ICE: An Evaluation Metric to Assess Symbolic Knowledge Quality

Federico Sabbatini[1] and Roberta Calegari[2]

[1] University of Urbino, Urbino, Italy f.sabbatini1@campus.uniurb.it
[2] University of Bologna, Bologna, Italy roberta.calegari@unibo.it

Abstract. The automated assessment of symbolic knowledge, derived, for instance, from extraction procedures, facilitates the autotuning of machine learning algorithms, obviating inherent biases in subjective human evaluations. Despite advancements, comprehensive metrics for evaluating knowledge quality are missing in the literature. To address this gap, our study introduces ICE, a novel evaluation metric designed to measure the quality of symbolic knowledge. This metric computes a score by considering three quality sub-indices, namely, predictive performance, human readability and completeness, and it can be tailored to suit the specific requirements of the case at hand by adjusting the weights assigned to each sub-index. We present here the mathematical formulation of the ICE score, and show its effectiveness through comparative analyses with existing quality scores applied to real-world tasks.

Keywords: Explainable artificial intelligence · Symbolic knowledge extraction · AutoML

1 Introduction

In the current landscape of artificial intelligence (AI), symbolic knowledge extraction (SKE) has gained widespread utilisation to address the interpretability challenges associated with sub-symbolic AI, characterised by efficacy in predictions, but often relying on complex models which pose challenges in terms of interpretability and explainability [14, 33]. SKE methodologies involve knowledge extraction from "black-box" models [15, 18], aiming to construct surrogate symbolic representations. These techniques play a crucial role in enhancing the interpretability and explainability of machine learning (ML) models, enabling human understanding and trust in decision-making processes.

The literature on SKE techniques emphasises the absence of universally optimum solutions [3, 6, 7, 26, 28, 34, 36, 37]. This inherent variability necessitates the systematic exploration of multiple SKE techniques to select the optimum approach for a given case. The extracted knowledge quality is linked to factors such as data distribution, pre-processing strategies, and feature selection techniques. Consequently, a rigorous evaluation of the knowledge is imperative to compare the efficacy of diverse techniques within the specific context of interest. Assessing the quality of knowledge derived through SKE involves several indices,

including accuracy, completeness, and readability [10, 20, 39]. However, manual evaluation of these indices is time-consuming and subject to subjective biases. Automation of this evaluation process aligns with an automated machine learning (AutoML) perspective [13], offering efficiency and objectivity in the selection of suitable SKE techniques.

While recent efforts have introduced metrics for automated evaluation, these metrics remain limited in scope, lacking the comprehensive coverage of necessary evaluation criteria and the integration of user feedback and customisation. Accordingly, in this paper we propose the Index for Complete quality Evaluation (ICE) as a scoring metric designed to comprehensively evaluate knowledge quality. ICE aims to advance the state of the art by facilitating the automated evaluation and comparison of symbolic knowledge, providing a complete, objective and quantitative assessment of the outputs from SKE procedures.

2 Related Works and Background Notions

SKE techniques currently find application in addressing a diverse exhibition of real-world challenges, particularly in critical domains where interpretability and human comprehension are imperative [2, 11]. These methods typically yield knowledge outputs represented symbolically, facilitating interpretable predictions. The literature encompasses numerous SKE techniques, necessitating the execution of various experiments to identify optimum approaches. The comparison of different outcomes – i.e., the knowledge extracted – is essential in this selection process to select the best approach.

Existing literature widely recognises that knowledge quality can be assessed based on predictive performance, human readability, and completeness [1, 7, 12, 19, 35, 40]. The results of these evaluations depend on both the chosen SKE algorithm and the user-defined parameters controlling the algorithm's behaviour. Consequently, comparisons can be conducted not only between distinct extraction procedures, but also between instances of the same extractor with varying parametrisations. For knowledge to be deemed of high quality, it must concurrently exhibit superior predictive performance, human readability, and completeness. Predictive performance concerns the knowledge's ability to provide accurate outputs. Readability quantifies the human effort required to comprehend the rationale behind the predictions. Completeness measures the proportion of predictions that the knowledge can provide in response to user queries. The conventional approach to knowledge comparisons typically involves a manual examination of individual quality indices. However, such a method is susceptible to human subjectivity (and possible biases) and lacks the capability for automated assessments. Moreover, the comparison of a set of knowledge is straightforward when there exists a candidate knowledge that maximises all three indices, rendering it the optimum knowledge within the set. Regrettably, real-world applications often present a fidelity/readability trade-off, wherein a comparison is made between knowledge exhibiting high predictive performance but limited readability and knowledge characterised by enhanced human read-

ability but diminished predictive performance [17]. In such scenarios, selecting the best knowledge necessitates a thorough consideration of all three quality indices, free from human bias or material errors. Nonetheless, it remains crucial to offer human users the ability to assign appropriate weights to different quality indices, enabling adaptation of the comparison to the sensitivity and objectives of the given task. This is particularly relevant in situations where the emphasis may vary, such as instances where predictive performance is prioritised over readability, as opposed to scenarios where readability is an imperative consideration.

To the best of our knowledge, to date, only two metrics have been established for the explicit purpose of evaluating the knowledge quality: FIRE [24] and Q_s [27]. FIRE, while overlooking knowledge completeness in its quality assessment, allows for the incorporation of a user-defined parameter to adjust the importance of knowledge readability and predictive performance, respectively. The metric is as a multiplicative scoring function, considering predictive performance and human readability as "losses", i.e., predictive loss as predictive error and readability loss as knowledge size. Consequently, smaller FIRE scores are indicative of higher knowledge quality, given that losses are essentially multiplied. Similarly, Q_s is grounded in the multiplication of index losses. Noteworthy distinctions from FIRE lie in the inclusion of knowledge completeness loss and the exclusion of user-defined capabilities to adjust relative loss weights. No other metrics evaluating symbolic knowledge quality have been proposed in the literature. Consequently, a comprehensive metric is lacking—one that incorporates predictive performance, human readability, and completeness indices while allowing for the customisation of their relative importance in overall score calculation.

