Minimal Learning Machine: A novel supervised distance-based approach for regression and classification

Article *in* Neurocomputing · January 2014

DOI: 10.1016/j.neucom.2014.11.073

CITATIONS
42

READS
785

5 authors, including:

Francesco Corona

76 PUBLICATIONS **1,254** CITATIONS

SEE PROFILE

Guilherme A. Barreto

Universidade Federal do Ceará

180 PUBLICATIONS **1,664** CITATIONS

SEE PROFILE

Yoan Miche

Nokia Bell Labs, Espoo, Finland

117 PUBLICATIONS **2,836** CITATIONS

SEE PROFILE

Amaury Lendasse

University of Houston

318 PUBLICATIONS **5,711** CITATIONS

SEE PROFILE

Neurocomputing ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Neurocomputing

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

Minimal Learning Machine: A novel supervised distance-based approach for regression and classification

Amauri Holanda de Souza Júnior[a,b,*], Francesco Corona[b,c], Guilherme A. Barreto[b], Yoan Miche[c], Amaury Lendasse[d,e]

[a] Federal Institute of Ceará, Department of Computer Science, Maracanaú, Ceará, Brazil
[b] Federal University of Ceará, Department of Teleinformatics Engineering, Av. Mister Hull, S/N - Center of Technology - Campus of Pici CP 6005, CEP 60455-760 Fortaleza, Ceará, Brazil
[c] Aalto University, Department of Computer Science, Konemiehentie 2, Espoo, Finland
[d] Department of Mechanical and Industrial Engineering and the Iowa Informatics Initiative, 3131 Seamans Center, The University of Iowa, Iowa City, IA 52242-1527, USA
[e] Arcada University of Applied Science, Helsinki, Finland

ARTICLE INFO

Article history:
Received 21 January 2014
Received in revised form
27 September 2014
Accepted 11 November 2014

Keywords:
Learning machines
Supervised learning
Regression
Pattern classification

ABSTRACT

In this work, a novel supervised learning method, the Minimal Learning Machine (MLM), is proposed. Learning in MLM consists in building a linear mapping between input and output distance matrices. In the generalization phase, the learned distance map is used to provide an estimate of the distance from K output reference points to the unknown target output value. Then, the output estimation is formulated as multilateration problem based on the predicted output distance and the locations of the reference points. Given its general formulation, the Minimal Learning Machine is inherently capable of operating on nonlinear regression problems as well as on multidimensional response spaces. In addition, an intuitive extension of the MLM is proposed to deal with classification problems. A comprehensive set of computer experiments illustrates that the proposed method achieves accuracies that are comparable to more traditional machine learning methods for regression and classification thus offering a computationally valid alternative to such approaches.

1. Introduction

Supervised machine learning methods for regression and classification have been designed mostly for data types that lie in vector spaces, i.e. response (i.e. output) and/or predictor (i.e. input) variables are often arranged into vectors of predefined dimensionality. There are other types of data, however, such as graphs, sequences, shapes, images, trees and covariance matrices, which are less amenable to being treated within standard regression/classification frameworks. These data types usually do not lie in a natural vector space, but rather in a metric space.

For this type of data, usually referred to as structured data [13,16], a more general approach to the characterization of the data items is to define a distance (or dissimilarity) measure between data items and to provide a learning algorithm that works with the resulting distance matrix. Since pairwise distance measures can be defined on structured objects (e.g. graphs, trees or strings), this procedure provides a bridge between the classical and the structural/syntactic approaches to pattern recognition [3,30].

Pairwise distance data occur frequently in empirical sciences, such psychology, economics, ecology and biochemistry, with most of the algorithms developed to handle this kind of data falling into the realm of unsupervised learning, predominantly as clustering [14,33,12,19] or multidimensional scaling algorithms [34].

For regression tasks, there are some prior works in which the response and/or the predictor variables are expressed as distance (i.e. dissimilarity) matrices. Cuadras and Arenas [8] proposed an approach to regression where only the predictors are expressed as a distance and classical multidimensional scaling (a.k.a principal coordinates analysis) [34] is used to generate scores. The response variable is then regressed on these scores. McArdle and Anderson [24] performed MANOVA[1] on ecological data with only knowledge of the distance matrix of the response variable. Finally, Lichstein [22] proposed a modeling approach where both the response and predictor variables are represented as distance matrices. However,

* Corresponding author.
E-mail addresses: amauriholanda@ifce.edu.br (A.H. de Souza Júnior), francesco.corona@aalto.fi (F. Corona), gbarreto@ufc.br (G.A. Barreto), yoan.miche@aalto.fi (Y. Miche), amaury-lendasse@uiowa.edu (A. Lendasse).

[1] Acronym for multivariate analysis of variance.

http://dx.doi.org/10.1016/j.neucom.2014.11.073
0925-2312/

since their method converts the distance matrices to vectors in a column-wise fashion, useful information provided by the geometry of the problem is lost.

For classification tasks, we refer the reader to the works of Hammer et al. [15], Zhu et al. [39] and Graepel et al. [11]. Roughly speaking, these works introduce extended versions of classification algorithms for data characterization by means of a matrix of pairwise similarities or more general dissimilarity measures, rather than explicit feature vectors. In Hammer et al. [15] the authors propose a general learning framework that unifies previous attempts of making LVQ algorithms capable of handling non-vectorial data, such as kernel GLVQ [31,28] and relational GLVQ [9]. This is possible by means of a pseudo-Euclidean embedding[2] of similarity (or dissimilarity) data, i.e. every finite data set which is characterized by pairwise similarities or dissimilarities can be embedded in a so-called pseudo-Euclidean vector space. In Zhu et al. [39] it is proposed an LVQ-based classifier that, in addition to its ability to directly deal with arbitrary symmetric dissimilarity matrices, provides confidence/reliability measures for the classification results. Finally, in the pioneering work of Graepel et al. [11], they suggested classification algorithms based on linear models which operate on distance data from both Euclidean and pseudo-Euclidean spaces.

As mentioned in the previous paragraphs, data characterization by means of pairwise dissimilarity measures have been associated with the processing of structured data, either for regression or classification purposes. However, we argue that the use of dissimilarity measures for data characterization may also be beneficial for the processing of unstructured data types, by allowing, for example, a nonlinear learning problem to be tackled by linear models. Bearing this in mind, we introduce a new supervised nonparametric method, called the Minimal Learning Machine (MLM), aiming at the efficient design of distance-based regression models or pattern classifiers for unstructured data types.

The single assumption of the MLM is the existence of a mapping between the geometric configurations of points in the input and output spaces. Based on a set of comprehensive computer experiments, we show that such a mapping can be accurately reconstructed by learning a multiresponse linear model between distance matrices. Under these conditions, for an input point with known configuration in the input space, its corresponding configuration in the output space can be easily estimated after learning a simple linear model between input and output distance matrices. The resulting estimate is then used to locate the output point and, thus, provide an estimate for the response variable.

One of the main advantages of the MLM is that it requires tuning of a single parameter, which is the number of reference points (i.e. training output samples) used to obtain an estimate of the response variable. Another advantage is that the MLM can nicely handle nonlinear problems, even being, in essence, a linear model between distance matrices. The analysis of the results allows us to conclude that the proposed distance-based method, when applied to standard vectorial (i.e. unstructured) data types, achieves accuracies that are comparable to those achieved by standard supervised nonlinear machine learning methods for regression and classification, such as the multilayer perceptron (MLP), radial basis functions (RBF) networks, support vector machine/regression (SVM/SVR) models, extreme learning machines (ELMs), and Gaussian processes (GP) methods, thus offering a simpler alternative to these nonlinear approaches.

The remainder of the paper is organized as follows. In Section 2, the Minimal Learning Machine is presented; the MLM is formulated (Section 2.1), its properties are discussed (Section 2.2), a simple extension for classification tasks is introduced (Section 2.3), the links with related works are briefly reported (Section 2.4), and two illustrative examples are presented (Section 2.5). In Section 3, a thorough experimental assessment of the Minimal Learning Machine is conducted to evaluate its performance and to compare it with state-of-the-art approaches in regression and classification problems.

2. Minimal Learning Machine

In this section, we start by introducing the basic formulation of the Minimal Learning Machine (MLM).

