Symbolic Knowledge-Extraction Evaluation Metrics: The FiRe Score

Anonymous submission

Abstract

Symbolic knowledge-extraction techniques are becoming of key importance for AI applications since they enable the explanation of opaque black-box predictors, enhancing trust and transparency. Existing techniques use to require tuning of hyper-parameters. Manual tuning is too time-expensive for users, that conversely could benefit from automatic procedures to perform the task. However, automatic procedures can compare different extraction algorithms only if an adequate metric – such as a scoring function resuming all the interesting features of the extractors – is provided. The definition of an evaluation metric for symbolic knowledge extractors is currently missing in the literature.

Accordingly, in this paper we introduce the FiRe score metric to assess the quality of a symbolic knowledge-extraction procedure, taking into account both its predictive performance and the readability of the extracted knowledge. A rigorous mathematical formulation is provided along with several practical examples to highlight its effectiveness to the end of being exploited inside automatic hyper-parameter tuning procedures.

Introduction

One of the main strengths of machine learning (ML) models is their ability to provide exceptionally accurate predictions when applied in (roughly) every conceivable scenario (Rocha, Papa, and Meira 2012). Unfortunately, the most powerful ML predictors (as deep neural networks, for instance) present a high price in terms of human-interpretability of their outputs. Indeed, they acquire knowledge during a training phase and store it in a sub-symbolic way, in the form of internal parameters. This common *opaque* behaviour constitutes a real barrier to the exploitation of such models, named *black boxes* (BBs), in critical areas, that are all those real-world applications heavily impacting human lives, e.g., in terms of safety, health and finance.

Different solutions have been proposed by the explainable artificial intelligence community to combine human interpretability with the predictive performance of BB models (Guidotti et al. 2018). Amongst the strategies available in the literature there is the choice of intrinsic explainable models (Rudin 2019), such as decision trees with a limited amount of internal nodes and leaves. When this option is not feasible or does not provide satisfying results, a different research branch suggests extracting the BB acquired knowledge by adhering to some *symbolic* representation, through a reverse-engineering of the BB behaviour (Kenny et al. 2021). This second strategy is the rationale behind symbolic knowledge-extraction (SKE) procedures.

In the years, a plethora of SKE techniques have been proposed in the literature, especially to tackle supervised classification tasks. Given the amount of available analogous algorithms applicable to the same tasks, it may be complex to find the most suitable. Furthermore, some procedures need the fine tuning of a set of hyper-parameters, usually requiring time and skills to be performed by users.

Comparisons between different instances of the same extractor, or different extractors, are usually carried out by observing *(i)* the predictive performance of the extractor, w.r.t. both the underlying BB predictions and the actual data set output variables; and *(ii)* the readability of the output human-intelligible knowledge.

The former can be easily assessed via the same metrics adopted to measure the predictive performance of the underlying BB (e.g., F_1 and accuracy scores for classification tasks and mean absolute/squared error and R^2 score for regression tasks). Conversely, the latter may be measured through different indicators, however, to the best of our knowledge, a widely-acknowledged, well-founded and sound definition has not been yet formulated. Comparisons performed by users, as well as automated algorithmic comparisons, can surely benefit from a unified scoring function encompassing both concepts of predictive performance and readability associated with the knowledge provided by SKE techniques. Accordingly, in this paper we propose the FiRe score as a compact and expressive metric to evaluate and compare different knowledge extractors, also in association with automated parameter tuning procedures.

Background and Related Works

SKE techniques have been applied in a wide variety of areas (Steiner et al. 2006; Hayashi, Setiono, and Yoshida 2000; Sabbatini and Grimani 2022). The process of knowledge extraction from opaque ML models can follow different paradigms, i.e., SKE algorithms can be either decompositional or pedagogical (Andrews, Diederich, and Tickle 1995). In both cases, the output knowledge is generally rep-

resented as a set of logic rules, sometimes translated into natural language. Decompositional techniques take into account the internal structure of the BB, that in turn is related to the BB nature. For instance, decompositional algorithms suitable for artificial neural networks cannot be applied to support-vector machines. Pedagogical techniques, on the other hand, are more general, since they are applicable to any kind of opaque model. They only consider the output response corresponding to a given input instance or set of instances, with the aim of creating a mimicking, human-interpretable model able to approximate the underlying model predictions in the most adherent way.

The predictive performance of the extracted knowledge may be assessed w.r.t. 2 different dimensions by using the same scoring function adopted for the underlying BB. These dimensions are: *(i)* the mimicking capabilities w.r.t. the underlying model predictions, usually called *fidelity*; and *(ii)* the predictive performance w.r.t. the data set output features.

The quality of the extracted knowledge, however, must consider also the actual readability from a human perspective, that in turn depends on a set of indicators. In particular, one should take into account: *(i)* the shape of the extracted knowledge, e.g., list or trees of rules, decision tables, etc.; and *(ii)* the readability of single atoms composing the knowledge, e.g., how individual rules or tree nodes and leaves are constructed. Both items may require further investigations to discern between more and less readable knowledge representations.

It is worthwhile to notice that, generally, rule lists and trees are equivalent, since starting from a tree it is possible to build a list of rules by converting each path from the tree root to a different leaf into a distinct rule. Analogously, decision tables usually represent a rule per row (or column), so they can be easily translated into a list of rules.

There exist remarkable differences, however, between *ordered* and *unordered* lists, since in the first case rules are evaluated from the top to the bottom and for this reason bottom rules may be simplified by assuming as trivially false the conditions represented in the top rules. This leads to more concise knowledge representations, but at the cost of human-readability, since to interpret the output of the bottommost rule it is necessary to acknowledge as false all the others. As a consequence, it is evident how the mere amount of output rules is a rough but reliable indicator of readability, but the readability of single rules presents tricky challenges that should be carefully investigated.

Furthermore, logic rules are usually implications having a set of preconditions (e.g., describing an input feature space subregion) and a postcondition (i.e., the output associated to inputs belonging to the subregion). Preconditions and postconditions present different degrees of intrinsic readability. In particular, examples of preconditions are conjunctions/disjunctions of *if-then*, *M-of-N*, oblique or fuzzy rules, in positive or negative form. Also in this case the more compact representations are denoted by a smaller human-readability and it is a difficult task to assign a numerical readability score to them. For instance, a trivial count of conditions and/or constants and/or variables contained in the rules is not suitable to assess readability, and this is easily demonstrated by considering the following example. A precondition expressed as

$$if(X \geq 0.5) \wedge (X \leq 0.75)$$

contains one variable, two conditions and two constants. An equivalent fuzzy rule in the form

$$if\ X\ is\ medium$$

has one variable as well, but only one condition and one constant. However, it presents the same degree of readability for humans, that need to know the semantics of $medium$. Finally, the same concept may be modelled as

$$if\ X \in [0.5, 0.75],$$

differing from the first representation in the number of conditions (1 vs. 2). But, actually, the readability extent is absolutely equivalent.

