Neural Computing and Applications (2020) 32:3633–3648
https://doi.org/10.1007/s00521-018-3921-3

ORIGINAL ARTICLE

Large-margin Distribution Machine-based regression

Reshma Rastogi[1] · Pritam Anand[1] · Suresh Chandra[2]

Received: 31 January 2018 / Accepted: 29 November 2018 / Published online: 12 December 2018

Abstract
This paper presents an efficient and robust Large-margin Distribution Machine formulation for regression. The proposed model is termed as 'Large-margin Distribution Machine-based Regression' (LDMR) model, and it is in the spirit of Large-margin Distribution Machine (LDM) (Zhang and Zhou, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2014) classification model. The LDM model optimizes the margin distribution instead of minimizing a single-point margin as is done in the traditional SVM. The optimization problem of the LDMR model has been mathematically derived from the optimization problem of the LDM model using an interesting result of Bi and Bennett (Neurocomputing 55(1):79–108, 2003). The resulting LDMR formulation attempts to minimize the ϵ-insensitive loss function and the quadratic loss function simultaneously. Further, the successive over-relaxation technique (Mangasarian and Musicant, IEEE Trans Neural Netw 10(5):1032−1037, 1999) has also been applied to speed up the training procedure of the proposed LDMR model. The experimental results on artificial datasets, UCI datasets and time-series financial datasets show that the proposed LDMR model owns better generalization ability than other existing models and is less sensitive to the presence of outliers.

Keywords Support vector machine · Regression · Large-margin Distribution Machine · ϵ-insensitive loss · Quadratic loss · Successive over-relaxation

1 Introduction

Support vector machine (SVM) is one of the most popular machine learning algorithms (Cortes and Vapnik [1]; Burges [2]; Cherkassky and Mulier [3]; Vapnik [4]). SVM has emerged from the research in statistical learning theory on how to regulate the trade-off between the structural complexity and empirical risk. It has outperformed the other existing tools in a wide variety of applications. Some of these applications can be found in Osuna et al. [5], Joachims [6], Schlkopf et al. [7] and Lal et al. [8]. SVM has been initially developed for the problem of pattern classification, but it has been extended to solve the problems of regression and clustering as well.

SVM classifier attempts to reduce the generalization error by maximizing the minimum margin, i.e. the minimum distance of the training points from the classification boundary (Cortes and Vapnik [1]; Burges [2]; Cherkassky and Mulier [3]; Vapnik [4]; Bradely and Mangasarian [9]). The maximization of margin in SVM classification problem amounts to the minimization of an upper bound on the VC dimension (Burges [2]; Vapnik [4]) of the classifying hyperplane. The maximum margin theory is not only relevant in the case of the SVM, but it has also been extended to interpret the good generalization ability of many other learning approaches such as AdaBoost (Freund and Schapire [10]) which is a major representative of ensemble methods (Zhou [11]). The theoretical studies which advocate the relevance of the maximum margin principle in these learning approaches can be found in Breiman [12] and Schapire et al. [13]. Moreover, some of the recent theoretical results (Reyzin and Schapire [14]; Wang et al.

✉ Reshma Rastogi
reshma.khemchandani@sau.ac.in

Pritam Anand
ltpritamanand@gmail.com

Suresh Chandra
chandras@maths.iitd.ac.in

[1] Faculty of Mathematics and Computer Science, South Asian University, New Delhi 110021, India

[2] Department of Mathematics, Indian Institute of Technology Delhi, New Delhi 110016, India

[15]; Gao and Zhou [16]) suggest that rather than simply considering a single-point margin, the margin distribution is important in these algorithms. Taking motivation from these studies, Teng and Zhou [17] have proposed Large-margin Distribution Machine (LDM) which tries to achieve better generalization performance by considering the margin mean as well as margin variance in the objective function of its optimization problem.

The support vector methodology has also been extended for handling the problem of regression (Vapnik et al. [18]; Drucker et al. [19]) The standard ϵ-SVR is an ϵ-insensitive model which sets an epsilon tube around the data points. The data points outside the epsilon tube contribute to the errors which are penalized in the objective function via a user-specified parameter. Bi and Bennett [20] have developed a geometric framework for SVR, showing that it can be related to an appropriate SVM problem. This result of Bi and Bennett [20] is very significant as it provides a classification eye to see the problem of regression.

This paper proposes an efficient Large-margin Distribution Machine-based Regression (LDMR) model which is similar in principle to LDM model for classification (Zhang and Zhou [17]). The proposed LDMR model is a more general regression model as the standard ϵ-SVR and LS-SVR models are special cases of LDMR model with particular choice of parameters. The optimization problem of LDMR model has been derived from the LDM model for classification using a well-known result of Bi and Bennett [20]. The resulting optimization problem simultaneously minimizes the quadratic loss function used in LS-SVR (Suykens et al. [21, 22]) and ϵ-insensitive loss function used in ϵ-SVR. It is noteworthy that the LS-SVR model fails to predict well on noisy datasets, whereas ϵ- SVR model totally ignores all data points which lie inside of ϵ-tube for the determination of the regressor. This strategy makes ϵ-SVR model sparse but does not minimize the scatter inside of the ϵ- tube. Our proposed LDMR model aims to take advantage of both of these models. Further, an effective successive over-relaxation (SOR) (Mangasarian and Musicant [25]) technique has also been applied for the efficient solution of the LDMR problem to reduce its training time complexity. Experimental results on the artificial datasets, UCI benchmark datasets (Blake and Merz [28]) and time-series datasets show that the LDMR model owns better generalization ability than the existing SVR models.

Taking forward the arguments of Huang et al. [29] to the regression analogue, it makes sense that apart form the sparsity, we should also minimize the scatter of data points which lie inside the ϵ-tube. Therefore, it is meaningful to have both of these loss functions in the proposed formulation so as to do a trade-off between sparsity and scatter minimization. It also enables the proposed model to utilize the full information of the training set and avoid over-fitting simultaneously.

We now briefly describe notations used in the rest of this paper. All vectors will be taken as column vectors, unless it has been specified otherwise. For any vector $x \in R^n$, $\|x\|$ will denote the l_2 norm. A vector of ones of arbitrary dimension will be denoted by e. Let (A, Y) denote the training set where $A = [A_1; A_2; \ldots; A_l]$ contains the l points in $\mathbb{R}^n$ represented by l rows of the matrix A and $Y = [y_1, y_2, \ldots, y_l] \in \mathbb{R}^{l\times 1}$ contains the corresponding response value of the row of matrix A.

The rest of this paper is organized as follows. Section 2 discusses LDM formulation proposed by Zhang and Zhou [17]. Section 3 briefly describes SVR models. Section 4 proposes the Linear Distribution Machine-based Regression (LDMR) and its extension for the nonlinear case. Section 5 contains the mathematical derivation of the optimization problem of the LDMR from the LDM formulation (Zhang and Zhou [17]) using a result of Bi and Bennett [20]. Section 6 describes the experimental results, while Sect. 7 is devoted to the conclusions.

2 Large-margin distribution machine

Inspired by the theoretical result of Gao and Zhou [16], the LDM model (Zhang and Zhou [17]) attempts to maximize the margin mean and minimize the margin variance simultaneously for optimizing the margin distribution. The margin mean $\bar{\gamma}$ and margin variance $\hat{\gamma}$ are given by

$$\bar{\gamma} = \frac{1}{l}\sum_{i=1}^{l} d_i\left(w^{\mathrm{T}}x_i\right) = \frac{1}{l}d^{\mathrm{T}}Xw, \tag{1}$$

$$\hat{\gamma} = \frac{1}{l}\sum_{i=1}^{l}\left(d_i(w^{\mathrm{T}}x_i) - \bar{\gamma}\right)^2. \tag{2}$$

Here the matrix $X = [x_1, x_2 \ldots, x_l]$ contains the given l data points in $\mathbb{R}^n$, and vector $d \in \mathbb{R}^{l\times 1}$ of ± 1 represents the corresponding class.

The linear LDM model finds the separating hyperplane $w^{\mathrm{T}}x = 0$ by solving the following optimization problem

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}w^{\mathrm{T}}w + \lambda_1\hat{\gamma} - \lambda_2\bar{\gamma} + C\sum_{i=1}^{l}\xi_i \\ &\text{subject to,} \\ &\quad y_i(w^{\mathrm{T}}x_i) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, \ldots, l, \end{aligned} \tag{3}$$

λ_1 and λ_2 are positive parameters for trading off the margin mean and margin variance. It is also notable that the LDM formulation reduces to the SVM when λ_1 and $\lambda_2 = 0$ in (3).

3 Support vector regression

3.1 ϵ-Support Vector Regression

Linear ϵ-SVR model finds a linear function $f(x) = w^{\mathrm{T}}x + b$, where $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$. To measure the empirical risk, it uses the ϵ-insensitive loss function

$$R^{\epsilon}_{\mathrm{emp}} = \frac{1}{l}\sum_{i=1}^{l} |y_i - f(x_i)|_{\epsilon}, \tag{4}$$

where $|y_i - f(x_i)|_{\epsilon} = \max(0, |y_i - f(x_i)| - \epsilon)$. The ϵ-SVR model minimizes the ϵ-insensitive loss function with a regularization term $\frac{1}{2}||w||^2$ in its optimization problem which is given as follows

$$\begin{aligned} &\min_{w,b,\xi,\xi^*} \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ &\text{subject to,} \\ &\quad y_i - (A_i w + b) \le \epsilon + \xi_i, (i = 1, 2, \ldots, l), \\ &\quad (A_i w + b) - y_i \le \epsilon + \xi_i^*, (i = 1, 2, \ldots, l), \\ &\quad \xi_i \ge 0, \xi_i^* \ge 0, (i = 1, 2, \ldots, l). \end{aligned} \tag{5}$$

Here $C > 0$ is the user-specified parameter that balances the trade-off between the fitting error and the flatness of the function $f(x) = w^{\mathrm{T}}x + b$.