2.1. Formulation

We are given a set of N input points $X=\{\mathbf{x}_i\}_{i=1}^{N}$, with $\mathbf{x}_i \in \mathbb{R}^D$, and the set of corresponding outputs $Y=\{\mathbf{y}_i\}_{i=1}^{N}$, with $\mathbf{y}_i \in \mathbb{R}^S$. Assuming the existence of a continuous mapping $f: \mathcal{X} \to \mathcal{Y}$ between the input and the output space, we want to estimate f from data with the multiresponse model

$$\mathbf{Y}=f(\mathbf{X})+\mathbf{R}.$$

The columns of the matrices $\mathbf{X}$ and $\mathbf{Y}$ correspond to the D inputs and S outputs respectively, and the rows to the N observations. The columns of the $N \times S$ matrix $\mathbf{R}$ correspond to the residuals.

The MLM is a two-step method designed to

1. reconstructing the mapping existing between input and output distances;
2. estimating the response from the configuration of the output points.

In the following, the two steps are discussed.

2.1.1. Distance regression

For a selection of reference input points $R=\{\mathbf{m}_k\}_{k=1}^{K}$ with $R \subseteq X$ and corresponding outputs $T=\{\mathbf{t}_k\}_{k=1}^{K}$ with $T \subseteq Y$, define $\mathbf{D}_x \in \mathbb{R}^{N \times K}$ in such a way that its kth column contains the distances $d(\mathbf{x}_i, \mathbf{m}_k)$ between the $i=1,\ldots,N$ input points $\mathbf{x}_i$ and the kth reference point $\mathbf{m}_k$. Analogously, define $\mathbf{\Delta}_y \in \mathbb{R}^{N \times K}$ in such a way that its kth column contains the distances $\delta(\mathbf{y}_i, \mathbf{t}_k)$ between the N output points $\mathbf{y}_i$ and the output $\mathbf{t}_k$ of the kth reference point. The mapping g between the input distance matrix $\mathbf{D}_x$ and the corresponding output distance matrix $\mathbf{\Delta}_y$ can be reconstructed using the multiresponse regression model

$$\mathbf{\Delta}_y=g(\mathbf{D}_x)+\mathbf{E}.$$

The columns of the matrix $\mathbf{D}_x$ correspond to the K input vectors and the columns of the matrix $\mathbf{\Delta}_y$ correspond to the K response vectors, the N rows correspond to the observations. The columns of the $N \times K$ matrix $\mathbf{E}$ correspond to the K residuals.

Assuming that mapping g between input and output distance matrices has a linear structure for each response, the regression model has the form

$$\mathbf{\Delta}_y=\mathbf{D}_x\mathbf{B}+\mathbf{E}. \tag{1}$$

The columns of the $K \times K$ regression matrix $\mathbf{B}$ correspond to the coefficients for the K responses. The matrix $\mathbf{B}$ can be estimated from data through a minimization of the multivariate residual sum of squares as loss function:

$$\mathrm{RSS}(\mathbf{B})=\mathrm{tr}\big((\mathbf{\Delta}_y-\mathbf{D}_x\mathbf{B})'(\mathbf{\Delta}_y-\mathbf{D}_x\mathbf{B})\big). \tag{2}$$

Under the normal conditions where the number of equations in Eq. (1) is larger than the number of unknowns, the problem is overdetermined and, usually, with no solution. This corresponds to the case where the number of selected reference points is smaller

[2] Non-Euclidean dissimilarities arise naturally when we want to build a measure that incorporates important knowledge about e.g. the relation between objects to be classified. Pseudo-Euclidean embedding allows one to embed such dissimilarities in a vector space in order to use standard (Euclidean) classification tools.

Please cite this article as: A.H. de Souza Júnior et al., Minimal Learning Machine: A novel supervised distance-based approach for regression and classification, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.11.073

than the number of available points (i.e. $K < N$). In this case, we must rely on the approximate solution provided by the usual least squares estimate of $\mathbf{B}$,

$$\hat{\mathbf{B}} = (\mathbf{D}_x'\mathbf{D}_x)^{-1}\mathbf{D}_x'\boldsymbol{\Delta}_y. \tag{3}$$

If in Eq. (1) the number of equations equals the number of unknowns (i.e. $K=N$ because all the learning points are also reference points), then the problem is uniquely determined and has a single solution if the matrix $\mathbf{D}_x$ is full-rank. In this case,

$$\hat{\mathbf{B}} = \mathbf{D}_x^{-1}\boldsymbol{\Delta}_y. \tag{4}$$

Clearly less interesting is the case where in Eq. (1) the number of equations is smaller than the number of unknowns (i.e. for $K > N$, corresponding to the situation where, after selecting the reference points, only a smaller number of learning points is used). This case leads to an underdetermined problem with, usually, infinitely many solutions.

It is worth mentioning that, since the MLM assumes a linear mapping between the distance matrices, other learning (or estimation) technique can be chosen instead of using the ordinary least squares method, such as the least mean squares (LMS) [36] or even the recursive least squares (RLS) [18] algorithms.

Given the possibility for $\mathbf{B}$ to be either uniquely solvable equation (4) or be estimated equation (3), for an input test point $\mathbf{x} \in \mathbb{R}^D$ whose distances from the K reference input points $\{\mathbf{m}_k\}_{k=1}^K$ are collected in the vector $\mathbf{d}(\mathbf{x}, R) = [d(\mathbf{x}, \mathbf{m}_1) \ldots d(\mathbf{x}, \mathbf{m}_K)]$, the corresponding distances between its unknown output $\mathbf{y}$ and the known outputs $\{\mathbf{t}_k\}_{k=1}^K$ of the reference points are

$$\hat{\boldsymbol{\delta}}(\mathbf{y}, T) = \mathbf{d}(\mathbf{x}, R)\hat{\mathbf{B}}. \tag{5}$$

The vector $\hat{\boldsymbol{\delta}}(\mathbf{y}, T) = [\hat{\delta}(\mathbf{y}, \mathbf{t}_1) \cdots \hat{\delta}(\mathbf{y}, \mathbf{t}_K)]$ provides an estimate of the geometrical configuration of $\mathbf{y}$ and the reference set T, in the $\mathcal{Y}$-space.

2.1.2. Output estimation

The problem of estimating the output $\mathbf{y}$, given the outputs $\{\mathbf{t}_k\}_{k=1}^K$ of all the reference points and estimates $\hat{\boldsymbol{\Delta}}(\mathbf{y}, T)$ of their mutual distances, can be understood as a multilateration problem [27] to estimate its location in $\mathcal{Y}$.

Numerous strategies can be used to solve a multilateration problem [26]. From a geometric point of view, locating $\mathbf{y} \in \mathbb{R}^S$ is equivalent to solve the overdetermined set of K nonlinear equations corresponding to S-dimensional hyper-spheres centered in $\mathbf{t}_k$ and passing through $\mathbf{y}$. Fig. 1 graphically depicts the problem for $S=2$.

Given the set of $k = 1, \ldots, K$ spheres each with radius equal to $\hat{\delta}(\mathbf{y}, \mathbf{t}_k)$

$$(\mathbf{y} - \mathbf{t}_k)'(\mathbf{y} - \mathbf{t}_k) = \hat{\delta}^2(\mathbf{y}, \mathbf{t}_k), \tag{6}$$

the location of $\mathbf{y}$ can be estimated from the minimization of the objective function

$$J(\mathbf{y}) = \sum_{k=1}^{K} \left((\mathbf{y} - \mathbf{t}_k)'(\mathbf{y} - \mathbf{t}_k) - \hat{\delta}^2(\mathbf{y}, \mathbf{t}_k) \right)^2. \tag{7}$$

The cost function has a minimum equal to 0 that can be achieved if and only if $\mathbf{y}$ is the solution of Eq. (6). If it exists, such a solution is thus global and unique. Due to the uncertainty introduced by the estimates $\hat{\delta}(\mathbf{y}, \mathbf{t}_k)$, a solution can be achieved by any minimizer $\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmin}}\, J(\mathbf{y})$ like the nonlinear least-squares estimates from standard gradient descent methods. In the following, the Levenberg–Marquardt (LM) method [23] is preferred.

2.2. Parameters and computational complexity

On the basis of the aforementioned overview, the number of reference points K is virtually the only hyper-parameter that the user needs to select in order to optimize a Minimal Learning Machine. As always, a selection based on conventional validation or on standard resampling methods for cross-validation could be adopted for the task [17].

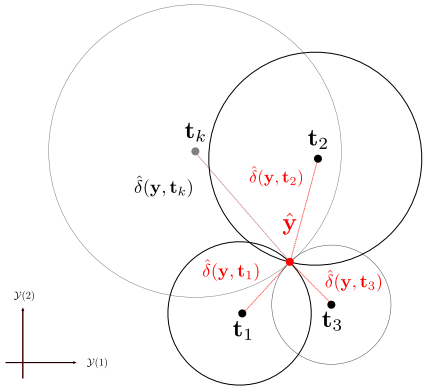

Fig. 1. Output estimation.