On the other hand, as a final consideration, postconditions may be constant values or some kind of function involving input features. In the latter case, readability is obviously reduced.

Given all these observations, the scoring function presented in the following only relies on the amount of extracted rules as readability indicator, since it appears to be the most straightforward and unambiguous, even thought there would actually be a number of additional parameters to be taken into consideration.

The FiRe score

The FiRe score is a multivariate function defined as follows:

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

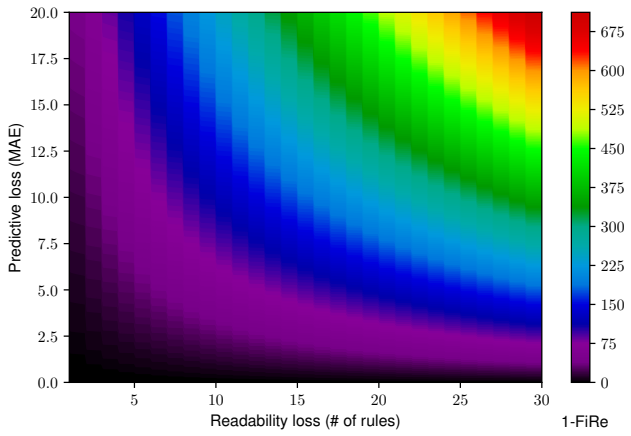
$$FiRe(\psi, p, r) = p \left\lceil \frac{r}{\psi} \right\rceil r^{0.05}, \tag{2}$$

where ψ, p and r are the fidelity/readability trade-off extent, a measure of the predictive loss of the extractor and a measure of its readability loss, respectively.

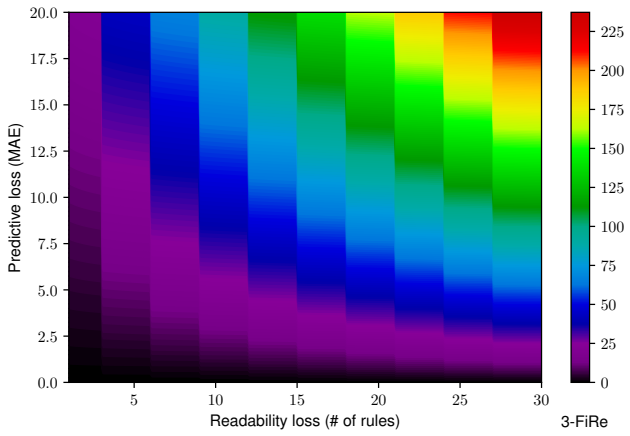
A good measure for the predictive loss p in regression tasks is the mean absolute error (MAE) of the extractor's predictions w.r.t. the underlying BB predictions or the data set outputs, depending on the need. For classification tasks, it is possible to use metrics anti-correlated with the accuracy score, e.g., $1 - accuracy$. The presented FiRe score adopts the mean absolute error metrics for regression, but it may be substituted with the mean squared error without substantial differences since both of them are generally correlated. Analogously, metrics inversely proportional to the R^2 value may be also exploited.

As for the readability loss r, the total amount of output rules is a suitable metric and thus it is the one adopted to calculate the FiRe score. More complex options will be evaluated in the future, e.g., taking into account the complexity of individual rules.

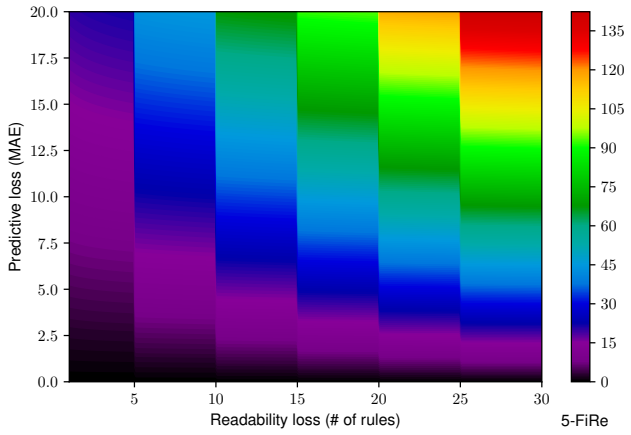
Finally, the ψ parameter describes how much the predictive loss is penalised w.r.t. the readability loss. It is important because, depending on the task at hand, the two losses may



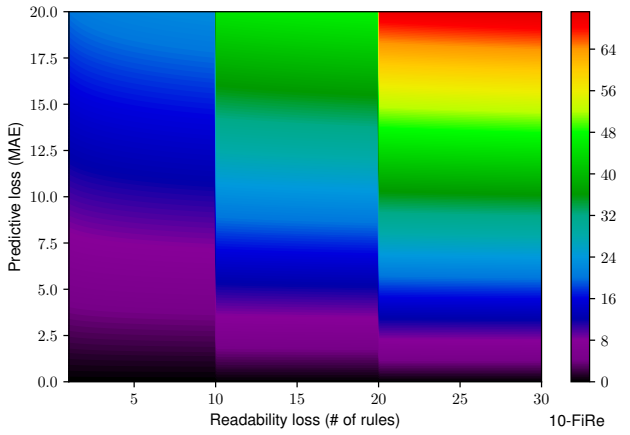
(a) 1-FiRe.



(b) 3-FiRe.



(c) 5-FiRe.



(d) 10-FiRe.

Figure 1: Graphs of different ψ-FiRe scoring functions, having $\psi \in \{1, 3, 5, 10\}$.

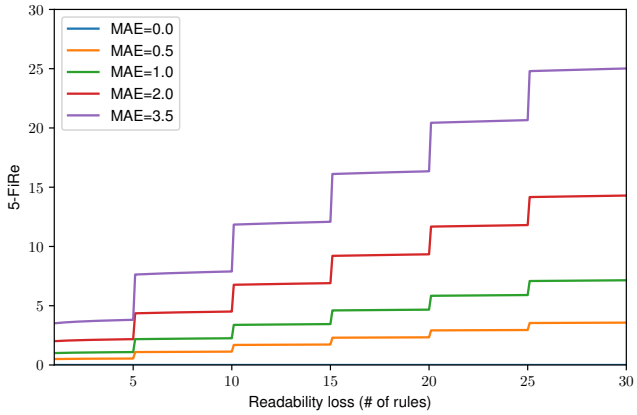
have different weights. In particular, $\psi = 1$ assigns the same importance to both losses. Growing ψ values tend to neglect the readability loss impact. In other words, given the aforementioned readability loss formulation and by assuming to have extracted m rules, if users set $\psi = n$ the FiRe score will consider only $\frac{m}{n}$ rules, rounded up to the nearest integer.

