Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 3267820, 16 pages
https://doi.org/10.1155/2023/3267820

Research Article

Exploring the Structure of IoT Data: A Symbolic Analysis Perspective

Yinghua Shen,[1] **Witold Pedrycz**,[2] **Wentao Li**,[3] **Zhi Xiao**,[1] **Tianhua Chen**,[4] **Xuan Hu**,[5] **and Yuan Chen**[6]

[1]*School of Economics and Business Administration, Chongqing University, Chongqing 400044, China*
[2]*Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada T6R 2G7*
[3]*College of Artificial Intelligence, Southwest University, Chongqing 400715, China*
[4]*Department of Computer Science, University of Huddersfield, Huddersfield HD1 3DH, UK*
[5]*School of Public Policy and Administration, Chongqing University, Chongqing 400044, China*
[6]*College of Management and Economics, Tianjin University, Tianjin 300072, China*

Correspondence should be addressed to Xuan Hu; shawn@cqu.edu.cn

Received 22 July 2022; Revised 26 November 2022; Accepted 18 January 2023; Published 9 February 2023

Academic Editor: Chuanwen Luo

With the development of different kinds of techniques, especially the Internet of Things (IoT), a large amount of quantitative (either numeric or categorical) data have been generated, transmitted, and stored in the modern society. People hope to understand the interested phenomenon from the collected quantitative data by utilizing different data analysis methods. Exploring the structure of data (e.g., the cluster centers or prototypes) has always been a hot spot in the domain of data mining and knowledge discovery, yet it seems that the modeling and analyzing process still focus on a low-level abstraction of the data because normally, the structure found is only represented by some numeric data points. In this study, we highlight that a low-level abstraction may not be a user-friendly way for people to grasp the knowledge contained in the data. Instead, we explore the structure of the data from a perspective of symbolic analysis. Specifically, two modes of abstraction are proposed. In the *vertical* mode (i.e., values of each feature are abstracted), the numeric prototypes are characterized by the symbolic prototypes such that people could get rid of being stuck in minor details of each feature. In the *horizontal* mode (i.e., values of each prototype are abstracted), the linguistic summarization is used to describe all the features of each symbolic prototype such that people could immediately grasp the essential information conveyed in the symbolic prototype. We conduct comprehensive experimental studies on the publicly available data to illustrate the feasibility and validity of the proposed symbolic analysis process.

1. Introduction

An increasing volume of data is being generated, transmitted, and stored nowadays due to the well-developed techniques such as Internet of Things (IoT) [1–8]. People prefer analyzing the data generated from the real-world phenomenon and building models to describe the phenomenon such that later, with the developed models, they could understand the surrounding environment and be more confident in making decisions. For a long time, quantitative models (constructed directly on a basis of quantitative data) have been favored due to the ubiquitous easy-to-collect quantitative (either numeric or categorical) data and the rigorous mathematical fundamentals behind them, say, linear regression models, polynomials, and neural networks. However, quantitative models are not always feasible or effective when the interested phenomenon is too complex to be completely understood. It happens that people only have some intuitive understanding, e.g., some common sense, of the phenomenon, which hinders building a sophisticated quantitative model. Besides, even though the quantitative models are effective, e.g., the neural networks constructed for pattern recognition, people today are still not quite clear how the model functions (resulting in the well-known black

box). All these drawbacks of the quantitative models bring the attention to the qualitative modeling techniques [9].

The crux of the qualitative modeling is that, instead of directly using the quantitative values, the qualitative concepts (e.g., describing the value of a variable or describing the changing trend of a variable) serve as the building blocks of such models, based on which further processing is conducted. This entire process could be regarded as a symbolic analysis process. Many different kinds of qualitative (or symbolic) models have been proposed so far, which could be roughly categorized into five groups.

(i) Physical process-based models [10, 11]. This kind of methods highlights the concept of process; i.e., many natural phenomena or physical situations could be regarded as a certain qualitative process. To describe this process, many new concepts such as the objects, quantity space representation of the objects, individual views, and histories have been proposed

(ii) Qualitative differential equation- (QDE-) based models. There are two branches of methods in this type of models. One branch revolves around the QSIM model proposed in [12], where the qualitative abstraction of the ordinary differential equation is pursued. The final QDE is represented by a set of qualitative constraints. Many QDEs [13–19] have been proposed to extend the applicable scenarios of QSIM. The other branch focuses on the construction QDE based on the so-called confluence [20]; here, the behavior of physical system is divided into different qualitative states, each of which is then described by a set of confluences

(iii) Logic-based models. Two branches of method are observed in this group. One group focuses on building an advanced expert system [21]; here, the quantity is first described by some qualitative symbols, say intervals; then, the logic is used to describe the relationships therein. Another group of methods [22–24] focus on applying the Inductive Logic Programming (ILP) to construct the qualitative models such that the background knowledge could be incorporated

(iv) Qualitative tree-based models. In this type of models, a structure similar to that of the conventional decision tree is formulated. The nodes of the trees are still some numeric thresholds of the features; however, the leaves are represented by some qualitative constraints. When the constraint describes the monotonicity of the output with all the features, we come with the QUIN model [9, 25]; when it only describes that with one feature (i.e., the partial derivative), we have either the Padé model [26] (corresponding to the numeric data) or the Qube model (corresponding to the categorical data) [27]

(v) Others. Here, we observe three other different qualitative models. In [28], the polynomials are used to approximate the numeric data, then from which the qualitative constraints given in QSIM model [12] are derived. In [29], the semiquantitative extension to the QSIM model is conducted; i.e., the fuzzy sets are used to build the quantity space, rules are used to realize the transition between the states, and fuzzy mathematics is used for qualitative simulation. In [30], the OLAP model based on the qualitative representation derived from the dynamic Bayesian networks (DBN) is constructed, making it possible to build high-level states and actions in the continuous environment for an intelligent agent

By summarizing all these proposed qualitative models, a common point observed is that the qualitative representation of the quantity stands in a key position during the entire modeling process. The qualitative representation is normally represented by a series of so-called landmarks (symbols); for example, we could use three symbols +, 0, and - to represent the sign of the magnitude of a variable or that of the changing trend of a variable (e.g., + represents an increasing trend). Albeit the significance of the qualitative representation, their derivation is kind of coarse, and there is a lack of a mechanism to obtain them automatically from the numeric data. This motivates our first objective to form a symbolic description of numeric data based on prototype-based clustering algorithms [31]. The symbolic prototypes (represented as a series of symbols, say, labels such as 1, 2, and 3) would be formed based on the numeric prototypes.

As another observation, it seems that the modeling process in the aforementioned literature still focuses on a low-level abstraction of the data. By low level, we mean the global abstraction of all the features of a prototype is missing in the current literature. In other words, people could not immediately grasp the most essential information contained in the data due to a lack of a human-centric manner of knowledge representation. This motivates the second objective of this research. We intend to build a higher level of qualitative model to describe each symbolic prototype. Specifically, given a symbolic prototype, we would address the issue of how to qualitatively describe the symbols of all the features based on the concept of linguistic summarization.

The aforementioned two objectives naturally bring the originality of the study as follows: the proposed symbolic description of the raw quantitative data delivers a high-level abstraction of the data in a *vertical* mode. By vertical, we mean that for each feature of the prototypes, the attention is paid on the relative locations of the prototypes rather than the detailed quantitative values. The proposed linguistic summarization of the symbolic prototype delivers a high-level abstraction of the data in a *horizontal* mode. By horizontal, we mean that for each symbolic prototype, we make a linguistic summarization of the symbols of the entire features. With both modes of abstraction, we deliver a more human-centric manner to understand the knowledge in the original raw data. The symbolic description followed by the

linguistic summarization forms the symbolic analysis highlighted in this paper.

The remainder of the paper is organized as follows: in Section 2, we propose the method of symbolic characterization of the numeric data (i.e., vertical mode of abstraction) such that the symbolic prototypes (information granules) are obtained. We investigate the property (i.e., stability) of the obtained symbolic prototypes in Section 3. In Section 4, the linguistic summarization of information granules (i.e., horizontal mode of abstraction) is proposed. We conduct experimental studies on the IoT data to illustrate the proposed two modes of abstraction in Section 5. The paper is concluded in Section 6.

