Optimally Selected Minimal Learning Machine

Conference Paper · November 2018

DOI: 10.1007/978-3-030-03493-1_70

4 authors:

Atilla Maia
Instituto Federal de Educação, Ciência e Tecnologia do Ceará

Madson Luiz Dantas Dias
Universidade Federal do Ceará

Joao Paulo Pordeus Gomes
Universidade Federal do Ceará

Ajalmar Rêgo da Rocha Neto
Instituto Federal de Educação, Ciência e Tecnologia do Ceará

Optimally Selected Minimal Learning Machine

Átilla N. Maia[1](✉), Madson L. D. Dias[2](✉), João P. P. Gomes[2], and Ajalmar R. da Rocha Neto[1]

[1] Graduate Program in Computer Science, Federal Institute of Ceará (IFCE), Fortaleza, Ceará, Brazil
atilla.negreiros@ppgcc.ifce.edu.br, ajalmar@ifce.edu.br

[2] Department of Computer Science, Federal University of Ceará (UFC), Fortaleza, Ceará, Brazil
{madson.dias,jpaulo}@lia.ufc.br

Abstract. This paper introduces a new approach to select reference points (RPs) to minimal learning machine (MLM) for classification tasks. A critical issue related to the training process in MLM is the selection of RPs, from which the distances are taken. In its original formulation, the MLM selects the RPs randomly from the data. We propose a new method called optimally selected minimal learning machine (OS-MLM) to select the RPs. Our proposal relies on the multiresponse sparse regression (MRSR) ranking method, which is used to sort the patterns in terms of relevance. After doing so, the leave-one-out (LOO) criterion is also used in order to select an appropriate number of reference points. Based on the simulations we carried out, one can see our proposal achieved a lower number of reference points with an equivalent, or even superior, accuracy with respect to the original MLM and its variants.

Keywords: Minimal Learning Machine · Reference points
Multiresponse sparse regression · Sparse models

1 Introduction

Minimal Learning Machine (MLM, [20]) is a recent supervised learning algorithm that can be used to handle classification and regression problems. This method has been employed to a diverse range of problems, such as fault detection in motors [4], ranking of documents [2] and location of mobile robots [13].

The basic idea behind MLM is the assumption about the existence of a mapping between the geometric configurations of points in the input space and the geometric configurations of respective points in the output space. This mapping is represented by two distance matrices (input and output), computed for all points in the training data set and a subset of it, named reference points. The learning of MLM is accomplished by determination of a regression linear model

Supported by Federal Institute of Ceará and Federal University of Ceará.

H. Yin et al. (Eds.): IDEAL 2018, LNCS 11314, pp. 670–678, 2018.
https://doi.org/10.1007/978-3-030-03493-1_70

between two distance matrices. Thus, given a point in the input space, the MLM can compute the location of this point in the output space through the learned regression model [5]. In other words, the MLM training algorithm consists in obtaining a multi-response linear system solution.

Several approaches have been applied to achieve sparse solutions in machine learning methods, such as greedy algorithms in orthogonal matching pursuit (OMP, [17]) for extreme learning machines (OMP-ELM, [1]) and multiresponse sparse regression (MRSR, [19]) for the least squares support vector machines (LSSVMs, [21]). Moreover, bayesian methods as sparse bayesian learning (SBL, [12]) have also been applied to single-hidden layer feedforward neural networks [11] and relevance vector machines (RVM, [22]).

In order to achieve sparsity for MLM, we propose a new method called optimally selected minimal learning machines (OS-MLM), which relies on a ranking method named multiresponse sparse regression (MRSR), used to sort the patterns in terms of relevance, and, after doing so, on the leave-one-out (LOO) criterion, applied to select an appropriate number of reference points. Our proposal is somewhat similar to a method called optimally pruned extreme learning machines (OP-ELM, [16]), which is used to prune neurons in hidden layers of extreme learning machines.

