On the interpretability of fuzzy cognitive maps

Nápoles, Gonzalo; Ranković, Nevena; Salgueiro, Yamisleydi

Published in:
Knowledge-Based Systems

DOI:
10.1016/j.knosys.2023.111078

Publication date:
2023

Document Version
Version created as part of publication process; publisher's layout; not normally made publicly available

Citation for published version (APA):
Nápoles, G., Ranković, N., & Salgueiro, Y. (2023). On the interpretability of fuzzy cognitive maps. *Knowledge-Based Systems*, *281*(111078). https://doi.org/10.1016/j.knosys.2023.111078

Knowledge-Based Systems 281 (2023) 111078

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

On the interpretability of Fuzzy Cognitive Maps

Gonzalo Nápoles [a,*], Nevena Ranković [a], Yamisleydi Salgueiro [b]

[a] Department of Cognitive Science & Artificial Intelligence Tilburg University, The Netherlands
[b] Department of Industrial Engineering, Faculty of Engineering, Universidad de Talca, Campus Curicó, Chile

ARTICLE INFO

Keywords:
Fuzzy Cognitive Maps
Decision making
Concept relevance
Interpretability

ABSTRACT

This paper proposes a post-hoc explanation method for computing concept attribution in Fuzzy Cognitive Map (FCM) models used for scenario analysis, based on SHapley Additive exPlanations (SHAP) values. The proposal is inspired by the lack of approaches to exploit the often-claimed intrinsic interpretability of FCM models while considering their dynamic properties. Our method uses the initial activation values of concepts as input features, while the outputs are considered as the hidden states produced by the FCM model during the recurrent reasoning process. Hence, the relevance of neural concepts is computed taking into account the model's dynamic properties and hidden states, which result from the interaction among the initial conditions, the weight matrix, the activation function, and the selected reasoning rule. The proposed post-hoc method can handle situations where the FCM model might not converge or converge to a unique fixed-point attractor where the final activation values of neural concepts are invariant. The effectiveness of the proposed approach is demonstrated through experiments conducted on real-world case studies.

1. Introduction

Fuzzy Cognitive Maps (FCMs) were introduced in [1] as a knowledge-based approach for modeling complex systems and performing what-if simulations. From a neural network perspective, FCMs are recurrent neural systems that consist of neural concepts and weighted connections [2]. Neural concepts represent variables, entities, or states related to the physical system under investigation and are interconnected by directed edges or connections assigned signed weights. The signed weight associated with each connection denotes the strength of causality or correlation between the corresponding variables. Despite their naming, FCMs do not involve any fuzzy characteristics as both their knowledge representation, reasoning process and inner states have crisp semantics [3].

FCMs have been widely recognized as being fairly interpretable, transferable, causal and transparent, thus making them appealing for decision-making processes [4,5]. Firstly, it is true that all components in the networks have a well-defined meaning for the problem domain being modeled. This is the case with neural concepts and causal relationships that are typically defined by human experts. Secondly, the system can either be visualized as a whole [6] or conveniently mined to extract useful from its knowledge structures. Finally, the transparency of these neural reasoning systems allows determining the most important concepts in the network from a static perspective where only the weights connecting the concepts are analyzed. This is often accomplished by computing centrality measures such as the incoming-weight and outgoing-weight degree measures (see Section 4 for further details).

However, analyzing the centrality of concepts as a quantitative measure of feature importance might be insufficient. On the one hand, FCMs are recurrent systems that support feedback loops and produce states with hidden patterns. While the *static* properties of FCM models are given by their weights only [7], the *dynamic* properties are defined by the weights, initial conditions, the activation function associated with neural concepts, and the reasoning rule used to update neurons' activation values. This fact challenges the traditional belief that FCM-based models provide for intrinsic interpretability. On the other hand, existing post-hoc approaches for computing feature importance such as LIME (Local Interpretable Model-Agnostic Explanations) [8] or SHAP (Shapley Additive Explanations) [9] might fail if directly applied to FCM-based models devoted to scenario analysis. The reason for this relies on the FCMs' convergence issues.

