_1 + \beta_2 - 1 = 0 \end{cases} \tag{7}$$

To solve (7) to get:

$$\begin{cases} \beta_1 = \dfrac{e_2 e_2' - e_1 e_2'}{e_1 e_1' + e_2 e_2' - e_1 e_2' - e_1 e_2'} \\ \beta_2 = \dfrac{e_1 e_1' - e_1 e_2'}{e_1 e_1' + e_2 e_2' - e_1 e_2' - e_1 e_2'} \end{cases} \tag{8}$$

In the above solutions, if $e_2 e_2' < e_1 e_2'$ or $e_1 e_1' < e_1 e_2'$, get minus solution. The combined SVM from different random training samples from the same sample space tends to have a positive solution rather than a minus one, thus needs to have a limiting condition: $\beta_1,\ \beta_2 \geq 0$, under $\beta_1,\ \beta_2 \geq 0$, the solution of optimization problem may get directly based on formula (8)

That is:

$$\beta_1 = \begin{cases} 0 & \beta_1 \leq 0 \\ \beta_1 & 0 < \beta_1 < 1 \\ 1 & \beta_1 \geq 1 \end{cases} \tag{9}$$

$$\beta_2 = 1 - \beta_1 \tag{10}$$

To divide the training sample set into two subsets to train LS-SVM respectively to bind them together lineally so to get the minimum derivation. To promote this idea: the training sample is m subsets, and each subset to training its LS-SVM, and to put them together lineally and to do sparse trimming, and finally get the whole training sample set LS-SVM.

2) The Algorithm Analysis and Improvement

According to the combinatorial optimization idea, we combine the *m* LS-SVMs of the training sample sets lineally one by one, and then do the sparse trimming after the combination. We can get the COLS-SVM by concluding the above process [3].

By extracting the training sample subsets, we train LS-SVM, and combinatorial optimization, sparseness, repeated cycles calculating to COLS-SVM algorithm, finally got the training samples of the LS–SVM. Each cycle time complexity is $O(m(\eta+2)(k+p)^3)$. Because $m = l/p$, $\eta = (p - 2\Delta)/\Delta$, each cycle time complexity rewrite for $O(l/\Delta)(k+p)^3)$. In practical application, Δ, l and k are constant, time complexity is increasing function for p. Based on the above improvement ideas, in the first round of circulation,

ordinal linear combination *m* disciplinal LS-SVM of the training sample subset, and at the same time sparse, then we combined the result of before the round cycle with LS-SVM of *l/p* of after a cycle training sample subset by linear, and at the same time sparse *p* training sample, just so cycle until the stability for SV_S vector, termination cycle. Summarize the process, we get improved accurate COLS-SVM algorithm.

Algorithm: Improved accurate COLS-SVM.

Step1: Given training sample set $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in(X,Y)$, among which $x_i\in X=R^n$, $y_i\in Y\in R, i=1,...,l$, to do unification with *T* using the formula (8), and arrange the training samples randomly in *T*;

Step2: empty *Ts1*, *Ts2*, *SVs*, and null $\alpha_{T1},\alpha_{T2},\alpha_{SV},Q_{optm},test_{SV}$ choose the proper kernel function $K(x_i, x_j)$ and parameter *C*, and initialize the finally reserved *k SV* and the training sample *m* among the training sample subsets;

Step3: to dram *k* and *p* training samples in order, and put into *Ts1*, *Ts2* respectively;

Step4: to use LELS-SVM to train $f_1(x)$, $f_2(x)$ of *Ts1*, *Ts2* of the training sample subsets of LS-SVM;

Step5: To use the formula (8--10) to calculate β_1,β_2 for the full training sample set;

Step6: To use $\beta_1\times\alpha_{T1}$ and $\beta_2\times\alpha_{T2}$ to replace α_{T1} and to put the training sample from *Ts2* into *Ts1* as supplement;

Step7: In *Ts1*, delete the *p* training samples with the least absolute value in α_{T1}, To train again LS-SVM:$f_1(x)$ of *Ts1* of the training samples subsets by using LELS-SVM algorithm;