Two figures of merit can be considered for selecting K, the Average Mean Squared Error for the output distances ($AMSE(\boldsymbol{\delta})$) and the Average Mean Squared Error for the responses $AMSE(\mathbf{y})$:

$$AMSE(\boldsymbol{\delta}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_v}\sum_{i=1}^{N_v}(\delta(\mathbf{y}_i, \mathbf{t}_k) - \hat{\delta}(\mathbf{y}_i, \mathbf{t}_k))^2 \tag{8}$$

$$AMSE(\mathbf{y}) = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N_v}\sum_{i=1}^{N_v}(y_i^{(s)} - \hat{y}_i^{(s)})^2 \tag{9}$$

For a set of N_v validation points $(\mathbf{x}_i, \mathbf{y}_i)$, the $AMSE(\boldsymbol{\delta})$ quantifies how well the distances $\delta(\mathbf{y}_i, \mathbf{t}_k)$ between the N_v output responses $\mathbf{y}_i$ and the outputs of the K selected reference points $\mathbf{t}_k$ are estimated $\hat{\delta}(\mathbf{y}_i, \mathbf{t}_k)$. The $AMSE(\mathbf{y})$, in turn, is only performed after both the distance regression and the estimation steps of the MLM are completed and, thus, it quantifies how well the S-dimensional outputs $\mathbf{y}_i = [y_i^{(1)} \ldots y_i^{(S)}]$ are estimated $\hat{\mathbf{y}}_i = [\hat{y}_i^{(1)} \ldots \hat{y}_i^{(S)}]$. In the case of univariate responses ($S=1$) the $AMSE(\mathbf{y})$ reduces to the conventional Mean Square Error for the outputs ($MSE(y)$).

Ideally, if we had enough data, we would set aside a validation set and use it to assess the performance of our model trained with a varying number K of reference points. The value of K that optimizes the chosen figure of merit is then used to learn the final model with all the data. However, since the data are often scarce, this is usually impossible and we need to resort to cross-validation. We split the data into a number F of roughly equal-sized parts and for the fth part, we train the model with a varying number K of reference points on data from the other $F-1$ parts and we calculate the corresponding figure of merit. This is repeated for $f = 1, 2, \ldots, F$ and the resulting figure of merit for the same value of K averaged over all the F folds. Again, the value of K that minimizes the chosen figure of merit is then used to learn the final model with all the data. The case $F=N$ is known as leave-one-out cross-validation.

The training procedure of the Minimal Learning Machine is sketched in Algorithm 1. The training computation can be roughly divided into two parts: (i) calculation of the pairwise distance matrices in the output and input space; (ii) calculation of the least-square solution for the multiresponse linear regression problem on distance matrices. The first part takes $\Theta(KN)$ time (see [5] for a review of algorithmic asymptotic analysis). The computational cost of the second part is driven by the calculation of the Moore–Penrose pseudoinverse matrix.

Algorithm 1. MLM training procedure.

Input: Training data sets X and Y, and K.

Please cite this article as: A.H. de Souza Júnior et al., Minimal Learning Machine: A novel supervised distance-based approach for regression and classification, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.11.073

Output: $\hat{\mathbf{B}}$, R and T.

1. Randomly select K reference points, R, from X and their corresponding outputs, T, from Y;
2. Compute $\mathbf{D}_x$: The distance matrix between X and R;
3. Compute $\mathbf{\Delta}_y$: The distance matrix between Y and T;
4. Calculate $\hat{\mathbf{B}} = (\mathbf{D}_x'\mathbf{D}_x)^{-1}\mathbf{D}_x'\mathbf{\Delta}_y$.

One of the most used methods for the calculation of Moore–Penrose pseudoinverses is the SVD [10], which runs in $\Theta(K^{2N})$ time. This method is very accurate but its drawback is that relies on computational time constants that make it time-intensive. In order to speed up the computation, several methods have been proposed (for example, see [21,7]). In [21], the computation is optimized by using a special type of tensor product and QR factorization, whereas the method proposed in [7] is based on a full-rank Cholesky decomposition. Even if such approaches improve significantly the computational time of computing the Moore–Penrose inverse matrix, the asymptotic time complexity is still equal to that provided by the SVD method. Even though, one might consider them for large datasets and real-time applications.

The time complexity of the MLM training phase is driven by the computation of the Moore–Penrose matrix and then it is given by $\Theta(K^2N)$. In order to establish a comparison, the MLM training computational cost is similar to what is presented by an extreme learning machine when the number of hidden neurons is equal to the number of reference points. It is worthy to note that the ELM is considered one of the fastest methods for nonlinear regression and classification tasks [25].

Algorithm 2. MLM test procedure.

Input: $\hat{\mathbf{B}}$, R, T and $\mathbf{x}$.
Output: $\hat{\mathbf{y}}$.

1. Compute $\mathbf{d}(\mathbf{x}, R)$;
2. Compute $\hat{\boldsymbol{\delta}}(\mathbf{y}, T) = \mathbf{d}(\mathbf{x}, R)\hat{\mathbf{B}}$;
3. Use T and $\hat{\boldsymbol{\delta}}(\mathbf{y}, T)$ to find an estimate for $\mathbf{y}$. This can be accomplished by any gradient descent algorithm over the cost function in Eq. (7).

Concerning the computational analysis of the generalization step in MLM (Algorithm 2), we consider the Levenberg–Marquardt method due to its fast and stable convergence, even though any gradient descent method can be used to minimize the objective function given in Eq. (7). For each iteration, the LM method involves the computation of the Jacobian matrix $\mathbf{J} \in \mathbb{R}^{K \times S}$ and the inverse of $\mathbf{J}^T\mathbf{J}$. In this regard, the computational complexity of the LM algorithm is about $\Theta(I(KS^2 + S^3))$, where S is the dimensionality of $\mathbf{y}$ and I denotes the number of iterations. In most of the regression and classification problems, S is a small number and then the complexity turns to be proportional to the number of reference points and the number of iterations. This is slightly worse than what is presented, for instance, by SVM methods with K support vectors, which is linear in K with small constant factor. Also, the ELM testing phase runs in $\Theta(KS)$ time, with K corresponding to the number of hidden units.

2.3. The Minimal Learning Machine for classification

An important class of problems is classification, where we are concerned with predicting categories usually denoted by qualitative outputs, also called class labels. For the task, we are still given a set of N input points $X = \{\mathbf{x}_i\}_{i=1}^N$, with $\mathbf{x}_i \in \mathbb{R}^D$, and the set of their corresponding class labels $L = \{l_i\}_{i=1}^N$, with $l_i \in \{C_1, \ldots, C_S\}$, where C_j denotes the jth class; for $S=2$, the problem is referred to as binary classification, whereas for $S > 2$ we have multi-class applications.

The Minimal Learning Machine can be extended to classification problems in a straightforward manner by representing the S class labels in a vectorial fashion through a 1-of-S encoding scheme. In such an approach, an S-level qualitative variable is represented by a vector of S binary variables or bits, only one of which is *on* at a time. Mathematically, the set of outputs $Y = \{\mathbf{y}_i\}_{i=1}^N$, with $\mathbf{y}_i \in \mathbb{R}^S$, that corresponds to the input points X is then defined in such a way that the jth component of $\mathbf{y}_i$ is set to α if $l_i = C_j$ and β otherwise, where α and β are integer scalars such as $\alpha > \beta$. An usual choice is $\alpha = 1$ and $\beta = -1$.

In classification of a test observation $\mathbf{x}$ with unknown class label $l \in \{C_1, \ldots, C_S\}$, the estimated class $\hat{l}$ associated to the output estimate $\hat{\mathbf{y}}$ is given by $\hat{l} = C_{s^*}$, where

$$s^* = \underset{s=1,\ldots,S.}{\operatorname{argmax}}\{\hat{y}^{(s)}\}. \tag{10}$$

As one can easily notice, for binary classification problems, we may simplify the approach by using a binary single output scheme where the outputs are represented by scalars $y_i \in \{\alpha, \beta\}$ in correspondence to the two classes.

Given this formulation, the Minimal Learning Machine provides unified implementation for regression, binary and multiclass applications.