2. Symbolic Characterization of Numeric Data

Description of the data of interest has always been a major target in the field of data mining and knowledge discovery. To find the representatives in the data, many clustering concepts and algorithms [31–33] have been proposed and applied to the data, resulting in the numeric representatives (e.g., the prototypes) of data. To make the obtained structure more descriptive, in recent years, the concept of granular computing [34–38] has been used to generalize these numeric prototypes to granular prototypes (e.g., hyperboxes). For either of these two categories of methods, the essence of these prototypes is values defined in the domain of those variables (features/attributes). However, these results are not always beneficial especially when they are directly provided to a human being for understanding the structure of the data. For example, suppose we have three prototypes in a two-dimensional space, which are represented as $\mathbf{v}_1 = [0.17, 10.32]$, $\mathbf{v}_2 = [0.01, 19.75]$, and $\mathbf{v}_3 = [0.28, 9.03]$. These numbers are not easy for a human being to grasp the structure of the data intuitively. Hence, this motivates us to make a symbolic characterization of the data structure. Specifically, instead of directly using these numbers, we represent the value of a feature of a prototype by a certain symbol (or label). And this label specifies the location (in an ascending order) of this value among the values of this feature for all the prototypes. Still using the above example, for 0.17 in $\mathbf{v}_1$, since we have $0.01 < 0.17 < 0.28$, it ranks the second place among all the values for the first attribute; thus, we replace it by a symbol 2. Similarly, we replace 0.01 by 1 and 0.28 by 3. With this mechanism, we have the new prototypes as $L_1 = [2]$, $L_2 = [1, 3]$, and $L_3 = [1, 3]$. In this case, for each prototype, we could immediately have a sense of what is the position of a value of a certain feature among all the prototypes. These new prototypes provide us with a high-level information which is much easier for a human being to understand.

With the illustrated example, in what follows, we briefly summarize the methods of symbolic characterization of data structure. Since it is not our focus to more reasonably devise the clustering algorithms, in this study, we specifically focus on the fuzzy clustering algorithms, Fuzzy C-Means (FCM) [31, 32, 39], in particular. Obviously, one may still resort to other prototype-based clustering methods (e.g., the K-means) to get the prototypes of the data. Suppose the data we are interested in are represented by $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\mathbf{N}\}$, where the k^{th} data point $\mathbf{x}_k$ is a vector in a n-dimensional space $\mathbf{R}^n$ spanned over n features $x_1, x_2, \cdots, x_n$. By setting the number of clusters c and fuzzification coefficient m, clustering data X with FCM is realized by minimizing the following objective function:

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{m} \|\mathbf{x}_k - \mathbf{v}_i\|^2, \tag{1}$$

with the squared weighted distance expressed as

$$\|\mathbf{x}_k - \mathbf{v}_i\|^2 = \sum_{j=1}^{n} \frac{\left(x_{kj} - v_{ij}\right)^2}{\sigma_j^2}, \tag{2}$$

where σ_j is the standard deviation of the j^{th} feature of the data, fuzzification coefficient m is greater than 1 (its commonly used value is set as 2), and unless otherwise specified, $\|\bullet\|$ stands for the weighted Euclidean distance. Here, this weighted distance is used to normalize features with different dimensions. The data set X is clustered into c clusters coming in the form of the partition matrix $U = [u_{ik}]_{c\times N}$, $i = 1, 2, \cdots, c, k = 1, 2, .., N$, and a collection of prototypes represented by a prototype matrix as $V = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_c)^T = [v_{ij}]_{c\times n}$. The k^{th} data is described by the k^{th} column of the partition matrix U. The FCM clustering algorithm is summarized as Algorithm 1.

Suppose that the structure of the data is finally represented as a series of prototypes represented as $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_c$. Focusing on the jth feature, the feature values of all the prototypes read as $v_{1j}, v_{2j}, \cdots, v_{cj}$. By sorting these values in ascending order, we easily know their positions in this ordered array. We then replace these feature values by their equivalent labels, that is, $v_{1j} \longrightarrow l_{1j}, v_{2j} \longrightarrow l_{2j}, \cdots, v_{cj} \longrightarrow l_{cj}$, where l_{ij} is an integer that ranges between 1 and c. Hence, finally, the numeric prototypes are transformed into the symbolic ones which are represented by $L_1, L_2, \cdots, L_c$. This transformation is shown in Figure 1 when only a two-dimensional data set is encountered. We see that the symbolic characterization delivers a qualitative description of the data, specifically of each feature because each label l_{ij} describes the position of the value of the jth feature of the ith prototype relative to other prototypes. This kind of description reflects the information from other prototypes and could be regarded as a vertical abstraction, in a sense that the abstraction is conducted for numeric values of the same feature but different prototypes. Hence, we term the symbolic characterization of numeric data as the vertical mode of abstraction.

3. Stability of Information Granules

Under the ideal condition, clustering algorithms are expected to be applied directly to the entire data to make use of the information provided by all the data (which reflects the entire phenomenon). However, in reality, quite often, only a part of the entire phenomenon is observed by

Input: X, c, m
Output: U, V
Initialize U
Repeat
$v_{ij} = \sum_{k=1}^{N} u_{ik}^{m} x_{kj} / \sum_{k=1}^{N} u_{ik}^{m}$
$u_{ik} = 1/\sum_{s=1}^{c} \left(\|\mathbf{x}_k - \mathbf{v}_i\| / \|\mathbf{x}_k - \mathbf{v}_s\| \right)^{2/(m-1)}$
Until $\|U_{iter+1} - U_{iter}\| < \varepsilon$, where ε is a small positive number.

Algorithm 1: FCM algorithm.

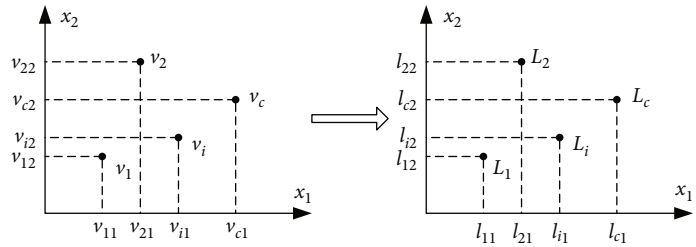


Figure 1: Obtain the symbolic prototypes from the numeric prototypes.

an organization or user, which results in a subset of the entire data. Since the phenomenon is often observed by multiple users, multiple subsets are generated. In this part, we are interested in analyzing the similarity or stability of the symbolic prototypes among the different subsets.

An intuitive illustration of the notion of stability is provided in Figure 2. Suppose that we have four data sets, and three numeric prototypes $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ are found for each data set along with their corresponding symbolic prototypes L_1, L_2, and L_3. The only difference among these four plots is the location of the third prototype $\mathbf{v}_3$ which is moved upward gradually from Figures 2(a) to 2(d). We may find that the structures between Figures 2(a) and 2(b) remain stable, because the relationships (relative positions) among the prototypes are the same. Those relationships are slightly changed in Figure 2(c) and are totally different in Figure 2(d) when three prototypes are nearly aligned. Although only one numeric prototype keeps changing, the labels of all the numeric prototypes may change depending on the stability of the data structure. As observed, since data structures in Figures 2(a) and 2(b) are most similar, their corresponding symbolic prototypes are exactly the same; structures in Figures 2(a) and 2(d) are significantly distinct, and their corresponding symbolic prototypes (especially the labels for feature x_2) are nearly different. In what follows, we study how the stability of the data structures could be measured by the symbolic prototypes.

Suppose that we generate two subsets which are represented as D_1 and D_2 from the entire data, respectively. By applying the symbolic characterization method to D_1 and D_2, we represent the symbolic prototypes as $\mathbf{L} = \{L_1, L_2, \cdots, L_i, \cdots, L_c\}$ and $\mathbf{M} = \{M_1, M_2, \cdots, M_i, \cdots, M_c\}$, where $L_i = (l_{i1}, l_{i2}, \cdots, l_{\text{in}})$and $M_i = (m_{i1}, m_{i2}, \cdots, m_{\text{in}})$. We propose the index to measure the distance between two symbolic prototypes.

$$\|L_i - M_i\| = \frac{1}{n} \sum_{j=1}^{n} f\left(l_{ij}, m_{ij}\right), \text{ where } f\left(l_{ij}, m_{ij}\right) = \begin{cases} 0, & \text{if } l_{ij} = m_{ij}, \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

The distance measure in (3) is a "coarse distance" because we only count how many the same-positioned entries are different. For example, if we have $L_i = \{1, 3, 3, 2\}$ and $M_i = \{1, 3, 2, 2\}$, then only the third entries are different and we have $\|L_i - M_i\| = (0 + 0 + 1 + 0)/4 = 0.25$. Obviously, in the extreme cases, if two symbolic prototypes are identical, the distance equals zero; and if they are different (i.e., the values of the same-positioned entries are always different), the distance is one.

With the distance measure for any pair of the symbolic prototypes, we define the stability index of the information granules as follows when two subsets are considered.

$$Z = 1 - \frac{1}{c} \sum_{i=1}^{c} \min_{s=1,2,\cdots,c} \|L_i - M_s\|. \tag{4}$$

By (4), for any given symbolic prototype L_i, we measure its distance to all the other symbolic prototypes M_s generated from the other data set and pair it with the one having the shortest distance to it. As it could be envisioned, if the distributions of the symbolic prototypes in $\mathbf{L}$ and $\mathbf{M}$ are similar (i.e., symbolic prototypes are stable under the different environment), Z could be a value close to one; otherwise, it is close to zero.

