Information Sciences 257 (2014) 313–330

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Genetic interval neural networks for granular data regression

Mario G.C.A. Cimino [a,*], Beatrice Lazzerini [a], Francesco Marcelloni [a], Witold Pedrycz [b,c]

[a] Dipartimento di Ingegneria dell'Informazione, University of Pisa, Largo Lucio Lazzarino 1, 56122 Pisa, Italy
[b] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2G7
[c] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

ARTICLE INFO

Article history:
Available online 16 January 2013

Keywords:
Function approximation
Genetic algorithm
Granular computing
Interval analysis
Interval order relation
Neurocomputing

ABSTRACT

Granular data and granular models offer an interesting tool for representing data in problems involving uncertainty, inaccuracy, variability and subjectivity have to be taken into account. In this paper, we deal with a particular type of information granules, namely interval-valued data. We propose a multilayer perceptron (MLP) to model interval-valued input–output mappings. The proposed MLP comes with interval-valued weights and biases, and is trained using a genetic algorithm designed to fit data with different levels of granularity. In the evolutionary optimization, two implementations of the objective function, based on a numeric-valued and an interval-valued network error, respectively, are discussed and compared. The modeling capabilities of the proposed MLP are illustrated by means of its application to both synthetic and real world datasets.

1. Introduction and background

Humans exploit abilities of processing non-numeric information clumps (*granules*) rather than individual numeric values [29]. Granulation is a suitable way for managing situations characterized by excess or a lack of data [13]. The first situation occurs, for instance, when there are collections of many objects that exhibit some similarity in terms of their properties or functional appearance [24]. Here, information granulation provides a vehicle to abstract the complexity of the data set that one can organize into hierarchies and convert the original problem into manageable subtasks. The second situation occurs, for instance, when there are noisy data or we encounter qualitative assessments provided by human experts. Here, granulation of information allows modeling the precision of indirect measurements, providing a computationally appealing view of knowledge [13].

There are a number of formal models of information granules including intervals, sets, rough sets, fuzzy sets, and shadowed sets just to name a few existing alternatives. The type of representation formalism is an essential issue to be tackled. Its choice depends on the available domain knowledge. In [27,28] the authors claim that the implementation of information granules in terms of *interval-valued* data is the easiest to comprehend and express by a domain expert, and the simplest to process when there is a great variability of granule sizes. In interval-valued data, all elements included in the same interval lose their identity in the sense they become fully indistinguishable. In a multidimensional domain, each granule is modeled by a hyperbox, which is a simple geometrical structure fully defined by its boundaries.

In the literature, two different areas are concentrated on interval-valued data, namely Symbolic Data Analysis and Interval Analysis [22]. Symbolic Data Analysis is a new development related to statistics, data mining and computer science. This paradigm offers a comprehensive approach that consists of summarizing a dataset by means of symbolic variables, e.g.

* Corresponding author. Tel.: +39 050 2217455; fax: +39 050 2217600.
E-mail addresses: m.cimino@iet.unipi.it (M.G.C.A. Cimino), b.lazzerini@iet.unipi.it (B. Lazzerini), f.marcelloni@iet.unipi.it (F. Marcelloni), wpedrycz@ualberta.ca (W. Pedrycz).

0020-0255/$ - see front matter
http://dx.doi.org/10.1016/j.ins.2012.12.049

interval variables, so as to transform the data set into a smaller and more manageable set of symbolic data. Symbolic data preserve the essential information and can be processed by means of symbolic methods. Interval Analysis introduces intervals as a fundamental means of representing real data, thus providing methods for numeric processing of these intervals.

In real-world scenarios, interval-valued data arise in several situations, such as recording monthly interval temperatures at meteorological stations, daily interval stock prices, inaccuracy of the measurement instruments, and range of variation of a variable through time [16].

The use of information granulation requires developing appropriate learning algorithms. Given this objective in mind, in the paper we propose a neural architecture to process information granules consisting of interval-valued data. Our focus is on an interval regression problem.

