ACSIJ Advances in Computer Science: an International Journal, Vol. 2, Issue 5, No.6 , November 2013
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Performance assessment among hybrid algorithms in tuning SVR parameters to predict pipe failure rates

Moosa Kalanaki[1], Jaber Soltani[2*]

[1] Department of Computer Engineering, Rouzbahan Higher Education Institute
Sari, Iran
Moses.kalanaki@gmail.com

[2*] Assistant Professor, Irrigation and Drainage Engineering Department, Aboureyhan Campus,
University of Tehran
Tehran, Iran
Jsoltani@ut.ac.ir

Abstract

Pipe failures often occur in water distribution networks and result in large water loss and social-economic damage. To reduce the water loss and maintain the conveyance capability of a pipe network, pipes that experienced a severe failure history are often necessary to be replaced. Several studies and methods have been introduced for predicting failure rates in urban water distribution network pipes by researchers, each of them has some special features regarding the effective parameters and many methods such as Classical and Intelligent methods are used, leading to some improvements. In this paper, the method incorporates hybrid support vector machine and heuristic algorithms techniques for efficient tuning of SVM *meta*-parameters for predicting water distribution network. Performance results are Compared with continuous genetic algorithm-based SVR (SVR-GA), continuous ant colony algorithm-based SVR (SVR-ACO), particle swarm optimization-based SVR (SVR-PSO), artificial neural networks (ANNs) and adaptive neuro-fuzzy inference systems (ANFIS).

Keywords: *Support vector regression, Heuristic algorithms, Kernel functions, Loss functions, Pipe failure rates.*

1. Introduction

In water supply systems, the accidents occurring in pipes are of the utmost importance and sensitivity. Great attention has been paid for reduction of the hydraulic losses as power consumption constitutes a substantial portion of operational costs for the overall pipeline transport. Failure of the pipes is not necessary with the end of their life and different factors namely age, diameter, material, stability and corrosion of soil and water, execution, installation and operational conditions such as hydraulic pressure are effective on it.

To show the importance of this issue, in 1998, about one million accidents have occurred, attributing to themselves more than 20 percent of the total Water and Wastewater Companies budget used for repairs and rehabilitation. Of course, 30 percent of the incidents have occurred to the distribution system pipes. In addition, studies show the maintenance costs of the traditional water sector has increased from 3 million dollars in 1999 to 10 million dollars in 2001 [1]. In most cases, accidents and pipe failures occur as a result of several factors some of which being measurable such as age, length, diameter, depth and pressure of the pipes [2]. Hence, order to provide a comprehensive model, all these factors should be considered. Many studies have employed with many methods, including ANN, ANFIS, fuzzy logic, and SVM in related field.

Another study with comparing among NLR, ANN, and ANFIS methods with some effective parameters results of the comparisons indicated that ANN and ANFIS methods are better predictors of failure rates compared with NLR. The results of the comparison between ANN and ANFIS showed that ANN model is more sensitive to pressure, diameter and age than ANFIS; So, ANN was more reliable [2].

SVM techniques used non-linear regression for environmental data and proposed a multi-objective strategy, MO-SVM, for automatic design of the support vector machines based on a genetic algorithm. MO-SVM showed more accurate in prediction performance of the groundwater levels than the single SVM [3].

Pressure sensitive and EPANET was used for estimating and Hydraulic modeling. EPANET results were used as SVM inputs. Research results showed that the leakage rate is predictable and the smallest changes are predictable using the employed sensors [4].

Rough set theory and support vector machine (SVM) was proposed to overcome the problem of false leak detection. For the computational training of SVM, used artificial bee colony (ABC) algorithm, the results are compared with those obtained by using particle swarm optimization (PSO).Finally; obtained high detection accuracy of 95.19% with ABC[5].