In what follows, we consider the case of multiple subsets. Suppose that we sample r subsets from X, which are represented as $D_1, D_2, \cdots, D_r$. Their symbolic prototypes are denoted by $\mathbf{L}(1) = \{L_1(1), L_2(1), \cdots, L_c(1)\}$, $\mathbf{L}(2) = \{L_1(2),$

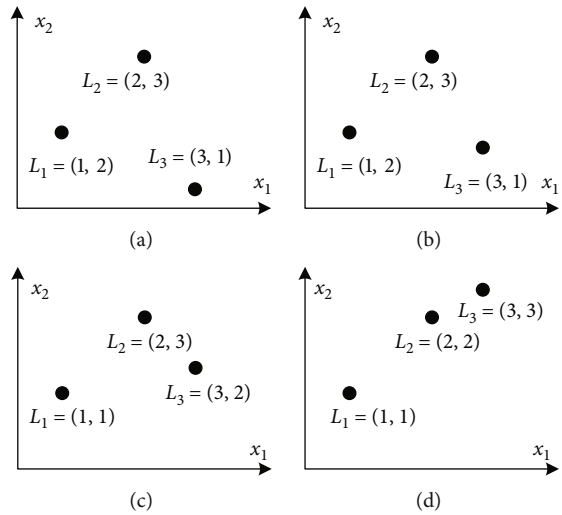


Figure 2: Data structures observed on four data sets.

$L_2(2), \cdots, L_c(2)\}, \cdots$, $\mathbf{L}(r) = \{L_1(r), L_2(r), \cdots, L_c(r)\}$. By using (4), we first measure the stability of information granules focusing on any pair of two subsets, that is,

$$Z(ii, jj) = 1 - \frac{1}{c}\sum_{i=1}^{c} \min_{s=1,2,\cdots,c} \|L_i(ii) - L_s(jj)\|. \quad (5)$$

Obviously, (5) could be regarded as a more detailed (indexed) version of (4). The overall stability of the information granules on all the subsets is defined as

$$Z = \frac{2}{r(r-1)} \sum_{jj=1}^{r-1} \sum_{ii=jj+1}^{r} Z(ii, jj), \quad (6)$$

which could be regarded as the average of the stability values of the related data pairs.

We use the four data structures shown in Figure 2 to illustrate the calculation of the stability of the data structures. Here, we have $c = 3$ and $n = 2$. Besides, to make it clear, we use L_i^{ii} to represent the ith prototype of the iith data set, $i = 1, 2, 3$; and $ii = 1, 2, 3, 4$. We show the process to obtain the stability index between any two data sets in columns in Table 1. For example, when considering the structures of D_1 and D_2 in Figures 2(a) and 2(b), by calculating the distance between L_1^1 and any symbolic prototype in D_1, we match L_1^1 with L_1^2. Similarly, L_2^1 is matched with L_2^2, and L_3^1 is matched with L_3^2. Afterwards, with (5), we have $Z(1, 2) = 1 - 1/3(0 + 0 + 0) = 1$ which demonstrates the high stability between the structures in D_1 and D_2. By the same process, the stability index values between other data pairs are also observed. Note that in Table 1, entry such as $L_1^1 \leftrightarrow L_1^3/L_3^3$ means that L_1^1 can be matched with either L_1^3 or L_3^3 because their distance to L_1^1 is the same. Now, if we consider the stability among the four data structures, with (6), we have $Z = 2/(4 \times 3) \times (1 + 2/3 + 1/2 + 2/3 + 1/2 + 2/3) = 2/3$. Since, the stability values are always located in [0, 1], the obtained values generally reflect our intuition about stability among the data structures.

4. Linguistic Summarization of Information Granules

With the same notation used in the previous section, the symbolic data structure is represented as $\mathbf{L} = \{L_1, L_2, \cdots, L_i, \cdots, L_c\}$, where $L_i = (l_{i1}, l_{i2}, \cdots, l_{in})$, $i = 1, 2, \cdots, c$. In this section, based on the concept of linguistic summarization [40, 41], we propose the horizontal mode of abstraction, in a sense that a higher level of qualitative description is applied to each of the formed symbolic prototypes to make a qualitative description of the values of a group of features.

For example, focusing on the symbolic prototype $L_i = (l_{i1}, l_{i2}, \cdots, l_{in})$, we could describe it by the alike following sentence.

most attributes of information granule L_i assume *high* values. (7)

Note that two linguistic terms with italic bold fonts are used in the above sentence. The term *most* is a linguistic "quantifier" of the number of attributes that satisfy a certain condition (status), while the term *high* is the linguistic "descriptor" of the symbol (label) of a certain feature. Obviously, for the linguistic quantifiers, instead of *most* quantifiers like *a few*, *around half*, etc. could also be used. These quantifiers intrinsically are fuzzy sets defined on the interval $[1, n]$; for example, the term *most* could be illustrated as a fuzzy set in Figure 3(a), while for the descriptors, terms like *low* and *median* could be used. These terms are intrinsically fuzzy sets defined on the interval $[1, c]$; for example, the term *low* could be illustrated as a fuzzy set in Figure 3(b). We may use $\{\tau_1, \tau_2, \cdots, \tau_q\}$ and $\{\mu_1, \mu_2, \cdots, \mu_s\}$, respectively, to represent all the available linguistic quantifiers and the possible linguistic descriptors.

Table 1: Stability of the structures between any data set pairs.

Data set pair	(D_1, D_2)	(D_1, D_3)	(D_1, D_4)	(D_2, D_3)	(D_2, D_4)	(D_3, D_4)
Prototype matching	$L_1^1 \leftrightarrow L_1^2$	$L_1^1 \leftrightarrow L_1^3/L_3^3$	$L_1^1 \leftrightarrow L_1^4/L_2^4$	$L_1^2 \leftrightarrow L_1^3/L_3^3$	$L_1^2 \leftrightarrow L_1^4/L_2^4$	$L_1^3 \leftrightarrow L_1^4$
	$L_2^1 \leftrightarrow L_2^2$	$L_2^1 \leftrightarrow L_2^3$	$L_2^1 \leftrightarrow L_2^4/L_3^4$	$L_2^2 \leftrightarrow L_2^3$	$L_2^2 \leftrightarrow L_2^4/L_3^4$	$L_2^3 \leftrightarrow L_2^4/L_3^4$
	$L_3^1 \leftrightarrow L_3^2$	$L_3^1 \leftrightarrow L_1^3/L_3^3$	$L_3^1 \leftrightarrow L_1^4/L_3^4$	$L_3^2 \leftrightarrow L_1^3/L_3^3$	$L_3^2 \leftrightarrow L_1^4/L_3^4$	$L_3^3 \leftrightarrow L_2^4/L_3^4$
$Z(ii, jj)$	1	2/3	1/2	2/3	1/2	2/3

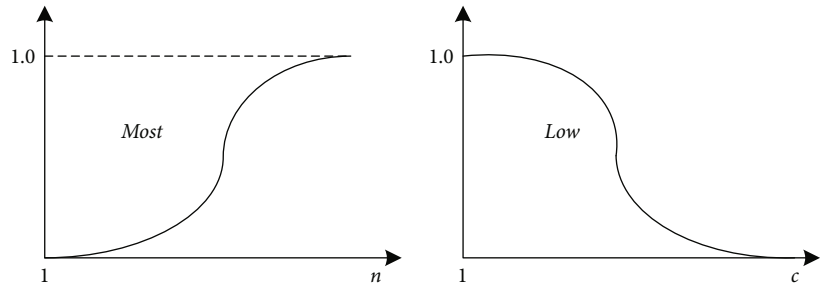


Figure 3: Examples of realization of (a) linguistic quantifier *most* and (b) linguistic descriptor *low*.

For simplicity, the linguistic characterization of an information granule (i.e., the symbolic prototype L_i) could be represented as a formula as

$$\tau(\{\text{symbols}\} = \mu) = \xi, \tag{8}$$

where $\xi \in [0, 1]$ indicates to what degree the summarization is valid.

To get the value of ξ, we follow the initial straightforward idea provided in [40], although many alternative methods could be found in the literature. Suppose we are focusing on the certain linguistic descriptor μ and quantifier τ, the detailed procedure is given as follows:

Step 1. For each symbol $l_{ij} \in L_i$, calculate its membership degree to the linguistic descriptor μ, as $\mu(l_{ij})$.

Step 2. Denote by $r = \sum_{j=1}^{n} \mu(l_{ij})$ the cumulated number of features that satisfy the given descriptor, where r is similar to the σ-count.

Step 3. Determine the membership degree of r with respect to the quantifier τ as $\tau(r)$.

