

A Rough-based Robust Support Vector Regression Network for Function Approximation

Chih-Ching Hsiao

Department of Electrical Engineering
Kao Yuan University
Kaohsiung City, Taiwan, R.O.C
t20051@cc.kyu.edu.tw

Shun-Feng Su, *Fellow, IEEE*

Department of Electrical Engineering
National Taiwan University of Science
and Technology
Taipei City, Taiwan, R.O.C
sfsu@mail.ntust.edu.tw

Chen-Chia Chuang

Department of Electrical Engineering,
National Ilan University
I-Lan County, Taiwan, ROC
ccchuang@niu.edu.tw

Abstract—Support vector regression (SVR) employs the support vector machine (SVM) to tackle problems of function approximation and regression estimation. SVR has been shown to have good robust properties against noise. Besides, in SVR, outliers may also possibly be taken as support vectors. Such an inclusion of outliers in support vectors may lead to seriously overfitting phenomena. The rough set theory is successful to deal with imprecise, incomplete or uncertain information system. In this paper, a novel regression approach, termed as the Rough Margin Support Vector Regression (RMSVR) network, is proposed to enhance the robust capability of SVR. The basic idea of the approach is to adopt the concept of rough sets to construct the model obtained by SVR and fine tune it with a robust learning algorithm. Simulation results of the proposed approach have shown the effectiveness of the approximated function in discriminating against outliers.

Keywords—support vector regression(SVR); outlier; rough sets; robust learning

I. INTRODUCTION

The support vector machine (SVM) is a universal approach for solving the problems of multidimensional function estimation. Those approaches are all based on the Vapnik–Chervonenkis (VC) theory [1-2]. When SVM is employed to tackle the problems of function approximation and regression estimation, the approaches are often referred to as the support vector regression (SVR) [3]. The SVR type of function approximation is very effective, especially for the case of having a high-dimensional input space. Another important advantage for using SVR in function approximation is that the number of free parameters in the function approximation scheme is equal to the number of support vectors. Such a number can be obtained by defining the width of a tolerance band. Thus, the selection of the number of free parameters can be directly related to the approximation accuracy and does not have to depend on the dimensionality of the input space or other factors as that in the cases of multilayer feedforward neural networks.

Outliers may occur due to various reasons, such as erroneous measurements or noisy phenomenon appearing in the tail portion of some noise distribution functions. When the obtained observations contain noise or outliers, the learning process being unaware of those situations may try to fit those unwanted data and this behavior may lead to a corrupted approximation function. This phenomenon is often called

overfitting [4-5]. In [6], SVR has been shown to have better robust properties for various signal-to-noise ratios. In [7], SVR has also been shown to have excellent performance for both the ε -insensitive function and Huber's robust function matching the correct type of noise in an application of time series prediction. Besides, a general cost function for SVR has also been proposed in considering error distributions [8]. In [9], the SVM with weighted least square (LS-SVM) is proposed to overcome the effects of outliers. However, when the parameters in those approaches are not properly chosen, the final results may be affected by its parameters. This property has also been mentioned in [9]. The selection of parameters in SVR is not straightforward. In [10], a robust SVR (RSVR) network is proposed to enhance the robust capability of SVR. The basic idea of the approach is to adopt the concept of traditional robust statistics to fine tune the model obtained by SVR.

In the last two decades, rough sets and fuzzy sets turned out to be two contemporary progresses in analyzing inexact, imprecise, uncertain, or vague knowledge. The former captures the distinct aspect of indiscernibility in knowledge, while the latter describes the inherent feature of vagueness in linguistics and decision-making. The rough set theory proposed by Pawlak [11-12] is a mathematical theory dealing with uncertainty in data. Rough sets rely on the notion of lower and upper approximations of a set and are applied to rough fuzzy control [13], modeling, system identification [14], discovery of decision rules from experimental data [15] and clustering [16-17]. Recently, combining both rough and SVM [18-24], Lingras [19] proposed a rough set approach to SVM classification removes the necessity of exact classification and is especially useful when dealing with noisy data. By utilizing the boundary region in rough sets shorten the training time and reduced storage requirements compared to the conventional techniques. In [20] proposed a so-called RST-SVM, the redundant attributes in multi-attribute information table can be reduced by RST approach, is utilized as a preprocessor to improve business failure prediction capability by SVM. Asharaf [18] incorporates rough set theoretic flavor in support vector clustering paradigm to achieve soft clustering. The user-defined parameters allow to control the extent of softness by controlling the width of boundary region and number of outliers. Chen [21] proposed fuzzy transitive kernels based fuzzy rough sets, and then reformulated hard margin support vector machines into fuzzy rough set based SVMs (FRSVMs)

