Applied Soft Computing 23 (2014) 546–555

Contents lists available at ScienceDirect

Applied Soft Computing

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

The generation of qualitative descriptions of multivariate time series using fuzzy logic

Juan Moreno-Garcia[a,*], Luis Rodriguez-Benitez[b], Juan Giralt[b], Ester del Castillo[b]

[a] Universidad de Castilla-La Mancha, Escuela de Ingenieria Industrial, Avda Carlos III s/n, 45071 Toledo, Spain
[b] Universidad de Castilla-La Mancha, Escuela Superior de Informatica, Paseo de la Universidad n°4, 13071 Ciudad Real, Spain

ARTICLE INFO

Article history:
Received 9 April 2013
Received in revised form 27 March 2014
Accepted 21 May 2014
Available online 13 June 2014

Keywords:
Qualitative description of data
Multivariate time series
Temporal knowledge acquisition
Knowledge representation

ABSTRACT

An approach to generating qualitative descriptions of multivariate time series using the 'Temporal Fuzzy Model' as an intermediate representation is presented. This model, based on fuzzy logic, allows of representing in a linguistic way, the information contained in multivariate time series. The extraction of qualitative descriptions can then be carried out directly. We present a formal description of the temporal fuzzy models, the induction and inference algorithms associated with them, and a process that acquires text from them. Finally, we present a set of experiments in order to establish a measure of how valid a representation is the model of the multivariate system, and also a second set of experiments that validate the qualitative description process.

1. Introduction

Qualitatively describing data is an attempt to generate a text from some numeric input, for example, from a multivariate time series (MTS). In this kind of problem, a knowledge acquisition process (KA) is needed to extract the relevant knowledge from the numerical series. Such acquisition processes are able to use a wide range of techniques, for example, automatic induction of rule-based systems [18,39]. The KA represents knowledge from data as a model and in order to represent these models in a comprehensible way, in this work, we use fuzzy logic and more concretely our work goes around the concept of linguistic variable.

Zadeh [67] proposed the use of linguistic variables [64–66] to express the domain knowledge in a way that is quite similar to how knowledge is represented in the human mind. Then, using fuzzy logic, understandable models can be obtained that will allow experts to make a better use of the knowledge. In general, these models are composed of a collection of fuzzy rules (Mandani rules [35] or TSK rules [58]) and some of them are based on the fuzzy inductive reasoning methodology [9], which is considered an appropriate methodology for modeling qualitative behaviour and for simulating physical systems.

In order to establish a classification of the techniques that study and interpret MTS, we distinguish between global and local techniques. Global techniques generate a single model representing the entirety of the input data, whereas in local techniques, the time series is preprocessed and is divided into segments or independent parts. Then a model for each segment is created [30]. Local techniques use time frames established by the user [19,20,42]. In this paper, we propose the use of a new local technique that employs a fuzzy model to acquire temporal knowledge from the MTS. This model is called the Temporal Fuzzy Model (TFM). After that, a qualitative description is generated from this TFM. In this paper, TFMs are used as an intermediate structure to represent the knowledge of a MTS in an appropriate way. The main aim of this paper is to generate qualitative descriptions of multivariate time series.

1.1. Contributions

The contributions of this paper are concerned with the definition of a new local technique for obtaining models from MTS, called the Temporal Fuzzy Model: the capability of representing the knowledge of this model, the particular characteristics of the induction algorithm, and the generation of qualitative descriptions from the fuzzy rules in the MTS.

More concretely, the induction algorithm generates the TFMs without establishing a priori the number of rules. It studies the MTS data only in a single run, determines the values of the membership functions of the variables before execution. One important contribution is the definition of the membership functions of the variables in design time: this allows an expert to determine them, while most induction techniques can obtain these functions only

* Corresponding author. Tel.: +34 902204100x5743; fax: +34 925268849.
E-mail addresses: juan.moreno@uclm.es, juan.moreno.garcia@gmail.com (J. Moreno-Garcia).

http://dx.doi.org/10.1016/j.asoc.2014.05.021
1568-4946/

during execution time. The induction algorithm uses only the output variable to find the rule that represents a consecutive set of examples. The value of each input variable is obtained using the union of intervals. TFMs need as input a MTS with at least one input variable and a single output variable. For this reason, the induction algorithm is not applicable to univariate time series since our proposal needs at least two input variables. A preliminary version of this induction algorithm was presented in [40].

Regarding the knowledge representation in the TFM, the concept of fuzzy rule is used but, unlike other fuzzy models, the rules in the TFMs are temporally ordered. Then, each fuzzy rule in the model expresses what is happening in the MTS in a specific period of time and this fact allows representing spatio-temporal information. The time component is represented by grouping consecutive data in the MTS with the same output value in a unique rule, while the spatial information is synthesised by means of the concept of a linguistic interval that represents fuzzy subsets in the domain of variables.

The knowledge acquisition is done from TFMs and not directly from the data, it is based on fuzzy logic and uses basic operations defined directly over labels and intervals, for example, interval differences and interval intersections. These operations are mathematically simple enough to be executed efficiently. Then a text that represents two consecutive rules is generated: this text reflects the cause–effect relation relative to time.

