Symbolic Knowledge Comparison: Metrics and Methodologies for Multi-Agent Systems[⋆]

Federico Sabbatini[1,*], Christel Sirocchi[1] and Roberta Calegari[2]

[1] *Department of Pure and Applied Sciences, University of Urbino Carlo Bo*

[2] *Department of Computer Science and Engineering (DISI), Alma Mater Studiorum–University of Bologna*

Abstract

In multi-agent systems, understanding the similarities and differences in agents' knowledge is essential for effective decision-making, coordination, and knowledge sharing. Current similarity metrics like cosine similarity, Jaccard similarity, and BERTScore are often too generic for comparing knowledge bases, overlooking critical aspects such as overlapping and fragmented boundaries, and varying domain densities. This paper introduces new specific similarity metrics for comparing knowledge bases, represented via symbolic knowledge. Our method compares local explanations of individual instances, preserving computational resources and providing a comprehensive evaluation of knowledge similarity. This approach addresses the limitations of existing metrics, enhancing the functionality and efficiency of multi-agent systems.

Keywords

Multi-agent systems, Knowledge similarity, Symbolic knowledge

1. Introduction

In multi-agent systems, the knowledge owned by each agent is the key to enabling autonomous decision-making [1]. Agents rely on their individual and collective knowledge to steer and react to complex environments, making understanding similarities and differences in their knowledge a fundamental aspect of their interaction. This understanding is pivotal for different applications within the system, including collaborative decision-making, knowledge sharing, and coordination [2].

To effectively compare the knowledge of different agents, similarity measurement metrics and techniques are needed, allowing agents to assess the extent to which their knowledge overlaps, identify areas of agreement and disagreement, and optimise their interactions accordingly [3]. Some examples include collaborative decision-making where agents can use similarity metrics to evaluate the compatibility of their internal knowledge or beliefs. By comparing their knowledge representations, agents can pinpoint areas of alignment and conflict, facilitating negotiations and enabling collaborative decisions that capitalise on the strengths and complementary aspects of each agent's knowledge. As for knowledge sharing, similarity metrics help determine which

WOA 2024: 25th Workshop "From Objects to Agents", July 8–10, 2024, Forte di Bard (AO), Italy

[⋆] Original research paper.

[*] Corresponding author.

f.sabbatini1@campus.uniurb.it (F. Sabbatini); c.sirocchi2@campus.uniurb.it (C. Sirocchi); roberta.calegari@unibo.it (R. Calegari)

0000-0002-0532-6777 (F. Sabbatini); 0000-0002-5011-3068 (C. Sirocchi); 0000-0003-3794-2942 (R. Calegari)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

pieces of knowledge are most relevant to share amongst agents. High similarity scores indicate significant overlap, suggesting that certain knowledge may be redundant if shared. In resource allocation and task assignment, especially in multi-agent systems with diverse expertise and capabilities, similarity metrics can match agents to tasks based on the similarity between their knowledge and the task requirements. Similarity metrics also guide adaptive learning and knowledge transfer processes by helping agents identify peers with similar knowledge profiles. Also, similarity metrics can help identify common ground when agents encounter conflicting beliefs or preferences. By focusing on areas of high similarity, agents can find mutually acceptable solutions and build consensus, facilitating smoother negotiation and collaboration.

State-of-the-art metrics to measure the similarity between two objects, e.g., vectors, exist, such as cosine similarity [4], Jaccard similarity [5], and semantic similarity using contextual embeddings like BERTScore [6]. However, these metrics are often generic distances, not specific for comparing knowledge bases, and thus they do not take into account some fundamental aspects, such as overlapping boundaries or fragmented, or regions with varying domain densities in their measurements. These are aspects that should be considered to effectively measure similarity between knowledge bases. Accordingly, in this paper, we define new similarity metrics designed to overcome these limitations.

Building on the premise that symbolic knowledge can be expressed through logical rules, e.g., approximated as hypercubes [7, 8], we propose similarity metrics for comparing two symbolic knowledge bases in this form. Instead of comparing each rule with all possible others, which would waste computational resources, our approach compares local explanations provided for the same instances by two specific knowledge pieces[1]. Since our approach considers explanations at a local level and aggregates similarity measurements across instances, it offers a comprehensive evaluation of knowledge similarity, addressing several limitations of existing metrics.

2. Explanation-Based Similarity Metrics

In this section, we formally present four formulations of metrics designed to express the similarity between pairs of knowledge pieces, highlighting their differences. These metrics are based on the following assumption: the similarity between two distinct knowledge pieces can be assessed by measuring the pairwise similarity of their local explanations when queried with the same instance. The core idea is to gather these similarity measurements across a sufficiently large set of instances and then aggregate them, for example, by averaging.

Leveraging local explanations simplifies the similarity assessment for several reasons. For instance, when comparing knowledge bases composed of predictive rules, it is necessary to develop a strategy for comparing these rules. Comparing all possible pairs of rules is computationally infeasible and often insignificant, as a rule may be similar to one but very different from all others in the knowledge base. This issue can be mitigated by considering only pairs of overlapping/adjacent/close rules, though defining formal notions for these concepts is complex. Exploiting local explanations bypasses these challenges and provides reliable proxies for knowledge similarity, enabling the creation of computationally feasible metrics.