3.2 Least-squares support vector regression model

Similar to ϵ-SVR model, least-squares support vector regression (LS-SVR) model (Suykens et al. [21, 22]) also finds a linear function $f(x) = w^{\mathrm{T}}x + b$, where $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$. The LS-SVR model minimizes the quadratic loss function

$$R_{\mathrm{emp}} = \frac{1}{l}\sum_{i=1}^{l}(y_i - f(x_i))^2, \tag{6}$$

in its optimization problem along with the regularization term $\frac{1}{2}||w||^2$. The optimization problem of the LS-SVR model can be expressed as

$$\begin{aligned} &\min_{w,b,\xi} \frac{c}{2}||w||^2 + C_1\sum_{i=1}^{l}(\xi_i^2) \\ &\text{subject to,} \\ &\quad y_i - (A_i w + b) = \xi_i, (i = 1, 2, \ldots, l), \end{aligned} \tag{7}$$

where $C_1 > 0$ is a user-defined parameter.

4 Large-margin distribution machine-based regression

In this section, we propose an efficient Large-margin Distribution Machine-Based Regression (LDMR) model.

4.1 Linear LDMR model

Given the training set (A, Y), the LDMR model finds a linear function $f(x) = w^{\mathrm{T}}x + b$, where $w \in R^n$ and $b \in R$. The proposed LDMR model minimizes the following generalized loss function

$$R^{f}_{\mathrm{emp}} = \frac{k}{2}\sum_{i=1}^{l}(y_i - f(x_i))^2 + C \cdot \sum_{i=1}^{l}|y_i - f(x_i)|_{\epsilon}. \tag{8}$$

Along with the regularization term, here $k > 0$ and $C > 0$. By introducing the regularization term $\frac{1}{2}||w||^2$ and slack variables ξ_1 and ξ_2, the primal form of the LDMR can be expressed as

$$\begin{aligned} &\min_{w,b,\xi_1,\xi_2} \frac{c}{2}||w||^2 + \frac{k}{2}||Y - (Aw + eb)||^2 + Ce^{\mathrm{T}}(\xi_1 + \xi_2) \\ &\text{subject to,} \\ &\quad Y - (Aw + eb) \le e\epsilon + \xi_1, \\ &\quad (Aw + eb) - Y \le e\epsilon + \xi_2, \\ &\quad \xi_1, \xi_2 \ge 0, \end{aligned} \tag{9}$$

where C, k, ϵ and c are user-defined positive parameters.

The proposed LDMR model minimizes three terms in its optimization problem. The first term $\frac{1}{2}w^{\mathrm{T}}w$ attempts to make the regressor as flat as possible. The second term $\sum_{i=1}^{l}(y_i - f(x_i))^2$ attempts to minimize the scatter of the data points, while the third term $\sum_{i=1}^{l}|y_i - f(x_i)|_{\epsilon}$ tries to have better sparsity. These three terms in the objective functions are traded off appropriately to make full use of the training set and avoid over-fitting of the data points simultaneously.

The ϵ-SVR model only minimizes the ϵ-insensitive loss function which ignores the error up to ϵ. The points lying on the bounding regressor $f(x) + \epsilon$ and $f(x) - \epsilon$ and outside the ϵ-insensitive zone are support vectors and participate in the construction of the final regressor. The data points which lie inside the ϵ-insensitive zone have been ignored to achieve the sparsity, but it also causes ϵ-SVR to lose the information contained in the training set. The outliers which are supposed to reside outside the ϵ-insensitive zone also affect the orientation and position of the regressor. On the other hand, the LS-SVR model minimizes the quadratic loss function where all of the data points are participating in the construction of the final regressor but cannot avoid

over-fitting of the regressor. That is why, the LS-SVR model fails to perform well on datasets which contain much noise. The proposed LDMR model obtains better generalization ability by finding a good trade-off between the ϵ-insensitive loss function and the quadratic loss function via the user-defined parameters c, k and C. Since the proposed LDMR model also assigns some weights to the points which are inside of the ϵ-insensitive zone in the construction of the final regressor, it is less sensitive to the presence of outliers and therefore can avoid the over-fitting as well. In this sense, the proposed LDMR model combines the benefits of both the SVR and LS-SVR models.

The proposed LDMR model is in the true spirit of the LDM classification model. The optimization problem of the LDMR model has been mathematically derived by the optimization problem of the LDM model using a result of the Bi and Bennett [20]. Unlike SVM, which minimizes the single-point margin only, the LDM model also minimizes the margin mean and margin variance together. It makes LDM formulation insensitive towards noises. These advantages of the LDM model have also been inherited in the LDMR model.

It is also noteworthy that the proposed LDMR model is a very general model. When $k = 0$, the primal problem (9) of proposed LDMR model reduces to primal problem (5) of ϵ-SVR. Also, when C becomes zero in the primal problem (9) of the proposed LDMR formulation, the variables $\xi_1, \xi_2 \geq 0$ are no more minimized in (9) and hence can take any values. Therefore, the constraints of the optimization problem (9) do not make any sense as these are always satisfied for any value of (w,b). Thus, these constraints become redundant. So with $C = 0$, the proposed LDMR model only minimizes $\frac{c}{2}||w||^2 + \frac{k}{2}||Y - (Aw + eb)||^2$ in its optimization problem (9) which is equivalent to solving the optimization problem (7) of the LS-SVR model with $C_1 = \frac{k}{2}$.

In order to find the solution of the primal problem (9), we need to derive its corresponding dual problem. Let us assume $H = [A, e]$ is a augmented matrix and $v = \begin{bmatrix} w \\ b \end{bmatrix}$, then $||w||^2$ can be written as $||w||^2 = v^{\mathrm{T}} I_0 v$, where $I_0 = \begin{bmatrix} I & 0 \\ & \cdot \\ & \cdot \\ 0 & \ldots 0 \end{bmatrix}$ and I is $n \times n$ identity matrix. Now the Lagrangian function for the primal problem (9) can be given by

$$\begin{aligned} L(v, \alpha_1, \alpha_2, \beta_1, \beta_2) = {} & \frac{c}{2} v I_0 v + \frac{k}{2}(Y - Hv)^{\mathrm{T}}(Y - Hv) \\ & + Ce^{\mathrm{T}}(\xi_1 + \xi_2) + \alpha_1^{\mathrm{T}}(Y - Hv - e\epsilon - \xi_1) \\ & + \alpha_2^{\mathrm{T}}(Hv - Y - e\epsilon - \xi_2) - \beta_1^{\mathrm{T}}\xi_1 - \beta_2^{\mathrm{T}}\xi_2, \end{aligned}$$

where $\alpha_1 = (\alpha_1^1, \alpha_1^2, \ldots, \alpha_1^l), \alpha_2 = (\alpha_2^1, \alpha_2^2, \ldots, \alpha_2^l), \beta_1$ and β_2 are the vector of Lagrangian multipliers. The KKT optimality conditions are given by

$$\frac{\partial L}{\partial v} = (cI_0 + kH^{\mathrm{T}}H)v - H^{\mathrm{T}}Y - H^{\mathrm{T}}\alpha_1 + H^{\mathrm{T}}\alpha_2 = 0, \quad (10)$$

$$\frac{\partial L}{\partial \xi_1} = Ce - \alpha_1 - \beta_1 = 0, \quad (11)$$

$$\frac{\partial L}{\partial \xi_2} = Ce - \alpha_2 - \beta_2 = 0, \quad (12)$$

$$Y - Hv \leq e\epsilon + \xi_1, \xi_1 \geq 0, \quad (13)$$

$$Hv - Y \leq e\epsilon + \xi_2, \xi_2 \geq 0, \quad (14)$$

$$\alpha_1^{\mathrm{T}}(Y - Hv - e\epsilon_1 - \xi_1) = 0, \quad (15)$$

$$\alpha_2^{\mathrm{T}}(Hv - Y - e\epsilon_1 - \xi_2) = 0, \quad (16)$$

$$\beta_1^{\mathrm{T}}\xi_1 = 0, \beta_2^{\mathrm{T}}\xi_2 = 0, \quad (17)$$

$$\alpha_1 \geq 0, \alpha_2 \geq 0, \beta_1 \geq 0, \beta_2 \geq 0. \quad (18)$$

Using the above KKT conditions, the dual problem of the primal problem (9) can be obtained as

$$\begin{aligned} \min_{\alpha_1, \alpha_2} & \frac{1}{2}(\alpha_1 - \alpha_2)^{\mathrm{T}} H(cI_0 + kH^{\mathrm{T}}H)^{-1}H^{\mathrm{T}}(\alpha_1 - \alpha_2) \\ & + Y^{\mathrm{T}}H(cI_0 + kH^{\mathrm{T}}H)^{-1}H^{\mathrm{T}}(\alpha_1 - \alpha_2) \\ & - Y^{\mathrm{T}}(\alpha_1 - \alpha_2) + \epsilon e^{\mathrm{T}}(\alpha_1 + \alpha_2) \quad (19) \\ \text{subject to,} & \\ & 0 \leq \alpha_1 \leq Ce, \\ & 0 \leq \alpha_2 \leq Ce. \end{aligned}$$

After obtaining the optimal value of the α_1 and α_2 from (19), we can obtain v using (10) as follows

$$v = \begin{bmatrix} w \\ b \end{bmatrix} = (cI_0 + kH^{\mathrm{T}}H)^{-1}H^{\mathrm{T}}(\alpha_1 - \alpha_2 + Y).$$

For the given $x \in R^n$, the estimated regressor is obtained as follows

$$f(x) = w^{\mathrm{T}}x + b$$

4.2 Nonlinear LDMR model

The nonlinear LDMR model will seek to estimate the function $f(x) = K(x^{\mathrm{T}}, A^{\mathrm{T}})u + b$, where K is an appropriately chosen positive definite kernel.

