Neurocomputing 78 (2012) 48–54

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Extracting linguistic rules from data sets using fuzzy logic and genetic algorithms

Dan Meng [a,*], Zheng Pei [b]

[a] School of Economics Information Engineering, Southwestern University of Finance and Economics, Chengdu 611130, China
[b] School of Mathematics & Computer Engineering, Xihua University, Chengdu 610039, China

ARTICLE INFO

Available online 25 August 2011

Keywords:
Computing with Words
Linguistic rules
Fuzzy logic
Genetic algorithms

ABSTRACT

Linguistic rules in natural language are useful and consistent with human way of thinking. They are very important in multi-criteria decision making due to their interpretability. In this paper, our discussions concentrate on extracting linguistic rules from data sets. In the end, we firstly analyze how to extract complex linguistic data summaries based on fuzzy logic. Then, we formalize linguistic rules based on complex linguistic data summaries, in which, the degree of confidence of linguistic rules from a data set can be explained by linguistic quantifiers and its linguistic truth from the fuzzy logical point of view. In order to obtain a linguistic rule with a higher degree of linguistic truth, a genetic algorithm is used to optimize the number and parameters of membership functions of linguistic values. Computational results show that the proposed method is an alternative method for extracting linguistic rules with linguistic truth from data sets.

1. Introduction

An abundance of data in database is often beyond human cognition and comprehension. In real life, information is commonly transmitted through statements in natural language, which is also called as linguistic information, e.g. "about half of employees are young" seems useful and consistent with human way of thinking. Linguistic information often involves uncertainty, formally, the most appropriate realistic models for dealing with linguistic information is Computing with Words (CWW) proposed by Zadeh in [48,49]. In uncertain information processing, extracting fuzzy rules and modeling with fuzzy rule-based systems is an important aspect and has been widely researched in [5,9,23,26,29,32,35–37,40,41]. Based on fuzzy logic [45,46], modeling with fuzzy rule-based systems can be performed depending on the desired degree of interpretability and accuracy of the final model. Unfortunately, interpretability and accuracy are contradictive properties directly depending on the learning process and model structure. When modeling some complex systems, fuzzy rule-based systems process accuracy but lack interpretability in fuzzy rules described by fuzzy sets, in which, genetic algorithms and/or neural network are main tools for optimizing the number of linguistic terms, membership function parameters and/or the number of rules [2,6,8,10,14,16,17,19,20,24,25,38–40,43]. For example, by using neural network or genetic algorithms, we extract the following fuzzy rule $\tilde{R}$: *If X is* μ_A, *then Y is* μ_B. However, we do not know which linguistic terms can be used to interpret μ_A and μ_B. Differently, a linguistic rule is expressed by $\tilde{R}_l$: *If X is big, then Y is small*, it owns interpretability. Such linguistic rules are very important in multi-criteria decision making, new product development, *etc.*

Linguistic rule-based systems composed of linguistic variables [47] taking values in a term set with a real-world meaning possess interpretability but lack accuracy. In recent years, many different possibilities to improve the accuracy of linguistic fuzzy rule-based systems while preserving its intrinsic interpretability have been considered, e.g. Alcalá et al. propose a new postprocessing approach to perform an evolutionary lateral tuning of membership functions and obtain linguistic models with higher levels of accuracy while maintaining good interpretability in [1]. In addition, based on 2-tuples linguistic representation model, Alcalá et al. present a multi-objective evolutionary approach to quickly learn the associated rule base and generate a set of linguistic fuzzy-rule based systems with different tradeoffs between accuracy and interpretability in regression problems in [3,4]. Cordón et al. use genetic process to learn the number of linguistic terms per variable, the membership function parameters that define their semantics and the number of rules and their composition in [11]. Ishibuchi et al. provide a three-objective genetics-based machine to extract linguistic rules for high-dimensional pattern classification problems in [18]. Broekhoven

* Corresponding author.
E-mail addresses: danmeng2006@gmail.com, mengd_t@swufe.edu.cn (D. Meng), zhengpei@mail.xhu.edu.cn (Z. Pei).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.05.029

et al. use a classic genetic algorithm with binary chromosomes, as well as a real-coded genetic algorithm to optimize the membership functions of the input variables while preserve their interpretability in fuzzy ordered classifiers in [7]. Fernandez et al. use the pairwise learning approach and preference relations to deal with multi-class classification for linguistic rule based classification systems, the method improves the performance of the linguistic rule based classification systems in [15]. Evsukoff et al. use spectral analysis with structure and parameters optimization to handle the interpretability of the rules and the model's accuracy such that it can be used as tool for data understanding in [13]. In [30], we have discussed extracting linguistic data summaries from personnel database. Linguistic data summaries is a linguistic statement investigated in [21,22,27,28,31,33,34,44], from the fuzzy logical point of view, we have analyzed membership functions of fuzzy quantifiers and linguistic truth, and provided two methods to extract simple and complex linguistic data summaries. One is based on *max* operator and the other is based on aggregation operator. To obtain a complex linguistic data summary with a higher degree of truth, we have also used genetic algorithms for optimizing the number and membership functions of linguistic terms.