Step8: To calculate sum of deviation square *Q* of $f_1(x)$ in training samples;

Step9: If $Q<Q_{optm}$, put training samples in *Ts1* into *SVs*, α_{T1} into α_{SV}, *Q* into Q_{optm}, $Test_{SV}$=1;

Step10: Judge whether the needle of training sample in *T* directs to the end or not, if yes, turn it to Step12;

Step11: To draw *p* training samples from *T* and put them into *TS2*, and use LELS-SVM algorithm to train LS-SVM: $f_2(x)$ of *TS2* of training sample subsets, then turn to step5;

Step12: Judge $Test_{SV}$, if $Test_{SV}$=1, arrange randomly the training samples in *T* and turn the needle to the first training sample, null $Test_{SV}$ and turn to Step11;

Step13: Training concluded.

C. Individual Credit Evaluation Applications Based on Improved COLS-SVM

Use improved COLS-SVM to classify individual credit evaluation. With the individual credit characteristic data in document [2] for reference, to make numerical experiments for construction bank, commercial bank and general individual credit characteristic data we designed. Experiment environment: Pentium IV3.0G CPU, 1024MB memory, Windows XP operating system, Matlab6.5.

Using improved COLS-SVM in designed 2), to compare with LSVM [5], KNN classification method to conduct data comparison experiment to get the following superficies: By experimenting with the marking data of the individual credit characteristic data of construction bank and commercial bank, to verify the effectiveness of COLS-SVM. To draw 3 groups of the individual credit characteristic data from the two banks to do training experiment to get *SV* and related parameters, then to draw another 3 groups to make testing experiment, among which LSVM algorithm degree of a polynomial *d*=3, KNN classification method *k*=15 The experiment as following: (Improved COLS-SVM is marked as COLS-SVM*)

TABLE I. THE MARKING DATA CLASSIFICATION ACCURACY RATE OF A CONSTRUCTION BANK

Samples	COLS-SVM*	LSVM (d=3)	KNN (K=15)
300	0.851	0.767	0.656
500	0.843	0.753	0.637
1000	0.829	0.746	0.609
Average	0.841	0.755	0.634

TABLE II. THE MARKING DATA CLASSIFICATION ACCURACY RATE OF A COMMERCIAL BANK

Sample	COLS-SVM*	LSVM (d=3)	KNN (K=15)
300	0.873	0.798	0.697
500	0.862	0.776	0.673
1000	0.859	0.732	0.645
Average	0.865	0.769	0.672

From the two tables, it is easy to see that COLS-SVM* has better classification ability than LSVM-polynomial kernel function and KNN classification. The average experiment accuracy rates are respectively 84.1% and 86.5% and has the characteristic of less number of dimensions, higher accuracy rate. (Number of dimensions 19 for construction bank, 16 for commercial bank)

Using general characteristic data in document [2] to continue the numerical experiment. To draw 3 groups respectively from the training sample and testing sample to conduct data experiment to get *SV* and related parameters, among which among which LSVM algorithm degree of a polynomial *d*=3, KNN classification method *k*=15 The experiment as in Table Ⅲ:

TABLE III. THE GENERAL DATA CLASSIFICATION ACCURACY RATE

Samples	COLS-SVM*	LSVM (d=3)	KNN (K=15)
300	0.926	0.857	0.718
500	0.915	0.849	0.709
1000	0.911	0.852	0.673
Average	0.917	0.853	0.700

According to the table Ⅲ, used in general characteristic data we designed, COLS-SVM* proves to have better classification ability with an average accuracy rate of 91.7%, comparing with LSVM algorithm degree of a polynomial and KNN classification method. Owing to less number of dimensions (10), the accuracy rate is higher.

D. Summary

Owing to Combinatorial Optimization Algorithm idea, put forward COLS-SVM and improve on it, and apply it to compare with LSVM polynomial kernel function algorithm and KNN classification method and to do the numerical experiment, thus get the classification result in individual credit grade, the numerical experiment shows: COLS-SVM* has better classification forecasting ability and has made new development in individual credit evolution.