The nonlinear LDMR model solves the following optimization problem

Table 1 Results on artificial datasets

Dataset	Regressor	SSE/SST	SSR/SST	RMSE
TYPE 1	LDMR	0.0106 ± 0.0058	0.9942 ± 0.0440	0.0323 ± 0.0092
	SVR	0.0115 ± 0.0046	0.9588 ± 0.0544	0.0344 ± 0.0071
	L_1-Norm SVR	0.0163 ± 0.0062	0.9876 ± 0.0608	0.0408 ± 0.0082
	LS-SVR	0.0158 ± 0.0080	0.9573 ± 0.0478	0.0395 ± 0.0106
TYPE 2	LDMR	0.0209 ± 0.0110	0.9956 ± 0.0653	0.0454 ± 0.0126
	SVR	0.0253 ± 0.0116	0.9978 ± 0.0750	0.0504 ± 0.0118
	L_1- Norm SVR	0.0284 ± 0.0131	0.9968 ± 0.0955	0.0534 ± 0.0128
	LS-SVR	0.0335 ± 0.0170	0.9239 ± 0.0665	0.0576 ± 0.0151
TYPE 3	LDMR	0.0328 ± 0.0158	1.0106 ± 0.0899	0.0574 ± 0.0134
	SVR	0.0366 ± 0.0185	1.0050 ± 0.0930	0.0605 ± 0.0146
	L_1-Norm SVR	0.0411 ± 0.0307	1.0132 ± 0.0990	0.0623 ± 0.0219
	LS-SVR	0.0560 ± 0.0290	0.9463 ± 0.0914	0.0744 ± 0.0200
TYPE 4	LDMR	0.0450 ± 0.0244	0.9498 ± 0.1044	0.0664 ± 0.0168
	SVR	0.0493 ± 0.0255	0.9107 ± 0.1094	0.0697 ± 0.0165
	L_1- Norm SVR	0.0563 ± 0.0250	0.9306 ± 0.1110	0.0753 ± 0.153
	LS-SVR	0.0471 ± 0.0287	0.9086 ± 0.0909	0.0676 ± 0.0188
TYPE 5	LDMR	0.0923 ± 0.0581	0.9380 ± 0.1363	0.0945 ± 0.0265
	SVR	0.0991 ± 0.0510	0.8299 ± 0.1290	0.0992 ± 0.0299
	L_1-Norm SVR	0.1059 ± 0.0571	0.9331 ± 0.1488	0.1023 ± 0.0259
	LS-SVR	0.0955 ± 0.0557	0.8519 ± 0.1262	0.0966 ± 0.0253
TYPE 6	LDMR	0.0130 ± 0.0057	1.0049 ± 0.0577	0.0915 ± 0.0181
	SVR	0.0188 ± 0.0188	0.9762 ± 0.0631	0.1046 ± 0.0428
	L_1- Norm SVR	0.0178 ± 0.0109	0.9911 ± 0.0685	0.1064 ± 0.0314
	LS-SVR	0.0156 ± 0.0072	0.9741 ± 0.0467	0.0996 ± 0.0209

$$\begin{aligned}
&\min_{u,b} \frac{c}{2}||u||^2 + \frac{k}{2}||Y - (K(A,A^{\mathrm{T}})u + eb)||^2 + Ce^{\mathrm{T}}(\xi_1 + \xi_2) \\
&\text{subject to,} \\
&Y - (K(A,A^{\mathrm{T}})u + eb) \le e\epsilon + \xi_1, \\
&(K(A,A^{\mathrm{T}})u + eb) - Y \le e\epsilon + \xi_2, \\
&\xi_1, \xi_2 \ge 0,
\end{aligned} \tag{20}$$

where C, k, ϵ and c_3 are user-supplied positive parameters. Let us assume $G = [K(A,A^{\mathrm{T}}), e]$ be a augmented matrix and $v = \begin{bmatrix} u \\ b \end{bmatrix}$, then $||u||^2$ can be written as $||u||^2 = v^{\mathrm{T}} I_0 v$, where $I_0 = \begin{bmatrix} I & 0 \\ & \cdot \\ & \cdot \\ 0 & \ldots 0 \end{bmatrix}$ and I is $m \times m$ identity matrix. Similar to the line of the linear case, the dual problem of the primal problem (20) can be obtained as follows

$$\begin{aligned}
&\min_{\alpha_1,\alpha_2} \frac{1}{2}(\alpha_1 - \alpha_2)^{\mathrm{T}} G(cI_0 + kG^{\mathrm{T}}G)^{-1}G^{\mathrm{T}}(\alpha_1 - \alpha_2) \\
&\quad + Y^{\mathrm{T}} G(cI_0 + kG^{\mathrm{T}}G)^{-1}G^{\mathrm{T}}(\alpha_1 - \alpha_2) \\
&\quad - Y^{\mathrm{T}}(\alpha_1 - \alpha_2) + \epsilon e^{\mathrm{T}}(\alpha_1 + \alpha 2) \\
&\text{subject to,} \\
&0 \le \alpha_1 \le Ce, \\
&0 \le \alpha_2 \le Ce.
\end{aligned} \tag{21}$$

After obtaining the optimal value of the α_1 and α_2 from (21), we can obtain v as $v = \begin{bmatrix} u \\ b \end{bmatrix} = (cI_0 + kG^{\mathrm{T}}G)^{-1}G^{\mathrm{T}}(\alpha_1 - \alpha_2 + Y)$. For the given $x \in R^n$, the estimated regressor is obtained as follows

$$f(x) = K(x^{\mathrm{T}}, A^{\mathrm{T}})u + b.$$

4.3 A fast LDMR model using successive over-relaxation technique

The dual problem of the proposed LDMR model (19) or (21) can be written in the following unified form

Table 2 Results on artificial datasets with outliers

Dataset	Regressor	SSE/SST	SSR/SST	RMSE
TYPE 3	LDMR	0.0136	1.10582	0.0392
With outliers	SVR	0.0184	1.1034	0.0455
	L_1-Norm SVR	0.0202	1.1242	0.0476
	LS-SVR	0.0273	1.0299	0.0554
TYPE 6	LDMR	0.0165	1.0239	0.1029
With outliers	SVR	0.0194	0.9969	0.1098
	L_1- Norm SVR	0.0188	1.0048	0.0534
	LS-SVR	0.0427	0.9825	0.1647

$$\begin{aligned} &\max_{\beta} p\beta - \beta^{\mathrm{T}} Q\beta \\ &\text{subject to,} \\ &\quad \beta \in S = \{0 \le \beta \le Ce\}. \end{aligned} \tag{22}$$

For example, the problem (22) becomes the problem (19) when $\beta = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$ $Q = \begin{bmatrix} H(cI_0 + kH^{\mathrm{T}}H)^{-1}H^{\mathrm{T}} & -H(cI_0 + kH^{\mathrm{T}}H)^{-1}H^{\mathrm{T}} \\ -H(cI_0 + kH^{\mathrm{T}}H)^{-1}H^{\mathrm{T}} & H(cI_0 + kH^{\mathrm{T}}H)^{-1}H^{\mathrm{T}} \end{bmatrix}$ and $p = -\big[\, Y^{\mathrm{T}}H(cI_0 + kH^{\mathrm{T}}H)H^{\mathrm{T}} - Y^{\mathrm{T}} + e^{\mathrm{T}}\epsilon, \quad -Y^{\mathrm{T}}H(cI_0 + kH^{\mathrm{T}}H)H^{\mathrm{T}} + Y^{\mathrm{T}} + e^{\mathrm{T}}\epsilon\big]$

Now problem (22) can be efficiently solved by following the successive over-relaxation (SOR) technique (Mangasarian OL and Musicant DR [25]) as follows.

Algorithm 1

We choose $t \in (0, 2)$ and start with any initial value of β say $\beta^0 \in R^n$. After computing β_i, we compute

$$\beta_{i+1} = (\beta_i - tE^{-1}(Q\beta_i - p + L(\beta_{i+1} - \beta_i))), \tag{23}$$

until $||\beta_{i+1} - \beta_i||$ is not less than some prescribed tolerance. Here nonzero elements of $L \in R^{l\times l}$ constitute the strictly lower triangular part of the symmetric matrix Q, and the nonzero elements of $E \in R^{l\times l}$ constitute the diagonal of Q (Shao et al. [23]).

It should be noted that it is well justified in [25] and [26] that the iterates $\{\beta_i\}$ converges R-linearly to the optimal solution $\bar{\beta}$ of the problem (22).

5 LDMR: regression via LDM

In this section, we derive the optimization problem of the proposed LDMR model from the optimization problem of the LDM model by making use of a result of the Bi and Bennett [20]. It has been shown in [20] that for a given $\bar{\epsilon} > 0$ and regression training set (A, Y), a regressor $y = \frac{w}{-\eta}^{\mathrm{T}} x + \frac{b}{-\eta}$, $(\eta > 0)$ is an ϵ-insensitive regressor if and only if the sets D^+ and D^- locate on different sides of $n+1$-dimensional hyperplane $w^{\mathrm{T}}x + \eta y + b = 0$, respectively, where

$$D^+ = \{(A_i, y_i + \bar{\epsilon}), i = 1, 2, \ldots, l\} \text{ and } D^- = \{(A_i, y_i - \bar{\epsilon}), i = 1, 2, \ldots, l\}.$$

In view of this result of Bi and Bennett [20], the regression problem is equivalent to the classification problem of sets D^+ and D^- in R^{n+1}. If we use the LDM methodology (Zhang and Zhou [17]) for the classification of these two sets D^+ and D^-, then we can find the LDMR formulation. For this we calculate the margin mean $\bar{\gamma}$ and margin variance $\hat{\gamma}$ as follows

$$\begin{aligned} \bar{\gamma} = \frac{1}{2l}\Bigg\{ &-\sum_{i=1}^{l}\big(A_i w + \eta(y_i - \bar{\epsilon}) + b\big) \\ &+ \sum_{i=1}^{l}(A_i w + \eta(y_i + \bar{\epsilon}) + b)\Bigg\} = \eta\bar{\epsilon}, \end{aligned} \tag{24}$$

$$\begin{aligned} \hat{\gamma} &= \frac{1}{2l}\Bigg\{\sum_{i=1}^{l}(-(A_i w + \eta(y_i - \bar{\epsilon}) + b) - \eta\bar{\epsilon})^2 \\ &\quad + \sum_{i=1}^{l}((A_i w + \eta(y_i + \bar{\epsilon}) + b) - \eta\bar{\epsilon})^2\Bigg\} \\ &= \frac{1}{2l}\Bigg\{\sum_{i=1}^{l}(-A_i w - \eta y_i - b)^2 + \sum_{i=1}^{l}(A_i w + \eta y_i + b)^2\Bigg\} \\ &= \frac{1}{l}\Bigg\{\sum_{i=1}^{l}(A_i w + \eta y_i + b)^2\Bigg\}, \end{aligned} \tag{25}$$

where $w \in R^n$, $\eta \in R$ and $b \in R$. Now the classification of sets D^+ and D^- using LDM model will result in the following QPP

$$\begin{aligned} &\min_{w,\eta,b,\xi,\xi^*} \frac{1}{2}(w^{\mathrm{T}}w + \eta^2) - \lambda_1\eta\bar{\epsilon} \\ &\quad + \lambda_2\frac{1}{l}\Bigg\{\sum_{i=1}^{l}(A_i w + \eta y_i + b)^2\Bigg\} + C\Bigg(\sum_{i=1}^{l}\xi_i + \sum_{i=1}^{l}\xi_i^*\Bigg) \\ &\text{subject to,} \\ &\quad A_i w + \eta(y_i + \bar{\epsilon}) + b \ge 1 - \xi_i \quad \text{for} \quad i = 1, 2, \ldots l, \\ &\quad -(A_i w + \eta(y_i - \bar{\epsilon}) + b) \ge 1 - \xi_i^* \quad \text{for } i = 1, 2, \ldots l, \\ &\quad \xi_i, \xi_i^* \ge 0 \quad \text{for} \quad i = 1, 2, \ldots, l. \end{aligned} \tag{26}$$