In this paper, effective parameters to predict pipe failure rates of water distribution network are taken into three models and compared with continuous genetic algorithm-based SVR (SVR-CGA), continuous ant colony algorithm-based SVR (SVR-CACO), particle swarm optimization-based SVR (SVR-PSO), artificial neural networks (ANNs) and adaptive neuro-fuzzy inference systems (ANFIS). This research tries to optimizing parameters related to SVR and selecting an optimal model of SVR to better pipes failure rate prediction. By comparing these results with the other methods such as ANFIS, and ANN-GA that had been done in the past results show the SVR-CACO model has better performance than the other models in time elapse.

2. Material and method

2.1. Support Vector Machine

Support vector machine is an algorithm to maximize a mathematical function based on data sets. To create maximum margin, at first two adjacent parallel planes and a separator is designed. They get away from each other until they hit the data. The plane farthest from the others is the best separator [6].

Support vector machine regression (SVR) is a method to estimate the mapping function from Input space to the feature space based on the training dataset [7].

In the SVR model, the purpose is estimating w and b parameters to get the best results. w is the weight vector and b is the bias, which will be computed by SVM in the training process.

In SVR, differences between actual data sets and predicted results is displayed by ε. Slack variables are (ξ_i, ξ_i^*) considered to allow some errors that occurred by noise or other factors. If we don't use slack variables, some errors may occur, and then the algorithm cannot be estimated. Margin is defined as margin=$\frac{1}{\|w\|}$. Then, to maximize the margin, through minimizing $\|w\|^2$, the margin becomes maximized. These operations give in Eqs (1- 3) and these are the basis for SVR [7].

$$\text{Minimize: } \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \quad (1)$$

$$\text{subject to: } y_i(w^T x_i+b)\geq 1-\xi_i,\ \xi_i\geq 0 \quad (2)$$

x_i is the input space and y_i is the feature space. w is the weight vector and b is the bias, which will be computed by SVR in the training process. *C is* a parameter that determines the trade-off between the margin size and the amount of error in training.

A kernel function is a linear separator based on inner vector products and is defined as follows:

$$k(x_i,x_j)=x_i^T x_j \quad (3)$$

If data points are moved using $\varphi: x \rightarrow \varphi(x)$ to the feature space (higher dimensional space), their inner products turn into Eq. (4) [5].

$$k(x_i,x_j)=\varphi(x_i)^T.\varphi(x_j) \quad (4)$$

x_i is the support vectors and x_j is the training data. Accordingly, with using kernel functions and determining derivatives of w and b, also using Lagrange multiplier the SVR function $F(\mathbf{x})$ becomes the following function.

$$F(x)=\sum_{i=1}(\bar{\alpha}_i-\alpha_i)K(x_i,x)+b \quad (5)$$

α_i is the vector of Lagrange multipliers and represent support vectors. If these multipliers are not equal to zero, they are multipliers; otherwise, they represent support vectors [11].

Loss function determines how to penalize the data while estimating. A Loss function implies to ignore errors associated with points falling within a certain distance. If ε-insensitive loss function is used, errors between $-\varepsilon$ and $+\varepsilon$ are ignored. If C=Inf is set, regression curve will follow the training data inside the margin determined by ε [6]. The related equation is shown in Equation 6.

$$|\xi|_\varepsilon=\begin{cases}0 & if\ |\xi|\leq\varepsilon\\ |\xi|-\varepsilon & otherwise.\end{cases} \quad (6)$$

2.2 Ant colony algorithm

The Ant colony optimization algorithm is an optimized technique for resolving computational problems which can be discovered good paths. The process by which ants could establish the shortest path between ant nests and food. Initially, ants leave their nest in random directions to search for food. This technique can be used to solve any computational problem that can be reduced to finding better paths in a graph these formulas have been shown in Eqs. (7) and (8). This method has been chosen from a paper [13].

