Extracting complex linguistic data summaries from personnel database via simple linguistic aggregations

Zheng Pei\textsuperscript{a},*, Yang Xu\textsuperscript{b}, Da Ruan\textsuperscript{c}, Keyun Qin\textsuperscript{b}

\textsuperscript{a}School of Mathematics and Computer Engineering, Xihua University, Chengdu, Sichuan 610039, China
\textsuperscript{b}Department of Mathematics, Southwest Jiaotong University, Chengdu, Sichuan 610031, China
\textsuperscript{c}Belgian Nuclear Research Centre (SCK\textsuperscript{c}CEN), Boeretang 200, B-2400 Mol, Gent University, Gent, Belgium

A linguistic data summary of a given data set is desirable and human consistent for any personnel department. To extract complex linguistic data summaries, the LOWA operator is used from fuzzy logic and some numerical examples are also provided in this paper. To obtain a complex linguistic data summary with a higher truth degree, genetic algorithms are applied to optimize the number and membership functions of linguistic terms and to select a part of truth degrees for aggregations, in which linguistic terms are represented by the 2-tuple linguistic representation model.

Crown Copyright © 2009 Published by Elsevier Inc. All rights reserved.

1. Introduction

An abundance of data is often beyond human cognitive and comprehension skills in reality. Natural language is thus used to communicate information by human beings. A linguistic data summary, which is expressed by a sentence or a small number of sentences in a natural language, would be desirable and human consistent, e.g., in a personnel department, linguistic data summary “almost all younger and well qualified employees are well paid” seems useful. To obtain linguistic data summaries from database is not a trivial act, but requires some more intelligent techniques\cite{18,22,23,26–29,32,33}. Linguistic data summaries have been extensively studied in\cite{19,20,24,34}. In this paper, from the fuzzy logical point of view, we discuss extracting linguistic data summaries from personnel database. In Section 2, we extract a simple linguistic data summary as a starting point. In Section 3, we then extract a complex linguistic data summary. In Section 4, we optimize the number and membership functions of linguistic terms and obtain a complex linguistic data summary with a higher truth degree based on genetic algorithms (GAs) are discussed. We conclude the paper in Section 5.

2. Linguistic data summary

Yager presented an approach to the linguistic data summary\cite{29}:
• \( V \) is a quality (attribute) of interest, e.g., age, salary, etc.;

• \( Y = \{y_1, y_2, \ldots, y_n\} \) is a set of objects (employees) that manifest quality \( V \), \( V(y_i) \) (\( i = 1, \ldots, n \)) are values of quality \( V \) for each object \( y_i \);

• \( D = \{V(y_i)| i = 1, \ldots, n\} \) is a set of data (the “database” on question).

A linguistic data summary of a data set consists of: (1) summarizer \( S \), e.g., young; (2) a quantity in agreement \( Q \), e.g., most; (3) truth \( T \). Hence, a linguistic data summary can be expressed, e.g., \( \text{most of employees are young is } T \). It can be formalized by \( Qy's \) are \( S \) is \( T \), in which \( Q \) is a fuzzy linguistic quantifier, \( S \) is a summarizer (property), and \( T \) is a truth degree.

\[
\text{most } (Q) \text{ of employees } (y's) \text{ are young } (S) \text{ is } (T).
\]

Eq. (1) is equal to decide the valuation of fuzzy statement \( are(Q'y's, S) \). Many researchers used \( T(are(Q'y's, S)) = a, a \in [0, 1] \) to express the valuation [19,20]. In this paper, \( Q = \{q_1, q_2, \ldots, q_m\}, S = \{s_1, s_2, \ldots, s_t\} \) and \( T = \{t_1, t_2, \ldots, t_k\}, \) for each \( q_m \in Q, s_r \in S, \) and \( t_k \in T \) are linguistic terms, respectively, e.g., \( \text{most, young, and very true} \). For the issue on how to decide the order and indexes of linguistic terms, we refer to [3,10–14].

2.1. Fuzzy sets of linguistic terms

Summarizer \( s_r \in S \) is a linguistic term semantically represented by a fuzzy set on \( D = \{V(y_i)| i = 1, \ldots, n\} \). For example, if \( s_r \) is young and \( D_{s_r} = \{1, \ldots, 90\} \), then the fuzzy set of \( s_r \) is \( \mu_{s_r} : D_{s_r} \rightarrow [0, 1] \). In this paper, the set \( Y \) of objects is a finite set, and \( s_r \in S \) has a fuzzy set on \( D_y \). From the logical point of view, the fuzzy sets of \( q_m \in Q \) and \( t_k \in T \) are different from the fuzzy set of \( s_r \) [21,25,31]. In fact, for the classical universal quantifier \( \forall \), numbers of objects are emphasized, i.e., \( (\forall u)p(u) \) means "all u such that p(u)". The situation of the truth degree is similar to that of the quantifier. Let \( P(Y) = \{A | A \subseteq Y\} \) be the power set of \( Y \). Define a binary relation on \( P(Y) : A \sim B \iff |A| = |B| \), where \( |A| \) is the potency of \( A \) and "~" is an equivalence relation on \( P(Y) \), denote \( P(Y)/\sim \). The fuzzy sets of \( q_m \in Q \) and \( t_k \in T \) can be defined as following:

- For each fuzzy linguistic quantifier \( q_m \in Q \), define
  \[
  \mu_{q_m} : P(Y) \rightarrow [0, 1].
  \] (2)

- For each fuzzy linguistic truth degree \( t_k \in T \), define
  \[
  \mu_{t_k} : [0, 1] \rightarrow [0, 1].
  \] (3)

Remark 1. Generally, the type of each fuzzy set of Eq. (2) needs to satisfy some properties, e.g., “most” is a non-decreasing function, etc. As an example, the following fuzzy sets are given

\[
\mu_{\exists}(A) = \begin{cases} 1 & \text{if } |A| \geq 1, \\ 0 & \text{if } A = \emptyset, \end{cases}
\]

\[
\mu_{\text{most}}(A) = \begin{cases} \frac{1}{|A|} & \text{if } |A| \geq \alpha, \\ 0 & \text{if } A = \emptyset, \end{cases}
\]

\[
\mu_{\text{about}}(A) = \begin{cases} \frac{|A|/\beta}{\alpha/\beta} & \text{if } |A| \leq \alpha, \\ 0 & \text{if } |A| \leq \beta, \end{cases}
\]

In which \( \alpha, \beta \) such that \( \alpha, \beta \in (0, |Y|) \), and decided by experts or deciders. Fuzzy linguistic truth degrees Eq. (3) are already defined by many researchers [21,25,31].

2.2. Extracting a simple linguistic data summary

Based on Eq. (1), extracting the linguistic data summary is equal to deciding \( Q, S, \) and \( T \) of Eq. (1). Let a simple linguistic data summary be “\( Q'y's \ are \ S \ is \ T \)”. Then \( Q, S, \) and \( T \) can be obtained automatically by the following steps:

- Fixing a summarizer \( s_r \in S \) (it can be one or several) and a level (threshold) \( \theta \), this can be done by experts or deciders. Let
  \[
  D_{s_r}^a = \mu_{s_r}^{-1}(V(y_i)) = \{V(y_i) | \mu_{s_r}(V(y_i)) \geq \theta\}. \] (4)

- Selecting a fuzzy linguistic quantifier \( q_m \in Q \). According to Eq. (4), \( q_m \) can be selected such that
  \[
  \mu_{q_m}(A) = \max(\mu_{q_1}(A), \mu_{q_2}(A), \ldots, \mu_{q_m}(A)).
  \] (5)

where \( A = \{y_i | V(y_i) \in D_{s_r}^a\} \). If \( \mu_{q_m}(A) \) is not only one, then one \( q_m \) can be selected by deciders or has a maximal index.

- Selecting a fuzzy linguistic truth degree \( t_k \in T \). From the logical point of view, the more objects satisfying statement with the quantifier is, the higher the truth degree \( t_k \in T \) is. On the other hand, the bigger \( \mu_{q_m}(A) \) is, the more objects satisfying the statement with the quantifier is. Hence,
\[
\mu_{i_1}(\mu_{te}(A)) = \max\{\mu_{i_1}(\mu_{te}(A)), \mu_{i_2}(\mu_{te}(A)), \ldots, \mu_{i_n}(\mu_{te}(A))\}.
\]

(6)

Example 2. A personnel database is in Table 1. 1.8 of \(y_1\) means the salary of the employee \(y_1\) is 1.8 thousand dollar.

Let \(S_{age} = \{\text{young(y), middle age(ma)}\}, S_{salary} = \{\text{low(l), high(h)}\}\). Select \(Q = \{\text{several(s), about half(ah), most(m)), T = \{\text{approximately true(at), true(t), very true(vt)}\}\). Their membership functions are as following:

\[
\begin{array}{c}
\mu_{y}(x) = \begin{cases} 
1 & \text{if } x \in [25, 30], \\
4 - \frac{1}{10}x & \text{if } x \in (30, 40], \\
0 & \text{if } x > 40,
\end{cases} \\
\mu_{sal}(x) = \begin{cases} 
1 & \text{if } x \geq 45, \\
\frac{1}{10}x - 3.5 & \text{if } x \in (35, 45), \\
0 & \text{if } x \leq 35.
\end{cases}
\end{array}
\]

\[
\begin{array}{c}
\mu_{l}(A) = \begin{cases} 
\frac{1}{2}(|A| - 1) & \text{if } |A| \in [1, 3], \\
2 - \frac{1}{2}|A| & \text{if } |A| \in (3, 6], \\
0 & \text{if } |A| > 6.
\end{cases} \\
\mu_{ah}(A) = \begin{cases} 
1 & \text{if } x \leq 0.5, \\
\frac{1}{2}x & \text{if } x \in (0.5, 0.8), \\
0 & \text{if } x \in [0.8, 1].
\end{cases}
\end{array}
\]