Let us assume that we apply LIME or SHAP on an FCM model such that we are concerned about the initial conditions and the outputs after performing a fixed number of iterations. If the model fails to converge, the feature importance values will change with the number of iterations defined by the modeler. In contrast, if the network converges to a unique fixed-point attractor, the outputs will be invariant w.r.t. initial conditions, thus causing the failure of these post-hoc methods since marginalizing a feature will not report any changes in the outputs.

[*] Corresponding author.
E-mail address: g.r.napoles@uvt.nl (G. Nápoles).

https://doi.org/10.1016/j.knosys.2023.111078
Received 13 June 2023; Received in revised form 1 September 2023; Accepted 13 October 2023
Available online 16 October 2023

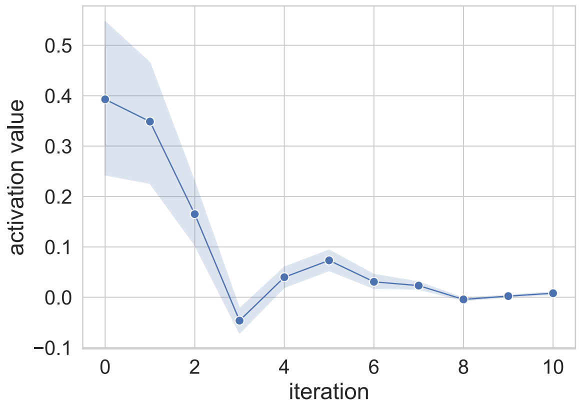


Fig. 1. Convergence of a neural concept to a unique fixed-point attractor for different initial activation vectors.

Fig. 1 depicts the issues of computing concept importance using SHAP or LIME in the presence of fixed-point attractors. This example shows the activation values of a neural concept initialized with random activation values. This neuron converges to a unique fixed point after 10 iterations. Note how variability in the neuron's activation values (visualized with the shadow) decreases until reaching the fixed-point attractor. The ramification of this behavior is that the final neurons' activation values will be the same regardless of the initial activation values used for starting the recurrent reasoning process. As a result, perturbation-based explanation methods will not be able to capture any variability in the model's outputs, thus failing to compute reliable concept importance scores.

This paper proposes a SHAP-based post-hoc method for computing concept attribution in FCM models devoted to scenario analysis. Unlike centrality measures that only use the model's static information to derive concept importance, the proposed SHAP method focuses on the model's dynamic properties. The main motivation supporting our method relies on the absence of post-hoc methods generating explanations from the more obscure yet important component of FCM models: their dynamic behavior. Therefore, such explanations will complement the intrinsic interpretability of FCM models and the explanation derived from their static information.

The proposed SHAP method uses the concepts' initial activation values as inputs and temporal states produced during the recurrent reasoning process as outputs. Hence, the relevance of neural concepts is computed considering the model's dynamic properties and hidden states, which are obtained from the interaction between the initial conditions, the weight matrix, the activation, and the reasoning rule. More importantly, the method will not fail even if the FCM model leads to chaos, cyclic patterns, or converges to a unique fixed-point attractor. This happens because the relevance of concepts is based on their hidden states rather than the final states produced by the model after performing a fixed number of iterations.

The rest of this paper is organized as follows. Section 2 introduces the theoretical foundations of cognitive mapping, while Section 3 revises existing approaches for determining concept importance in FCM-based models. Section 4 presents classic measures based on the degree centrality notion, while Section 5 describes the proposed SHAP-based post-hoc method. The case studies and the numerical comparison between centrality-based measures and the proposed SHAP method are discussed in Section 6. Towards the end, Section 7 provides some concluding remarks and future work directions to be explored.