$$P_k(i,j)=\begin{cases}\arg\max_{s\in M_k}\{[\tau(i,s)]^\alpha[\eta(i,s)]^\beta & \text{, if } q\leq q0\\ \text{Eq. (8)}\end{cases} \quad (7)$$

$$P_k(i,j)=\begin{cases}\frac{[\tau(i,s)]^\alpha[\eta(i,s)]^\beta}{\sum_{s\in M_k}[\tau(i,s)]^\alpha[\eta(i,s)]^\beta} & j\notin M_k\\ 0 & \text{O.w}\end{cases} \quad (8)$$

where $\tau(i, j)$ is the pheromone level between node i and node j, $\mu(i, j)$ is the inverse of the distance between nodes i and j. In this study, the forecasting error represents the distance between nodes. The α and β are parameters determining the relative importance of pheromone level and M_k is a set of nodes in the next column of the node matrix for ant k. q is a random uniform variable [0, 1] and the value q_0 is a constant between 0 and 1, i.e., $q_0 \in [0, 1]$. The local and global updating rules of pheromone are expressed as Eqs. (9) and (10), respectively

$$\tau(i,j)=(1-\rho)\,\tau(i,j)+\rho\tau_0 \quad (9)$$

$$\tau(i,j)=(1-\delta)\tau(i,j)+\delta\Delta\tau(i,j) \quad (10)$$

The δ is the global pheromone decay parameter, $0 < \delta < 1$, and, based on authors' experiments.

The $\Delta\tau$ (i, j), expressed as Eq. (11), is used to increase the pheromone on the path of the solution.

$$\Delta\tau(i,j) = \begin{cases} \frac{1}{L} & \text{, if } (i,j)\epsilon \text{globalbest route} \\ 0 & \text{O.W} \end{cases} \quad (11)$$

where L is the length of the shortest route.

At first to get SVR related parameters, each parameters show by 10 nodes, so the range of numbers limited between [0, 9]. For getting more accurate computed parameters 5 numbers considered.
The values ρ of and τ0 are set to be 0.2 and 1, respectively. Assume the limits of parameters σ, C, and ε are 1, 100,000, and 1, respectively. Numbers of nodes for each ant set to 50, so total nodes are equal to 150.

2.3 Continuous Genetic Algorithm

The Continuous Genetic Algorithm is inherently faster than the binary GA, because the chromosomes do not have to be decoded prior to the evaluation of the cost function [14]. Thus, using the aforementioned variables like kernel parameters as decision variables in a population-based optimization strategy may be a way of constructing an optimal SVR. To cover the entire search space, the initial population was considered randomly, commensurate with the best fitness function Eq. (12) of each population; the best of them has been selected. Some properties of GA, such as the ability of solving hard problems, noise tolerance, easy to interface and hybridize, make them a suitable and quite workable technique for parameter identification of fermentation models.

Minimize cost: $\left|\overline{Y}_{pred}-Y_{train}\right|$ (12)

Where Y_{pred} and Y_{train} are predicting and training output, respectively. In this algorithm, the parameters must be optimized and determined by GA that includes iterations to reach convergence. At first, number of chromosomes, mutation rates, and crossovers must be correctly determined to reach the best results. The next step, Objective function, decision variables and their constraints must be determined in this model. To start the optimization process, initial GA variables like mutation, crossover and selection rates must be determined. Also required parameters for SVR such as kernel and loss functions must be determined.

2.4 Particle Swarm Optimization

The particle swarm optimization (PSO) was designed by [15]. This algorithm simulates the moving of social behaviour among individuals (particles) through a multi-dimensional search space, each particle represents a potential solution and has a position represented by a position vector [15].
A swarm of particles moves through the problem space, with the moving velocity of each particle represented by a velocity vector [16]. PSO operates in three steps at first Define each particle as a potential solution to a problem and best positions have been selected. Each particle has a velocity and by selecting, the best of them these velocities will update.

$$v_i(t+1)=wv_i(t)+c_1r_1\left[\hat{x}_i(t)-x_i(t)\right]+c_2r_2[g(t)-x_i(t)] \quad (13)$$

$v_i(t)$ is the particle's velocity at time t and $x_i(t)$ is the particle's position at time t and $\hat{x}_i(t)$ is the particle's individual best solution in time t. g(t) is the swarm's best solution in time t and w is inertia weight.