The domain of the function is explained by the following observations. The ψ parameter is a positive real value for design. Limiting the admissible ψ values to $\mathbb{N}_1$ may be also reasonable, but an extension to $\mathbb{R}_{>0}$ makes the FiRe score more flexible. On the other hand, the p parameter is a measurement of a predictive error, so it may be equal to 0 in the best case, or arbitrarily larger otherwise since there is no upper bound to the predictive error of a model. Finally, r is an integer number greater or equal to 1, since it represents a discrete quantity. However, the admissible values for this parameter have been extended from $\mathbb{N}_1$ to $\mathbb{R}_{\geq 1}$ for the sake of flexibility, analogously to the range for ψ. This choice enables, for instance, the FiRe score calculation for averaged sets of extractors trained with the same hyper-parameters, resulting in a more robust score. Similarly to the p parameter, there is no upper bound for r.

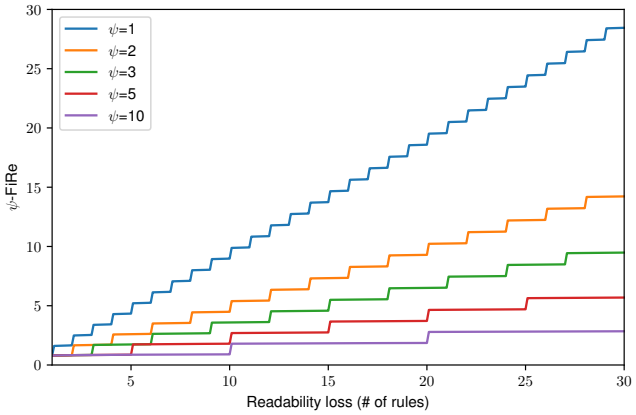
As a result of the observations above, the FiRe score is defined as a continuous (yet non-differentiable) function in the aforementioned domain and it may assume any non-negative value. Therefore, the score is a function bounded from below by 0.

In the following, we use the notation ψ-FiRe(p, r) as a clearer alias of FiRe(ψ, p, r) and we consider without loss of generality the ψ-FiRe($\cdot$) function as a bivariate function, by assuming the ψ parameter fixed *a priori*.

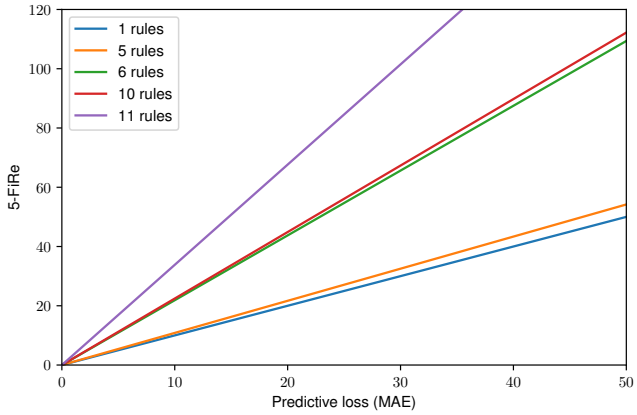
The FiRe score has been formulated to assign low scores to desirable extractors. It assumes that a good extractor should exhibit a low predictive loss and a low readability loss. For this reason, it is a multiplicative score between the two parameters. The ceiling function appearing as the second factor of the score has the purpose to give a step-function shape to the FiRe score. The exact shape of



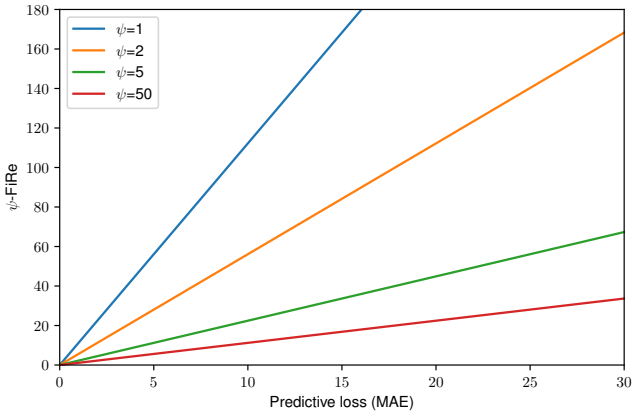
(a) Trend of 5-FiRe w.r.t. readability loss for different predictive loss values.



(b) Trends of several ψ-FiRe functions having equal predictive loss w.r.t. readability loss.



(c) Trend of 5-FiRe w.r.t. predictive loss for different readability loss values.



(d) Trends of several ψ-FiRe functions having equal readability loss w.r.t. predictive loss.

Figure 2: Projections of several ψ-FiRe functions w.r.t. different values of ψ, readability loss and predictive loss.

the steps is regulated through the ψ parameter. By setting $\psi = n$, users impose these steps to have a length equal to n. Since a flat step would assign the same ψ-FiRe score to extractors having the same predictive loss and a different but similar readability loss (e.g., r = 1 and 2, respectively, and $\psi = 10$), a third factor is queued to the score definition to discern amongst the extractors lying on the same step which one has to be considered the best. In this way the FiRe score keeps the step-function shape, but becomes an increasing monotonic function (for any $p > 0$, since $\psi\text{-}FiRe(0, r) = 0, \forall r, \forall \psi$). Examples of ψ-FiRe graphs are reported in Figure 1, for different values of ψ, p and r.

The monotonicity of the ψ-FiRe score is ensured by the following conditions:

monotonicity w.r.t. the projection of p (cf. Figure 2a)

$$r_1 < r_2 \iff \psi\text{-}FiRe(p, r_1) < \psi\text{-}FiRe(p, r_2), \quad \forall p \in \mathbb{R}_{>0}, \quad \forall r_1, r_2 \in \mathbb{R}_{\geq 1} \tag{3}$$

monotonicity w.r.t. the projection of r (cf. Figure 2c)

$$p_1 < p_2 \iff \psi\text{-}FiRe(p_1, r) < \psi\text{-}FiRe(p_2, r), \quad \forall r \in \mathbb{R}_{\geq 1}, \quad \forall p_1, p_2 \in \mathbb{R}_{>0} \tag{4}$$

Alternatively, Equations (3) and (4) can be substituted by the following condition:

monotonicity w.r.t. a partial order on the domain

$$(p_1 < p_2) \wedge (r_1 < r_2) \iff \iff \psi\text{-}FiRe(p_1, r_1) < \psi\text{-}FiRe(p_2, r_2), \quad \forall p_1, p_2 \in \mathbb{R}_{>0}, \quad \forall r_1, r_2 \in \mathbb{R}_{\geq 1} \tag{5}$$

The increasing trend of the score may be observed in Figure 2 and it is demonstrated through its partial derivatives. Equations (3) to (5) hold for any possible $\psi > 0$ that may be assigned to the ψ-FiRe score, as it is possible to notice from Figures 2b and 2d.