[1] We point out here that we use the terms “instance”, “sample” and “individual” as synonyms representing a single entry of the data set at hand

2.1. Notation

Let us briefly introduce the notation used in the following sections. Let $\mathcal{S}$ represent a data set consisting of n pairs $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i$ is an instance described by k input attributes $(x_i^1, x_i^2, \dots, x_i^k)$, and y_i is the corresponding outcome. Formally:

$$\mathcal{S} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_n, y_n)\},$$

$$\boldsymbol{x}_i = (x_i^1, x_i^2, \dots, x_i^k).$$

Let D_x and D_y be the domains of the instances' inputs and outputs, respectively:

$$(\boldsymbol{x}_i \in D_x) \ \wedge \ (y_i \in D_y), \quad \forall i = 1, 2, \dots, n.$$

Let us assume the existence of a predictive function f defined as follows:

$$f : D_x \to D_y,$$

$$f(\boldsymbol{x}_i) = \hat{y}_i,$$

where $\hat{y}_i$ is the value predicted by f for the instance $\boldsymbol{x}_i$. Any entity with predictive capabilities, such as a machine learning model or symbolically represented knowledge, can be modelled with f. In this paper, we consider without loss of generality the comparison between two symbolic knowledge bases, e.g., knowledge encoded by domain experts or obtained with symbolic knowledge-extraction (SKE) techniques applied to black-box machine learning predictors [9]. Nonetheless, our approach can be employed to compare any pair of objects providing local explanations.

Finally, let us assume the existence of an explaining function e^f, which maps data set instances $\boldsymbol{x}_i$ to their corresponding *local explanations* ξ_i^f, derived by analysing f:

$$e^f : D_x \to D_x,$$

$$e^f(\boldsymbol{x}_i) = \xi_i^f, \quad \boldsymbol{x}_i \in \xi_i^f \subseteq D_x, \quad \forall i = 1, 2, \dots, n.$$

Specifically, ξ_i^f defines a subregion within the domain D_x that encloses $\boldsymbol{x}_i$ and provides a local explanation for the prediction $f(\boldsymbol{x}_i)$.

The similarity metrics introduced in the following subsections are represented by the operators $\overset{\mathcal{S}}{\approx}$, $\overset{\mathcal{S}}{\sim}$, $\overset{\mathcal{S}}{\approx}_{\subset}$, and $\overset{\mathcal{S}}{\sim}_{\subset}$. These metrics are functions evaluated on the instances of a data set $\mathcal{S}$, mapping pairs of predictive functions f_1 and f_2 to the $[0, 1]$ interval:

$$\overset{\mathcal{S}}{\approx}, \overset{\mathcal{S}}{\sim}, \overset{\mathcal{S}}{\approx}_{\subset}, \overset{\mathcal{S}}{\sim}_{\subset} : (D_x \to D_y) \times (D_x \to D_y) \to [0, 1].$$

The symbol $\approx$ signifies approximate equality between operands, while $\sim$ denotes similarity that is not necessarily bound to equality. Additionally, the subscript $\subset$ indicates asymmetric relations, while its absence indicates symmetric ones. These symbols are used as infix binary operators (e.g., $f_1 \overset{\mathcal{S}}{\sim} f_2$).

The specific task at hand determines the nuances of similarity assessment, which must satisfy specific sets of properties. For example, when comparing a knowledge base encoded by human experts with symbolic knowledge extracted via SKE, one may desire the latter to precisely align with the expert knowledge, focusing on the exact coincidence between corresponding decision boundaries. Conversely, a different scenario may involve replacing a knowledge base with a newer, more accurate one without altering the decision boundaries that led to correct predictions in the past (backward compatibility; [10]). In this case, the similarity assessment should only consider the subset of data set instances that received the expected outcomes from the current knowledge piece. Accordingly, we define two subsets of $\mathcal{S}$ suitable for similarity assessment based on user needs: $\mathcal{S}^{f_1=f_2}$ and $\mathcal{S}^f$.

Definition 1 (Congruence set)**.** *The congruence set $\mathcal{S}^{f_1=f_2}$ is defined as the subset of $\mathcal{S}$ instances receiving the same predictions from both f_1 and f_2, regardless of the prediction correctness:*

$$\mathcal{S}^{f_1=f_2} = \{\boldsymbol{x}_i \mid (\boldsymbol{x}_i, y_i) \in \mathcal{S} \ \wedge \ f_1(\boldsymbol{x}_i) = f_2(\boldsymbol{x}_i)\}\,.$$

Definition 2 (Backward compatibility set)**.** *The backward compatibility set $\mathcal{S}^f$ is defined as the subset of $\mathcal{S}$ instances receiving correct predictions from f:*

$$\mathcal{S}^f = \{\boldsymbol{x}_i \mid (\boldsymbol{x}_i, y_i) \in \mathcal{S} \ \wedge \ f(\boldsymbol{x}_i) = y_i\}\,.$$

If necessary, the intersection of the two sets can be taken to fulfill both definitions:

Definition 3 (Backward-compatible congruence set)**.** *The backward-compatible congruence set $\mathcal{S}^{f_1,f_2}$ is defined as the subset of $\mathcal{S}$ instances receiving correct predictions from both f_1 and f_2:*

$$\mathcal{S}^{f_1,f_2} = \{\boldsymbol{x}_i \mid (\boldsymbol{x}_i, y_i) \in \mathcal{S} \ \wedge \ f_1(\boldsymbol{x}_i) = f_2(\boldsymbol{x}_i) = y_i\}\,.$$

In the following we use the symbol $\mathcal{S}^\star$ as a generic notation for any of these subsets.

