Transport and Communications Science Journal, Vol. 75, Issue 07 (09/2024), 2081-2092

Transport and Communications Science Journal

A DESIGN OF COMPUTATIONAL FUZZY SET-BASED SEMANTICS FOR EXTRACTING LINGUISTIC SUMMARIES

Pham Dinh Phong, Nguyen Duc Du*

University of Transport and Communications, No 3 Cau Giay Street, Hanoi, Vietnam

ARTICLE INFO

TYPE: Research Article
Received: 24/07/2024
Revised: 04/09/2024
Accepted: 10/09/2024
Published online: 15/09/2024
https://doi.org/10.47869/tcsj.75.7.6
* *Corresponding author*
Email: nducdu@utc.edu.vn; Tel: 0912363245

Abstract. Linguistic summarization of data is to extract a set of summary sentences in the form of natural language, so-called linguistic summaries, from numeric data. The extracted linguistic summaries should be compact and diverse, and have a validity measure greater than a given threshold, so genetic algorithms are applied to extract such linguistic summaries. Besides, the interpretability of linguistic summary content is considered in recent studies in such a way that enlarged hedge algebras are applied to generate multi-semantic structures for linguistic words of linguistic variables ensuring the interpretability of the content of the linguistic summaries. However, the membership function of computational fuzzy-set-based semantics of linguistic words is usually in the shape of trapezoid. In this paper, a membership function of the form *S*-function is applied to improve the quality of extracted linguistic summaries. Besides, the applied algorithm is parallelized to reduce running time. The experimental results with the creep dataset have demonstrated the effectiveness of the proposed method.

Keywords: linguistic summary, hedge algebras, fuzzy set, computational semantics, multi-semantic structure.

1. INTRODUCTION

Nowadays, numeric data in all sectors of our social life is increasing rapidly. Most of us do not easily understand that digital data. Therefore, the need of extracting useful information hidden in numeric data for decision making is extremely urgent, requiring researchers to propose effective data mining methods. Among those methods, extracting linguistic summaries (LSs) in the form of sentences in natural language according to a given structural format from numeric data is an effective and useful data mining method. It has practical application significance because each LS describes knowledge about real-world objects stored as numeric data in a dataset. Knowledge expressed in natural language makes it easier for human users to understand than numbers. The structure of LS used in this study is a sentence with Yager's quantifier word [1] of the form: "*Q y* are *S*" or "*Q F y* are *S*" [1-11]. For example, "*Very few* (*Q*) sales of printers (*y*) is with *high* commission (*S*)" [7], "*Most* (*Q*) hospitals (*y*) with *very high* average hospital stay (*F*) have *very low* computer (*S*)" [5]. Human users read summary sentences to understand information and knowledge stored in the dataset through the semantics of the linguistic words such as '*very few*', '*most*', '*high*', '*very low*', and '*very high*' in the linguistic summary structure. The quantifier word *Q* represents a proportion that satisfies the summarizer *S* compared to all objects in the dataset in the first sample sentence or objects in the group that satisfy the filter criterion *F* in the second sample sentence.

Each linguistic summary is evaluated by the validity measure or truth measure. The validity measure is calculated by the membership function value of fuzzy sets representing the computational semantics of the corresponding linguistic words in the sentence structure. Each linguistic summary in a set of linguistic summaries extracted from a given dataset has its validity (*T*) greater than a given threshold. One of the big challenges of extracting linguistic summaries from numeric data is that when all three components *Q*, *F*, and *S* are completely unknown in terms of attributes and linguistic words. This case is the most general level, so the number of extracted linguistic summaries is very large leading to tremendous computational volume. However, human users can discover useful and interesting knowledge hidden in the numeric dataset. In practice, human users cannot read all the huge number of extracted linguistic summaries, so they just read some certain useful ones. To extract a set of useful linguistic summaries, so-called the optimal set of linguistic summaries, genetic algorithms are applied based on some constraints and quality assessments [5, 12-14].

Despite having quality assessments and adding two additional genetic operators such as *Propositions Improver operator* and *Cleaning operator*, the genetic algorithm models in [13, 14] have not eliminated all linguistic summaries with the validity value $T = 0$ and still have three linguistic summaries with the value of $T < 0.8$. These disadvantages of those algorithm models may result from no data pattern satisfying the filter criterion *F*, and using too few words in the quantifier set, which has only five words '*none*', '*few*', '*half*', '*much*', and '*most*', leading to not fully describe the data elements. In addition, the computational fuzzy set-based semantics of linguistics words are designed human experts, so they may depend on the intuition recognition of them.

