Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (16): 3263-3268, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.3263.3268

An Integrated Data Mining Method and its Apllication

S. Chuanhe, L. Zhongwen and L. Ying
Institute of Financial Engineering of Shandong Women's University, 250300, Jinan, China

Abstract: In order to address nonlinearity and instability in financial time series, this study is aimed at structuring an integrated data mining method based on Support Vector Machines (SVM) and Wavelet Neural Networks (WNN) and at exploring its application in analyzing financial market interactions. In the proposed methodology, a kind of WNN is employed to select input features for the SVM using variance rating analysis, with the SVM also improved in feature weighting through innovated discounted least square. This model proposed is capable of mining more information contents implied in samples and also enhancing generation ability of the model by taking use of the advantages of WNN and SVM. Empirical results show that the proposed hybrid approach can capture unique interaction mechanisms between financial markets in China more efficiently than other analysis methods.

Key words: Nonlinearity and instability, the integration of SVMs to WNNs, feature selection, discounted least square, market interaction analysis

INTRODUCTION

For financial risk management and investment/portfolio diversification, it is important to understand the mechanism of how risk spillover occurs across different markets. In order to capture and mine the mechanism, many methods including econometrics and statistics methodologies have already been employed, such as GARCH (Generalized Autoregressive Conditional Heteroskedasticity), cointegration, artificial intelligence methods and their combinations, most of which can deal with nonlinearity and instability in financial issues to some extent.

This study will only employ artificial intelligence methods to analyze the interaction mechanism between financial markets and focus on Support Vector Machines (SVM) and Wavelet Neural Networks (WNN). As a kind of effective analysis tools, SVM and WNN have played an important role in addressing nonlinearity and instability in financial issues (Kim, 2003; Becerra *et al*., 2005; Amornwattana *et al*., 2007; Tsai and Wu, 2008; Boyacioglu *et al*., 2009). And the existing combinations of SVMs and WNNs have also proved to be more efficient in model performance than other methods (Whitman *et al*., 2001; Chandra *et al*., 2010) but the SVMs and WNNs in these combinations were just integrated in their original patterns respectively which will certainly weaken the prediction accuracy of the combinations.

Consequently, this study will attempt to introduce a novel hybrid approach based on innovated SVM and WNN and subsequently explore its application in measuring financial market interactions. And the remainder of this study is organized as follow. In section 2, the methodologies about support vector machines, wavelet neural networks and their integration are described, respectively. Section 3 presents data analysis and out-of sample results of the interaction between financial markets in China, with the conclusion in the final section.

AN INTEGRATED DATA MINING METHOD

An innovated sample-weighting SVM: The support vector machine is a new generation learning system for small samples and is constructed according to the statistical learning theory by Vapnik (1995) which employs the structural risk minimization principle other than the empirical risk minimization principle generally used by the neural network methodology. It therefore overcomes the shortcomings of the neural network recognition algorithm, such as large samples, "the curse of dimensionality", local optimization, over-fitness, etc. and has a better model generalization performance. It is now being established as one of the standard tools for machine learning and data mining, such as pattern recognition, regression analysis and probability density estimation (Taylor and Cristianini, 2000). Here, the support vector regression machine (ε-SVR hereafter), as one kind of support vector machines, is introduced below (Vapnik *et al*., 1997; Drucker *et al*., 1997).

The ε-SVR is involved in the approximation of function, that is, the selection of a special function in an image space through machine learning, with the function expressed as follows:

Corresponding Author: S. Chuanhe, Institute of Financial Engineering of Shandong Women's University, 250300, Jinan, China

Given a training set, $\{X_i, y_i\}$, i = 1, 2,…, l, where the input data X is assumed to be a compact domain in a Euclidean space R^n and the output data y is assumed to be a closed subset of R. Learning from data can be viewed as an approximation of the multivariate function f(X) which represents the relation between the input X and the output y. By some nonlinear mapping $\phi(X)$, the input X is mapped onto a hypothesis space (or feature space) in which the learning machine (algorithm) selects a certain function f(X).

According to the learning theory, for constructing a nonlinear support vector machine, the decision function takes the following form:

$$f(X) = W \bullet \phi(X) + b \tag{1}$$

Considering the ε-insensitive cost function:

$$M(y, f(X)) = L(|y\text{-}f(X)|_\varepsilon)$$

Where:

$$|y - f(X)|_\varepsilon = \begin{cases} 0, & \text{if } |y - f(X)| \le \varepsilon \\ |y - f(X)| - \varepsilon, & \text{others} \end{cases}$$

and ε represents the insensitivity range of the ε-SVR.