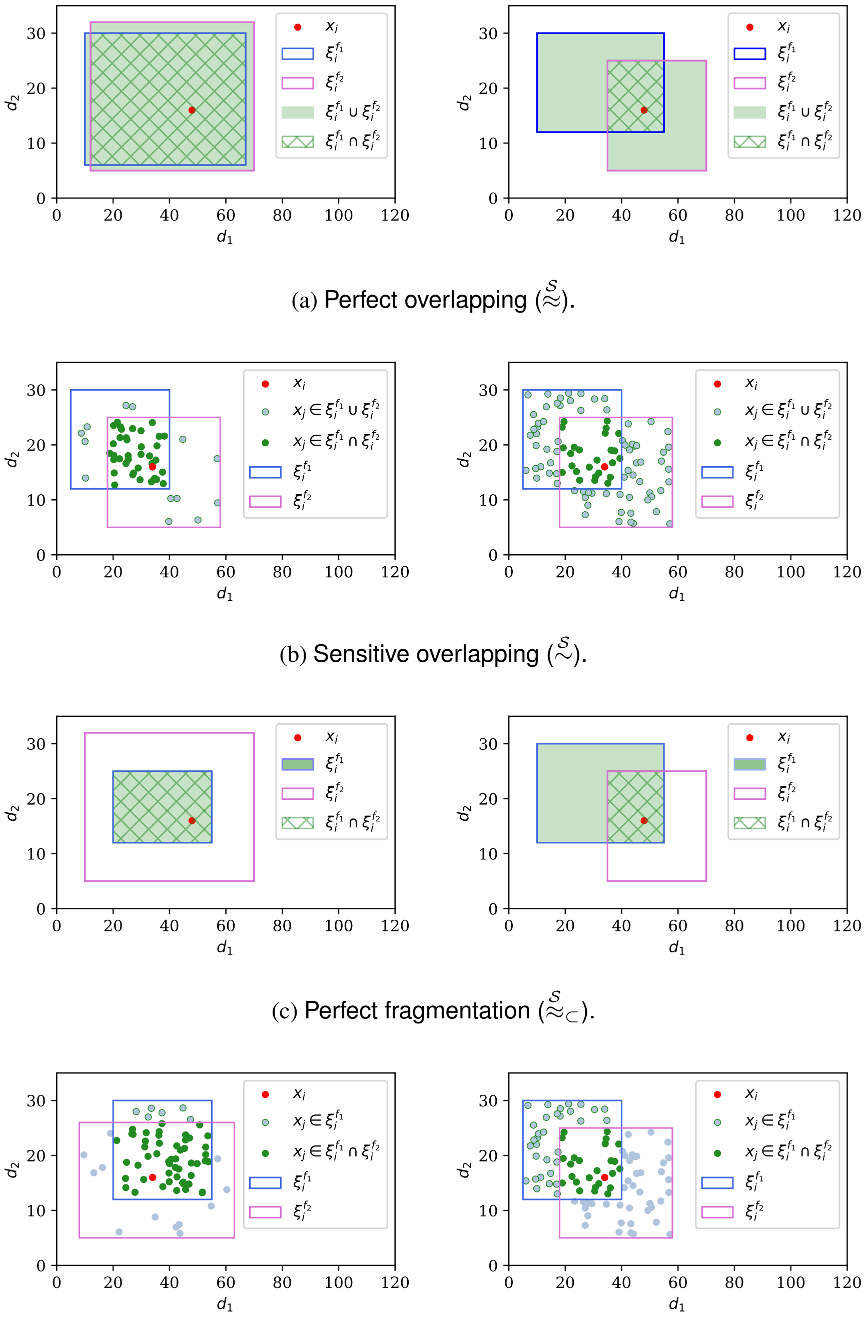
2.2. Perfect Overlapping

Two knowledge bases are considered *perfectly overlapping* if they provide identical local explanations for the same queries. This definition can be extended to produce not just a Boolean indicator of similarity, but also a numerical evaluation of the degree of overlap between the explanations provided for a set of instances.

Definition 4 (Perfect overlapping)**.** *Let f_1, f_2 be two functions representing the predictive process of two symbolic knowledge bases and e^{f_1}, e^{f_2} be the explaining functions for f_1 and f_2, respectively. The knowledge bases are* perfectly overlapping *if the explanations provided by them for each instance $\boldsymbol{x}_i$ of $\mathcal{S}^\star$ are perfectly overlapping, i.e., if $f_1 \overset{\mathcal{S}^\star}{\approx} f_2$ = 1:*

$$f_1 \overset{\mathcal{S}^\star}{\approx} f_2 = \frac{1}{|\mathcal{S}^\star|} \sum_{\boldsymbol{x}_i \in \mathcal{S}^\star} \frac{volume\left(e^{f_1}(\boldsymbol{x}_i) \cap e^{f_2}(\boldsymbol{x}_i)\right)}{volume\left(e^{f_1}(\boldsymbol{x}_i) \cup e^{f_2}(\boldsymbol{x}_i)\right)},$$

where $volume(\cdot)$ is a function expressing the volume of an input space subregion.



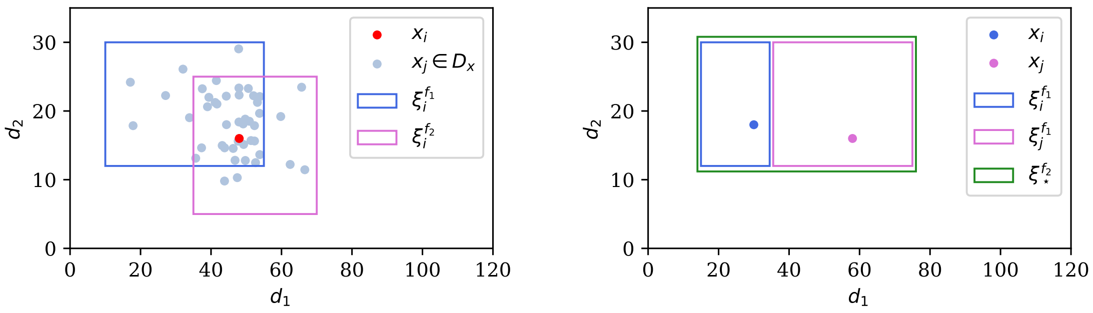
(a) Perfect overlapping ($\overset{\mathcal{S}}{\approx}$).

(b) Sensitive overlapping ($\overset{\mathcal{S}}{\sim}$).

