Applied Soft Computing 11 (2011) 2902–2905

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Support vector machine approach for longitudinal dispersion coefficients in natural streams

H. Md. Azamathulla [a,*], Fu-Chun Wu [b]

[a] REDAC, Universiti Sains Malaysia, Engineering Campus, 14300 Nibong Tebal, Pulau Pinang, Malaysia
[b] Department of Bio-Environmental Systems Engineering, National Taiwan University, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 13 March 2010
Received in revised form 1 October 2010
Accepted 28 November 2010
Available online 4 December 2010

Keywords:
Support vector machine
Rivers
Dispersion
Streams

ABSTRACT

This paper presents the support vector machine approach to predict the longitudinal dispersion coefficients in natural rivers. Collected published data from the literature for the dispersion coefficient for wide range of flow conditions are used for the development and testing of the proposed method. The proposed SVM approach produce satisfactory results with coefficient of determination = 0.9025 and root mean square error = 0.0078 compared to existing predictors for dispersion coefficient.

1. Introduction

The longitudinal dispersion of pollutants in rivers is significant to practicing hydraulic and environmental engineers for designing outfalls or water intakes and for evaluating risks from accidental releases of hazardous contaminants [1]. Many researchers have contributed to the understanding of the mechanisms of longitudinal dispersion in rivers, beginning with the simplest dispersion of dissolved contaminants in pipe flow [2]. Later, the concept of dispersion was extended to the mixing in open channels and further to natural streams. Many theoretical and empirical formulations have been proposed to determine the longitudinal dispersion coefficient. This paper presents an alternative approach to estimate longitudinal dispersion coefficient in natural streams using support vector machine (SVM). Fitness of models has been tested using the observed dispersion coefficient as available in literature. Data corresponding to various natural streams has been used for this purpose. From the published results, it has been shown that the longitudinal dispersion coefficients vary within a wide range (1.9–2883.5).

Accurate estimation of longitudinal dispersion coefficient is required in several applied hydraulic problems such as: river engineering, environmental engineering, intake designs, estuaries problems and risk assessment of injection of hazardous pollutant and contaminants into river flows [3,4]. Investigation of quality condition of natural rivers by one-dimensional (1D) mathematical model requires the best estimations for longitudinal dispersion coefficient [5]. When measurements and real data of mixing processes in river are available, the longitudinal dispersion coefficient is determined simply, but in rivers that the mixing and dispersing data is not available and these phenomena are not known, should use alternative methods for estimation of dispersion coefficient values [6]. In these cases, because of the complexity of mixing phenomena in natural rivers, the best estimations of dispersion coefficients are not possible and usually these values are determined by several simple regressive equations [1]. There are several empirical equations for estimation of longitudinal dispersion coefficient in natural rivers that have presented in next sections [7]. Estimation of longitudinal dispersion coefficient in rivers using equations of Table 1 requires hydraulic and geometry of data sets. These equations are valid only in their calibrated ranges of flow and geometry conditions and for larger or smaller ranges have not good results [17,18].

The main aim of this note is to develop the SVM for dispersion coefficient and assessing the accuracy of these methods in comparisons with real data and at least not at end, developing a new and accurate methodology for dispersion coefficient determination. Therefore, the present study applies a soft computing technique SVM.

2. Support vector regression

When support vector machines were first used for classification, in 1996, another version of SVMs was proposed by Drucker

* Corresponding author.
E-mail addresses: redacazamath@eng.usm.my, mdazmath@gmail.com (H.Md. Azamathulla), fcwu@ntu.edu.tw (F.-C. Wu).

1568-4946/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2010.11.026

Table 1
Empirical equations for estimation of longitudinal dispersion coefficient [14].