This means that the cost is equal to 0 if the deviation of the expected value from the observed value is smaller than the ε.

Then, solving regression problem is equivalent to optimizing the following problem:

$$\begin{aligned} &\min \ \frac{1}{2}\|W\|^2 + C\Big(\sum_{i=1}^{l}\xi_i^* + \sum_{i=1}^{l}\xi_i\Big) \\ &\text{s.t.} \ \ y_i - (W \cdot \varphi(X_i)) - b \le \varepsilon + \xi_i^*, i = 1,2,\cdots,l \\ &(W \cdot \varphi(X_i)) + b - y_i \le \varepsilon + \xi_i, i = 1,2,\cdots,l \\ &\xi_i^* \ge 0, i = 1,2,\cdots,l \\ &\xi_i \ge 0, i = 1,2,\cdots,l \end{aligned} \tag{2}$$

where, C is the regularization constant which plays a trading-off between the regularization performance and the empirical error and ξ_i, ξ^*_i represent slack variables.

The corresponding dual problem can be expressed by:

$$\begin{aligned} &\min \ \frac{1}{2}\sum_{i=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(X_i, X_j) \\ &\quad + \varepsilon\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) - \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i) \\ &\text{s.t.} \ \ \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0 \\ &\quad 0 \le \alpha_i^*, \alpha_i \le C \qquad i = 1,2,\cdots,l \end{aligned} \tag{3}$$

By solving this quadratic programming problem above, the w and b in (1) can be obtained:

$$w = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)\varphi(X_i), \ b = y_j - \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)K(X_i, X_j) + \varepsilon$$

Then, the decision function represented in (1) is correspondingly transformed into:

$$f(X) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)K(X_i, X_j) + b \tag{4}$$

Obviously, it isn't difficult to find in (2) that all of the samples are given same weights, hereby ignoring important information implied in some samples and similarly overrating some samples irrelevant to the output of the SVM.

For the solution of the drawback above, a sample-weighting method based on discounted least squares, pioneered by Cao and Tay (2003), is innovated here so that both time discounting and information content are all put into account. The ideology of the innovated discounted least squares here is expressed as follows:

$$C_i = C\frac{2}{1 + \exp(p - 2pi/l)}\left(1 + \frac{m_i}{l}\right)$$

where C_i is penalization coefficient of the weighted i^{th} sample, p is the parameter to control the ascending rate and the weight of the sample is signified by:

$$\frac{2}{1 + \exp(p - 2pi/l)}\left(1 + \frac{m_i}{l}\right)$$

Obviously, the new weight is a product of two terms, that is:

$$\frac{2}{1 + \exp(p - 2pi/l)}$$

and $1+m_i/l$. The first term reflects the time discounting meaning of the sample through the changes in the parameters p and i and the second term is another coefficient which considers information importance implied in the i^{th} sample, with the definition as follows:

Given a sample x_i (i = 1, 2….l), consider its some domain $N(x_i, \delta)$, where δ is a constant specified in advance. If there exist m_i samples in $N(x_i, \delta)$, m_i/l is then defined as an indicator that reflects the importance of i^{th} sample.

Under the sample weighting framework above, the dual problem in (3) is changed into (5):

$$\begin{aligned}\min\ & \frac{1}{2}\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(X_i,X_j)\\ & +\varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i)\\ \text{s.t. }& \sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=0\\ & 0\le\alpha_i^*,\alpha_i\le C_i \qquad i=1,2,\cdots,l\end{aligned} \tag{5}$$

By now, the difference between (3) and (5) is that only the penalization coefficient changes from C to C_i.

Wavelet neural networks: As a feed forward neural network based on Wavelet Transform (WT), the essence of WNN is to find a family of wavelet in the characteristic space so that the complex function relationship contained in the original signal might be exactly expressed. Actually, the wavelet transform is a kind of integral transformation (Zhang and Benveniste, 1992):

$$w_f(a,b)=\int_{-\infty}^{+\infty}f(t)h(a,b,t)dt$$

where f(t) is a function with the dense set:

$$(a,\ b,\ t)=|a|^{-1/2}h\left(\frac{t-b}{a}\right)$$

which represents a family of translated and expanded wavelet functions.

Here, h(t) is a basic wavelet and is represented by Morlet function with the expression as:

$$h(t)=\cos(1.75t)\exp(-t^2/2)$$

where $|a|^{-1/2}$ is a normalized coefficient and b and a are translation factor and expansion factor, respectively which enable networks to be provided with flexible function approximation ability.