(c) Perfect fragmentation ($\overset{\mathcal{S}}{\approx}_{\subset}$).

(d) Sensitive fragmentation ($\overset{\mathcal{S}}{\sim}_{\subset}$).

Figure 1: Examples of the proposed similarity metrics associated with high scores (left) and low scores (right).



(a) Explanations with non-uniform density.

(b) Fragmented explanations.

Figure 2: Different scenarios of overlapping explanations.

Definition 4 essentially represents an intersection-over-union operation, also known as Jaccard similarity, applied to pairs of local explanations for each instance in a given data set. This metric reaches a score of exactly 1 only if the explanations are perfectly equivalent, meaning the intersection of the explanations matches their union. The calculation of $\overset{\mathcal{S}}{\approx}$ is illustrated in Figure 1a.

While this definition serves as an ideal proxy for detecting perfect overlap between distinct knowledge bases, it presents two potential issues. The first issue arises when applying the knowledge bases to a data set with a non-homogeneous sample distribution, as illustrated in Figure 2a. In the figure, explanations $\xi_i^{f_1}$ and $\xi_i^{f_2}$ for the instance $\boldsymbol{x}_i \in D_x$ overlap but are not identical. The intersection-over-union measurement between $\xi_i^{f_1}$ and $\xi_i^{f_2}$ would yield a low score, indicating low perfect overlap for f_1 and f_2. However, the intersection of the two explanations includes the most relevant portion of the explanation union, specifically the region with the highest sample density $\boldsymbol{x}_j \in D_x$. Thus, even if the explanations are not identical, the regions of the input space within the union but outside the intersection may be considered less significant, and therefore less penalising in the similarity assessment. This concept is formalised into a scoring metric in Definition 5.

The second issue arises when comparing a knowledge base providing many distinct explanations with small coverage (f_1) to another generating fewer explanations with larger coverage (f_2), as illustrated in Figure 2b. In the figure, two instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j \in D_x$ have different local explanations $\xi_i^{f_1}$ and $\xi_j^{f_1}$, respectively, according to knowledge base f_1, while they share the same explanation $\xi_\star^{f_2}$ according to knowledge base f_2. However, $\xi_\star^{f_2}$ essentially represents the union of $\xi_i^{f_1}$ and $\xi_j^{f_1}$. Thus, the perfect overlap score for f_1 and f_2 would be low, even though the knowledge bases are nearly identical except for the fragmentation in the explanations provided by f_1. Since this fragmentation may not be a significant factor in assessing knowledge similarity in some applications, we formalise a scoring metric accordingly in Definition 6.

2.3. Sensitive Overlapping

The definition of perfect overlap can be relaxed to a notion of *sensitive overlap*, which measures the extent to which a pair of knowledge bases produce overlapping explanations *in the most relevant subregions of the input space* when queried with a set of instances.

Definition 5 (Sensitive overlapping)**.** *Let f_1, f_2 be two functions representing the predictive process of two symbolic knowledge bases and e^{f_1}, e^{f_2} be the explaining functions for f_1 and f_2, respectively. The knowledge bases are* sensitively overlapping *if the explanations provided by them for each instance $\boldsymbol{x}_i$ of $\mathcal{S}^\star$ are overlapping in the input space regions characterised by high sample density, i.e., if $f_1 \overset{\mathcal{S}^\star}{\sim} f_2$ = 1:*

$$f_1 \overset{\mathcal{S}^\star}{\sim} f_2 = \frac{1}{|\mathcal{S}^\star|} \sum_{\boldsymbol{x}_i \in \mathcal{S}^\star} \frac{\left| \{\boldsymbol{x}_j \mid \boldsymbol{x}_j \in e^{f_1}(\boldsymbol{x}_i) \ \cap \ e^{f_2}(\boldsymbol{x}_i)\} \right|}{\left| \{\boldsymbol{x}_j \mid \boldsymbol{x}_j \in e^{f_1}(\boldsymbol{x}_i) \ \cup \ e^{f_2}(\boldsymbol{x}_i)\} \right|}.$$

Definition 5 employs an intersection-over-union operation applied to pairs of local explanations, evaluated based on the number of enclosed data set samples rather than the explanation volumes. The metric reaches a score of 1 only if all data set samples are consistently enclosed within the intersection of the examined explanations, even if their union is larger. The calculation of $\overset{\mathcal{S}}{\sim}$ is illustrated in Figure 1b.

2.4. Perfect Fragmentation

The definition of perfect overlap can also be relaxed into a notion of *perfect fragmentation*, representing the extent to which a knowledge base provides explanations that are *perfectly enclosed within broader explanations* generated by a different knowledge base when both are queried with the same instances.

Definition 6 (Perfect fragmentation)**.** *Let f_1, f_2 be two functions representing the predictive process of two symbolic knowledge bases and e^{f_1}, e^{f_2} be the explaining functions for f_1 and f_2, respectively. The knowledge base f_1 is* perfectly fragmented *with respect to the knowledge f_2 if the explanation provided by f_1 for each instance $\boldsymbol{x}_i$ of $\mathcal{S}^\star$ is entirely enclosed within the explanation of f_2 for the same instance, i.e., if $f_1 \overset{\mathcal{S}^\star}{\approx}_{\subset} f_2$ = 1:*

$$f_1 \overset{\mathcal{S}^\star}{\approx}_{\subset} f_2 = \frac{1}{|\mathcal{S}^\star|} \sum_{\boldsymbol{x}_i \in \mathcal{S}^\star} \frac{volume\left(e^{f_1}(\boldsymbol{x}_i) \cap e^{f_2}(\boldsymbol{x}_i)\right)}{volume\left(e^{f_1}(\boldsymbol{x}_i)\right)},$$

where $volume(\cdot)$ has the same meaning as in Definition 4.