2.4. Related work

Weston et al. [35] introduced the kernel dependency estimation (KDE) approach, which is a kernel based method for learning a dependency between two classes of objects – one class being the input and the other class the corresponding output. The objects can be defined in terms of vectors, images, strings, trees or graphs. In a like manner to MLM, KDE learning employs similarity measures in both input and output spaces, and it also requires an output estimation, which is in this case called pre-image problem. There are, however, notable differences between the MLM and KDE. Firstly, the KDE is a kernel method and, as such, it relies on kernel functions in order to embed the objects into vector spaces, whereas any dissimilarity or proximity measure can be used in MLM. Secondly, the KDE requires a PCA step to be applied to the kernel output similarity matrix. In MLM, one uses the distance matrix directly in order to learn the input–output mapping. Cortes et al. [6] proposed the Regression for Learning Transductions (RLT) as an alternative to KDE in the learning of string–string mappings. In RLT, the authors simplify the KDE framework by not requiring the prior dimensionality reduction step. Thirdly, the pre-image problem (i.e. output estimation) in KDE requires an exhaustive pre-image search, where the closest output sample is chosen from the available training set. In MLM, an optimization procedure is carried out as described in Algorithm 2. Finally, the KDE was evaluated mainly on classification problems, while the MLM is evaluated comprehensively in both regression and classification tasks.

2.5. Two illustrative examples

In this section, we illustrate the Minimal Learning Machine using two synthetic problems – one for regression and one for classification. The regression problem consists in the estimation of a smoothed and nonlinear version of the parity function. The classification problem consists in the estimation of the nonlinearly separable and nonconvex classes of the Tai Chi symbol.

2.5.1. The smoothed parity

To illustrate the behavior of the Minimal Learning Machine for regression, we generated 2^{13} bi-dimensional input points uniformly distributed in the unit-square, $\mathbf{x} \in [0, 1]^2$, and built the

Please cite this article as: A.H. de Souza Júnior et al., Minimal Learning Machine: A novel supervised distance-based approach for regression and classification, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.11.073

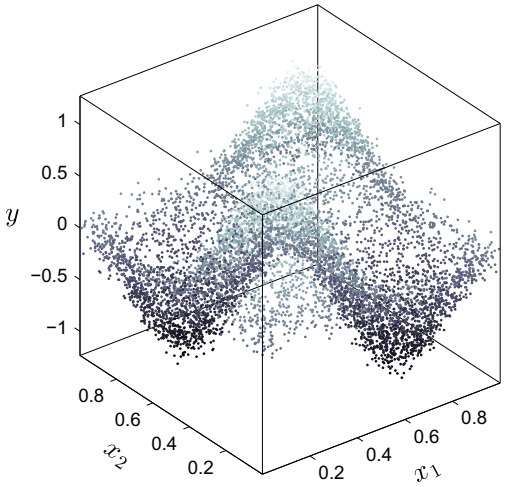

Fig. 2. The *smoothed* parity function: data.

response using the smoothed parity function model $y = f(\mathbf{x}) + \varepsilon$, where $f = \sin(2\pi x_1)\sin(2\pi x_2)$ and $\varepsilon \sim \mathcal{N}(0, 0.1^2)$, Fig. 2.

We analyzed the performance of the MLM for N learning points ranging from 2^1 to 2^{12} and K randomly selected reference points, such that $K \leq N$. An independent set of $N_v = 2^{12}$ points is used for validating the MLM in terms of its hyper-parameter K. Two figures of merit are used – the $AMSE(\boldsymbol{\delta})$ for the output distances and the $MSE(\mathbf{y})$ for the single response. In our experiments, the Minimal Learning Machines are trained with all the different N-sized learning sets and for all possible numbers of reference points. For each size N of the learning set and for a varying number K of reference points, the figures of merit of the distance regression and the output estimation steps are then evaluated on the validation set (Fig. 3).

Fig. 3(a) shows the $AMSE(\boldsymbol{\delta})$ performance of the Minimal Learning Machine in the distance estimation step, for different combinations of N learning points and K reference points. For a given number N of learning points, the number of reference points that leads to the best performances is denoted by a red dot and, then, a red circle is used to denote the MLM with the overall smallest $AMSE(\boldsymbol{\delta})$. Related to that, Fig. 3(b) illustrates the $AMSE(\boldsymbol{\delta})$ achieved by the best performing MLMs for different sizes of the learning set. Analogously, for the output estimation step, Fig. 3(c) shows the $MSE(\mathbf{y})$ performances, where again the best performing MLMs are denoted by red dots, and a red circle denotes the best MLM overall. By the same token, Fig. 3(d) shows the performance of the best MLMs for different sizes of the learning set. Based on the results depicted in Fig. 3(b) and (d), it is possible to observe that the MLM performance improves as the number N of learning points increases; an expected result. On the other hand, the optimal number K of reference points does not necessarily grow at the same rate of the number of learning points. After some point, including more reference points decreases the generalization performances of the MLM, both in terms of distance regression and output estimation, a clear indication that the MLM starts overfitting. Considering the nature of the sets of equations that characterize the two steps of the MLM learning, this is again an expected result, further confirmed by experimental evidence.

The best Minimal Learning Machine overall was found to be the one trained using $N = 2^{12}$ learning points and $K = 2^8$ reference points. It is worth noting that for both the distance regression and the output estimation step, the optimal number of reference points is found to be the same. Fig. 4 illustrates the validation results when estimating the response with the best performing MLM. Interestingly, the $MSE(\mathbf{y})$ achieved by this MLM is equal to 0.011, which tends to the variance of the additive noise in the response ($\mathrm{Var}(\varepsilon) = 0.010$) and thus also to the smallest Mean Square Error that any regression model can achieve without over-fitting.

2.5.2. The Tai Chi

We illustrate the Minimal Learning Machine on a binary classification problem, the Tai Chi symbol, where the Yin and Yang regions are the two non-convex and nonlinearly separable classes. For the task, we generated 2^{13} bidimensional input points uniformly distributed in the Tai Chi symbol. After assigning the class labels to the Yin and Yang regions, we purposely mislabeled 10% of the observations, Fig. 5. Half of the dataset is used for training the Minimal Learning Machines with a varying number N of learning points and a number K of randomly selected reference points, again with always $K \leq N$. The 2^{12} remaining samples are used for validation.

We evaluated the performance of the Minimal Learning Machine on the validation set using the $AMSE(\boldsymbol{\delta})$ and the $AMSE(\mathbf{y})$, Fig. 6. Fig. 6(a) and (c) depicts with red dots the configuration of the best MLMs for given sizes of the learning set and the circle is used to depict the best model overall. As expected, for each size N of the learning set it is again possible to select an optimal number K of reference points that minimizes the validation error, from the point of view of both the distance regression and the output estimation step. The overall best configuration of the MLM is found to be the one that is based on the largest number of learning point ($N = 2^{12}$) and a number of reference points equal to $K = 2^8$.

With respect to the estimation step, Fig. 7 shows the estimated classes in validation using the overall best MLM configuration. The classification accuracy achieved by this MLM is equal to 88%, which tends again to the percentage of purposely mislabeled data.

3. Experiments

In this section, we present the results achieved by the Minimal Learning Machine on twelve real-world datasets commonly used for benchmarking purposes in regression and classification. The performance of the MLM is then compared to what is achieved by five other reference methods: the extreme learning machine (ELM, [20]), the radial basis function network (RBF, [2,38]), the support vector machine (SVM, [32]), Gaussian processes (GP, [29]) and the multilayer perceptron (MLP, [1]).

The datasets are available from the University of California at Irvine (UCI) Repository (www.ics.uci.edu/~mlearn/) and a description of the datasets is summarized in Table 1. The datasets have been chosen to object heterogeneity in the number of samples and inputs. All the datasets have been preprocessed in the same way. Categorical variables have been removed as well as samples containing missing values. Ten different random permutations of the whole datasets are taken, and two thirds are used to create the training set and the remaining for the test set. Then, the training set is normalized to zero mean and unit variance, and the test set is normalized using the same mean and variance from the training set. It may also be noticed that the proportions of the classes, for classification cases, have been kept balanced: each class is represented in an equal proportion, in both training and test sets.