Since q quantifiers and s descriptors are used, we may finally obtain qs pieces of linguistic summarizations, attached with the corresponding validity values. Then, the summarizations with the values of validity exceeding some given threshold t, say 0.8, could be reserved due to the high value of reliability.

Example 1. We use an example to illustrate how the validity degree of a summarization is calculated and what the summarization results look like when multiple descriptors and multiple quantifier are used. Suppose we are provided with a symbolic prototype of $L = (1, 4, 2, 1, 1, 2)$ with the cluster number c equals 5, the used descriptors are {low, intermediate, high}, and the quantifiers are {a few, around half, most}. The corresponding membership functions of these linguistic terms are provided by users, which are shown in Figure 4. Here, for simplicity, only piecewise linear membership functions are considered, and we only show the formulas for membership functions of the three descriptors as (9)–(11); those formulas for the quantifiers are obtained in a similar way (replacing c by n).

Descriptor *low* for a label

$$A(x) = \begin{cases} 1, & x \le 1, \\ \dfrac{(c + 1 - 2x)}{(c - 1)}, & 1 < x \le \dfrac{(c + 1)}{2}, \\ 0, & x > \dfrac{(c + 1)}{2}. \end{cases} \tag{9}$$

Descriptor *intermediate* for a label

$$A(x) = \begin{cases} \dfrac{(2x - 2)}{(c - 1)}, & 1 \le x < \dfrac{(c + 1)}{2}, \\ \dfrac{(2c - 2x)}{(c - 1)}, & \dfrac{(c + 1)}{2} \le x \le c, \\ 0, & \text{otherwise}. \end{cases} \tag{10}$$

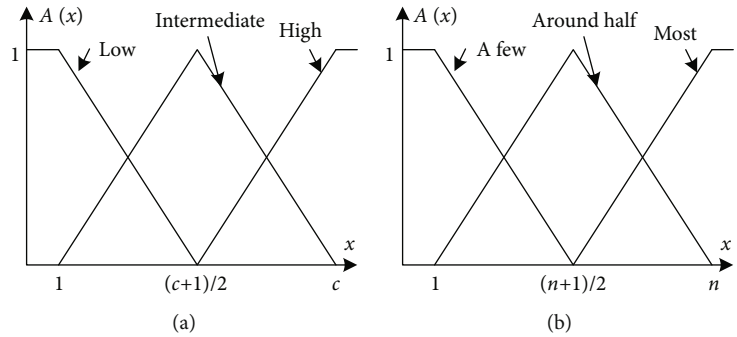


Figure 4: User-defined linguistic terms for (a) descriptors and (b) quantifiers.

Table 2: Results of linguistic summarization of a certain symbolic prototype.

No.	Linguistic summarizations	Validity degree
1	*A few* attributes of information granule assume *high* values	**1.0**
2	*A few* attributes of information granule assume *intermediate* values	**0.8**
3	*Around half* attributes of information granule assume *low* values	**0.8**
4	*Around half* attributes of information granule assume *intermediate* values	0.2
5	*Most* attributes of information granule assume *low* values	0.2
6	*A few* attributes of information granule assume *low* values	0.0
7	*Around half* attributes of information granule assume *high* values	0.0
8	*Most* attributes of information granule assume *intermediate* values	0.0
9	*Most* attributes of information granule assume *high* values	0.0

Table 3: Publicly available IoT data: a summary.

Data #	Data name	# of observations	# of features
1	Banknote	1372	4
2	CCPP	9568	5
3	CBM	11,934	16
4	Quake	2178	4
5	Stock	950	10
6	Stulong	1417	5

Descriptor *high* for a label

$$A(x) = \begin{cases} 0, & x \le \dfrac{(c+1)}{2}, \\ \dfrac{(2x-c-1)}{(c-1)}, & \dfrac{(c+1)}{2} < x \le c, \\ 1, & x > c. \end{cases} \tag{11}$$

Now, let us focus on the summarization of "*most* attributes of information granule assume *low* values" for the symbolic prototype $L = (1, 4, 2, 1, 1, 2)$ when $c = 5$. We follow the steps mentioned above to get the validity degree of this sentence.

Step 1. We calculate the membership degree of each label to the linguistic descriptor *low*. With (9), we have $A_{\text{low}}(x = 1) = 1.0$, $A_{\text{low}}(x = 2) = 0.5$, and $A_{\text{low}}(x = 4) = 0$.

Step 2. We summarize the obtained membership values of all the elements of the symbolic prototype, resulting in $r = 3^* A_{\text{low}}(x = 1) + 2^* A_{\text{low}}(x = 2) + A_{\text{low}}(x = 4) = 4.0$.

Step 3. With the membership function of *most* in Figure 2(b), we easily obtain that $A_{\text{most}}(x = 4.0) = 0.2$.

By exploring all the possible descriptors and quantifiers, we finally obtain 9 pieces of summarizations, along with their validity degrees as in Table 2. All the used linguistic terms for the descriptors and quantifiers are shown in italics. We sort them in descending order according to the validity degree. One may set a threshold beyond which the linguistic summarizations could be reserved or simply claim that the top k (say, three) summarizations could be reserved. Throughout this paper, we adopt the latter strategy; hence, in the example, we reserve the first three linguistic summarizations (with their values of validity degree shown in boldface). These summarizations generally are consistent with our intuition of the symbolic prototype.

5. Experimental Studies

In this section, we check the stability of the information granules and see the performance of the linguistic summarizations on some publicly available IoT data.

5.1. Stability of Information Granules. We use some publicly available IoT data sets to demonstrate the stability of information granules. These data sets could be found in either

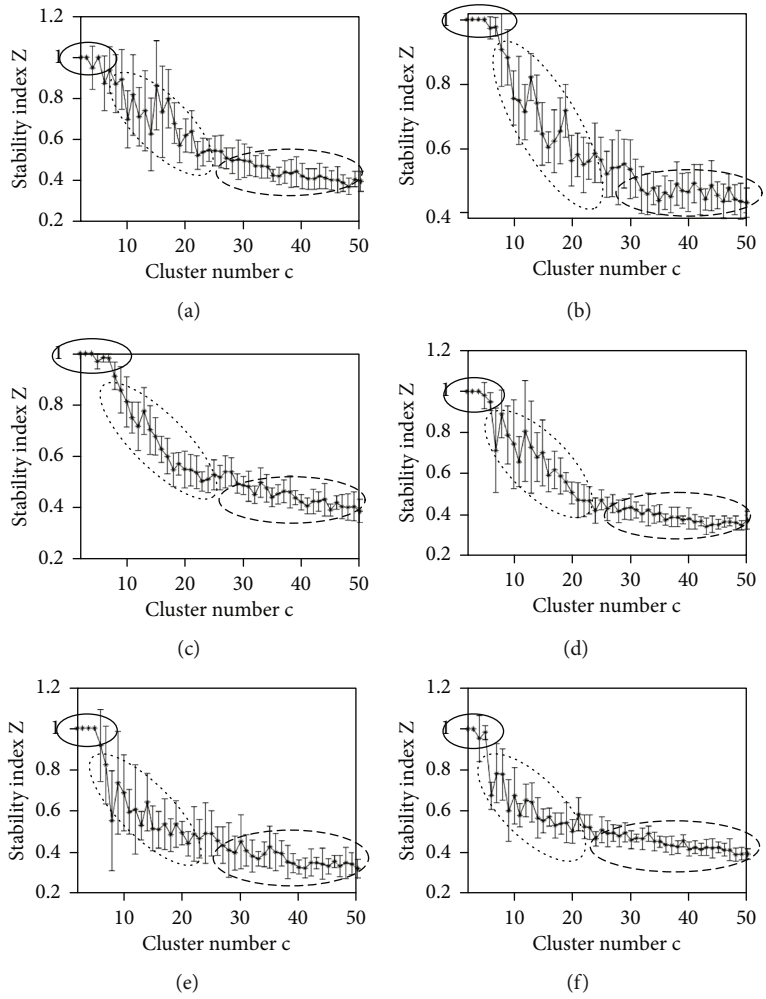


Figure 5: Trends of the stability index with the increasing number of clusters when $P = 2$: (a) banknote; (b) CCPP; (c) CBM; (d) quake; (e) stock; (f) Stulong.

the UCI machine learning repository (https://archive.ics.uci.edu/ml/index.php) or the KEEL data repository (https://sci2s.ugr.es/keel/datasets.php). We list the basic information of these data sets in Table 3. From each given data set, a sampling rate p is used to get P subsets, and we are interested in how the structures among these subsets change (the stability of these structures) with the increasing number of clusters c. The flow of our experiments is shown as follows: the FCM algorithm is performed on each of the P subsets to get the numeric prototypes; these prototypes will be transformed into the symbolic prototypes according to the method introduced in Section 2, accompanied with the process to get the value of the stability index proposed in Section 3. The sequential processes of FCM clustering, symbolic prototype obtaining, and stability index calculating will be repeated k times (i.e., the k-fold experiment) so that the mean and standard deviation of the stability index are obtained. During all the experiments, the fuzzification coefficient m is set as 2, k is set as 10, and c ranges from 2 to 50 with a step size of one.

