

A Novel Linguistic Summarization of Time Series Data Based on Enlarged Hedge Algebra Formalism and Genetic Algorithm

Tran Xuan Thanh¹, Pham Dinh Phong², Pham Thi Lan³

¹ Faculty of Information Technology, East Asia University of Technology, Bacninh, Vietnam | Graduate University of Sciences and Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam

² Faculty of Information Technology, University of Transport and Communications, Hanoi, Vietnam

³ Faculty of Information Technology, Hanoi National University of Education, Hanoi, Vietnam

Article Info

Article history:

Received Jun 27, 2025

Revised Mar 25, 2026

Accepted Mar 31, 2026

fromKeywords:

Linguistic summarization

Hedge algebras

Interpretability

Time series data

Genetic algorithm

ABSTRACT

The linguistic summarization of time series data (TSD) has been examined extensively because the extracted knowledge represented as summary sentences in natural language is interpretable for all people. The existing extracting methods use manually designed fuzzy partitions of value domains, so the word semantic depends on the subjective opinions of designers. Besides, the number of linguistic words with the fuzzy set-based computational semantics used to describe the TSD, the quantifier, and the summarizer is usually limited to 7 ± 2 . That cardinality is not rich enough to describe the special characteristics in a certain period in the TSD. In this paper, enlarge hedge algebra is applied to create a mathematical formalism for automatically designing interpretable and scalable multi-level semantic structures for the corresponding value domains of linguistic variables and these structures can be arbitrarily extended as needed. The objectives of the applied genetic algorithm were also adjusted to improve the optimization goals. The experimental results on the patient admission data have shown that our proposed methods obtain the outstanding results in terms of accuracy, conciseness, and coverage.

Copyright © 2026 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Pham Dinh Phong

Faculty of Information Technology, University of Transport and Communications, Hanoi, Vietnam

Email: phongpd@utc.edu.vn

1. INTRODUCTION

Time series data (TSD) is continuously generated from daily activities such as the number of daily hospital admissions, stock exchange rates, average daily temperature, number of deaths due to traffic accidents, etc. TSD contains valuable information and hidden knowledge, which plays an important role in making decision and business plan. This information and knowledge are the basis for businesses to make the right decisions. Therefore, the problem of extracting information and hidden knowledge contained in TSD has important practical significance and has received the much attention of scientist community. Using diagrams and charts to visualize information and knowledge contained in TSD is one of the good choices. However, reading and understanding diagrams and graphs also requires certain specialized knowledge, sometimes requiring the data analyses of professional experts. On the other hand, there may be some special data values on the time series corresponding to certain time periods, change very little, bringing great challenges to graphs-based data analyses. Extracting information and knowledge represented as summary sentences, namely linguistic summaries (LSs), first introduced by Yager in 1982 [1] is easier to understand than numerical data or graphs. LSs represented as sentences in natural language help managers discover patterns from both database [2-13] and time series data [14-24] to make informed decisions to address problems that may arise in the future. For example, the sentence “*Most of the days in August, the patient inflow is very high or high*” [17] help hospital managers proactively arrange human resources and materials to meet high or very high medical examination and treatment needs in August, thereby optimizing resources.

Many existing studies in extracting LSs from TSD have shown good results in terms of three aspects: accuracy, conciseness, and time coverage [14-24]. In [17], Rita et al. proposed using 5 linguistic words to partition the TSD, the quantifier, and the summarizer. The TSD was represented in the form of a chart in such a way that the time dimension is hierarchically partitioned and represented by the horizontal axis, and the values of time series are represented by the vertical axis. The time dimension is partitioned by season, climate, and month. In that study, the authors also introduced the prototype of summary sentence in the form "*Q of T are S*", where

- *Q* is the quantifier word such as *most*, *a half*, *few*, etc.
- *T* is the element that represents a specific time or period of time.
- *S* is a statement that describes the characteristics of a set of numerical values in a time interval *T* corresponding to a quantifier *Q*. The computational semantics of *S* is often a fuzzy set in the form of trapezoid.

The truth value of a LS is specified by the membership values of fuzzy sets associated with the linguistic words appeared in the summaries such as "*most*", "*a half*", "*few*", ... To ensure the summary quality, a given threshold of truth value is imposed, e.g. 0.8 [17], so only the summary sentences having its truth value greater than and equal to that given threshold are extracted. However, the cardinality of extracted LSs in general form, i.e. the elements of *Q*, *T*, and *S* are not prespecified, is still large, so it is a huge challenge for the computational volume. In this case, there may be a lot of trivial extracted LSs without condensed information, and they do not bring practical value to users. Because users usually have not enough time to read and understand all large sentences, there is a need of extracting a compact set of sentences containing useful information. Therefore, the optimization algorithms (for example, genetic and greedy [17]) are applied to optimize summary sentences based on the constraints and summary quality evaluation parameters such as brevity, accuracy, and coverage. The genetic algorithm is usually applied to select a compact set of linguistic summaries from the unlabelled time series data instead of modern neural network models such as LSTM (Long-short term memory), Transformer because those neural network models need data to train, and they are not interpretable to human.

Despite the notable achievements, the existing methods of extracting LSs from TSD have the following shortcomings: first, the parameters in the multi-objective optimization problem are contradictory, when the coverage is larger, the conciseness tends to decrease and vice versa, or the increased truth value can cause the coverage to decrease. Second, the number of linguistic words used to describe the TSD, the quantifier, and the summarizer is usually limited to 7 ± 2 . That cardinality is not rich enough to describe the special characteristics in a certain period in the TSD. For example, in a certain month, the number of hospital admissions spikes, while the quantifier at the highest semantic level is only "*very high*", using the word "*very very high*" instead would better describe the data characteristics during this period. Therefore, there is a need of arbitrarily extending word set by adding the more specific words to describe special features of TSD more accurately.