3. Case Study

A part of a water distribution network of a city in Iran is considered as the study area. This city is one of the cities being frequently visited by travellers. The area of this district is 2,418 hectares, with a population of 93719 people, supplied with 579,860 meters of distribution pipes including steel pipes 800, 700 and 600 millimetres in diameter, asbestos cement and cast iron pipes 400, 300, 250, 200, 150, 100 and 80 millimetres in diameter.
The installation and execution of the network pipelines in this area were generally started in 1981. According to statistical records, this region has the highest failure rate especially on asbestos cement. In this study, due to incomplete data on steel and cast iron pipes, asbestos cement pipes are only used in the modelling process.
In order of modelling the failure rate of the asbestos cement pipes, the daily events have been recorded from 2005 to 2006 and analysed as to 2438 record data including some information such as diameter, year of implementation, installation depth, total accidents happened and the average of hydraulic pressure. These data have been collected from local water and water waste company.

4. Model Construction

There are different measures by which SVM performance is assessed. The total available data are divided as training data (70% of data) and test data (20% data) and the rest of them used for validation data are chosen randomly. SVM algorithm was trained on training data and its performance has been estimated on test data.

As we know, diversity in the initial solution of optimization algorithms such as GA and PSO is vital for preventing premature phenomena.
This research has been developed by MATLAB (version 7.12(R 2011a)) and SVM Toolbox and parameters were localized by searching algorithms to solve these problems. Eq. (14) is used for normalizing the Input values to the models.

$$x_n=0.8\frac{(x-x_{min})}{(x_{max}-x_{min})}+0.1 \quad (14)$$

x is the original value, x_{min} is the minimum value and x_{max} is the maximum value between input values, and x_n shows normalized values. So that, input results are between [0.1, 0.9].Also, In this paper, the root of mean squared error (RMSE), normal root of mean squared error (NRMSE) and coefficient of determination(R^2) are used as assessment criteria of the reliability of the model.

$$R^2=\frac{\left(\sum_{i=1}^{n}(y_{actual}-\bar{y}_{actual})(y_{pred}-\bar{y}_{pred})\right)^2}{\sum_{i=1}^{n}(y_{actual}-\bar{y}_{actual})^2\sum_{i=1}^{n}(y_{pred}-\bar{y}_{pred})^2} \quad (15)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{actual_i}-y_{pred_i}\right)^2} \quad (16)$$

$$NRMSE=\frac{RMSE}{var(y_{actual})} \quad (17)$$

Where y_{actual} is the observed data, $y_{prediction}$ is the predicted data, $y_{average}$ is the average of data, and n is the number of observations. Also, var(y_{actual}) is the variance of actual data.

Table 1- Related kernel functions

Kernel Type	Formula	Related variables
Gaussian RBF	$k=e^{-\frac{(u-v)(u-v)'}{2p_1^2}}$	P_1 defines RBF function width, like as.δ
Exponential RBF	$k=e^{-\sqrt{\frac{(u-v)(u-v)'}{2p_1^2}}}$	P_1 defines eRBF function width like as RBF.
Polynomial	$k=(UV+1)^{p1}$	P_1 determines Polynomial degree.
Spline	$Z=1+UV+\left(\frac{1}{2}\right)UV\min(u,v)-\left(\frac{1}{6}\right)\min(u,v)^3$ $k=Prod(z)$	

5. Result

The results obtained by mentioned algorithms and appropriate kernel and loss functions have been shown in Fig.[1-6], SVR-CACO and SVR-PSO results extracted from [17] and [18] respectively.
Time and error rate parameters considered as a comparison parameters. Fig.1 notes CACO searching algorithm has a better performance in time-consuming than the CGA. Results from ε-insensitive loss function and related kernel functions present in Fig. [1-6]. According to the figure's results can be found CACO searching algorithm as same as CGA in accuracy but it has a better performance in time-consuming than the CGA.