Such a metric would serve as a foundational element for enabling an unbiased, standardised, and concise evaluation of symbolic knowledge quality. It would be also crucial for AutoML procedures, as it facilitates the automatic selection of high-quality symbolic knowledge representations, leading to more precise and efficient ML-based systems. Without such an evaluation metric, AutoML algorithms may inadvertently choose suboptimum symbolic knowledge representations, resulting in subpar model performance and resource wastage.

In this study we introduce ICE as a comprehensive scoring function addressing the current literature gap for knowledge quality assessment. The ICE score enhances the efficacy of previously introduced metrics (FIRE and Q_s) by incorporating all three indices from the literature to evaluate symbolic knowledge (predictive performance, readability, and completeness). As a result it provides a quantitative assessment of knowledge quality, also empowering users to customise weight parameters, assigning varying importance to the three indices based on the task and user requirements. Consequently, different symbolic knowledge instances can be easily compared using the ICE metric.

Quality Indices The evaluation of knowledge quality commonly relies on three primary indices: predictive performance, human-readability extent, and completeness [10, 39]. Each index lacks a unique definition since it often depends

on the task at hand. Predictive performance evaluation mirrors approaches applicable to black-box models or other predictors. The assessment may involve comparing the ground truth of a data set or the outputs of an opaque model emulated by the symbolic knowledge. Task-dependent evaluations prevail, with accuracy and F_1 scores being standard for classification tasks, while mean absolute/squared error (MAE/MSE) and the R^2 score are commonplace for regression tasks. Readability is often associated with knowledge size [8]. For instance, an SKE algorithm generating a list of n rules may be deemed more human-readable than another procedure presenting a list or tree with $2n$ rules or leaves. While readability information can extend to the complexity of individual knowledge items, quantitative and formal assessments of this complexity are currently unavailable. For example, comparing a tree leaf's readability describing an M-of-N logic rule to a decision table entry associated with a fuzzy rule lacks established techniques [24]. Consequently, knowledge size is generally considered sufficient to express readability due to its straightforward interpretation. Completeness is estimated as the percentage of the input feature space covered by the knowledge, representing the subspace where predictions can be made. In instances where this measurement proves impractical, such as data sets with a multitude of input features, completeness estimation can be achieved by querying the knowledge with a set of instances and calculating the percentage of provided responses.

3 The ICE Score

The ICE (Index for Complete quality Evaluation) score provides an evaluation encompassing predictive performance, human readability and completeness of the analysed knowledge. To achieve this goal the first two indices are squashed in the $(0, 1)$ open interval with a generalised sigmoid function (parametrised by the user) and then they are multiplied together and by the completeness index. The user-defined parameters used inside the generalised sigmoid functions enable the ICE score to assign a customised relative relevance to predictive performance and human readability. No parameters are required for the completeness importance since this latter is assumed equal to 1 by default to avoid an unnecessary complex score formulation. The completeness does not need to be normalised in the specified interval, because it is naturally expressed as a percentage, and thus it always lies within the $[0, 1]$ closed interval.

ICE is a multiplicative metric assuming that good quality knowledge is associated with high values of all the three underlying indices and, conversely, bad quality is associated with small values of at least one index. Therefore, the ICE score can assume values in the $[0, 1)$ half-open interval. Ideally, good knowledge should have ICE score as close as possible to 1. The ICE score is defined as the following continuous and differentiable function:

$$ICE : (\mathbb{R}_{\leq 1} \times \mathbb{R}_{>0} \times [0, 1] \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}) \mapsto [0, 1), \tag{1}$$

$$ICE(p,\ r,\ c,\ \varphi,\ \rho) \quad = \quad P(p,\ \varphi) \cdot R(r,\ \rho) \cdot c, \tag{2}$$

where p, r and c are the raw knowledge predictive performance, size and completeness measurements, respectively, φ is the relative importance assigned to the raw predictive performance, and ρ is the one assigned to the knowledge size.$P(\cdot)$ and $R(\cdot)$ are the continuous and differentiable functions expressing the knowledge *accuracy* and *readability*, respectively, defined as follows:

$$P \ : \ (\mathbb{R}_{\leq 1} \times \mathbb{R}_{>0}) \mapsto (0,1), \tag{3}$$

$$P(p,\varphi) = \left(1 + e^{5(\varphi(1-p)-1)}\right)^{-1}, \tag{4}$$

$$R \ : \ (\mathbb{R}_{>0} \times \mathbb{R}_{>0}) \mapsto (0,1), \tag{5}$$

$$R(r,\rho) = \left(1 + e^{0.3\rho r - 5}\right)^{-1}. \tag{6}$$

The accuracy function $P(p,\varphi)$ used for the ICE calculation is an ad-hoc function representing the knowledge raw predictive performance p squashed in $(0,1)$ and weighted with respect to the user-defined importance parameter φ. Analogously, the readability function $R(r,\rho)$ squashes in the same interval the raw human-readability extent (i.e., the knowledge size) r and weights it according to the user-defined importance parameter ρ. Since the completeness importance is always considered equal to 1 in the ICE score calculation, users can act on the φ and ρ parameters to decide the relative importance of knowledge predictive performance and size with respect to each other and the completeness by selecting values smaller or larger than 1. If both φ and ρ are equal to 1, then the ICE score equally weights the three indices. The ICE score trend for different values of its parameters is reported in Figure 1.