Reference	Equation	Author
Tayfour and Singh [21]	$K_x = 5.93HU^*$	Elder [22]
Deng et al. [1]	$K_x = 0.58(H/U)^2UB$	McQuivey and Keefer [23]
Fisher et al. [5]	$K_x = 0.011U^2B^2/HU^*$	Fisher et al. [5]
Seo and Bake [4]	$K_x = 0.55BU^*/H^2$	Li et al. [24]
Seo and Bake [4]	$K_x = 0.18(U/U^*)^{0.5}(B/H)^2HU^*$	Liu [25]
Tavakollizadeh and Kashefipur [26]	$K_x = 2.0(B/H)^{1.5}HU^*$	Iwasa and Aya [27]
Seo and Cheong [7]	$K_x = 5.92(U/U^*)^{1.43}(B/H)^{0.62}HU^*$	Seo and Cheong [7]
Sedighnezhad et al. [3]	$K_x = 0.6(B/H)^2HU^*$	Koussis and Rodriguez-Mirasol [15]
FaghforMaghrebi and Givehchi [28]	$K_x = 0.2(B/H)^{1.3}(U/U^*)^{1.2}HU^*$	Li et al. [24]
Rajeev and Dutta [13]	$K_x/HU^* = 2(W/H)^{0.96}(U/U^*)^{1.25}$	Rajeev and Dutta [13]

et al. [8]. The new SVM version contains all of the main features that characterize the maximum margin algorithm, including a nonlinear function that is leaned by linear learning machine mapping into high dimensional kernel induced feature space. The capacity of the system is controlled by parameters that do not depend on the dimensionality of the feature space.

In the same way as with a classification approach, there is motivation to seek and optimize the generalization bounds given for regression. They rely on defining the loss function that ignores errors, which are situated within a certain distance of the true value. This type of function is often called epsilon intensive loss function. In SVR, the input x is first mapped onto a m-dimensional feature space using some fixed (nonlinear) mapping, and then a linear model is constructed in this feature space. Using mathematical notation, the linear model (in the feature space) $f(x, w)$ is given by

$$f(x, w) = \sum_{j=1}^{n} w_i g_i(x) + b \tag{1}$$

where $gj(x), j = 1, \ldots, n$ are a set of nonlinear transformations, and w and b are the weight vector and the bias terms. The quality of estimation is measured by the loss function $L(y, f(x, w))$. SVM regression uses a new type of loss function called ϵ the insensitive loss function proposed by Vapnik [19,20]:

$$L_\varepsilon(y, f(x, w)) = \begin{cases} 0 & \text{if } |y - f(x, w)| \le \varepsilon \\ |y - f(x, w)| - \varepsilon & \text{otherwise} \end{cases} \tag{2}$$

The empirical risk is

$$R_{emp}(w) = \frac{1}{m}\sum_{i=1}^{m} L_\varepsilon(y_i, f(x_i, w)) \tag{3}$$

SVR performs linear regression in the high-dimension feature space using ε insensitive loss and, at the same time, tries to reduce model complexity by minimizing $||w||^2$. This can be described by introducing (non-negative) slack variables, $\xi_i, \xi_i^*, = 1, \ldots, m$ to measure the deviation of training samples outside the ε-insensitive zone. Thus, SVR is formulated as the minimization of the following function:

$$\min \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*)$$
$$\text{such that} \begin{cases} y_i - f(x_i, w) \le \varepsilon + \xi_i^* \\ f(x_i, w) - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0, \quad i = 1, \ldots, m \end{cases} \tag{4}$$

This optimization problem can transformed into the dual problem and its solution is given by

$$f(x) = \sum_{i=1}^{n_{sv}} (\alpha_i - \alpha_i^*) k(x_i, x)$$
$$\text{subject to } 0 \le \alpha_i^* \le C, \quad 0 \le \alpha_i \le C$$

where n_{sv} is the number of support vectors (SVs) and the $k(x_i, x)$ is the kernel function.

This optimization model can be solved using the Lagrangian method, which is almost equivalent to the method used to solve the optimization problem in the separable case.

Accordingly, the coefficients α_i can be found by solving the following convex quadratic programming problem.

The kernel function is formulated as

$$k(x, x_i) = \sum_{j=1}^{n} g_j(x) g_j(x_i) \tag{5}$$

It is well known that SVM generalization performance (estimation accuracy) depends on a good setting of meta-parameters parameters C and ε and the kernel parameters. The choices of C and ε control the prediction (regression) model complexity. The problem of optimal parameter selection is further complicated because the SVM model complexity (and hence its generalization performance) depends on all three parameters Smola and Schölkopf [9]. Kernel functions are used to change the dimensionality of the input space to perform the classification (or regression) task with more confidence.