with new constraints in which the memberships are taken into account. It has certain capacity to deal with noise. In [22], a called rough support vector regression (RSVR), is proposed to improve the modeling of rough patterns. rough insensitive loss function by applying separate SVR for patterns of lower and upper values. The main advantage of rough-SVR was that it almost never over-predicted the upper values. Based on ν -SVM(SVR) [25], Zhang [23] gave a rough margin based support vector machine for classification to escape from the overfitting problem due to outliers. In [24] focus on the rough margin based ν -SVR which constructs the rough margin with double tolerance bands. In this paper, a novel regression approach, termed as the Rough Margin Support Vector Regression (RMSVR) network, is proposed to enhance the robust capability of SVR. The basic idea of the approach is to adopt the concept of rough sets to construct the model obtained by SVR and fine tune it with a robust learning algorithm. Simulation results of the proposed approach have shown the effectiveness of the approximated function in discriminating against outliers.

The reminder of this paper is outlined as follows. Section II describes the concept of rough sets and the fundamental ideas for SVR. In Section III, the RMSVR network is proposed and its learning process is introduced. Section IV gives experimental results. Concluding remarks are given in Section V.

II. SUPPORT VECTOR REGRESSION AND ROUGH SETS

A. Support vector regression

A function approximation problem can be formulated as to obtain a function from a set of observations $\{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_N, y_N)\}$ with $\vec{x}_i \in R^m$ and $y_i \in R$, where N is the number of training data, $\vec{x}_i$ is the i -th input vector, and y_i is the desired output for the input $\vec{x}_i$. Based on the SVM theory, SVR is to approximate the given observations in an m -dimensional space by a linear function in another feature space F . The function in SVR is of the form

$$f(\vec{x}, \vec{w}) = \langle \vec{w}, \Phi(\vec{x}) \rangle + b \quad (1)$$

where $\langle \cdot, \cdot \rangle$ is an inner product defined on F , $\Phi(\cdot)$ is a nonlinear mapping function from R^m to F , $\vec{w} \in F$ is a weight vector to be identified in the function, and b is a threshold. Usually, the considered cost function [26] is

$$R_{SV}[f] = R_{emp}[f] + C \cdot \|\vec{w}\|^2 \quad (2)$$

$$\text{where } R_{emp}[f] = \frac{1}{N} \sum_{i=1}^N L(y_i - f(\vec{x}_i, \vec{w})) = \frac{1}{N} \sum_{i=1}^N L(e_i) \quad [27],$$

$L(y - f(\vec{x}, \vec{w}))$ is the loss function measuring the error between y and the estimated output $f(\vec{x}, \vec{w})$ for a given input, and $C > 0$ is a regular constant. In traditional SVR, the ε -insensitive function is used as the loss function in (2).

It was shown in [2] that the solution of the above problem can be expressed in terms of support vectors $\vec{w} = \sum_{i=1}^N \theta_i \Phi(\vec{x}_i)$, and the function f is then written as.

$$f(\vec{x}, \vec{w}) = \sum_{i=1}^N \theta_i \langle \Phi(\vec{x}_i), \Phi(\vec{x}) \rangle + b \quad (3)$$

In (3), the inner product $\langle \Phi(\vec{x}_i), \Phi(\vec{x}) \rangle$ in the feature space is usually considered as a kernel function [28]. The kernel function determines the smoothness properties of solutions and should reflect a prior knowledge on data. The choice of the kernel function is usually left for users. The coefficients θ_i s in (3) can be solved by quadratic programming methods with suitable transformation of the above problem into constraint optimization problems. In (3), the constant b is unknown. Various forms can be found in the literature [6].