We point out that the fixed values used to parametrise the exponentials within the $P(\cdot)$ and $R(\cdot)$ functions have been carefully tuned to obtain "well-behaved" sigmoid functions suitable to be customised by users only by providing a single additional importance parameter. In this context, a well-behaved sigmoid *(i)* tends to 1 when representing desirable knowledge properties (e.g., high predictive performance or human readability), *(ii)* tends to 0 when denoting poor knowledge quality, and *(iii)* exhibits a growth rate tunable via a single user-defined parameter, representing the importance of the sigmoid function in the overall ICE score calculation. The fixed values of Equations (4) and (6) (e.g. 0.3, 5) were chosen after a thorough study of the aforementioned properties and utilizing optimisation tools that provide sigmoid curve fitting.

In the following, we analyse the ICE function domain, the properties of the accuracy and readability functions and those resulting in the ICE score.

3.1 ICE Function Domain

The ICE scoring function is defined in the domain reported in Equation (1), given the following assumptions: (i) knowledge raw predictive performance (p) assessed via task-dependent scores equal to 1 in the best case, may have no lower-bound in the worst case; (ii) knowledge size (r) an integer number greater or equal to 1 (knowledge contains at least one item). To enhance score flexibility the corresponding considered range has been extended to all positive real numbers;

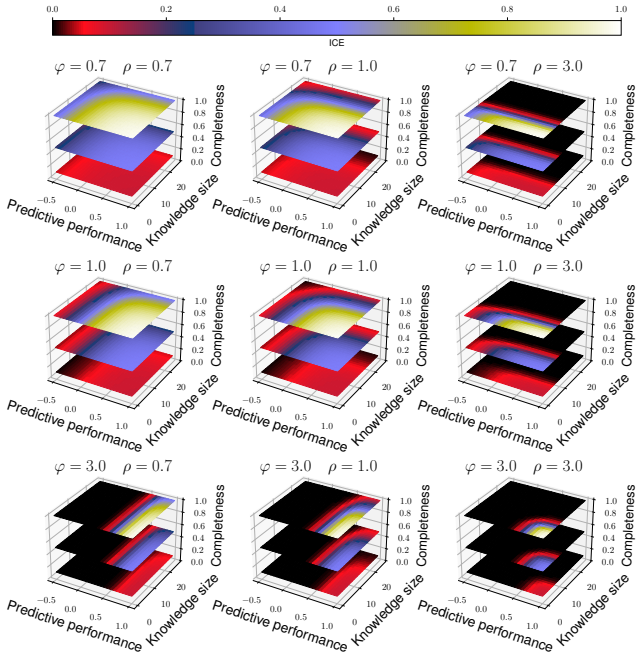

Fig. 1: ICE score for different values of knowledge predictive performance p (x-axis), size r (y-axis) and completeness c (z-axis), predictive performance importance φ (rows) and size importance ρ (columns). Different colours represent different ICE score values (top colourbar).

(iii) knowledge coverage (c) expressed as a percentage and mapped in the $[0, 1]$ interval; (iv) predictive performance importance (φ) a positive real number by design; (v) knowledge size importance (ρ) a positive real number by design.

3.2 ICE Accuracy Function

ICE adopts the $P(\cdot)$ accuracy function shown in Equations (3) and (4) and depicted in the left panel of Figure 2 to apply the user-defined weight φ to the raw predictive performance p measured for the knowledge. The left part of Figure 3 reports the accuracy for different fixed values of the φ weight and the function's first and second partial derivatives with respect to p. The function is bounded in $(0, 1)$ for any possible value of p and φ. Indeed, from Equation (4):

$$\lim_{p \to -\infty} P(p, \varphi) = 0, \quad \forall \varphi > 0, \tag{7}$$

$$\lim_{p \to 1} P(p, \varphi) = 1, \quad \forall \varphi > 0, \tag{8}$$

$$\lim_{\varphi \to \infty} P(p, \varphi) = 0, \quad \forall p < 1, \tag{9}$$

$$\lim_{\varphi \to 0} P(p, \varphi) = 1, \quad \forall p < 1. \tag{10}$$

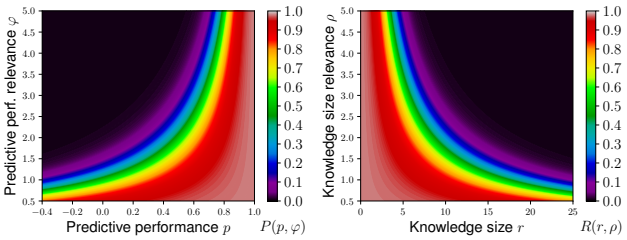

Fig. 2: ICE accuracy and readability function trends in their domains. Functions' input parameters are reported in the axes, the corresponding value is represented by the colour (small values are associated with dark colours).

The accuracy function is monotonically increasing with respect to p and decreasing with respect to φ, indeed:

$$0 < \frac{\partial P}{\partial p} = \frac{5\ \varphi\ e^{5(\varphi(1-p)-1)}}{\left(1+e^{5(\varphi(1-p)-1)}\right)^2}, \tag{11}$$

$$0 > \frac{\partial P}{\partial \varphi} = -\frac{5\ (1-p)\ e^{5(\varphi(1-p)-1)}}{\left(1+e^{5(\varphi(1-p)-1)}\right)^2}, \tag{12}$$

$$p_1 < p_2 \iff P(p_1, \varphi) < P(p_2, \varphi), \tag{13}$$

$$\varphi_1 < \varphi_2 \iff P(p, \varphi_1) > P(p, \varphi_2). \tag{14}$$

Equations from (11) to (14) hold $\forall p, p_1, p_2 \in \mathbb{R}_{\leq 1}$, $\forall \varphi, \varphi_1, \varphi_2 \in \mathbb{R}_{>0}$, with the only exception of Equation (12) strictly requiring $p < 1$.