Here we note that $\eta \neq 0$ and therefore, without loss of generality, we can assume that $\eta > 0$. The constraint of (26) can then be rewritten as

Table 3 Results on commonly used benchmark datasets

Dataset	Regressor	SSE/SST	SSR/SST	RMSE	MAE	CPU time (in seconds)
Yatch hydro dynamics						
308×7	SVR	0.0230 ± 0.0080	0.8983 ± 0.0716	2.2212 ± 0.4703	1.5544 ± 0.2620	3.70
	LS-SVR	0.0168 ± 0.0059	0.9581 ± 0.0620	1.8875 ± 0.03784	1.4574 ± 0.2574	0.36
	L_1-Norm SVR	0.0166 ± 0.0062	0.9435 ± 0.0558	1.8664 ± 0.3253	1.4131 ± 0.2165	40.48
	LDMR	0.0133 ± 0.0114	0.9836 ± 0.0424	1.5889 ± 0.4195	1.2173 ± 0.2009	3.74
Concrete Slump						
103×8	SVR	0.0069 ± 0.0033	0.9979 ± 0.0739	0.5792 ± 0.1348	0.4526 ± 0.1162	0.84
	LS-SVR	0.0075 ± 0.0076	0.9816 ± 0.0615	0.5529 ± 0.2312	0.4518 ± 0.1699	0.30
	L_1-Norm SVR	0.0079 ± 0.0067	0.9937 ± 0.0488	0.5949 ± 0.2038	0.4764 ± 0.1789	5.00
	LDMR	0.0066 ± 0.0059	0.9957 ± 0.0520	0.5393 ± 0.1889	0.4373 ± 0.1685	0.85
Pyrims						
74×28	SVR	0.4053 ± 0.1501	0.6215 ± 0.0344	0.0705 ± 0.0346	0.0519 ± 0.0192	0.60
	LS-SVR	0.3601 ± 0.2193	0.7192 ± 0.2495	0.0670 ± 0.0484	0.0471 ± 0.0196	0.29
	L_1-Norm SVR	0.3576 ± 0.1732	0.8418 ± 0.4101	0.0649 ± 0.0307	0.0464 ± 0.0149	1.80
	LDMR	0.3305 ± 0.1437	0.7796 ± 0.3291	0.0648 ± 0.0391	0.0471 ± 0.0165	0.60
Motorcycle						
133×2	SVR	0.2247	0.8284	22.8209	16.6075	11.35
	LS-SVR	0.2316	0.7356	23.1667	17.4959	0.18
	L_1-Norm SVR	0.2182	0.8954	22.4860	16.2270	45.67
	LDMR	0.2192	0.8956	22.5393	16.1859	11.46
NO2						
500×8	SVR	0.4778 ± 0.1108	0.6837 ± 0.1674	0.5089 ± 0.0638	0.4041 ± 0.0496	4.67
	LS-SVR	0.4437 ± 0.1134	0.6020 ± 0.1694	0.4899 ± 0.0678	0.3901 ± 0.0493	0.26
	L_1-Norm SVR	0.4845 ± 0.0981	0.5708 ± 0.1632	0.5139 ± 0.0620	0.4023 ± 0.0432	172.97
	LDMR	0.4566 ± 0.1225	0.6256 ± 0.1936	0.4964 ± 0.0679	0.3930 ± 0.0530	6.28
Chwirut						
214×3	SVR	0.0215 ± 0.0106	0.9677 ± 0.0792	3.2882 ± 1.0788	2.3358 ± 0.6187	1.69
	LS-SVR	0.0214 ± 0.0115	0.9793 ± 0.0880	3.2774 ± 1.1400	2.3523 ± 0.6832	0.09
	L_1-Norm SVR	0.0213 ± 0.0106	0.9800 ± 0.0818	3.2701 ± 1.0973	2.3315 ± 0.6282	15.21
	LDMR	0.0213 ± 0.0104	0.9766 ± 0.0857	3.2759 ± 1.0175	2.3242 ± 0.6202	1.57
Auto MPG						
398×8	SVR	0.1142 ± 0.0637	0.8892 ± 0.0975	2.5231 ± 0.7177	1.8570 ± 0.3800	5.14
	LS-SVR	0.1154 ± 0.0586	0.9021 ± 0.0763	2.5380 ± 0.6683	1.8812 ± 0.3555	0.18
	L_1 -Norm SVR	0.1170 ± 0.0604	0.8868 ± 0.1009	2.5485 ± 0.6482	1.8780 ± 0.3746	109.57
	LDMR	0.1140 ± 0.0586	0.8902 ± 0.0989	2.5214 ± 0.6533	1.8733 ± 0.3554	4.91
Boston Housing						
506×14	SVR	0.2294 ± 0.0622	0.7574 ± 0.1012	3.3628 ± 0.6404	2.4022 ± 0.3012	23.74
	LS-SVR	0.2360 ± 0.0622	0.8031 ± 0.1138	3.3960 ± 0.5710	2.3865 ± 0.2505	0.24
	L_1 -Norm SVR	0.2430 ± 0.0634	0.7676 ± 0.1072	3.4519 ± 0.5645	2.4581 ± 0.2888	176.33
	LDMR	0.2348 ± 0.0577	0.8086 ± 0.1188	3.3982 ± 0.5747	2.4052 ± 0.2500	6.77

Table 4 Average ranks of SVR, LS-SVR, L_1-Norm SVR and LDMR models on SSE/SST values

Dataset	SVR	LS-SVR	L_1-Norm SVR	LDMR
Yacht hydrodynamics	4	3	2	1
Concrete Slump	2	3	4	1
Pyrims	4	3	2	1
Motorcycle	3	4	1	2
NO2	3	1	4	2
Chwirut	4	3	1.5	1.5
Auto MPG	2	3	4	1
Boston Housing	1	3	4	2
Average	2.8750	2.8750	2.8125	1.4375

$$
\begin{aligned}
&A_i\left(\frac{w}{-\eta}\right)-(y_i+\bar{\epsilon})+\left(\frac{b}{-\eta}\right)\le\frac{1}{-\eta}-\left(\frac{\xi_i}{-\eta}\right),\\
&\quad i=1,2,\ldots,l\\
&-A_i\left(\frac{w}{-\eta}\right)+(y_i-\bar{\epsilon})-\left(\frac{b}{-\eta}\right)\le\frac{1}{-\eta}-\left(\frac{\xi_i^*}{-\eta}\right),\\
&\quad i=1,2,\ldots,l\\
&\frac{-\xi_i}{\eta}\le 0,\frac{-\xi_i^*}{\eta}\le 0,\quad i=1,2,\ldots,l
\end{aligned}
$$

Further, the objective function of (26) can also be written as

$$
\begin{aligned}
\min_{w,\eta,b,\xi,\xi^*}\ &\eta^2\left[\frac{1}{2}\left(\left(\frac{w}{-\eta}\right)^{\mathrm{T}}\left(\frac{w}{-\eta}\right)+1\right)-\frac{\lambda_1}{\eta}\bar{\epsilon}\right.\\
&+\lambda_2\frac{1}{l}\left\{\sum_{i=1}^{l}\left(A_i\left(\frac{w}{-\eta}\right)-y_i+\left(\frac{b}{-\eta}\right)\right)\right.\\
&\left.\left.\left(A_i\left(\frac{w}{-\eta}\right)-y_i+\left(\frac{b}{-\eta}\right)\right)\right\}\right.\\
&\left.+\frac{C}{\eta}\left(\sum_{i=1}^{l}\frac{\xi_i}{\eta}+\sum_{i=1}^{l}\frac{\xi_i^*}{\eta}\right)\right]
\end{aligned}
$$

On replacing $\epsilon:=\bar{\epsilon}-\frac{1}{\eta}$, $w:=(\frac{w}{-\eta})$, $b:=(\frac{b}{-\eta})$, $\xi_i:=\frac{\xi_i}{\eta}$, $\xi_i^*:=\frac{\xi_i^*}{\eta}$, $C:=\frac{C}{\eta}$ and $\lambda:=\frac{\lambda}{\eta}$ and noting that $\eta\ge 0$, the objective function of the optimization problem (26) can be rewritten as

$$
\begin{aligned}
\min_{w,b,\xi,\xi^*}\ &\eta^2\left[\frac{1}{2}w^{\mathrm{T}}w-\lambda_1\epsilon+\lambda_2\frac{1}{l}\left\{\sum_{i=1}^{l}(y_i-(A_iw+b))^2\right\}\right.\\
&\left.+C\left(\sum_{i=1}^{l}\xi_i+\sum_{i=1}^{l}\xi_i^*\right)+\frac{1}{2}\right].
\end{aligned}
$$

Also the constraints of the optimization problem (26) can be expressed as

$$
\begin{aligned}
&(A_iw+b)-y_i\le\epsilon+\xi_i,\qquad \text{for } i=1,2,\ldots l,\\
&y_i-(A_iw+b)\le\epsilon+\xi_i^*,\qquad \text{for } i=1,2,\ldots l,\\
&\xi_i,\xi_i^*\ge 0\qquad \text{for } i=1,2,\ldots,l,\qquad \text{where}\qquad \epsilon\ge 0.
\end{aligned}
\tag{27}
$$

We further need to show that $\epsilon=\bar{\epsilon}-\frac{1}{\eta}$ is always non-negative. We prove this assertion as follows.