According to the SVM theory, SVR has the advantage of self-determining its structure. Hence, there are no initialization problems for SVR. For training data sets with certain noise distributions, SVR also has shown excellent performance under the ε -insensitive function or Huber's robust function [3,5,7]. However, the robust effects against training data sets with outliers are not obvious in SVR.

B. Rough sets

The theory of rough sets [11-12] has recently emerged as another major mathematical tool for managing uncertainty that arises from granularity in the domain of discourse—that is, from the indiscernibility between objects in a set. The intention is to approximate a rough (imprecise) concept in the domain of discourse by a pair of exact concepts, called the lower and upper approximations. These exact concepts are determined by an indiscernibility relation in the domain, which, in turn, may be induced by a given set of attributes ascribed to the objects of the domain. The lower approximation is the set of objects definitely belonging to the vague concept, whereas the upper approximation is the set of objects possibly belonging to the same. A rough set X is characterized by its lower and upper approximations $\underline{B}X$ and $\overline{B}X$, respectively, with the following properties.

- 1) An object $\vec{x}_k$ can be part of at most one lower approximation.
- 2) If $\vec{x}_k \in \underline{B}X$ of cluster X , then simultaneously $\vec{x}_k \in \overline{B}X$.
- 3) If $\vec{x}_k$ is not a part of any lower approximation, then it belongs to two or more upper approximations.

III. ROUGH MARGIN SUPPORT VECTOR REGRESSION ALGORITHM

A novel approach, termed as Rough Margin Support Vector Regression (RMSVR) Algorithm, is proposed. Support vector regression (SVR) [3] which is constructed by minimizing Vapnik's ε -insensitive loss function residuals between the outputs of SVR and the target values and the

model complexity. Instead of selecting an appropriate ε , Schölkopf et al [25] proposed a ν -support vector regression (ν -SVR), which introduces a new parameter ν which can control the number of support vectors and training errors without defining ε a priori. They proved that ν is an upper bound on the fraction of margin errors and lower bound of the fraction of support vectors.

In [24] focus on the rough margin based ν -SVR, which constructs the rough margin with double ε instead of single ε for ν -SVR such as Fig.1. Following the rough theory, ε_l and ε_u is the lower margins and the upper margins, respectively. The regions between those two margins are equivalent to the boundary regions; the region inside the lower margins is equivalent to the positive region; and the regions outside the upper margins are equivalent to the negative regions. When the datum lies in the positive region, it is regarded as non-support vector. The datum lying in the boundary and negative regions are all support vectors. The farther the datum deviates from the target function, the more important it is, so it emphasizes the residuals introduced by the data lying in the negative regions. Although the experimental results on benchmark datasets have been conformed the effectiveness of this idea, but if the outliers exist that will result in overfitting phenomena. In our approach, the datum lying in the boundary regions are considered as support vectors, i.e. ε_l and ε_u is the lower bound and upper bound of the fraction of support vectors, respectively. The datum lying in negative regions are viewed as outliers.

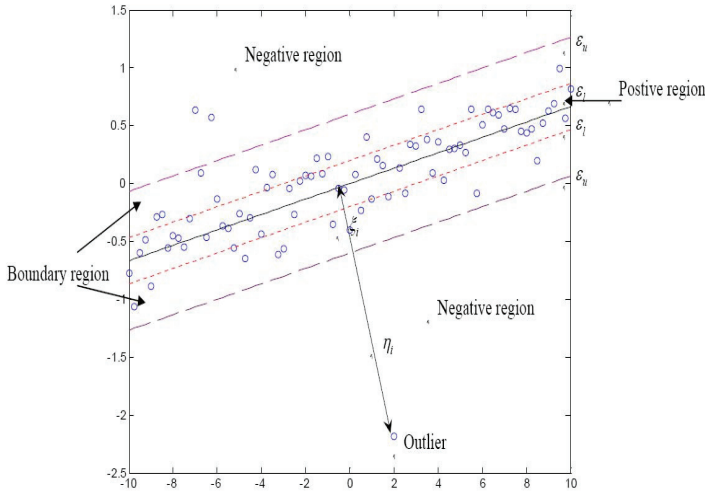


Fig.1 Rough margin SVR

Similar to the classical ν -SVR and rough ν -SVR, the primal problem can be described as:

$$\min_{\vec{w}, \varepsilon_l, \varepsilon_u, \xi_i, \xi_i^*, \eta_i, \eta_i^*} \frac{1}{2} \|\vec{w}\|^2 + C \left(\nu (\varepsilon_l + \delta^b \varepsilon_u) + \frac{1}{l} \sum_{i=1}^l (\xi_i + \xi_i^*) + \frac{\delta^n}{l} \sum_{i=1}^l (\eta_i + \eta_i^*) \right) \quad (4)$$

subject to

$$\begin{aligned} (\vec{w} \cdot \Phi(\vec{x}_i) + b) - y_i &\leq \varepsilon_l + \xi_i + \eta_i, \\ y_i - (\vec{w} \cdot \Phi(\vec{x}_i) + b) &\leq \varepsilon_l + \xi_i^* + \eta_i^*, \\ 0 &\leq \xi_i \leq \varepsilon_u - \varepsilon_l, \quad 0 \leq \xi_i^* \leq \varepsilon_u - \varepsilon_l, \\ \eta_i &\geq 0, \quad \eta_i^* \geq 0, \quad \varepsilon_l \geq 0, \quad \varepsilon_u \geq 0, \end{aligned}$$

The Lagrange function is defined as

$$\begin{aligned} L = & \frac{1}{2} \|\vec{w}\|^2 + C \left(\nu (\varepsilon_l + \delta^b \varepsilon_u) + \frac{1}{l} \sum_{i=1}^l (\xi_i + \xi_i^*) + \frac{\delta^n}{l} \sum_{i=1}^l (\eta_i + \eta_i^*) \right) \\ & - \sum_{i=1}^l \alpha_i (\varepsilon_l + \xi_i + \eta_i + y_i - (\vec{w} \cdot \Phi(\vec{x}_i) + b)) \\ & - \sum_{i=1}^l \alpha_i^* (\varepsilon_l + \xi_i^* + \eta_i^* - y_i - (\vec{w} \cdot \Phi(\vec{x}_i) + b)) \\ & - \sum_{i=1}^l \beta_i (\varepsilon_u - \varepsilon_l - \xi_i) - \sum_{i=1}^l \beta_i^* (\varepsilon_u - \varepsilon_l - \xi_i^*) \\ & - \sum_{i=1}^l \gamma_i \xi_i - \sum_{i=1}^l \gamma_i^* \xi_i^* - \sum_{i=1}^l \sigma_i \eta_i - \sum_{i=1}^l \sigma_i^* \eta_i^* - \lambda_l \varepsilon_l - \lambda_u \varepsilon_u \end{aligned} \quad (5)$$

where $\alpha_i \geq 0, \alpha_i^* \geq 0, \beta_i \geq 0, \beta_i^* \geq 0, \gamma_i \geq 0, \gamma_i^* \geq 0, \sigma_i \geq 0, \sigma_i^* \geq 0, \lambda_l \geq 0, \lambda_u \geq 0$ are Lagrangian multipliers. According to KKT conditions and then the dual problem can be written as

$$\min_{\alpha_i, \alpha_i^*} \frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\vec{x}_i, \vec{x}_j) - \sum_{i=1}^l (\alpha_i - \alpha_i^*) y_i \quad (6)$$

subject to

$$\begin{aligned} \sum_{i=1}^l (\alpha_i - \alpha_i^*) &= 0 \\ \alpha_i, \alpha_i^* &\in \left[0, \delta_n \frac{C}{l} \right] \\ \sum_{i=1}^l (\alpha_i + \alpha_i^*) &\leq (1 + \delta_b) C \nu \end{aligned}$$

The regression estimate then takes the form,

$$f(x) = \sum_{i=1}^l \theta_i K(\vec{x}_i, \vec{x}_j) + b, \quad (7)$$

where $\theta_i = \alpha_i - \alpha_i^*$ and $K(\vec{x}_i, \vec{x}_j)$ is a kernel function.