Definition 6 is consequently more adaptable than Definition 4, as it compares the intersection between explanations with a single explanation rather than with their union.

The computation of $\overset{\mathcal{S}}{\approx}_{\subset}$, depicted in Figure 1c, is asymmetrical and results in a score of 1 only if all the data set samples under consideration receive an explanation from the first operand that is entirely enclosed within the broader explanation generated by the second operand.

2.5. Sensitive Fragmentation

The concepts of sensitive overlap and perfect fragmentation can be combined into the definition of *sensitive fragmentation*, which assesses the similarity by considering the extent to which knowledge generates explanations whose most relevant subregion is perfectly enclosed within the broader explanation provided by a different knowledge.

Definition 7 (Sensitive fragmentation)**.** *Let f_1, f_2 be two functions representing the predictive process of two symbolic knowledge bases and e^{f_1}, e^{f_2} be the explaining functions for f_1 and f_2, respectively. The knowledge base f_1 is* sensitively fragmented *with respect to the knowledge f_2 if the explanation provided by f_1 for each instance $\boldsymbol{x}_i$ of $\mathcal{S}^\star$ is overlapping with the corresponding explanation provided by f_2 in the input space regions characterised by high sample density, i.e., if $f_1 \overset{\mathcal{S}^\star}{\sim}_{\subset} f_2$ = 1:*

$$f_1 \overset{\mathcal{S}^\star}{\sim}_{\subset} f_2 = \frac{1}{|\mathcal{S}^\star|} \sum_{\boldsymbol{x}_i \in \mathcal{S}^\star} \frac{\left| \{\boldsymbol{x}_j \mid \boldsymbol{x}_j \in e^{f_1}(\boldsymbol{x}_i) \, \cap \, e^{f_2}(\boldsymbol{x}_i)\} \right|}{\left| \{\boldsymbol{x}_j \mid \boldsymbol{x}_j \in e^{f_1}(\boldsymbol{x}_i)\} \right|}.$$

Definition 7 represents the least rigid metric amongst those proposed in this study, as it compares the intersection of explanations with a single explanation (rather than their union) and evaluates based on the number of enclosed data instances (rather than volumes).

The sensitive fragmentation metric is asymmetrical and yields a score of 1 only when all the data set samples enclosed within the explanation provided by the first operand are also enclosed within the explanation provided by the second operand, for every pair of explanations produced for the considered data set instances. The calculation of $\overset{\mathcal{S}}{\sim}_{\subset}$ is illustrated in Figure 1d.

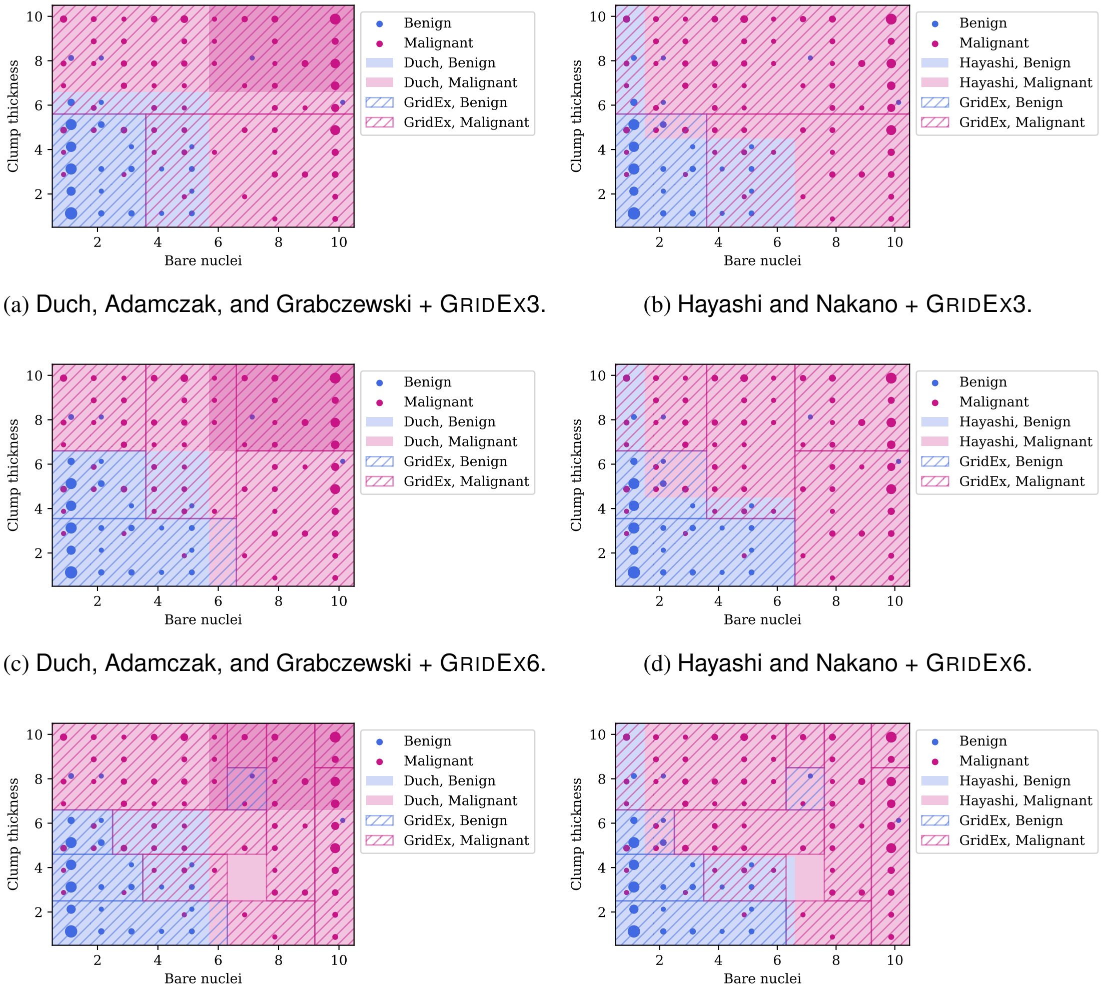
3. Experiments

We conducted experiments using the presented similarity metrics on the original Wisconsin Breast Cancer data set [11], which comprises nine input features ranging from 1 to 10 and represents cases of breast cancer. The data set's binary output class indicates whether the tumour is benign or malignant.

