On the Dissimilarity of Fuzzy Information Granules

KACZMAREK-MAJER Katarzyna[1,2] and DAŇKOVÁ Martina[2]

[1] *Systems Research Institue Polish Academy of Sciences, Newelska 6, 01-147 Warsaw, Poland*
E-mail: kaczmar@ibspan.waw.pl

[2] *University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, 30. dubna 22, 701 03 Ostrava 1, Czech Republic*
E-mail: martina.dankova@osu.cz

1 Introduction

Granular Computing (GC) has evolved dynamically over recent decades, substantially enhancing methods to improve our understanding of large numerical datasets, as demonstrated, for example, by Pedrycz and Bargiela [1]. Recent developments have advanced areas such as fuzzy association rule mining and linguistic summarization, among others. Despite significant theoretical and applied achievements and numerous successful practical applications, one of the main challenges that remains is adequate validation of fuzzy information granules, and the task becomes even more complex if various granular computing approaches are confronted.

Let us consider the following two examples of fuzzy information granules:

I_1: *Almost every recording in depression has low energy.*

I_2: *If we observe low energy, then the patient is most likely suffering from depression.*

In practice, there is often a need to compare or confront different types of information granules (often imprecise) that are coming from various sources. To the best of the authors' knowledge, while there are significant contributions to particular areas (e.g., association rules mining), including approaches to assess their quality, there is not much research on the comparative analysis of different granules, which should thoroughly consider various quality criteria, types of quantifiers, and representations of linguistic expressions.

In this work, we pose the question of how to assess the dissimilarity of pairs of information granules that may be exemplified with I_1 and I_2. We focus on two representative types of information granules, namely fuzzy association rules (FAR) and fuzzy linguistic summaries, and aim to (1) propose a unified notation for the construction and selection of the most meaningful fuzzy information granules, and (2) analyze and discuss the assessment of dissimilarity across the considered types. We consider an illustrative example of real-life data collected within the mental health monitoring application scenario, which served as inspiration for this research. This work will conclude with a discussion about open challenges.

Acknowledgement This work is supported from the project "Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583" which is co-financed by the European Union. Katarzyna Kaczmarek-Majer acknowledge support from the project "ExplainMe: Explainable Artificial Intelligence for Monitoring Acoustic Features extracted from Speech" (FENG.02.02-IP.05-0302/23) carried out within the First Team programme of the Foundation for Polish Science co-financed by the European Union under the European Funds for Smart Economy 2021-2027 (FENG).

https://www.doi.org/10.15452/978-80-7599-515-5.2026.11

2 Fuzzy Linguistic Summaries

Fuzzy linguistic summarization (FLS) provides natural-language descriptions of large numerical datasets and has proven to be human-consistent [2]. Originating from Yager's protoforms ('$\mathcal{Q}$ y's are $\mathcal{P}$' and '$\mathcal{Q}R$ y's are $\mathcal{P}$'), numerous approaches have since been developed. While studies (see e.g., [2, 3]) confirm their interpretability and value as information granules, clear guidelines for their practical evaluation are still lacking. This contribution builds upon previous achievements in the theories of generalized and intermediate quantifiers introduced by Mostowski [5]. We adopt the general definition of fuzzy linguistic summaries (FLS) [4] with a linguistic quantifier, qualifiers, and summarizers combined by logical connectives such as AND or OR. FLS can be exemplified with the following sentence: *Almost every sample with low energy and low variability has low speech quality.* Attributes are described by linguistic expressions represented through fuzzy membership functions that refer to relative quantities, such as *low, medium, or high.* FLS is defined as a 5-tuple (quintuple) $S = (A, L, (\mathcal{P}, \diamond), (\mathcal{R}, \star), Q)$ consisting of a set of attributes, linguistic expressions, summarizers, qualifiers, and a quantifier. For further details, we refer interested readers to [4].