Let $(\bar{w},\bar{\eta},\bar{\xi}\ \bar{\xi^*})$ be the solution of QPP (26) which finds the classifier for the sets D^+ and D^-. There would always exists an index j such that

$$(A_jw+\eta(y_j+\bar{\epsilon})+b)\ge 1, \tag{28}$$

$$-(A_jw+\eta(y_j-\bar{\epsilon})+b)\ge 1. \tag{29}$$

Adding (28) and (29), we get $\bar{\epsilon}\ge\frac{1}{\eta}$, which proves that $\epsilon=\bar{\epsilon}-\frac{1}{\eta}\ge 0$.

Now for $\eta>0$, the classifier $w^{\mathrm{T}}x+\eta y+b=0$ for the classes D^+ and D^- gives the regressor $y=(w^{\mathrm{T}}x+b)$ with $w:=\left(\frac{w}{-\eta}\right)$, $b:=\left(\frac{b}{-\eta}\right)$. Since constraints in (27) do not involve η, η does not play any role now in the determination of the regressor. Therefore, problem (26) becomes

$$
\begin{aligned}
\min_{w,b,\xi,\xi^*}\ &\frac{1}{2}w^{\mathrm{T}}w-\lambda_1\epsilon+\lambda_2\frac{1}{l}\left\{\sum_{i=1}^{l}(y_i-(A_iw+b))^2\right\}\\
&+C\left(\sum_{i=1}^{l}\xi_i+\sum_{i=1}^{l}\xi_i^*\right)\\
\text{subject to,}\\
&(A_iw+b)-y_i\le\epsilon+\xi_i\qquad \text{for } i=1,2,\ldots l,\\
&y_i-(A_iw+b)\le\epsilon+\xi_i^*\qquad \text{for } i=1,2,\ldots l,\\
&\xi_i,\xi_i^*\ge 0\qquad \text{for } i=1,2,\ldots,l.
\end{aligned}
\tag{30}
$$

Also, since ϵ has been taken as constant, it can be removed from the objective function of (30). Further, after replacing $k:=2\lambda_2\frac{1}{l}$, the problem (30) can be written in the vector form as follows

Table 5 Tuned parameter values of SVR models on UCI datasets

Dataset	q	C	c	k	ϵ
Yatch					
SVR	1	1024			1
LS-SVR	1	1024			
L-1 Norm SVR	1	256	0.0313		1
LDMR	1	1024	0.0009	1	1
Concrete Slump					
SVR	1	1024			0.1
LS-SVR	1	1024			
L-1 Norm SVR	1	128	0.0009		0.1
LDMR	1	1024	0.0009	1	0.1
Pyrims					
SVR	4	4			0.05
LS-SVR	2	16			
L-1 -Norm SVR	4	0.5	0.0313		0.05
LDMR	4	8	0.0313	1	0.3
Motorcycle					
SVR	0.0078	32			0.1
LS-SVR	0.0078	4			
L-1 Norm SVR	0.0078	2	1		0.1
LDMR	0.0078	128	4	1	0.1
NO2					
SVR	0.25	8			0.3
LS-SVR	0.25	8			
L-1 -Norm SVR	2	128	1		0.1
LDMR	2	8	0.5	1	0.6
Chwirut					
SVR	0.0078	32			0.1
LS-SVR	0.0078	16			
L-1 -Norm SVR	0.0078	32	2		0.1
LDMR	0.0078	32	2	1	0.1
Auto MPG					
SVR	0.5	128			1
LS-SVR	0.5	32			
L_1-Norm SVR	1	8	0.0313		0.8
LDMR	0.5	64	0.0313	1	1
Boston Housing					
SVR	2	128			2
LS-SVR	2	16			
L-1 -Norm SVR	2	16	1		2
LDMR	2	16	0.0078	1	1.5

$$\min_{w,b,\xi,\xi^*} \frac{c}{2}||w||^2 + \frac{k}{2}||Y-(Aw+eb)||_2 + Ce^{\mathrm{T}}(\xi+\xi^*)$$
$$\text{subject to,}$$
$$Y-(Aw+eb) \le e\epsilon + \xi,$$
$$(Aw+eb)-Y \le e\epsilon + \xi^*,$$
$$\xi, \xi^* \ge 0. \tag{31}$$

6 Experimental results

We have performed a number of experiments to verify the efficacy of the our proposed LDMR model. For this, we have compared the LDMR model with existing SVR models, namely standard SVR, SMO-based SVR, LS-SVR and L_1-Norm SVR (Tanveer [24]) on certain artificial datasets, UCI benchmark datasets (Blake and Merz [28]) and time-series financial datasets.

All the simulations have been performed in MATLAB 12.0 environment (http://in.mathworks.com/) on Intel XEON processor with 16.0 GB RAM. The L_1-Norm SVR model has been solved by using the 'linprog' function of the MATLAB. For small- and medium-scale datasets, the proposed LDMR and standard SVR models have been solved by using the 'quadprog' function of MATLAB. For the large-scale datasets, we have used the SOR technique (Mangasarian and Musicant [25]) to solve the proposed LDMR model efficiently, whereas the SMO (Chang and Lin [27]) method has been used for obtaining the solution of standard SVR for large-scale datasets. For this, we have downloaded its code form (https://www.csie.ntu.edu.tw/~cjlin/libsvm/). Throughout these experiments, we have used RBF kernel $\exp(\frac{-||x-y||^2}{q})$ where q is the kernel parameter.

The optimal values of parameters in SVR models have been obtained using the exhaustive search method (Hsu and Lin [30]) with cross-validation. For all SVR models, the value of the kernel parameter q has been searched in the set $\{2^i, i=-10,-2,\ldots,10\}$. The value of the parameter ϵ in ϵ-SVR, L_1-Norm SVR and proposed LDMR model has been searched in the set of $\{0.05, 0.1, 0.2, 0.3\ldots, 1, 1.5, 2\}$. The value of the parameter k of the proposed LDMR has been fixed to 1 throughout the experiments. The value of parameters C in ϵ-SVR, LS-SVR, L_1-Norm SVR and proposed LDMR model has been searched in the set $\{2^i, i=-10,-2,\ldots,12\}$. The value of parameter c in L_1-Norm SVR and proposed LDMR model has also been searched in the set $\{2^i, i=-10,-2,\ldots,12\}$.

Table 6 Results on large-scale datasets

Dataset	Regressor	SSE/SST	SSR/SST	RMSE	MAE	CPU time(s)
Parkinsons Telemonitoring	SMO SVR	0.2289 ± 0.0070	0.7593 ± 0.0191	0.0446 ± 0.0011	0.0353 ± 0.0010	0.78
$(2000 + 3847 \times 22)$	SOR LDMR	0.1978 ± 0.0014	0.7862 ± 0.0032	0.0408 ± 0.0009	0.0285 ± 0.0002	22.03
Wine Quality Red	SMO SVR	0.6371 ± 0.0071	0.4245 ± 0.0149	0.6433 ± 0.0162	0.4928 ± 0.0104	0.75
$(1000 + 599 \times 22)$	SOR LDMR	0.6139 ± 0.0027	0.4371 ± 0.0080	0.6339 ± 0.0149	0.4903 ± 0.0099	5.94
Wine Quality White	SMO SVR	0.6792 ± 0.0026	0.3737 ± 0.0078	0.7321 ± 0.0064	0.5690 ± 0.0048	1.97
$(1000 + 3898 \times 22)$	SOR LDMR	0.6736 ± 0.0036	0.3605 ± 0.0106	0.7267 ± 0.0043	0.5648 ± 0.0046	7.01

Table 7 Results on financial dataset

Dataset	Regressor	SSE/SST	SSR/SST	RMSE	MAE	CPU time (in seconds)
IBM						
(244×4)	SVR	0.1009	0.9825	0.1760	0.1226	0.58
	LS-SVR	0.0995	0.9873	0.1748	0.1220	0.37
	L_1-Norm SVR	0.1025	1.0141	0.1774	0.1242	1.12
	LDMR	0.0989	0.9959	0.1743	0.1210	0.59
	SOR LDMR	0.0989	0.9958	0.1742	0.1210	0.40
SBI						
(513×4)	SVR	0.0202	0.9520	0.0742	0.0564	0.69
	LS-SVR	0.0197	0.9605	0.0733	0.0552	0.36
	L_1-Norm SVR	0.0204	0.9436	0.0746	0.0572	3.71
	LDMR	0.0193	0.9622	0.0725	0.0545	0.68
	SOR LDMR	0.0193	0.9620	0.0725	0.0546	0.38

6.1 Performance criteria

In order to evaluate the performance of the regression methods, we first summarize some commonly used evaluation criteria. Without the loss of generality, let l and k be the number of the training samples and testing samples, respectively. Furthermore, for $i = 1, 2, \ldots k$, let y_i' be the predicted value for the response value y_i and $\bar{y} = \frac{1}{k}\sum_i^k y_i$ is the average of $y_1, y_2, \ldots, y_k$. The definition and significance of the some evaluation criteria have been listed as follows.

(i) *SSE* Sum of squared error of testing, which is defined as $\text{SSE} = \sum_{i=1}^{k}(y_i - y_i')^2$. SSE represents the fitting precision.

(ii) *SST* Sum of squared deviation of testing samples, which is defined as $\text{SST} = \sum_{i=1}^{k}(y_i - \bar{y})^2$. SST shows the underlying variance of the testing samples.

(iii) *SSR* Sum of square deviation of the testing samples which can be explained by the estimated regressor. It is defined as $\text{SSR} = \sum_{i=1}^{k}(y_i' - \bar{y})^2$.

(iv) *RMSE* Root mean square of the testing error, which is defined as $\text{RMSE} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_i - y_i')^2}$.

(v) *SSE/SST* SSE/SST is the ratio between the sum of the square of the testing error and sum of the square of the deviation of testing samples. In most cases, small SSE/SST means good agreement between estimations and real values.