To overcome the limitations of the genetic algorithm models proposed in [13, 14] described above, Ho et al. proposed an extraction model [15] that the multi-semantic structures of linguistic variables are generated automatically by utilizing enlarged hedge algebras [16]. After that, Lan et al. proposed a genetic algorithm model combining the greedy strategy for extracting an optimal set of linguistic summaries [17]. This hybrid algorithm model extracts a set of

linguistic summaries without the value of T less than 0.8. However, the set of values of the fuzzy parameters (FPs) of enlarged hedge algebras used to generate multi-semantic structures for dataset's attributes is determined based on the experience of human experts, so they may not be optimal. Therefore, Phong et al. proposed an algorithm to optimize the set of values of fuzzy parameters to improve the quality of the set of linguistic summaries extracted from numeric dataset, in which the swarm optimization algorithm (PSO) combines with genetic algorithm and greedy strategy to simultaneously optimize the set of values of FPs and the set of extracted linguistic summaries [18]. The computational fuzzy set-based semantics of linguistic words used to construct the multi-semantic structures in [15, 17, 18] have their membership function in the form of trapezoid that can represent the interval semantic core of linguistic words. However, this type of membership function has its edges represented by a linear function with large slope, so it is not flexible and causes large information loss. The computational fuzzy set-based semantics in the form of *S*-function was first proposed in [19] and effectively applied to regression [19] and classification problems [20, 21]. It is a nonlinear function, so it is suitable for both representing the variation of the inherent semantics and the interval semantic core of linguistic words. In this paper, the membership function in the form of *S*-function is applied to construct multi-semantic structures for linguistic variables to improve the quality of the extracted set of linguistic summaries. Besides, the applied hybrid genetic algorithm in [18] is parallelized to reduce running time. Experimental results with the creep database have demonstrated the effectiveness of the proposed method.

2. RESEARCH METHOD

2.1. Extracting linguistic summaries from numeric data with quantifier word

To summarize numeric datasets using sentences in natural language, Yager [1] proposed a sentence structure extracted as a fuzzy clause with quantifiers. The problem of extracting a set of summary sentences, so-called linguistic summaries, from a numeric dataset is stated as follows:

Let $Y = \{y_1, y_2, \ldots, y_n\}$ be a set of objects (records) in the dataset such as a set of customers of a bank; $A = \{A_1, A_2, \ldots, A_m\}$ is the set of attributes to be considered of objects in set Y such as *AGE*, *SALARY*, *MARITAL*, … Denote $A_i(y_j)$ the value of the attribute A_i of object y_j. The dataset is given by the set $D = \{\{A_1(y_1), A_2(y_1), \ldots, A_m(y_1)\}, \ldots, \{A_1(y_n), A_2(y_n), \ldots, A_m(y_n)\}\}$ is the input to the problem of extracting linguistic summaries. The output of the problem is a set of linguistic summaries containing quantifier words in one of the following general forms:

$$Q\ y \text{ are } S \tag{1}$$

$$Q\ F\ y \text{ are } S \tag{2}$$

where S is the summarizer of the linguistic summary expressed by one word in the value domain of the linguistic variable. Q is a quantifier with semantics that represents the proportion of objects that satisfy summarizer S in the entire dataset D as in sentences like the form (1) or in the group of objects that meet the filter criterion F as in sentences like the form (2). The filter criterion F is optional to identify a group of objects in the set of objects Y considered in the summary statement. For example, a fuzzy filter criterion like *AGE* = '*young*' means only considering subjects in the age group '*young*'. The validity value T is a value in the interval [0, 1] that evaluates the correctness of the linguistic summary. The value of T is considered the truth value of a fuzzy proposition with a quantifier and is calculated according to one of the following two formulas [14, 17, 18]:

$$T(Q \text{ y are } S) = \mu_Q\left[\frac{1}{n}\sum_{i=1}^{n}\mu_S(y_i)\right] \quad (3)$$

$$T(Q\,F \text{ y are } S) = \mu_Q\left[\frac{\sum_{i=1}^{n}\left(\mu_F(y_i)\wedge\mu_S(y_i)\right)}{\sum_{i=1}^{n}\mu_F(y_i)}\right] \quad (4)$$

where μ_Q, μ_F, and μ_S are the membership function values of the fuzzy sets representing the semantics of linguistic words in the components *Q*, *F*, and *S*, respectively.

Among the linguistic summary candidates, only the ones which have their value of *T* greater than a given threshold δ (e.g., δ = 0.8 [14]) are put into the set of extracted linguistic summaries to ensure the quality of the extracted set. In addition, some other evaluation measures such as imprecision, coverage, focus, and appropriateness [11, 14] can be used to evaluate the goodness of the linguistic summary.