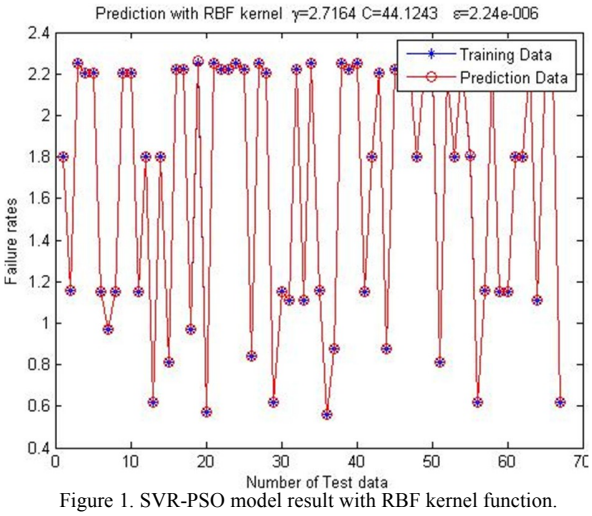

Figure 1. SVR-PSO model result with RBF kernel function.

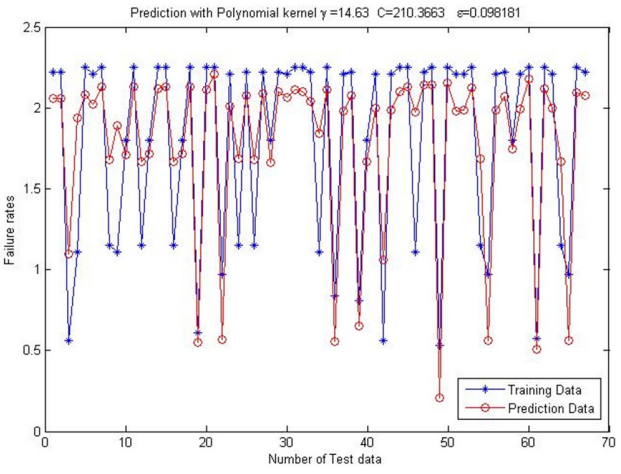

Figure 2. SVR-PSO model result with polynomial kernel function

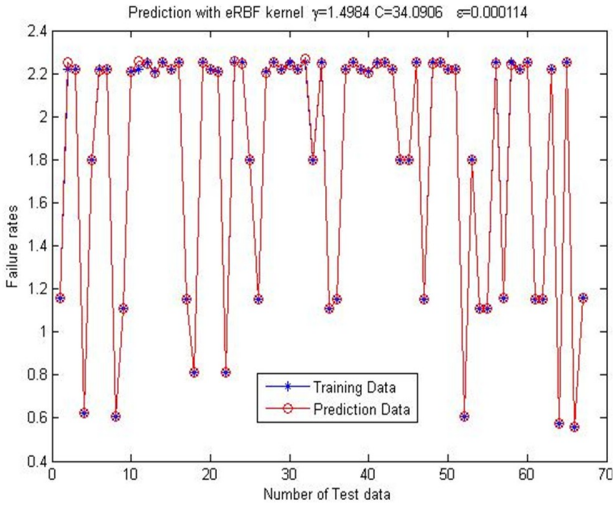

Figure 3. SVR-CGA model result with eRBF kernel function

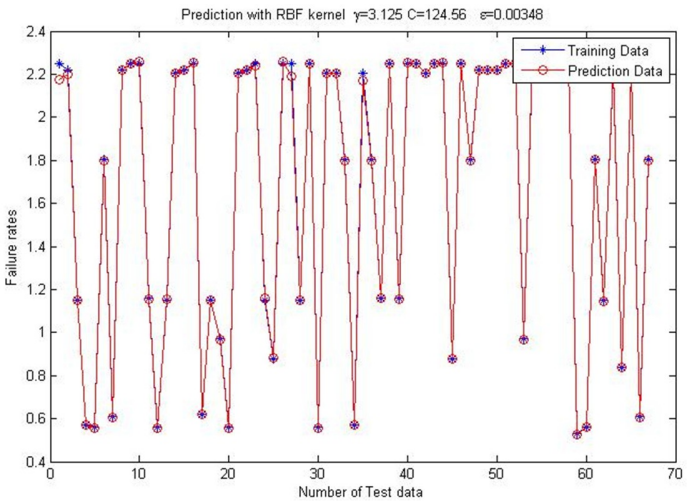

Figure 4. SVR-CACO model result with RBF kernel function.