The remaining part of the paper is organized as follows. In Sect. 2 we describe the regular MLM. Section 3 briefly presents some strategies to select RPs. In Sect. 4, we present a multi-response sparse regression method, while our proposal is detailed in Sect. 5. In Section 6, we present the simulations carried out. Finally, the conclusions are outlined in Sect. 7.

2 Minimal Learning Machine

The Minimal Learning Machine is a supervised method whose training step consists of fitting a multiresponse linear regression model between distances computed from the input and output spaces. Output prediction for new inputs is achieved by estimating distances in output spaces using the underlying linear model followed by a optimization procedure in the space of possible outputs.

Let a training dataset $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$, a set $\mathcal{R} = \{(\boldsymbol{r}_m, \boldsymbol{t}_m)\}_{m=1}^{M} \subseteq \mathcal{D}$ of reference points, such that $\boldsymbol{x}_n, \boldsymbol{r}_m \in \mathbb{R}^D$ and $\boldsymbol{y}_n, \boldsymbol{t}_m \in \mathbb{R}^S$. Furthermore, let $\mathbf{D}, \boldsymbol{\Delta} \in \mathbb{R}^{N \times M}$ are distance matrices such that their m-th columns are respectively $[\|\boldsymbol{x}_1 - \boldsymbol{r}_m\|_2, \cdots, \|\boldsymbol{x}_N - \boldsymbol{r}_m\|_2]^T$ and $[\|\boldsymbol{y}_1 - \boldsymbol{t}_m\|_2, \cdots, \|\boldsymbol{y}_N - \boldsymbol{t}_m\|_2]^T$. The key idea behind MLM is the assumption of a linear mapping between $\mathbf{D}$ and $\boldsymbol{\Delta}$, giving rise to the following regression model:

$$\boldsymbol{\Delta} = \mathbf{D}\mathbf{B} + \mathbf{E} \tag{1}$$

where $\mathbf{B} \in \mathbb{R}^{M \times M}$ is the matrix of regression coefficients and $\mathbf{E} \in \mathbb{R}^{N \times K}$ is a matrix of residuals. Under the normal conditions where the number of selected reference points is smaller than the number of training points (i.e., $M < N$), the matrix $\mathbf{B}$ can be approximated by the usual least squares estimate

$$\hat{\mathbf{B}} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\boldsymbol{\Delta}. \tag{2}$$

Given a new input point $\boldsymbol{x}$, the vector $\hat{\boldsymbol{\delta}} = [\hat{\delta}_1, \cdots, \hat{\delta}_M]$ of the distances between the output $\boldsymbol{y}$ of point $\boldsymbol{x}$ and the M output reference points, is given by

$$\hat{\boldsymbol{\delta}} = [\|\boldsymbol{x} - \boldsymbol{r}_1\|_2, \cdots, \|\boldsymbol{x} - \boldsymbol{r}_M\|_2]\, \hat{\mathbf{B}}. \tag{3}$$

Therefore, an estimate $\hat{\boldsymbol{y}}$ of $\boldsymbol{y}$ can be obtained by the following minimization problem:

$$\hat{\boldsymbol{y}} = \arg\min_{\boldsymbol{y}} \left\{ \sum_{m=1}^{M} \left((\boldsymbol{y} - \boldsymbol{r}_m)^T (\boldsymbol{y} - \boldsymbol{r}_m) - \hat{\delta}_m^2 \right)^2 \right\}, \tag{4}$$

which can be approached via any gradient-based optimization algorithm. In the original paper, the regular MLM applies the Levenberg-Marquardt method [14].

For the classification case, where outputs $\boldsymbol{y}_n$ are represented using the 1-of-S encoding scheme[1]. It was showed in [15] that under the assumption that the classes are balanced, the optimal solution to Eq. (4) is given by $\hat{\boldsymbol{y}} = \boldsymbol{t}_{m^*}$, where $m^* = \arg\min_m \hat{\boldsymbol{\delta}}$. It means that output predictions for new incoming data can be carried out by simply selecting the output of the nearest reference point in the output space, estimated using the linear model $\hat{\mathbf{B}}$. This method was named Nearest Neighbor MLM (NN-MLM).