Although the threshold δ has been set, the number of extracted linguistic summaries is still very large. Therefore, Donis-Diaz and his colleagues in [13, 14] applied genetic algorithms to extract a set of the optimal linguistic summaries based on goodness and diversity measures.

In [14], the goodness *Gn* of a linguistic summary is evaluated by formula (5), where St(*Q*) is the weight of the pre-selected quantifier word *Q* based on the priority of the quantifier words. In [13, 14], the values of St(*Q*) are St("*Most*") = 1, St("*Much*") = 0.75, St("*Half*") = 0.20, St("*Some*") = 0.15, St("*Few*")=0.05. In both studies, the goodness *Gd* of a set of linguistic summaries is the average of the goodness of each linguistic summary in the set according to formula (6), where *l* is the number of linguistic summaries in the set.

$$Gn = T \cdot St(Q) \quad (5)$$

$$Gd = \frac{\sum_{i=1}^{l} Gn_i}{l} \quad (6)$$

In [13, 14], the diversity of a set of linguistic summaries is computed by formula (7):

$$De = \frac{C}{l} \quad (7)$$

where *l* is the number of linguistic summaries and *C* is the number of clusters when clustering the linguistic summary set specified by the similarity function *L* as follows:

$$L(p1, p2) = \begin{cases} Yes & if \ \sum_{k=0}^{m} H(p1_k, p2_k) < 2 \\ No & otherwise \end{cases} \quad (8)$$

In formula (8), if the function $L(p1, p2)$ has its value '*Yes*', two linguistic summaries $p1$ and $p2$ are similar. $p1_k$ and $p2_k$ are two vectors which have $m + 1$ elements, where their first elements $p1_0$ and $p2_0$ are the indexes of the quantifier word *Q* in *Dom*(*Q*) and the elements $p1_i$ and $p2_i$ are the indexes of the words in the linguistic value domain $Dom(A_i)$ of the linguistic variable associated with the attribute A_i of the vector representing two linguistic summaries $p1$ and $p2$. If the attribute A_i is not in the linguistic summary, the i^{th} element in the vector representing the linguistic summary takes the value of 0. The function $H(p1_k, p2_k)$ is used to compare the k^{th} elements of two vector and computed by formula (9) as follows:

$$H(p1_k, p2_k) = \begin{cases} 1 & if \quad |p1_k - p2_k| > round\left(20\% * size\left(Dom(A_k)\right)\right) or \\ & if \quad p1_k = 0 \; and \; p2_k \neq 0 \quad or \\ & if \quad p1_k \neq 0 \; and \; p2_k = 0 \\ 0 & otherwise \end{cases} \tag{9}$$

Based on the goodness *Gd* and diversity *De* of the set of linguistic summaries, the fitness function *Fit* of each individual representing a set of linguistic summaries is computed as follows:

$$Fit = m_g Gd + m_d De \tag{10}$$

where m_g and m_d are the weight of *Gd* and *De*, respectively, satisfying the condition $m_g + m_d = 1$. The authors in [14] chose the value of m_g is 0.7 and m_d is 0.3.

2.2. Extracting a set of optimal linguistic summaries ensuring the interpretability of their contents

The fuzzy partitions manually designed for linguistic variables in the proposals in [13, 14] are singular and the number of linguistic words corresponding to each variable is limited to 7 ± 2. To overcome those limitations of the existing proposals, the authors in [15] proposed a method to design fuzzy partitions in the form of an interpretable and scalable multi-semantic structure, which ensures two relationships based on the inherent semantics of linguistic words: the semantic order relationship and the generality and specificity relationship of linguistic words by utilizing the mathematical formalism of the algebras theory. These relationships are preserved when mapping from a set of linguistic words to a set of computational fuzzy set-based semantics. When users need to expand the number of linguistic words used to design the multi-semantic structures, they only need to add more specific linguistic words at the level $k + 1$ while preserving the semantics of the linguistic words being used. This approach ensures the interpretability of the content of the extracted linguistic summaries. Besides, a new genetic algorithm combining with greedy strategy is proposed to limit the number of extracted linguistic summaries that have their validity value $T = 0$. The sentence structure of the extracted linguistic summaries is in the form of (11) as in [15].

"*Qo*s are $o(E_s)$," and "*Qo*s that are $o(F_q)$ is $o(E_s)$" (11)

where $o(E_s)$ is the summarizer, $o(F_q)$ is the filter criterion.