5.1.1. Identical Subsets. We start the experiment with the simplest case; i.e., we set $p = 1$ and $P = 2$. In other words, for a given data set, the two sampled subsets are identical with the given data set. The trends of the stability of the structures of the two subsets with the increasing number of clusters are shown in Figure 5. Obviously, for each data set, three phases of the trends could be identified. The structures between the two subsets remain quite stable when the number of clusters c is small (e.g., around 5), because the value of the stability index is equal to or quite close to one. Those high values of stability are highlighted with the solid ellipses. Then, the stability of the structures drops significantly with the increasing cluster number, whose trends are highlighted by the dotted ellipses. Afterwards, these trends will generally remain in a low level, which are illustrated by the dashed ellipses. Now, if we consider more subsets, i.e., we set $p = 1$ and $P = 10$, the trends of the stability of the structures are shown in Figure 6. Similar observations are observed to the case where $P = 2$.

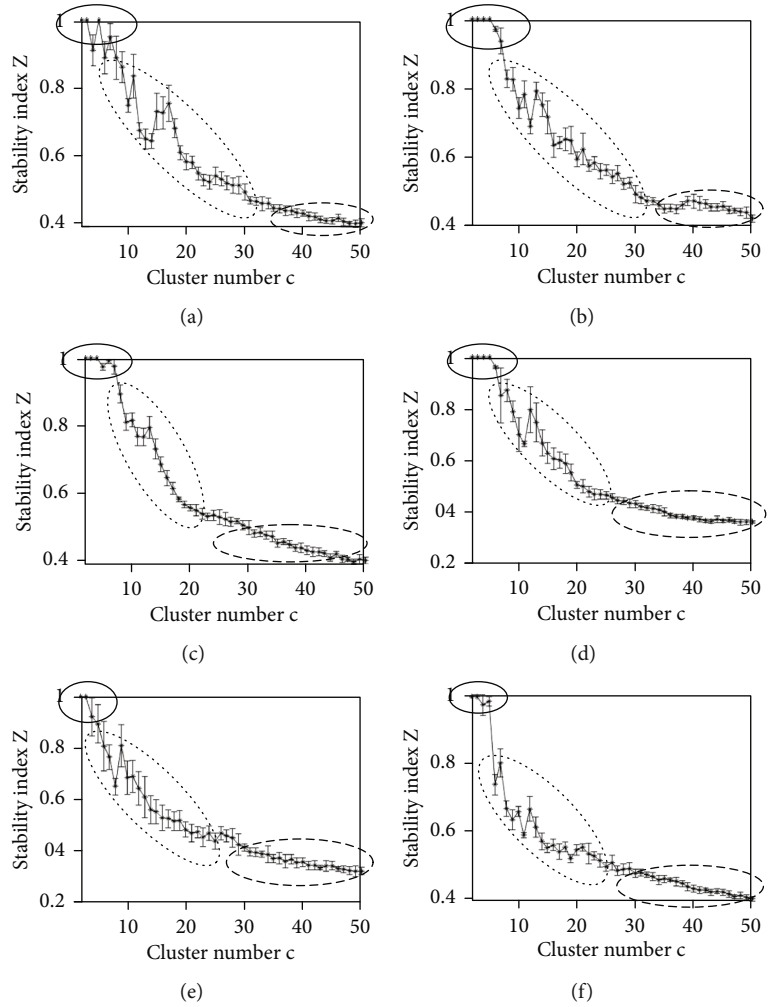


FIGURE 6: Trends of the stability index with the increasing number of clusters when $P = 10$: (a) banknote; (b) CCPP; (c) CBM; (d) quake; (e) stock; (f) Stulong.

It seems reasonable for us to obtain such a type of changing trends (with three phases) of the stability; we give the reason as follows:

(a) The phase of high stability. It seems that when the number of clusters is low, it is easier to obtain the similar cluster prototypes. If two data subsets are similar (in this experiment, they are the same) to each other, then there is a large possibility that the obtained high-level abstraction is similar (i.e., a high value of stability close to one)

(b) The phase of sharp decreasing of stability. When the level of detail increases, it happens frequently that the obtained structures start to be distinct with each other (even if the two subsets are the same). We illustrate this by a simple example. Suppose that a two-dimensional synthetic data set is generated as three Gaussian clusters with the cluster centers as $\mathbf{v}_1 = (6, 4)$, $\mathbf{v}_2 = (6, 10)$, and $\mathbf{v}_3 = (10, 10)$, and the spread (the covariance matrix) of each cluster is represented as

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{12}$$

The synthetic data are shown as the black dots in any plot in Figure 7. Now, if we cluster those data into three clusters, we obtain the data structure (three prototypes denoted by the circles) in Figure 7(a). If we repeat our clustering process two more times, the derived prototypes are shown in Figures 7(b) and 7(c), respectively. Clearly, the structures among the different data sets are nearly the same. Now, if we cluster the data into nine clusters (i.e., to explore more details of the data), the obtained prototypes (derived from three repeated clustering process on the data set) are shown from Figures 7(d) to 7(f). Note that the locations of these prototypes among

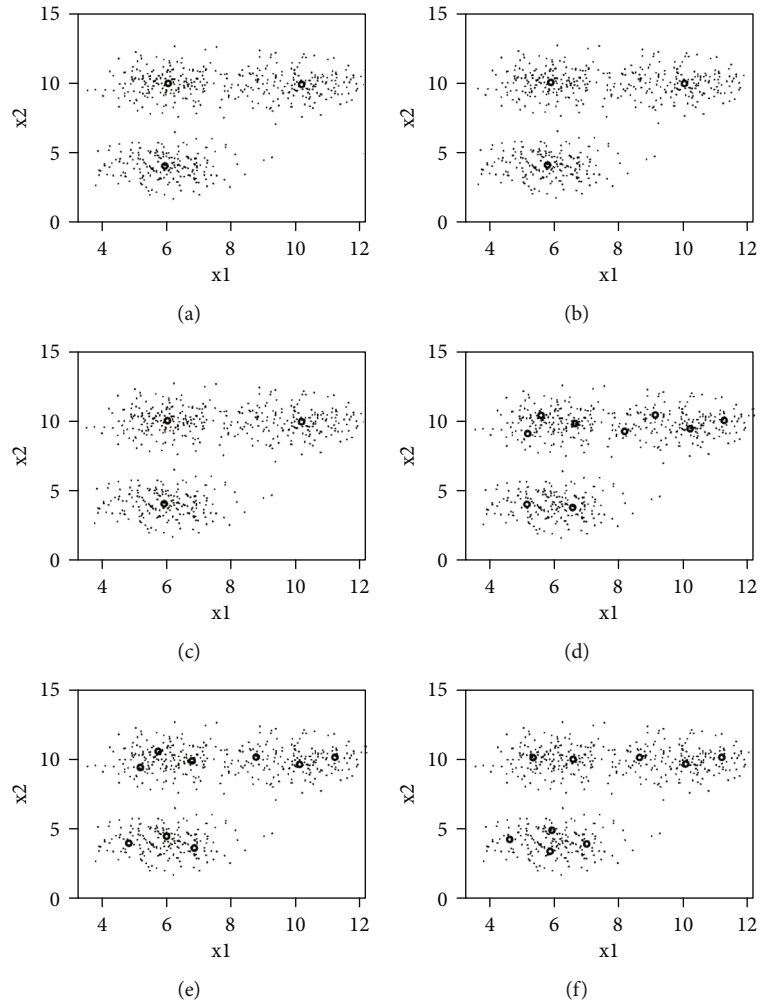


Figure 7: Data structures of a synthetic data set when the FCM algorithm is implemented three times: structures derived from the (a) first time running, (b) second time running, and (c) third time running, when $c = 3$; and structures of (d) first time running, (e) second time running, and (f) third time running, when $c = 9$.

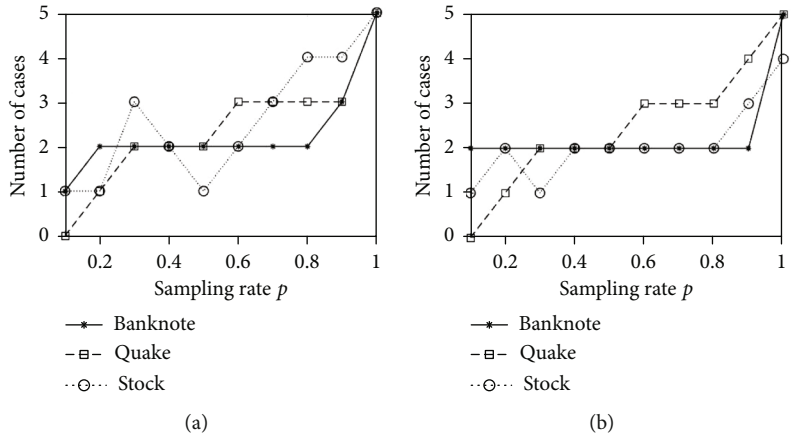


Figure 8: Sensitivity analysis on the sampling percentage for three data sets when (a) $P = 2$ and (b) $P = 10$.