Hedge algebra was first proposed by Nguyen and Wechler in 1990 [25, 26] and it is enlarged in 2014 [27] to represent the interval semantic core of words. The word domain of linguistic variables (LVs) can be automatically generated by acting linguistic hedges on two primary words [27]. For example, given two primary words "*low*" and "*high*" (denoted by c^- and c^+ , respectively), two linguistic hedges *very* (*V*) and *rather* (*R*), and the artificial hedge h_0 , the word set X_2 with the specific level k of 2 generated by acting the linguistic hedges on two primary words is $\{0, VVc^-, Vc^-, RVc^-, h_0c^-, RRc^-, Rc^-, VRc^-, W, VRc^+, Rc^+, RRe^+, h_0c^+, RVc^+, Vc^+, VVc^+, 1\}$. Two primary words and the hedges are linearly ordered, so the generated words in X_2 are linearly ordered by the inherent semantics of words. Noted that $h_0c^- = c^-$, it means that h_0c^- becomes a fixed point and its fuzziness measure values can be used to represent the interval semantic core of c^- and it is used to represent the small base of the associated trapezoidal fuzzy set. This is also true for the words h_0c^+ and h_0x , where x is any non-constant word. By utilizing enlarged hedge algebras (EHAs), a mathematical formalism for automatically generating the interpretable and scalable multi-level semantic structures for the corresponding value domains of LVs are constructed. The generated multi-level semantic structures ensure the generality and specificity of their corresponding word sets, so they can be extended as needed by arbitrarily adding the more specific words while the fuzzy set-based computational semantics (FSCS) of the existing words are still retained [28, 29].

In this paper, the above-mentioned formalism is applied to automatically generate such structures of multi-level semantics for LVs associated with the TSD, the quantifier (*Q*), and the summarizer (*S*). The cardinality of the generated word set of each linguistic domain of LVs at a given period is specified by users and it can be extended when needed to increase the chance of obtaining interesting summary sentences. The number of words used in each linguistic domain of LVs generated by EHAs is rich enough to describe its universe of discourse (UoD), so the truth values of the extracted summaries are usually greater than 0.8. To achieve a compact set of summaries, a genetic algorithm is applied to optimize the candidate set of summary sentences. We assume that the coverage objective is always satisfied, so the fitness function of the genetic algorithm only has a combination of two objectives: brevity and accuracy. Experimental results with a historical TSD of the number of patients admitted per day in one year of a hospital demonstrate the efficiency of the proposed

method. However, in this study, the trapezoidal type-1 fuzzy sets are used to represent the computational semantics of words instead of interval-valued fuzzy sets because applying type-1 fuzzy set can significantly improve the performance of the proposed method and a mathematical formalism for automatically generating interval-valued fuzzy sets requires further studies and mathematical formalism proofs. Therefore, applying interval-valued fuzzy sets will be examined in our future studies.

2. BACKGROUND KNOWLEDGE

2.1. Linguistic summarization of time series data

The linguistic summarization of data stored in a database was first proposed by Yager in 1982 [1]. The output of this summarization method is a summary sentence set represented as fuzzy clauses with quantifiers, so-called linguistic summaries (LSs). Then, this method was applied to mine the TSD by Chiang et al. in 2000 [14]. The problem of extracting LSs from TSD is summarized as follows:

The historical TSD occurred within one year is stored in a list of $DS = \{d_1, d_2, \dots, d_n\}$, where each d_i is a historical data value at the time t_i of $T = \{t_1, t_2, \dots, t_n\}$. The history data is partitioned by a set of linguistic words $S = \{s_1, s_2, \dots, s_h\}$. The time dimension is hierarchically partitioned by month and season as in [17], $\mathfrak{T} = \{T_1, T_2, \dots, T_l\}$. To extract the features of TSD corresponding to a time point or a period, a set of quantifier words $Q = \{q_1, q_2, \dots, q_h\}$ is used, where $q_1 \leq q_2 \leq \dots \leq q_h$ by their inherent semantic order relation.

The algorithm output of linguistic summarization of TSD is a set of LSs of the form ***Q of T are S*** [17]. This set can be represented as $QTS = \{\{Q_1, T_1, s_1\}, \{Q_2, T_2, s_2\}, \dots, \{Q_m, T_m, s_m\}\}$, where

- $Q_i \in Q$ is a quantifier that can be selected from a set of quantifiers such as: “very many”, “a half”, “few”, ... the computational semantics of these words are represented by fuzzy sets.
- $T_i \in \mathfrak{T}$ is a word that indicates a time point, a period by month or by season such as: *August, Hot to Cold, Summer, ...*
- $s_i \in S$ is a conclusion about the value in a period T_i according to the quantifier Q_i , so-called summarizer. The summarizer words can be “low”, “high”, “very high”, ...

2.2. Enlarge hedge algebras for representing interval semantic core of words

Humans use natural language to communicate, store, and process knowledge. The nature of natural language is weakly structured based on the inherent semantic order of linguistic words. Therefore, it can be formalized by a mathematical structure, so-called the hedge algebra.

Hedge algebra (HA) theory was first introduced by C. H. Nguyen and W. Wechler in 1990 [25, 26]. Based on actual application requirements, HA was enlarged to model the interval semantic core of words, namely enlarged hedge algebra (EHA), by C. H. Nguyen et al. in 2014 [27]. Based on HA and EHA, a formalism to generate the FSCS from the inherent word's semantics has been constructed. Therefore, it is the mathematical foundation for the bridge of the computational semantics and the inherent word's semantics. Then, HA has been efficiently applied to different application areas such as data mining [27, 29, 30], image processing [31], fuzzy control [32, 33], ...

Assume that $\mathcal{A} = (X, G, \mathcal{C}, \mathcal{H}, \leq)$ is a hedge algebra, where X is the linguistic domain of a LS $\mathcal{X}$; $G = \{c^-, c^+\}$ is the primary word set including the negative and positive words, respectively; $\mathcal{C} = \{0, W, 1\}$ is the word constant set, where 0, W , and 1 are the smallest, the neutral, and the largest elements, respectively. For example, $\mathcal{C} = \{\text{“absolutely young”, “medium”, “absolutely old”}\}$; $\mathcal{H} = \mathcal{H}^- \cup \mathcal{H}^+$ is the hedge set including negative hedge set $\mathcal{H}^-$ and positive hedge set $\mathcal{H}^+$. For example, $\mathcal{H}^- = \{\text{“little”, “rather”}\}$, $\mathcal{H}^+ = \{\text{“more”, “very”}\}$; $\leq$ expresses the word semantic order relation in the linguistic domain of $\mathcal{X}$.