The greedy strategy of extracting a linguistic summary **Random-Greedy-LS** proposed in [16] is summarized as follows:

- *Step 1*: Randomly generate a filter criterion $o(F_q)$ including the attributes and their linguistic words. Compute the support of $o(F_q)$ using the formula $supp(o(F_q)) = \sum_{i=1}^{n} \mu_{F_q}(y_i) / n$, where n is the number of records in the dataset. If $supp(o(F_q)) > \beta$ (a threshold input by user), the filter criterion $o(F_q)$ is accepted and goto Step 2, otherwise rondomly generate another filter criterion $o(F_q)$.
- *Step 2*: Randomly select an attribute in the filter criterion $o(E_s)$ according to the pre-specified number of attributes, scan the combinations of used linguistic words of the attributes in $o(E_s)$ to get a combination of lingusitic words that has the maximum value of the expression $r = \frac{\sum \mu_{o(Fq)} \wedge \mu_{o(Es)}}{\sum \mu_{o(Fq)}}$.

- *Step 3*: Select a quantifer word Q^* in the used word set that makes $T = \mu_{Q^*}(r)$ reaches the maximum value. If there are many words Q^*, select the one that has the largest semantic order.

The output of the above algorithm is a linguistic summary with large value of goodness *Gn* among the linguistic summaries that have the same filter criterion $o(F_q)$ and structure $o(E_s)$.

To extract a set of linguistic summaries which has its goodness and diversity as large as possible, the authors in [17] proposed a genetic algorithm combining with the greedy strategy described above, so-called **Greedy-GA**. That hybrid algorithm is implemented based on three genetic operators: Selection operator chooses a certain proportion of the best individuals based on the value of the fitness function *Fit* according to formula (10) to the next generation; Crossover operator exchanges the linguistic summary between the individuals; Mutation operator replaces several linguistic summaries of an individual with the new ones generated by the procedure **Random-Greedy-LS**. Thus, the Crossover operator does not change the goodness of each linguistic summary but changes the goodness of the set of linguistic summaries, and the Mutation operator changes both the goodness *Gd* and diversity *De* of the set of linguistic summaries. The experimental results showed that the **Greedy-GA** method outperformed the **Hybrid-GA** method in [14]. However, the fuzzy parameter values of hedge algebras associated with linguistic variables are specified by human experts and depend on their experience. Therefore, the fuzzy parameter values used in the experiments may not be good enough. The authors in [18] proposed an optimization scheme for concurrently optimizing the fuzzy parameter values and the extracting linguistic summary sets. The performance of the proposed model was shown by the experimental results.

2.3. The proposed method of extracting a set of optimal linguistic summaries

It can be seen that the membership function of computational fuzzy-set-based semantics of linguistic words used to construct the multi-semantic structures in the existing models in [15, 17, 18] is in the shape of trapezoid. Although this type of membership function can represent the interval semantic core of linguistic words, it has two edges represented by a linear function with a large slope. Therefore, it is not flexible and may cause large information loss. The computational fuzzy set-based semantics in the form of *S*-function was first proposed in [19] and effectively applied to regression [19] and classification problems [20, 21]. It is shown in Figure 1 that this type of function is nonlinear, so it is suitable for both representing the variation of the inherent semantics and the interval semantic core of linguistic words.

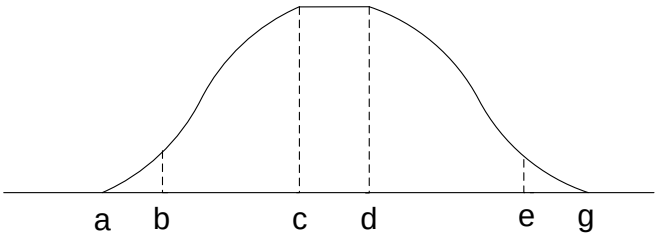

Figure 1. A type of membership function of fuzzy set in the form of *S*-function [19].

Assume that *v* is a datum, the function represents the membership degree of *v* to the left part of the *S*-function, S_{left} as follows:

$$S_{left} = \begin{cases} 0, & 0 \le v \le a \\ \dfrac{(v-a)^2}{(b-a)(c-a)}, & a \le v \le b \\ 1 - \dfrac{(v-c)^2}{(c-b)(c-a)}, & b \le v \le c \\ 1, & v \ge c \end{cases} \quad (12)$$

and the function represents the membership degree of v to the right part of the S-function, S_{right} as follows:

$$S_{right} = \begin{cases} 1, & 0 \le v \le d \\ 1 - \dfrac{(v-d)^2}{(d-e)(d-g)}, & d \le v \le e \\ \dfrac{(v-g)^2}{(e-d)(g-d)}, & e \le v \le g \\ 0, & v \ge g \end{cases} \quad (13)$$

In this paper, to make use of the advantages of S-function, it is applied to construct multi-semantic structures for linguistic variables to improve the quality of the extracted set of linguistic summaries. We also applied the multi-semantic structure used in [21] that ensures the interpretability in the sense of Tarski [22] as shown in Figure 2.