Another interesting property of the accuracy function is the flex point observed for a fixed value of φ. The flex point analysis gives an insight about the relationship between p and φ within the accuracy function. Equating to 0 the second partial derivative of $P(\cdot)$ with respect to p we obtain:

$$\frac{\partial^2 P}{\partial p^2} = 25\ \varphi^2\ \frac{e^{5(\varphi(1-p)-1)} - e^{10(\varphi(1-p)-1)}}{\left(1+e^{5(\varphi(1-p)-1)}\right)^3} = 0, \qquad p = 1 - \frac{1}{\varphi}. \tag{15}$$

It is possible to obtain the same result through the second partial derivative for φ. Anyway, given the flex point properties of generic sigmoid functions ranging in $(0, 1)$, $P(p, \varphi) = 0.5\ \forall p, \varphi \mid p = 1 - \frac{1}{\varphi}$.

3.3 ICE Readability Function

ICE adopts the $R(\cdot)$ readability function reported in Equations (5) and (6) to apply the user-defined weight ρ to the measured knowledge size r. The function is shown in the right panels of Figures 2 and 3. In Figure 3 the readability score obtained for a set of ρ weight values is reported. Similarly to the accuracy,

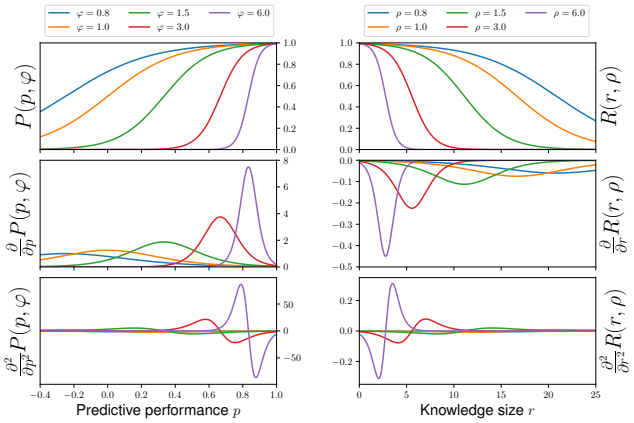

Fig. 3: ICE accuracy and readability functions with respect to predictive performance p, knowledge size r, φ and ρ (top panels). First and second partial derivatives (middle and bottom panels, respectively) for p and r.

also the readability function is bounded in $(0, 1)$ for any possible value of its parameters, as ensured by Equation (6), indeed:

$$\lim_{r \to \infty} \; R(r, \rho) = 0, \quad \forall \rho > 0, \tag{16}$$

$$\lim_{r \to 0} \; R(r, \rho) = 1, \quad \forall \rho > 0, \tag{17}$$

$$\lim_{\rho \to \infty} \; R(r, \rho) = 0, \quad \forall r > 0, \tag{18}$$

$$\lim_{\rho \to 0} \; R(r, \rho) = 1, \quad \forall r > 0. \tag{19}$$

The readability function is monotonically decreasing with respect to r and ρ:

$$0 \; > \; \frac{\partial R}{\partial r} \; = \; -\frac{0.3 \; \rho \; e^{0.3\rho r - 5}}{\left(1 + e^{0.3\rho r - 5}\right)^2}, \tag{20}$$

$$0 \; > \; \frac{\partial R}{\partial \rho} \; = \; -\frac{0.3 \; r \; e^{0.3\rho r - 5}}{\left(1 + e^{0.3\rho r - 5}\right)^2}, \tag{21}$$

$$r_1 \; < \; r_2 \iff R(r_1, \rho) > R(r_2, \rho), \tag{22}$$

$$\rho_1 \; < \; \rho_2 \iff R(r, \rho_1) > R(r, \rho_2). \tag{23}$$

Equations from (20) to (23) hold $\forall r, r_1, r_2, \rho, \rho_1, \rho_2 \in \mathbb{R}_{>0}$.

Also in this case it is interesting to study the flex point observed for a fixed value of ρ. By equating to 0 the second partial derivative of $R(\cdot)$ with respect to r we obtain that $R(r, \rho) = 0.5 \; \forall r, \rho \mid r = \frac{5}{0.3\rho}$.

3.4 ICE Function Properties

The overall ICE score is calculated by multiplying the accuracy function, the readability function and the completeness measurement. Given the properties

of the two functions and the fact that completeness may be considered as a multiplicative constant, several properties may be demonstrated for the ICE score. Information on the ICE function range may be derived:

$$0 \leq ICE(p, r, c, \varphi, \rho) < 1, \forall\ p,\ r,\ c,\ \varphi,\ \rho. \tag{24}$$

From Equations (7) to (10) and (16) to (19) the ICE asymptotic behaviour may be inferred. In particular, the ICE score tends to 0 if the knowledge predictive performance tends to $-\infty$, or if at least one amongst knowledge size, predictive performance importance or size importance tends to ∞. Formally,

$$\lim_{P(p,\varphi)\to 0} ICE(p, r, c, \varphi, \rho) = 0, \quad \forall r,\ \rho,\ c, \tag{25}$$

$$\lim_{R(r,\rho)\to 0} ICE(p, r, c, \varphi, \rho) = 0, \quad \forall p,\ \varphi,\ c. \tag{26}$$

Furthermore, it is trivial to demonstrate that

$$ICE(p, r, c, \varphi, \rho) = 0 \quad \iff \quad c = 0. \tag{27}$$

The ICE score tends to 1 if *(i)* the knowledge predictive performance also tends to 1 or the corresponding user-defined relevance tends to 0 and, at the same time, *(ii)* the knowledge size or its importance tends to 0, and *(iii)* the knowledge completeness is equal to 1. Formally,

$$ICE(p, r, c, \varphi, \rho) \simeq 1 \quad \iff \quad P(p, \varphi) \simeq 1,\ \ R(r, \rho) \simeq 1,\ \ c = 1. \tag{28}$$

ICE score values near 0 and 1 descend from the elementary properties of multiplication. Indeed *at least one* amongst accuracy, readability or completeness terms near 0 is sufficient to drag the ICE score towards 0. Conversely, *all of them* need to be near 1 to enable an ICE score close to 1. Monotonicity of the ICE function may be trivially deduced for individual projections of the involved variables, except when $c = 0$ (in this case the ICE score is always 0).