The topological structure of WNN is expressed in Fig. 1 and can be expressed as follows:

Let h(a, b, t) represent a activation function of neural network, $\{x_{i1}, x_{i2},\ldots, x_{ik},\ldots, x_{im}|\ i=1, 2,\ldots, l; k=1, 2,\ldots, m\}$ as samples of the input vectors and $\{y_i\}$ as samples of the output labels, m and 1 signify numbers of input neurons and output neurons respectively.

At the same time, let k and j(k = 1, 2..., m; j = 1, 2..., J) represent input neurons and output neurons, respectively, where m and J signify numbers of their nodes. And b_j and a_j represent the translation factor and the expansion factor of WNN, respectively and w_{kj} represents the weight of connections between the node of the input lay, k and the node of the hidden layer, j and r_j represents the weight of connections between the hidden layer, j and the output layer. Then, the value of the output node of the network takes as follows:

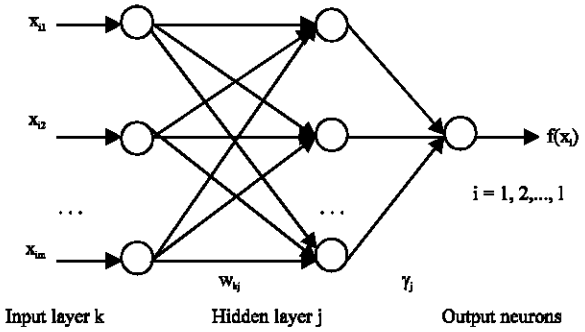

Fig. 1: Topological structure of WNN

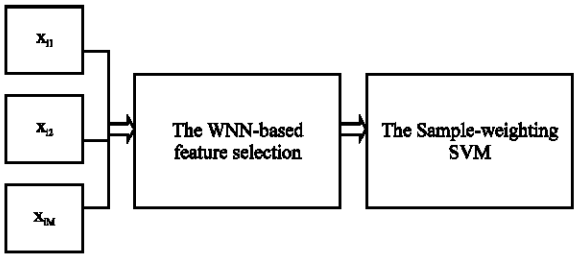

Fig. 2: Integration of sample-weighting SVM and WNN

$$f(x_i)=\sum_{j=1}^{J}r_j h(\lambda_{ji})$$

Where:

$$\lambda_{ji}=\frac{1}{a_j}\left(\sum_{k=1}^{m}w_{kj}x_{ik}-b_j\right)$$

Next, the error function of the network is defined by:

$$\frac{1}{2}\sum_{i=1}^{l}(y_i-f(x_i))^2$$

Under the framework of WNN above, the error function of the network will be increasingly minimized through repeated model training, with the training process stopped until the error is smaller than the accuracy degree given in advance.

The integration of the sample-weighting SVM and WNN: As for the combination of SVMs and WNNs, Chandra *et al.* (2010) proposed the support vector machine and wavelet neural network hybrid. Being contrary to Chandra's methodology, a new hybrid will be constructed underneath, with a rough framework portrayed in Fig. 2.

The main steps prescribed in Fig. 2 are partitioned into two tasks, that is, the feature selection based on WNN and the model constructing for SVM.

First, the feature selection based on WNN shown in Fig. 1 is implemented using a method called "variance rating analysis", pioneered by Yang and Gao (1999). The method can be described as follows:

$$\rho_{jk} = \frac{Cov(x_{ik}, \lambda_{ji})}{Var(x_{ik})Var(\lambda_{ji})},\ \rho_j = \frac{Cov(h(\lambda_{ji}), f(x_i))}{Var(h(\lambda_{ji}))Var(f(x_i))}$$

where, ρ_{jk} and ρ_j represent degrees to which the input nodes influence the hidden nodes and the hidden nodes influence the output nodes, respectively.

Thus, the variance rate, denoted by Z_k, can be taken:

$$Z_k = \sum_{j=1}^{J} \rho_{jk} \cdot \rho_j$$

And the sum of Z_k can be calculated as follows:

$$Z = \sum_{k=1}^{m} |Z_k| \quad (6)$$

Obviously, the Z represents the total influence on the network output from all input variables (i.e., features). If there exist n (n<m) variables and the ration of the corresponding total influence of the n variables to Z is 85% or above, this means that the n variables turn into "important variables" and that the remainders are removed.

Secondly, taking the n variables selected above as components of the input vectors of the SVM in (5), along with their corresponding original output samples as output labels, the innovated sample-weighting SVM can be constructed using (5).

EMPIRICAL ANALYSIS

The proposed hybrid based on the innovated SVM and WNN will be applied in measuring financial market interaction in China below, with the comparison of the proposed hybrid with the classical pattern of SVM exercised first.