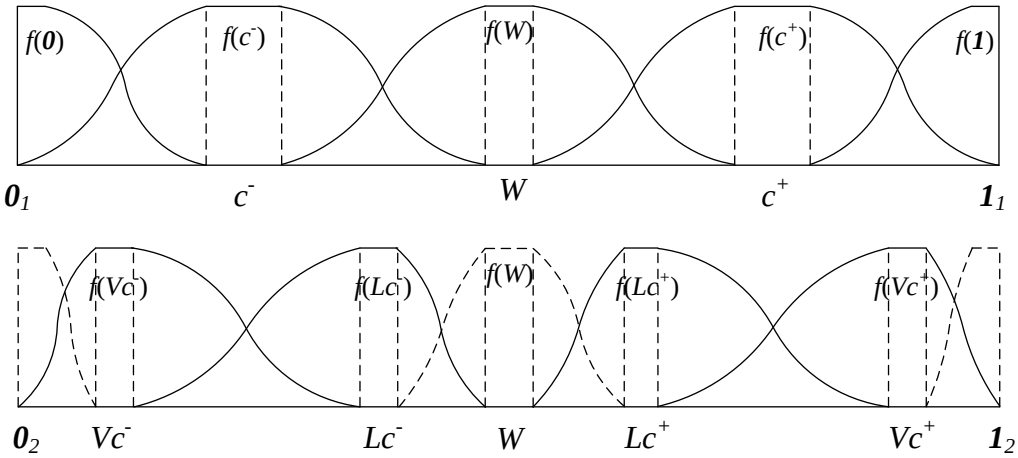

Figure 2. A multi-semantic structure using the membership function in the form of S-function [19].

Besides, the hybrid genetic algorithm **FPO_GreadyGA** in [18] is structured by two nested optimization algorithms. The outer algorithm is a particle swarm optimization (PSO) [23] used to optimize fuzziness parameter values and the inner algorithm is a genetic algorithm combining with greedy strategy used to optimize the set of linguistic summaries. Because of being a nested algorithm, it runs quite slowly. Therefore, in this paper, it is parallelized to reduce running time and its new structure is described in the algorithm **Parallel_FPO_GreadyGA**.

Algorithm Parallel_FPO_GreadyGA

Input: Dataset $\mathcal{D}$;

Parameters: G_{max}, N, syntactic semantics of hedge algebras;

//G_{max}: the number of cycles, N: the number of particles;

Output: The optimal fuzziness parameter values *bestGlobalFit*;

Begin

Randomly generate initial swarm $S_0 = \{X_i^0 \mid i = 1, \ldots, N\}$ with $X_i^0 = \{m(c^-), \mu(Little)\}$;

Initialize PG^t and P_i^t to 0 to store the global and personal best position;

bestGlobalFit = 0;

For each particle x_i **parallel do begin**

F_i^0 = **Greedy-GA**(X_i^0); //Call **Greedy-GA** algorithm

If F_i^0 > *bestGlobalFit* **then begin**

bestGlobalFit = F_i^0;

$PG^0 = X_i^0$;

End;

Update the personal best position P_i^0 of particle x_i based on F_i^0;

End;

t = 1;

Repeat

For each particle x_i **parallel do begin**

Update new velocity V_i^t of particle x_i;

Update new position X_i^t of particle x_i;

F_i^t = **Greedy-GA**(X_i^t); //Call genetic algorithm combing with greedy strategy

If F_i^t is greater than F_i^{t-1} **then begin**

Update the personal best position P_i^t of particle x_i based on F_i^t;

If F_i^t > *bestGlobalFit* **then begin**

bestGlobalFit = F_i^t;

$PG^t = X_i^t$;

End;

End;

End;

$t = t + 1$;

Until $t = G_{max}$;

Return *bestGlobalFit* và $PG^{G_{max}-1}$;

End.