3 RPs Selection Methods

In the original MLM proposal, the choice of reference points is random, leaving just the number of points by user's taste [5]. In this paper, the MLM with RP random selection will be called random MLM (RN-MLM). A particular case of this approach is the use of all points in the dataset as RPs. In this case, this approach will be called full MLM (FL-MLM). Another approach to select reference points is the Fuzzy C-means MLM (FCM-MLM) [8], which uses the fuzzy C-means algorithm as main method. In this case, the K parameter represents the maximum number of RPs (i.e. $|\mathcal{R}| \leq K$). This is possible by removing RPs in heterogeneous regions.

4 Multiresponse Sparse Regression

The MRSR is an extension of least angle regression (LARS, [7]) that enables the most accurate prediction averaged over all the target variables rather than only one as LARS does. LARS is a less greedy version of traditional forward selection methods. Forward stagewise linear regression, an interative technique, builds up the regression function in successive small steps k in direction of the target whose correlation with the current residuals is maximal, so that $0 < k < |c_{\hat{j}}|$; where $c_{\hat{j}}$ is the current correlation between the targets and the regressor $\hat{j}$ with

[1] A S-level qualitative variable is represented by a vector of S binary variables or bits, only one of which is *on* at a time. Thus, the j-th component of an output vector $\boldsymbol{y}$ is set to 1 if it belongs to class j and 0 otherwise.

higher absolute correlation. A big step size $k = |c_{\hat{j}}|$ leads to the classic forward selection, which can be overly greedy. In order to improve the process, LARS determines analytically the optimal step size in direction of target such that another regressor has as much correlation as possible with the current predictor. The MRSR technique is shown below.

Let the approximation of the linear system presented in Eq. (1) as $\mathbf{D}\mathbf{B}^t = \mathbf{\Delta}^t$, where, in terms of the MRSR, $\mathbf{D}$ is the regressor matrix, $\mathbf{B}^t$ is the solution for the linear system and, of course, $\mathbf{\Delta}^t$ is the t-th approximation of $\mathbf{\Delta}$. The matrix $\mathbf{B}^t$ is updated by

$$\mathbf{B}^{t+1} = (1 - \gamma^t)\mathbf{B}^t + \gamma^t \bar{\mathbf{B}}^{t+1}, \tag{5}$$

where $\gamma^t = \min\{\gamma : \gamma \geq 0 \text{ and } \gamma \in \Gamma_j \text{ for some } j \notin \mathcal{A}\}$ is the step size at t-th iteration, so that Γ_j is the set

$$\Gamma_j = \left\{ \frac{c^t_{\max} + \boldsymbol{s}^T(\mathbf{\Delta} - \mathbf{\Delta}^t)^T \boldsymbol{d}_j}{c^t_{\max} + \boldsymbol{s}^T(\bar{\mathbf{\Delta}}^{t+1} - \mathbf{\Delta}^t)^T \boldsymbol{d}_j} \right\}, \tag{6}$$

and $c^t_{\max} = \max_j \left\{ c^t_j = \left\| (\mathbf{\Delta} - \mathbf{\Delta}^t)^T \boldsymbol{d}_j \right\|_1 \right\}$ is the maximum cumulative correlations from the set of regressors that satisfy the maximum $\mathcal{A} = \{j : c^t_j = c^t_{\max}\}$, $\boldsymbol{s}$ is a vector of ± 1, and, finally, the matrix $\bar{\mathbf{B}}^{t+1}$ is computed by the usual least squares estimate

$$\bar{\mathbf{B}}^{t+1} = \mathbf{D}_{\mathcal{A}}(\mathbf{D}^T_{\mathcal{A}}\mathbf{D}_{\mathcal{A}})^{-1}\mathbf{D}^T_{\mathcal{A}}\mathbf{\Delta}, \tag{7}$$

where $\mathbf{D}_{\mathcal{A}} = [\ \cdots\ \boldsymbol{d}_j\ \cdots\ ]_{j \in \mathcal{A}}$ is an $N \times |\mathcal{A}|$ matrix of collected regressors that belong to $\mathcal{A}$. Note that $\boldsymbol{d}_j$ is the i-th column of $\mathbf{D}$.