Formally, linguistic rule $\tilde{R}_l$: *If X is big*, *then Y is small* is a fuzzy statement. In fuzzy logic system, every fuzzy statement is given a linguistic truth [34], e.g. very true, rather true, almost false, quite false, *etc*. In uncertain inference, the more true of fuzzy statements, the more confident of their conclusion. From the inference point of view, truth of linguistic rule can be also used to explain accuracy of linguistic rule, hence, obtaining linguistic rule with higher linguistic truth from database is desired. Obviously, truth of linguistic rule $\tilde{R}_l$: *If X is big*, *then Y is small* is determined completely once the truth of linguistic data summaries '*X is big*' and '*Y is small*' in fuzzy logic system. Hence, the following three steps are used to extract linguistic rule with linguistic truth from database:

(1) Extract (complex) linguistic data summaries with linguistic truth from database.
(2) Obtain linguistic rules based on complex linguistic data summaries.
(3) Obtain truth of linguistic rule based on truth of linguistic data summaries in fuzzy logic system.

In this paper, we provide an alternative method to extract linguistic rules with linguistic truth from decision tables based on linguistic data summaries, in which, linguistic quantifiers and linguistic truth are obtained from the fuzzy logical point of view. Genetic algorithms will be used for optimizing the number and membership functions of linguistic terms. The rest of this paper is arranged as follows: In Section 2, we make a review of linguistic data summaries. In Section 3, we formalize linguistic rules based on complex linguistic data summaries and present a method for obtaining linguistic quantifiers and linguistic truth of linguistic rules. In Section 4, we provide the objective function for optimizing the number and parameters of linguistic rules with higher fuzzy linguistic quantifier and linguistic truth based on GAs. In Section 5, we give computational results for evaluation of red wine. We conclude in Section 6.

2. Linguistic data summary

A simple linguistic data summary can be expressed, e.g. *'most of employees are young' is true*. It can be formalized by *'Qys are S' is T*, in which, Q is a fuzzy linguistic quantifier, $Y=\{y_i|i=1,\ldots,n\}$ is a set of objects, S is a summarizer (a fuzzy linguistic value) of a (an) quality (attribute) for Y, e.g. *young* is summarizer of ages of employees, and T is linguistic truth for the fuzzy statement *'Qys are S'*. Denote $D=\{v(y_i)|i=1,\ldots,n\}$ the values of quality v for objects Y, then a summarizer S of v is semantically represented by a fuzzy set $\mu_S : D\longrightarrow[0,1]$. From the logical point of view, the fuzzy sets of a fuzzy linguistic quantifier and a linguistic truth are different from the fuzzy set of a summarizer in a linguistic data summary. In fact, for the classical universal quantifier $\forall$, numbers of objects are emphasized, *i.e.*, $(\forall u)p(u)$ means "every u satisfies $p(u)$." Let $P(Y)=\{A|A\subseteq Y\}$ be the power set of Y. Define a binary relation on $P(Y)$: $A\sim B \Longleftrightarrow |A|=|B|$, where $|A|$ is the cardinality of A and "$\sim$" is an equivalence relation on $P(Y)$, denote $\overline{P}(Y)=P(Y)/\sim$. Then the fuzzy sets of Q and T can be defined as $\mu_Q : \overline{P}(Y)\longrightarrow[0,1]$ and $\mu_T : \mu_Q(\overline{P}(Y))\longrightarrow[0,1]$, respectively. Accordingly, a simple linguistic data summary can be extracted automatically at level θ as follows [30]:

- Fixing a linguistic value S (it can be one or several) and a level (threshold) θ decided by experts or users. Let

$$D_S^\theta=\mu_S^{-1}(v(y_i))=\{v(y_i)|\mu_S(v(y_i))\geq\theta\}. \tag{1}$$

- Selecting a fuzzy linguistic quantifier Q, *i.e.*, can be selected such that

$$\mu_Q(C)=max\{\mu_{Q_1}(C),\mu_{Q_2}(C),\ldots,\mu_{Q_m}(C)\}, \tag{2}$$

in which $C=\{y_i|v(y_i)\in D_S^\theta\}$.
- Selecting linguistic truth T, *i.e.*,

$$\mu_T(\mu_Q(C))=max\{\mu_{T_1}(\mu_Q(C)),\mu_{T_2}(\mu_Q(C)),\ldots,\mu_{T_k}(\mu_Q(C))\}. \tag{3}$$

The so-called complex linguistic data summary has the form: *'Qys are S_1 and (or) $\cdots$ and (or) S_r' is T*, in which, S_1 is a summarizer of v_1 for Y, ..., S_r is a summarizer of v_r for Y, respectively. Based on (1), (2) and (3), we can extract simple linguistic data summaries *'Q_1ys are S_1' is T_1*, $\cdots$ and *'$Q_r$ys are S_r' is T_r*, respectively. Intuitively, extracting a complex linguistic data summary is equal to combining $\{Q_1,\ldots,Q_r\}$ and $\{T_1,\ldots,T_r\}$ to obtain Q and T, respectively.

Example 1 (*Pei et al. [30]*). Given a database (Table 1). Let $S_{\mathbf{age}}$={young (y), middle age (ma)}, $S_{\mathbf{salary}}$={low (l), high (h)}, Q={several (s), about half (ah), most (m)}, T={approximately true (at), true (t), very true (vt)}. Membership functions are given as follows:

$$\mu_y(x)=\begin{cases}1, & \text{if } x\in[25,30],\\ 4-\dfrac{x}{10}, & \text{if } x\in(30,40],\\ 0, & \text{if } x>40,\end{cases}\qquad \mu_{ma}(x)=\begin{cases}1, & \text{if } x\geq 45,\\ \dfrac{x}{10}-3.5, & \text{if } x\in(35,45),\\ 0, & \text{if } x\leq 35,\end{cases}$$

Table 1
Personnel database.

$V\backslash Y$	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
Age	25	48	31	35	28	51	37	43	34	27	53	45
Salary	1.8	2.0	2.8	3.0	2.8	3.0	2.3	2.5	3.5	2.9	3.0	3.1