Table 4: Linguistic summarizations of the symbolic prototypes for Stock when $c = 5$.

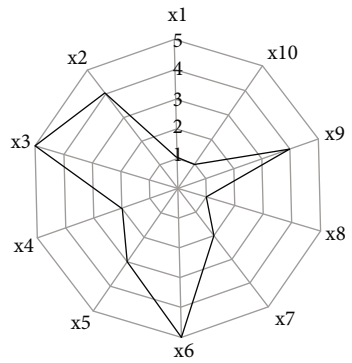
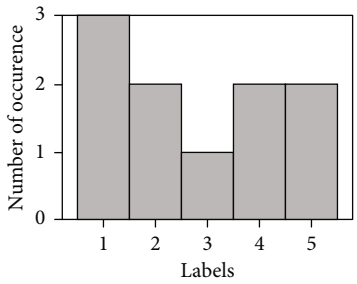
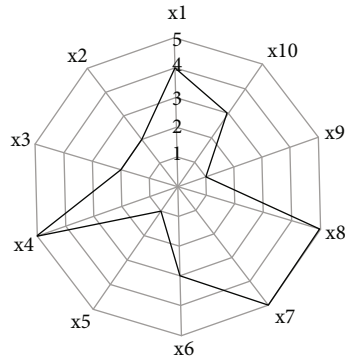
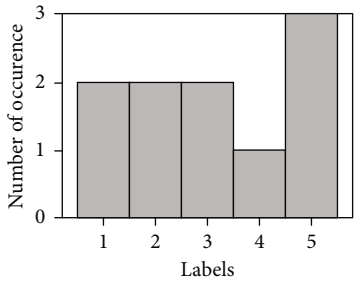
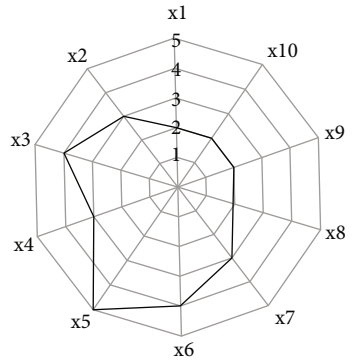
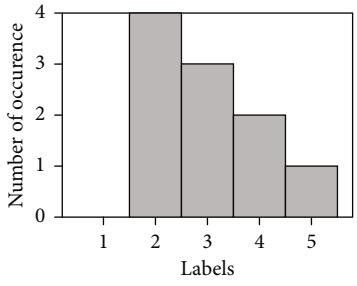
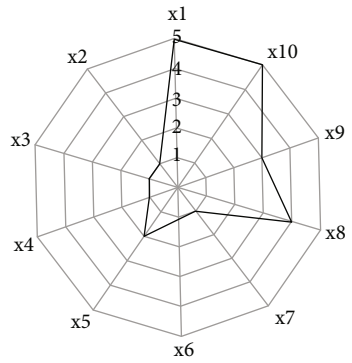
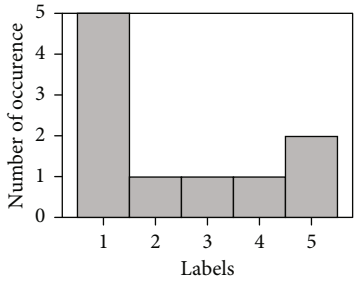
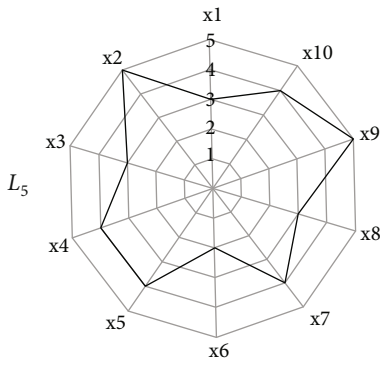
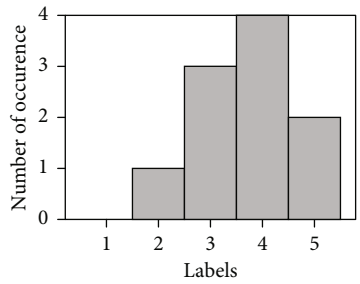
L_i	Radar plot	Histogram plot	Linguistic summarization
L_1			(1) *Around half* attributes of information granule assume *low* values (0.67) (2) *A few* attributes of information granule assume *intermediate* values (0.56) (3) *A few* attributes of information granule assume *high* values (0.56)
L_2			(1) *A few* attributes of information granule assume *low* values (0.56) (2) *Around half* attributes of information granule assume *intermediate* values (0.56) (3) *Around half* attributes of information granule assume *high* values (0.56)
L_3			(1) *Around half* attributes of information granule assume *intermediate* values (0.89) (2) *A few* attributes of information granule assume *low* values (0.78) (3) *A few* attributes of information granule assume *high* values (0.78)
L_4			(1) *Around half* attributes of information granule assume *low* values (1) (2) *A few* attributes of information granule assume *intermediate* values (0.78) (3) A *few* attributes of information granule assume *high* values (0.67)

Table 4: Continued.

L_i	Radar plot	Histogram plot	Linguistic summarization
L_5			(1) *A few* attributes of information granule assume *low* values (1) (2) *Around half* attributes of information granule assume *intermediate* values (1) (3) *Around half* attributes of information granule assume *high* values (0.67)

different plots are significantly different. Specifically, the number of prototypes in the bottom left cluster ranges from 2 to 4 in the repeated experiments. This example illustrates why when more details of the data are explored, the less stability among the data structures (even of the same data set) could be encountered

(c) The phase of converging to a stable level. This phenomenon is also explainable; after all, we only considered a rough index in (3) to measure the difference between two symbolic prototypes. For instance, when the cluster number $c = 4$, we may have two symbolic prototypes as $L = (2, 3, 2, 4)$ and $M = (1, 4, 3, 3)$; their distance equals one according to (3). However, suppose that now we have $c = 5$ and the two symbolic prototypes are $L = (2, 3, 2, 5)$ and $M = (1, 5, 3, 3)$, we obtain the same distance as the case where $c = 4$. Here, we see that although c has been increased, the distance between the pair of symbolic prototypes may remain the same, which may further lead to the unchanged value of the stability

5.1.2. Different Subsets. In the former section, we only considered the case where all the subsets are identical. It is interesting to check the results when different subsets are encountered. By setting the sampling rate to other values rather than one, we could obtain the different subsets. Then, the stability of the information granules among these subsets is calculated in a similar way to the previous section. Here, the point we are interested in is that how many times we could observe the high value of the stability (the threshold for this high value is set as 0.9) even though the subsets are different. Taking Figure 6(f) as an example, we observe four points whose index values are greater than 0.9; then, we conclude that the number of cases with a high value of stability is 4. By ranging the sampling rate p from 0.1 to 1, in Figure 8, we show the obtained results on three UCI data sets when the number of subsets P is set as 2 and 10, respectively. The general observation is that with the increasing sampling rate (hence, the subsets become more and more similar with each other), we obtain more cases where a high stability of information granules is achieved.

We show some possible insights obtained from the experiments in this section as follows. In reality, the same phenomenon is observed by many different agents. It is difficult for each of them to obtain all the related data. However, even with that, if they adopt the symbolic description of the structure of their own data and as long as the number of the prototypes is kept in a low level, there is a large possibility that the symbolic prototypes across the different agents are similar to each other. In other words, without utilizing the outside data, one agent could still nicely grasp the essential structure of the entire phenomenon with the method of symbolic description.

5.2. Linguistic Summarization of Information Granules. In this part, we conduct experiments on two publicly available data stock and CBM (see their basic information in Table 3) to comprehensively show the linguistic summarizations of the symbolic prototypes of the two data sets. For each data set, the FCM algorithm is used to cluster the data into $c = 5$ clusters and these numeric prototypes are transformed into the symbolic prototypes as $L_1, L_2, \cdots, L_5$. These symbolic prototypes for each data set are shown as the radar plots in Tables 4 and 5. The histogram for the labels in each symbolic prototype is also provided in these tables, which later would be used as auxiliary tools to validate the formed linguistic summarizations. The linguistic descriptors and quantifiers defined in Figure 4 are still used here to form the linguistic summarizations; thus, the same as the illustrated example in Section 4, there should be 9 pieces of summarizations for each symbolic prototype. We sort these summarizations in descending order in terms of their values of the validity degree and report those with the first three largest values in Tables 4 and 5. If we compare the obtained three linguistic summarizations with their corresponding symbolic prototype and histogram, most of the time, we find they make sense. For example, if we check the second linguistic summarization for L_1 derived on stock, i.e., "*a few* attributes of information granule assume *intermediate*

Table 5: Linguistic summarizations of the symbolic prototypes for CBM when $c = 5$.