The idea behind the MRSR is to have, at beginning, the matrix $\mathbf{B}^t$ with zero values and then to add a new nonzero row at each new step. Therefore, the weight matrix has k nonzero rows at k-th step of the MRSR. As each row is added one after another, the sequence created represents the rank of rows. The row to be added is chosen by computing the cumulative correlation between the regressor (vector $\boldsymbol{d}_j : j \notin \mathcal{A}$) and the current residuals (the difference $\mathbf{\Delta} - \mathbf{\Delta}^t$). As the linear system output is a vector, the MRSR coincides with the LARS algorithm [7]. In fact, as stated, MRSR is an extension of LARS. More details about MRSR can be found in [19].

5 Our Proposal: OS-MLM

In a nutshell, the optimally selected minimal learning machine (OS-MLM)[2] relies on three main steps. The first step is to build the matrix of distances $\mathbf{D}$ from the data using all patterns as RPs. The second one is to rank columns of $\mathbf{D}$ (which also means to rank patterns) by MRSR in order to obtain the most relevant columns (i.e., reference points). After that, the last step is the leave-one-out optimization so that the best set of columns is achieved. The main idea is to leave less important columns (patterns) out of the solution by eliminating the columns

[2] The use of therm "optimally" is based on the our inspiration, the optimally pruned extreme learning machines method [16].

with low rank from matrix $\mathbf{D}$, then to solve the problem by the pseudo-inverse. We highlight the $\mathbf{D}$'s rows, also associated with a certain reference point, are not removed because its elimination would lead to a loss of labeling information and performance [23]. Despite the MRSR ranks the rows, we can use the row ranking as column ranking since the n-th row equals the n-th column (i.e., $\mathbf{D}^T = \mathbf{D}$). After ranking the patterns, the decision for the best set of reference points in the model is taken by LOO validation method. As such, LOO is computed by the prediction sum of squares (PRESS, [3]), since the LOO in its default version is very time consuming. At last, the matrix $\mathbf{B}$ is calculated in order to obtain the final model. We highlight that our proposal is inspired by recent methodologies, called OP-ELM and OP-LSSVM, proposed to prune hidden layer neurons in extreme learning machines (ELM, [9]) and support vectors in least squares support vector machines (LSSVM, [18]), respectively.

5.1 OS-MLM Algorithm

We present the proposed algorithm for OS-MLM below.

Algorithm 1. OS-MLM

Input: $\mathcal{D}$: training dataset
Output: Regression model ($\hat{\mathbf{B}}$), set of RPs ($\mathcal{R}$)
1: Build input distance matrix $\mathbf{D}$ from the data using all samples as RPs
2: Rank the columns of $\mathbf{D}$ (reference points) by MRSR

$$\boldsymbol{r} \leftarrow \textsc{Ranking-By-MRSR}(\mathbf{D})$$

3: Using the rank $\boldsymbol{r}$, select the best RP set by PRESS LOO

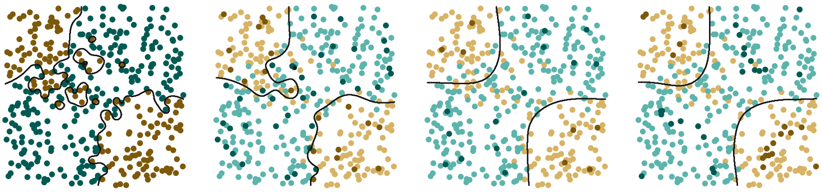
$$\mathcal{R} \leftarrow \textsc{Selecting-By-LOO}(\mathbf{D}, \boldsymbol{r})$$