Two common kernel functions are radial basis function (RBF):

$$k(x, x') = \exp(-\gamma||x - x'||^2) \tag{6}$$

and a polynomial function:

$$k(x, x') - (xx' + 1)^p \tag{7}$$

The radial parameters $\gamma > 0$ and p are the kernel specific parameters; they are set to values priory and used throughout the training process. Other kernel functions are also introduced that are to be used for specific purposes [10].

An algorithm for solving the problem of regression with support vector machines was proposed by Platt [11] called sequential minimal optimization (SMO). It puts chunking to the extreme by iteratively selecting subsets only of size 2 and optimizing the target function with respect to them. This algorithm has a much simpler background and is easier to implement. The optimization sub problem can be solved analytically solved, without the need to use a quadratic optimizer. Shevade et al. [12] proposed an improvement that enhances the algorithm such that it performs significantly faster.

3. Model development

The scenarios considered in building the SVM model inputs (flow width (W)/flow depth (H)), flow velocity (U)/shear velocity (U^*)) and output (longitudinal dispersion coefficient (m^2/s) K_x/flow depth (H) × shear velocity (U^*). From the collected data sets (Table 2) used in this study, around 60% (58 data set) of these patterns were used for training (chosen randomly until the best training performance was obtained), while the remaining patterns about 20% (20 data set) were used for testing, and about 20% (18 data set) for validating, the SVM model. Software was developed to perform the analysis, and can be obtained from the first author.

The Neurosolutions 5.0 toolbox, developed by Nerodimension Inc. [16], is used while developing SVM model. The model parameters α_i and ε were initially fixed as 1 and 0. A genetic algorithm was used to obtain the optimal value of ε. During the genetic

Table 2
Range of collected data [29].

	Flow width, W (m)	Flow depth, H (m)	Flow velocity, U (m/s)	Shear velocity, U^* (m/s)	K_x (m^2/s)
Max value	711.20	25.1	2.23	0.553	2883.5
Min value	11.89	0.22	0.034	0.0024	1.9
Avg. value	59.86	3.69	0.71	0.095	223.1

search, an initial population of chromosomes (ε values) was created and the fitness of each candidate solution (chromosome) was evaluated against the fitness function (MSE of a threefold cross-validation set). Then the population is evolved through multiple generations (through mutation, crossover and selection), and the optimal solution (chromosome) was selected. Optimal ε is found to be 0.0001 for the present problem. The optimal values of kernel parameters C and σ are found to be 0.35 and 20.0, respectively.

4. Results and discussion of SVM

The performance of the SVM model was compared with the traditional longitudinal dispersion coefficient equations. Overall, particularly for field measurements, the SVM model gives better predictions than the existing models. The SVM model produced the least errors ($R = 0.95$, $R^2 = 0.9025$ and RMSE = 0.00780) and Fig. 1 show the observed and estimated K_x/HU^* of the unseen training data. From Fig. 2 (validation set) it is clear that the traditional predictor [13] under or over estimate the longitudinal dispersion coefficient. SVM produced for test data correlation coefficient, $R = (0.93)$, coefficient of determination R^2 (=0.8641) and root mean square error, (RMSE = 2.234). It can be concluded that for all the data sets the SVM model give either better or comparable results.

The above result are not astonishing, since the most significant advantage of the proposed SVM compared to classical regression analysis based models (traditional equations) is that it is capable of mapping the data into a high dimensional feature space, where a variety of methods (described in the previous section) are used to find relations in the data. Since the mapping is quite general, the relations found in this way are accordingly very general.