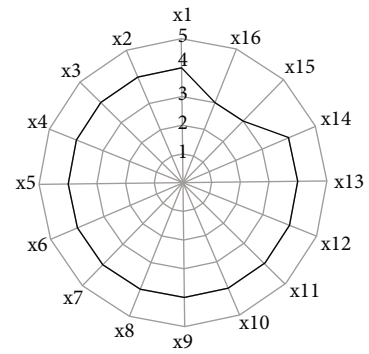
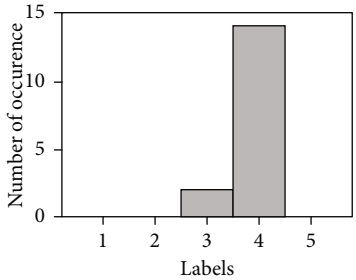
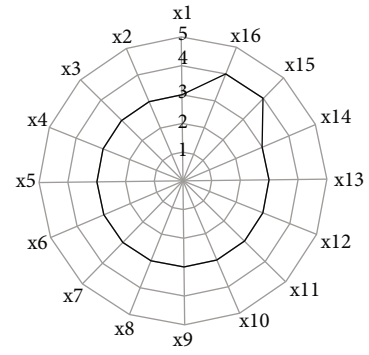
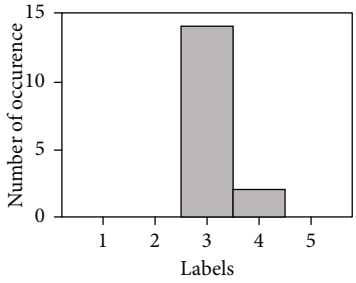
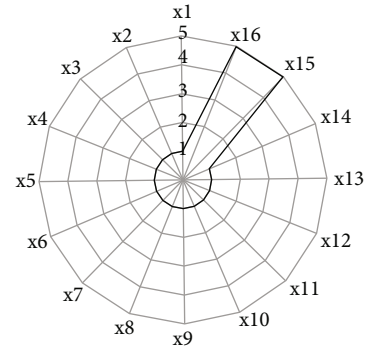
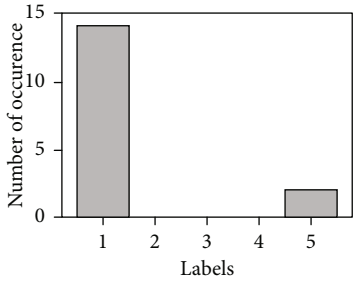
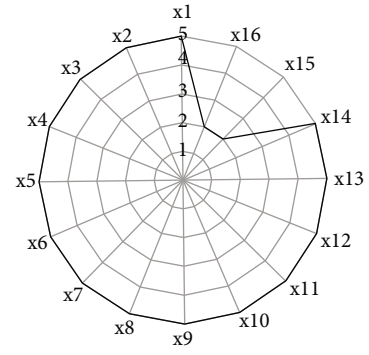
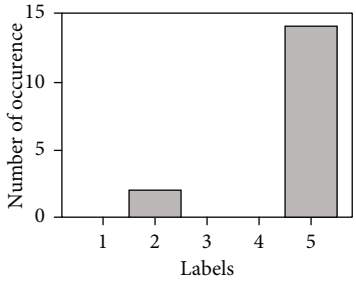
L_i	Radar plot	Histogram plot	Linguistic summarization
L_1			(1) *A few* attributes of information granule assume *low* values (1) (2) *Around half* attributes of information granule assume *intermediate* values (0.93) (3) *Around half* attributes of information granule assume *high* values (0.80)
L_2			(1) *A few* attributes of information granule assume *low* values (1) (2) *A few* attributes of information granule assume *high* values (1) (3) *Most* attributes of information granule assume *intermediate* values (0.87)
L_3			(1) *A few* attributes of information granule assume *intermediate* values (1) (2) *A few* attributes of information granule assume *high* values (0.87) (3) *Most* attributes of information granule assume *low* values (0.73)
L_4			(1) *A few* attributes of information granule assume *low* values (1) (2) *A few* attributes of information granule assume *intermediate* values (1) (3) *Most* attributes of information granule assume *high* values (0.67)

Table 5: Continued.

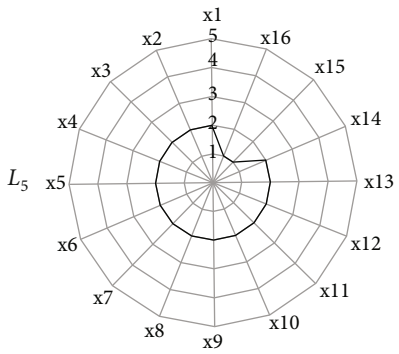
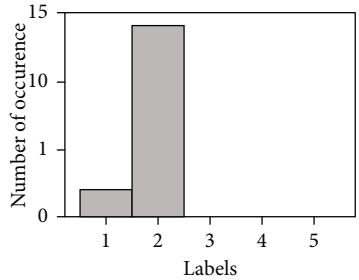
L_i	Radar plot	Histogram plot	Linguistic summarization
L_5			(1) *A few* attributes of information granule assume *high* values (1) (2) *Around half* attributes of information granule assume *low* values (0.93) (3) *Around half* attributes of information granule assume *intermediate* values (0.80)

Table 6: Validity degree of each linguistic summarization of the symbolic prototypes.

Quantifier	Descriptor	For stock					For CBM				
		L_1	L_2	L_3	L_4	L_5	L_1	L_2	L_3	L_4	L_5
A few	*Low*	0.33	**0.56 (1)**	**0.78 (2)**	0	**1 (1)**	**1 (1)**	**1 (1)**	0	**1 (1)**	0
	Intermediate	**0.56 (2)**	0.44	0	**0.78 (2)**	0	0	0	**1 (1)**	**1 (1)**	0.2
	High	**0.56 (2)**	0.44	**0.78 (2)**	**0.67 (3)**	0.33	0.20	**1 (1)**	**0.87 (2)**	0	**1 (1)**
Around half	*Low*	**0.67 (1)**	0.44	0.22	**1 (1)**	0	0	0	0.27	0	**0.93 (2)**
	Intermediate	0.44	**0.56 (1)**	**0.89 (1)**	0.22	**1 (1)**	**0.93 (2)**	0.13	0	0	**0.80 (3)**
	High	0.44	**0.56 (1)**	0.22	0.33	**0.67 (2)**	**0.80 (3)**	0	0.13	0.27	0
Most	*Low*	0	0	0	0	0	0	0	**0.73 (3)**	0	0.07
	Intermediate	0	0	0.11	0	0	0.07	**0.87 (2)**	0	0	0
	High	0	0	0	0	0	0	0	0	**0.73 (2)**	0

values (0.56)," we find that it is consistent with the corresponding histogram because there is only one label "3," which could be regarded as *a few* to some extent. The more detailed information of the validity degree of each linguistic summarization of a specific symbolic prototype is documented in Table 6. Here, for each data set, one may read the table column wise, and in each column, we show the values of validity degree of all the possible linguistic summarizations for each symbolic prototype. Then, we highlight the largest three values of validity degree in boldface along with their orders among all the nine values of validity.

6. Conclusions

To have a better understanding of the knowledge contained in the ubiquitous quantitative (either numeric or categorical) data generated with the novel techniques such as IoT, in this study, we proposed a symbolic analysis method to represent the data structure in a more human-centric manner. Specifically, two different abstraction modes have been proposed. With the vertical mode, the numeric prototypes could be represented by the symbolic ones, and for each feature, prototypes are arranged in ascending trend and represented by labels such as 1, 2,..., c. In this way, people could pay more attention to the relative locations of the prototypes rather than the detailed quantitative values. With the horizontal mode, each symbolic prototype will be described by a linguistic summarization considering all the labels of the features; a sentence could be used to reveal the information contained in the symbolic prototype. With both modes of abstraction, we deliver a more human-centric manner to understand the knowledge in the original raw data. Furthermore, although we focus on the symbolic data description in this study, high-level causality analysis with symbolic rules could be an interesting research direction. Obviously, a detailed study of the "symbolic" fuzzy rule-based model [42–45] deserves a future investigation.

Data Availability

The data sets could be found in either the UCI machine learning repository or the KEEL data repository.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 72001032, Grant 72002152, Grant 72071021, and Grant 71904020 and in part by China Postdoctoral Science Foundation under Grant 2020M673148.