(1) Fixing a summarizer \(s' = \text{young} \in S_{age}\) and \(s'' = \text{high} \in S_{salary}\). Let threshold \(\theta = 0.5\). Then according to \(\mu_{y}\) and \(\mu_{sal}\), we obtain

\[
\begin{align*}
D^{0.5}_{y} &= \{V(y_1) | \mu_{y}(V(y_1)) \geq 0.5\} = \{25, 31, 35, 28, 34, 27\}, \\
D^{0.5}_{sal} &= \{V(y_1) | \mu_{sal}(V(y_1)) \geq 0.5\} = \{2.8, 3.0, 3.5, 2.9, 3.1\}, \\
A_y &= \{y_1 | V(y_1) \in D^{0.5}_{y}\} = \{y_1, y_3, y_4, y_5, y_9, y_{10}\}, \\
A_{sal} &= \{y_1 | V(y_1) \in D^{0.5}_{sal}\} = \{y_3, y_4, y_5, y_9, y_{10}, y_{11}, y_{12}\}.
\end{align*}
\]

(2) By \(\mu_{l}, \mu_{ah}, \mu_{m}\) and Eq. (9), we obtain \(\mu_l(A_y) = \mu_m(A_y) = 0, \mu_{ah}(A_y) = 1\),

\[
\max\{\mu_l(A_y), \mu_m(A_y), \mu_{ah}(A_y)\} = \mu_{ah}(A_y).
\]

(3) Selecting a fuzzy linguistic truth degree which has a maximal index, i.e., the fuzzy linguistic truth degree is \(\mu_{te}(\mu_{ah}(A_y))\).

Hence, according to Eqs. (11) and (12), we obtain the following linguistic data summary:

about half of employees are young is very true.

(13)

(3) Similar (2), we also obtain the following linguistic data summary

most of employees have high salary is approximately true.

(14)

3. Extracting a complex linguistic data summary based on either the Max operator or the LOWA operator

Here, the so-called complex linguistic data summary has the form “Q’s are \(S_1\) and \(\ldots\) and \(S_n\) is \(T\)” Based on Eq. (4)–(6), it is easy to obtain “Q’s are \(S_1\) is \(T_1\), \ldots, “Q’s are \(S_n\) is \(T_n\). Hence, for a complex linguistic data summary, the problem is how to combine \(\{Q_1, \ldots, Q_n\}\) and \(\{T_1, \ldots, T_n\}\) to obtain \(Q\) and \(T\), respectively.

Table 1

<table>
<thead>
<tr>
<th>V (Y)</th>
<th>(y_1)</th>
<th>(y_2)</th>
<th>(y_3)</th>
<th>(y_4)</th>
<th>(y_5)</th>
<th>(y_6)</th>
<th>(y_7)</th>
<th>(y_8)</th>
<th>(y_9)</th>
<th>(y_{10})</th>
<th>(y_{11})</th>
<th>(y_{12})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>25</td>
<td>48</td>
<td>31</td>
<td>35</td>
<td>28</td>
<td>51</td>
<td>37</td>
<td>43</td>
<td>34</td>
<td>27</td>
<td>53</td>
<td>45</td>
</tr>
<tr>
<td>Salary</td>
<td>1.8</td>
<td>2.0</td>
<td>2.8</td>
<td>3.0</td>
<td>2.8</td>
<td>3.0</td>
<td>2.3</td>
<td>2.5</td>
<td>3.5</td>
<td>2.9</td>
<td>3.0</td>
<td>3.1</td>
</tr>
</tbody>
</table>
Example 3. Assume \( S_r \) \( \in \mathbb{R} \) \((r = 1, \ldots, r)\), and

\[
D^\mu_{r'} = \mu_{r'}(V(y_i)) = \{ V(y_i) \mid \mu_{r'}(V(y_i)) \geq b_r \}.
\]

