Knowledge-Based Systems 67 (2014) 244–262

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Modeling of a semantics core of linguistic terms based on an extension of hedge algebra semantics and its application

Cat Ho Nguyen [a,b,*], Thai Son Tran [a], Dinh Phong Pham [c]

[a] Institute of Information Technology, Vietnam Academy of Science and Technology, Hoang Quoc Viet str. 18, Cau Giay, 10 000 Ha Noi, Viet Nam
[b] Research and Development Center, Duy Tan University, Da Nang, 182 Nguyen Van Linh, Da Nang, Viet Nam
[c] Information Technology Faculty, University of Engineering and Technology, VNU, 144 Xuan Thuy, Cau Giay, Ha Noi, Viet Nam

ARTICLE INFO

Article history:
Received 7 November 2013
Received in revised form 30 March 2014
Accepted 29 April 2014
Available online 2 June 2014

Keywords:
Fuzzy rule based systems
Hedge algebras
Fuzziness measure
Fuzziness intervals
Interval-valued mappings

ABSTRACT

Computing with words and fuzzy linguistic rule based systems play important roles as they can find various significant applications based on simulating human capability. In fuzzy set approaches, words are mapped to fuzzy sets, on which work operations of the developed methodologies. The interpretability of the methodologies depends on how well word semantics is represented by fuzzy sets, which in practice are designed based on human-user's intuition. In these approaches there is no formal linkage of fuzzy sets with the inherent semantics of words to ensure the interpretability of fuzzy sets and, hence, fuzzy rules. Hedge algebras, as models of linguistic domains of variables, provide a formalism to generate triangular fuzzy sets of terms from their own semantics. This permits for the first time to design genetically terms along with their integrated triangular fuzzy sets and to construct effective fuzzy rule based classifiers. To answer the question if trapezoidal fuzzy sets can be used instead of triangular fuzzy sets in the above design method, in this study we introduce and develop the so-called enlarged hedge algebras, in which the concept of semantics core of words can be modeled. We show that these algebras provide a formal mechanism to design optimal words integrated with their trapezoidal fuzzy sets as well as fuzzy linguistic rule based classifiers to solve classification problems. Two case studies are examined to show the usefulness of the proposed algebras.

1. Introduction

The main aim of the study is to develop the so-called enlarged hedge algebras, in which the semantics core of fuzzy linguistic terms can be modeled. They establish a formalism to generate trapezoidal fuzzy set based semantics of terms from their qualitative semantics and can be applied to solve classification problems by using fuzzy linguistic rule based systems (FLRBSs) effectively.

The concept of linguistic variables and its applications are introduced and examined by Zadeh [48]. Based on these works, many fuzzy scientific and engineering disciplines, including the one of fuzzy systems, are founded and developed. Fuzzy systems attract much attention and research effort of fuzzy community, because of their application capacity in various areas. For example, some of which are the fuzzy control by Mamdani [34], King and Mamdani [31], Battaini et al. [10] and Rao and Sivasubramanian [45], the industry by Yen et al. [47], ... In particular, over more than last decade, FLRBSs have intensively been investigated and achieved many significant successes, as we can see, e.g. [22,20,12,13,17]. Based on the human user-centric viewpoint, one should deal with the interpretability, the comprehensiveness and the complexity of the designed FLRBSs incorporated closely with their accuracy in a reasonable manner, e.g. see [24,25,27,15,20]. This leads to the development of multi-objective optimization methods to design FLRBSs, including their main task to extract fuzzy rules from given datasets. It is a complicated research task as the above objectives of such a design are in general conflict and their performance depends on many factors:

(1) The linguistic labels as elementary semantic elements of fuzzy rules;
(2) The semantic representation (fuzzy sets) of linguistic labels;
(3) The form of fuzzy rules and their semantics;
(4) The selected fuzzy reasoning method;
(5) The features (complexity, distribution of data, ...) of the very given dataset such as its high dimensionality or its spatial distribution (e.g. imbalanced datasets).