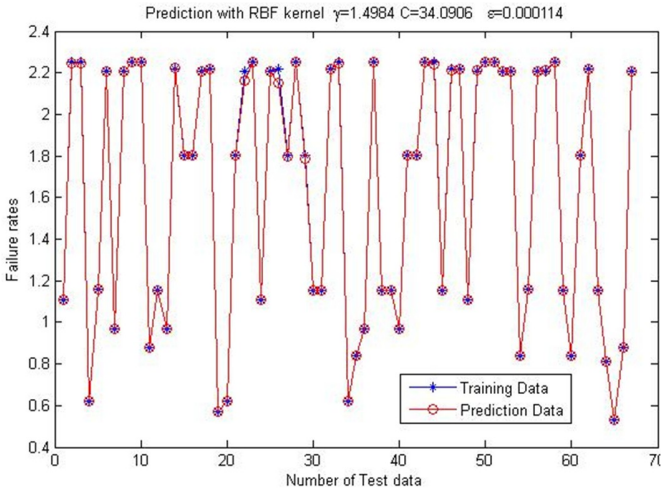

Figure 5. SVR-CGA model result with RBF kernel function.

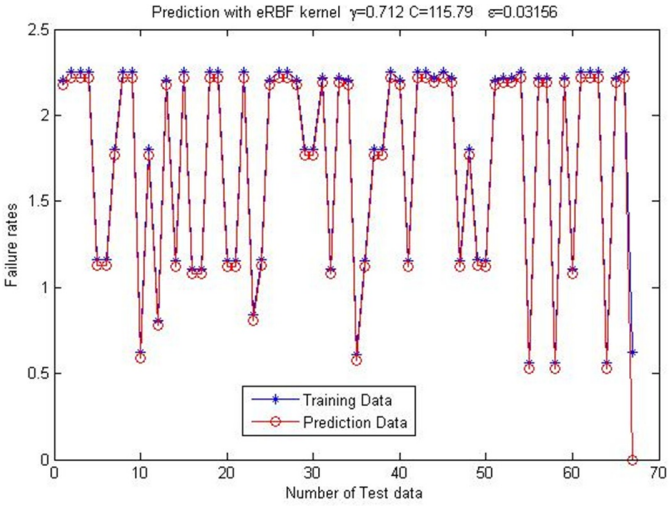

Figure 6. SVR-CACO model result with eRBF kernel function.

So, in order of modeling pipe failures in urban water distribution systems, heuristic algorithms-based SVR model with RBF kernel and quadratic loss function were used and the results were determined. In the following part, Table 2 consists of ε-insensitive and quadratic loss function and related kernel functions. These results indicate that RBF kernels presented better results, because in each of them predicted data are very similar to actual data it means predicting process has a high accuracy.

Table 2. Notes in RBF kernel function and both (quadratic & ε-insensitive) loss function SVR-PSO presented the best result and time-consuming in SVR-CACO was less.

In Figure 2. Predicted result shows low accuracy because the predicted data is not fitted to actual ones.

Table results indicate that RBF function offered the best results, because it has acted faster and showed better performance compared with other kernel functions as regards the data correlation and error parameters.