The informativeness of each summary is assessed using qualitative criteria. In this work, we consider the degree of truth, expressing how accurately the summary reflects the data [6], and the degree of support, indicating how many data objects it covers. These measures are used to select the most meaningful summaries within each meta-category, as depicted in Algorithm 1.

Algorithm 1 Top-K Fuzzy Linguistic Summary (per Meta-Category)

Require: Quantifiers Q with expert ranking; feature levels; K meta-categories; measures T_1, T_2

```
for i = 1 to K do
    S ← all summaries using features in meta-category i
    Compute T_1(s) and T_2(s) for all s ∈ S
    S* ← arg max_{s∈S} T_1(s)
    Break ties in S* by, in order:
   (i) maximize # of summarizers,
   (ii) maximize # of quantifiers,
   (iii) prefer higher-ranked quantifiers in Q,
   (iv) maximize T_2
    Output the remaining s*
end for
```

3 Fuzzy association rules

Fuzzy association rules (FAR) [7] are derived from fuzzified data using quality measures such as fuzzy support, fuzzy confidence, and Łukasiewicz-based implicational quantifiers [9]. We employ the implementation from the `nuggets` R package, which extracts pairs of fuzzy antecedents and consequents, evaluates them using selected quantifiers, and filters rules with sufficiently high support and confidence. Multiple antecedents are aggregated using the minimum t-norm, chosen for its stability in high-dimensional settings where other continuous t-norms (e.g., product or Łukasiewicz) rapidly reduce activation values. To reduce redundancy, a subset-based filtration removes overly specific rules within the same consequent group, retaining only the most informative ones.

The resulting rules serve as the basis for an Implicational Model with Quantifiers (IMQ) [8], a fuzzy inference framework where each rule of the form IF A THEN B is evaluated using the Łukasiewicz implication and weighted by its corresponding quantifier. The inference process is described in Algorithm 2.

Algorithm 2 Minimal IMQ Inference from Fuzzy Association Rules

Require: Fuzzified data $\mathcal{D}$; measures (support, confidence, q); thresholds $\tau_{\text{supp}}, \tau_{\text{conf}}$
Ensure: Predicted class label(s) with ties preserved

$\mathcal{R} \leftarrow$ mine rules from $\mathcal{D}$; keep if support $\geq \tau_{\text{supp}}$ and confidence $\geq \tau_{\text{conf}}$
For each consequent B, remove rules whose antecedents are proper subsets of another rule's antecedents
for each instance $\bar{x}$ **do**
  for each class C **do**
    $\text{Score}(C) := 1$
  end for
  for each rule $(A \Rightarrow B) \in \mathcal{R}$ **do**
    $u \leftarrow \text{Imp}_L(A(\bar{x}), [B = C]); \quad w \leftarrow q(A, B)$
    $\text{Score}(C) \leftarrow \min\big(\text{Score}(C), \text{Imp}_L(w, u)\big)$ for the corresponding C
  end for
  Output $\arg\max_y \text{Score}(C)$ (report all ties)
end for

4 Assessment of dissimilarity between information granules

Having defined the two types of granules, we now consider the challenging problem of assessing the dissimilarity score between the outcomes of FLS and FAR.

Let us denote FLS as $\mathcal{S} = \{(S_i, t_{i,1}, t_{i,2})\}_{i \in I}$, and analogously, let FAR, together with the values of the quantifier, be denoted by $\mathcal{R} = \{(R_j, q_j)\}_{j \in J}$.The list of qualitative criteria for both granules needs to be established at the beginning. For FLSs, degrees of truth $t_{\cdot,1}$ and support $t_{\cdot,2}$ are considered to assess their validity. For FARs, the degree of the quantifier q is taken into account, as it represents the degree of truth of the corresponding rule. Moreover, the degree of support s, as well as any other relevant characteristics of a rule, can be computed for each rule. Consequently, $\mathcal{R}$ can be extended to a trichotomious representation $\mathcal{R}_t = \{(R_j, q_j, s_j)\}_{j \in J}$

These requirements may serve as the foundation for defining an aggregation operator that combines individual attribute-based dissimilarities into a unified score, ensuring consistency with the trichotomous representation and monotonicity (the monotonicity of the scoring indicates that the greater the difference, the higher the value of dissimilarity).