III. The Improved Forecast Model Coal Mine Ground-water-level Based on SVR

A. The Coal Mine Ground-water-level forecasting Background

In the energy industry, coal exploitation is a high risk industry, which often has some mine flooding accidents and brings heavy loss to the life and property security of our country and people. So the research on the forecast of the coal mine underground water level has important theoretical and practical significance, which is an issue with many influencing factors, highly non-linear and temporal [6][7][8] series.

Support Vector Regression Algorithm (SVRA) is a method to regression prediction and function approximation. SVRA has not only strict theory base [9][10][11], but also can well resolve such practical problem as non-linearity, high dimension and local minima. So, it is a good method to forecast Coal Mine Ground-water-level with SVRA.

B. The theory of prediction with SVRA and improvement

SVM is originally proposed for classification. It is a new Machine Learning method based on Statistical Learning Theory. Because of the use of kernel function and Structural Risk Minimization, SVM has good performance to classification with limited samples. SVM is formulated as a convex quadratic programming, so the minima which are found are global optimal solutions. SVRA is a method to regression prediction and function approximation.

SVM is good at solving regression prediction problem by utilizing an appropriate loss function [12]. The ε-insensitive loss function is used usually. When $|y-f(x)|\le\varepsilon$, there is no loss. The ε-insensitive loss function is that

$$L_\varepsilon(y)=\max\{0,|y-f(x)|-\varepsilon\} \tag{11}$$

We consider a given train data set $D=\{(x_1,y_1)...,\ (x_l,y_l)\}, x\in R^n, y\in R$ .For linear regression problem, SVR would like to find regression prediction function:

$$f(x)=(\omega\cdot x)+b \tag{12}$$

Where, $(\omega\cdot x)$ is inner product, $\omega\in R^n, b\in R$.

By utilizing largest margin principle, ε-insensitive loss function and Lagrange function, we get the optimization problem [11]:

$$\min \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(x_i\cdot x_j)-$$

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)y_i+\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)\varepsilon \tag{13}$$

$$s.t.\ 0\le\alpha_i,\alpha_i^*\le C\ ,\ i=1,2,...,l \tag{14}$$

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \tag{15}$$

Where, C is penalty parameter, α,α^* are Lagrange multipliers.

By solving convex quadratic programming (13)-(15), we get the optimal solution

$$\bar{\alpha}=(\bar{\alpha}_1,\bar{\alpha}_1^*,\bar{\alpha}_2,\bar{\alpha}_2^*,...,\bar{\alpha}_l,\bar{\alpha}_l^*)^T\ ,$$

So, we get the regression prediction function

$$f(x)=\sum_{i=1}^{l}(\bar{\alpha}_i^*-\bar{\alpha}_i)(x_i\cdot x)+b \tag{16}$$

For nonlinear regression problem, by utilizing an appropriate kernel function, we can get nonlinear regression function. The samples can be project to high dimension space by inner product where the linear approximation can be done. Kernel function

$$K(x_i,x_j)=\varphi(x_i)\varphi(x_j) \tag{17}$$

So, we get the regression prediction function for nonlinear regression problem

$$f(x)=\sum_{i=1}^{l}(\bar{\alpha}_i^*-\bar{\alpha}_i)K(x_i,x_j)+b \tag{18}$$

The advantages of utilizing ε-insensitive loss function in support vector regression machine is that only the Lagrange multipliers $\alpha_i-\alpha_i^*$ of the samples outside the Strip region are not to equal zero. The vectors are named as support vector that their Lagrange multipliers are not equal to zero.