Then Q can be selected as

\[
\mu_Q(B) = \max\{\mu_{Q_1}(B), \mu_{Q_2}(B), \ldots, \mu_{Q_r}(B)\},
\]

where \( B = \{ y_i \mid V(y_i) \in D^\mu_{r'} \} \cap \cdots \cap \{ y_i \mid V(y_i) \in D^\mu_{r'} \} \). T can be selected as

\[
\mu_T(\mu_Q(B)) = \max\{\mu_{T_1}(\mu_Q(B)), \mu_{T_2}(\mu_Q(B)), \ldots, \mu_{T_r}(\mu_Q(B))\}.
\]

Method 2: Based on the LOWA operator \[10\], linguistic terms Q and T can be selected by

\[
Q = \phi(Q_1, Q_2, \ldots, Q_r) = \text{lwb} = \{ w_k, b_k, k = 1, \ldots, r \}
\]

\[
= w_1 \oplus b_1 \oplus (1 - w_1) \oplus C^{-1}\{ b_1, b_2, h = 2, \ldots, r \},
\]

\[
T = \phi(T_1, T_2, \ldots, T_r) = \text{lwb}^T = \{ w_k', b_2', k = 1, \ldots, r \}
\]

\[
= w'_1 \oplus b'_1 \oplus (1 - w'_1) \oplus C^{-1}\{ b_1', b_2, h = 2, \ldots, r \},
\]

where \( \text{lwb} = \{ w_1, \ldots, w_r \} \) and \( \text{lwb} = \{ w_1, \ldots, w_r \} \) are weighting vectors such that \( w_i, w_i' \in [0, 1] \), \( \sum_{i=1}^{r} w_i = 1 \) and \( \sum_{i=1}^{r} w_i' = 1 \). \( \beta_0 = \sum_{i=1}^{r} w_i \), \( \beta_0' = \sum_{i=1}^{r} w_i' \). \( B = \{ b_1, b_2, \ldots, b_r \} \) and \( B' = \{ b_1', b_2', \ldots, b_r' \} \) are vectors associated with \( \{ Q_1, \ldots, Q_r \} \) and \( \{ T_1, \ldots, T_r \} \) such that

\[
B = \sigma'(\{ Q_1, \ldots, Q_r \}) = \{ Q_{\sigma(1)}, Q_{\sigma(2)}, \ldots, Q_{\sigma(r)} \},
\]

\[
B' = \sigma'(\{ T_1, \ldots, T_r \}) = \{ T_{\sigma(1)}, T_{\sigma(2)}, \ldots, T_{\sigma(r)} \},
\]

where \( Q_{\sigma(j)} \leq Q_0 \) and \( T_{\sigma(j)} \leq T_{\sigma(j)} \) \((\forall i \leq j)\), \( \sigma \) and \( \sigma' \) are permutations over \( \{ Q_1, \ldots, Q_r \} \) and \( \{ T_1, \ldots, T_r \} \), respectively. \( C^2 \) and \( C^2 \) are defined by

\[
C^2(\lambda_i, b_i, i = 1, 2) = \lambda_i \oplus b_i \oplus (1 - \lambda_i) \oplus b_2 = Q_k \in \{ Q_1, \ldots, Q_r \},
\]

\[
C^2(\lambda', b', i = 1, 2) = \lambda' \oplus b' \oplus (1 - \lambda') \oplus b_2' = Q_k' \in \{ T_1, \ldots, T_r \},
\]

where \( b_1 = Q_2, b_2 = Q_j \) \((j \geq i)\) and \( b'_1 = T_{\bar{j}}, b'_2 = T_{\bar{j}} \) \((j \geq i)\) then \( k = \min\{m, i + \text{round}(\lambda_i(j - i))\} \) and \( k' = \min\{m, i + \text{round}(\lambda'_i(j - i))\} \), where \( \text{round} \) is the usual round operation. In Eqs. (18) and (19), weighting vectors \( W \) and \( W' \) can be computed by \[30\],

\[
Q(x, a, b) = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } a \leq x < b, \\ 1 & \text{if } x \geq b. 
\end{cases}
\]

where \( x, a, b \in [0, 1] \). Some examples of \( Q(r, a, b) \) are most, at least half and as many as possible, their parameters \((a, b)\) are \((0.3, 0.8), (0.0, 5), \) and \((0.5, 1)\), respectively. By \( Q(r, a, b) \), weighting vectors \( W \) and \( W' \) are

\[
w_i = Q(\frac{i}{r}, a, b) - Q(\frac{i-1}{r}, a, b), \quad i = 1, \ldots, r.
\]

\[
w_i' = Q(\frac{i}{r}, a, b') - Q(\frac{i-1}{r}, a, b'), \quad i = 1, \ldots, r.
\]