Let us start with a small illustrative example, and let us consider a simple dissimilarity d_1 defined by

$$d_1(\mathcal{S}, \mathcal{R}) = \sum_{l \in L} |\{1 \mid (l \in \mathcal{R}) \not\equiv (l \in \mathcal{S})\}|,$$

where $L = L_{FLS} \cup L_{FAR}$, L_{FLS}, and L_{FAR} are sets of all linguistic expressions included by FLS and FAR methods, respectively. Note that in this particular example, we do not include any numerical characteristics (i.e., $\{t_{\cdot,1}, t_{\cdot,2}, q_{\cdot}, s_{\cdot}\}$); therefore, it is not necessary to specify which representation we are dealing with. Applying this formula to an illustrative example from Table 1, we receive $d_1(\mathcal{S}, \mathcal{R}) = 1$ because the only difference in linguistic expressions involved is “short sleep”.

Table 1: Illustrative examples of FLSs and FAR.

No.	**Fuzzy information granules**
S_1	Almost every recording in depression has low voice energy.
S_2	Almost every recording in depression from a day with short sleep also has low voice energy variability.
R_1	If we observe low voice energy and low voice energy variability, then the patient is most likely suffering from depression.

In this contribution, we will investigate extensions of this definition to involve a weighting scheme based on the quality criteria, advanced handling to the missing information about the attributes, and we will attempt the semantic proximity between different linguistic expressions included in the considered two information granules (summaries and rules).

In future work, we will further extend the scoring to incorporate linguistic descriptions provided by domain experts into the comparative analyses, examining the limitations of their potential integration within the present framework, and developing advanced formulas for defining the dissimilarity score.

References

[1] A. Bargiela and W. Pedrycz. *From Numbers to Information Granules* The Springer International Series in Engineering and Computer Science, Springer, Boston, **717** (2003) 125-160.

[2] J. Kacprzyk, R. R. Yager, and J. M. Merigo. *Towards human-centric aggregation via ordered weighted aggregation operators and linguistic data summaries: A new perspective on Zadeh's inspirations.* IEEE Computational Intelligence Magazine, **14(1)** (2019) 16-30.

[3] K. Kaczmarek-Majer et al. *Plenary: Explaining black-box models in natural language through fuzzy linguistic summaries.* Information Sciences, **614** (2022) 374-399.

[4] K. Miś, K. Kaczmarek-Majer, M. Baczyński: *Fuzzy Linguistic Summaries and the Double Negation Property.* TechRxiv. July 24 (2025).

[5] A. Mostowski: *On a generalization of quantifiers.* Fund. Math. **44** (1957) 12–36.

[6] L. A. Zadeh: *A computational approach to fuzzy quantifiers in natural languages.* Computers & Mathematics with Applications, **9(1)** (1983) 149-184.

[7] G. Chen and E. E. Kerre. *Fuzzy Logic in Discovering Association Rules: An Overview* (2006) Springer US, Boston, 459-493. `https://doi.org/10.1007/0-387-34296-6_14`

[8] M. Daňková, *Weighted Fuzzy Rules Based on Implicational Quantifiers*, in: V.-N. Huynh et al. (eds.), *Integrated Uncertainty in Knowledge Modelling and Decision Making*, Springer, 2023, pp. 27–36. `https://doi.org/10.1007/978-3-031-46775-2_3`

[9] J. Ivánek, *Construction of implicational quantifiers from fuzzy implications*, Fuzzy Sets and Systems 151(2), 381–391 (2005). `https://doi.org/10.1016/j.fss.2004.07.002`