Table 2.Comparison results among heuristic algorithm-based SVR

	Kernel Type	Algorithm Type	Time(s)	R^2	RMSE	NRMSE	C	P1	ε
ε-insensitive	RBF	CACO- SVR	7170.8	0.999714	0.01395395	0.0374732	124.56	3.125	0.00348
		SVR-CGA	10416	0.9998343	0.00795713	0.0220281	34.0906	1.4984	0.000114
		SVR-PSO	7793.7	0.9999842	0.00248263	0.0069736	44.12426	2.7163	0.00000224
	eRBF	CACO-SVR	6896.5	0.9908480	0.07701178	0.20325	115.79	0.712	0.03156
		SVR-CGA	8798	0.97392	0.01407516	0.17277	176.021	1.1031	0.0007262
		SVR-PSO	69١6.5	0.98846625	0.07598939	0.2010738	123.8638	3.4919	2.22E-16
	Polynomial	CACO-SVR	7919.4	0.4876852	0.4423224	1.24770	54.754	6.3154	0.02346
		SVR-CGA	8516	0.49546	0.73222312	1.90157	78.7583	9.2783	0.842413
		SVR-PSO	8071.8	0.86154729	0.31666793	0.8624184	210.3663	14.63	0.098181
	Spline	CACO-SVR	7311.7	0.6629965	0.3697717	1.01165	43.57	-	0.12680
		SVR-CGA	9880	0.68674	0.339090	0.90198	4.2508	-	0.142851
		SVR-PSO	6987.9	0.68602510	0.36729896	0.9721290	50.9547	-	0.0001
quadratic	RBF	CACO-SVR	317.23	0.99997672	0.002861	0.00812	72.739	5.4153	--------
		SVR-CGA	549.4	0.99998413	0.002755	0.00691	344.49	3.4962	---------
		SVR-PSO	498.3	0.99998713	0.002664	0.00584	194.79	4.1342	
	eRBF	CACO-SVR	259.82	0.9859318	0.07599	0.20145	344.58	3.83	--------
		SVR-CGA	314.9	0.9989343	0.08333	0.04175	143.13	1.77	--------
		SVR-PSO	287.5	0.9979697	0.08471	0.04516	161.13	2.41	
	Polynomial	CACO-SVR	711.35	0.9623311	0.114152	0.30398	1456.89	5.62	--------
		SVR-CGA	992.35	0.55049	0.906832	0.461017	54.760	8.5465	--------
		SVR-PSO	1018.41	0.74113	0.8166124	0.39916	54.760	8.5465	
	Spline	CACO-SVR	206.03	0.7856149	0.31790	0.81007	356.8	-	--------
		SVR-CGA	391.9	0.7841079	0.784107	0.27634	18.779	-	--------
		SVR-PSO	401.6	0.7131167	0.714619	0.29143	91.119	-	

In Table 3, different types of models were implemented in the past on the same data set.
According to the Table 3; it shows SVR-PSO model has the best performance in all factors except time spent.
The ANN and ANFIS models are significantly different from the SVR-based models.
ANN-GA model has a good performance but has spent a long time [19]. Elapsed time in ANFIS model has an acceptable but performance has not enough accuracy.
Overall, based on RMSE and R^2 measures, we conclude that SVR-based models perform best in terms of prediction accuracy. In addition,its implementation is much easier than that of traditional models such as ANFIS and ANN.

Table 3-Comparison result among other model

Model Type	RMSE	R2	NRMSE	Elapsed-time(s)
SVR-PSO	0.002482634	0.999984	0.006973621	7793.7
ANN-GA	0.009775	0.999232	0.026433	1526.2
ANFIS	0.020785	0.99802	0.17101	469.632

6. Conclusions

In the recent decades, public health and health care has been more attention by responsible in each region.
Development of a pipe break rate prediction model by intelligent models and determination of an optimal replacement time are explored in the research. Part of the water distribution system of Mashhad City is selected as the case study, and the data on pipe properties as well as pipe break records are collected.
SVM methodology with a robust parameter tuning procedure has been described in this work, which can be used effortlessly where phenomenological model is difficult to develop.
The method employs hybrid heuristic algorithms approach for minimizing the generalization error.
Superior prediction performances were obtained for the case study. The results indicate that SVM based technique with the parameters tuning approach by using heuristic algorithm described in this work can yield excellent generalization and can be advantageously employed for a large class of problems encountered in process engineering.