3.5 Flexibility of the ICE Score

ICE has been designed to satisfy flexibility requirements for the fidelity/readability trade-off, that may not be sufficiently handled by the existing Q_s and FiRe scores, despite their dedicated customisation parameters. For instance, let us consider the comparison of a knowledge described by 4 rules covering the whole input space having accuracy score = 0.95 and another exhaustive knowledge with only one rule having accuracy = 0.75. Both alternatives are equivalent if considering completeness, whereas the former maximises the predictive performance and the latter maximises the human-readability extent. Depending on the specific application (e.g., if human readability is more or less important than the predictive performance), end-users may prefer one knowledge or the other. A flexible metric to evaluate knowledge quality should allow users to tune the importance of the underlying quality indices to reflect this necessity. However,

the knowledge with 4 rules is the best according to the Q_s score, without any provision for human intervention to alter the rating based on fidelity/readability trade-off preferences. Furthermore, it is trivial to demonstrate analytically that with the FiRe metric it is not possible to find a value for the trade-off parameter such that the single-ruled knowledge is considered better than the other. Conversely, the ICE importance parameters for the predictive performance and readability may be set equal to 0.5 and 2, respectively, to privilege the knowledge with fewer rules, or equal to 2 and 0.5, respectively, to privilege the one with highest predictive performance. ICE is thus more flexible than FiRe, which is not capable of satisfying all possible users' needs.

4 Experiments

The effectiveness of the ICE scoring function in evaluating symbolic knowledge quality has been assessed through the comparison of the outputs provided by different SKE algorithms based on clustering, trees and/or hypercubes [21, 29, 31]: ExACT [22], CREAM [23], Iter [12], GridEx [32] and CART [4]. We relied on the ML models and SKE techniques implemented within the scikit-learn and PSyKE[3] Python libraries [5, 16, 25, 30]. The ICE score is thus applied to give a quantitative assessment of the outputs provided by this pool of extractors. Other analogous metrics (i.e., the FiRe and Q_s scores) have also been applied to the same outputs as benchmarks.

Experiments are carried out on the Wine [9] and Wisconsin breast cancer (WBC) [38] classification data sets. SKE algorithms have been applied to unbounded decision trees (DTs) previously trained on them. For each combination of data set and extraction technique, the corresponding output symbolic knowledge has been evaluated based on its raw predictive performance via the F_1 score, its size and its completeness. Given that Iter and GridEx provide knowledge in the form of a logic rule list, the knowledge size corresponds to the number of rules. ExACT, CREAM and CART, conversely, provide tree-structured symbolic knowledge and therefore the size corresponds to the number of leaf nodes. Completeness has been calculated as the percentage of the input feature space volume covered by the extracted rules with respect to the whole volume. Tree-based knowledge has always completeness = 1.

Since the Q_s and FiRe metrics require knowledge quality indices to be expressed as losses, we calculated the predictive loss as $1 - F_1$, the readability loss as the knowledge size and the completeness loss as $2 - \mathit{completeness}$, as suggested in [27]. We recall here that completeness ranges in [0, 1] and therefore a loss calculated as $1 - \mathit{completeness}$ is not suitable for multiplicative quality evaluations, as the case of Q_s, since exhaustive completeness (i.e., a loss equal to 0) would zero the score regardless of the predictive and readability loss values.

Q_s does not require user-defined parameters, so it is applied to the loss measurements without customisations. For the fidelity-readability trade-off parameter (ψ) required by the FiRe score, we selected $\psi = 1$ and $\psi = 3$ to test the

[3] https://github.com/psykei/psyke-python

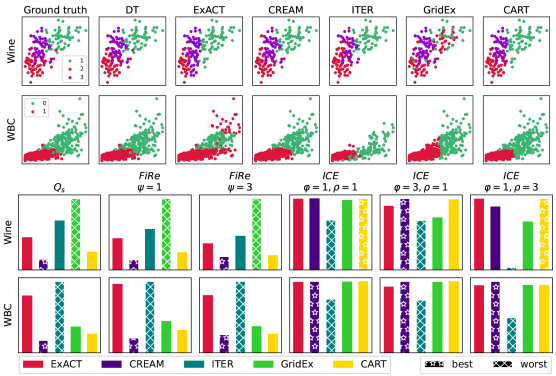

Fig. 4: Experiment results.

Table 1: Experimental results (EX = ExACT, CR = CREAM, G = GridEx).