The hyper-parameters for the SVM and the MLP are selected using 10-fold cross-validation. The SVM is learned using the SVM toolbox [4] with default settings for the hyper-parameters and grid search: the grid is logarithmic between 2^{-2} and 2^{10} for each hyper-parameter; nu-SVC has been used for classification and epsilon-SVR for regression, with radial basis function kernel. The MLP is trained using Levenberg–Marquardt optimization and a range of hidden units from 1 to 20. The learning of GP is based on the default settings in the Matlab Toolbox [29]. The ELM network uses sigmoid kernel

Please cite this article as: A.H. de Souza Júnior et al., Minimal Learning Machine: A novel supervised distance-based approach for regression and classification, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.11.073

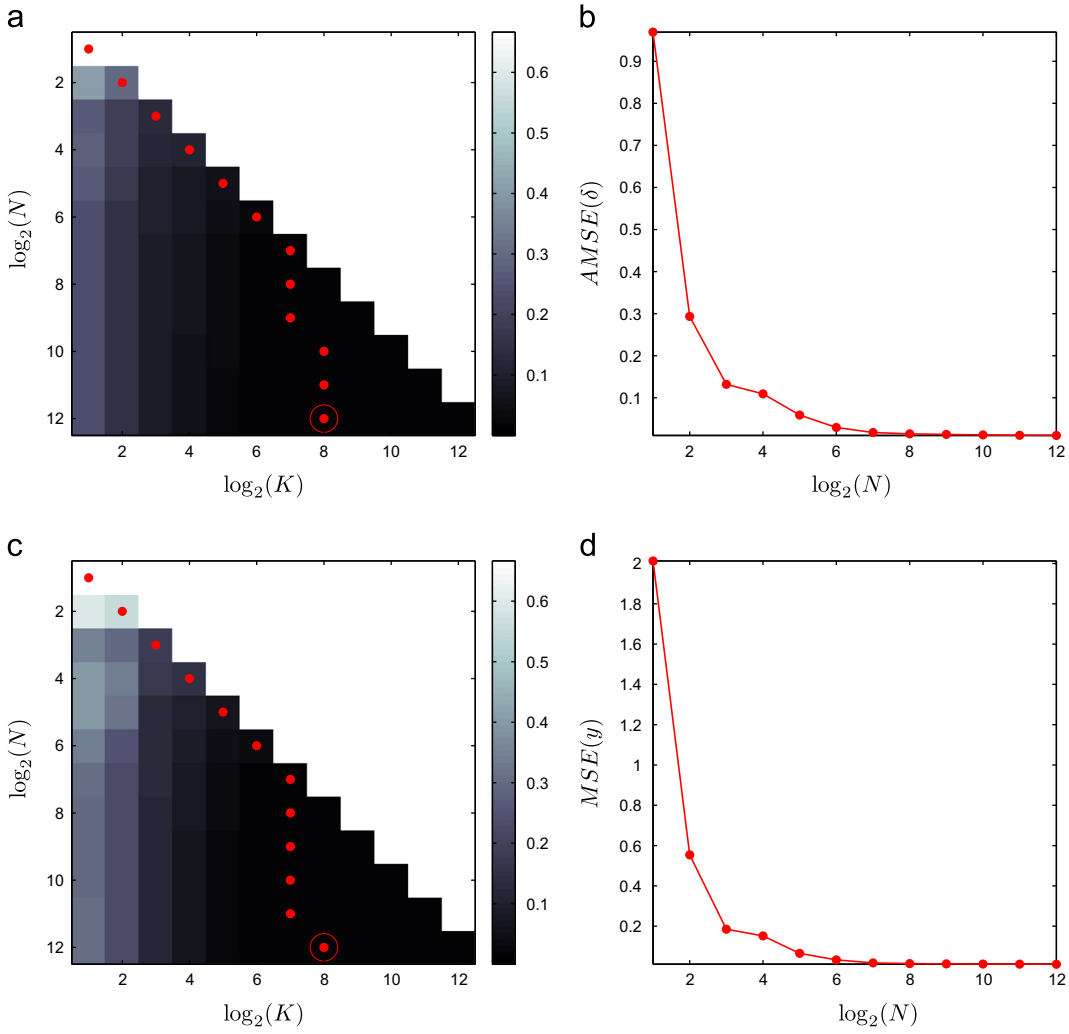

Fig. 3. The *smoothed* parity function: figures of merit. (a) *AMSE*(δ). (b) Optimal *AMSE*(δ). (c) *MSE*(**y**). (d) Optimal *MSE*(**y**). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

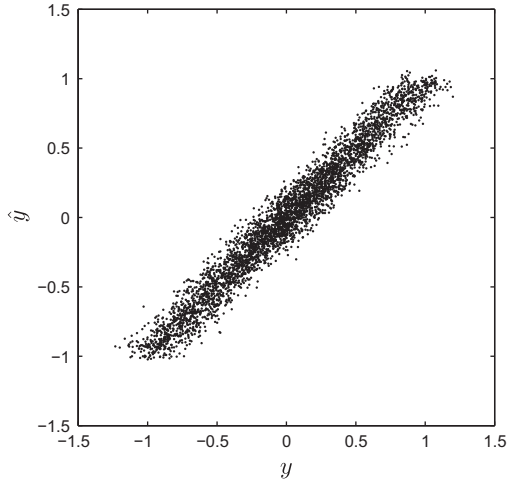

Fig. 4. The *smoothed* parity function: output estimation with $N = 2^{12}$ and $K = 2^8$.

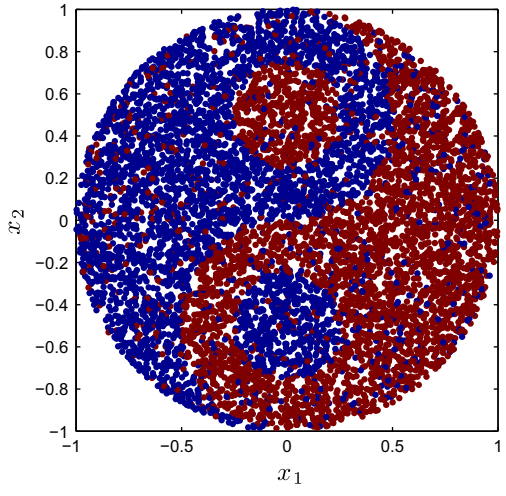

Fig. 5. The Tai Chi symbol: data.

and it has been validated with the number of hidden units ranging from 10 to 100 with increments of 10. For the experiments with the RBF network, the centers of the Gaussian basis functions are selected by the *k*-means algorithm, with the number of centers varying from 5% to 50% (step size of 5%) of the number of learning points. We also applied 10-fold cross-validation to select the optimal number of centers. The only hyper-parameter of the Minimal Learning Machine, the number of reference points *K*, has been optimized using 10-fold cross-validation with reference points randomly selected in a range of 5–100% (with a step size of 5%) of the available training samples.

Please cite this article as: A.H. de Souza Júnior et al., Minimal Learning Machine: A novel supervised distance-based approach for regression and classification, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.11.073