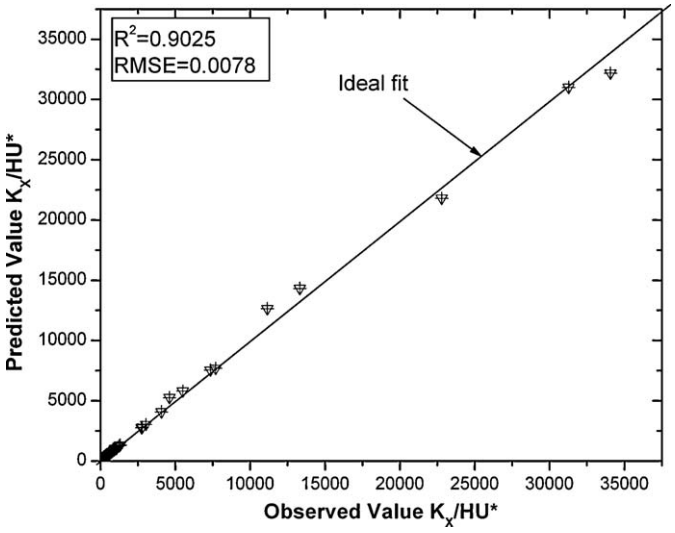

Fig. 1. Comparison of observed versus predicted K_x/HU^* for training data using SVM.

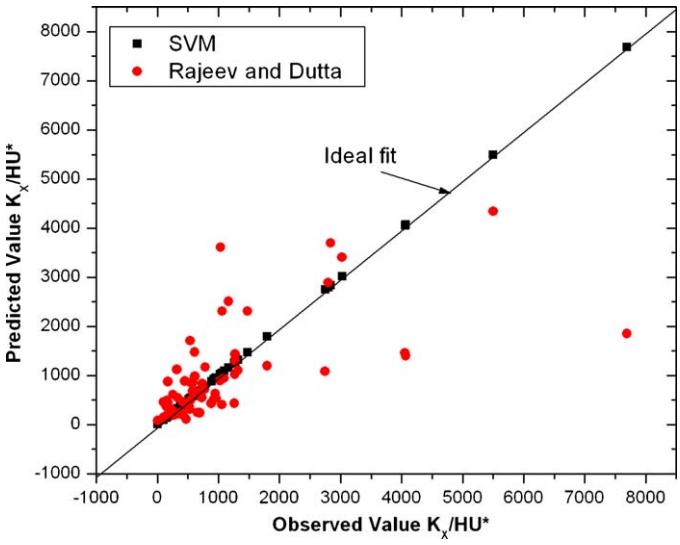

Fig. 2. Comparison of observed versus predicted K_x/HU^* by SVM and Rajeev and Dutta for validation data set.

5. Conclusions

Longitudinal dispersion in rivers is a complex phenomenon. Natural channels have bends, changes in shape, pools and many other irregularities, all of which contribute significantly to the dispersion process. To overcome the complexity and uncertainty associated with the dispersion, this research demonstrates that an SVM model can be applied for accurate prediction of longitudinal dispersion coefficients. The genetic programming will be used to predict longitudinal dispersion coefficient in the future with more database.

Notations

B, W	flow width (m)
H	flow depth (m)
U	flow velocity (m/s)
U^*	shear velocity (m/s)
K_x	longitudinal dispersion coefficient (m^2/s)

References

[1] Z.Q. Deng, V.P. Singh, L. Bengtsson, Longitudinal dispersion coefficient in single channel streams, Journal of Hydraulic Engineering 128 (10) (2001) 901–916.
[2] N. Ahsan, Estimating the coefficient of dispersion for a natural stream, World Academy of Science, Engineering and Technology 44 (2008) 131–135.
[3] H. Sedighnezhad, H. Salehi, D. Mohein, Comparison of different transport and dispersion of sediments in mard intake by FASTER model, in: Proceedings of the Seventh International Symposium River Engineering, 16–18 October, Ahwaz, Iran, 2007, pp. 45–54.
[4] I.W. Seo, K.O. Bake, Estimation of the longitudinal dispersion coefficient using the velocity profile in natural streams, Journal of Hydraulic Engineering 130 (3) (2004) 227–236.
[5] H.B. Fisher, E.J. List, R.C.Y. Koh, J. Imberger, N.H. Brooks, Mixing in Inland and Costal Waters, Academic Press Inc., San Diego, 1979, pp. 104–138.
[6] S.M. Kashefipur, A. Falconer, Longitudinal dispersion coefficients in natural channels, in: Proceedings of the Fifth International Hydro Informatics Conference, 1–5 July, Cardiff University, 2002, pp. 95–102.
[7] I.W. Seo, T.S. Cheong, Predicting longitudinal dispersion coefficient in natural stream, Journal of Hydraulic Engineering 124 (1) (1998) 25–32.
[8] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, in: M. Mozer, M. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems, vol. 9, MIT Press, Cambridge, MA, 1997, pp. 155–161.
[9] A.J. Smola, B. Schölkopf, A Tutorial on Support Vector Regression, Royal Holloway College, London, UK, NeuroCOLT Tech., Rep. TR 1998-030, 1988.
[10] B. Uestuen, W.J. Melssen, L.M.C. Buydens, Facilitating the application of support vector regression by using a universal Pearson VII function