The comparison of the integration model with the classical pattern of SVM: Now, the test is used for comparing the integration model shown in Fig. 2 with the classical pattern of SVM in (2) or (3). The prediction performance is evaluated using the following statistical metrics, namely, the normalized mean squared error (NMSE) and the mean absolute error (MAE), described, respectively as follows:

$$NMSE = \frac{1}{\sigma^2 N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2,\ MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (7)$$

where, σ^2 is the normalized squared error of the data, not the same as that in Eq. 8.

Here, the Overnight Call Rate (O/N), one-week monetary market rate (1W), two-week monetary market rate (2W), one-month monetary market rate (1M), three-month monetary market rate (3M), six-month monetary market rate (6M), nine-month monetary market rate (9M), one-year monetary market rate (1Y) and money supply M_0, M_1 and M_2 are chosen as the monetary market price indexes and constitute the input vector of the network which means that the input vector is 11-dimentional. At the same time, the output of the network is just the total stock market trading volume.

The data of all variables above are monthly collected from October 2006 to November 2009 in China, with 39 data patters totally gathered. In order to eliminate the difference in the order of magnitude of samples among different variables, the normalization process is exercised simultaneously.

For the integration model shown in Fig. 2, the initial variation interval, $[a_j, b_j]$, of WNN is determined using trial-and-error techniques and its original value is set [-1, 1]. And the parameters C and sig2 (σ^2) of SVM are initially set 10 and 0.2, respectively and let $p_1 = p_2 = 10$ and δ=0.05, the decision function about the total stock market trading volume comes from the solution of (5). All of implementations are taken using MATLAB software and the testing results are shown in Table 1.

The test results in Table 1 certified the effectiveness of the novel hybrid based on the innovated SVM and WNN, with the values of NMSE and MAE in (7) all being smaller than those of the classical pattern of SVM.

The application of the integration model in measuring market interactions: The interaction between the monetary market and the stock market in China will be still taken as example and the task is to analyze the sensitivity in their price changes.

In order to analyze the sensitivity of changes in the total stock market trading volume to changes in every influencing factor (i.e., every the monetary market price index above), take underneath the partial derivative of the decision function of the SVM, that is the output $f(x_i)$, with respect to the input x_{ik} according to (4) (Cao *et al*., 2003; Shen and Wang, 2011):

$$\frac{\partial y_i}{\partial x_{ik}} = \frac{\partial(\sum_{i=1}^{l}(\alpha_i^* - \alpha_i)K(X_i, X_j) + b)}{\partial x_{ik}},\ k = 1, 2\cdots, 11$$

Table 1: Comparison of the integration model with the classical pattern of SVM

Models	NMSE	MAE
The integration model	0.0839	0.6417
The classical pattern of SVM	0.1284	0.8753

Table 2: Result of feature selection based on WNN

Total stock market trading volume	M_2	1Y	6M	M_1	9M
	46.17	62.23	71.42	80.85	87.65

Table 3: Outcome of the sensitivity analysis (Δ_k)

Δ_k	M_2	1Y	6M	M_1	9M
	0.213	0.101	0.063	0.021	0.008

The radial basis function is chosen as the kernel function of the ε-SVR in (4), with the expression as follows:

$$K(X_i, X_j) = \exp(-\|X_i-X_j\|^2/\sigma^2$$

where, σ^2 is the width of the radial basis function.

$$\frac{\partial y_i}{\partial x_{ik}} = -\frac{2}{\sigma^2}\sum_{j=1}^{l}(\alpha_j^* - \alpha_j)(x_{ik} - x_{jk})\exp(-\frac{1}{\sigma^2}\sum_{l=1}^{n}(x_{il} - x_{jl})^2) \quad (8)$$

where, x_{ik} (i = 1, 2,…, l; k = 1, 2,…, 11) represents the kth component of the sample input vector, X_i and l_s is the number of support vectors.

For every sample of the kth influencing factor, all the values of the sensitivity of the decision function can be obtained by (8) and the averaged sensitivity of the kth influencing factor, denoted by Δ_k, can be expressed in (9):

$$\Delta_k = \frac{1}{1}\sum_{i=1}^{1}\left|\frac{\partial y_i}{\partial x_{ik}}\right| \quad (9)$$

After comparison of Δ_k (k = 1, 2…, 11), the most possible pathway in the interaction between the monetary market and the stock market can be found, with the mechanism of interaction easily established.

All of the data and the values of parameters above are used again in this section underneath. Under the condition that the value of Z in (6) reaches 85% or above, the contribution rates of the selected input variables to the variance of the total stock market trading volume are shown in Table 2.

It is found in Table 2 that there are distinct differences among the five input variables selected, that is, M_2, 1Y, 6M, M_1 and 9M, with the influence of M_2 on the total stock market trading volume being biggest.