In the proposed algorithm described above, the enlarged hedge algebras [16] is applied to generate multi-semantic structures, so each particle x_i at generation t of the swarm represents a set of fuzziness parameter values $X_i^t = \{m(0),\ m(c^-),\ m(W),\ m(1),\ \mu(Little),\ \mu(h_0)\}$, where $m(0)$, $m(c^-)$, $m(W)$, $m(1)$, $\mu(L)$, and $\mu(h_0)$ are the fuzziness measures of the smallest word constant 0, the generator word c^-, the neutral element W, the largest word constant 1, the negative hedge *Little* (L), the artificial hedge h_0, respectively. Because there are only two hedges *Little* and *Very* applied to the model, the fuzziness measure of the positive generator c^+ is computed as $m(c^+) = 1 - m(0) - m(c^-) - m(W) - m(1)$ and the one of positive hedge *Very* (V) is computed as $\mu(Very) = 1 - \mu(Little) - \mu(h_0)$. Each set of given fuzziness parameter values represented by X_i^t is the input of the algorithm **Greedy-GA**. The multi-semantic structures designed using fuzzy sets of the shape of *S*-function of linguistic variables are automatically generated for the reasoning process. During the reasoning process, computational fuzzy set-based semantics of the shape of *S*-function are interacted with data to specify the parameter values of the algorithm model. Furthermore, the algorithm is parallelized in such a way that at each iteration of outer loop, the inner loop separates particles in the swarm for parallel execution. Therefore, each particle is independently executed on a single core of a computer. It means that each inner loop independently executes the algorithm **Greedy-GA** on a core. The disadvantage of this algorithm is that the maximum number of particles is the number of cores of the computer used to experiment. However, modern processors have a lot of cores, so it is not too much of a concern.

The output of the proposed algorithm is a set of values of the fuzziness parameters stored in the global variable $PG^{G_{max}-1}$ corresponding to the value of the best fitness function *bestGlobalFit* in the last iteration. In a practical implementation, the variable *bestGlobalFit* can be an object variable containing the optimal set of linguistic summaries.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1. Experiment setup

In this paper, the experimental program is implemented by C# programming language, .NET Framework 4.0 and Task Parallel Library, running on a computer with Intel Core i7-1360P 2.2GHz (16 cores), 16GB RAM, and Windows 11 64-bit. The experimental data set is creep [14]. The parameter values of the experimental algorithm model are as follows:

- The number of iterations of the PSO algorithm in the outer loop is 5, the number of particles is 10 (*less than the number of cores of the computer*), the Inertia coefficient is 0.7, the self-cognition and social-cognition coefficients are both 1.5.

- The number of generations of the genetic algorithm in the inner loop is 100, the number of individuals per generation is 20, the selection rate is 0.15, the crossover rate is 0.8, and the mutation rate is 0.1.

- The number of sentences extracted in each set of linguistic summaries is 30.

- The constraints on the fuzziness parameter values are set as in [18] and through the experimental process as follows:

+ The constraints on the attributes of *creep*: $0{,}001 \le m(0) \le 0{,}05;\ 0{,}2 \le m(c^-) \le 0{,}35;\ 0{,}02 \le m(W) \le 0{,}05; 0{,}35 \le m(1) \le 0{,}4; 0{,}2 \le \mu(L) \le 0{,}5;\ 0{,}05 \le \mu(h_0) \le 0{,}3.$

+ The constraints on the quantifier word *Q*: $0{,}001 \le m(0) \le 0{,}05;\ 0{,}2 \le m(c^-) \le 0{,}45;\ 0{,}001 \le m(W) \le 0{,}15; 0{,}002 \le m(1) \le 0{,}05$; $0{,}2 \le \mu(L) \le 0{,}5;\ 0{,}05 \le \mu(h_0) \le 0{,}3.$

+ The constraints on the other attributes: $0{,}001 \le m(0) \le 0{,}05;\ 0{,}2 \le m(c^-) \le 0{,}45;\ 0{,}002 \le m(W) \le 0{,}15; 0{,}001 \le m(1) \le 0{,}05$; $0{,}2 \le \mu(L) \le 0{,}5;\ 0{,}05 \le \mu(h_0) \le 0{,}3.$

+ The specificity of the word domains of all attributes and quantifier *Q*: $k = 3$.

3.2. Experimental results and comparisons

The creep dataset [14] is used for our experiments. The experimental algorithm models, which are **Greedy-GA** in [17], **FPO_GreadyGA** in [18], and our proposed algorithm **Parallel_FPO_GreadyGA** in this paper were run experimentally 10 times. The results of the 10 runs were averaged and the last figures are shown in Table 1.

Table 1. The comparison of the experimental results of the proposed model and the existing models FPO_GreadyGA [18], Greedy-GA [17], and Hybrid-GA [14].