Our objective was to compare existing knowledge bases in the literature with alternatives obtained by applying SKE techniques. We compared these based on accuracy, readability, and completeness of the knowledge pieces against the data set ground truth, through both individual evaluations and the $FiRe$ aggregating scoring metric [12]. Additionally, we measured the similarity between the extracted knowledge bases and those existing in the literature, taken as reference, as described in this study.

The knowledge bases used as reference are from the works of Duch et al. (2001) and Hayashi and Nakano (2015). The corresponding classification rules are displayed in Listings 1 and 2, respectively. Notably, both sets of rules base predictions on two input features: "Bare Nuclei" (BN) and "Clump Thickness" (CT). Decision boundaries identified by these knowledge bases are illustrated in Figure 3 as coloured rectangles.

The GridEx SKE algorithm [15] was utilised to extract symbolic knowledge from a gradient boosting (GB) classifier trained on the WBC data set. The hyper-parameters of the GB, namely



(a) Duch, Adamczak, and Grabczewski + GRIDEX3.

(b) Hayashi and Nakano + GRIDEX3.

(c) Duch, Adamczak, and Grabczewski + GRIDEX6.

(d) Hayashi and Nakano + GRIDEX6.

(e) Duch, Adamczak, and Grabczewski + GRIDEX12.

(f) Hayashi and Nakano + GRIDEX12.

Figure 3: Comparison of decision boundaries of the knowledge rules proposed by Duch, Adamczak, and Grabczewski (2001) and Hayashi and Nakano (2015) for the original WBC data set (highlighted with coloured backgrounds) with the knowledge extracted using multiple GRIDEX instances applied to a gradient boosting classifier trained on the same data set (represented by hatched regions). Only the two input features used to make predictions on the data set are displayed. Samples are depicted as circles indicating the quantity of positive and negative classes, with the radius of the circles corresponding to the count of instances across both classes. Notably, denser background colours signify overlapping rules, only present in the knowledge base by Duch, Adamczak, and Grabczewski (left panels), while small uncovered regions are considered negligible due to the absence of training samples in the knowledge extracted with GRIDEX12 (bottom panels).

Listing 1 Knowledge base adapted from Duch, Adamczak, and Grabczewski (2001) for the WBC data set.

```
Malignant if BN > 5.5
Malignant if CT > 6.5
Benign if BN <= 5.5 ∧ CT <= 6.5
```

Listing 2 Knowledge base adapted from Hayashi and Nakano (2015) for the WBC data set.

```
Benign if BN <= 1.5
Malignant if BN > 1.5 ∧ CT > 4.5
Malignant if BN > 6.5 ∧ CT <= 4.5
Benign if 1.5 < BN <= 6.5 ∧ CT <= 4.5
```

Listing 3 Knowledge base extracted with GRIDEX3.

```
Benign if BN <= 4.0 ∧ CT <= 5.5
Malignant if BN > 4.0 ∧ CT <= 5.5
Malignant if CT > 5.5
```

Listing 4 Knowledge base extracted with GRIDEX6.

```
Benign if BN <= 4.0 ∧ 4.0 < CT <= 7.0
Malignant if BN <= 4.0 ∧ CT > 7.0
Malignant if BN > 7.0 ∧ CT > 7.0
Benign if BN <= 7.0 ∧ CT <= 4.0
Malignant if BN > 7.0 ∧ CT <= 7.0
Malignant if 4.0 < BN <= 7.0 ∧ CT > 4.0
```

the number of estimators and learning rate, were tuned via a grid search and set to 133 and 0.2033, respectively. The GB was trained by using 75% of the WBC data instances as the training set, while the remaining samples were reserved for the test set, resulting in an observed test accuracy of 0.99 for the trained model. Several GRIDEX instances with varying parameters were applied to the trained GB. We leveraged the implementation provided within the PSYKE framework [16, 17, 18]. Three instances, employing an adaptive splitting strategy with a recursion depth of 1, were chosen to generate knowledge bases containing 3, 6, and 12 items, respectively, for subsequent analysis and comparison with reference knowledge pieces. These instances are henceforth referred to as GRIDEX3, GRIDEX6, and GRIDEX12, respectively. The corresponding classification rules are summarised in Listings 3 to 5, and their decision boundaries are depicted in Figure 3 as hatched regions superimposed onto the reference knowledge boundaries.

Quality evaluations regarding classification accuracy, human readability (quantified as knowledge size), and completeness (represented by the percentage of covered test samples; [19]) for all

Listing 5 Knowledge base extracted with GridEx12.

```
Benign if BN <= 2.3 ∧ 4.6 < CT <= 6.4
Benign if 6.1 < BN <= 7.4 ∧ 6.4 < CT <= 8.2
Malignant if 6.1 < BN <= 7.4 ∧ CT > 8.2
Malignant if BN > 8.7 ∧ CT > 8.2
Malignant if 6.1 < BN <= 8.7 ∧ CT <= 2.8
Benign if BN <= 3.6 ∧ 2.8 < CT <= 4.6
Malignant if 3.6 < BN <= 6.1 ∧ 2.8 < CT <= 4.6
Benign if BN <= 6.1 ∧ CT <= 2.8
Malignant if BN > 8.7 ∧ CT <= 8.2
Malignant if 7.4 < BN <= 8.7 ∧ CT > 2.8
Malignant if 2.3 < BN <= 7.4 ∧ 4.6 < CT <= 6.4
Malignant if BN <= 6.1 ∧ CT > 6.4
```

Table 1
Quality assessments for the knowledge bases adopted in our experiments.