References

[1] M. Zheng, S. Chen, W. Liang, and M. Song, "NSAC: a novel clustering protocol in cognitive radio sensor networks for Internet of Things," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5864-5865, 2019.

[2] K.-K. R. Choo, S. Gritzalis, and J. H. Park, "Cryptographic solutions for industrial Internet-of-Things: research challenges and opportunities," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3567–3569, 2018.

[3] C. Luo, J. Yu, D. Li, H. Chen, Y. Hong, and L. Ni, "A novel distributed algorithm for constructing virtual backbones in wireless sensor networks," *Computer Networks*, vol. 146, pp. 104–114, 2018.

[4] C. Luo, M. N. Satpute, D. Li, Y. Wang, W. Chen, and W. Wu, "Fine-grained trajectory optimization of multiple UAVs for efficient data gathering from WSNs," *IEEE/ACM Transactions on Networking*, vol. 29, no. 1, pp. 162–175, 2020.

[5] M. Babar, M. D. Alshehri, M. U. Tariq et al., "IoT-enabled big data analytics architecture for multimedia data communications," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 5283309, 2021.

[6] Z. Hu, G. Xue, Y. Chen, M. Li, M. Wang, and F. Lv, "City-wide NB-IoT network monitoring and diagnosing," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 3153274, 2022.

[7] J. Li, M. Dai, and Z. Su, "Energy-aware task offloading in the Internet of Things," *IEEE Wireless Communications*, vol. 27, no. 5, pp. 112–117, 2020.

[8] Y. Su, J. Li, Y. Li, and Z. Su, "Edge-enabled: a scalable and decentralized data aggregation scheme for IoT," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1854–1862, 2023.

[9] I. Bratko and D. Suc, "Learning qualitative models," *AI Magazine*, vol. 24, no. 4, p. 107, 2003.

[10] K. D. Forbus, "Qualitative process theory," *Artificial Intelligence*, vol. 24, no. 1–3, pp. 85–168, 1984.

[11] T. Hinrichs and K. Forbus, "Learning qualitative models by demonstration," *Twenty-Sixth AAAI Conference on Artificial Intelligence*, pp. 207–213, 2012, http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/download/4959/5140.

[12] B. Kuipers, "Qualitative simulation," *Artificial Intelligence*, vol. 29, no. 3, pp. 289–338, 1986.

[13] H. Kay, B. Rinner, and B. Kuipers, "Semi-quantitative system identification," *Artificial Intelligence*, vol. 119, no. 1–2, pp. 103–140, 2000.

[14] B. Kuipers, "Qualitative reasoning: modeling and simulation with incomplete knowledge," *Automatica*, vol. 25, no. 4, pp. 571–585, 1989.

[15] D. Berleant and B. J. Kuipers, "Qualitative and quantitative simulation: bridging the gap," *Artificial Intelligence*, vol. 95, no. 2, pp. 215–255, 1997.

[16] W. Pang and G. M. Coghill, "QML-Morven: a novel framework for learning qualitative differential equation models using both symbolic and evolutionary approaches," *Journal of Computer Science*, vol. 5, no. 5, pp. 795–808, 2014.

[17] W. Pang and G. M. Coghill, "QML-AiNet: an immune network approach to learning qualitative differential equation models," *Applied Soft Computing*, vol. 27, pp. 148–157, 2015.

[18] W. Pang and G. M. Coghill, "An immune-inspired approach to qualitative system identification of biological pathways," *Natural Computing*, vol. 10, no. 1, pp. 189–207, 2011.

[19] A. Varšek, "Qualitative model evolution," in *Twelfth International Joint Conference on Artificial Intelligence*, pp. 1311–1316, Sydney, Australia, 1991.

[20] J. De Kleer and J. S. Brown, "A qualitative physics based on confluences," *Artificial Intelligence*, vol. 24, no. 1–3, pp. 7–83, 1984.

[21] I. Bratko, I. Mozetič, and N. Lavrač, *KARDIO: A Study in Deep and Qualitative Knowledge for Expert Systems*, MIT Press, 1990.

[22] S. Dzeroski and L. Todorovski, "Discovering dynamics: from inductive logic programming to machine discovery," *Journal of Intelligent Information System*, vol. 4, no. 1, pp. 89–108, 1995.

[23] G. M. Coghill, A. Srinivasan, and R. D. King, "Qualitative system identification from imperfect data," *Journal of Artificial Intelligence Research*, vol. 32, pp. 825–877, 2008.

[24] A. Srinivasan and R. D. King, "Incremental identification of qualitative models of biological systems using inductive logic programming," *Journal of Machine Learning Research*, vol. 9, pp. 1475–1533, 2008.

[25] D. Šuc and I. Bratko, "Induction of qualitative trees," in *European Conference of Machine Learning in 2001 (LNCS 2167)*, pp. 442–453, Freiburg, Germany, 2001.

[26] J. Žabkar, M. Možina, I. Bratko, and J. Demšar, *Learning qualitative models from numerical data*, 2013.

[27] J. Žabkar, I. Bratko, and J. Demšar, "Extracting qualitative relations from categorical data," *Artificial Intelligence*, vol. 239, pp. 54–69, 2016.

[28] R. K. Gerçeker and A. Say, *Using polynomial approximations to discover qualitative models*, 2006.

[29] Q. Shen and R. Leitch, "Fuzzy qualitative simulation," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 4, pp. 1038–1061, 1993.

[30] J. Mugan and B. Kuipers, "Autonomous learning of high-level states and actions in continuous environments," *IEEE Transactions on Autonomous Mental Development*, vol. 4, no. 1, pp. 70–86, 2012.

[31] Y. Shen, W. Pedrycz, and X. Wang, "Clustering homogeneous granular data: formation and evaluation," *IEEE transactions on cybernetics*, vol. 49, no. 4, pp. 1391–1402, 2019.

[32] X. Jing, Z. Yan, Y. Shen, W. Pedrycz, and J. Yang, "A group-based distance learning method for semisupervised fuzzy clustering," *IEEE transactions on cybernetics*, vol. 52, no. 5, pp. 3083–3096, 2022.

[33] Y. Shen and W. Pedrycz, "Collaborative fuzzy clustering algorithm: some refinements," *International Journal of Approximate Reasoning*, vol. 86, pp. 41–61, 2017.

[34] X. Hu, Y. Shen, W. Pedrycz, Y. Li, and G. Wu, "Granular fuzzy rule-based modeling with incomplete data representation," *IEEE transactions on cybernetics*, vol. 52, no. 7, pp. 6420–6433, 2022.

[35] Y. Shen, W. Pedrycz, and X. Wang, "Approximation of fuzzy sets by interval type-2 trapezoidal fuzzy sets," *IEEE Transactions on Cybernetics*, vol. 50, no. 11, pp. 4722–4734, 2020.

[36] W. Li, H. Zhou, W. Xu, X. Z. Wang, and W. Pedrycz, "Interval dominance-based feature selection for interval-valued ordered data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, 2022.

[37] W. Li, W. Xu, X. Zhang, and J. Zhang, "Updating approximations with dynamic objects based on local multigranulation rough sets in ordered information systems," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 1821–1855, 2022.

[38] W. Li, W. Pedrycz, X. Xue, W. Xu, and B. Fan, "Distance-based double-quantitative rough fuzzy sets with logic operations," *International Journal of Approximate Reasoning*, vol. 101, pp. 206–233, 2018.

[39] W. Li, S. Zhai, W. Xu et al., "Feature selection approach based on improved fuzzy C-means with principle of refined justifiable granularity," *IEEE Transactions on Fuzzy Systems*, pp. 1–15, 2022.

[40] R. R. Yager, "A new approach to the summarization of data," *Information Sciences*, vol. 28, no. 1, pp. 69–86, 1982.

[41] J. Kacprzyk and S. Zadrożny, "Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools," *Information Sciences*, vol. 173, no. 4, pp. 281–304, 2005.

[42] X. Hu, Y. Shen, W. Pedrycz, X. Wang, A. Gacek, and B. Liu, "Identification of fuzzy rule-based models with collaborative fuzzy clustering," *IEEE transactions on cybernetics*, vol. 52, no. 7, pp. 6406–6419, 2022.

[43] Y. Shen, W. Pedrycz, X. Jing, A. Gacek, X. Wang, and B. Liu, "Identification of fuzzy rule-based models with output space knowledge guidance," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 11, pp. 3504–3518, 2021.

[44] T. Chen, C. Shang, J. Yang, F. Li, and Q. Shen, "A new approach for transformation-based fuzzy rule interpolation," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 12, pp. 3330–3344, 2020.

[45] X. Hu, X. Liu, W. Pedrycz et al., "Multi-view fuzzy classification with subspace clustering and information granules," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–14, 2022.