By adding b to problem (18), we get the optimization problem [13][14][15]:

$$\begin{aligned}&\min_{w,b,\xi^{(*)}}\ \frac{1}{2}\|(w,b)\|^2+\frac{C}{l}\sum_{i=1}^{l}(\xi_i+\xi_i^*)\\ &s.t.\ \ ((w,b)(x_i,1)-y_i\le\varepsilon+\xi_i,\ i=1,2,\cdots,l\\ &\quad y_i-((w,b)\cdot(x_i,1)\le\varepsilon+\xi_i^*,\ i=1,2,\cdots,l\\ &\quad \xi_i^{(*)}\ge 0,\ i=1,2,\cdots,l\end{aligned} \tag{19}$$

Let $z_i=(x_i,1)^T\in R^{n+1}$, $h=(w,b)^T\in R^{n+1}$. Problem (19) turns into that

$$\begin{aligned} &\min_{w,b,\xi^{(*)}} \quad \frac{1}{2}\|h\|^2 + \frac{C}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ &s.t. \quad (h\cdot z_i) - y_i \le \varepsilon + \xi_i,\ i = 1,2,\cdots,l \\ &y_i - (h\cdot z_i) \le \varepsilon + \xi_i^*,\ i = 1,2,\cdots,l \\ &\xi_i^{(*)} \ge 0,\ i = 1,2,\cdots,l \end{aligned} \tag{20}$$

The dual problem of problem (19) is as follows:

$$\begin{aligned} \min\ & \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(z_i \cdot z_j) \\ & + \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \\ s.t.\ & 0 \le \alpha_i, \alpha_i^* \le C\ , i = 1,2,...,l \end{aligned} \tag{21}$$

C. Establishing improved forecast model of Coal Mine Ground-water-level

1) Determining forecast factor and output factor

Dynamic underground water level is a direct reflection of equilibrium status of the regional underground water level, which is an integrated externalization of the aquifer structure, properties, circulating conditions and the interaction between the supplies and the drainage factors. The actual measurement data of a mine area of Shandong are taken to do the forecast model of underground water level experiment. Based on the relevant data analysis, the main factors that affect the underground water level of digging are amount of precipitation, evaporation capacity, exploitation and riverway flux. Besides, with a comprehensive consideration of the hysteresis of all the factors and the self-correlation of the underground water level and regarding the amount of precipitation and measured underground water level in the former period as the impact factors, six impact factors are obtained as hefts of the input vectors supporting vector regression machine: amount of precipitation X_1, evaporating capacity X_2, exploitation X_3, river flux X_4, precipitation in the former period X_5 and measured underground water level in the former period X_6. Forecasting underground water level Y is taken as the output value of the model. So the forecasting issue of underground water level in the digging can be transformed to a support vector regression forecasting one with six characteristic variables and one output.

2) The improved forecasting model of Coal Mine Ground-water-level based on SVR

We are given the actual measurements of ground-water-level $\{(x_1, y_1),...,(x_l, y_l)\}$, where pattern $x_i \in R^6$ consists of the factors which effect the forecast of ground-water-level and $y_i \in R$ is the value relating to ground-water-level, $i = 1,2,...,l$. We can train the given sets $\{(x_1, y_1),...,(x_l, y_l)\}$ by SVR. Then we can get the improved forecast model of ground-water-level:

$$y = \sum_{i=1}^{l}(\bar{\alpha}_i^* - \bar{\alpha}_i)K(x_i, x_j) + b \tag{22}$$

D. Numerical experiments

We proceed the numerical experiments under the conditions that the CPU is Pentium IV3.0G, Memory is 1024MB, and operating system is Windows XP. The testing program of improved forecast model of ground-water-level utilizes Matlab6.5. The numerical experiments are based on the history data of a mine area of Shandong.

All data is distributed into train set and test set firstly. We get the regression prediction function based on train set. The accuracy of regression prediction function is tested based on test set.

The data of numerical experiments consist of observation value of long-time observation hole *O6-2*. History data 1 to 30 are used as train samples. History data 31 to 66 are used as test samples. The comparative results of the forecast values of ground-water-level of hole *O6-2* with factual values are showed in Table Ⅳ.

