Expert Systems with Applications 41 (2014) 2035–2042

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A hybrid model of genetic algorithm with local search to discover linguistic data summaries from *creep* data

C.A. Donis-Díaz [a,*], A.G. Muro [a], R. Bello-Pérez [a], E.V. Morales [b]

[a] *Informatics Studies Center, Universidad Central Marta Abreu de Las Villas, C. Camajuaní, km 5 1/2, CP 54830, Villa Clara, Cuba*
[b] *Physics Dept., Universidad Central Marta Abreu de Las Villas, Cuba*

ARTICLE INFO

Keywords:
Genetic algorithms
Linguistic data summarization
Creep rupture stress
Data mining
Fuzzy logic

ABSTRACT

A hybrid model of Genetic Algorithm (GA) with local search to discover linguistic summaries and its application into the *creep* data analysis is proposed in this paper. Two specifics operator and a called *Diversity* term in the fitness function are introduced by the model to guarantees summaries with high quality and a wide range of information respectively. The experiments show that the hybrid model improves the results compared to those obtained using the classical model of GA. The quality of the summaries was verified by the interpretation of some of them from the theoretical point of view.

1. Introduction

Now days, it is usual to face an abundance of data that is beyond human cognitive, perceptional and comprehension skills. Data summarization is one of the basic capabilities needed now by any "intelligent" system that is meant to operate in real life situations. Many available techniques used for data summarization are not "intelligent enough", and not human consistent, partly due to a little use of natural language (e.g. summarizing statistics as average, median, minimum, maximum, α-percentile, etc.). In this context, the linguistic data summarization (LDS) has emerged as an interesting and promising approach to produce summaries from numerical data using natural language.

Several approaches or philosophies have been used to derive linguistic summaries (Bosc, Dubois, Pivert, Prade, & Calmes, 2002; Dubois & Prades, 1992; Raschia & Mouaddib, 2002; Rasmussen & Yager, 1997, 1999). The use of the fuzzy logic with linguistic quantifiers is one of the most conceptually simple, developed and used approaches. It was introduced by Yager (1982, 1991, 1996), and then considerably advanced by Kacprzyk (2000), (Kacprzyk & Yager, 2001), (Kacprzyk, Yager, & Zadrożny, 2000, 2001), (Kacprzyk & Zadrożny, 1998; Zadrożny & Kacprzyk, 1999) and implemented in Kacprzyk and Zadrożny (2000a,b,c,d).

In LDS, the fully automatic determination of the best linguistic summary has been addressed as a problem that may be infeasible in practice due to the high number of possible summaries and it is revealed as a direction for future research in some works (Kacprzyk & Zadrożny, 2009a). Several approaches have been used to deal with:

- in Kacprzyk and Zadrożny (1998) an interactive approach was proposed with a *user assistance* in the definition of summarizers by the indication of attributes and their combinations of interest. This was a first step in the definition of *protoforms* (Kacprzyk & Zadrożny, 2002, 2005, 2009b), a set of templates that can be used to specify the form of the summaries to be obtained reducing the complex search,
- the use of some sophisticated search techniques which limit the size of the search space as exemplified by George and Srikanth's use of a genetic algorithm (George & Srikanth, 1996) and other works (mainly in time series applications) (Castillo-Ortega & et al., 2011a,b; Kacprzyk, Wilbik, & Zadrożny, 2006) that use evolutive heuristics.

LDS has been used in several fields. Its success for the description of trends in *creep* data has been exposed in (Díaz, Perez, & Morales, 2011). The creep rupture stress (*creep*) is one of the most important mechanical properties considered in the design of new steels used in aeronautical, energy and petrochemical industries. It measures the stress level in which the steel structure fails when it is exposed to quite aggressive conditions (like high steam temperatures) over periods of time as long as 30 years. The objective in Díaz et al. (2011) was not making a complete mining of linguistic summaries over the *creep* data but to study the effectiveness of LDS to describe trends in such data. For this purpose was used a small group of linguistic summaries that were designed intentionally to be compared with results obtained using a neural network model proposed in Brun et al. (1999), Masuyama and Bhadeshia

* Corresponding author. Address: Marino Cabreras, 19. Camajuani, CP 52500, Villa Clara, Cuba. Tel.: +53 53123411.
E-mail addresses: cadonis@uclv.edu.cu (C.A. Donis-Díaz), amuro@uclv.cu (A.G. Muro), rbellop@uclv.edu.cu (R. Bello-Pérez), valencia@uclv.edu.cu (E.V. Morales).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.09.002

(2007). The study concluded that LDS effectively describe trends that exist in *creep* data and promoted the use of the technique for widely mining this data.

In the present study, a complete mining of linguistic data summaries over *creep* data is performed by using a proposed hybrid model that combines the heuristic search of the Genetic Algorithms (GAs) with a local search.