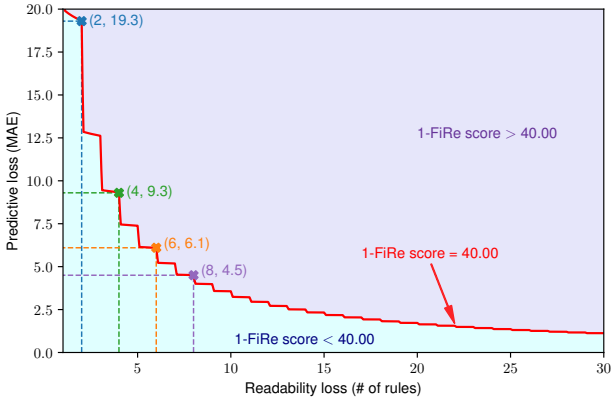


Figure 3: Graphical representation of the boundaries identified by the 1-FiRe score (isoline for 1-$FiRe = 40.0$).

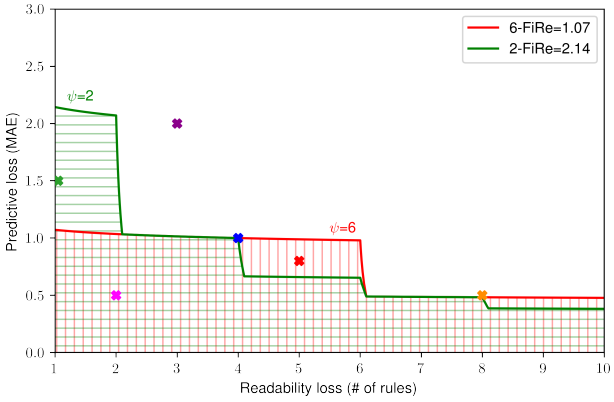


Figure 4: Different boundaries associated to the 2-FiRe and 6-FiRe scores.

The partial derivative w.r.t. p is the following:

$$\frac{\partial \psi\text{-}FiRe}{\partial p} = \left\lceil \frac{r}{\psi} \right\rceil r^{0.05} \tag{6}$$

always positive and defined in the whole domain. The partial derivative w.r.t. r is the following:

$$\frac{\partial \psi\text{-}FiRe}{\partial r} = \frac{0.05p \left\lceil \frac{r}{\psi} \right\rceil}{r^{0.95}} \tag{7}$$

always positive for $p > 0$ and defined in the whole domain except for $\frac{r}{\psi} \in \mathbb{Z}$. The derivative is 0 for $p = 0$, indeed in this case the ψ-FiRe score is always 0 regardless of the values of ψ and r.

Comparing Algorithms with FiRe

Given all the aforementioned properties about the FiRe scoring function, we exemplify here some applicative scenarios from a theoretical point of view. Let us assume to have an extraction procedure providing as output knowledge a single human-interpretable rule. The mean absolute error associated with this rule is equal to 40.0 and we chose to adopt $\psi = 1$. As a consequence, 1-$FiRe(40.0, 1) = 40.0$.

In Figure 3 the isoline corresponding to a 1-FiRe score equal to 40.0 is represented in red. Readability loss and predictive loss are reported on the x-axis and y-axis, respectively, as the number of extracted rules and mean absolute error. The extractor under study, with 1 rule and a predictive error equal to 40.0, lies on the red isoline. The same condition holds for all extractors having the same 1-FiRe score value. This is the case, for example, of extractors providing 2, 4, 6 and 8 rules associated with MAE of 19.3, 9.3, 6.1 and 4.5, respectively. All these models are considered *equivalent* on the basis of the 1-fire score. Conversely, a model able to extract 4 rules with a predictive error of 5.0 is considered *better*, since it has a smaller 1-FiRe score and thus it lies in the graph under the red isoline. More precisely, 1-$FiRe(5.0, 4) = 21.4$. On the other hand, an extractor providing 6 rules with MAE = 12.0 is considered *worse*, since its 1-FiRe score is greater than 40.0 and thus it graphically lies above the red isoline. Indeed, 1-$FiRe(12.0, 6) = 78.7$.

Figure 3 clearly highlights how the FiRe score identifies an exact boundary separating, w.r.t. a given extractor, the sets of equivalent, worse and better extractors. Furthermore, the isoline depicts the fidelity/readability trade-off correlated to $\psi = 1$. By observing the red isoline it is noticeable how a doubling of the readability loss (e.g., from 2 to 4) is accepted only if it is approximately balanced with a halving of the predictive loss. The curve can be also read in the opposite sense, e.g., a doubling of the predictive loss is accepted only when (approximately) compensated by a readability loss halving.

Finally, we exploit the same Figure to stress the fact that the isoline presents an asymptotic trend when the p and r parameters tend to infinity. This behaviour reflects the actual quality of the knowledge provided by SKE techniques. Indeed, when the number of rules or the predictive error are very high, the evaluated knowledge has low quality and it is no more a relevant task to have a fine-grained measure of *how* a loss should be compensated by the other.

How to Exploit the ψ Parameter

It is of fundamental importance to carefully choose the ψ parameter of FiRe. Figure 2b already shown that larger values of ψ reduce the impact of the readability loss. However, it is important to know that different ψ values may lead to *opposite* results when applied to compare the same extractors. This peculiarity is depicted in Figure 4, representing the separating boundaries identified by the isolines obtained via the 2-FiRe and 6-FiRe scores w.r.t. a given extractor described in the following. The boundaries associated with the two scoring functions are represented as green and red isolines, respectively. The hatched area below each isoline highlights the parameter space region denoting "more desirable" extractors, providing knowledge with better quality w.r.t. extractors lying on the isoline.