EHA is enlarged by adding an artificial hedge $\mathfrak{h}_0$ to model the interval semantic core of words. Specifically, when a word x is acted by $\mathfrak{h}_0$, it becomes a fixed point and $\mathfrak{h}_0(x)$ is so-called the semantic core of x . Given a linear hedge algebra $\mathcal{A} = (X, G, \mathcal{C}, \mathcal{H}, \leq)$ of a LS $\mathcal{X}$, an artificial hedge $\mathfrak{h}_0 \notin \mathcal{H}$. On the syntactic aspect, $\mathfrak{h}_0 x \notin X$ and let $X_{en} = X \cup \{\mathfrak{h}_0 x : x \in X\}$, we have $\mathcal{A}_{en} = (X_{en}, G, \mathcal{C}, \mathcal{H}_{en}, \leq)$ is an enlarged hedge algebra of $\mathcal{A}$, where $\mathcal{H}_{en} = \mathcal{H} \cup \{\mathfrak{h}_0\}$, $X_{en} = \mathcal{C} \cup \mathcal{H}_{en}(G) = \mathcal{C} \cup \{\mathfrak{h}_n \dots \mathfrak{h}_1 c : c \in G, \mathfrak{h}_j \in \mathcal{H}_{en}\}$, for $j = 1 \dots n$. Therefore, $X = \mathcal{C} \cup \mathcal{H}(G) \subseteq X_{en} = \mathcal{C} \cup \mathcal{H}_{en}(G)$.

Let $\mathcal{A}_{en,k} = \{x \in X_{en} : k = |x|\}$, where $|x|$ is the length of x , the set of words having the specificity level of k , and $X_{en,(k)} = \{x \in X_{en} : |x| \leq k\} = \bigcup_{1 \leq i \leq k} X_{en,i}$, the union of the words having the specificity level less than and equal to k . Then, $X_{en,1} = G \cup \mathcal{C}$ and for $\forall k \geq 2$ then $X_{en,k} = X_k \cup \{\mathfrak{h}_0 u : u \in X_{en,(k-1)}\}$, i.e., for $k > 0$, $X_{en,(k)}$ include all words having the specificity level k , less than k , and their semantic cores.

Definition 1. [27] A function $m: X_{en} \rightarrow \mathbb{P}[0,1]$ is called the fuzziness measure of $\mathcal{A}_{en}$ provided that:

$$(F1): m(0) + m(c^-) + m(W) + m(c^+) + m(1) = 1;$$

$$(F2): \sum_{x \in \mathcal{H}_{en}} m(\mathfrak{h}_x) = m(x) \text{ với } \forall x \in \mathcal{H}(G);$$

(F3): $\forall x, y \in \mathcal{H}(G), \forall h \in \mathcal{H}_{en}$, the proportion $\frac{m(hx)}{m(x)} = \frac{m(hy)}{m(y)}$ does not depend on any words in X_{en} , so-called the fuzziness measure of the hedge h and denoted by $\mu(h)$.

The fuzziness measure of a linguistic word of $\mathcal{A}_{en}$ satisfies the following properties:

- (1) $\sum_{x \in X_{(k)}} m(x) = 1, k > 0$. For $k = 1, m(0) + m(c^-) + m(W) + m(c^+) + m(1) = 1$;
- (2) $\sum_{h \in \mathcal{H}_{en}} \mu(h) = 1$;
- (3) $m(hx) = \mu(h)m(x)$, for $h \in \mathcal{H}_{en}, \forall x \in \mathcal{H}(\{c^-, c^+\})$ and $hx \neq x$;
- (4) $m(x) = \mu(h_n) \dots \mu(h_1)m(c)$, where $x = h_n \dots h_1 c, c \in \{c^-, c^+\}$, is the canonical representation of $x \in X_{en}$.

Definition 2. [27] Given a $\mathcal{A}_{en}$, a mapping $f_{en}: X_{en} \rightarrow \mathbb{P}([0,1])$ is an interval semantically quantifying mapping provided that:

(Q1) f_{en} always preserves the order relation of words in X_{en} , that is, it is an isomorphism of $(X_{en}, \leq)$ in the set of ordered words. It means that $(\forall x, y \in X_{en}) (x \leq y \Rightarrow f_{en}(x) \leq f_{en}(y))$.

(Q2) $\forall x \in X_{en}, f_{en}(h_0 x) \subseteq m(x)$, i.e., it preserves the order relation of the hedge h_0 .

(Q3) $C(f_{en}(h_0 X_{en})) = [0, 1]$, where C is closure operator $[0, 1]$, $h_0 X = \{h_0 x: x \in X_{en}\}$ and $m(Y) = \{m(x): x \in Y\}$, for $\forall Y \subseteq X_{en}$.

Definition 3. Given a fuzziness measure $m: X_{en} \rightarrow [0, 1]$ of an EHA $\mathcal{A}_{en}$ of the LS $\mathcal{X}$ and each word $x \in X_{en}$ is associated with an interval $\mathfrak{I}(x) \subseteq [0, 1]$. These intervals are called fuzziness intervals of the words of $\mathcal{X}$ provided that:

(FI1): $|\mathfrak{I}(x)| = m(x)$ for $\forall x \in X_{en}$, $|\mathfrak{I}(x)|$ is the length of interval $\mathfrak{I}(x)$;

(FI2): The set $\{\mathfrak{I}(hx) | x \in X_{en}\}$ constitutes a fuzzy partition of $|\mathfrak{I}(x)|$ and its elements have the same order as their associated linguistic words.

Noted that the index of k is usually used to represent the fuzziness interval of x at level k , denoted by $\mathfrak{I}_k(x)$.

The interval semantically quantifying mapping (ISQM) $f_{en}(x)$ of a word x is specified by $f_{en}(x) = \mathfrak{I}(h_0 x)$, $x \in X_{en}$ and this assertion has been proven in [27].

The generality-specificity structure of a LS $\mathcal{X}$, $G^x = (X^x, g)$: the domain of X^x also includes other semantic structure, generality-specificity relation, that is, a word x is more general than a word y and denoted by $g(x, y)$ and vice versa, y is said to be more specific than x . Therefore, the relation g is called a generality-specificity relation.

The word domain X^x consists of two structures, the order-based semantic structure $S^x = (X_{en}^x, \leq)$ and the generality-specificity structure, $G^x = (X_{en}^x, g)$, that is, a variable $\mathcal{X}$ has not only one semantic structure but also many complex issues in it. $G^x = (X_{en}^x, g)$ is called a multi-level semantic structure and is denoted by $S^x = (X_{en}^x, \leq, g)$.

2.3. Representing multi-level semantic structure of word domain based on enlarged hedge algebras

To make sure that the structure $\mathfrak{I}(X^x)$ represents the structure $S^x = (X_{en}^x, \leq, g)$ preserving the structure of S^x or in other words, $\mathfrak{I}(X^x)$ to be interpretable, we need to define two notation relations $\leq$ and $\subset$ on $\mathfrak{I}(X^x)$ because S^x has the order-based relations $\leq$ and the generality-specificity relations g . Each trapezoidal fuzzy set is denoted by a triple $(a, \mathbf{b}, c)$, where two points $a, c \in [0, 1]$, and $\mathbf{b}$ is a subinterval of $[0, 1]$ acting as the core of the triple and $a < \mathbf{b} < c$.