2. Theoretical background

2.1. Linguistic data summarization using fuzzy logic with linguistic quantifiers and their protoforms

In this paper is considered the LDS approach that uses fuzzy logic with linguistic quantifiers. Here is presented a basic description:

Given: (1) $Y = \{y_1, \ldots, y_n\}$ a set of objects (records) in a database, e.g., the set of workers, (2) $A = \{A_1, \ldots, A_m\}$ a set of attributes characterizing objects from Y, e.g., humidity, temperature, etc. in a database of weather prediction, and (3) $A_j(y_i)$ denotes a value of attribute A_j for object y_i, a linguistic summary of a data set D consists of:

- a summarizer S, i.e. an attribute together with a linguistic value (fuzzy predicate) defined on the domain of attribute A_j (e.g. '*low rain likely*' for attribute 'rain likely');
- a quantity in agreement Q, i.e. a linguistic quantifier (e.g. most);
- truth (validity) T of the summary, i.e. a number from the interval [0,1] assessing the truth of the summary (e.g. 0.7); usually, only summaries with a high value of T are interesting;
- optionally, a qualifier R, i.e. another attribute together with a linguistic value (fuzzy predicate) defined on the domain of attribute A_k determining a (fuzzy subset) of Y (e.g. '*high*' for attribute '*temperature*').

Thus, linguistic summaries may be exemplified by:

$$T(\text{most days have low rain likely}) = 0.7 \quad (1)$$

$$T(\text{much days with high temperature have high rain likely}) = 0.75 \quad (2)$$

and their foundation is Zadeh (1983) linguistically quantified proposition corresponding to either, for (1) and (2):

$$Qy's \text{ are } S \quad (3)$$

$$QRy's \text{ are } S \quad (4)$$

The T, i.e. the truth value of (3) or (4), may be calculated by using either original Zadeh's calculus of linguistically quantified statements (Zadeh, 1983), or other interpretations of linguistic quantifiers. Using the first, a (proportional, non decreasing) linguistic quantifier Q is assumed to be a fuzzy set in [0,1] and the values of T are calculated as

$$T(Qy's \text{ are } S) = \mu_Q\left[\frac{1}{n}\sum_{i=1}^{n}\mu_S(y_i)\right]$$

$$T(QRy's \text{ are } S) = \mu_Q\left[\frac{\sum_{i=1}^{n}(\mu_R(y_i) \wedge \mu_S(y_i))}{\sum_{i=1}^{n}\mu_R(y_i)}\right]$$

In some works a linguistic summary is intended as an individual linguistically quantified proposition while in others it is intended as a set or collection of such propositions. For the purpose of the present work it will be treated as the last one, i.e. the term *summary* will be used to refer the set of sentences or linguistically quantified propositions.

In Kacprzyk and Zadrożny (2002, 2005, 2009b) is presented the concept of a *protoform* as a more or less abstract prototype (template) of a linguistically quantified proposition. The most abstract *protoform* corresponds to (3) and (4), while (1) and (2) are examples of fully instantiated *protoforms*. Thus, *protoforms* form a hierarchy, where higher/lower levels correspond to more/less abstract *protoforms*. Going down this hierarchy one has to instantiate particular components of (3) and (4), i.e., Q, S and R. In Table 1, basic types of *protoforms* are shown, of a more and more abstract form (Zadrożny, Kacprzyk, & Gola, 2005). Each of the fuzzy predicates S and R may be defined by listing their atomic fuzzy predicates (pairs of "attribute/linguistic value") and structure, i.e., how these atomic predicates are combined.

2.2. Mining of linguistic data summaries

In the process of mining linguistic summaries, at one extreme, the system may be responsible for both, the construction and verification of linguistically quantified propositions (which corresponds to Type 5 *protoforms* in Table 1). At the other extreme, the user proposes a proposition and the system only verifies its validity (which corresponds to Type 0 *protoforms* in Table 1). The latter approach, obviously secures a better interpretability of the results. On the other hand, the former approach seems to be more attractive and in the spirit of data mining meant as the discovery of interesting, unknown regularities in data. The use of heuristic searches is more useful while going down the hierarchy.

2.3. LDS over creep data

As mentioned in the introduction, LDS has proven effective to describe *creep* trends regarding specific variables that are used in the process of designing new ferritic steels. However, the use of the technique in this paper (Díaz et al., 2011) was reduced to obtain a small group of completely described summaries (using *protoform* 0) to be compared with results obtained with a neural network model. Unlike this, a complete mining of linguistic summaries (using *protoform* 5) over the *creep* data is performed in the present study.

When working on *creep* data, LDS must discover linguistically quantified propositions with a high degree of T having a form (*protoform*) with the following features:

- the sought summarizer (S) must refer to the fuzzy variable *creep*, then the structure of S is known and composed only by one fuzzy predicate in the form of: "*creep is <linguistic value>*"; here it is only necessary to generate the linguistic value,
- the label of the quantifier Q is a sought element,
- the structure of the qualifier R is known in the sense that will be composed by none, one or several fuzzy predicates related by the fuzzy operator *"and"* used as the algebraic

Table 1
Classification of protoforms.

Type	Protoform	Given	Sought
0	*QRy's* are *S*	All	Validity *T*
1	*Qy's* are *S*	*S*	*Q*
2	*QRy's* are *S*	*S and R*	*Q*
3	*Qy's* are *S*	*Q* and structure of *S*	Linguistic values in *S*
4	*QRy's* are *S*	*Q*, *R* and structure of *S*	Linguistic values in *S*
5	*QRy's* are *S*	Nothing	*S*, *R* and *Q*