Models	Fitness function value *Fit*	The average truth values *T*	The number of summaries with *Q* > *a half*	The number of summaries with *T* > 0,8	The number of summaries with *T* = 0
Hybrid-GA [14]	0.6659	0.9139	17.8	27.0	1.0
Greedy-GA [17]	0.7905	0.9951	18.8	30.0	0.0
FPO_GreadyGA [18]	0.8828	0.9970	21.9	30.0	0.0
Parallel_FPO_GreadyGA	0.8872	0.9997	22.7	30.0	0.0

It can be seen in Table 1 that with the same experimental values of models' parameters, the proposed algorithm model **Parallel_FPO_GreadyGA** has its average fitness function value *Fit* is 0.9997, the average truth values *T* is 0.8872, the number of linguistic summaries with *Q* > *a haft* is 22.7, all greater than those of **FPO_GreadyGA** in [18], **Greedy-GA** in [17], and **Hybrid-GA** in [14]. Besides, the proposed model has a maximum number of summary sentences with truth value *T* > 0.8 of 30 summary sentences and no summary sentences with truth value *T* = 0, equivalent to the results of both **FPO_GreadyGA** and **Greedy-GA** models. Furthermore, considering the diversity aspect of the extracted linguistic summary set, the average diversity values of the proposed model is 0.91668, greater than the one of **FPO_GreadyGA** that is 0.90667. It proves that the increasement of the fitness function value is contributed by both goodness and diversity values. With the above comparison results, it can be stated that the proposed algorithm model with the computational fuzzy set-based semantics in the form of *S*-function gives better experimental results than the other three compared models.

Table 2. The comparison of running times of the proposed model between running sequentially and running in parallel.

Comparison criterion	Running sequentially	Running in parallel	Percentage of reduction
Running time (in second)	19772	5714	-71.1%

Considering the running time, the comparison results are shown in Table 2. Specifically, when running sequentially on a single core, the average running time of a cycle of the outer

PSO is 19772 seconds. Whereas, when the number of particles is set to 10, they will be run separately on 10 cores of the computer and the average running time of a cycle of the outer PSO will be reduced to 5714 seconds, according to a reduction of 71.1%. Therefore, we can state that our new proposed algorithm model is efficient in both aspects of running time and the goodness and diversity measures.

4. CONCLUSION

Linguistic summarization of data is an essential problem in data mining field because it extracts useful knowledge in the form of sentences of natural language, so-called linguistic summaries, from numeric data. The existing studies focuses on extracting a set of compact linguistic summaries with high validity and diversity by utilizing fuzzy set theory and genetic algorithm. Recently, enlarged hedge algebra was utilized to automatically generate multi-semantic structures for linguistic variables ensuring the interpretability of the content of the linguistic summaries. However, the membership function of computational fuzzy-set-based semantics of linguistic words is usually in the shape of trapezoid. This type of membership function has its edges represented by a linear function with large slope, so it is not flexible and causes large information loss. In this paper, a membership function of the form *S*-function is applied to improve the quality of extracted linguistic summaries because this type of function is nonlinear, so it is suitable for both representing the variation of the inherent semantics and the interval semantic core of linguistic words. Besides, the applied algorithm is parallelized to reduce running time. The experimental results with the creep dataset have demonstrated the effectiveness of the proposed method on both aspects of running time and the goodness and diversity measures.

ACKNOWLEDGMENT

This research is funded by University of Transport and Communications (UTC) under grant number T2023-CN-008TD.

REFERENCES

[1]. R. R. Yager, A new approach to the summarization of data, Information Sciences, 28 (1982) 69-86). https://doi.org/10.1016/0020-0255(82)90033-0

[2]. J. Kacprzyk, R. R. Yager, S. Zadrożny, A fuzzy logic based approach to linguistic summaries of databases, International Journal of Applied Mathematics and Computer Science, 10 (2000) (813-834).

[3]. J. Kacprzyk, S. Zadrożny, Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools, Information Sciences, 173 (2005) 281-304. https://doi.org/10.1016/j.ins.2005.03.002

[4]. C. A. D. Díaz, R. B. Pérez, E. V. Morales, Using Linguistic Data Summarization in the study of creep data for the design of new steels, in Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on,160-165. https://doi.org/10.1109/ISDA.2011.6121648

[5]. T. Altintop, R. R. Yager, D. Akay, F. E. Boran, M. Ünal, Fuzzy Linguistic Summarization with Genetic Algorithm: An Application with Operational and Financial Healthcare Data, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 25 (2017) 599-620. https://doi.org/10.1142/S021848851750026X

[6]. R. J. Almeida, M.-J. Lesot, B. Bouchon-Meunier, U. Kaymak, G. Moyse, Linguistic summaries of categorical time series for septic shock patient data, Fuzz-IEEE 2013-IEEE International Conference on Fuzzy Systems, Hyderabad, India. IEEE, (2013), 1-8. https://doi.org/10.1109/FUZZ-IEEE.2013.6622581