TABLE IV. COMPARATIVE RESULT OF THE FORECAST VALUES OF GROUND-WATER-LEVEL OF HOLE *O6-2* WITH FACTUAL VALUES

Time	Factual values(m)	Forecast values(m)	Relative error (%)
2009.1	12.57	12.48	-0.72
2009.2	13.72	13.79	+0.51
2009.3	13.85	13.90	+0.36
2009.4	14.26	14.37	+0.77
2009.5	14.71	14.82	+0.75
2009.6	15.69	15.77	+0.51
2009.7	16.35	16.43	+0.49
……	……	……	……

From Table Ⅳ, we conclude that forecast values are consistent well with factual values. Average relative error is 0.38%, which verifies that the improved forecast model (22) is effective.

Table Ⅳ only includes forecast values of one long-time observation hole. When we forecast the observation values of other long-time observation hole, the input factor and the long-time observation data of observation hole which is affected geographically should be considered.

The comparative result of the forecast values of ground-water-level of 6 holes with factual values are showed in Table Ⅴ.

TABLE V. COMPARATIVE RESULT OF THE FORECAST VALUES OF GROUND-WATER-LEVEL OF 6 HOLES WITH FACTUAL VALUES

Hole number	Factual values(m)	Forecast values(m)	Relative error (%)
O1-1	16.38	16.45	+0.43
O2-3	16.59	16.63	+0.24
O3-2	16.73	16.87	+0.84
O4-3	16.69	16.72	+0.18
O5-2	16.85	16.81	-0.24
O6-2	16.72	16.79	+0.42

From Table Ⅴ, we conclude that the improved forecast model (22) is effective for different observation hole. The forecast values are consistent well with factual values. Average relative error is 0.31%.

The real-time forecast values of observation hole *O3-1* are listed in Table Ⅵ.

TABLE VI. COMPARATIVE RESULT OF THE REAL-TIME FORECAST VALUES OF GROUND-WATER-LEVEL OF HOLE *O3-1* WITH FACTUAL VALUES

Time	Factual values(m)	Forecast values(m)	Relative error (%)
2009.10	13.79	13.82	+0.22
2009.11	13.26	13.34	+0.60
2009.12	12.84	12.76	-0.62
2010.01	——	12.35	——
2010.02	——	12.87	——
……	……	……	……

From Table Ⅵ, we conclude that the improved forecast model (22) is effective.

E. Summary

There are two innovation points in this paper. The Forecasting Model of Coal Mine Ground-water-level is proposed based on SVR algorithm, and improves on it. Choosing the appropriate kernel function and the parameters of Forecasting Model in this paper based on cross validation methods via parallel grid search in SVR algorithm.

By factual testing and analysis to the Ground-water-level of a digging in one mine area of Shandong, we conclude that:

① Cross validation methods via parallel grid search is used in choosing the parameters of regression model, which avoids aimlessness and randomness of choice, and raises the forecast accuracy.

②The numerical experiments show that the improved Forecasting Model of Coal Mine Ground-water-level proposed basing on SVR is effective. Experiments and forecast are stable. We propose and improve on a new effective method to the Forecasting Coal Mine Ground-water-level in this paper.

Ⅳ. THE CHOLESTEROL MEASUREMENT BASED ON IMPROVED SORR

A. The Cholesterol Measurement Background

The measurement of cholesterol in human's blood is of great significance to the clinical diagnosis of cerebrovascular disease. If the content of cholesterol is over low or over high, there may be a disorder in cholesterol metabolism; therefore a method which is more accurate and effective for the content determination of cholesterol is needed in clinical assay. This work is our continuative research and has new advances.

According to the algorithm SOR put forward by Mangasarian for classification problem [16], we can do a generalization in the regression problem, and improve the standard algorithm SVR (support vector regression), then we get support vector regression algorithm SORR, and improve on it, and apply it to the medical field -- cholesterol content determination in three kinds of plasma lipoprotein (VLDL、LDL、HDL), providing a new method for the clinical content determination of cholesterol.