(vi) *SSR/SST* It is the ratio between the variance obtained by the estimated regressor on testing samples and actual underlying variance of the testing samples.

6.2 Experiment 1 (artificial datasets)

To compare the performance of the proposed methods with the existing methods, we have synthesized some artificial datasets. For the training samples (x_i, y_i) for $i = 1, 2, \ldots, l$, datasets have been generated as follows.

TYPE 1

$$y_i = \frac{\sin(x_i)}{x_i} + \lambda_i, \lambda_i \sim U[-0.2, 0.2]$$

and x_i is from $U[-4\pi, 4\pi]$.

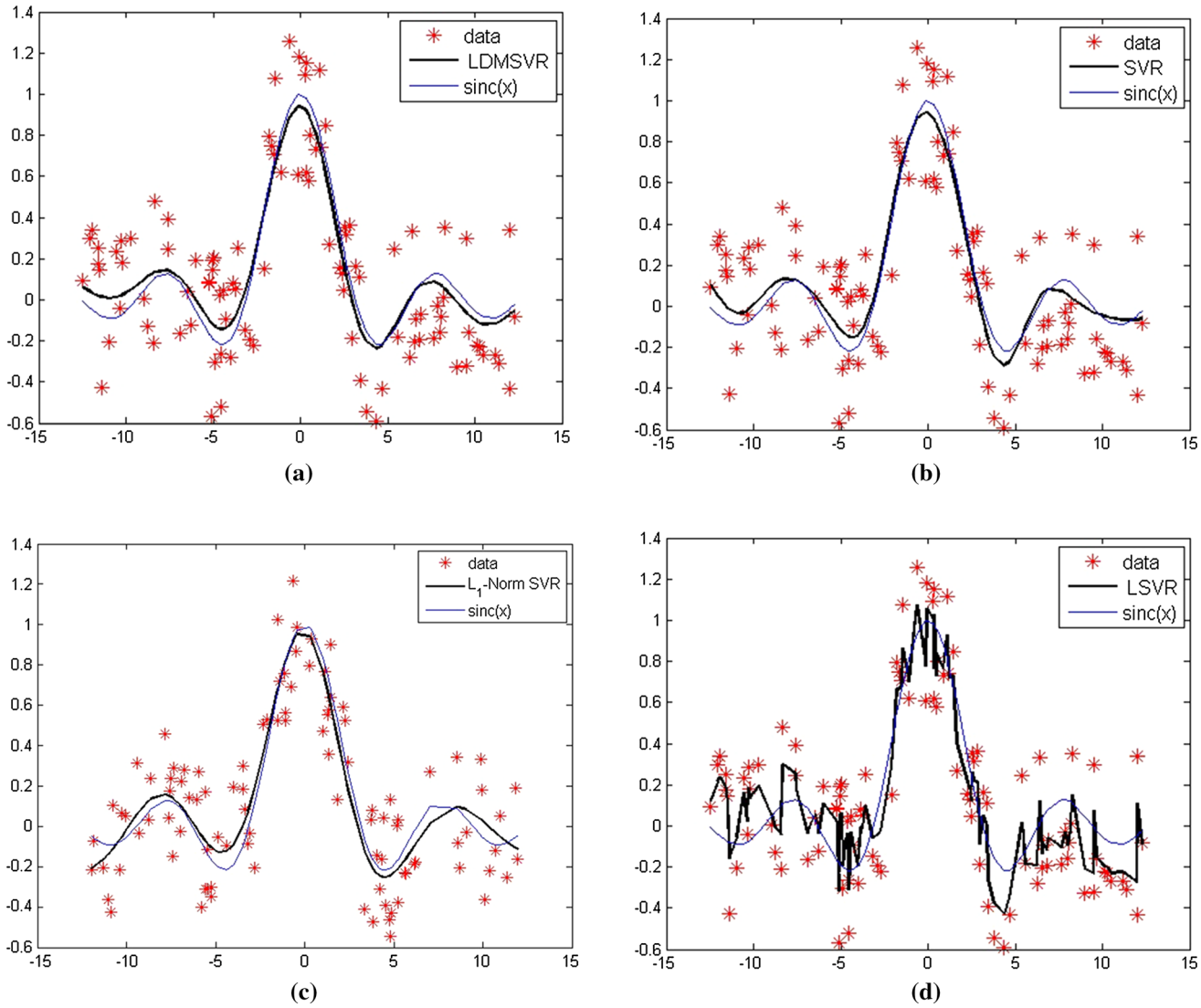

Fig. 1 Performance of **a** LDMR **b** SVR **c** L_1-Norm SVR and **d** LS-SVR on TYPE 3 dataset

TYPE 2

$$y_i = \frac{\sin(x_i)}{x_i} + \lambda_i, \lambda_i \sim U[-0.3, 0.3]$$

and x_i is from $U[-4\pi, 4\pi]$.

TYPE 3

$$y_i = \frac{\sin(x_i)}{x_i} + \lambda_i, \lambda_i \sim U[-0.4, 0.4]$$

and x_i is from $U[-4\pi, 4\pi]$.

TYPE 4

$$y_i = \frac{\sin(x_i)}{x_i} + \lambda_i, \lambda_i \sim N[0, 0.2]$$

and x_i is from $U[-4\pi, 4\pi]$.

TYPE 5

$$y_i = \frac{\sin(x_i)}{x_i} + \lambda_i, \lambda_i \sim N[0, 0.3]$$

and x_i is from $U[-4\pi, 4\pi]$.

TYPE 6

$$y_i = \left|\frac{x_i - 1}{4}\right| + \left|\sin(\pi(1 + \frac{x_i - 1}{4}))\right| + 1 + \lambda_i,$$

$\lambda_i \sim U[-0.5, 0.5]$ and x_i is from $U[-10, 10]$.

TYPE 6 dataset contains 200 training samples and 400 non-noise testing samples, while other datasets contain 100 training samples and 500 non-noise testing samples. To avoid the biased comparison, ten independent groups of

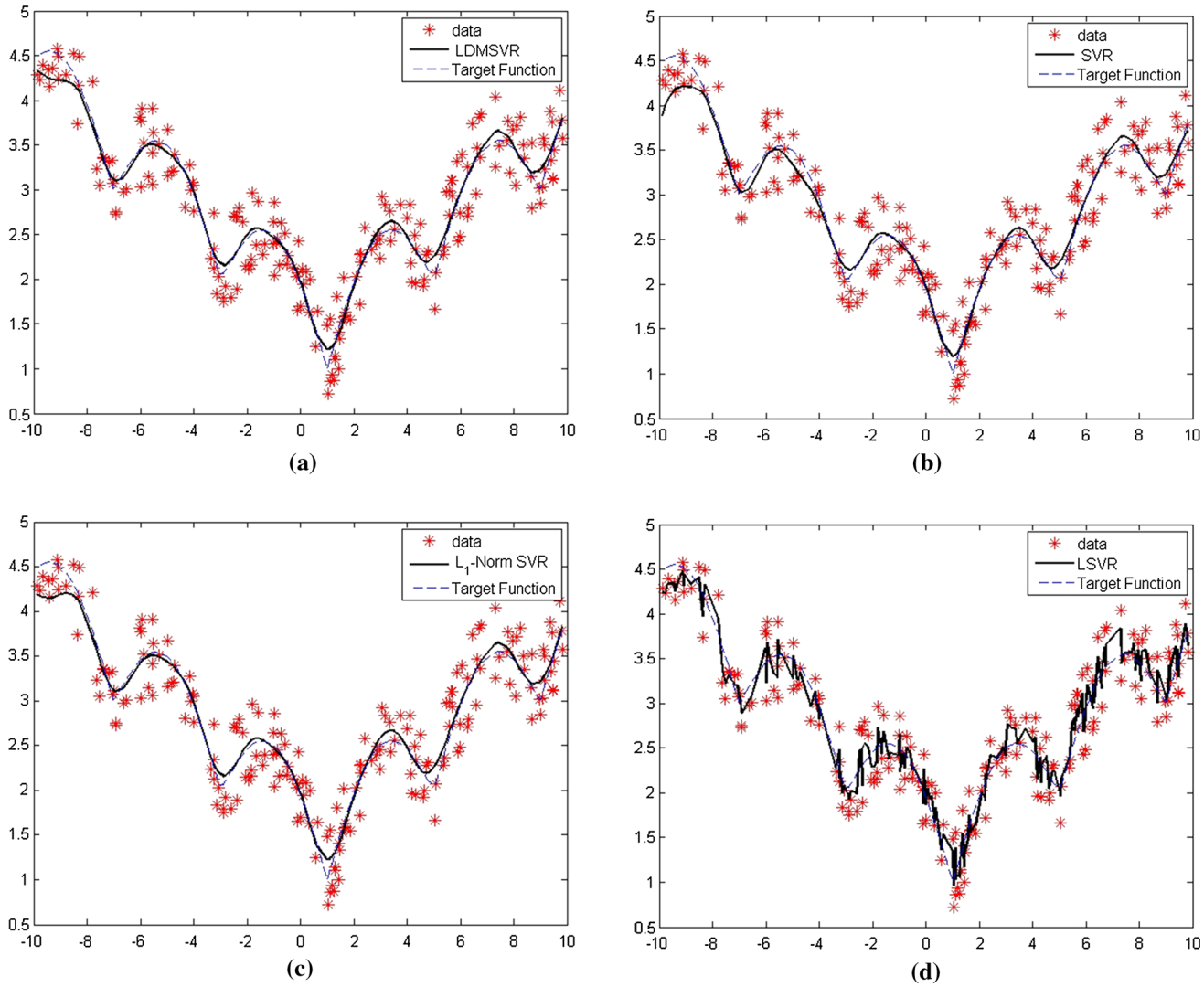

Fig. 2 Performance of **a** LDMR **b**SVR **c** L_1-Norm SVR and **d** LS-SVR on TYPE 6 dataset

noisy samples were generated randomly using MATLAB toolbox for all types of training sets.

Figures 1 and 2 illustrate the estimated function obtained by the LDMR, standard SVR, L_1-Norm SVR and LS-SVR models for TYPE 3 and TYPE 6 datasets, respectively. Table 1 shows the comparison of the proposed LDMR with SVR, L_1-Norm SVR and LS-SVR models on artificial datasets. It can be observed that the proposed LDMR, irrespective of the nature of noises present in the training set, owns always better generalization ability than other regression methods.