1.2. The structure of the paper

This paper is organised as follows. Section 2 exposits the background, Section 3 describes the formal representation of the TFMs, the induction method is given in detail in Section 4, and Section 5 treats the generation of the qualitative descriptions. Finally, Section 6 tests the proposed method and Section 7 presents some conclusions and prospects for future research.

2. Background

This paper is related to the automatic induction of rule-based system (RBS) and the automatic techniques for generating qualitative descriptions of data [15]. The application of computational methods to induce RBS is an interesting way to learn from data efficiently and comprehensibly. Moreover, fuzzy logic can improve on the regression and classification methods by using fuzzy sets to make overlapping class definitions [38]. This research is more related with the description of systems although it is also linked to prediction systems. RBSs can represent both classification and regression functions, and different types of fuzzy models have been used for these purposes [23].

The classification functions are represented by an RBS. A single rule has a class assignment as consequent. Rule evaluation consists of studying the activation degree of each rule and selecting the rule with maximal support for each class. Activation degrees are also a measure of the uncertainty involved in a classification decision. There has been a lot of research into this subject. Fuzzy rule-based systems (FRBSs) have been applied to pattern classification problems [3,7,21,25,33]. They model the uncertainty in obtaining a good performance in complex classification applications. There have been many methods proposed for designing fuzzy classifiers, such as heuristic approaches [25,36], clustering methods [11,31,50], neural networks [10,34,41], and neuro-fuzzy methods [10,37].

Fuzzy regression analysis is a powerful method for obtaining the relation between the explanatory variables and the response variable in complex systems [8,26,60]. It has been successfully applied to various problems. Recently, fuzzy regression models have been used in different types of systems, for example: thermal comfort forecasting [12], insurance [1], housing [2], R&D project evaluation [24]. The first work about the fuzzy linear regression (FLR) model was presented by Tanaka et al. [51]. This work uses crisp explanatory variables and fuzzy response variables. There, the FLR problem is solved as a linear programming model to determine the regression coefficients as fuzzy numbers. Later, this research was improved [52–54]. In the literature, there have been many other research works in this line.

The automatic qualitative description techniques of data can be divided into linguistic summary reports and natural language generation systems [45]. Linguistic summarisation processes complex information describing emerging patterns through linguistic expressions and the manipulation of information granules in the form of words [15]. Yager introduced the linguistic fuzzy summary concept [62] and later Kacprzyk expanded this concept [27]. A fuzzy linguistic summary 'is a set of sentences which express knowledge about a situation through the use of fuzzy linguistic summarisers and fuzzy linguistic quantifiers' [15]. In recent years, this concept has been used for different applications, e.g., data mining [63], database query [27], and for describing temporal series [6]. Kacprzyk presents an interesting review of this subject in [28].

Previous research on knowledge acquisition for natural language generation (NLG) can be found in Reiter et al. [45–48], who present a prominent research line in this subject. Walker [59] have attempted to learn NLG rules from user ratings of the generated texts, which can perhaps be considered a type of experiment-based KA. Portet et al. [44] present a background of NLG that shows the most important works on visualisation and data-to-text (NLG) systems. *Data-to-text systems* summarise numeric data using text that makes the data more accessible to human users than do time-series plots [49]. Weather forecasting presents some interesting applications of data-to-text. *FoG* [17] automatically produces English or French texts using input data that has previously been manipulated by human users through a graphical user interface. MULTIMETEO [14] is another multilingual generator based on structured input data that provides the user with the possibility of editing the produced output. *SumTime system* [57,48] shows the potential of this technology. Other data-to-text systems summarise small data sets, including summaries of statistical data [16], air quality reports [4], and financial data [44]. Two works use large data sets: *SumTime-Turbine* summarises data sets from gas turbines [61], and *RoadSafe* summarises large meteorological datasets [56]. Also used as input is raw data. Hueske-Kraus presents in detail some works in this line of research [22]. Kahn et al. [29] propose TOPAZ, a decision support system that summarises data related to blood cell counts and drug dosages of lymphoma patients over a period of time.

3. Temporal fuzzy models

This section presents the behaviour of TFMs and a set of preliminary definitions [40]. First, the formal definition of an MTS is given. An MTS is an ordered data sequence, $E=\{e_1, e_2, \ldots, e_n\}$, where i denotes the instant when the example e_i is registered and n represents the total number of observations [43]. Its representation is $e_i=(x_1^i, \ldots, x_m^i, s^i)$, where $x_j^i \in X_j$ and $s^i \in Y$. Let us suppose that each example has a set of m real input variables defined over the domain $X=X_1 \times X_2 \times \cdots \times X_m \subset \Re^m$, and an output variable Y defined over the domain $S \subset \Re$ $(\Xi : X \longrightarrow Y)$. An MTS represents a collection of quantitative data.

TFMs use linguistic variables and each input variable X_i takes its values from an ordered set of linguistic labels $A_i=\{A_i^1, A_i^2, \ldots, A_i^{n_i}\}$ with n_i being the number of labels in A_i. The output variable also takes its values from an ordered set of linguistic labels $B=\{B^1, B^2,$