Example 3. Continue Example 2, by Eqs. (13) and (14), combining \{about half, most\} and \{approximately true, very true\} are needed. Using Method 1, we obtain \( B = A_6 \cap A_7 \cap \{ y_3, y_4, y_5, y_9, y_{10} \} \). So, \( \mu_i(B) = \frac{1}{2}, \mu_{ab}(B) = 0.5, \mu_m(B) = 0, \mu_{ab}(\mu_{ab}(B)) = 1, \mu_{ab}(\mu_{ab}(B)) = 0, \) and a complex linguistic data summary is

\[
\text{about half of employees are young and have high salary is approximately true.}
\]

Using Method 2, we select the linguistic quantifier 'most' and the weighting vectors \( W = W' = (0.4, 0.6) \), and

\[
C^2(\lambda_i, b_i, i = 1, 2) = 0.4 \odot \text{most} \oplus 0.6 \odot \text{about half} = \text{about half}, C^2(\lambda_i, b_i, i = 1, 2) = 0.4 \odot \text{very true} \oplus 0.6 \odot \text{approximately true}, \text{true} = \text{true2}. \text{ Hence, the complex linguistic data summary is}
\]

\[
\text{about half of employees are young and have high salary is true.}
\]

In Example 3, the conclusions of Method 1 and Method 2 are almost the same. In Method 1, by the logical relationship among fuzzy linguistic quantifiers, the fuzzy linguistic truth degrees and summarizer, \( B \) of Eq. (16) can be obtained. Since the truth degree of the complex linguistic data summary is maximal, the \text{Max} operator is used. In Method 2, the \text{round}(·)
operator is used, and the round operator in combination of linguistic indexes is very sensitive to the values around 0.5 (round(0.499) = 0 and round(0.501) = 1), the difference is occurred between Method 1 and Method 2. From the simple computing point of view, Method 2 is better to aggregate linguistic terms.

4. Optimizing a complex linguistic data summary based on GAs

Genetic algorithms (GAs) are search algorithms that use operations found in natural genetics to guide the trek through a search space [4, 15]. A number of papers have been devoted to the automatic generation of the knowledge base of a fuzzy rule-based system (FRBS) using GAs [1, 2, 5–9, 16]. In this section, we optimize the number and membership functions of linguistic terms and obtain a complex linguistic data summary with a higher truth degree based on GAs.

4.1. Optimizing the number and membership functions of linguistic terms

Personnel database is an information system, in which employee (in Y) is an object and a quality (in V) is an attribute. In a real-world personnel database, generally, V(yj) (yj∈Y) is a numerical one. In practice, a numeral value is hard to obtain for a good decision. Hence, a linguistic data summary is needed, the process is equal to classifying on Y, especially fuzzy classifying. Naturally, selecting the number and membership functions of linguistic terms is a problem. Here, suppose there exist L qualities (attributes) in V, and each domain of attribute is denoted by Di∈R, i = 1, . . . , L, then each object (employee) yj∈Y is understood as a point on space D1 × · · · × Di, i.e., yj = (d1j, . . . , dimj), dij ∈ Di. Suppose that n′(n′ < n) patterns yj = (d1j, . . . , dimj), i = 1, . . . , n′ are given as training patterns from M classes: class 1 (C1), . . . , class M (CM). The problem is to generate the number and membership functions of linguistic terms on Di that divide the pattern space into M disjoint decision areas. Assuming each axis Di is partitioned into kI fuzzy subsets {Aij | kI = 1, . . . , KI}, then D1 × · · · × Di is divided into K1 × · · · × KI fuzzy subspaces, and each fuzzy subspace can be expressed by an If–Then rule:

\[ R_{k1 \times \cdots \times kI} : \text{if } d_1 \text{is } A_{1j} \text{ and } \cdots \text{and } d_i \text{is } A_{ij}, \text{ then } y_j \text{ belongs to class } C_m \text{ with } CF = CF_{k1 \times \cdots \times kI}. \]  

(23)

where \( R_{k1 \times \cdots \times kI} \) is a label of an If–Then rule. \( A_{ij}(l = 1, \ldots, L) \) are fuzzy subsets on \( D_i \), \( C_m(m = 1, \ldots, M) \) is the consequent, and \( CF_{k1 \times \cdots \times kI} \) is the grade of certainty of the If–Then rule and determined by the following procedure.