4: Compute the distance matrices $\mathbf{D}$ and $\boldsymbol{\Delta}$ for data poins and the new $\mathcal{R}$ set.
5: Calculate the regression model $\hat{\mathbf{B}}$ through Eq. (2)
6: **return** $\hat{\mathbf{B}}$, $\mathcal{R}$

6 Simulations and Discussion

For a qualitative analysis, we have also applied OS-MLM, RN-MLM, FL-MLM and FCM-MLM to solve an artificial problem. The problem named simple checkerboard (SCB) consists of 400 points taken from a 2×2 checkerboard (Fig. 1).

Based on Fig. 1, one can see that the number of RPs for OS-MLM is lower than the number of RPs for RN-MLM and very close to the FCM-MLM. Moreover, the decision boundary generated from the FCM-MLM is more smoothed than the other models. Additionally, one can note, in the qualitative analysis, the OS-MLM method avoids RPs on overlapping regions. Thus, the decision boundaries are not overfitted.



(a) $|\mathcal{R}| = 400$ (b) $|\mathcal{R}| = 60$ (c) $|\mathcal{R}| = 21$ (d) $|\mathcal{R}| = 38$

Fig. 1. Decision surface and number of RPs for (a) FL-MLM, (b) RN-MLM, (c) FCM-MLM, and (d) OS-MLM, when applied to SCB problem.

Tests with real-world benchmark datasets were also evaluated in this work. We used some UCI datasets [10]: Haberman's Survival (HAB) with 3 features and 306 patterns; Vertebral Column Pathologies (VCP) with 4 features and 310 patterns; Liver Disorders (LID) with 6 features and 345 patterns; Ionosphere (ION) with 34 features and 351 patterns; Breast Cancer Winconsin (BCW) with 9 features and 683 patterns; Pima Indians Diabets (PID) with 8 features and 768 patterns; Car Evaluation (CAR) with 6 features and 1728 patterns; and Human Immunodeficiency Virus protease cleavage (HIV) with 8 features and 3272 patterns. In addition, two well-known artificial data sets were also used in our simulations, Ripley (RIP) with 2 features and 1250 patterns and Two Moon (TMN) with 2 features and 1001 patterns.

In our simulations, 80% of the data was randomly selected for training purposes and the remaining 20% was used for assessing the classifiers' generalization performance. We carried out 30 executions on each dataset.

The adjustment of the parameter K for the FCM-MLM and the RN-MLM model was performed using grid search with 10-fold cross-validation. The RPs were selected in the range of 5–100% (with a step of 5%) of the available training samples. The accuracy was used to choose the best value of K.

In order to verify the possible equivalence between the classifier accuracies, we perform a two-sample Friedman test [6] with a significance level of 1%. In the hypothesis test, the null and the alternative hypothesis means that the accuracy is equivalent or not, respectively.

In Table 1, we report performance metrics for the aforementioned 30 independent runs. We show the accuracy (ACC), the average of the percentage of reduction (RED) in the RPs, compared to the FL-MLM model, and the results of the Friedman statistic test, in which ✓ is a relation of equivalence between the methods and ✗ indicates a considerable difference among them.