6.3 Experiment 2 (artificial datasets with outliers)

As compared to other regression methods, the proposed LDMR is a robust method and is less sensitive to the presence of outliers. To realize this, we have generated datasets by adding five and ten outliers points in the TYPE 3 and TYPE 6 datasets, respectively. Figure 3 shows the estimated function obtained by LDMR, standard SVR, L_1-Norm SVR and LS-SVR models on TYPE 3 dataset with outliers. For L_1-Norm SVR and standard SVR models, the optimal RMSE has been found at $\epsilon = 0.3$ but, still at this value of the ϵ, outliers lie outside of the ϵ-insensitive zone and affect the orientation and position of the estimated function. The LDMR model reduces the effect of these outliers by also assigning some weight to the points lying inside of the ϵ-insensitive zone in the optimization problem. Table 2 shows the comparison of the proposed LDMR, standard SVR, L_1-Norm SVR and LS-SVR models on artificial datasets with outliers. In these datasets, 200 points were used for training and 400 non-noise points were used for testing.

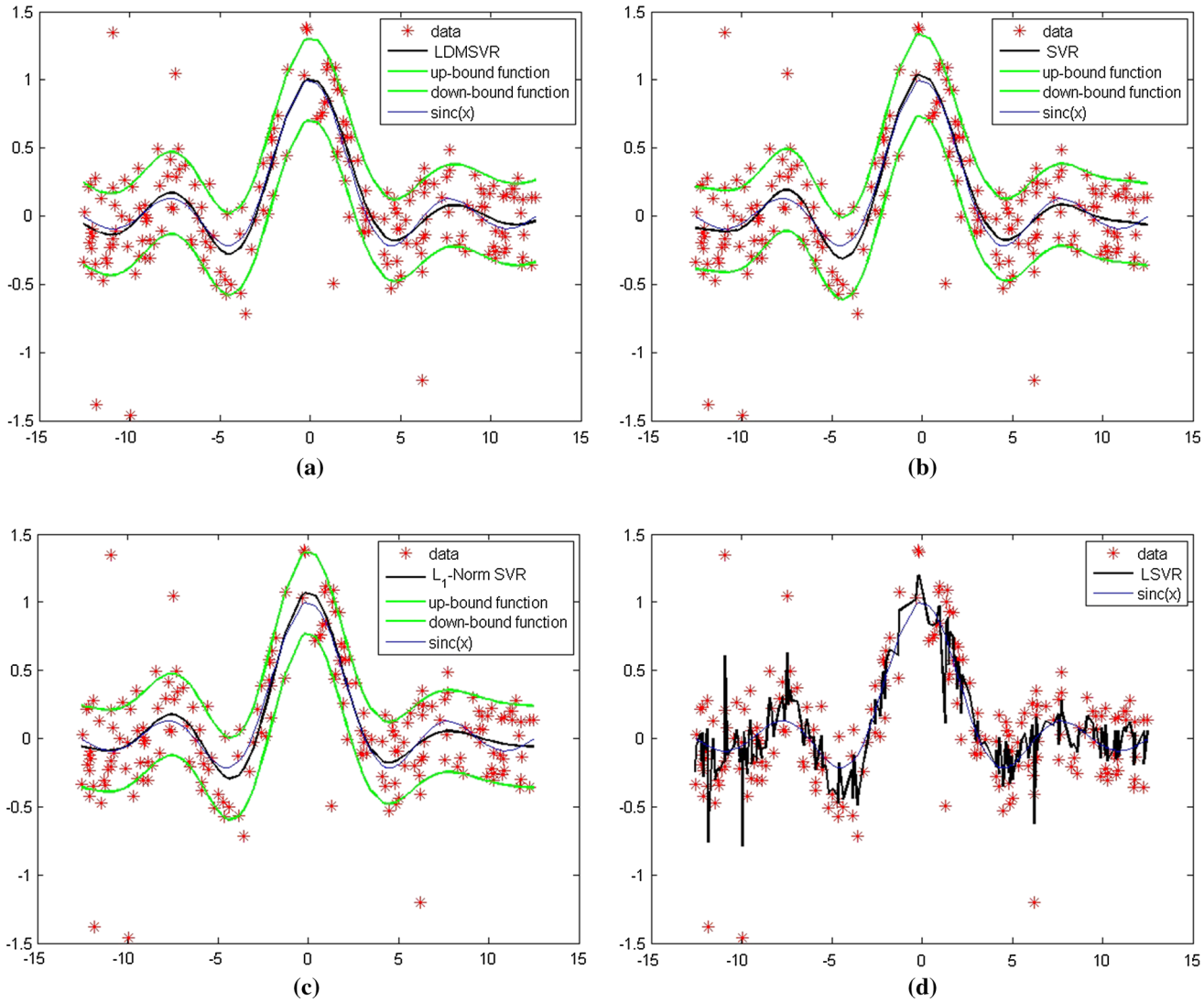

Fig. 3 Performance of **a** LDMR **b** SVR **c** L_1-Norm SVR and **d** LS-SVR on TYPE 3 dataset with outliers

6.4 Experiment 3 (benchmark datasets)

We have checked the performance of the proposed methods on eight UCI benchmark datasets, namely Yacht Hydrodynamics, Concrete Slump, Pyrims, Motorcycle, NO2, Chwirut, Auto MPG and Boston Housing which are commonly used in evaluating a regression method. For the Motorcycle dataset, the criterion leave-one out (Kohavi [32]) was used to report the numerical results. For other datasets, we have used tenfold cross-validation (Duda and Hart [31]) to report the numerical results. For all the datasets, only feature vectors were normalized in the range of [0,1].

Table 3 lists the performance of LDMR, standard SVR, LS-SVR and L_1-Norm SVR using different evaluation criteria described in Sect. 6.1 for eight different UCI datasets. It can be observed that the proposed LDMR model outperforms the other existing regression methods in most of the datasets. For the statistical analysis of performance of the regression methods, their average rank has been computed using the values of SSE/SST. The obtained average ranks are summarized in Table 4. It can be observed that on an average, the proposed LDMR models obtain better ranks than existing regression methods.

For all SVR models, we have tuned their parameters to obtain their best choice in each datasets. We list the tuned parameter values of SVR models for UCI datasets in Table 5. Figure 6 shows the effect of the parameters C and c on SSE/SST values in LDMR model for Pyrims and Auto MPG datasets. It shows that the performance of the LDMR model is sensitive to the choice of parameters c and C.

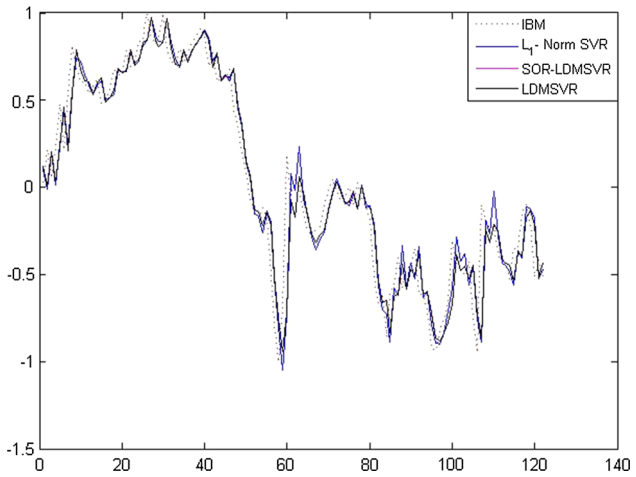

Fig. 4 SSE/SST values obtained by LDMR model on different values of C and c parameter on **a** Pyrims and **b** Auto MPG dataset

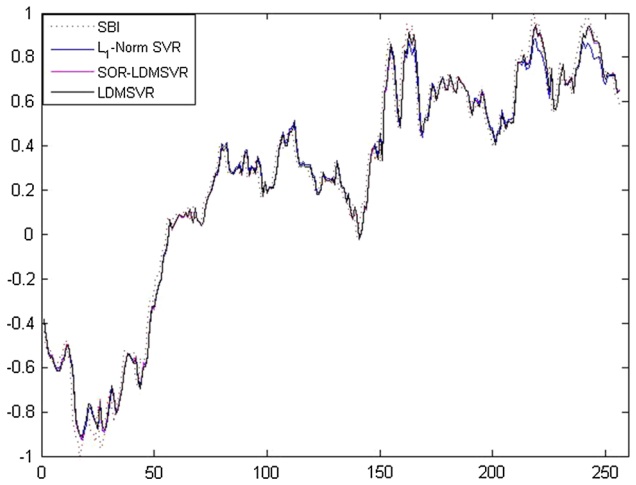

Fig. 5 Performance of L_1-Norm SVR, LDMR and SOR LDMR on IBM dataset

6.5 Experiment 4 (large-scale datasets)

We have also compared the proposed model with SVR model on large–scale datasets. Since the 'quadprog' implementation of the proposed LDMR model and SVR model is not efficient for the large-scale datasets, we have used the SOR method in the proposed LDMR model for these datasets. For the SVR model, we have used its SMO implementation. We have downloaded Parkinsons Telemonitoring (5847 $\times$ 22), Wine Quality Red (1599 $\times$ 12) and Wine Quality White (4898 $\times$ 12) datasets from the UCI Repository [28]. For all these datasets, we have normalized their input feature vectors in the range of [0,1]. For each dataset, we have fixed the number of training points and testing points and randomly permuted the data points in training set and testing set in ten different trials. The regression methods have been evaluated for each trial.

Table 6 shows the comparison of the SMO SVR and proposed SOR LDMR on large-scale datasets using different evaluation criteria. It can be observed that the proposed LDMR model with its SOR implementation outperforms the existing SVR model with SMO implementation.

6.6 Experiment 5 (time-series financial dataset)

For further evaluation of the proposed methods, we have simulated the existing as well as proposed regression methods on real-world time-series financial datasets. We have downloaded the daily stock prices of the IBM and SBI.INS for the period of 27 March 2016 to 27 March 2017 and 27 March 2015 to 27 March 2017, respectively, from the Yahoo financial website. The datasets were generated by taking the last three days closing stock prices as input feature and next day closing stock price as response value. In experiments, the first 50% data points of the datasets were used for training and rest of them have been used for the testing. The input features were normalized in the range of $[-1, 1]$.

