ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

A Robust Minimal Learning Machine based on the M-Estimator

João P. P. Gomes[1], Diego P. P. Mesquita[1] Ananda L. Freire[2], Amauri H. Souza Junior[2], and Tommi Karkkainen[3]

[1] Federal University of Ceará - UFC,
Department of Computer Science , Fortaleza-CE, Brazil
jpaulo@lia.ufc.br,diego@diegoparente.com
[2] Federal Institute of Ceará,
Department of Teleinformatics, Fortaleza-CE, Brazil
amauriholanda@ifce.edu.br,anandalf@gmail.com
[3] University of Jyvaskyla,
Department of Math. Information Technology, Finland
tommi.karkkainen@jyu.fi

Abstract. In this paper we propose a robust Minimal Learning Machine (R-RLM) for regression problems. The proposed method uses a robust M-estimator to generate a linear mapping between input and output distances matrices of MLM. The R-MLM was tested on one synthetic and three real world datasets that were contaminated with an increasing number of outliers. The method achieved a performance comparable to the robust Extreme Learning Machine (R-RLM) and thus can be seen as a valid alternative for regression tasks on datasets with outliers.

1 Introduction

Machine learning methods have been extensively applied in various problems, achieving remarkable performances. Despite their popularity, it is known that the performance of the least-squares methods strongly depends on the quality of the data [1]. If a supervised data set contains outliers, a robust approach can provide a final model that properly maps the inputs to the desired outputs [2].

The design of outlier robust machine learning methods has been addressed in many works and is still an active research topic. In [3], the authors proposed a robust Extreme Learning Machine (ELM) where the $l1$-norm loss function was used to enhance the robustness. A robust ELM was also proposed in [4], based on batch and sequential learning strategies for various forms of M-estimators.

Recently, a new supervised learning method, called Minimal Learning Machine (MLM, [5]), was proposed. The technique achieved promising results on applications like face recognition [6] and systems identification [7]. MLM is based on the idea of the existence of a mapping between the geometric configurations of points in the input and output space. The main advantages of MLM include fast training, simple formulation, and only one hyperparameter to be optimized (number of reference points).

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

This paper proposes a robust MLM for data sets with outliers. The proposed method uses the Iteractively Reweighted Least Squares (IRLS) algorithm to provide a robust estimate of a linear model that maps MLM's input and output distance matrices. The proposed approach is compared to the standard MLM, ELM and a recently proposed robust ELM [4], achieving promising results.

2 Minimal Learning Machine

Let $\mathcal{X} = \{X_i\}_{i=1}^N$ be a set of training inputs, $\mathcal{Y} = \{Y_i\}_{i=1}^N$ the set of their respective outputs, $\mathcal{R} = \{R_k\}_{k=1}^K$ be a non-empty subset of $\mathcal{X}$ and $\mathcal{T} = \{T_k\}_{k=1}^K$ be such that T_i is the output of R_i. Furthermore, let $D_x, \Delta_y \in \mathbb{R}^{N\times K}$ be euclidean distance matrices such that the their k-th columns are respectively $[\|X_1 - R_k\|_2, \cdots, \|X_N - R_k\|_2]^T$ and $[\|Y_1 - T_k\|_2, \cdots, \|Y_N - T_k\|_2]^T$. The key idea behind MLM is the assumption of a linear mapping between D_x and Δ_y, giving rise to the following regression model:

$$\Delta_y = D_x\beta + E \tag{1}$$

where $E \in \mathbb{R}^{N\times K}$ is a matrix of residuals and $\beta \in \mathbb{R}^{K\times K}$ is the matrix of regression coefficients. It turns out β can be estimated using Ordinary Least-Squares (OLS):

$$\hat{\beta} = (D_x^T D_x)^{-1} D_x^T \Delta_y \tag{2}$$

Thus, given a new input point X, we can obtain an estimate $\hat{\delta} = [\hat{\delta}_1, \cdots, \hat{\delta}_K]$ of the distances between the output of X and the K points in T, given by:

$$\hat{\delta} = [\|X - T_1\|_2, \cdots, \|X - T_K\|_2]\,\hat{\beta}. \tag{3}$$

In other words, we expect that the output Y of X to be such that:

$$\|Y - T_i\|_2 \approx \hat{\delta}_i \quad \forall i \in \{1, \cdots, K\}. \tag{4}$$

Therefore, an estimate $\hat{Y}$ of Y can be obtained by minimizing

$$J(Y) = \sum_{k=1}^{K} \left((Y - T_k)^T (Y - T_k) - \hat{\delta}_k^2 \right)^2, \tag{5}$$

which can be done employing any gradient-based optimization algorithm.