Let us assume to have an extractor able to obtain 4 rules from a BB model with a mean absolute error equal to 1.0

Algorithm	Accuracy, p	r	ψ-FiRe		
			$\psi = 1$	$\psi = 2$	$\psi = 3$
9-NN	0.97, -	-	-	-	-
CART	0.95, 0.05	3	0.17	0.11	0.06
ITER	0.89, 0.11	3	0.36	0.24	0.12
OMITTED (4 feat.)	0.93, 0.07	3	0.22	0.14	0.07
OMITTED (2 feat.)	0.91, 0.09	3	0.27	0.18	0.09
GridEx	0.94, 0.06	3	0.18	0.12	0.06
GridEx	0.81, 0.19	4	0.80	0.40	0.40
GridEx	0.97, 0.03	8	0.26	0.13	0.11

Table 1: Quality assessments for the knowledge extracted by different SKE algorithms from a 9-NN for the Iris data set.

(blue cross in the Figure). The ψ-FiRe scores associated to this model are:

$$2\text{-}FiRe(1.0, 4) = 2.14,$$

$$6\text{-}FiRe(1.0, 4) = 1.07.$$

A SKE algorithm extracting 8 rules with MAE = 0.5 (orange cross) has the same FiRe scores for both values of ψ. The models are thus equivalent according to both of the considered scoring functions. Analogously, by assuming two extractors providing 2 and 3 output rules with predictive errors equal to 0.5 and 2.0, respectively (fuchsia and purple cross in the Figure), both scores are unanimous in evaluating the former as a better extraction procedure and in considering worse the latter.

Different behaviours can be observed, for instance, by selecting an extracted knowledge composed of a single rule with MAE = 1.5 (green cross). In this case, the scores are evaluated as follows:

$$2\text{-}FiRe(1.5, 1) = 1.5 < 2.14 = 2\text{-}FiRe(1.0, 4),$$

$$6\text{-}FiRe(1.5, 1) = 1.5 > 1.07 = 6\text{-}FiRe(1.0, 4),$$

and their interpretation leads to opposite conclusions. In particular, the single-ruled knowledge is considered better than the others lying on the isoline if considering $\psi = 2$. On the other hand, it is worse if considering $\psi = 6$.

The dual situation can be encountered with knowledge having 5 rules and MAE = 0.8 (red cross). In this case the knowledge quality is considered better when evaluated through the 6-FiRe score and worse with the 2-FiRe score.

Given all these remarks, we suggest selecting the most adequate value of the ψ parameter for the task at hand after having observed the corresponding isolines.

Experiments

The effectiveness of the FiRe score to evaluate and compare the quality of SKE techniques' extracted knowledge has been assessed by running several experiments. In particular, the omitted framework (Anonymous 2021, Anonymous 2022) has been used to train a BB predictor and a set of extractors on the well-known Iris dataset[1] (Fisher 1936).The adopted extractors are the following: CART (Breiman et al. 1984), ITER (Huysmans, Baesens, and Vanthienen 2006), OMITTED (Anonymous 2022) and GridEx (Sabbatini, Ciatto, and Omicini 2021). All these techniques have been applied to a k-nearest neighbour (k-NN) classifier, having $k = 9$. Since they all are pedagogical algorithms, the provided output knowledge has been extracted only by observing the input/output response of the 9-NN.

CART induces a decision tree classifier on the 9-NN predictions and it has been executed with the default parameters.

On the other hand, ITER, OMITTED and GridEx produce a hypercubic partitioning of the input feature space according to different strategies. ITER creates and expands cubes in a bottom-up iterative fashion. It relies on 4 hyper-parameters: *(i)* the number of starting cubes, set to 1; *(ii)* the minimum amount of instances to consider inside each cube, set to 75; *(iii)* the size of cube updates, set to 7% of each input feature range interval; *(iv)* the maximum number of iterations to be performed, set to 600.

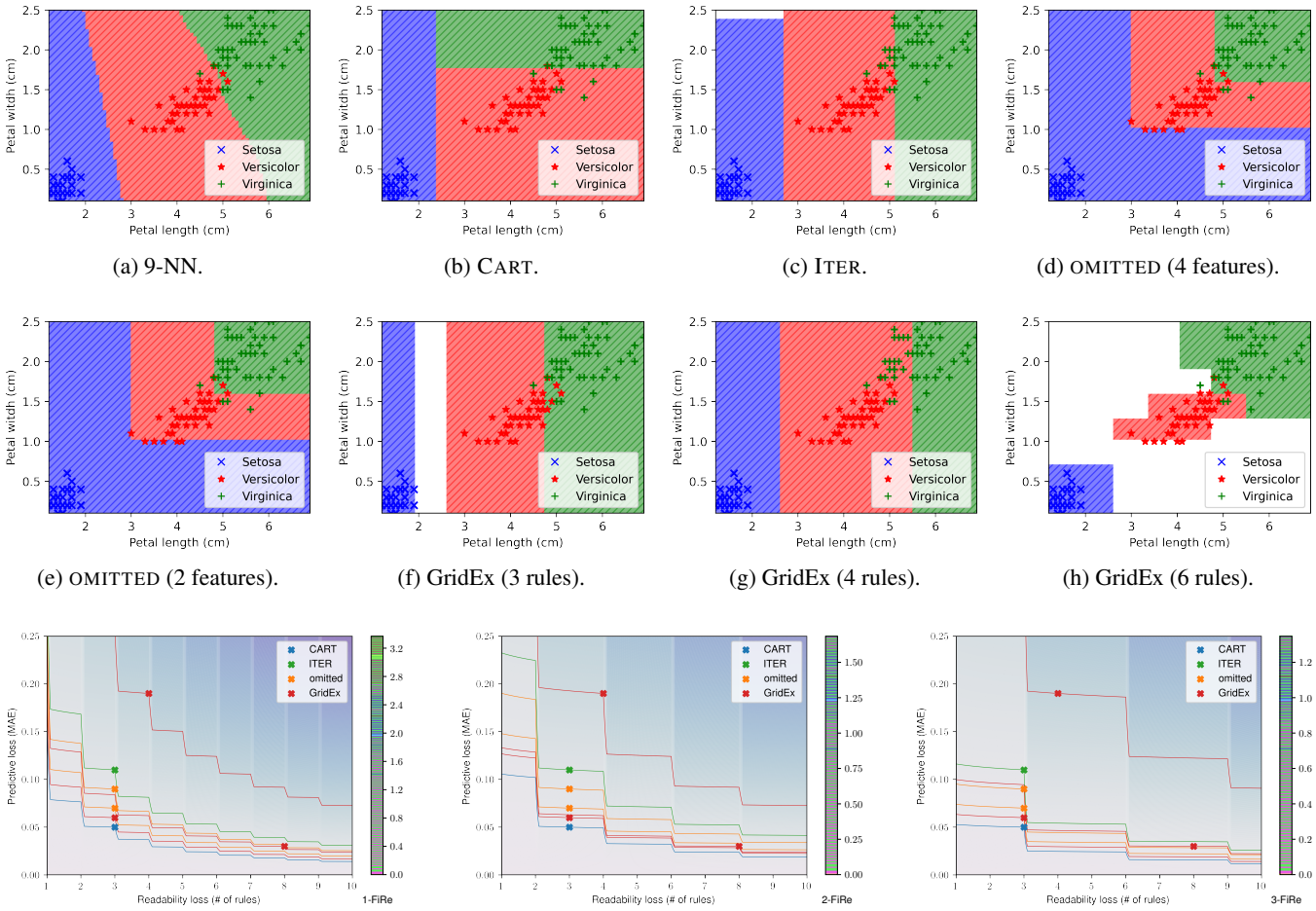
GridEx partitions the input feature space in a top-down recursive and symmetric manner, starting from the whole space. It relies on 4 hyper-parameters: *(i)* the maximum depth of the recursive splitting; *(ii)* the minimum amount of instances to consider inside each cube, set to 1; *(iii)* the number of slices to perform at each iteration; *(iv)* the error threshold to decide if a hypercubic region should be further partitioned, set to 0.1. An error threshold equal to 0.1 means that all cubes having an accuracy smaller than 0.9 are further split. The number of slices to perform has been adaptively chosen. In particular, our experiments, resumed in Table 1, consider 3 GridEx instances. The first and the second perform 8 and 2 slices, respectively, only along the most relevant input feature. The third performs 4 slices on the 2 most relevant input dimensions. As for the maximum depth, the first instance has a value equal to 1, the others equal to 2.