Definition 4. For any constructed trapezoidal fuzzy set $\mathfrak{I}(X^x)$, define:

1) The order-based relation $\leq$ on $\mathfrak{I}(X^x)$: two triples t and t' with $t = (a, \mathbf{b}, c)$ and $t' = (a', \mathbf{b}', c')$ satisfies $t \leq t'$ if and only if their cores satisfy $\mathbf{b} = \mathbf{b}'$ hoặc $\mathbf{b} < \mathbf{b}'$ and satisfy at least one of the inequations $a \leq a'$ and $c \leq c'$.

2) The inclusion relation $\subset$ on $\mathfrak{I}(X^x)$: the two triples t and t' above are said to satisfy $t \subset t'$ if and only if the large base of t is included in the large base of t' , i.e., $(a, c) \subset (a', c')$.

The set $\mathfrak{I}(X^x)$ with two relations $\leq$ and $\subset$ is denoted as $M_{Gr}^x = (\mathfrak{I}(X^x), \leq, \subset)$, which is called the trapezoidal multi-granularity structure of $\mathcal{X}$. In practical applications, the domain of words used on each variable is often limited to a maximum level of specificity level of k .

Definition 5. With every integer number $k > 1$, k -section $\mathfrak{B}_k^A$ of the semantic structure $S^X = (X_{en}^X, \leq, g)$ is a substructure $S_k^X = (X_{(k)}^X, \leq_k, g_k)$ satisfying the following conditions:

- (i) $X_{(k)}^X = \{x \in X^X: |x| \leq k\}$, the set of words having their specificity level not greater than k ;
- (ii) The relations $\leq_k$ and g_k are respectively the relations $\leq$ and g restricted on the word set $X_{(k)}^A$.

Definition 6. For every integer number $k > 1$, a k -section of the trapezoidal multi-granularity structure $M_{Gr}^X = (\mathfrak{F}(X^X), \leq, \subset)$ of X is the structure $M_{Gr,k}^X = (\mathfrak{F}(X_{(k)}^X), \leq_k, \subset_k)$, which is called a level k trapezoidal multi-granularity substructure satisfying the following conditions:

- (i) $\mathfrak{F}(X_{(k)}^X)$, where $X_{(k)}^X$ defined as in Definition 5 is a set of trapezoidal fuzzy sets of the words $X_{(k)}^X$ designed in levels from $l = 1$ to k ;
- (ii) The relations $\leq_k$ and $\subset_k$ are respectively relations $\leq$ and $\subset$ are limited on $\mathfrak{F}(X_{(k)}^X)$.

As proved in [29] that the structure M_{Gr}^X as shown in Figure 1 is an isomorphic image of the semantic structure $S^X = (X_{en}^X, \leq, g)$, meaning that S^X can be interpreted in M_{Gr}^X .

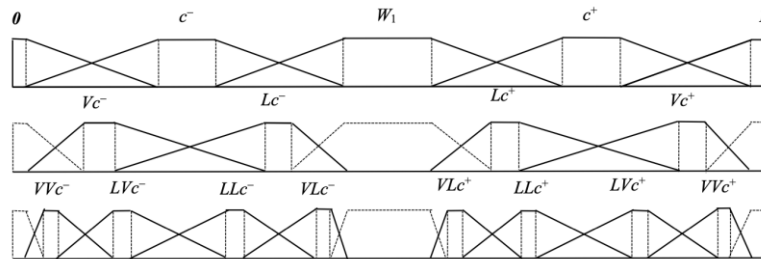


Figure 1. The interpretable and scalable multi-level semantic structure M_{Gr}^X with $k = 3$ [29].

3. THE PROPOSED METHOD OF LINGUISTIC SUMMARIZATION OF TIME SERIES DATA

This section presents our proposed method of extracting an optimal set of LSs from TSD by utilizing the enlarged hedge algebra formalism combining with genetic algorithm. The experimental results and the comparisons with the existing methods are also presented.

3.1. The interpretable and scalable multi-level semantic core structures for linguistic summarization of time series data

As traditional approaches of designing FSCS for linguistic words in the word domains of LVs, the fuzzy sets are manually specified by human experts. Therefore, the number of fuzzy sets used for constructing fuzzy partitions are fixed and limited to 7 ± 2 and when adding more words to fuzzy partitions, the semantics of the existing words will be changed leading to having uninterpretable models. As stated above, the enlarged hedge algebra provides a mathematical formalism for automatically generating the interpretable and scalable multi-level semantic structure M_{Gr}^X for the word domains of LVs. It means that the generated multi-level semantic structures M_{Gr}^X ensure the generality and specificity of their corresponding word sets, so they can be extended as needed by arbitrarily adding the more specific words while the FSCS of the existing words are still retained [28, 29]. In this paper, this mathematical formalism is applied to automatically generate such multi-level semantic structures for LVs associated with the TSD, the quantifier (Q), and the summarizer (S). Based on the practical application contexts, the level of specificity of words can be specified by users. They just set the value of k to 1, 2, 3, or higher. The higher value of k , the more specific words are used, so the algorithm model can detect more special characteristics of data. Specifically, the value of k is 1, the number of used words is 5, the value of k increases to 2, the number of words is 9, the value of k increases to 3, the number of words is 17, and so on.

The syntactical semantics design of quantifier Q : Two primary words $G = \{\text{"few"}, \text{"many"}\}$, $H = \{\text{little (L), very (V)}\}$, and artificial hedge h_0 .

The syntactical semantics design of summarizer S : Two primary words $G = \{\text{"low"}, \text{"high"}\}$, $H = \{\text{little (L), very (V)}\}$, and artificial hedge h_0 .

The time dimension is hierarchically partitioned by month, season in [17].

3.2. The criteria of evaluation of a set of linguistic summaries

For each time interval $T_i \in \mathfrak{I}$, DS_{T_i} denotes a subset of values in DS (historical TSD) corresponding to T_i . The truth value of each LS l is computed as:

$$Tr_l = \mu_{q_l} \left[\frac{\sum_{d_t \in DS_{T_l}} \mu_{s_l}(d_t)}{|DS_{T_l}|} \right] \quad (1)$$

where μ_{q_l} is the membership function of the quantifier word $q_l \in Q$, $\mu_{s_l}(d_t)$ is the membership degree of the historical data value $d_t \in DS_{T_l}$ to the fuzzy set associated with the summarizer word $s_l \in S$, and $|DS_{T_l}|$ is the number of historical data values in the time interval T_l . The word $q_l \in Q$ is chosen so that the value of Tr_l reaches its maximum value.