Data set	Wine					WBC				
Quality metric	EX	CR	Iter	G	CART	EX	CR	Iter	G	CART
F_1 score	0.80	**0.95**	0.92	*0.73*	0.89	*0.82*	**0.94**	0.89	0.92	**0.94**
Knowledge size	**3**	4	*9*	5	**3**	3	**2**	*5*	3	3
Completeness	**1.00**	**1.00**	*0.75*	**1.00**	**1.00**	**1.00**	**1.00**	*0.76*	**1.00**	**1.00**
Q_s	0.608	**0.203**	0.926	*1.340*	0.338	0.536	**0.118**	0.672	0.246	0.178
FiRe, $\psi=1$	0.643	**0.217**	0.829	*1.452*	0.357	0.566	**0.122**	*0.590*	0.260	0.188
FiRe, $\psi=3$	0.214	**0.109**	0.276	*0.581*	0.119	0.189	**0.061**	*0.236*	0.087	0.063
ICE, $\varphi=1,\rho=1$	0.966	0.970	*0.678*	0.946	**0.972**	0.968	**0.979**	*0.733*	0.974	0.975
ICE, $\varphi=3,\rho=1$	0.862	**0.964**	*0.669*	0.706	0.949	0.896	**0.972**	*0.717*	0.962	0.968
ICE, $\varphi=1,\rho=3$	0.892	0.795	*0.032*	0.607	**0.898**	0.894	**0.952**	*0.470*	0.900	0.901

metric under different conditions (equal importance for predictive and readability losses and higher importance for the predictive loss since high values of ψ tend to neglect the impact of the readability loss). Finally, for the ICE score we tested three different cases: $\varphi = 1$ and $\rho = 1$; $\varphi = 3$ and $\rho = 1$; $\varphi = 1$ and $\rho = 3$. Results of the experiments are reported in Figure 4, i.e., a comparison of the data sets' ground truth with the predictions of DTs and SKE algorithms in the top panels and the corresponding knowledge quality evaluations in the bottom ones. Star- and cross-hatched bars highlight the best and worst score values, respectively. We recall that differently from ICE, knowledge evaluated via the Q_s and FiRe metrics has good quality if associated with *small* scores. Table 1 reports for each case study the quality indices for all the adopted SKE extractors and the corresponding Q_s, FiRe and ICE scores calculated upon these indices. Corresponding index losses are not reported since they can be trivially obtained as described above. For each index and metric, the best(worst) values are highlighted in bold(italic) font.

Case Studies In the Wine Data Set case study there is no candidate knowledge having all the best quality indices at the same time: CREAM has the best F_1 score and complete coverage, but it provides one rule more than CART and ExACT. These two alternatives have comparable completeness but smaller F_1 scores, especially ExACT. However, this evaluation is somehow limiting because there may be situations where higher readability (fewer rules) is preferred over accuracy. It is worth noticing that in these scores readability (as per literature) only considers the size of the knowledge, but it could include more advanced evaluations based on its human interpretability. The FiRe and Q_s scores are unanimous in declaring the symbolic knowledge extracted via CREAM the one having the best quality. ICE accepts human customisation and assigns the highest quality to CART when setting $\varphi < \rho$ (knowledge readability has a predominant role with respect to predictive performance). Conversely, when $\varphi > \rho$ the best SKE algorithm is CREAM since its output knowledge has the highest predictive performance, which in this case is the prevalent term in the ICE calculation. Finally, with $\varphi = \rho$ both terms are equally weighted, resulting in a very slight difference in the evaluation of CART and CREAM.

For the WBC data set there is an SKE technique minimising knowledge size and maximising its completeness and F_1 score at the same time. Thus, CREAM is considered the best technique according to all quality metrics. Iter is clearly the algorithm providing the worst knowledge, given that its individual quality indices are all suboptimum. Of particular interest in this case study is the comparison between ExACT and GridEx. They produce knowledge having the same completeness and size, however, GridEx has a higher F_1 score. This latter is obviously considered better than ExACT according to all quality metrics, however, the difference is slight if the quality is evaluated through ICE with $\varphi \leq \rho$, since it assigns more relevance to the knowledge size than to the predictive performance. On the contrary, the difference in quality is far more evident when assessed via ICE with $\varphi > \rho$, assigning larger weights to the F_1 score impact.

5 Conclusions

ICE is a new metric evaluating the quality of symbolic knowledge according to a set of relevant indices, such as predictive performance, human-readability extent, and completeness. It proves to be effective in carrying out automated assessments and comparisons, also enabling human tuning of the weights to be assigned to individual quality indices. For this reason, ICE results are more flexible than existing alternatives. We believe that complete scoring functions for symbolic knowledge as ICE may be effective in developing algorithmic solutions for automated tuning of parameters required by SKE procedures, therefore avoiding time-consuming manual selection performed by humans, possibly leading to suboptimum results. Future works will be devoted to a deeper investigation of interpretability including readability information about the individual knowledge items. Furthermore, we plan to study a more sound approach to avoid subjectivity in the ICE score's weight adjustment, currently still lacking unambiguity.

6 Acknowledgments

This work has been supported by PNRR – M4C2 – Investimento 1.3, Partenariato Esteso PE00000013 – "FAIR—Future Artificial Intelligence Research" – Spoke 8 "Pervasive AI", funded by the European Commission under the NextGenerationEU programme and by the European Union's Horizon Europe AEQUITAS research and innovation programme under grant number 101070363.