As expected, the performance of the OS-MLM was equivalent to or higher than that achieved by the RN-MLM, FL-MLM, and the FCM-MLM for each evaluated dataset. Particularly, the OS-MLM achieved the best results (in terms of accuracy) in 4 of the datasets. In terms of RPs reduction, our proposal achieved best results in 9 of the datasets. In other words, our proposal achieves accuracies

Table 1. Performance metrics for OS-MLM, RN-MLM, FL-MLM, and FCM-MLM; Accuracy (ACC) and reduction percentage in comparison with the training set (RED); and results of statistical tests.

dataset	*metric*	OS-MLM	RN-MLM	FL-MLM	FCM-MLM
HAB	ACC	**73.44** ± 4.09	71.97 ± 4.27 ✓	68.09 ± 4.98 ✗ ✗	71.36 ± 6.03 ✓ ✓ ✗
	RED	**94.46** ± 3.32	80.20 ± 14.22		59.44 ± 16.50
VCP	ACC	87.20 ± 3.57	**87.80** ± 3.92 ✓	87.37 ± 3.98 ✓ ✓	84.78 ± 4.48 ✓ ✓ ✓
	RED	**79.87** ± 13.66	49.17 ± 27.95		80.60 ± 5.99
LID	ACC	**71.01** ± 5.16	68.84 ± 5.59 ✓	68.16 ± 6.76 ✗ ✓	68.70 ± 5.44 ✓ ✓ ✓
	RED	**86.32** ± 6.12	59.82 ± 26.51		63.61 ± 20.79
ION	ACC	92.33 ± 3.04	93.10 ± 2.73 ✓	**94.14** ± 2.74 ✗ ✗	93.33 ± 2.87 ✗ ✓ ✓
	RED	**76.57** ± 18.93	42.18 ± 21.05		57.77 ± 12.40
BCW	ACC	91.33 ± 12.75	96.98 ± 1.40 ✗	96.96 ± 1.27 ✗ ✓	**97.00** ± 1.31 ✓ ✓ ✓
	RED	74.25 ± 27.37	62.61 ± 22.71		**86.91** ± 4.65
PID	ACC	**75.15** ± 2.96	74.59 ± 2.58 ✓	73.16 ± 2.38 ✗ ✗	73.44 ± 2.86 ✗ ✗ ✓
	RED	**95.04** ± 0.87	75.92 ± 16.10		84.61 ± 8.11
CAR	ACC	92.89 ± 1.95	97.00 ± 0.76 ✗	**98.02** ± 0.69 ✗ ✗	96.12 ± 1.29 ✗ ✗ ✗
	RED	**90.69** ± 3.48	17.64 ± 10.05		27.14 ± 6.96
TMN	ACC	99.52 ± 0.48	99.82 ± 0.28 ✗	**99.87** ± 0.22 ✗ ✓	**99.87** ± 0.22 ✗ ✗ ✗
	RED	**93.10** ± 6.79	61.92 ± 20.72		63.63 ± 23.71
RIP	ACC	**90.49** ± 1.83	89.75 ± 1.77 ✓	88.32 ± 1.61 ✗ ✗	89.81 ± 1.88 ✓ ✓ ✗
	RED	**96.02** ± 0.80	76.64 ± 18.83		87.30 ± 11.01
HIV	ACC	86.62 ± 1.26	86.50 ± 1.30 ✓	85.99 ± 1.14 ✗ ✗	**86.68** ± 1.30 ✓ ✗ ✗
	RED	**90.92** ± 2.14	75.32 ± 23.16		90.68 ± 17.02

that are comparable to others variants of MLM, but with a lower number of RPs. It is also important to notice that for most datasets the OS-MLM achieved a low standard deviation.

7 Conclusions

Motivated by the poor generalization capability of the random selection, this paper presents an alternative algorithm to select reference points of the MLM for classification tasks based on multiresponse sparse regression and PRESS statistics. Four strategies of MLM RPs selection are evaluated. The results of the simulations indicate the OS-MLM works very well, providing a competitive classifier while maintaining its simplicity. We also achieved sparseness in our classifier.