B. The Improved SORR Algorithm

To apply the algorithm SOR put forward by Mangasarian for classification problem into the regression problem and get support vector regression algorithm SORR, and improve on it. The detail method is shown below:

If we get the nonlinear regression function is:

$$f(x) = w' \cdot \phi(x) + \mathrm{b} \tag{23}$$

In the nonlinear instances, the optimization problem:

$$\min_{w} \ \frac{1}{2} w' w + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \tag{24}$$

$$\text{s.t.} \quad \begin{aligned} & y_i - w \cdot \phi(x_i) - b \le \varepsilon + \xi_i \\ & w \cdot \phi(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ & \xi_i \ge 0, \xi_i^* \ge 0, \quad i = 1,2,...,l \end{aligned} \tag{25}$$

Can be transformed into the optimization problem:

$$\min_{w} \ \frac{1}{2} w' w + \frac{1}{2} b^2 + \frac{C}{2} \sum_{i=1}^{l} (\xi_i^2 + \xi_i^{*2}) \tag{26}$$

$$\text{s.t.} \quad \begin{aligned} & y_i - w \phi(x_i) - b \le \varepsilon + \xi_i \\ & w \phi(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ & i = 1,2,...,l \end{aligned} \tag{27}$$

Using Lagrange function shown below to solve them (26--27), specific reference [17].

By learning, we get the regression function as shown:

$$f(x) = \sum_{x_i \in SV} (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{28}$$

Observing the analysis before, we find that the support vector regression algorithm we put forward requires one less constraint condition than standard SVR, and there is not an upper bound for the parameter in optimization problem, thereby the computational complexity is reduced and the speed of study is improved. Later we will prove it by numerical experiments, and apply it to the content determination of cholesterol in plasma lipoproteins.

C. The Numerical Experiments of Cholesterol Measurement

The data source is data file Choles_all.mat in Matlab, the data include 264 patients' blood samples and the corresponding cholesterol content in the three plasma lipoproteins--HDL, LDL and VLDL, every blood sample is made up of spectral data which contains 21 kinds wavelength in the electrophoresis belt. Using blood samples, we can construct a support vector regression to regression estimate the content of cholesterol in every plasma lipoprotein.

Testing environment: Pentium IV processor, 3.0G CPU, 1024MB RAM, Windows XP operation system. Using the Matlab6.5 programming to prove the improved SORR mentioned before.

In experiment we use regression to estimate the content of cholesterol in each plasma lipoprotein corresponding to each blood sample, use the content and actual content to obtain the related coefficient *R*, *R* can reflect the property of regression estimation [18].

$$R = \frac{Cov[f(x), y]}{\| f(x) \| \cdot \| y \|} \tag{29}$$

Among $|R| \leqslant 1$, $f(x)$ is the regression estimated blood data corresponding to each plasma lipoprotein (264×1 column vector), y is the corresponding actual content (264×1 column vector), the closer R is to 1, the more accurate the regression estimation is.

1) The experiment of study speed

To train with the whole parameters, and take ε=0.1 in both algorithms SVR and improved SORR, using the Gaussian kernel function: $K(x, x_i) = \exp(-\| x - x_i \|^2 / \sigma^2)$. Take parameter σ^2=1, when changing between C=10^0, 10^1, 10^5, 10^{10},…,10^{50}, Using both algorithms SVR and improved SORR to regression estimate the cholesterol content in plasma lipoprotein LDL, the study speed is shown in the diagram Ⅶ as below: (Improved SORR is marked as SORR*)

TABLE VII. COMPARISON OF REGRESSION STUDY TIME OF THE LDL CHOLESTEROL CONTENT IN TWO METHODS ('S)

C	1	10	10^5	10^{10}	10^{20}	10^{30}	10^{40}	10^{50}
SVR	1.96	2.01	2.35	2.46	5.27	5.48	5.51	5.62
SORR*	1.85	1.97	2.03	2.12	2.65	2.87	2.95	3.07

As we can see from Table Ⅶ, when C=10^0~10^{10}, two algorithms average learning time is little difference, but with the increase of C(C=10^{20}~10^{50}), an average of study time is 5.6 s about SVR algorithms, but improved SORR algorithms is 3 s, improved SORR than SVR fast learning about 46%. This also has concerned with us machine performance. For the same HDL, VLDL cholesterol content regression estimation, basically the same conclusion.