1. For each class \( C_m \) and rule \( R_{k1 \times \cdots \times kI} \), we have

\[ \forall C_m \ni \sum_{y_j \in C_m} A_{ij}(d_1j) \times \cdots \times A_{ij}(d_im). \]  

(24)

2. Selecting

\[ CF_{k1 \times \cdots \times kI} = \max(\forall C_m \ni \sum_{y_j \in C_m} A_{ij}(d_1j) \times \cdots \times A_{ij}(d_im)). \]  

(25)

Remark 4. If \( CF_{k1 \times \cdots \times kI} = 0 \), the rule \( R_{k1 \times \cdots \times kI} \) is useless to classify \( y_j \), and the consequent of the rule \( R_{k1 \times \cdots \times kI} \) is modified by \( C_m = \emptyset \). If two or more \( \forall C_m \) are equal to \( CF_{k1 \times \cdots \times kI} \), then the rule is not good to classify \( y_j \) and also \( C_m = \emptyset \).

Based on the rule set R, a new pattern \( y' = (d_1j, \ldots, d_im) \) is classified by

1. Calculate \( \forall k1 \times \cdots \times kI \) for each rule \( R_{k1 \times \cdots \times kI} \),

\[ \forall k1 \times \cdots \times kI = A_{ij}(d_1j) \times \cdots \times A_{ij}(d_im) \times CF_{k1 \times \cdots \times kI}. \]  

(26)

2. Find the class \( C_{m'} \) such that

\[ \forall C_{m'} = \max(\forall k1 \times \cdots \times kI | R_{k1 \times \cdots \times kI} \in R). \]  

(27)

If \( \forall C_{m'} = 0 \) or \( C_{m'} = \emptyset \) of the rule, then \( y' \) is left as an unclassified pattern, else assign \( y' \) to the class \( C_{m'} \) determined by Eq. (27).

The main components of optimizing the number and membership functions of linguistic terms based on GAs are described as follows [6, 7]:

(1) Encoding the solution: The two components of the solution to be encoded are the number of linguistic terms for variable (granularity) and the membership functions that define their semantics. (1) Number of labels (S1), in this paper, there are L variables (qualities), the number of labels per variable is stored into an integer array of length L. In this contribution, the possible values considered are the set \( \{3, \ldots, 9\} \). (2) Membership functions (S2), in this paper, we deal with triangular functions, a real number array of \( L \times 9 \times 3 \) positions is used to store the membership functions. If \( S_1 \) is the granularity of variable \( l, S_2 \in \{3, \ldots, 9\}, P_{l1}, P_{l2}, P_{l3} \) are the definition points of the label \( j \) of the variable \( l \), and \( S_2l \) is the information about the fuzzy partition of variable \( l \) in \( S_2 \), then a graphical representation of the chromosome is shown
as following: \( S_1 = (s_1, s_2, \ldots, s_L) \), \( S_2 = (P_{i_1}^1, P_{i_2}^1, \ldots, P_{i_n}^1, P_{i_1}^2, P_{i_2}^2, \ldots, P_{i_n}^2) \), \( S_3 = (s_{L+1}, s_{L+2}, \ldots, s_{2L}) \), \( S = S_1S_2 \). Uniform fuzzy partitions are denoted by \( (V_i^0, V_i^1, \ldots, V_i^{L-1}) \) for each variable. For general fuzzy partition, a variation interval is defined for each one of the membership function definition points [6]. i.e., \( P_i^0 \in [L_0, R_0] = \left[ V_i^0 - \frac{V_i^1 - V_i^0}{2}, V_i^0 + \frac{V_i^1 - V_i^0}{2} \right], P_i^0 \in [L_0, R_0] = \left[ V_i^0 - \frac{V_i^1 - V_i^0}{2}, V_i^0 + \frac{V_i^1 - V_i^0}{2} \right] \).

(2) Initial gene pool: The initial population is composed of four groups [6]: (1) Each chromosome will have the same number of labels in all its variables and the membership functions are uniformly distributed across the domain of variable. (2) Each chromosome can have a different granularity per variable (different values in \( S_1 \)) and the membership functions are uniformly distributed as in the first part. (3) Each chromosome will have the same number of labels in all its variables. Then a uniform fuzzy partition is built for each variable as in the first group and the variation intervals of all the definition points are calculated. Finally, a value for all the definition points is randomly chosen from the correspondent variation interval. (4) Each chromosome can have different number of labels per variable as in the second group and the membership functions are calculated in the same way as in the third group, a random value is in the variation interval.

(3) Evaluating the chromosome: Maximizing the number of correctly classified pattern and minimizing the number of If-Then rule are the goals, so, we have

\[
\text{Minimize}: f(s) = \omega_1 \text{DCP}(s) + \omega_2 |s|, \tag{28}
\]

where \( s \) is a chromosome, \( \text{DCP}(s) \) is the number of unclassified patterns by \( s \), \( |s| \) is the number of If-Then rules in \( s \). The weights in Eq. (28) should be specified as \( 0 < \omega_2 < \omega_1 \) [17]. \( f(s) \) is treated as the fitness function in GAs.