based kernel, Chemometrics and Intelligent Laboratory Systems 81 (2006) 29–40.
[11] J. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Schölkopf, C.J.C. Burges, A.J. Smola (Eds.), Advances in Kernel Methods – Support Vector Learning, MIT Press, Cambridge, MA, 1999, pp. 185–208.
[12] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy, Improvements to the SMO algorithm for SVM regression, IEEE Transactions on Neural Networks (2000).
[13] R.S. Rajeev, S. Dutta, Prediction of longitudinal dispersion coefficients in natural rivers using genetic algorithm, Hydrology Research 40 (6) (2009) 544–552.
[14] R. Hossien, A. Seyed Ali, K. Ehsan, E. Mohammad Mehdi, An expert system for predicting longitudinal dispersion coefficient in natural streams by using ANFIS, Expert Systems 36 (4) (2009) 8589–8596.
[15] A.D. Koussis, J. Rodriguez-Mirasol, Hydraulic estimation of dispersion coefficient for streams, Journal of Hydraulic Engineering, ASCE 124 (1998) 317–320.
[16] NEUROSOLTIONS 5.0, www.neurosolutions.com. NeuroDimension, Inc., 2009.
[17] Z. Ahmad, Estimation of longitudinal dispersion coefficients, Journal of Hydraulic Engineering, Indian Society for Hydraulics 9 (1) (2003) 14–28, ISH publishers.
[18] Z. Ahmad, Mixing length for establishment of longitudinal dispersion in streams, International Journal of Modelling & Simulation 29 (2) (2009) 1–10, Acta Press.
[19] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[20] V. Vapnik, Statistical Learning Theory, Springer, NY, 1998.
[21] G. Tayfour, V.P. Singh, Predicting longitudinal dispersion coefficient in natural streams by artificial neural network, Journal of Hydraulic Engineering 131 (11) (2005) 991–1000.
[22] J.W. Elder, The dispersion of marked fluid in turbulent shear flow, Journal of Fluid Mechanics 5 (1959) 544–560.
[23] R.S. McQuivey, T.N. Keefer, Simple method for predicting dispersion in streams, Journal of Environmental Engineering Division, American Society of Civil Engineering 100 (4) (1974) 997–1011.
[24] Z.H. Li, J. Huang, J. Li, Preliminary study on longitudinal dispersion coefficient for the gorges reservoir, in: Proceedings of the Seventh International Symposium Environmental Hydraulics, 16–18 December, Hong Kong, China, 1998.
[25] H. Liu, Predicting dispersion coefficient of stream, Journal of Environment Engineering Division, ASCE 103 (1) (1977) 59–69.
[26] A. Tavakollizadeh, S.M. Kashefipur, Effects of dispersion coefficient on quality modeling of surface waters, in: Proceedings of the Sixth International Symposium River Engineering, 16–18 October, Ahwaz, Iran, 2007, pp. 67–78.
[27] Y. Iwasa, S. Aya, Predicting longitudinal dispersion coefficient in open channel flows, in: Proceedings of International Symposium on Environmental Hydraulics, Hong Kong, 1991, pp. 505–510.
[28] M. FaghforMaghrebi, M. Givehchi, Using non-dimensional velocity curves for estimation of longitudinal dispersion coefficient, in: Proceedings of the Seventh International Symposium River Engineering, 16–18 October, Ahwaz, Iran, 2007, pp. 87–96.
[29] Z.F. Toprak, H.K. Cigizoglu, Predicting longitudinal dispersion coefficient in natural streams by artificial intelligence methods, Hydrological Processes 22 (2008) 4106–4129.