3 Robust Minimal Learning Machine Using M-Estimators(R-MLM)

It is known that OLS can lead to bad estimates when the error distribution is not Gaussian. This may happen when data is contaminated with outliers. One remedy to this problem is the use of robust regression techniques such as the widely used M-estimators. The robustness of M-estimators is achieved by

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

minimizing a surrogate function instead of the sum of the squared errors [4]. Based on the work of Huber [9], a general M-estimator used to estimate the distance to the k-th reference point on the output space minimizes the following objective function:

$$J(\hat{\mathbf{b}}_k) = \sum_{i=1}^{N} \rho(\delta(\mathbf{y}_i, m_k) - \mathbf{d}(\mathbf{x}_i, R)\hat{\mathbf{b}}_k), \tag{6}$$

where $\hat{\mathbf{b}}_k$ is the k-th column of $\hat{\mathbf{B}}$ and the function $\rho(\cdot)$ gives the contribution of each error $e_{i,k} = \delta(\mathbf{y}_i, m_k) - \hat{\delta}(\mathbf{y}_i, m_k)$ to the objective function. OLS is a particular case of M-estimator, characterized by $\rho(e_{i,k}) = e_{i,k}^2$. The function ρ must be such that $\rho(e) \geq 0$, $\rho(0) = 0$, $\rho(e) = \rho(-e)$ and, if $|e| > |e'|$, $\rho(e) \geq \rho(e')$. For more details, the reader may recur to [10].

Let ϱ denote the derivative of ρ. Differentiating ρ with respect to $\hat{\mathbf{b}}_k$ and setting to zero, we have:

$$\sum_{i=1}^{N} \varrho\left(\delta(\mathbf{y}_i, m_k) - \mathbf{d}(\mathbf{x}_i, R)\hat{\mathbf{b}}_k\right) \mathbf{d}(\mathbf{x}_i, R)^T = \mathbf{0}, \tag{7}$$

where $\mathbf{0} \in \mathbb{R}^k$ is a row vector of zeros. Consider the weight function $\omega(e_{i,k}) = \varrho(e_{i,k})/e_{i,k}$, and let $\omega_{i,k} = \omega(e_{i,k})$. Then, the estimating equations are given by:

$$\sum_{n=1}^{N} \omega_{i,k} \left(\delta(\mathbf{y}_i, m_k) - \mathbf{d}(\mathbf{x}_i, R)\hat{\mathbf{b}}_k\right) \mathbf{d}(\mathbf{x}_i, R)^T = \mathbf{0}. \tag{8}$$

Thus, solving the estimating equations corresponds to solving a weighted least-squares problem, minimizing Eq. (9).

$$\sum_{i=1}^{N} \omega_{i,k}^2 e_{i,k}^2 = \sum_{i=1}^{N} \omega^2(e_{i,k}) e_{i,k}^2. \tag{9}$$

One should bear in mind that there is no closed-form equation to estimate $\hat{\mathbf{b}}_k$, once the weights depend on the estimated errors, these errors depend upon the estimated coefficients, and these coefficients depend upon the weights [4]. In such context, the *Iteratively Reweighted Least-Squares* (IRLS) [10] emerges as an alternative estimation method. Its algorithm can be described as follows:

Step 1. Provide an initial estimate $\hat{\mathbf{B}}(0)$ using the OLS solution.

Step 2. At each iteration j, compute the residuals from the previous iteration $e_{i,k}(j-1)$, $i = 1, \ldots, N$, associated with the distance to the k-th reference point on the output space, and then compute the corresponding weights $\omega_{i,k}(j-1) = \omega[e_{i,k}(j-1)]$.

Step 3. Solve for new weighted-least-squares estimate of $\hat{\mathbf{b}}_k(j)$:

$$\hat{\mathbf{b}}_k(j) = \left[\mathbf{D}_x^T \mathbf{W}(j-1)\mathbf{D}_x\right]^{-1} \mathbf{D}_x^T \mathbf{W}(j-1)\boldsymbol{\delta}(Y, \mathbf{t}_k), \tag{10}$$

where $\mathbf{W}(j-1) = \text{diag}\{\omega_{i,k}(j-1)\}$ is an $N \times N$ weight matrix.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

Step 4. Repeat Steps 2 and 3 until the convergence of $\hat{\mathbf{b}}_k(j)$.