We let $\nu \geq 0, \delta^b \geq 0, \delta^n \geq 0$. If $\delta^b = 0, \delta^n = 1$, RMSVR is equivalent to the classical ν -SVR. In other word, let $\delta^b = 1, \delta^n \geq 1$, RMSVR becomes the rough ν -SVR. In this function approximation approach, we select $\delta^b > 1$ to emphasize those support vectors lying in the boundary regions, and select $\delta^n < 1$ give the penalty to the datum lying in negative regions. Hence, b can be computed as:

$$\begin{aligned} b = & \frac{1}{2} \left\{ \min_{i \notin NG} \left(y_i - \sum_{j=1}^N (\alpha_j - \alpha_j^*) K(\vec{x}_i, \vec{x}_j) \right) \right. \\ & \left. + \max_{i \in NG} \left(y_i - \sum_{j=1}^N (\alpha_j - \alpha_j^*) K(\vec{x}_i, \vec{x}_j) \right) \right\} \end{aligned} \quad (8)$$

where NG expressed the datum lying in negative regions.

The initial phase is to determine the network structure and the corresponding initial network weights through the RMSVR approach. In this paper, the kernel function is chosen as Gaussian function and defined as

$$K(\vec{x}, \vec{x}_i) = \exp\left[-\|\vec{x} - \vec{x}_i\|^2 / 2\tau^2\right] \quad (9)$$

where is τ a constant. After applying the RMSVR approach, an initial network is obtained which has been with a level of accuracy against outliers. Furthermore, to improve the accuracy, it should to adjust those weights by the robust learning approach. A robust learning algorithm called ARBP [5] is proposed, in this algorithm, it is simply embedding an annealing process into the learning process, and a deterministic annealing process is used to replace the scale estimator. Based on this idea, the robust cost function of ARBP is defined as:

$$E_{ARBP} = \sum_{i=1}^l \chi(e_k, \kappa(\tau)), \quad (10)$$

where $\chi(\cdot)$ is the loss function. $e_k = y_k - \hat{y}_k$. The annealing schedule, $\kappa(\tau)$ is usually selected to decrease toward zero as the epoch number increases to infinite; that is $\lim_{\tau \rightarrow \infty} \kappa(\tau) = 0$.

In the learning algorithm, $\bar{\theta}$ can be updated as

$$\bar{\theta}(t+1) = \bar{\theta}(t) + \Delta \bar{\theta}(t) \quad (11)$$

with

$$\Delta \bar{\theta}(t) = \varsigma \frac{\partial E_{ARBP}(t)}{\partial \bar{\theta}}, \quad (12)$$

where ς is the learning constant.

The proposed Algorithm is described in the following:

- [Step 1] Initialize the RMSVR structure with the given kernel functions, the loss function, and the constant C, v, δ^b, δ^n .
- [Step 2] For each training pattern, compute the estimated result by Equ.(7) and its error.
- [Step 3] Update the weight vector incrementally by (11) and (12).
- [Step 4] Compute the robust cost function defined by (10).
- [Step 5] If the termination conditions are not satisfied, then go to Step 2; otherwise, terminate the learning process.

IV. SIMULATION EXAMPLE

A sinc function is considered and defined as

$$y = \frac{\sin(x)}{x}, \quad -10 \leq x \leq 10.$$

80 input-output data are used. In this example, outliers are added artificially by moving some points away from designated locations. The parameters selected as $\tau = 3, C = 0.5$. Based on the above procedure, the approximated results of RMSVR are presented in Fig. 1 and RMSVR with robust learning are shown

as Fig.2. It can be found that the proposed RMSVR can reduce the overfitting phenomena.