Table 7 lists the comparison of the proposed methods with other existing regression methods using different evaluation criteria on IBM and SBI.INS finance datasets along with their training time. Figures 4 and 5 show the performance of the proposed regression methods along with existing regression methods on IBM and SBI.INS finance dataset, respectively. It can be easily observed in Table 5 and Figures 5 and 6 that the proposed LDMR formulations outperform the existing regression methods.

7 Conclusion

An efficient LDM-based regression model has been proposed in this paper which can be mathematically derived from the optimization problem of the LDM (Zhang and Zhou [17]) by making use of a result of Bi and Bennett [20]. The proposed LDMR model simultaneously minimizes the ϵ-insensitive loss function as well as the quadratic loss function. In this sense, the proposed LDMR formulation combines the benefits of the LS-SVR model (Suykens et al. [21, 22]) as well as the ϵ-SVR model. The proposed LDMR model obtains better generalization ability by finding a trade-off between the ϵ-insensitive loss and the quadratic loss via the user-defined parameters k and

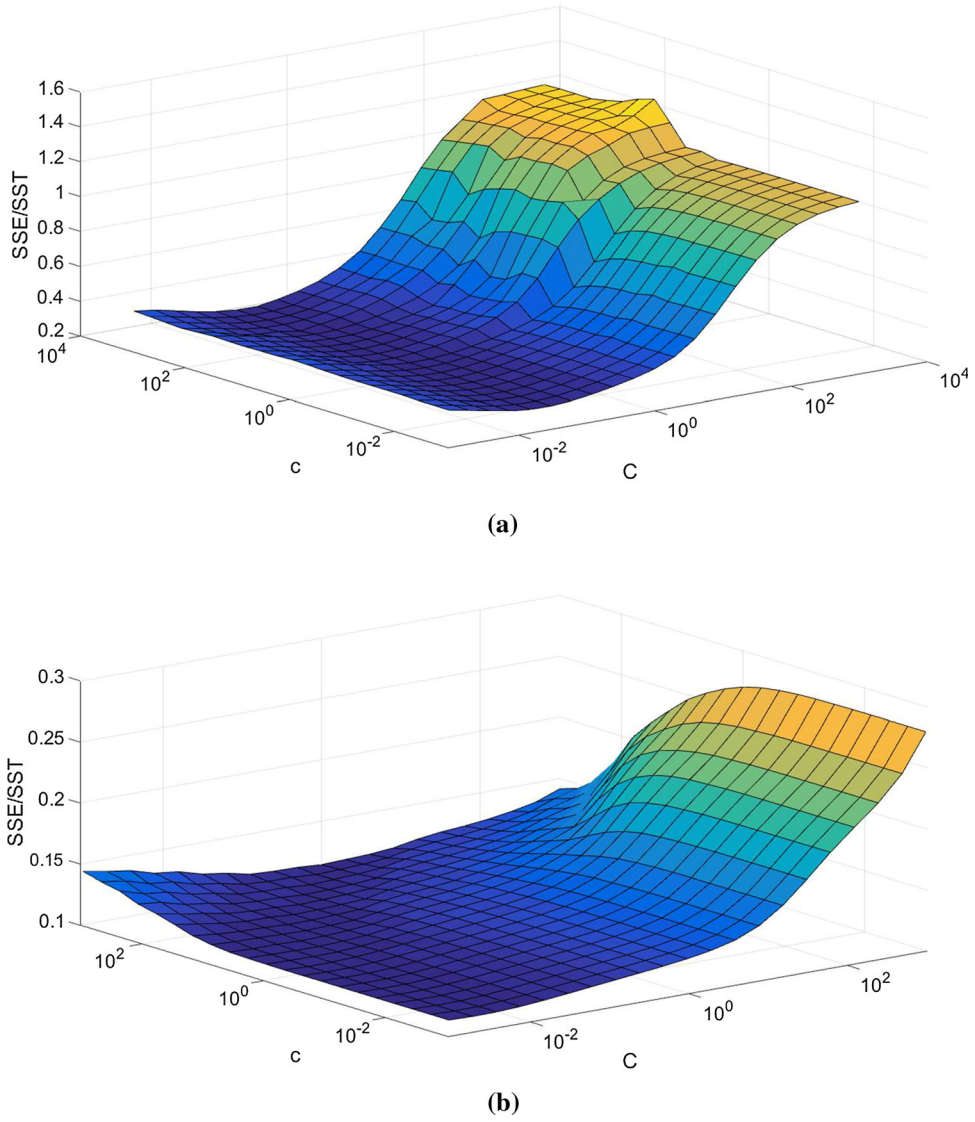

Fig. 6 Performance of L_1-Norm SVR, LDMR and SOR LDMR on SBI dataset

C. The proposed LDMR model has also been observed to be less sensitive to the presence of outliers. Further, the application of the SOR technique (Mangaserian and Musicant [25]) significantly reduces the training time of the proposed LDMR model.

One of the major difficulties with the proposed LDMR model is that it requires more parameters to be tuned. This requires future study so as to develop a model which can autolearn parameters ϵ, c and C from the data. Further, we have observed that the SOR technique works well in the proposed LDMR formulation but, as compared to SMO method in SVR model, it takes more time to train the model. So we would like to test the SMO method for solving the optimization problem in the proposed LDMR model in future.

Acknowledgements We would like to thank the learned referees for their valuable comments and suggestions which has substantially improved the contents and presentation of the manuscript. We would also like to acknowledge Ministry of Electronics and Information Technology, Government of India, as this work has been funded by them under Visvesvaraya Ph.D. Scheme for Electronics and IT, Order No. Phd-MLA/4(42)/2015-16.

Compliance with ethical standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Cortes C, Vapnik V (1995) Support vector networks. Mach Learn 20(3):273–297
2. Burges JC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2(2):121–167
3. Cherkassky V, Mulier F (2007) Learning from data: concepts, theory and methods. Wiley, New York

4. Vapnik V (1998) Statistical learning theory, vol 1. Wiley, New York
5. Osuna E, Freund R, Girosit F (1997) Training support vector machines: an application to face detection. In: Proceedings of IEEE computer vision and pattern recognition, San Juan, Puerto Rico, pp 130–136
6. Joachims T (1998) Text categorization with support vector machines: learning with many relevant features, European conference on machine learning. Springer, Berlin
7. Schlkopf B, Tsuda K, Vert JP (2004) Kernel methods in computational biology. MIT Press, Cambridge
8. Lal TN, Schroder M, Hinterberger T, Weston J, Bogdan M, Birbaumer N, Scholkopf B (2004) Support vector channel selection in BCI. IEEE Trans Biomed Eng 51(6):10031010
9. Bradley P, Mangasarian OL (2000) Massive data discrimination via linear support vector machines. Optim Methods Softw 13(1):1–10
10. Freund Y, Schapire RE (1995) A decision-theoretic generalization of on-line learning and an application to boosting. In: Proceedings of the 2nd European conference on computational learning theory, Barcelona, Spain, p. 2337
11. Zhou ZH (2012) Ensemble methods: foundations and algorithms. CRC Press, Boca Raton
12. Breiman L (1999) Prediction games and arcing classifiers. Neural Comput 11(7):14931517
13. Schapire RE, Freund Y, Bartlett PL, Lee WS (1998) Boosting the margin: a new explanation for the effectives of voting methods. Annu Stat 26(5):16511686
14. Reyzin L, Schapire RE (2006) How boosting the margin can also boost classifier complexity. In: Proceedings of 23rd international conference on machine learning, Pittsburgh, PA, p 753–760
15. Wang L, Sugiyama M, Yang C, Zhou ZH, Feng J (2008) On the margin explanation of boosting algorithm. In: Proceedings of the 21st annual conference on learning theory, Helsinki, Finland, p 479–490
16. Gao W, Zhou ZH (2013) On the doubt about margin explanation of boosting. Artif Intell 199–200:2244
17. Zhang T, Zhou ZH (2014) Large margin distribution machine. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM
18. Vapnik V, Golowich SE, Smola AJ (1997) Support vector method for function approximation, regression estimation and signal processing. In: Mozer M, Jordan M, Petsche T (eds) Advances in neural information processing systems. MIT Press, Cambridge, pp 281–287
19. Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V (1997) Support vector regression machines. In: Mozer MC, Jordan MI, Petsche T (eds) Advances in neural information processing systems. MIT Press, Cambridge, pp 155–161
20. Bi J, Bennett KP (2003) A geometric approach to support vector regression. Neurocomputing 55(1):79–108
21. Suykens JAK, Lukas L, van Dooren P, De Moor B, Vandewalle J (1999) Least squares support vector machine classifiers: a large scale algorithm. In: Proceedings of European conference of circuit theory design, pp 839–842
22. Suykens JAK, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293300
23. Shao YH, Zhang C, Yang Z, Deng N (2013) An ϵ-twin support vector machine for regression. Neural Comput Appl 23(1):175–185
24. Tanveer M, Mangal M, Ahmad I, Shao YH (2016) One norm linear programming support vector regression. Neurocomputing 173:1508–1518
25. Mangasarian OL, Musicant DR (1999) Successive overrelaxation for support vector machines. IEEE Trans Neural Netw 10(5):1032–1037
26. Luo ZQ, Tseng P (1993) Error bounds and convergence analysis of feasible descent methods: a general approach. Ann Oper Res 46(1):157–178
27. Chang CC, Lin CJ (2011) LIBSVM, a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):27
28. Blake CI, Merz CJ (1998) UCI repository for machine learning databases [http://www.ics.uci.edu/*mlearn/MLRepository.html]
29. Huang X, Shi L, Suykens JA (2014) Support vector machine classifier with pinball loss. IEEE Trans Pattern Anal Mach Intell 36(5):984–997
30. Hsu CW, Lin CJ (2002) A comparison of methods for multi class support vector machines. IEEE Trans Neural Netw 13:415–425
31. Duda RO, Hart PR, Stork DG (2001) Pattern classification, 2nd edn. Wiley, Hoboken
32. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai, Vol. 14, No. 2