2) The regression precision experiment

Draw randomly 1/3 of the data samples mentioned before to exercise, and use the whole sample to test, and take ε=0.1 in both SVR and SORR*, using the Gaussian kernel function, take parameter σ^2=1. When apply different value to C, the regression estimation results by two algorithms are compared in diagram Ⅷ as shown.

TABLE VIII. COMPARISON OF THE PRECISION OF REGRESSION ESTIMATE (VALUE **R**) ON DIFFERENT ***C*** NUMERICAL VALUES

	HDL		LDL		VLDL	
	SVR	SORR*	SVR	SORR*	SVR	SORR*
C=10^0	0.735	0.901	0.791	0.916	0.612	0.811
C=10^1	0.810	0.912	0.802	0.923	0.601	0.826
C=10^2	0.901	0.928	0.869	0.925	0.632	0.833
C=10^8	0.841	0.932	0.672	0.927	0.567	0.843
C=10^9	0.831	0.943	0.634	0.933	0.581	0.836
C=10^{10}	0.832	0.941	0.635	0.945	0.579	0.841

As we can see from Table Ⅷ, different values for C, SORR* compare with SVR, SORR*'s Learning precision is improved. Equally, when C is given a constant value, we give parameter σ^2 different values for experimental comparisons, and get the same conclusion. To a breakthrough in the algorithm theory, we must make innovations.

D. Summary

Using the improved SORR to regression estimate the content of cholesterol in three different plasma lipoproteins, we compare it with SOR in study speed and regression precision, the experiment data proves: improved SORR is more accurate, comparing with the standard support vector regression algorithm SVR, it guarantees the regression precision and improves the study speed, provides a new method for clinical determination of cholesterol. As we can see, to apply this method to real clinical diagnosis, we need to combine it with other methods and do further research.

Ⅳ. CONCLUSION

This paper is about the support vector machine (SVM) classification problem and regression problems algorithm and its application research. TO the Financial Calculation and the Forecasting Coal Mine Ground-water-level and the Plasma Lipoprotein Cholesterol Measurement, the COLS-SVM algorithm and the Model and SORR algorithm are put forward, and improve on them. The numerical experiment results show that the improved COLS-SVM and mine ground-water-level improved Forecasting Model and improved SORR algorithm are effective.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (50534050, 50774080/E040101, 50811120111, 40971275), and Program of Shandong Taian Science and Technology (20082025).

REFERENCES

[1] C. H. Shen, "Personal Credit Evaluation in Consumer Lending Based on Support Vector Machine," China Agricultural University, Beijing City, 2005, pp. 68-133.

[2] T. A. Liu, "SVM and Application Study in Personal Credit Scoring," Shandong University of Science and Technology, Taian City, 2005, pp. 43-53.

[3] T. A. Liu, Y. J. Wang, W. T. Liu, "The Individual Credit Evaluation Based on COLS-SVM," In Proc. 2009 International Forum on Computer Science-Technology and Applications (IFCSTA 2009), IEEE Computer Society CPS Press, Dec. 2009, pp. 455-457.

[4] L. Z. Gan, Z. H. Sun, and Y. X. Sun, "Sparse least squares support vector machine," Journal of Zhejiang University (Engineering Science), 2007, 41(2) pp. 245-248.

[5] O. L. Mangasarian, David R. Musicant, "Lagrangian Support Vector Machines," Journal of Machine Learning Research, 2001, (1), pp. 161-177.

[6] X. H. Wu, H. L. Shi, and Z. L. Li, "The Forecasting Method of Coal Mine Ground-water-level Based on BP Neural Network," Industry and Mine Automation, Changzhou City, 2006,10, pp. 21-23.

[7] B. R. Chen, "Groundwater Level trends and Prediction," Science Publication, Beijing City, 1998.

[8] J. Q. Yang, X. X. Luo, "A RBF Artificial Neural Network Model for the Prediction of Regional Groundwater Level," Journal of China Hydrology, Beijing City, 2001,8, pp. 1-3.

[9] X. F. Song, D. J. Chen, and H. J. Yu, "A Survey of Optimization Algorithms in Support Vector Machine," Computer Science, Chongqing City, 2003, Vol.30, No.3, pp. 12-20.