V. CONCLUSION

Rough set theory turned out to be particularly adequate for the analysis of various types of data with inexact, uncertain or vague knowledge. In this paper, a novel regression approach, termed as the Rough Margin Support Vector Regression (RMSVR) network, is proposed to enhance the robust capability of SVR. The basic idea of the approach is to adopt the concept of rough sets to construct the model obtained by SVR and fine tune it with a robust learning algorithm. Simulation results of the proposed approach have shown the effectiveness of the approximated function in discriminating against outliers.

ACKNOWLEDGMENT

This work was supported in part by National Science Council of Taiwan under Grant NSC 98-2221-E-244 -004 -.

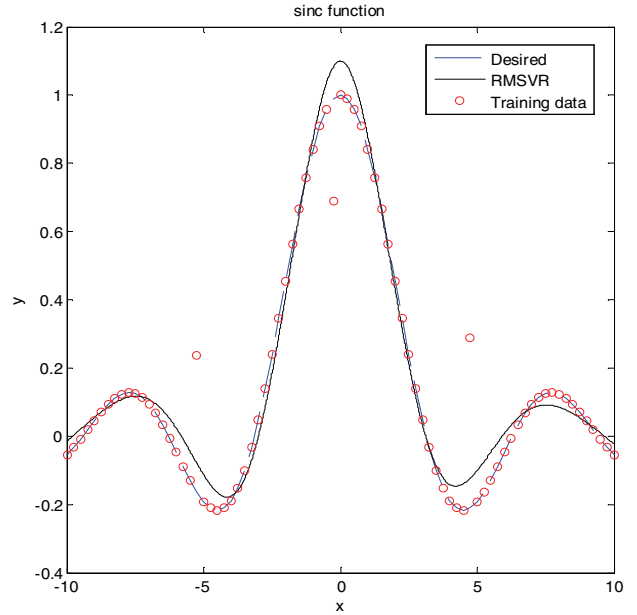


Fig.1 The approximated results

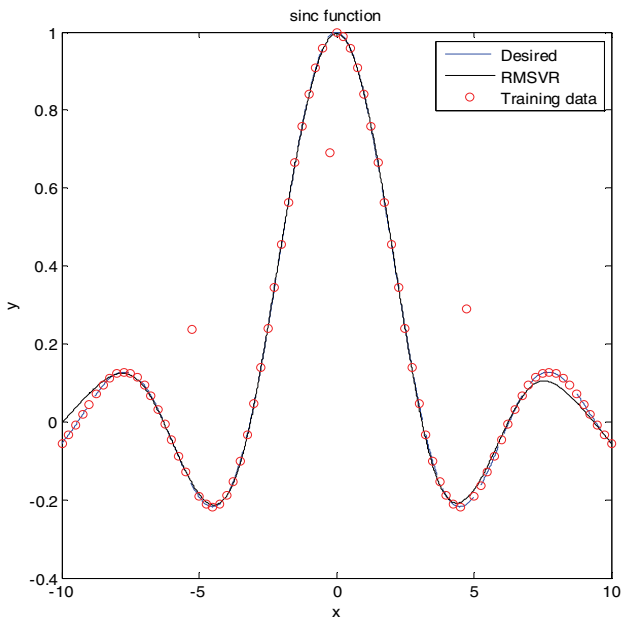


Fig.2 The approximated results with robust learning

REFERENCES

- [1] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.
- [2] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
- [3] V. Vapnik, S. Golowich, and A. J. Smola, "Support vector method for function approximation, regression estimation, and signal processing," in *Neural Information Processing Systems*. Cambridge, MA: MIT Press, 1997, vol. 9.
- [4] D. S. Chen and R. C. Jain, "A robust back-propagation learning algorithm for function approximation," *IEEE Trans. Neural Networks*, vol. 5, pp. 467–479, May 1994.
- [5] C.-C. Chuang, S.-F. Su, and C.-C. Hsiao, "The annealing robust backpropagation (ARBP) learning algorithm," *IEEE Trans. Neural Networks*, vol. 11, pp. 1067–1077, Sept. 2000.
- [6] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Royal Holloway College, London, U.K., Neuro COLT Tech. Rep. TR-1998-030, 1998.
- [7] S. Mukherjee, E. Osuna, and F. Girosi, "Nonlinear prediction of chaotic time series using a support vector machine," in *Proc. NNSP*, 1997, pp. 24–26.
- [8] A. J. Smola, "Regression estimation with support vector learning machines," Master's thesis, Technical Univ. Munchen, Munich, Germany, 1998.
- [9] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: Robustness and sparse approximation," *Neurocomput.*, 2001.
- [10] C. C. Chuang, S. F. Su, J. T. Jeng and C. C. Hsiao, "Robust support vector regression networks for function approximation with outliers," *IEEE Trans. Neural Networks*, vol. 13, no. 6, pp. 1322–1330, 2002.
- [11] Pawlak. Z, "Why rough set?," *Proceedings of IEEE international conference on Fuzzy systems*, New Orleans, Louisiana. USA, 1996, pp. 738–743.
- [12] Pawlak. Z, "Rough sets," *Int. J. Comput. Inf. Sci.*, 1982, vol.11, pp. 341–356.