(4) Genetic operators: Selection, crossover and mutation are as follows:

- **Selection**: Let the current population be \( \mathcal{P} \). The selection probability \( P(s) \) of chromosome \( s \) is

\[
P(s) = \left( f_{\text{max}}(\mathcal{P}) - f(s) \right)/ \sum_{r \in \mathcal{P}} (f_{\text{max}}(\mathcal{P}) - f(s)), \tag{29}
\]

where \( f_{\text{max}}(\mathcal{P}) = \max(f(s)) \forall s \in \mathcal{P} \).

- **Crossover**: Two different crossover operators are considered [15]. (1) Crossover when both parents have the same granularity level per variable, in this case, the crossover operator in \( S_2 \) and maintaining the parent \( S_1 \) values in the offspring. If \( S_1^{w1} = ((P_{i_1}^{1w}), \ldots, (P_{i_n}^{1w})) \) and \( S_1^{w2} = ((P_{i_1}^{2w}), \ldots, (P_{i_n}^{2w})) \) are to be crossed, the following four offspring are generated:

\[
S_2^{w1,w2} = ((P_{i_1}^{1w}), \ldots, (P_{i_n}^{1w})), (P_{i_n}^{1w}) = \min((P_{i_n}^{1w})), (P_{i_n}^{2w}) = \max((P_{i_n}^{2w})),
\]

where \( i = 1, 2, 3 \). This operator uses a parameter that is either a constant or a variable whose value depends on the age of the population [15].

- **Mutation**: Two different operators are used. (1) Mutation on \( S_1 \), in this case, once a new value \( s'_i \in \{3 \ldots 9\} \) at the point \( i \) of \( S_1 \) is selected, a uniform fuzzy partition for this variable is stored in its corresponding zone of \( S_2 \). (2) Mutation on \( S_2 \): Let \( S_2^{w1} = ((P_{i_1}^{1w}), \ldots, (P_{i_n}^{1w})), \ldots, (P_{i_n}^{1w})) \) and the element \( (P_{i_n}^{1w}) \) be selected for this mutation (the domain of \( (P_{i_n}^{1w}) \) is \( (P_{i_n}^{1w}), (P_{i_n}^{1w}), \ldots, (P_{i_n}^{1w})) \), the result is a vector \( S_2^{w1,w2} = ((P_{i_1}^{1w}), \ldots, (P_{i_n}^{1w})), \ldots, (P_{i_n}^{1w})) \), and

\[
\left\{ \begin{array}{ll}
(P_{i_n}^{1w}) + \Delta(t, (P_{i_n}^{1w}), (P_{i_n}^{1w})) & \text{if } e = 0, \\
(P_{i_n}^{1w}) + \Delta(t, (P_{i_n}^{1w}), (P_{i_n}^{1w})) & \text{if } e = 1.
\end{array} \right.
\]

with \( t \) being the current generation, \( e \) being a random number in \( \{0,1\} \), and the function \( \Delta(t, y) = \Delta(t, y) = y(1 - r^{1-\delta})^{\delta} \), with \( r \) being a random number in the interval \([0,1] \), \( T \) the maximum number of generations and \( b \) a parameter chosen by users [6].

4.2. Obtaining a complex linguistic data summary with a higher truth degree

To make the complex linguistic data summary with a higher truth degree, parts of \( \{Q_1, \ldots, Q_r\} \) and \( \{T_1, \ldots, T_r\} \) can be selected to obtain \( Q \) and \( T \), respectively. In this subsection, GAs are used to select parts of \( \{Q_1, \ldots, Q_r\} \) and \( \{T_1, \ldots, T_r\} \). To avoid the loss of information, the 2-tuple linguistic representation model is used to represent \( Q_r \) and \( T_r \) (\( r = 1, \ldots, r \)).
Definition 5 [14]. Let $s_i \in S$ be a linguistic term. Then its equivalent 2-tuple linguistic representation is obtained by means of the function $\theta$ as:

$$\theta : S \rightarrow (S \times [-0.5, 0.5]), \quad \theta(s_i) = (s_i, 0).$$

Definition 6 [14]. Let $\{s_0, s_1, \ldots, s_g\}$ be a linguistic term set and $\beta \in [0, g]$ a value supporting the result of a symbolic aggregation operator. Then the 2-tuple that expresses the equivalent information to $\beta$ is obtained with the following function:

$$A : [0, g] \rightarrow (S \times [-0.5, 0.5]),$$

$$A(\beta) = (s_i, \alpha), \quad \text{with} \quad s_i = \text{round}(\beta),$$

$$\alpha = \beta - i, \quad \alpha \in [-0.5, 0.5].$$

where round$(\cdot)$ is the usual round operation, $s_i$ has the closest index label to $\beta$ and $\alpha$ is the value of symbolic translation.