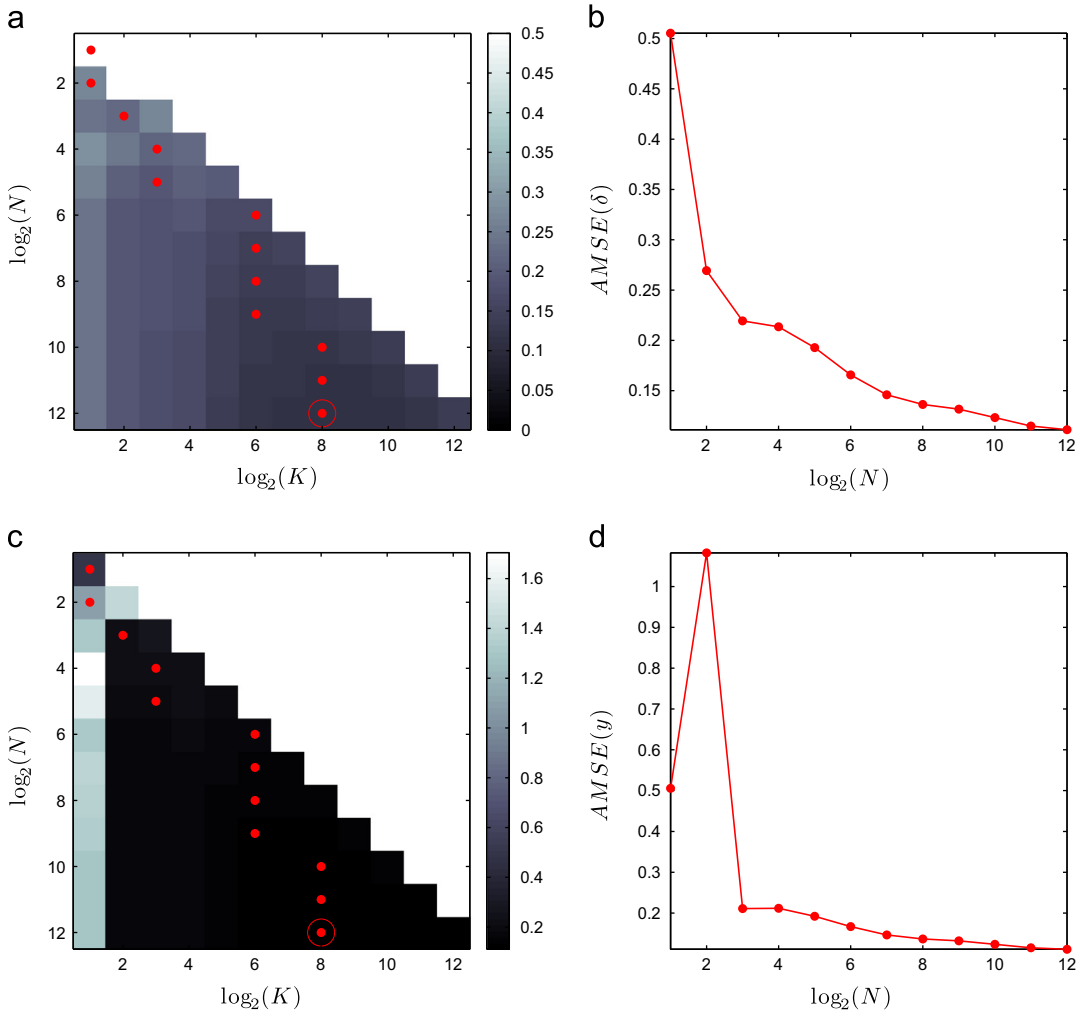

Fig. 6. The Tai Chi symbol: figures of merit. (a) *AMSE*(δ). (b) Optimal *AMSE*(δ). (c) *AMSE*(**y**). (d) Optimal *AMSE*(**y**). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

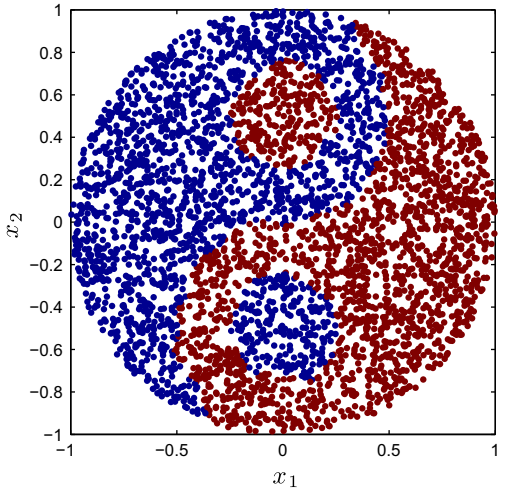

Fig. 7. The Tai Chi symbol: output estimation, with $N = 2^{12}$ and $K = 2^8$.

3.1. Results on regression

All the models are evaluated using the mean and standard deviation of the resulting MSE over 10 independently drawn test sets. We also carried out a statistical evaluation of the MLM performance against those achieved by the other models using the Wilcoxon signed-ranks test [37] with significance level equal to 5%.

Table 1
Description of the datasets: type (regression or classification), input/output dimensionality and number of training/test samples.

Dataset	Type	Dim		# Samples	
		In	Out	Train	Test
Ailerons	R	5	1	4752	2377
Elevators	R	6	1	6344	3173
Breast Cancer	R	32	1	129	65
Boston Housing	R	13	1	337	169
Servo	R	4	1	111	56
Abalone	R	8	1	2784	1393
Stocks	R	9	1	633	317
Auto Price	R	15	1	106	53
Wisconsin Breast Cancer	C	30	2	379	190
Pima Indians Diabetes	C	8	2	512	256
Iris	C	4	3	100	50
Wine	C	13	3	118	60

The null hypothesis is that the difference between MSE values comes from a distribution with zero median. In our case, we compare the MLM performance against each other method for each dataset.

On the basis of the experimental results reported in Table 2, we can observe that the state-of-the-art models seem to be able to achieve similar accuracies. In this regard, the MLM also achieves performances that are comparable to such methods. The MLM presents the smallest MSE (average) for five out of eight regression problems. Also, even

Please cite this article as: A.H. de Souza Júnior et al., Minimal Learning Machine: A novel supervised distance-based approach for regression and classification, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.11.073

Table 2
Test results: MSE, standard deviations (below the MSE) and Wilcoxon signed-ranks test results (✓: fail to reject, and ×: reject). The best performing models are in boldface.

Datasets	Models					
	MLM	ELM	RBF	SVM	GP	MLP
Ailerons	**2.7**e−**8** 1.6e−9	2.9e−8 1.5e−9 ×	3.0e−8 1.8e−9 ×	1.3e−7 2.6e−8 ×	**2.7**e−**8** 1.9e−9 ✓	2.7e−7 4.4e−9 ×
Elevators	**2.0**e−**6** 6.1e−8	2.1e−6 5.5e−8 ✓	2.1e−6 6.8e−8 ×	6.2e−6 6.8e−7 ×	**2.0**e−**6** 5.0e−8 ×	2.6e−6 9.0e−8 ✓
Breast Cancer	**1.1**e+**3** 1.8e+2	1.2e+3 1.4e+2 ✓	1.2e+3 1.8e+2 ✓	1.2e+3 7.2e+1 ✓	1.3e+3 1.9e+2 ×	1.5e+3 4.4e+2 ×
Boston	1.9e+1 9.0	2.2e+1 7.1 ✓	2.0e+1 6.6 ✓	3.4e+1 3.1e+1 ✓	**1.1**e+**1** 3.5 ×	2.2e+1 8.8 ✓
Servo	**4.6**e−**1** 3.0e−1	7.0 e−1 2.5 e−1 ×	6.1e−1 3.4e−1 ×	6.9e−1 3.2e−1 ×	4.8e−1 3.5e−1 ✓	6.0e−1 3.2e−1 ✓
Abalone	4.7 3.3e−1	4.6 2.4e−1	4.7 2.3e−1	**4.5** 2.7e−1	**4.5** 2.4e−1	4.6 5.0e−1
Stocks	**4.1**e−**1** 5.8e−2	9.0e−1 7.3e−2 ×	7.1e−1 2.0e−1 ×	5.1e−1 9.8e−2 ×	4.4e−1 5.0e−2 ✓	8.8e−1 2.1e−1 ×
Auto Price	2.6e+7 2.7e+7	1.3e+7 4.1e+6 ✓	1.1e+7 5.4e+6 ✓	9.8e+7 8.4e+6 ✓	2.0e+7 1.0e+7 ✓	**1.0**e+**7** 3.9e+6 ✓

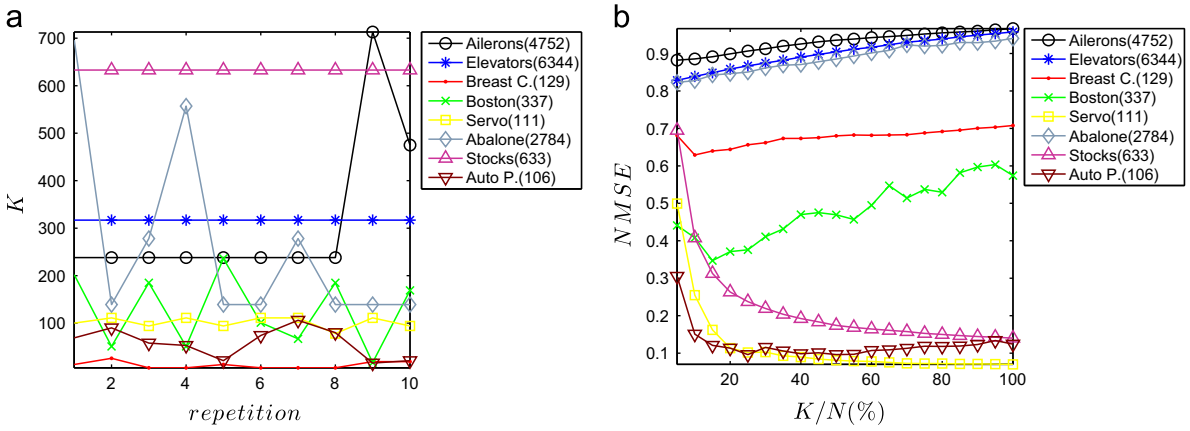

Fig. 8. MLM validation performance on regression. Legends also contain the total number of training samples. (a) Optimal K value over different runs. (b) Error per number of reference points.

though we have applied random selection of reference points, the standard deviation of MSEs reported by MLM is rather similar to the smallest values achieved by the state-of-the-art methods. The hypothesis tests have shown that the MLM presents statistical difference to the other methods. For the Stocks and Ailerons datasets, the MLM was the best performing model and the null hypothesis cannot be rejected only for the GP model. In contrast, the MLM results are not statistically distinct from the other methods for the Auto Price and Abalone datasets – cases in which the MLM is not the best performing model.