[10] S. W. Ji, W. T. Ji, "A Tutorial Survey of Support Vector Machine Training Algorithms," Computer Technology and Development, Xian City,2004,1, pp.18-20.

[11] N. Y. Deng and Y. J. Tian , New Method in Data Mining – Support Vector Machines, Science Publication, Beijing City, 2004.

[12] V. Vapnik, S. Golowich and A. Smola, "Support vector method for function approximation, regression estimation and signal processing," In M. Jordan and T. Petsche, editors, Advances in Neural Information Processing System 9, MIT Press, Cambridge, MA, 1997, pp. 281-287.

[13] Gunnar Ratsch, Member, IEEE, Sebastian Mika, Bernhard Scholkopf, and Klaus-Robert Muller, "Constructing Boosting Algorithms form SVMs: An Application to One-Class Classification," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(9), pp. 1-16

[14] B. Scholkopf, J. Platt and J. Shawe-Taylor, "Estimating the Supper of a High-Dimensional Distribution," Journal of Neural Computation, 2001, 13(7), pp. 1443-1471.

[15] V. Franc, V. Hlavc, "Kernel repression of the Kessler construction for Multi-class SVM classification," In:H.Wildenauer and W. Kropatsch, editor, Proceedings of the CWWW'02, page 7, Wien, Austria, February 2002. PRIP. Available at ftp://cmp.felk.cvut.cz/pub/cmp/articles/franc/ francmulitiKernel102.pdf.

[16] O. L. Mangasarian, David R. Musicant, "Successive Overrelaxation for Support Vector Machines," Journal of IEEE Transactions on Neural Networks, 1999, 10(5),pp. 1032-1037.

[17] T. A. Liu and B. C. Yang, "Application of Successive Overrelaxation for Support Vector Regression Arithmetic in Cholesterol Measurements," Journal of Computers and Applied Chemistry, Beijing City, 2007, 24(9), pp. 1163-1165.

[18] L. Ding, L. Tao, "New Learning Algorithm of the Improved SVM for Regression," Journal of Computer Engineering and Applications, Beijing City, 2005, 19, pp. 44-46.

LIU Tai-an was born in Taian city of Shandong province, China in 1963. He is a Ph. D. Candidate, in China University of Mining and Technology, Xuzhou China, major study in data mining and software engineering.

He is an associate professor in department of information and engineering of Shandong University of Science and Technology. His current research interests include data mining and software engineering, published articles: ① "New multiclassification algorithm based on fuzzy support vectormachines", Chengdu, China, Journal of Application Research of Computers, Jul. 2008 Vol. 25 No. 7. ② "Research on Minimal Kernel Classifiers of Feature Selection", Beijing, China, Journal of Computer Engineering and Applications, Jun. 2007 Vol. 43 No. 16. (The INSPEC database indexed).

He is a senior member of China Computer Federation, is an excellent reviewer of "Chinese Sciencepaper Online".

WANG Yun-jia was born in Jianhu city of Jiangsu province, China in 1960. He is a professor and Ph. D. supervisor, in China University of Mining and Technology, Xuzhou China, is a director of Institute of Natural Resource Information and Economy, is a dean of School of Environmental Science and Spatial Informatics. His research interests include data mining and digital mines. He is a committee member of international society of mine surveying, editorial committee member of Acta Geodaetica et Cartographica Sinica (AGCS).

WANG Yin-lei was born in Xintai city of Shandong province, China in 1972. He is a master's degree, major study in teaching methodology of physics subject.

He is an assistant in department of basic courses of Shandong University of Science and Technology. His current research interests include physics teaching and software engineering, published articles:① "The solution of the harmonic oscillator with electric charge at the electric field in the coordinate basis", Yibin, China, Journal of Yibin University, Dec. 2008 Vol. 12 No. 12. ② "The universe conception of Einstein", Leshan, China, Journal of Leshan Normal College, May. 2009 Vol. 24 No. 5.

LIU Wen-tong is a student of Nanyang Technological University in Singapore.