Definition 7 [12]. Let $A = \{(r_1, \alpha_1), \ldots, (r_m, \alpha_m)\}$ be a set of 2-tuples to be aggregated. Then the extended LOWA operator, $\phi^e$, is defined as

$$\phi^e((r_1, \alpha_1), \ldots, (r_m, \alpha_m)) = WB^T = EC^m \{w_i, (r_{\sigma(i)}, \alpha_{\sigma(i)})\}, \quad i = 1, \ldots, m = A \left( \sum_{i=1}^{m} w_i A^{-1}((r_{\sigma(i)}, \alpha_{\sigma(i)})) \right) = A \left( \sum_{i=1}^{m} w_i \beta_{\sigma(i)} \right),$$

(30)

where $\beta_{\sigma(i)} = A^{-1}((r_{\sigma(i)}, \alpha_{\sigma(i)}))$, $\sigma$ is a permutation over $\{1, \ldots, m\}$ such that $r_{\sigma(i)} \leq r_{\sigma(j)}$ ($\forall i \leq j$).

We just discuss how to obtain a part of $\{(T_1, 0), \ldots, (T_s, 0)\}$ based on GAs for aggregating a complex linguistic data summary with a higher truth degree.

1. **Encoding the solution**: The length of each chromosome is $r$, and

$$S = t_1 t_2 \cdots t_r, \quad \text{and} \quad \forall t_r \in \{0, 1\}, \quad (31)$$

2. **Initial gene pool**: By Eq. (31), the initial population is randomly selected in $2^r$ solutions as usual [7].

3. **Evaluating the chromosome**: For a solution $s = t_1 t_2 \cdots t_r$ and fixed a linguistic quantifier $Q(r, a, b)$ which is defined by Eq. (20), using the operator $\phi^e$ which is defined by Eq. (30), we obtain

$$(T_j, \alpha_j) = \phi^e(s) = EC^m \{w_i, (T_{\sigma(i)}, \alpha_{\sigma(i)})\} \quad i = 1, \ldots, r, \quad \text{and} \quad t_{\sigma(i)} = 1 = A \left( \sum_{i=1}^{m} w_i A^{-1}((r_{\sigma(i)}, \alpha_{\sigma(i)})) \right),$$

(32)

where $w_i$ is decided by Eq. (20), in this paper, suppose $Q(r, a, b)$ is fixed, $T_j \in \{T_1, \ldots, T_r\}$ and $\alpha_{\sigma(i)} = 0$. According to Eq. (32), the fitness function is obtained as follows:

Maximize: $f(s) = A \left( \sum_{i=1}^{m} w_i A^{-1}((r_{\sigma(i)}, 0)) \right) = (T_j, \alpha_j).$\quad (33)

4. **Genetic operators**: **Selection** is defined as following: Let the current population be $\Psi$. Then the selection probability $P(s)$ of chromosome $s$ is

$$P(s) = \left( f(s) - f_{\min}(\Psi) \right) \bigg/ \sum_{s \in \Psi} \left( f(s) - f_{\min}(\Psi) \right),$$

(34)

where $f_{\min}(\Psi) = \min \{ f(s) | s \in \Psi \}$, assume the index of $f_{\min}(\Psi)$ is $j$, then $f(s) - f_{\min}(\Psi) = f_j + \alpha_j - f_{j'} - \alpha_{j'}$. The generation of the offspring population is put into effect by using the classical binary multi-point crossover and uniform mutation operators [9,16]. Obviously, based on these genetic operators and fitness function Eq. (33), the optimal solution can be obtained, the optimal solution makes a complex linguistic data summary with a higher truth degree. Correspondingly, using $\phi^e$, aggregating $\{(Q_1, 0), \ldots, (Q_s, 0)\}$ can be obtained.

5. **Conclusions**

From the formalization point of view, a linguistic data summary presented by Yager is equal to a fuzzy statement with a fuzzy linguistic quantifier. In this paper, based on the structure and valuation of a fuzzy statement, the method
to extract Q, S and T of a linguistic data summary is discussed. For a complex linguistic data summary, the LOWA operator is used to obtain Q, S and T. In practice, optimizing a complex linguistic data summary is needed. In this paper, a formal optimizing method based on GAs is given, in which linguistic terms are represented by the 2-tuple linguistic representation model.

Acknowledgements

The authors thank professor Francisco Herrera for improving this paper significantly. This work is supported by the excellent young foundation of Sichuan Province (Grant No. 06ZQ026-037) and SZD, the National Natural Science Foundation of China (Grant No. 60875034), the Special Research Funding to Doctoral Subject of Higher Education Institutes in China (Grant No. 20060613007).

References