As noticed, the MLM performance is quite similar to the state-of-the-art methods. Thus, the computational complexity takes an important role in the decision making process of selecting the most appropriate method. In this regard, an essential aspect for fast MLM training is the number of reference points, or more specifically, the property that the optimal number of reference points does not grow at the same rate of the number of learning points (dataset size). In order to illustrate such a property, we report in Fig. 8 the results of the cross-validation phase in terms of selecting K for all the 10 independent runs. Fig. 8(a) shows that the rate between the optimal number of reference points per number of learning points is almost equal to one for small datasets (up to about 600 samples), and it is stable number. Concerning large datasets (Abalone dataset), our experiments have shown that it is not needed as many reference points as learning points and similar behavior is shown in the synthetic regression example in Section 2.5. Fig. 8(b) illustrates how the validation errors (Normalized Mean Squared Error, NMSE[3]) change as a function of the proportion of reference points. Based on Fig. 8(b), using 20% of the learning points as reference points seems to be a good choice for most datasets.

[3] The mean squared errors have been normalized to be between 0 and 1.

Table 3
Test performance: accuracies (%), the corresponding standard deviations and Wilcoxon signed-ranks test results (✓: fail to reject, ×: reject). For each dataset, the best performing models are in boldface.

Datasets	Models					
	MLM	ELM	RBF	SVM	GP	MLP
Wisconsin B.C.	**97.7**	95.7	95.6	91.6	97.3	96.6
	0.6	1.2	1.4	1.7	0.9	1.9
		×	×	×	✓	×
Pima I.D.	74.2	74.6	74.5	72.7	**76.3**	75.2
	1.7	2.1	2.3	1.5	**1.8**	1.9
		✓	✓	✓	×	✓
Iris	95.0	96.0	**96.4**	95.4	95.6	94.8
	1.4	2.3	**1.8**	1.9	2.3	3.8
		✓	×	✓	✓	✓
Wine	**99.0**	97.2	98.0	95.8	96.2	96.0
	1.2	3.5	2.0	2.9	2.1	2.4
		✓	×	×	×	×

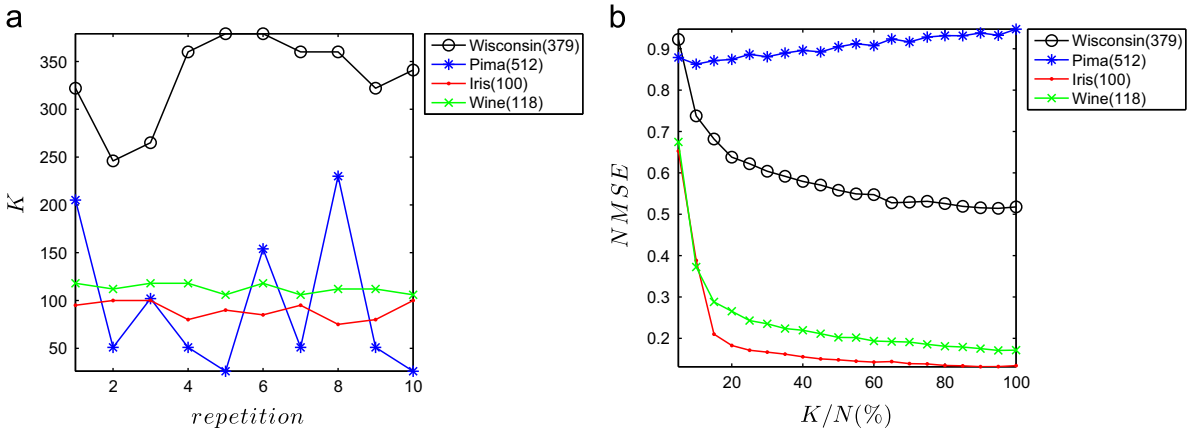

Fig. 9. MLM validation performance on classification. Legends also contain the total number of training samples. (a) Optimal K value over different runs. (b) Error per number of reference points.

3.2. Results on classification

In order to evaluate the MLM performance for classification problems, we use the mean success classification rate and the corresponding standard deviations. To assess statistical significance for the results achieved by the MLM in comparison to the other methods, we carried out the nonparametric Wilcoxon signed-ranks hypothesis test. Again, the null hypothesis is that the difference between MSE values comes from a distribution with zero median. The results are reported in Table 3.

From Table 3 we can observe that the MLM exhibits an equivalent or even better generalization performance in comparison to the other models. Moreover, the MLM has shown a stable performance since its standard deviations are smaller than those of the other methods, specially on the Wisconsin B.C. and Wine datasets. In opposite, the SVM model presented the worst performances overall. Based on the hypothesis tests, the MLM achieves performances that are statistically different from the other methods, particularly for the Wine and Wisconsin B.C. datasets. With regard to Iris and Pima I.D. sets, the MLM provides results mostly equivalent to the state-of-the-art methods.

We report in Fig. 9 the MLM performance during the validation procedure. From Fig. 9, one may notice that it is not needed as many reference points as learning points, particularly for large datasets, e.g. the Pima I.D. dataset. Although the optimal number of reference points is about 100% of the learning points ($K=N$) for Wine and Iris datasets (Fig. 9(a)), we observe that, again, 20–40% of the number of learning points has provided a good threshold for selecting K, where the error measures (NMSE) stabilize (Fig. 9(b)), and then increasing K does not reduce the error considerably.

4. Conclusions

This work presents a new supervised learning method, the Minimal Learning Machine or MLM. Learning an MLM consists in reconstructing the mapping existing between input and output distance matrices and then exploiting the geometrical arrangement of the output points for estimating the response. Based on our experiments, a multiresponse linear regression model is capable of reconstructing the mapping existing between the aforementioned distance matrices. Given its general formulation, the Minimal Learning Machine is also inherently capable of operating on multi-dimensional responses and it can be extended to classification problems in a straightforward fashion.

A significant advantage of the MLM over other supervised learning methods is that the MLM has only one hyper-parameter to be optimized using standard resampling methods, like leave-one-out cross-validation. The computational complexity of the MLM training procedure is very low, competing with the fastest machine learning approaches. Regarding the test procedure, we have reported that combining a random selection of reference points from the training set and Levenberg–Marquardt optimization allows us to achieve state-of-the-art performance.

On a large number of synthetic and real-world problems, the Minimal Learning Machine has achieved accuracies that are comparable to what is obtained using state-of-the-art classification and regression methods. We have reported the performances on twelve datasets from the UCI Repository and comparisons with five reference approaches. The results highlight the potentiality of the MLM on supervised learning tasks.

Please cite this article as: A.H. de Souza Júnior et al., Minimal Learning Machine: A novel supervised distance-based approach for regression and classification, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.11.073

Acknowledgments

The authors would like to thank the financial support received from the Brazilian Agency of Post-Graduate Studies (CAPES) under the Grant no. 9147-12-8.