References

1. Augasta, M.G., Kathirvalavakumar, T.: Reverse engineering the neural networks for rule extraction in classification problems. Neural Process. Lett. **35**(2), 131–150 (2012). https://doi.org/10.1007/s11063-011-9207-8, https://doi.org/10.1007/s11063-011-9207-8
2. Baesens, B., Setiono, R., Mues, C., Vanthienen, J.: Using neural network rule extraction and decision tables for credit-risk evaluation. Management Science **49**(3), 312–329 (2003). https://doi.org/10.1287/mnsc.49.3.312.12739
3. Barakat, N., Diederich, J.: Eclectic rule-extraction from support vector machines. International Journal of Computer and Information Engineering **2**(5), 1672–1675 (2008). https://doi.org/10.5281/zenodo.1055511
4. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. CRC Press (1984)
5. Calegari, R., Sabbatini, F.: The PSyKE technology for trustworthy artificial intelligence **13796**, 3–16 (Mar 2023). https://doi.org/10.1007/978-3-031-27181-6_1, https://doi.org/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
6. Castillo, L.A., González Muñoz, A., Pérez, R.: Including a simplicity criterion in the selection of the best rule in a genetic fuzzy learning algorithm. Fuzzy Sets Syst. **120**(2), 309–321 (2001). https://doi.org/10.1016/S0165-0114(99)00095-0, https://doi.org/10.1016/S0165-0114(99)00095-0
7. Craven, M.W., Shavlik, J.W.: Extracting tree-structured representations of trained networks. In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (eds.) Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference, pp. 24–30. The MIT Press (Jun 1996), http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf
8. Czarnowski, I., Caballero, A.M., Howlett, R.J., Jain, L.C.: Intelligent Decision Technologies 2016: Proceedings of the 8th KES International Conference on Intelligent Decision Technologies (KES-IDT 2016)–Part I, vol. 56. Springer (2016)
9. Forina, M., Leardi, R., Armanino, C., Lanteri, S., Conti, P., Princi, P.: Parvus: An extendable package of programs for data exploration, classification and correlation. Journal of Chemometrics **4**(2), 191–193 (1988)
10. d'Avila Garcez, A.S., Broda, K., Gabbay, D.M.: Symbolic knowledge extraction from trained neural networks: A sound approach. Artificial Intelligence **125**(1-2), 155–207 (2001)
11. Hofmann, A., Schmitz, C., Sick, B.: Rule extraction from neural networks for intrusion detection in computer networks. In: 2003 IEEE International Conference on Systems, Man and Cybernetics. vol. 2, pp. 1259–1265. IEEE (2003). https://doi.org/10.1109/ICSMC.2003.1244584

12. Huysmans, J., Baesens, B., Vanthienen, J.: ITER: An algorithm for predictive regression rule extraction. In: Data Warehousing and Knowledge Discovery (DaWaK 2006). pp. 270–279. Springer (2006). https://doi.org/10.1007/11823728_26
13. Karmaker, S.K., Hassan, M.M., Smith, M.J., Xu, L., Zhai, C., Veeramachaneni, K.: Automl to date and beyond: Challenges and opportunities. ACM Computing Surveys (CSUR) **54**(8), 1–36 (2021)
14. Kenny, E.M., Ford, C., Quinn, M., Keane, M.T.: Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. Artificial Intelligence **294**, 103459 (2021). https://doi.org/10.1016/j.artint.2021.103459
15. Lipton, Z.C.: The mythos of model interpretability. Queue **16**(3), 31–57 (Jun 2018). https://doi.org/10.1145/3236386.3241340
16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research (JMLR) **12**, 2825–2830 (2011), https://dl.acm.org/doi/10.5555/1953048.2078195
17. Puiutta, E., Veith, E.M.: Explainable reinforcement learning: A survey. In: International cross-domain conference for machine learning and knowledge extraction. pp. 77–95. Springer (2020)
18. Rocha, A., Papa, J.P., Meira, L.A.A.: How far do we get using machine learning black-boxes? International Journal of Pattern Recognition and Artificial Intelligence **26**(02), 1261001–(1–23) (2012). https://doi.org/10.1142/S0218001412610010
19. Saad, E.W., Wunsch II, D.C.: Neural network explanation using inversion. Neural Networks **20**(1), 78–93 (2007). https://doi.org/10.1016/j.neunet.2006.07.005, https://doi.org/10.1016/j.neunet.2006.07.005
20. Sabbatini, F., Calegari, R.: Achieving complete coverage with hypercube-based symbolic knowledge-extraction techniques. In: Nowaczyk, S., Biecek, P., Chung, N.C., Vallati, M., Skruch, P., Jaworek-Korjakowska, J., Parkinson, S., Nikitas, A., Atzmüller, M., Kliegr, T., et al. (eds.) Artificial Intelligence. ECAI 2023 International Workshops – XAI^3, TACTIFUL, XI-ML, SEDAMI, RAAIT, AI4S, HYDRA, AI4AI, Kraków, Poland, September 30 – October 4, 2023, Proceedings, Part I. Communications in Computer and Information Science, vol. 1947, pp. 179–197. Springer (2023). https://doi.org/10.1007/978-3-031-50396-2_10, https://doi.org/10.1007/978-3-031-50396-2_10
21. Sabbatini, F., Calegari, R.: Bottom-up and top-down workflows for hypercube- and clustering-based knowledge extractors. In: Calvaresi, D., Najjar, A., Omicini, A., Aydogan, R., Carli, R., Ciatto, G., Främling, K. (eds.) Explainable and Transparent AI and Multi-Agent Systems. Fifth International Workshop, EXTRAAMAS 2023, London, UK, May 29, 2023, Revised Selected Papers. LNCS, vol. 14127, pp. 116–129. Springer Cham, Basel, Switzerland (2023). https://doi.org/10.1007/978-3-031-40878-6_7
22. Sabbatini, F., Calegari, R.: ExACT explainable clustering: Unravelling the intricacies of cluster formation. In: Baker, C.K., Gómez Álvarez, L., Heyninck, J., Meyer, T., Peñaloza, R., Vesic, S. (eds.) Joint Proceedings of the 2nd Workshop on Knowledge Diversity and the 2nd Workshop on Cognitive Aspects of Knowledge Representation co-located with 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023), Rhodes, Greece, September 3–4, 2023. CEUR Workshop Proceedings, vol. 3548. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3548/paper3.pdf