The evaluation of a set of LSs bases on the following criteria:

The coverage of a set of LSs: each data value of $d_i \in DS$ is considered in a LS with a given coverage degree $\alpha > 0$. It means that $\mu_{s_j}(d_i) > \alpha$, where $\mu_{s_j}(d_i)$ is membership degree of d_i to the word s_j . The set of values $\{d_i\}$ is called the set of values covered by the set of LSs, denoted by DS' . The coverage of a set of LSs is specified by the following formula:

$$Cv = \frac{|DS'|}{|DS|} \quad (2)$$

where $|DS'|$ and $|DS|$ are the number of covered data values and the total number of data values in the historical TSD, respectively.

The brevity of a set of LSs: The brevity is the number of LSs m in the set. However, when becoming an objective of the genetic algorithm, it conflicts the accuracy objective. Therefore, its value (Br) is computed as the inverse ratio of m as follows:

$$Br = \frac{1}{m} \quad (3)$$

The accuracy of a set of LSs: The accuracy of a set of LSs with m sentences is computed by averaging the truth values Tr_l of all LSs, denoted by Ac , as follows:

$$Ac = \frac{\sum_{l=1}^m Tr_l}{m} \quad (4)$$

3.3. The proposed genetic algorithm for selecting optimal set of linguistic summaries

The fitness function of the genetic algorithm is the combination of three criteria of linguistic summary evaluation: coverage, brevity, and accuracy. The coverage criterion is important, so to cover all historical data values, the coverage criterion is fixed by 100%. Consequently, the fitness function is just the combination of two objectives: brevity and accuracy as follows:

$$F = \lambda \times Br + \beta \times Ac \quad (5)$$

where $\lambda + \beta = 1$.

On the other hand, the accuracy of a LS is very important rather than the brevity, so in the experiments, the values of λ and β are set to 0.1 and 0.9, respectively.

The proposed genetic algorithm model for selecting optimal set of LSs, named as LS-HA-GA, is a genetic algorithm with two individuals in the population. In the genetic algorithm, the on-point crossover operator and the mutation operator are applied as follows:

- The on-point crossover operator: for two parent individuals p_1 and p_2 , take half of the LSs from p_1 and combine them with half of the LSs from p_2 to create two new individual s_1 and s_2 .
- The mutation operator: for each individual, randomly select and apply the following three types of mutation:
 - + Replace an existing LS by a new one: randomly replace a LS with a new generated one that meets the coverage criterion and the minimum truth value threshold η .
 - + Split a LS into some ones: select a LS in which the time dimension T is a season, replace it by some new LSs in which the time dimension T is a month according to that season.
 - + Merge some LSs into one: select some new LSs in which time dimension T is a month in the same season, merge them into a new LS in which the time dimension T is a season.

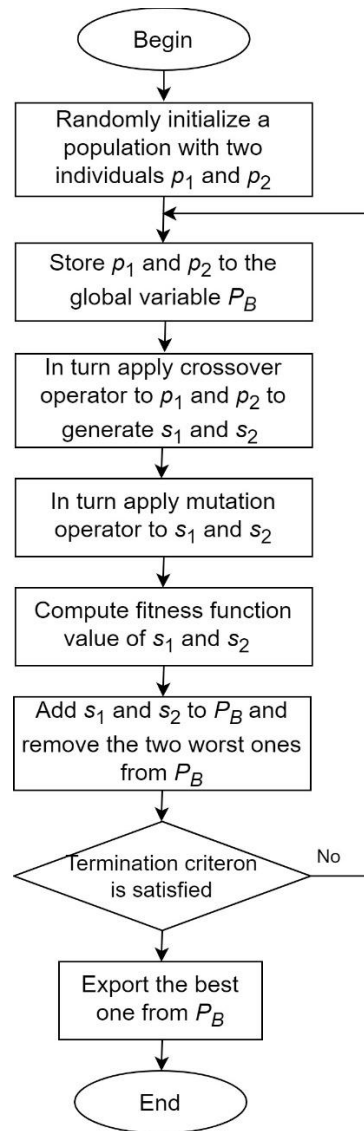


Figure 2. Genetic algorithm model for selecting optimal set of LSs.

The proposed LS-HA-GA model is described as follows and summarily visualized in Figure 2.

Step 1: Randomly initialize a population with two individuals.

Randomly generate two individuals o_1 and o_2 , where each individual represents N candidate LSs in the form of ***Q of T are S*** that satisfy the coverage criterion and the minimum truth value threshold η . Compute the fitness values F_1 and F_2 of o_1 and o_2 , respectively, by Equation (5).

Step 2: Store two initial population to a global variable P_B .

Set P_B to empty.

Add o_1 and o_2 to P_B .

Step 3: For $k = 1$ to $MaxGen$

Step 3.1. Choose two individuals p_1 and p_2 from P_B .

Step 3.2. Apply the one-point crossover operator to p_1 and p_2 to generate two children s_1 and s_2 .

Step 3.3. In turn apply the mutation operator to s_1 and s_2 .

Step 3.4. Compute the fitness values F_1 and F_2 of s_1 and s_2 , respectively, by Equation (5).

Step 3.5. In turn add s_1 and s_2 to P_B and then remove the two worst ones from P_B based on their fitness function values.

End For

Step 4: Return a set of LSs represented by the best individual having the best fitness function value in P_B .

As other genetic-based optimization models, the computational complexity them is not fixed and depends on many parameters. The most time-consuming part of the proposed LS-HA-GA model is the calculation of fitness values (F). Other factors contribute to the computational complexity are the number of generations ($MaxGen$) and the population size. The LS-HA-GA model only has two individuals, and each individual applies the simple one-point crossover and mutation operators. Therefore, the computational complexity mainly depends on $MaxGen$ and F . It can be seen in Equation (5) that F depends on the number of sentences (m), the number of words in Q ($|Q|$) and the number of data point in $T_i \in \mathfrak{S}$. If T_i is large enough, the number of data point in T_i is approximately equal to the number of TSD n . Consequently, the time complexity of the LS-HA-GA model is $O(MaxGen \times m \times |Q| \times n)$.

4. EXPERIMENTAL STUDIES AND DISCUSSION

4.1. Experiment setup

The experiments in this study were implemented by *Python* programming language with the support of libraries such as *numpy* for data processing, *scipy* for calculations, and *matplotlib* for data visualization. The experimental system used a personal computer (laptop) with Intel Core i5 2.60 GHz Dual Core CPU configuration, 8GB RAM, 64-bit MacOS operating system. The experimental data is a simulated time series consisting of 365 historical data values, representing the number of patients admitted daily in a year at a medical facility, extracted from the study in [17].