* Corresponding author at: Institute of Information Technology, Vietnam Academy of Science and Technology, Hoang Quoc Viet str. 18, Cau Giay, 10 000 Ha Noi, Viet Nam. Tel.: +84 0437569948.

E-mail addresses: ncatho@gmai.com (C.H. Nguyen), trn_thaison@yahoo.com (T.S. Tran), dinhphongpham@gmail.com (D.P. Pham).

http://dx.doi.org/10.1016/j.knosys.2014.04.047
0950-7051/

There are several approaches to solve effectively the FLRBS design problem. In general, one may first exploit Evolutionary Multi-objective Optimization (EMO) or machine learning approaches, e.g. the ones for solving classification problems developed by Ishibuchi et al. [23,28], Cordon et al. [12], Ishibuchi and Yamamoto [24–26], Ishibuchi and Nojima [17], Narukawa et al. [36], Nojima et al. [43], and by Alcalá et al. [5,6]. In accordance with this trend, many methods for identification of Mamdani FLRBSs to solve regression problems are also developed, e.g. by Jimenez et al. [29], Wang et al. [50,51], Cococcioni et al. [14], Alcalá et al. [3], Gacto et al. [19], Botta et al. [11], Márquez et al. [33]. Also, other methods to design FLRBSs are examined to determine suitable semantics of linguistic terms by tuning their fuzzy set parameters or by concurrently learning rules and parameters, (e.g. [6,7,4,18]), or to deal with imbalanced or high dimensional datasets (e.g. [21,32]). Related to the representation of rule semantics, [8,9], propose a novel and simplified form to represent the semantics of the antecedent parts of FLRBSs without the use of explicit aggregation with the help of t-norm or s-norm. Instead their innovative idea is to exploit the data density distribution of data clouds of datasets and based on this the fuzzy membership of a particular data sample to a cloud is derived from the data density spatial distribution of the data associated with that cloud.

Thus, we see that fuzzy systems have intensively been studied by many distinct approaches. However, it can be seen that no any of them can immediately handle the terms that convey their own semantics, because there lacks a formalized bridge between the terms and their fuzzy sets. In practice, we observe that when human being draws a fuzzy rule representing a piece of his knowledge from his data collected from the reality, he should carefully select from his *term-vocabulary* certain terms corresponding to the piece of his knowledge. In this process, terms with their inherent semantics are interacted with his collected data and, hence, all terms in *his term-vocabulary* are, in principle, under his consideration. But, in any present approaches, a small set of terms of interest for each variable must be pre-specified and the inherent qualitative semantics of terms are still not be taken into account based on a formalized formalism. Consequently, such small pre-specified term-vocabulary is a limitation and, in such case, one can only design the fuzzy set based semantics of the pre-specified terms based on heuristic criteria for preserving term semantics.

To overcome these difficulties hedge algebras (HAs) were developed to provide a formalized formalism to link the designed fuzzy sets of terms with their (qualitative) semantics, in which the term semantics is interpreted as the order-based semantics and it determines the fuzzy sets of terms. This permits to simulate human process of deriving fuzzy rules from the reality. Relying upon this, [40], develop a two-phase method to design FLRBSs to solve classification problems effectively, in which terms along with their triangles for the first time can be genetically designed.

Trapezoids are usually used to represent the term semantics, see e.g. [46]. In the HA trend, a mathematical formalism is needed to generate trapezoidal fuzzy set based semantics from the qualitative term semantics. The difference of trapezoids from triangles is that the core of trapezoids, as fuzzy sets, are interval-cores. They may be regarded as representing the core of the term semantics expressed by trapezoids, as the numeric values of the interval-cores are understood as most compatible with the respective terms or, equivalently, with their semantics core.