References

[1] N. A. Elahi Panah, Subsequent water distribution networks in the country until 1400 (In Persian). J. Water and Environment, vol 28, pp. 4-15, 1998.
[2] M. Tabesh, J. Soltani, R. Farmani, D.A. Savic, Assessing Pipe failure Rate and Mechanical Reliability of water Distribution Networks Using Data Driven Modeling, J.Hydroinf. Vol 11, 2009, pp. 1-17.
[3] O. Giustolisi, Using a multi-objective genetic algorithm for SVM construction, J. Hydroinf. Vol 8, pp. 2006, 125-139.
[4] J. Mashford, D.D. Silva, D. Marny, S. Burn. An approach to leak detection in pipe networks using analysis of monitored pressure value by support vector machine, IEEE Comp. Society, Vol. 3, 2009, pp. 534 –539.
[5] S.K. Mandal, M.K. Tiwari, F.T.S. Chan, Leak detection of pipeline: An integrated approach of rough set theory and artificial bee colony trained SVM. Expert Systems with Applications, Vo.l 39, pp. 3071–308, 2012.
[6] E. Carrizosa, D. Romero Morales, Supervised classification and mathematical optimization, Computers & OperationsResearch. Vol. 40, 2013, pp. 150–165.
[7]V. N.Vapnik, Principles of risk minimization for learning theory. Advances in Neural Information Processing Sys., Vol 4, 1992, pp. 831-838.
[8] Z. Qi, Y. Tian, Y. Shi, Robust twin support vector machine for pattern classification, J. Pattern Recognition. Vol. 46, 2013, pp. 305-316.
[9] T. Hofmann, B. Scholkopf, A.Smola, Kernel methods in machine learning, The annals of statistics. Vol. 36, 2008, pp. 1171-1220.
[10] I. Tsang, T. Kwok, P. Cheung, Core Vector Machines: Fast SVM Training on Very Large Data Sets, J. Machine Learning Research ,Vol. 6, 2005, pp. 363-392.
[11] V.N. Vapnik, and O. Chapelle, Bounds on error expectation for support vector machines. J. Neural Computation, Vol. 12, pp. 2013-2036,2000.
[12] A.J. Smola, and B. Scholkopf, A Tutorial on Support Vector Regression,NeuroCOLT,Royal Holloway College, University of London, 1998.
[13] W.C. Hong, Y. Dong, F. Zheng, C.Y. Lai. Forecasting urban traffic flow by SVR with continuous ACO, J. Applied Mathematical Modeling, Vol 35, 2011, pp. 1282–129.
[14] R. Haupt, and S. Haupt, Practical Genetic Algorithms, Second Edition, John Wiley and Sons, USA, 2004.
[15] C.F. Juang, A Hybrid of Genetic Algorithm and Particle Swarm Optimization for Recurrent Network Design, IEEE TRANSACTIONS ON SYS, Vol. 34 , 2004, pp. 997-1006.
[16] J. Kennedy, and R. Eberhart, Particle Swarm Optimization, IEEE Int Conf on Neural Networks , 1995, pp.1942–1948.
[17] M. Kalanaki, J. Soltani, S. Tavassoli "Using CACO-SVR in pipe failure rates prediction", 12th Iranian hydraulic conference, 2013.
[18] M. Kalanaki, J. Soltani, S. Tavassoli, "The use of hybrid SVR-PSO model to predict pipes failure rates", International journal of science and engineering research ,Vol.4 , No.11, 2013, pp. 1022-1025.
[19] J. Soltani, and M. Pour Tabari, determination of effective parameters in pipe Failure rate in water distribution system using the combination of Artificial Neural Networks and Genetic Algorithm, J Water & Wastewater Vol.83, 2012, 2-18.