The first conceptualization of neural networks for processing granular data was introduced by Pedrycz and Vukovich [28]. Here, several design approaches are discussed, together with a number of architectures of granular neural networks and associated training methods. Also, the authors tackle a number of fundamental issues of these networks, such as specificity of information granules, learning complexity and generalization capabilities. Neural architectures based on interval arithmetic have been proposed in [7,10,11,22,24,26]. In particular, the model developed in [22] uses a standard multilayer perceptron (MLP) with numeric weights and biases, and a neuron transfer function able to operate with interval-valued inputs and outputs. Here, the training process uses an error function based on a weighted Euclidean distance between intervals, and a Quasi Newton method for the minimization of the error function. Other minimization methods such as genetic algorithms (GA) and evolutionary strategies have been also discussed [10].

In its general architecture proposed in the literature, an MLP that processes interval-valued data is characterized by weights and biases expressed in terms of intervals, and maps an interval-valued input vector to an interval-valued output. However, very often, in the design of the training algorithms some simplifying assumptions are being made, e.g., it is assumed that inputs, weights and biases may be real numbers, or the error function formed for the intervals is not compliant with the rules of the interval arithmetic. As regards the error function, other strategies exploit a modified quadratic error function, based on upper and lower bounds [10,14]. In [11] we proposed a new genetic algorithm-based learning method for a general interval-valued neural architecture. In this study, we adopted a numeric-valued error function derived from basic properties of interval arithmetic. However, due to the interval nature of the network output, the error of the model should be viewed as an interval to quantify the precision of the model (in addition to the quantification of its accuracy [4]). Indeed, to assess the extent to which a set of outputs of the model satisfy the requirements is not sufficient to know the proximity of a representative output to a desired value. We should also know how much other outputs are close to the representative. Such information is often important to guarantee an effective optimization, especially when the model exhibits low resolution. Hence, a numeric-valued error represents a design constraint which strongly limits the capabilities of the GA optimization.

Unlike the approaches proposed in the literature, in our MLP architecture, weights and biases are intervals, and each operation performed in the network is based on interval arithmetic. Furthermore, the parameters of the neural network are learned by using a GA which optimizes an error function expressed in terms of intervals. The use of network errors expressed as intervals requires defining an ordering relation between the chromosomes based on the interval arithmetic.

The resulting network allows forming mappings at different levels of granularity and therefore at different model resolutions. Resolution concerns the smallest change in the data that produces a modification in the model. Since the level of granularity is problem-oriented and user-dependent, it is a parameter of our neural architecture. We show the effectiveness of the method by using various interval-valued datasets.

The paper is organized as follows. In Section 2, we introduce some basic notions of interval arithmetic. The architecture of the interval neural network and the problem requirements are formally defined in Sections 3 and 4. Sections 5 and 6 describe the design of the GA, especially its search space representation issues. A procedure for parameter setting and experimental results are shown in Section 7. Finally, Section 8 draws some conclusions and discusses some future work.

2. Interval arithmetic: some definitions

We employ a generic implementation of information granules in terms of conventional interval-valued schemes. An interval-valued variable $\ddot{X}$ is defined as:

$$\ddot{X} = [\underline{x}, \bar{x}] \in \mathbb{IR}, \quad \underline{x}, \bar{x} \in \mathbb{R}, \tag{1}$$

where $\mathbb{IR}$ is the set of all closed and bounded intervals in the real line, and $\underline{x}$ and $\bar{x}$ are the boundaries of the intervals. An F-dimensional granule (hyperbox [15]) is then represented by a vector of interval-valued variables as follows:

$$\ddot{\mathbf{X}} = \left[\ddot{X}_1, \ldots, \ddot{X}_F\right] \in \mathbb{IR}^F, \quad \ddot{X}_i \in \mathbb{IR}. \tag{2}$$

Alternatively, an interval variable can be described in terms of its midpoint $\dot{x}$ and half-width $\widehat{x}$, as follows [10]:

$$\widehat{\dot{X}} = \langle \dot{x}, \widehat{x} \rangle \in \mathbb{IR}, \quad \dot{x}, \ \widehat{x} \in \mathbb{R}, \quad \dot{x} = (\underline{x} + \bar{x})/2, \quad \widehat{x} = (\bar{x} - \underline{x})/2. \tag{3}$$