[7]. J. Kacprzyk R. R. Yager, Linguistic summaries of data using fuzzy logic, International Journal of General System, 30 (2001) 133-154. https://doi.org/10.1080/03081070108960702
[8]. M. D. Peláez-Aguilera, M. Espinilla, M. R. Fernández Olmo, J. Medina, Fuzzy linguistic protoforms to summarize heart rate streams of patients with ischemic heart disease, Complexity, 2019 (2019) 1-11. https://doi.org/10.1155/2019/2694126
[9]. A. Duraj, P. S. Szczepaniak, L. Chomatek, Intelligent Detection of Information Outliers Using Linguistic Summaries with Non-monotonic Quantifiers, International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, (2020) 787-799. https://doi.org/10.1007/978-3-030-50153-2_58
[10]. A. Jain, M. Popescu, J. Keller, M. Rantz, B. Markway, Linguistic summarization of in-home sensor data, Journal of biomedical informatics, 96 (2019) 103240. https://doi.org/10.1016/j.jbi.2019.103240
[11]. A. Wilbik, I. Vanderfeesten, D. Bergmans, S. Heines, W. van Mook, Linguistic summaries for compliance analysis of a glucose management clinical protocol, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), (2018) 1-7. https://doi.org/10.1109/FUZZ-IEEE.2018.8491449
[12]. F. E. Boran, D. Akay, A generic method for the evaluation of interval type-2 fuzzy linguistic summaries, IEEE transactions on cybernetics, 44 (2013) 1632-1645. https://doi.org/10.1109/TCYB.2013.2291272
[13]. C. Donis-Diaz, R. Bello, J. Kacprzyk, Linguistic data summarization using an enhanced genetic algorithm, Technical Transactions – Automatic Control, 2013 (2013) 3-12.
[14]. C. Donis-Diaz, A. Muro, R. Bello-Pérez, E. V. Morales, A hybrid model of genetic algorithm with local search to discover linguistic data summaries from creep data, Expert Systems with Applications, 41 (2014) 2035-2042. https://doi.org/10.1016/j.eswa.2013.09.002
[15]. C. H. Nguyen, T. L. Pham, T. N. Nguyen, C. H. Ho, T. A. Nguyen, The linguistic summarization and the interpretability, scalability of fuzzy representations of multilevel semantic structures of word-domains, Microprocessors and Microsystems, 81 (2021) 103641. https://doi.org/10.1016/j.micpro.2020.103641
[16]. C. H. Nguyen, T. S. Tran, D. P. Pham, Modeling of a semantics core of linguistic terms based on an extension of hedge algebra semantics and its application, Knowledge-Based Systems, 67 (2014) 244-262. https://doi.org/10.1016/j.knosys.2014.04.047
[17]. T. L. Pham, C. H. Nguyen, D. P. Pham, Extracting an optimal set of linguistic summaries using genetic algorithm combined with greedy strategy, Journal on Information Technologies & Communications, 2020 (2020) 75-87. https://doi.org/10.32913/mic-ict-research.v2020.n2.954
[18]. D. P. Pham, T. L. Pham, X. T. Tran, A fuzzinness parameter optimization method to extract the optimal set of linguistic summaries from numeric data, TNU Journal of Science and Technology, 229 (2024) 49-57. https://doi.org/10.34238/tnu-jst.9824
[19]. V. T. Hoang, D. D. Nguyen, C. H. Nguyen, A Method to design semantic of linguistics based on the enlarged hedge algebra and applied to building FRBS for solving regression, Journal on Information Technologies & Communications, 38 (2017) 51-57 (In Vietnamese).
[20]. D. D. Nguyen, D. P. Pham, D. V. Pham, D. T. Nguyen, A design method of computational semantics of linguistic words for fuzzy rule-based classifier, Journal on Information Technologies & Communications, 2020 (2020) 9-18 (In Vietnamese). https://doi.org/10.32913/mic-ict-research-vn.v2020.n1.914
[21]. D. D Nguyen, A method for designing fuzzy rule-based classifier using *S*-function based fuzzy set guarantee interpretability, Transport and Communications Science Journal, 75 (2024) 1335-1347 (In Vietnamese). https://doi.org/10.47869/tcsj.75.3.2
[22]. A. Tarski, A. Mostowski, R. Robinson, Undecidable Theories. North-Holland, 1953.
[23]. J. Kennedy, R. C. Eberhart, Particle Swarm Optimization, Proceedings of the IEEE International Conference on Neural Networks, Piscataway, New Jersey. IEEE Service Center, (1995) 1942-1948.