From a variety of weighting functions available in the literature for M-estimators, we chose the bisquare weighting function:

$$\omega(e_{i,k}) = \begin{cases} \left[1 - \left(\frac{e_{in}}{\kappa}\right)^2\right]^2, & \text{if } |e_{i,k}| > \kappa \\ 1, & \text{otherwise.} \end{cases} \tag{11}$$

Smaller values of κ leads to more resistance to outliers, but at the expense of lower efficiency when the errors are normally distributed [4]. In particular, we choose $\kappa = 4.685\sigma$, where σ is a robust estimate of the standard deviation of the errors. A common approach is to take $\sigma = \text{MAR}/0.6745$, where MAR is the median absolute residual.

4 Experiments

To asses the performance of R-MLM, we conducted experiments using four regression problems: SinC, Battery, Dow Jones Index (Stocks) and Boston Housing. SinC is an artificial data set composed by 2000 samples with 1 input and 1 output generated from the sinc function. Battery is a time series prediction task where future voltage values shall be predicted given previous ones. The dataset is composed by 3000 samples with 5 inputs and 1 output. A detailed description of this dataset can be seen in [11]. The remaining data sets are real-world problems available at UCI machine learning repository [12]. Stocks comprises 750 samples with 16 inputs and 1 output, while Boston has 506 samples with 13 inputs and 1 output. For training and testing, the sets were divided: SinC (1000/1000), Battery (1500/1500), Stocks (400/350) and Boston (379/127). The R-MLM was compared to ELM, MLM and the robust ELM proposed in [4]. For all methods, we performed a 10-fold cross validation to select the hyper-parameters (number of reference point in MLM and the number of hidden neurons in ELM).

The outlier robustness of the methods was investigated by randomly contaminating several training data targets with one-sided or two-sided outliers. For the two-sided case, normally distributed errors were added to a percentage of the target values. In the one-sided case the absolute value of the errors was added. This approach is described in details in [13]. It is worth noting that no testing data was corrupted with outliers.

Table 1 presents the mean and standard deviation of the Root Mean Squared Error (RMSE) for all data sets and methods for 10% and 30% of outliers. Each entry in Table 1 correspond to the outcome of 20 similar trials. In each of these trials, the training and test samples drawn randomly without replacement from the original data sets. All implementations were executed using MATLAB.

As expected, the robust variants of MLM and ELM achieved lower RMSE values when compared to its standard versions. This performance gap is even more noticeable when the number of outliers is increased. Analyzing the results of R-ELM and R-MLM, it is possible to see that R-ELM performed better on the SinC example while R-MLM achieved better results in Boston Housing. For the Stocks and Battery datasets, the results were very similar.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

Table 1. Comparison between ELM, MLM, R-ELM and R-MLM.

Data	**Method**	**Mean testing RMSE and Standard Deviation**	
		Contamination rate (%)	
		10	30
SinC (1 sided)	ELM	0.11337 ± 0.00622	0.27326 ± 0.00874
	MLM	0.09440 ± 0.01050	0.24440 ± 0.01790
	R-ELM	0.00595 ± 0.00490	0.00835 ± 0.00065
	R-MLM	0.03950 ± 0.03190	0.04680 ± 0.02880
SinC (2 sided)	ELM	0.09168 ± 0.00331	0.13190 ± 0.00778
	MLM	0.06530 ± 0.00800	0.11720 ± 0.02400
	R-ELM	0.00801 ± 0.00511	0.00710 ± 0.00240
	R-MLM	0.02260 ± 0.00890	0.04111 ± 0.02480
Stocks (1 sided)	ELM	0.1573 ± 0.0148	0.3102 ± 0.0241
	MLM	0.1261 ± 0.0148	0.2720 ± 0.0230
	R-ELM	0.0758 ± 0.0060	0.0828 ± 0.0061
	R-MLM	0.0701 ± 0.0039	0.0874 ± 0.0211
Stocks (2 sided)	ELM	0.1325 ± 0.0146	0.2087 ± 0.0267
	MLM	0.1096 ± 0.0117	0.1616 ± 0.0020
	R-ELM	0.0763 ± 0.0059	0.0804 ± 0.0049
	R-MLM	0.0777 ± 0.0163	0.0780 ± 0.0054
Boston (1 sided)	ELM	0.1962 ± 0.02608	0.3270 ± 0.0311
	MLM	0.1606 ± 0.0255	0.2787 ± 0.0336
	R-ELM	0.1470 ± 0.0218	0.1865 ± 0.0413
	R-MLM	0.1178 ± 0.0237	0.1523 ± 0.0332
Boston (2 sided)	ELM	0.1924 ± 0.0284	0.2380 ± 0.0526
	MLM	0.1630 ± 0.0208	0.2026 ± 0.0326
	R-ELM	0.1505 ± 0.0277	0.1990 ± 0.0359
	R-MLM	0.1186 ± 0.0283	0.1689 ± 0.0374
Battery (1 sided)	ELM	0.1152 ± 0.0115	0.2750 ± 0.0333
	MLM	0.1222± 0.0321	0.2812 ± 0.0314
	R-ELM	0.0912 ± 0.0222	0.1913 ± 0.0421
	R-MLM	0.0901 ± 0.0246	0.1878 ± 0.0401
Battery (2 sided)	ELM	0.1089 ± 0.0277	0.2380 ± 0.0445
	MLM	0.1130 ± 0.0212	0.2536 ± 0.0319
	R-ELM	0.0943 ± 0.0278	0.1879 ± 0.0456
	R-MLM	0.0911 ± 0.0276	0.1912 ± 0.0419