- [13] T. Y. Lin, "Rough Logic Formalism For Fuzzy Controllers-A Hard and Soft Computing Approach," *Jouranal of Approximate Reasoning*, pp.395–414, 1996.
- [14] C. C. Hsiao, "A Rough-Set-Based for fuzzy modeling with outlier," *2008 International Conference on Instrumentation, Control and Information Technology(SICE 2008)*, pp.364–368, August, 2008.
- [15] X. Wang, Eric C.C. Tsang, S. Zhao, D. Chen and Daniel S. Yeung, "Learning fuzzy rules from fuzzy samples based on rough set technique", *Information Sciences* , vol.177, pp.4493–4514, 2007.
- [16] Sushmita Mitra, Haider Banka, and Witold Pedrycz, "Rough-Fuzzy Collaborative Clustering", *IEEE Trans. Syst., Man, Cybern-Part B*, vol. 36, no. 4, pp.795–805, 2006.
- [17] Pradipta Maji, Sankar K. Pal, "Rough Set Based Generalized Fuzzy C-Means Algorithm and Quantitative Indices", *IEEE Trans. Syst., Man, Cybern-Part B*, vol. 37, no. 6, pp.1529–1540, 2007.
- [18] S. Asharaf, S.K. Shevade, M. and Narasimha Murty, "Rough support vector clustering", *Pattern Recognition*, vol.38 , pp.1779 – 1783, 2005.
- [19] P. Lingras, C.J. Butz, "Rough set based 1-v-1 and 1-v-r approaches to supportvector machine multi-classification," *Information Sciences*, vol.177, pp. 3782–3798, 2007.
- [20] C.C. Yeh , D.J. Chi and M.F. Hsu, "A hybrid approach of DEA, rough set and support vector machines for business failure prediction," *Expert Systems with Applications*, vol. 37, pp.1535–1541, 2010.
- [21] D. Chen, Q. He and X. Wang, "FRSVMs: Fuzzy rough set based support vector machines," *Fuzzy Sets and Systems* , vol.161, pp.596–607, 2010.
- [22] P. Lingras, C.J. Butz, "Rough support vector regression," *European Journal of Operational Research*, 2009.
- [23] Zhang, J., & Wang, Y., "A rough margin based support vector machine," *Information Sciences*, vol.178, pp.2204–2214, 2008.
- [24] Y. Zhao and J. Sun, "Rough ν -support vector regression," *Expert Systems with Applications*, vol.36, pp.9793–9798, 2009.
- [25] Schölkopf, B., Smola, A., Williamson, R., and Bartlett, P., "New support vector algorithms," *Neural Computation*, vol.12, pp.1207–1245, 2000.
- [26] A. J. Smola and B. Schölkopf, "From regularization operators to support vector kernels," in *Neural Information Processing Systems*. Cambridge, MA: MIT Press, 1997, vol. 9.
- [27] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York: Springer-Verlag, 1982.
- [28] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoér, "Theoretical foundations of the potential function method in pattern recognition learning," *Automat. Remote Contr.*, pp. 821–837, 1964.