Then, the five selected variables displayed in Table 2 constitute the input vector of SVM. The outcome of the sensitivity analysis in (9) is then calculated and revealed in Table 3.

The results in Table 3 display that the sensitivity, Δ_k, in the change of the total stock market trading volume to that of M_2 achieves 0.2130 and is bigger than that of the one-year monetary market rate (1Y). This means that the interaction pathway based on the price of the monetary market, that is, the interest rate, is far premature. Subsequently, the real practice that the money supply is chosen as an intermediary tool for monetary policies sill has important realism significance under existing development conditions of financial markets in China.

CONCLUSION

The integration of the innovated sample-weighting SVM and WNN is indeed an exciting innovation in addressing nonlinearity and instability in financial issues, with various combinations having been constructed so far. Here the innovated sample-weighting SVM based on discounted least squares proposed can not only allow for time discounting meaning but also put the information importance of samples into account, hereby enhancing robustness of the SVM. Furthermore, a kind of WNN is employed beforehand to select features for the SVM using the variance rating methodology so as to eliminate some input variables irrelevant to the output of the artificial intelligence network.

The empirical results certified the effectiveness of the novel hybrid based on innovated SVM and WNN proposed and the outcomes of empirical analysis here have constructive direction significance in financial market development in China.

ACKNOWLEDGMENT

The study is sponsored by Shandong Province Natural Science Foundations (Grant No. ZR2011GM012 and ZR2012GM006) and the National Statistical Science Research Program (Grant No. 2012LY022 and 2012LY054).

REFERENCES

Amornwattana, S., D. Enke and C.H. Dagli, 2007. A hybrid option pricing model using a neural network for estimating volatility. Int. J. Gen. Syst., 36: 558-573.

Becerra, V.M., R.K.H. Galvao and M. Abou-Seada, 2005. Neural and wavelet network models for financial distress classification. Data Mining Knowl. Discovery, 11: 35-55.

Boyacioglu, M.A., Y. Kara and O.K. Baykan, 2009. Predicting bank ?nancial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. Expert Syst. Appli., 36: 3355-3366.

Cao, L., H.P. Lee, C.K. Seng and Q. Gu, 2003. Saliency analysis of support vector machines for gene selection in tissue classification. Neural Comput. Appli., 11: 244-249.

Cao, L.J. and F.E.H. Tay, 2003. Support vector machine with adaptive parameters in financial time series forecasting. IEEE Trans. Neural Networks, 14: 1506-1518.

Chandra, D.K., V. Ravi and P. Ravisankar, 2010. Support vector machine and wavelet neural network hybrid: Application to bankruptcy prediction in banks. Int. J. Data Mining Modelling Manage., 2: 1-21.

Drucker, H., C.J.C. Burges, L. Kaufman, A. Smola and V. Vapnik, 1997. Support Vector Regression Machines. In: Advances in Neural Information Processing Systems 9, Mozer, M.C., J.I. Jordan and T. Petscbe (Eds.). MIT Press, New Jersey, pp: 155-161.

Kim, K.J., 2003. Financial time series forecasting using support vector machines. Neurocomputing, 55: 307-319.

Shen, C.H. and X.R. Wang, 2011. Analysis of convertible bond value based on integration of support vector machine and copula function. Communi. Stat. Simulation Comput., 40: 1563-1575.

Taylor, J.S. and N. Cristianini, 2000. Support Vector Machines and Other Kernel-Based Kearning Methods. Cambridge University Press, London.

Tsai, C.F. and J.W. Wu, 2008. Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Syst. Appli., 34: 2639-2649.

Vapnik, V.N., 1995. The Nature of Statiscal Learning Theory. Springer, New York.

Vapnik, V.N., S.E. Golowich and A. Smola, 1997. Support Vector Nethod for Function Approximation, Regression Estimation and Signal Processing. In: Advances in Neural Information Processing Systems 9, Mozer, M. and M. Jordan and T. Petsche (Eds.). The MIT Press, Cambridge, MA., pp: 281-287.

Whitman, B., G. Flake and S. Lawrence, 2001. Artist detection in music with minnowmatch. Proceedings of the 2001 IEEE Workshop on Neural Networks for Signal Processing, September 10-12, 2001, North Falmouth, MA., pp: 559-568.

Yang, L. and Z.Y. Gao, 1999. Variable selection based on wavelet neural network. J. Northern Jiaotong Univ. (In Chinease).

Zhang, Q.H. and A. Benveniste, 1992. A wavelet networks. IEEE Trans. Neural Networks, 3: 889-898.