References

1. Alcin, O., Sengur, A., Qian, J., Ince, M.: OMP-ELM: orthogonal matching pursuit-based extreme learning machine for regression. J. Intell. Syst. **24**(1), 135–143 (2015)
2. Alencar, A.S.C., et al.: MLM-rank: a ranking algorithm based on the minimal learning machine. In: 2015 Brazilian Conference on Intelligent Systems, BRACIS 2015, Natal, Brazil, 4–7 November 2015, pp. 305–309. IEEE (2015)
3. Allen, D.M.: The relationship between variable selection and data agumentation and a method for prediction. Technometrics **16**(1), 125–127 (1974)
4. Coelho, D.N., Barreto, G.D.A., Medeiros, C.M.S., Santos, J.D.A.: Performance comparison of classifiers in the detection of short circuit incipient fault in a three-phase induction motor. In: 2014 IEEE Symposium on Computational Intelligence for Engineering Solutions, CIES 2014, Orlando, FL, USA, 9–12 December 2014, pp. 42–48. IEEE (2014)
5. de Sousa, L.S., Dias, M.L.D., Rocha Neto, A.R.: Máquinas de vetores-suporte de mínimos quadrados esparsas via recozimento simulado. In: Simpósio Brasileiro de Automação Inteligente (SBAI). SBA, Rio Grande do Norte, Brasil, October 2015
6. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
7. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**(2), 407–499 (2004)
8. Florêncio, J.A.V., Dias, M.L.D., da Rocha Neto, A.R., de Souza Júnior, A.H.: A fuzzy C-means-based approach for selecting reference points in minimal learning machines. In: Barreto, G.A., Coelho, R. (eds.) NAFIPS 2018. CCIS, vol. 831, pp. 398–407. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95312-0_34
9. Huang, G., Zhu, Q., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing **70**(1–3), 489–501 (2006)
10. Lichman, M.: UCI machine learning repository (2013)
11. Luo, J., Vong, C., Wong, P.: Sparse Bayesian extreme learning machine for multi-classification. IEEE Trans. Neural Netw. Learn. Syst. **25**(4), 836–843 (2014)
12. MacKay, D.J.: Bayesian interpolation. Neural Comput. **4**(3), 415–447 (1992)
13. Marinho, L.B., Almeida, J.S., Souza, J.W.M., de Albuquerque, V.H.C., Filho, P.P.R.: A novel mobile robot localization approach based on topological maps using classification with reject option in omnidirectional images. Expert Syst. Appl. **72**, 1–17 (2017)
14. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Ind. Appl. Math. **11**(2), 431–441 (1963)
15. Mesquita, D.P.P., Gomes, J.P.P., Junior, A.H.S.: Ensemble of efficient minimal learning machines for classification and regression. Neural Process. Lett. **46**, 1–16 (2017)
16. Miche, Y., Sorjamaa, A., Bas, P., Simula, O., Jutten, C., Lendasse, A.: OP-ELM: optimally pruned extreme learning machine. IEEE Trans. Neural Netw. **21**(1), 158–162 (2010)
17. Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.S.: Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In: Proceedings of 27th Asilomar Conference on Signals, Systems and Computers, vol. 1, pp. 40–44, November 1993
18. da Silva Vieira, D.C., da Rocha Neto, A.R., Rodrigues, A.W.D.O.: Sparse least squares support vector regression via multiresponse sparse regression. In: 2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, 24–29 July 2016, pp. 3218–3225. IEEE (2016)

19. Similä, T., Tikka, J.: Multiresponse sparse regression with application to multidimensional scaling. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3697, pp. 97–102. Springer, Heidelberg (2005). https://doi.org/10.1007/11550907_16
20. de Souza Junior, A.H., Corona, F., Barreto, G.D.A., Miché, Y., Lendasse, A.: Minimal learning machine: a novel supervised distance-based approach for regression and classification. Neurocomputing **164**, 34–44 (2015)
21. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
22. Tipping, M.E.: Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. **1**, 211–244 (2001)
23. Valyon, J., Horvath, G.: A sparse least squares support vector machine classifier. In: Proceedings of IEEE International Joint Conference on Neural Networks, vol. 1, pp. 543–548 (2004)