Thus, according to the aim of this study, we have to enlarge develop HAs to obtain their enlarged hedge algebra (EnHAs) in which the semantics core of their terms can be modeled. The enlargement of a given (ordinary) HA can be realized by introducing a special hedge, denoted by h_0, to represent the semantics core of its terms and their semantics cores can be simulated in an axiomatic way. Then, we have to examine how the quantification characteristics of the HAs will be changed to model the quantitative term semantics in the new context. To exhibit the applicability of the EnHAs, we apply the same two-phase method examined in Nguyen et al. [40], except only trapezoids used instead of triangles, to design FLRBCSs to solve classification problems. We make two case studies of computer simulation, the one dealing with 17 classification datasets, which shows mainly that the use of trapezoids outperforms the use of triangles, and the other dealing with 24 classification datasets, which shows that the FLRBCSs designed by the proposed method outperform those examined by Alcalá-Fdez et al. [4], and by Fazzolari et al. [18].

The paper is organized as follows. In Section 2, necessary concepts and properties of the ordinary HAs are recalled to form a basis to present understandably the rest sections. Section 3 is devoted to the enlargement of ordinary HAs, the definition and the examination of elementary properties of the enlarged HAs. Fuzziness of terms is essential to describe fuzzy information and, hence, fuzziness measure and fuzziness intervals of terms are investigated in Section 4. They are basic concepts to determine other concept of the quantification of terms. In Section 5, a concept of interval-valued quantifying mappings, one of quantification characteristics of linguistic terms, is examined. Closed relationships of fuzziness measure, fuzziness intervals and interval-valued quantifying mappings are discoursed there. To show the practical soundness of the development of the concept of semantics core of terms, Section 6 is devoted to examine two case studies to prove the usefulness of the use of trapezoidal fuzzy sets produced by EnHA-based method in designing FLRBCSs. Some conclusions are presented in Section 7.

2. Necessary notions and properties of hedge algebras

Term-domains of variables can obviously be ordered based on their inherent term semantics and, thus, Nguyen et al. [38] discussed more detailedly that it seems natural to define the qualitative semantics of terms to be an inherent order-based semantics of variable term-domains. Then, the qualitative semantics of a term x of a variable $\mathcal{X}$ is represented as a collection of order relationships of x with the other terms in the set $Dom(\mathcal{X})$ of all the terms of $\mathcal{X}$. Since HAs are still not familiar for many readers, we give a short review why $Dom(\mathcal{X})$ can be viewed as HAs in order for them to more easily understand the rest of the paper (see more details in Nguyen and Wechler [41,42], Nguyen [37], Nguyen and V.L. Nguyen [39].) Consider an HA $\mathcal{AX} = (X, G, C, H_I, \leqslant)$ of $\mathcal{X}$, where

- $(X, \leqslant)$ is an order based structure, where X is a term-set of a linguistic variable $\mathcal{X}, X \subseteq Dom(\mathcal{X})$, and $\leqslant$ is an order relation induced by the inherent semantics of the terms of $\mathcal{X}$.
- G is a set of two generators c^- and c^+. c^- is called the negative primary term and c^+ is called the positive primary term, where we have $c^- \leqslant c^+$, for instance, *small* $\leqslant$ *big*;
- $C = \{\mathbf{0}, W, \mathbf{1}\}$ is the set of constants satisfying $\mathbf{0} \leqslant c^- \leqslant W \leqslant c^+ \leqslant \mathbf{1}$, whose meanings state that $\mathbf{0}$ and $\mathbf{1}$ are, respectively, the least and the greatest term in the structure $(X, \leqslant)$, W is the neutral term, e.g. "*medium*", "*middle-age*", ... Example for $\mathbf{0}$ and $\mathbf{1}$ may be "*absolutely small*" and "*absolutely big*", respectively;
- $H_I = H \cup \{I\}$, where H is a set of hedges of $\mathcal{X}$ and I is the identity of X.

For example, let us consider linguistic variable SPEED with its atomic terms *slow*, *medium* and *fast*, abbreviated as s, m and f, respectively, and with its hedges *Little*, *Rather* and *Very*, abbreviated respectively as L, R and V. In this case, $c^- = s$, $c^+ = f$, $W = m$ and $H = \{L, R, V\}$. Assuming X the set of all possible terms of SPEED