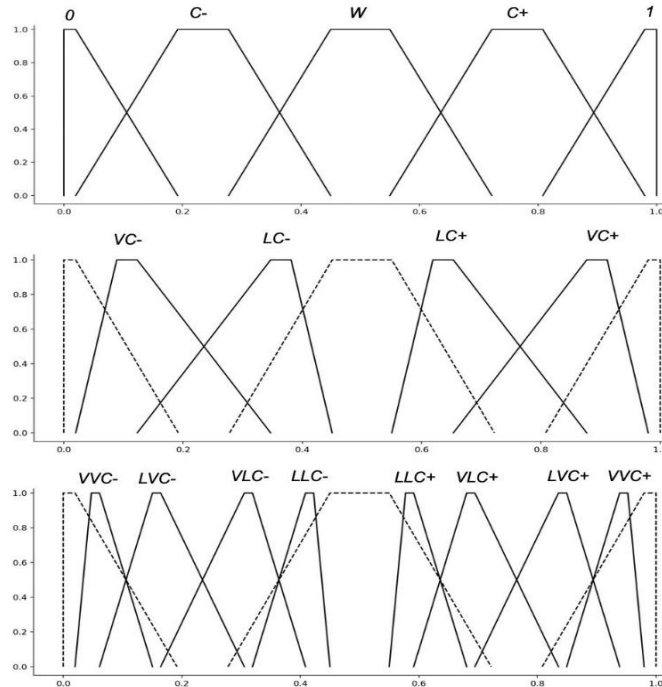


Figure 3. The generated multi-level semantic structure with $k=3$.

After doing some experiments and evaluation, the given values of parameters of the algorithm models are chosen as follows:

- The syntactical semantics design is as in Section 2.3.
- The quantitative semantics of the quantifier Q : $m(\text{"few"}) = 0.43$, $m(\text{"many"}) = 0.43$.
- The quantitative semantics of the summarizer S : $m(\text{"low"}) = 0.43$, $m(\text{"high"}) = 0.43$.
- For both Q and S : the fuzziness parameter values of three word constants: $m(0) = 0.02$, $m(W) = 0.1$, $m(1) = 0.02$. The fuzziness parameter values of the hedges: $\mu(h_0) = 0.02$, $\mu(V) = 0.3$, and $\mu(L) = 0.5$. The level of specificity: $k = 1, 2$, and 3 .
- For the applied genetic algorithm model: the number of generations: $MaxGen = 100$, the number of candidate LSs represented by an initial individual: $N = 15$.

- The coefficients of the brevity and accuracy in (4) are 0.1 and 0.9, respectively.
- The minimum truth threshold: $\eta = 0.6$.

With the given syntactical semantics and quantitative semantics set as above, the multi-level semantic structure with $k = 3$ is generated as shown in Figure 3.

4.2. Results and comparison

First, the proposed algorithm model LS-HA-GA was executed with the specificity level of $k = 1$. Therefore, the number of words generated in the linguistic domain of LVs associated with both quantifier Q and summarizer S is 5. As a result, only the words with the length of 1 are used to describe the UoD of the historical TSD and summarizer. As shown in the highest part of Figure 3, the trapezoidal FSCS of five words institute the fuzzy partition of the UoD of variables, where 0, c^- , W , c^+ , and 1 stands for the words “*absolutely few*” or “*absolutely low*”, the negative primary word “*few*” or “*low*”, “*medium*”, the primary word “*many*” or “*high*”, and “*absolutely many*” or “*absolutely high*”.

Next, the linguistic domains are extended by increasing the specificity level of k to 2 in such a way that two linguistic hedges *Little* (L) and *Very* (V) simultaneously act on two primary words. Consequently, four new words Vc^- , Lc^- , Lc^+ , and Vc^+ with the length of 2 are added while the FSCS of the existing words (0, c^- , W , c^+ , and 1) are still preserved, as shown in the highest and the middle parts of Figure 3.

Finally, the linguistic domains are extended by increasing the specificity level of k to 3 and eight new words are added as shown in Figure 3.

With each specificity level of k , the proposed algorithm model LS-HA-GA was executed 10 times independently, so we received 10 running results of brevity and accuracy values. These values were averaged accordingly. The coverage value is always 365. Then, the averaged values of brevity and accuracy are compared with the corresponding ones of the existing algorithm models, genetic and greedy ones. The comparison results are shown in Table 1.

Table 1. The comparison of the results of the proposed algorithm models with the ones of existing algorithm models in [17]

Algorithm model	Brevity	Accuracy	Coverage
Genetic [17]	14	0.9148	362
Greedy [17]	6	0.7596	363
LS-HA-GA ($k=1$)	8	0.9312	365
LS-HA-GA ($k=2$)	4	0.9700	365
LS-HA-GA ($k=3$)	4	0.9725	365

As shown in Table 1, with the specificity level of 1, the proposed algorithm model LS-HA-GA ($k=1$) gives the accuracy value of 0.9312, much better than the ones of the genetic and greedy algorithm models in [17], respectively are 0.9148 and 0.7596. The coverage value of it is also better. The brevity value of it is 8 which is better than that of the genetic algorithm model and worse than that of the greedy algorithm model. With the specificity level of 2 and 3, the LS-HA-GA has both brevity and accuracy values better than those of the genetic and greedy algorithm models in [17]. On the other hand, the accuracy of the LS-HA-GA with the specificity level of 3 has the accuracy value of 0.9725, better than that of the LS-HA-GA with the specificity level of 2. Therefore, when the linguistic domain is extended, the LS-HA-GA gives more accuracy value. Obviously, the more specific words are added, the more special characteristics of the historical TSD are detected.

Some set of LSs achieved from the experiments with the specificity level of 1, 2, and 3 are shown in Table 2. As shown in Table 2, in case the specificity level is 1, there are 8 extracted LSs in which 6 of them with the time dimension T is a month. In case the specificity level is 2 or 3, there are only 4 extracted LSs, in which 3 of them with the time dimension T is a season.

About the experimental execution time analysis, the total execution time depends on the specificity level of words k because at the higher level of specificity, more words with more specific semantics are added to describe special features of TSD more accurately. Therefore, with the computer configuration for the experiments shown in Subsection 4.1 and the number of generations is 100, the total execution time (in minute) of the proposed LS-HA-GA model according to the specificity levels from $k = 1$ to 3 are 0.12, 1.8 and 2.4. With the more complex problems, the early stopping mechanism should be applied because the algorithm model usually converges in the early generations.