OMITTED adopts an underlying clustering technique to divide the input space into hypercubic hierarchical regions. We set equal to 2 the maximum depth parameter and equal to 0.1 the error threshold, which has the same semantics than that of GridEx. For our experiments we trained 2 OMITTED instances, one considering all the 4 input features and the other only the 2 most relevant.

The classification accuracy of each extractor, as well as that of the 9-NN, has been reported in Table 1. The Table also shows the number of extracted rules, representing the readability loss r of the extractors. Analogously, the predictive loss p is reported as $1 - accuracy$. Finally, the last three columns report the ψ-FiRe scores associated with each extractor for different values of ψ. Data has been gathered on single executions since the goal of this Section is to highlight how to exploit the FiRe score to carry out comparisons.

A graphical representation of the decision boundaries given by the 9-NN and the extractors are reported in Figure 5. The bottom row of the Figure reports the isolines for

[1] https://archive.ics.uci.edu/ml/datasets/iris



(a) 9-NN. (b) CART. (c) ITER. (d) OMITTED (4 features).

(e) OMITTED (2 features). (f) GridEx (3 rules). (g) GridEx (4 rules). (h) GridEx (6 rules).

(i) 1-FiRe. (j) 2-FiRe. (k) 3-fire.

Figure 5: Decision boundaries for the Iris data set obtained with different extractors applied to a 9-NN and corresponding 1-FiRe, 2-FiRe and 3-FiRe score isolines.

the ψ-FiRe scores adopted in the experiments. It is important to focus on the fact that it is not possible to obtain less than 3 rules, since the Iris data set describes 3 output classes and in the best case an extracted knowledge contains a rule per distinct class.

From Table 1 and Figure 5 it is evident that CART is the SKE algorithm providing the best output knowledge in terms of both readability and predictive performance. Indeed, it has the lowest ψ-FiRe score regardless of the adopted ψ. This result is true and acceptable since CART is the algorithm providing the smallest amount of rules with the smallest predictive loss.

Different conclusions may be drawn by comparing the GridEx instance providing 8 output rules and the 2 OMITTED instances. Indeed, by observing the corresponding ψ-FiRe scores and the equivalent isolines in the Figure, GridEx may be considered better than one, both or none of the OMITTED instances when adopting $\psi = 1$, 2 or 3, respectively.

Conclusions

In this paper we present FiRe, a scoring function to evaluate and compare SKE algorithms. More precisely, it is a compact score encompassing both a readability assessment and a predictive performance evaluation and it may be exploited to help users choosing the best extraction procedure w.r.t. a specific fidelity/readability trade-off, expressed as a parameter. The FiRe score may also be applied together with automatic parameter-tuning procedures. We showed the properties of the scoring function and a rigorous mathematical formulation has been provided.

Our future works will be focused on the enhancement of the FiRe score concerning its readability loss parameter, with a more expressive formulation than the mere amount of rules provided as output by an extractor. Furthermore, we plan to exploit this score inside the OMITTED procedure (Anonymous 2022) to automatically find the best parameter values for GridEx and omitted (Anonymous 2022).

References

Andrews, R.; Diederich, J.; and Tickle, A. B. 1995. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6): 373–389.

Breiman, L.; Friedman, J.; Stone, C. J.; and Olshen, R. A. 1984. *Classification and Regression Trees*. CRC Press.

Fisher, R. A. 1936. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2): 179–188.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5): 1–42.

Hayashi, Y.; Setiono, R.; and Yoshida, K. 2000. A comparison between two neural network rule extraction techniques for the diagnosis of hepatobiliary disorders. *Artificial intelligence in Medicine*, 20(3): 205–216.

Huysmans, J.; Baesens, B.; and Vanthienen, J. 2006. ITER: An Algorithm for Predictive Regression Rule Extraction. In *Data Warehousing and Knowledge Discovery (DaWaK 2006)*, 270–279. Springer.

Kenny, E. M.; Ford, C.; Quinn, M.; and Keane, M. T. 2021. 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.

Rocha, A.; Papa, J. P.; and Meira, L. A. A. 2012. How far do we get using machine learning black-boxes? *International Journal of Pattern Recognition and Artificial Intelligence*, 26(02): 1261001–(1–23).

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215.

Sabbatini, F.; Ciatto, G.; Calegari, R.; and Omicini, A. 2021. On the Design of PSyKE: A Platform for Symbolic Knowledge Extraction. In Calegari, R.; Ciatto, G.; Denti, E.; Omicini, A.; and Sartor, G., eds., *WOA 2021 – 22nd Workshop "From Objects to Agents"*, volume 2963 of *CEUR Workshop Proceedings*, 29–48. Sun SITE Central Europe, RWTH Aachen University. 22nd Workshop "From Objects to Agents" (WOA 2021), Bologna, Italy, 1–3 September 2021. Proceedings.