Knowledge	Accuracy	Size	Coverage	6-$FiRe$
Duch et al.	0.94	**3**	**1.00**	0.061
Hayashi et al.	0.94	4	**1.00**	0.062
GridEx3	0.93	**3**	**1.00**	0.070
GridEx6	**0.96**	6	**1.00**	**0.045**
GridEx12	0.94	12	**1.00**	0.131

knowledge pieces utilised in the experiments are detailed in Table 1 for the test set (with the best values highlighted in bold). Additionally, we introduced the $FiRe$ metric with a fidelity/readability trade-off parameter $\psi = 6$ (6-$FiRe$) to offer a more concise quality assessment. It is worth noting that high-quality knowledge corresponds to low $FiRe$ scores, and higher ψ values prioritise predictive accuracy while disregarding readability. Furthermore, similarity assessments between the extracted and reference knowledge bases, evaluated on the backward-compatible congruence set (Definition 3), are summarised in Table 2. It is important to mention that the robustness of the results presented in Tables 1 and 2 has been confirmed through 10-fold cross-validation repeated 10 times. Given the low variability of the results, we provide here only the average values without their corresponding standard deviations.

Amongst the reference knowledge bases considered, the one presented by Duch et al. exhibits identical accuracy and coverage compared to that offered by Hayashi and Nakano. However, the former comprises fewer items, leading to knowledge that is more readable and consequently of higher quality. This is evidenced by a lower 6-$FiRe$ score, as anticipated.

Similarly, amongst the extracted knowledge bases, the one provided by GridEx6 emerges as the best according to the $FiRe$ score, boasting the highest accuracy and completeness, despite its suboptimal size. Indeed, any loss in human readability is offset by the gain in accuracy. This

Table 2
Similarity measurements between the GridEx outputs (f_1) and the reference knowledge bases (f_2).

$f_1 \downarrow \quad f_2 \rightarrow$	Duch et al.	Hayashi et al.
GridEx3	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx} f_2$ = 0.48	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx} f_2$ = 0.34
	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim} f_2$ = 0.82	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim} f_2$ = 0.86
	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx_\subset} f_2$ = 0.83	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx_\subset} f_2$ = 0.46
	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim_\subset} f_2$ = 0.92	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim_\subset} f_2$ = 0.92
GridEx6	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx} f_2$ = 0.34	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx} f_2$ = 0.14
	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim} f_2$ = 0.50	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim} f_2$ = 0.44
	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx_\subset} f_2$ = 0.84	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx_\subset} f_2$ = 0.42
	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim_\subset} f_2$ = 0.99	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim_\subset} f_2$ = 0.87
GridEx12	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx} f_2$ = 0.23	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx} f_2$ = 0.08
	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim} f_2$ = 0.40	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim} f_2$ = 0.32
	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx_\subset} f_2$ = 0.97	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx_\subset} f_2$ = 0.42
	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim_\subset} f_2$ = 1.00	$f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim_\subset} f_2$ = 0.90

knowledge piece also outperforms the reference ones for the same reasons. Upon comparing the knowledge bases of GridEx6 (f_1) and Duch et al. (f_2), it becomes apparent that their decision boundaries are not identical. This enables GridEx to achieve superior predictive performance. However, the diversity in boundaries is reflected in a low perfect-overlapping score: $f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx} f_2$ = 0.34. Given the distribution of data set samples across the classification rules, the sensitive-overlapping score is also low: $f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim} f_2$ = 0.50. From these perspectives, the knowledge bases are dissimilar, and substituting the reference knowledge with the extracted one does not ensure coherence and continuity.

A more thorough examination of the disparities between the two knowledge bases reveals that their boundaries differ because GridEx6 comprises more, albeit smaller, classification rules. Consequently, the perfect-fragmentation score is elevated, as the GridEx6 rules are predominantly enclosed within the reference ones: $f_1 \overset{\mathcal{S}^{f_1,f_2}}{\approx_\subset} f_2$ = 0.84. Given that non-overlapping rule regions typically exhibit a sparse sample density, it follows that the sensitive-fragmentation score is even higher: $f_1 \overset{\mathcal{S}^{f_1,f_2}}{\sim_\subset} f_2$ = 0.99. Hence, continuity and coherence are maintained if preserving boundaries is not imperative.

In summary, aligning with the visual examination, it can be concluded that the knowledge base of GridEx6 does not exhibit perfect or sensitive overlap with the reference provided by Duch et al.. However, it is nearly perfectly fragmented compared to the latter, earning an optimal score when assessing its degree of sensitive fragmentation.

4. Discussion and conclusion

Our study introduces novel similarity metrics tailored for comparing knowledge bases within multi-agent systems, addressing the critical need of effective knowledge assessment in autonomous decision-making environments. Through our experiments and analyses, we have demonstrated the practical applicability and significance of these metrics in enhancing various aspects of multi-agent systems' functionality and efficiency.

The results of our experiments highlight the utility of our proposed similarity metrics in facilitating collaborative decision-making, knowledge sharing, coordination, and resource allocation within multi-agent systems. Specifically, our metrics enable agents to assess the compatibility of their internal knowledge, identify areas of agreement and conflict, and optimise their interactions accordingly. Moreover, they aid in determining which pieces of knowledge are most relevant to share amongst agents and match agents to tasks based on their knowledge profiles and task requirements.