Table 2. The extracted linguistic summaries with the specificity level of 1, 2, and 3

No.	Specificity level	Summary sentences	Truth value
1	$k = 1$	<i>Many</i> days in <i>cool season</i> , the patient inflow is <i>medium</i>	0.91
2		<i>A half</i> of days in <i>September</i> , the patient inflow is <i>medium</i>	1.0
3		<i>A half</i> of days in <i>October</i> , the patient inflow is <i>medium</i>	1.0
4		<i>Most</i> of days in <i>November</i> , the patient inflow is <i>low</i>	0.62
5		<i>A half</i> of days in <i>December</i> , the patient inflow is <i>low</i>	0.99
6		<i>A half</i> of days in <i>hot season</i> , the patient inflow is <i>many</i>	1.0
7		<i>Many</i> days in <i>January</i> , the patient inflow is <i>low</i>	0.96
8		<i>A half</i> of days in <i>February</i> , the patient inflow is <i>low</i>	0.97
1	$k = 2$	<i>Many</i> days in <i>cool season</i> , the patient inflow is <i>medium</i>	0.91
2		<i>A half</i> of days in <i>hot season</i> , the patient inflow is <i>Little many</i>	1.0
3		<i>Little many</i> days in <i>cold season</i> , the patient inflow is <i>low</i>	0.97
4		<i>A half</i> of days in <i>March</i> the patient inflow is <i>Little many</i>	1.0
1	$k = 3$	<i>A half</i> of days in <i>hot season</i> , the patient inflow is <i>Little many</i>	1.0
2		<i>Little many</i> days in <i>cold season</i> , the patient inflow is <i>low</i>	0.97
3		<i>Little very many</i> days in <i>cool season</i> , the patient inflow is <i>medium</i>	0.92
4		<i>A half</i> of days in <i>March</i> the patient inflow is <i>Little many</i>	1.0

To evaluate the proposed algorithm model with the real-world data, we have executed it with the dataset of the number of patients daily visiting a hospital in northern Vietnam as shown in Figure 4.

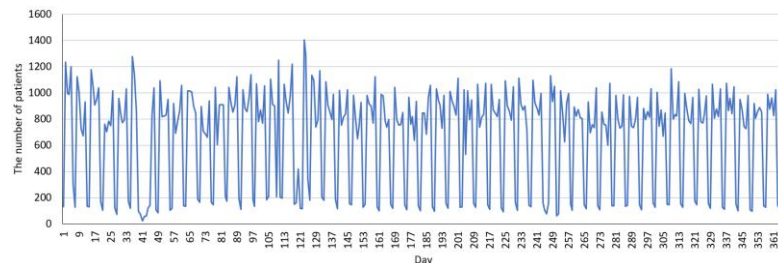


Figure 4. The number of patients daily visiting a hospital in northern Vietnam

The specificity level was set to 1, 2, and 3, and the experimental results are as follows:

- For the specificity level $k = 1$, the accuracy of the extracted set of LSs is 0.612975 and the brevity is 4.
- For the specificity level $k = 2$, the accuracy of the extracted set of LSs set is 1.0 and the brevity is 4.
- For the specificity level $k = 3$, the accuracy of the extracted set of LSs is 1.0 and the brevity is 4.

As we can see that when the specificity level increases from 2 to 3, the quality of the extracted LSs does not change. It is shown that there is not much special characteristics in the TSD.

5. CONCLUSIONS

This paper presents a method of automatically generating interpretable and scalable multi-semantic structures for extracting a compact set of LSs from TSD based on enlarged hedge algebra formalism combining with a genetic algorithm with two individuals. The proposed method allows users to extract interpretable LSs and extend the word domains to extract LSs describing the more specific characteristics of TSD when required. With the help of the interpretable multi-level semantic structures, the FSCS of the existing words are still preserved when adding more words to the linguistic domains. The comparisons of the experimental results

show the superior performance of our proposed method. This not only increases the applicability of enlarged hedge algebras in extracting LSs from TSD in practice but also states the great potential applications of enlarged hedge algebras in solving real-word problems.