Sabbatini, F.; Ciatto, G.; Calegari, R.; and Omicini, A. 2022. Symbolic Knowledge Extraction from Opaque ML Predictors in PSyKE: Platform Design & Experiments. *Intelligenza Artificiale*, 16(1): 27–48.

Sabbatini, F.; Ciatto, G.; and Omicini, A. 2021. GridEx: An Algorithm for Knowledge Extraction from Black-Box Regressors. In Calvaresi, D.; Najjar, A.; Winikoff, M.; and 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*, volume 12688 of *LNCS*, 18–38. Basel, Switzerland: Springer Nature. ISBN 978-3-030-82016-9.

Sabbatini, F.; and Grimani, C. 2022. Symbolic knowledge extraction from opaque predictors applied to cosmic-ray data gathered with LISA Pathfinder. *Aeronautics and Aerospace Open Access Journal*, 6(3): 90–95.

Steiner, M. T. A.; Steiner Neto, P. J.; Soma, N. Y.; Shimizu, T.; and Nievola, J. C. 2006. Using neural network rule extraction for credit-risk evaluation. *International Journal of Computer Science and Network Security*, 6(5A): 6–16.

View Reviews

Paper ID
2349

Paper Title
Symbolic Knowledge-Extraction Evaluation Metrics: The FiRe Score

Track Name
Main Track

Reviewer #2

Questions

1. {Summary} Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
The authors design a metric, FiRe score, to evaluate the effectiveness of symbolic knowledge extraction. The score considers accuracy and readability trade-off. The authors provide some math analysis on the score design.

2. {Strengths and Weaknesses} Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: novelty, quality, clarity, and significance.
The authors discuss a relatively novel question in this paper: how to balance the fidelity-readability trade-off in symbolic knowledge extraction. However, I cannot see too much novelty for the score design.

The paper needs to improve readability. It would be better to immediately introduce the key concepts when mentioning them, like decompositional and pedagogical in the first page. The logic between paragraphs also takes much time to understand. Besides, some properties about FiRe score has better way to demonstrate, and it does not need 4 sub-figures (about half page).

The experiment and results are not sufficient. We can see the extraction with better readability and predictive performance leads to lowest FiRe score, but not on the contrary. Besides, the authors do not discuss the situation that both readability and predictive performance are not improved at the same time.

3. {Questions for the Authors} Please carefully describe questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.
1. Why there is a "0.05" in the FiRe score formula? Could you explain all the magic numbers mentioned in this paper?

2. If readability and predictive performance are not improved at the same time, is the FiRe score still effective?

4. {Evaluation: Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?
Fair: The paper contributes some new ideas or represents incremental advances.

5. {Evaluation: Quality} Is the paper technically sound?
Fair: The paper has minor technical flaws. For example, the proof of a theorem has some fixable errors or the experimental evaluation is weak.

6. {Evaluation: Significance} How do you rate the likely impact of the paper on the AI research community?
Fair: The paper is likely to have modest impact within a subfield of AI.

7. {Evaluation: Clarity} Is the paper well-organized and clearly written?
Poor: The paper is unclear and very hard to understand.

8. (Evaluation: Reproducibility) Are the results (e.g., theorems, experimental results) in the paper easily

reproducible? (It may help to consult the paper's reproducibility checklist.)checklist.)
Good: key resources (e.g., proofs, code, data) are available and sufficient details (e.g., proofs, experimental setup) are described such that an expert should be able to reproduce the main results.

9. {Evaluation: Resources} If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper's reproducibility checklist)
Fair: The shared resources are likely to be of some use to other AI researchers.

10. {Evaluation: Ethical considerations} Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?
Fair: The paper addresses some applicable ethical considerations but fails to address some important ones.

11. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.
Borderline reject: Technically solid paper where reasons to reject, e.g., poor novelty, outweigh reasons to accept, e.g. good quality. Please use sparingly.

13. (CONFIDENCE) How confident are you in your evaluation?
Somewhat confident, but there's a chance I missed some aspects. I did not carefully check some of the details, e.g., novelty, proof of a theorem, experimental design, or statistical validity of conclusions.

14. (EXPERTISE) How well does this paper align with your expertise?
Mostly Knowledgeable: This paper has little overlap with my current work. My past work was focused on related topics and I am knowledgeable or somewhat knowledgeable about most of the topics covered by the paper.

16. I acknowledge that I have read the author's rebuttal (if applicable) and made changes to my review as needed.
Agreement accepted

Reviewer #7

Questions

1. {Summary} Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
The paper provides an evaluation metric for symbolic knowledge extraction named FiRe as a metric to evaluate and compare different knowledge extractors, which defines an evaluation metric for symbolic knowledge extractors. In performing the evaluation, two main components are considered: fidelity and readability of the extracted knowledge.

2. {Strengths and Weaknesses} Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: novelty, quality, clarity, and significance.
Strengths
1. This paper proposes an important study that attempts to develop a widely accepted, well-founded, and reasonable definition and measurement of the assessment criteria of interpretable AI through different metrics.
2. When constructing the evaluation function, some key evaluation criteria were considered.

Weaknesses
1. The choice of many parameters is not explained in detail in the text, which is difficult to follow this formula.
2. Only the small Iris dataset was used for the training test, rather than training on the larger dataset. Generally, more knowledge will be extracted with larger datasets, and this part needs further explanation.

3. {Questions for the Authors} Please carefully describe questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.
1. It is mentioned in this paper that readability is related to the form of rule representation, the readability of individual atoms that constitute knowledge, and the form of rules. But the formula uses only the number of rules as a readability indicator. Why are the other parameters discarded, and how would they affect the overall assessment?
2. As described in the formula, only the number of rules is considered a readability indicator. Does the same number

of different kinds of parameters yield the same ψ value (e.g., choosing rule 1,2, or rule 3,4) and does each rule have the same expressiveness? And how would this affect the final evaluation?
3. For different classification tasks (regression and classification tasks), different prediction losses p seem to be used, how does this affect the generic evaluation criteria?

4. {Evaluation: Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?
Fair: The paper contributes some new ideas or represents incremental advances.

5. {Evaluation: Quality} Is the paper technically sound?
Good: The paper appears to be technically sound. The proofs, if applicable, appear to be correct, but I have not carefully checked the details. The experimental evaluation, if applicable, is adequate, and the results convincingly support the main claims.

6. {Evaluation: Significance} How do you rate the likely impact of the paper on the AI research community?
Fair: The paper is likely to have modest impact within a subfield of AI.

7. {Evaluation: Clarity} Is the paper well-organized and clearly written?
Good: The paper is well organized but the presentation has minor details that could be improved.