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

5 Conclusion

This work presented a variant of the MLM algorithm with improved robustness for datasets with outliers. The so called R-MLM, uses robust estimates to find the mapping between input and output distance matrices of MLM. R-MLM was tested on two real world and one synthetic dataset and was compared to MLM, ELM and a robust version of ELM (R-ELM). Result showed the R-MLM outperformed ELM and MLM on its standard versions and achieved comparable results when compared to R-ELM. On the basis of the achieved results, we can state that R-MLM is a valid alternative for regression problems with outliers.

Acknowledgments

The authors acknowledge the support of CNPq (Grant 402000/2013-7).

References

1. Chen D, Jain R.: A robust backpropagation learning algorithm for function approximation. IEEE Trans Neural Net. 5, 467–479, 1994
2. Kärkkäinen T., Heikkola E.: Robust formulations for training multilayer perceptrons. Neural Computation, 16, 837–862, 2004
3. Kai Zhang, Minxia Luo: Outlier-robust extreme learning machine for regression problems. Neurocomputing. 151, 1519–1527, 2015
4. A. L. B. Barros, G. A. Barreto: Building a robust extreme learning machine for classification in the presence of outliers in Hybrid Artificial Intelligent Systems, ser. Lecture Notes in Computer Science . 8073, 588–597, 2013
5. Souza Junior A.H., Corona F., Miché Y., Lendasse A., Barreto G., Simula O.: Minimal Learning Machine: A New Distance-Based Method for Supervised Learning. Proceedings of the 12th International Work Conference on Artificial Neural Networks (IWANN'2013). 7902, 408–416, 2013
6. Mesquita, D. P. P.; Gomes, J. P. P., Junior, A. H. S.: Ensemble of Minimal Learning Machines for Pattern Classification in Ignacio Rojas; Gonzalo Joya Caparros; Andreu Catala, ed., 'IWANN (2)',Springer . 164, 142–152, 2015
7. Souza Junior A.H.S, Corona F., Barreto G. A., Miche Y., Lendasse A.: Minimal Learning Machine: A novel supervised distance-based approach for regression and classification Neurocomputing . 164, 34–44, 2015
8. Marquardt D. W.: An Algorithm for Least-Squares Estimation of Nonlinear Parameters Journal of the SIAM. 11, 431–441, 1963
9. P. J. Huber: Robust estimation of a location parameter Annals of Mathematical Statistics . 35, 73–101, 1964
10. Fox J.F: Applied Regression Analysis, Linear Models, and Related Methods Sage Publications, 1997.
11. Darielson Souza, Vandilberto Pinto, Luis Nascimento, Joao Torres, Joao Gomes, Jarbas Sa-Junior and Romulo Almeida: Battery Discharge forecast applied in Unmanned Aerial Vehicle Przeglad Elektrotechniczny . 2, 185–192, 2016
12. Frank A., Asuncion A.: UCI Machine Learning Repository University of California, Irvine, School of Information and Computer Sciences, 2010
13. Horata P., Chiewchanwattana S., Sunat K.: Robust extreme learning machine Neurocomputing . 102, 31–24, 2013