Neurocomputing 164 (2015) 34–44

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
Available online 27 March 2015

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:

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

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.