8. (Evaluation: Reproducibility) Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)checklist.)
Fair: key resources (e.g., proofs, code, data) are unavailable and/or some key details (e.g., proof sketches, experimental setup) are unavailable which make it difficult to reproduce the main results.

9. {Evaluation: Resources} If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper's reproducibility checklist)
Fair: The shared resources are likely to be of some use to other AI researchers.

10. {Evaluation: Ethical considerations} Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?
Not Applicable: The paper does not have any ethical considerations to address.

11. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.
Borderline reject: Technically solid paper where reasons to reject, e.g., poor novelty, outweigh reasons to accept, e.g. good quality. Please use sparingly.

13. (CONFIDENCE) How confident are you in your evaluation?
Somewhat confident, but there's a chance I missed some aspects. I did not carefully check some of the details, e.g., novelty, proof of a theorem, experimental design, or statistical validity of conclusions.

14. (EXPERTISE) How well does this paper align with your expertise?
Mostly Knowledgeable: This paper has little overlap with my current work. My past work was focused on related topics and I am knowledgeable or somewhat knowledgeable about most of the topics covered by the paper.

Reviewer #8

Questions

1. {Summary} Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
The study of interpretability evaluation metric is an important step in XAI. Previous papers focus more on the readability of predictors. This paper proposes a novel metric combining model predictiveness and interpretability. The authors analyze some mathematical properties and conduct experiments to show the interpretability of the metric. This novel metric is parameterized and the authors discuss the hyperparmaeter tuning as well.

2. {Strengths and Weaknesses} Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: novelty, quality, clarity, and significance.
Strengths:

1. The study of interpretability evaluation metric is an important step in XAI.
2. The authors propose a novel metric with analytical properties.

Weaknesses:
1. Writing should be improved. e.g. "Existing techniques use to require tuning of hyper-parameters."
2. The proposal of this metric form is not well motivated. Why these three variables? Why multiplicative form?
3. The advantages of the metric is not well demonstrated. What could we know in terms of interpretability with this new metric compared to previous metric?

3. {Questions for the Authors} Please carefully describe questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.
It would be much helpful if the authors could provide further evidence and justifications on the weaknesses mentioned above.

4. {Evaluation: Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?
Fair: The paper contributes some new ideas or represents incremental advances.

5. {Evaluation: Quality} Is the paper technically sound?
Fair: The paper has minor technical flaws. For example, the proof of a theorem has some fixable errors or the experimental evaluation is weak.

6. {Evaluation: Significance} How do you rate the likely impact of the paper on the AI research community?
Fair: The paper is likely to have modest impact within a subfield of AI.

7. {Evaluation: Clarity} Is the paper well-organized and clearly written?
Good: The paper is well organized but the presentation has minor details that could be improved.

8. (Evaluation: Reproducibility) Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)checklist.)
Fair: key resources (e.g., proofs, code, data) are unavailable and/or some key details (e.g., proof sketches, experimental setup) are unavailable which make it difficult to reproduce the main results.

9. {Evaluation: Resources} If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper's reproducibility checklist)
Fair: The shared resources are likely to be of some use to other AI researchers.

10. {Evaluation: Ethical considerations} Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?
Fair: The paper addresses some applicable ethical considerations but fails to address some important ones.

11. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.
Reject: For instance, a paper with poor quality, inadequate reproducibility, incompletely addressed ethical considerations.

13. (CONFIDENCE) How confident are you in your evaluation?
Very confident. I have checked all points of the paper carefully. I am certain I did not miss any aspects that could otherwise have impacted my evaluation.

14. (EXPERTISE) How well does this paper align with your expertise?
Very Knowledgeable: This paper significantly overlaps with my current work and I am very knowledgeable about most of the topics covered by the paper.

16. I acknowledge that I have read the author's rebuttal (if applicable) and made changes to my review as needed.
Agreement accepted

Reviewer #9

Questions

1. {Summary} Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
The authors introduce the FiRe score metric in this paper to evaluate the effectiveness of a symbolic knowledge-extraction technique while also considering the readability of the extracted knowledge. It can be used to assist users in selecting the appropriate extraction process for a certain fidelity/readability trade-off, stated as a parameter. To put it more accurately, it is a compact score combining both a readability assessment and a predictive performance evaluation.

2. {Strengths and Weaknesses} Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: novelty, quality, clarity, and significance.
This paper is difficult to follow. It is not clearly stated the contribution and significance. The author state they proposed the FiRe which can be used to evaluate and compare SKE algorithms, but not clearly showing what is the advantages.

3. {Questions for the Authors} Please carefully describe questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.
What are the questions or conclusions the experiments want to show? Is it only want to show the effectiveness of the newly proposed methods? What is the advantage of FiRe?

4. {Evaluation: Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?
Fair: The paper contributes some new ideas or represents incremental advances.

5. {Evaluation: Quality} Is the paper technically sound?
Fair: The paper has minor technical flaws. For example, the proof of a theorem has some fixable errors or the experimental evaluation is weak.

6. {Evaluation: Significance} How do you rate the likely impact of the paper on the AI research community?
Fair: The paper is likely to have modest impact within a subfield of AI.

7. {Evaluation: Clarity} Is the paper well-organized and clearly written?
Poor: The paper is unclear and very hard to understand.

8. (Evaluation: Reproducibility) Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)checklist.)
Fair: key resources (e.g., proofs, code, data) are unavailable and/or some key details (e.g., proof sketches, experimental setup) are unavailable which make it difficult to reproduce the main results.

9. {Evaluation: Resources} If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper's reproducibility checklist)
Fair: The shared resources are likely to be of some use to other AI researchers.

10. {Evaluation: Ethical considerations} Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?
Fair: The paper addresses some applicable ethical considerations but fails to address some important ones.

11. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.
Reject: For instance, a paper with poor quality, inadequate reproducibility, incompletely addressed ethical considerations.

13. (CONFIDENCE) How confident are you in your evaluation?
Somewhat confident, but there's a chance I missed some aspects. I did not carefully check some of the details, e.g., novelty, proof of a theorem, experimental design, or statistical validity of conclusions.

14. (EXPERTISE) How well does this paper align with your expertise?
Knowledgeable: This paper has some overlap with my current work. My recent work was focused on closely related topics and I am knowledgeable about most of the topics covered by the paper.

16. I acknowledge that I have read the author's rebuttal (if applicable) and made changes to my review as needed.

Agreement accepted

Authors comment

The paper has been completely rewritten taking into account all the comments of the reviewer. In particular, motivations have been better highlighted as well as an explanation of all the parameters contained in the definition of the metric.