References

[1] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Inc., New York, NY, 1995.
[2] M.D. Buhmann, Radial Basis Functions, Cambridge University Press, Cambridge, UK, 2003.
[3] H. Bunke, Structural and syntactic pattern recognition, Handbook of Pattern Recognition & Computer Vision, World Scientific, River Edge, NJ, USA (1993) 163–209.
[4] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans., Intell. Syst. Technol. 2 (27) (2011) 1–27.
[5] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms, 3rd edition, The MIT Press, Cambridge, MA, USA, 2009.
[6] Cortes, C., Mohri, M., Weston, J., 2005. A general regression technique for learning transductions, in: Proceedings of the 22Nd International Conference on Machine Learning. ICML '05, 2005, pp. 153–160.
[7] P. Courrieu, Fast computation of Moore–Penrose inverse matrices, Neural Inf. Process. Lett. Rev. 8 (2005) 25–29.
[8] C. Cuadras, C. Arenas, A distance based regression model for prediction with mixed data, Commun. Stat.—Theory Methods 19 (6) (1990) 2261–2279.
[9] A. Gisbrecht, B. Mokbel, F.-M. Schleif, X. Zhu, B. Hammer, Linear time relational prototype based learning, Int. J. Neural Syst. 22 (5) (2012) 1–11.
[10] G.H. Golub, C.F.V. Loan, Matrix Computations, 3rd edition, Johns Hopkins University Press, Baltimore, MD, USA, 1996.
[11] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, K. Obermayer, Classification on pairwise proximity data, in: M.I. Jordan, M.J. Kearns, S.A. Solla (Eds.), NIPS Proceedings, vol. 11, MIT Press, Cambridge, MA, 1999, pp. 438–444.
[12] T. Graepel, K. Obermayer, A stochastic self-organizing map for proximity data, Neural Comput. 11 (1) (1999) 139–155.
[13] M. Hagenbuchner, A. Sperduti, A.C. Tsoi, A self-organizing map for adaptive processing of structured data, IEEE Trans. Neural Netw. 14 (3) (2003) 491–505.
[14] B. Hammer, A. Hasenfuss, Topographic mapping of large dissimilarity data sets, Neural Comput. 22 (9) (2010) 2229–2284.
[15] B. Hammer, D. Hofmann, F.-M. Schleif, X. Zhu, Learning vector quantization for (dis-)similarities, Neurocomputing 131 (2014) 43–51.
[16] B. Hammer, A. Micheli, M. Strickert, A. Sperduti, A general framework for unsupervised processing of structured data, Neurocomputing 57 (2004) 3–35.
[17] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining Inference, and Prediction, 2nd edition, Springer, New York, NY, USA, 2009.
[18] S. Haykin, Adaptive Filter Theory, 4th edition, Prentice-Hall, New Jersey, NJ, 2001.
[19] T. Hofmann, J. Buhmann, Pairwise data clustering by deterministic annealing, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1) (1997) 1–14.
[20] G.B. Huang, Q.Y. Zhu, C.K. Ziew, Extreme learning machine: theory and applications, Neurocomputing 70 (1–3) (2006) 489–501.
[21] V.N. Katsikis, D. Pappas, A. Petralias, An improved method for the computation of the Moore–Penrose inverse matrix, Appl. Math. Comput. 217 (2011) 9828–9834.
[22] J.W. Lichstein, Multiple regression on distance matrices: a multivariate spatial analysis tool, Plant Ecol. 188 (2) (2006) 117–131.
[23] D.W. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, J. Soc. Ind. Appl. Math. 11 (2) (1963) 431–441.
[24] B. McArdle, M. Anderson, Fitting multivariate models to community data: a comment on distance-based redundancy analysis, Ecology 82 (1) (2001) 290–297.
[25] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, A. Lendasse, OP-ELM: optimally pruned extreme learning machine, IEEE Trans. Neural Netw. 21 (1) (2010) 158–162.
[26] W. Navidi Jr., W.S.M. Hereman, W, Statistical methods in surveying by trilateration, Comput. Stat. Data Anal. 27 (2) (1998) 209–227.
[27] E. Niewiadomska-Szynkiewicz, M. Marks, Optimization schemes for wireless sensor network localization, Int. J. Appl. Math. Comput. Sci. 19 (2) (2009) 291–302.
[28] A.K. Qin, P.N. Suganthan, A novel kernel prototype-based learning algorithm, in: Proceedings of the 17th International Conference on Pattern Recognition (ICPR'2004), vol. 4, 2004, pp. 621–624.
[29] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, Cambridge, MA, 2006.
[30] R.J. Schalkoff, Pattern Recognition Statistical, Structural and Neural Approaches, John Wiley & Sons, New York, NY, USA, 1992.
[31] F.-M. Schleif, T. Villmann, B. Hammer, P. Schneider, Efficient kernelized prototype based classification, Int. J. Neural Syst. 21 (6) (2012) 443–457.
[32] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (2004) 199–222.
[33] P. Somervuo, T. Kohonen, Self-organizing maps and learning vector quantization for feature sequences, Neural Process. Lett. 10 (2) (1999) 151–159.
[34] J. Wang, Geometric Structure of High-Dimensional Data and Dimensionality Reduction, Springer, Berlin, Germany, 2011.
[35] J. Weston, O. Chapelle, A. Elisseeff, B. Schölkopf, V. Vapnik, Kernel dependency estimation, in: S. Becker, S. Thrun, K. Obermayer (Eds.), NIPS ProceedingsMIT Press, Cambridge, MA, 2002, pp. 873–880.
[36] B. Widrow, Thinking about thinking: the discovery of the LMS algorithm, IEEE Signal Process. Mag. 22 (1) (2005) 100–106.
[37] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (6) (1945) 80–83.
[38] Y. Wu, H. Wang, B. Zhang, K.-L. Du, Using radial basis function networks for function approximation and classification, ISRN Appl. Math. 2012 (2012) 1–34 (ID324194), 1–34. URL http://dx.doi.org/10.5402/2012/324194.
[39] X. Zhu, F.-M. Schleif, B. Hammer, Adaptive conformal semi-supervised vector quantization for dissimilarity data, Pattern Recognit. Lett. 49 (2014) 138–145.

Amauri Holanda de Souza Junior received his Ph.D. and M.Sc. degrees in Teleinformatics Engineering from the Federal University of Ceará in 2014 and 2009, respectively, and his B.Sc. degree in Teleinformatics from the Federal Institute of Education, Science and Technology of Ceará in 2007. He is a professor of the Department of Computer Science at the Federal Institute of Education, Science and Technology of Ceará, Brazil. His main research interests are signal processing, nonlinear system identification, machine learning and computer programming.

Francesco Corona is a Docent in Information and Computer Science at the Department of Computer Science of the Aalto University (Finland) and a senior researcher at Department of Teleinformatics Engineering of the Federal University of Ceará (Brazil). He received the Laurea degree in Chemical Engineering and the Ph.D. degree in Industrial Engineering from the University of Cagliari (Italy). He joined the Aalto University/Helsinki University of Technology in 2007 and the Federal University of Ceará in 2014. His interest concentrates on Statistical Machine Learning and Information Visualization technologies, with application to full-scale process modelling, visualization, control and optimization.

Guilherme Barreto was born in Fortaleza, Ceará, Brazil, in 1973. He received his B.S. degree in Electrical Engineering from the Federal University of Ceará in 1995, and both the M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Sao Paulo in 1998 and 2003, respectively. Currently, he is full professor of the Department of Teleinformatics Engineering, Federal University of Ceará (UFC), Fortaleza, Ceará, Brazil. At this institution, Prof. Guilherme Barreto leads the Group of Advanced Machine Learning (GRAMA), whose members pursue a variety of research topics, such as neural networks & computational intelligence, pattern recognition and machine learning, nonlinear system identification, time series prediction, and intelligent robotics. More recently, members of GRAMA have been collaborating extensively with outstanding research groups in Portugal (FEUP), Spain (University of Granada), Germany (University of Bielefeld) and Finland (Aalto University). Prof. Barreto has been serving as reviewer for several international journals (IEEE TNNLS, Neural Networks, Neurocomputing, Neural Processing Letters, Computers and Electrical Engineering, Biomedical Signal Processing and Control, and IEEE Sensors) and conferences (IJCNN, ESANN, IbPRIA, IDEAL, among others). He is also serving as the editor-in-chief of the journal Learning & Nonlinear Models (L& NLM) published by the Brazilian Computational Intelligence Society (http://sbic.ct.ufrn.br/) and as an associate editor of the Springer's International Journal of Machine Learning and Cybernetics (IJMLC).

Yoan Miche received an Engineer's Degree from Instituti National Polytechnique de Grenoble (INPG, France), and more specifically from TELECOM, INPG, on September 2006. He also graduated with a Master's Degree in Signal, Image, Speech and Telecom from ENSERG, INPG, at the same time. He is currently working in the ICS lab at Aalto University as a postdoctoral researcher, after obtaining a D.Sc. from INPG (France) and Aalto University (Finland) in 2010. His main research interests are currently in machine learning for classification/regression tasks in Internet Security.

Please cite this article as: A.H. de Souza Júnior et al., Minimal Learning Machine: A novel supervised distance-based approach for regression and classification, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.11.073

Amaury Lendasse was born in 1972 in Belgium. He received the M.S. degree in Mechanical Engineering from the Universite Catholique de Louvain (Belgium) in 1996, M.S. in control in 1997 and Ph.D. in 2003 from the same university. In 2003, he has been a post-doctoral researcher in the Computational Neurodynamics Lab at the University of Memphis. Since 2004, he is a Chief Research Scientist and a Docent in the Adaptive Informatics Research Centre in the Aalto University School of Science (previously Helsinki University of Technology) in Finland. He has created and is leading the Environmental and Industrial Machine Learning (previously Time Series Prediction and Chemoinformatics) Group. He is the Chairman of the annual ESTSP conference (European Symposium on Time Series Prediction) and member of the editorial board and program committee of several journals and conferences on machine learning. He is the author or the coauthor of around 160 scientific papers in international journals, books or communications to conferences with reviewing committee. His research includes time series prediction, chemometrics, variable selection, noise variance estimation, determination of missing values in temporal databases, nonlinear approximation in financial problems, functional neural networks and classification.

Please cite this article as: A.H. de Souza Júnior et al., Minimal Learning Machine: A novel supervised distance-based approach for regression and classification, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.11.073