In conclusion, our study underscores the importance of effective knowledge assessment in multi-agent systems and introduces novel similarity metrics designed to meet this need. By leveraging these metrics, agents can better understand the landscape of their collective knowledge, leading to more effective collaboration, decision-making, and coordination within the system. We believe that our proposed metrics represent a significant step forward in advancing the capabilities of multi-agent systems and pave the way for further research and development in this field.

Acknowledgments

This work has been supported by the EU ICT-48 2020 project TAILOR (No. 952215) and the European Union's Horizon Europe AEQUITAS research and innovation programme under grant number 101070363.

References

[1] V. Marik, M. Pechoucek, O. Štepánková, Social knowledge in multi-agent systems, Multi-Agent Systems and Applications: 9th ECCAI Advanced Course, ACAI 2001 and Agent Link's 3rd European Agent Systems Summer School, EASSS 2001 Prague, Czech Republic, July 2–13, 2001 Selected Tutorial Papers 9 (2001) 211–245.

[2] P. Panzarasa, N. R. Jennings, T. J. Norman, Formalizing collaborative decision-making and practical reasoning in multi-agent systems, Journal of logic and computation 12 (2002) 55–117.

[3] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, Semantic measures for the comparison of units of language, concepts or instances from text and knowledge base analysis, arXiv preprint arXiv:1310.1285 (2013).

[4] C. D. Manning, P. Raghavan, H. Schütze, Introduction to information retrieval, Cambridge University Press, 2008. URL: https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf. doi:10.1017/CBO9780511809071.

[5] A. H. Murphy, The Finley affair: A signal event in the history of forecast verification, Weather and forecasting 11 (1996) 3–20.

[6] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[7] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, Hypercube-based methods for symbolic knowledge extraction: Towards a unified model, in: A. Ferrando, V. Mascardi (Eds.), WOA 2022 – 23rd Workshop "From Objects to Agents", volume 3261 of *CEUR Workshop Proceedings*, Sun SITE Central Europe, RWTH Aachen University, 2022, pp. 48–60. URL: http://ceur-ws.org/Vol-3261/paper4.pdf.

[8] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, Towards a unified model for symbolic knowledge extraction with hypercube-based methods, Intelligenza Artificiale 17 (2023) 63–75. URL: https://doi.org/10.3233/IA-230001. doi:10.3233/IA-230001.

[9] G. Ciatto, F. Sabbatini, A. Agiollo, M. Magnini, A. Omicini, Symbolic knowledge extraction and injection with sub-symbolic predictors: A systematic literature review, ACM Computing Surveys 56 (2024) 161:1–161:35. URL: https://doi.org/10.1145/3645103. doi:10.1145/3645103.

[10] J. L. Haggerty, R. J. Reid, G. K. Freeman, B. H. Starfield, C. E. Adair, R. McKendry, Continuity of care: A multidisciplinary review, Bmj 327 (2003) 1219–1221.

[11] W. H. Wolberg, O. L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology., Proceedings of the national academy of sciences 87 (1990) 9193–9196.

[12] F. Sabbatini, R. Calegari, Symbolic knowledge-extraction evaluation metrics: The FiRe score, in: K. Gal, A. Nowé, G. J. Nalepa, R. Fairstein, R. Rădulescu (Eds.), Proceedings of the 26th European Conference on Artificial Intelligence, ECAI 2023, Kraków, Poland. September 30 – October 4, 2023, 2023. URL: https://ebooks.iospress.nl/doi/10.3233/FAIA230496. doi:10.3233/FAIA230496.

[13] W. Duch, R. Adamczak, K. Grabczewski, A new methodology of extraction, optimization and application of crisp and fuzzy logical rules, IEEE Transactions on Neural Networks 12 (2001) 277–306. URL: https://doi.org/10.1109/72.914524. doi:10.1109/72.914524.

[14] Y. Hayashi, S. Nakano, Use of a recursive-rule extraction algorithm with J48graft to achieve highly accurate and concise rule extraction from a large breast cancer dataset, Informatics in Medicine Unlocked 1 (2015) 9–16.

[15] F. Sabbatini, G. Ciatto, A. Omicini, GridEx: An algorithm for knowledge extraction from black-box regressors, in: D. Calvaresi, A. Najjar, M. Winikoff, K. Främling (Eds.), Explainable and Transparent AI and Multi-Agent Systems. Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers, volume 12688 of *LNCS*, Springer Nature, Basel, Switzerland, 2021, pp. 18–38. doi:10.1007/978-3-030-82017-6_2.

[16] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, On the design of PSyKE: A platform for symbolic knowledge extraction, in: R. Calegari, G. Ciatto, E. Denti, A. Omicini, G. Sartor (Eds.), WOA 2021 – 22nd Workshop "From Objects to Agents", volume 2963 of *CEUR Workshop Proceedings*, Sun SITE Central Europe, RWTH Aachen University,

2021, pp. 29–48. 22nd Workshop “From Objects to Agents” (WOA 2021), Bologna, Italy, 1–3 September 2021. Proceedings.

[17] R. Calegari, F. Sabbatini, The PSyKE technology for trustworthy artificial intelligence 13796 (2023) 3–16. URL: https://doi.org/10.1007/978-3-031-27181-6_1. doi:10.1007/978-3-031-27181-6_1, xXI International Conference of the Italian Association for Artificial Intelligence, AIxIA 2022, Udine, Italy, November 28 – December 2, 2022, Proceedings.

[18] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, Symbolic knowledge extraction from opaque ML predictors in PSyKE: Platform design & experiments, Intelligenza Artificiale 16 (2022) 27–48. URL: https://doi.org/10.3233/IA-210120. doi:10.3233/IA-210120.

[19] F. Sabbatini, R. Calegari, On the evaluation of the symbolic knowledge extracted from black boxes, AI and Ethics 4 (2024) 65–74. doi:https://doi.org/10.1007/s43681-023-00406-1.