23. Sabbatini, F., Calegari, R.: Explainable clustering with CREAM. In: Marquis, P., Tran, C.S., Kern-Isberner, G. (eds.) 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023). pp. 593–603. IJCAI Organization, Rhodes, Greece (September 2–8 2023). https://doi.org/10.24963/kr.2023/58
24. Sabbatini, F., Calegari, R.: Symbolic knowledge-extraction evaluation metrics: The FiRe score. In: Gal, K., Nowé, A., Nalepa, G.J., Fairstein, R., Rădulescu, R. (eds.) Proceedings of the 26th European Conference on Artificial Intelligence, ECAI 2023, Kraków, Poland. September 30 – October 4, 2023 (2023). https://doi.org/10.3233/FAIA230496, https://ebooks.iospress.nl/doi/10.3233/FAIA230496
25. Sabbatini, F., Calegari, R.: Unlocking insights and trust: The value of explainable clustering algorithms for cognitive agents. In: Falcone, R., Castelfranchi, C., Sapienza, A., Cantucci, F. (eds.) Proceedings of the 24th Workshop "From Objects to Agents", Roma, Italy, November 6–8, 2023. CEUR Workshop Proceedings, vol. 3579, pp. 232–245. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3579/paper18.pdf
26. Sabbatini, F., Calegari, R.: Unveiling opaque predictors via explainable clustering: The CReEPy algorithm. In: Boella, G., D'Asaro, F.A., Dyoub, A., Gorrieri, L., Lisi, F.A., Manganini, C., Primiero, G. (eds.) Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AI*IA 2023), Rome, Italy, November 6, 2023. CEUR Workshop Proceedings, vol. 3615, pp. 1–14. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3615/paper1.pdf
27. Sabbatini, F., Calegari, R.: On the evaluation of the symbolic knowledge extracted from black boxes. AI and Ethics **4**(1), 65–74 (January 2024). https://doi.org/https://doi.org/10.1007/s43681-023-00406-1
28. Sabbatini, F., Calegari, R.: Untying black boxes with clustering-based symbolic knowledge extraction. Intelligenza Artificiale **18**(1), 21–34 (2024). https://doi.org/10.3233/IA-240026, https://doi.org/10.3233/IA-240026
29. Sabbatini, F., Ciatto, G., Calegari, R., Omicini, A.: Hypercube-based methods for symbolic knowledge extraction: Towards a unified model. In: Ferrando, A., Mascardi, V. (eds.) WOA 2022 – 23rd Workshop "From Objects to Agents", CEUR Workshop Proceedings, vol. 3261, pp. 48–60. Sun SITE Central Europe, RWTH Aachen University (Nov 2022), http://ceur-ws.org/Vol-3261/paper4.pdf
30. Sabbatini, F., Ciatto, G., Calegari, R., Omicini, A.: Symbolic knowledge extraction from opaque ML predictors in PSyKE: Platform design & experiments. Intelligenza Artificiale **16**(1), 27–48 (2022). https://doi.org/10.3233/IA-210120, https://doi.org/10.3233/IA-210120
31. Sabbatini, F., Ciatto, G., Calegari, R., Omicini, A.: Towards a unified model for symbolic knowledge extraction with hypercube-based methods. Intelligenza Artificiale **17**(1), 63–75 (2023). https://doi.org/10.3233/IA-230001, https://doi.org/10.3233/IA-230001
32. Sabbatini, F., Ciatto, G., Omicini, A.: GridEx: An algorithm for knowledge extraction from black-box regressors. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) Explainable and Transparent AI and Multi-Agent Systems. Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers, LNCS, vol. 12688, pp. 18–38. Springer Nature, Basel, Switzerland (2021). https://doi.org/10.1007/978-3-030-82017-6_2

33. Sabbatini, F., Grimani, C., Calegari, R.: Bridging machine learning and diagnostics of the esa lisa space mission with equation discovery via explainable artificial intelligence. Advances in Space Research **74**(1), 505–517 (2024). https://doi.org/https://doi.org/10.1016/j.asr.2024.04.041, https://www.sciencedirect.com/science/article/pii/S0273117724003880
34. Saito, K., Nakano, R.: Extracting regression rules from neural networks. Neural Networks **15**(10), 1279–1288 (2002). https://doi.org/10.1016/S0893-6080(02)00089-8
35. Schmitz, G.P.J., Aldrich, C., Gouws, F.S.: ANN-DT: an algorithm for extraction of decision trees from artificial neural networks. IEEE Transactions on Neural Networks **10**(6), 1392–1401 (1999). https://doi.org/10.1109/72.809084
36. Setiono, R., Liu, H.: NeuroLinear: From neural networks to oblique decision rules. Neurocomputing **17**(1), 1–24 (1997). https://doi.org/10.1016/S0925-2312(97)00038-6, https://doi.org/10.1016/S0925-2312(97)00038-6
37. Setiono, R., Thong, J.Y.L.: An approach to generate rules from neural networks for regression problems. Eur. J. Oper. Res. **155**(1), 239–250 (2004). https://doi.org/10.1016/S0377-2217(02)00792-0, https://doi.org/10.1016/S0377-2217(02)00792-0
38. Street, W.N., Wolberg, W.H., Mangasarian, O.L.: Nuclear feature extraction for breast tumor diagnosis. In: Acharya, R.S., Goldgof, D.B. (eds.) Biomedical Image Processing and Biomedical Visualization. vol. 1905, pp. 861 – 870. International Society for Optics and Photonics, SPIE (1993). https://doi.org/10.1117/12.148698, https://doi.org/10.1117/12.148698
39. Tran, S.N., d'Avila Garcez, A.S.: Knowledge extraction from deep belief networks for images. In: IJCAI-2013 workshop on neural-symbolic learning and reasoning (2013)
40. Zhou, Z., Jiang, Y., Chen, S.: Extracting symbolic rules from trained neural network ensembles. AI Commun. **16**(1), 3–15 (2003), http://content.iospress.com/articles/ai-communications/aic272