Support vector machine approach for longitudinal dispersion coefficients in natural streams

H. Md. Azamathulla\textsuperscript{a,}\textsuperscript{*}, Fu-Chun Wu\textsuperscript{b}

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

\textbf{A B S T R A C T}

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 produces satisfactory results with coefficient of determination $= 0.9025$ and root mean square error $= 0.0078$ compared to existing predictors for dispersion coefficient.

\begin{itemize}
  \item Article history:
  \begin{itemize}
    \item Received 13 March 2010
    \item Received in revised form 1 October 2010
    \item Accepted 28 November 2010
    \item Available online 4 December 2010
  \end{itemize}
  \item Keywords:
  \begin{itemize}
    \item Support vector machine
    \item Rivers
    \item Dispersion
    \item Streams
  \end{itemize}

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...
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 learned by linear learning machines 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 degree 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

\[ f(x, w) = \sum_{j=1}^{n} w_j g_j(x) + b \]  

where \( g_j(x), j=1, \ldots, n \) 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 \( \varepsilon \) the insensitive loss function proposed by Vapnik [19, 20]:

\[ L_\varepsilon(y, f(x, w)) = \begin{cases} 0 & \text{if } |y - f(x, w)| < \varepsilon \\ |y - f(x, w)| & \text{otherwise} \end{cases} \]  

The empirical risk is

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

SVR performs linear regression in the high-dimension feature space using \( \varepsilon \) 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 \( \varepsilon \)-insensitive zone. Thus, SVR is formulated as the minimization of the following function:

\[ \min \frac{1}{2} ||w||^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) \]  

such that

\[ \begin{cases} y_i - f(x_i, w) \leq \varepsilon + \xi_i^* \\ f(x_i, w) - y_i \leq \varepsilon + \xi_i \\ \xi_i, \xi_i^* \geq 0, \quad i = 1, \ldots, m \end{cases} \]  

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

\[ f(x) = \sum_{i=1}^{n} \alpha_i - \alpha_i^* \]  

subject to \( 0 \leq \alpha_i^* \leq C, \quad 0 \leq \alpha_i \leq C \)

where \( n_{sv} \) is the number of support vectors (SVs) and the \( k(x, x) \) 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 \( \alpha_i \) can be found by solving the following convex quadratic programming problem.

The kernel function is formulated as

\[ k(x, x_i) = \sum_{j=1}^{n} g_j(x)g_j(x_i) \]  

### 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 \( (m^2/s) \)/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 \( \alpha_i \) and \( \varepsilon \) were initially fixed as 1 and 0. A genetic algorithm was used to obtain the optimal value of \( \varepsilon \). During the genetic
search, an initial population of chromosomes ($\varepsilon$ 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 $\varepsilon$ is found to be 0.0001 for the present problem. The optimal values of kernel parameters $C$ and $\sigma$ 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 $K_x/\text{HU}^*$ 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.

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

[10] B. Uestuen, W.J. Melssen, L.M.C. Buylens, Facilitating the application of support vector regression by using a universal Pearson VII function