REFERENCES

- [1] R. R. Yager, "A new approach to the summarization of data," *Information Sciences*, vol. 28, no. 1, pp. 69-86, 1982.
- [2] J. Kacprzyk, R. R. Yager, and S. Zadrozny, "A fuzzy logic based approach to linguistic summaries of databases," *International Journal of Applied Mathematics and Computer Science*, vol. 10, no. 4, pp. 813-834, 2000.
- [3] J. Kacprzyk and S. Zadrozny, "Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools," *Information Sciences*, vol. 173, no. 4, pp. 281-304, 2005.
- [4] C. Donis-Diaz, R. Bello, and J. Kacprzyk, "Linguistic data summarization using an enhanced genetic algorithm," *Technical Transactions – Automatic Control*, vol. 2013, no. 2, pp. 3-12, 2013.
- [5] C. Donis-Diaz, A. Muro, R. Bello-Pérez, and E. V. Morales, "A hybrid model of genetic algorithm with local search to discover linguistic data summaries from creep data," *Expert Systems with Applications*, vol. 41, no. 4, pp. 2035-2042, 2014.
- [6] T. Altıntop, R. R. Yager, D. Akay, F. E. Boran, and 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*, vol. 25, no. 04, pp. 599-620, 2017.
- [7] M. D. Peláez-Aguilera, M. Espinilla, M. R. F. Olmo, and J. Medina, "Fuzzy linguistic protoforms to summarize heart rate streams of patients with ischemic heart disease," *Complexity*, vol. 2019, pp. 1-11, 2019.
- [8] A. Duraj, P. S. Szczepaniak, and 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, pp. 787-799.
- [9] A. Jain, M. Popescu, J. Keller, M. Rantz, and B. Markway, "Linguistic summarization of in-home sensor data," *Journal of biomedical informatics*, vol. 96, Art. no. 103240, 2019.
- [10] A. Wilbik, I. Vanderfeesten, D. Bergmans, S. Heines, and W. van Mook, "Linguistic summaries for compliance analysis of a glucose management clinical protocol," *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018, pp. 1-7.
- [11] C. A. Calderón, I. P. Pupo, R. Y. Herrera, P. Y. P. Pérez, R. P. Pulgarón, L. A. Acuña, "Sport Customized Training Plan Assisted by Linguistic Data Summarization," *Computational Intelligence Applied to Decision-Making in Uncertain Environments, Studies in Computational Intelligence*, vol. 1195. Springer, Cham, 2025, pp. 283–309.
- [12] K. K. Majer, G. Casalino, G. Castellano, M. Dominiak, O. Hryniewicz, O. K. ska, G. Vessio, N. D. Rodríguez, "PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries," *Information Sciences*, vol. 614, 374–399, 2022.
- [13] G. Serkan, D. Akay, F. E. Boran, R. R. Yager, "Linguistic summarization of fuzzy social and economic networks: an application on the international trade network," *Soft Computing*, vol. 24, 1511–1527, 2020. <https://doi.org/10.1007/s00500-019-03982-9>
- [14] D.-A. Chiang, R. L. Chow, and Y.-F. Wang, "Mining time series data by a fuzzy linguistic summary system," *Fuzzy Sets and Systems*, vol. 112, pp. 419-432, 2000.
- [15] J. Kacprzyk, A. Wilbik, and S. Zadrozny, "On some types of linguistic summaries of time series," *International IEEE Conference Intelligent Systems, September 2006*, 373-378.
- [16] J. Kacprzyk, A. Wilbik, S. Zadrozny, "Linguistic summarization of time series using a fuzzy quantifier driven aggregation," *Fuzzy Sets and Systems*, vol. 159, pp. 1485–1499, 2008.
- [17] C. O. Rita, Marin N., S. Daniel, G.B. T. Andrea, "Linguistic Summarization of Time Series Data using Genetic Algorithms," *Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-11)*, Aix-les-Bains, France, July 2011, 416-423.
- [18] A. Wilbik and J. Kacprzyk, "Towards a multi-criteria analysis of linguistic summaries of time series via the measure of informativeness," *Int. J. Data Analysis Techniques and Strategies*, Vol. 4, No. 2, 2012.
- [19] R. J. Almeida, M.-J. Lesot, B. Bouchon-Meunier, U. Kaymak, and 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*, Jul. 2013, pp.1-8.
- [20] M.-G. Juan, R.-B. Luis, G. Juan, C. Ester, "The generation of qualitative descriptions of multivariate time series using fuzzy logic," *Applied Soft Computing*, vol. 23, pp. 546–555, 2014.
- [21] M. Gilles, M.-J. Lesot, "Linguistic summaries of locally periodic time series," *Fuzzy Sets and Systems*, vol. 285, pp. 94–117, 2016.
- [22] V. Novák, "Mining information from time series in the form of sentences of natural language," *International Journal of Approximate Reasoning*, vol. 78, pp. 192–209, 2016.
- [23] Ö. Ilyas, E. B. Fatih, A. Diyar, "A possibilistic approach for interval type-2 fuzzy linguistic summarization of time series," *Artificial Intelligence Review*, vol. 54, 3991–4018, 2021.
- [24] B. Khedidja, H. Allel, L. Mohand, "Trends summarization of times series: a multi-objective genetic algorithm-based model," *Journal of Smart Environments and Green Computing*, vol. 2, pp. 19-33, 2022.
- [25] C. H. Nguyen, W. Wechler, "Hedge algebras: an algebraic approach to structures of sets of linguistic domains of linguistic truth variables", *Fuzzy Sets and Systems*, vol. 35, no. 3, pp. 281-293, 1990.
- [26] C. H. Nguyen, W. Wechler, "Extended hedge algebras and their application to fuzzy logic," *Fuzzy Sets and Systems*, vol. 52, pp. 259–281, 1992.

- [27] C. H. Nguyen, T. S. Tran, and 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*, vol. 67, pp. 244-262, 2014.
- [28] C. H. Nguyen, T. L. Pham, T. N. Nguyen, C. H. Ho, and T. A. Nguyen, "The linguistic summarization and the interpretability, scalability of fuzzy representations of multilevel semantic structures of word- domains," *Microprocessors and Microsystems*, vol. 81, Art. no. 103641, 2021.
- [29] V. T. Hoang, C. H. Nguyen, D. D. Nguyen, D. P. Pham, V. L. Nguyen, "The interpretability and scalability of linguistic-rule-based systems for solving regression problems," *International Journal of Approximate Reasoning*, vol. 149, pp. 131–160, 2022.
- [30] D. P. Pham, T. L. Pham, X. T. Tran, "A Fuzziness Parameter Optimization Method To Extract The Optimal Set Of Linguistic Summaries From Numeric Data," *TNU Journal of Science and Technology*, vol. 229, no. 07, pp. 49 - 57, 2024.
- [31] N. H. Huy, N. C. Ho, N. V. Quyen, "Multichannel image contrast enhancement based on linguistic rule-based intensifiers," *Applied Soft Computing Journal*, vol. 76, pp. 744–762, 2019.
- [32] B. H. Le, L. T. Anh, B. V. Binh, "Explicit formula of hedge-algebras-based fuzzy controller and applications in structural vibration control," *Applied Soft Computing*, vol. 60, pp. 150–166, 2017.
- [33] H. Q. Cao, N. A. Tran, H. L. Bui, "Hedge-Algebras-based hybrid control of earthquake-induced buildings using upgraded tuned liquid column dampers," *Soil Dynamics and Earthquake Engineering*, vol. 182, 108728, 2024.

BIOGRAPHY OF AUTHORS



Tran Xuan Thanh received a Bachelor degree of Education in Technical Education from Hanoi National University of Education in 2003, a Bachelor degree of Engineering in Information Technology from Hanoi University of Science and Technology in 2006, and a Master degree in Computer Science from the Posts and Telecommunications Institute of Technology in 2012. From 2008 to 2020, he worked at the Faculty of Information Technology, Thanh Do University. Since 2021, he has been a lecturer at the Faculty of Information Technology, East Asia University of Technology, Vietnam. His research interests include optimization algorithms, and the applications of fuzzy logic and hedge algebra in linguistic summarization problems.



Pham Dinh Phong received a Master degree in Information Technology and a Doctor of Philosophy degree in Computer Science from University of Engineering and Technology, Vietnam National University, Hanoi in 2011 and 2018, respectively. Now, he is a lecturer at University of Transport and Communications, Vietnam. His research interests include hedge algebras, fuzzy systems, soft computing, data mining and machine learning.



Pham Thi Lan received a Bachelor degree in Informatics Teaching from Hanoi National University of Education in 2006, Master degree in Computer Science from Hanoi National University of Education in 2008, and Doctor of Philosophy degree in Computer Science at the Graduate University of Science and Technology, Vietnam Academy of Science and Technology in 2022. Since 2008, she has been working as a lecturer in the Faculty of Information Technology, Hanoi National University of Education, Vietnam. Her research interests include data structure and algorithms, applications of hedge algebra, and approaches to extracting linguistic summaries.