2. Fuzzy cognitive maps

FCM models are graph-theoretic tools for modeling complex systems composed of interconnected concepts and feedback loops [10]. A pivotal component when resenting knowledge is the weight matrix $\mathbf{W}_{N\times N}$ where N represents the number of neural concepts depicted as nodes in the digraph. As such, the weight attached to the edge departing from the ith concept and arriving at the jth concept is represented as $w_{ij} \in [-1,1]$. These causal weights can be viewed as measures of the degree of influence that one concept has on another in the network.

Moreover, each neuronal concept C_i in the network is attached with an activation value $a_i^{(t)}$ denoting the extent to which that concept is active in the tth iteration. The initial activation vector $a_i^{(0)}$ is provided by the modeler when performing what-if simulations. Hence, the activation vector $\mathbf{A}^{(t)} = (a_1^{(t)}, \ldots, a_i^{(t)}, \ldots, a_N^{(t)})$ gives the state of the system in the current iteration $t \in \{1,2,\ldots,T\}$, while the initial activation vector $\mathbf{A}^{(0)} = (a_1^{(0)}, \ldots, a_i^{(0)}, \ldots, a_N^{(0)})$ encodes the scenario to be simulated.

Eq. (1) shows the generalized reasoning rule [11] used to compute the activation value of the ith neural concept in the $(t+1)$th iteration given initial condition provided by the domain expert,

$$a_i^{(t+1)} = \underbrace{\phi \cdot f\left(\sum_{j=1}^{N} a_j^{(t)} w_{ji}\right)}_{\text{nonlinear component}} + \underbrace{(1-\phi)\cdot a_i^{(0)}}_{\text{linear component}}, \quad (1)$$

such that $f(\cdot)$ is the activation function and $0 \leq \phi \leq 1$ is a parameter used to control the nonlinearity degrees of the reasoning rule. When $\phi = 1$, the model performs as a closed system where the activation value of a neuron depends on the activation values of connected neurons in the previous iteration. When $0 < \phi < 1$, we add a linear component to the reasoning rule devoted to preserving the initial activation values of neurons when updating their activation values in the current iteration. When $\phi = 0$, the model narrows down to a linear regression where the initial activation values serve as regressors. The reasoning rule of FCM models in Eq. (1) stops when either (i) the model converges to a fixed point or (ii) a maximal number of iterations T is reached. Overall, we have three possible states:

- **Fixed point** $(\exists t_\alpha \in \{1,\ldots,(T-1)\} : a_i^{(t+1)} = a_i^{(t)}, \forall i, \forall t \geq t_\alpha)$: the FCM produces the same state vector after t_α, thus $a_i^{(t_\alpha)} = a_i^{(t_\alpha+1)} = a_i^{(t_\alpha+2)} = \cdots = a_i^{(T)}$. If the fixed point is unique, the FCM model will produce invariant states (i.e., the same state vector regardless of the initial conditions).
- **Limit cycle** $(\exists t_\alpha, P, j \in \{1,\ldots,(T-1)\} : a_i^{(t+P)} = a_i^{(t)}, \forall i, \forall t \geq t_\alpha)$: the FCM produces the same state vector periodically after the period P, thus $a_i^{(t_\alpha)} = a_i^{(t_\alpha+P)} = \cdots = a_i^{(t_\alpha+jP)}$, where $t_\alpha + jP \leq T$.
- **Chaos**: the FCM produces different state vectors for successive iterations with no clear pattern.

The activation function $f(\cdot)$ deserves further discussion since it provides the FCM model with its nonlinear capabilities [12]. This bounded function transforms the raw activation values that a neuron receives during the reasoning process and outputs activation values that belong to a pre-defined interval, usually $[0,1]$ or $[-1,1]$. In this paper, we will focus on activation functions that produce values in the $[-1,1]$ intervals as these functions do not lead to misleading activation values. For example, neural concepts having no incoming connections will become active if we use a sigmoid activation function, which might affect the reliability of the system being modeled.

Eq. (2) shows a straightforward continuous activation function referred to as saturation,

$$f(\bar{a}_i^{(t+1)}) = \begin{cases} -1, & \text{if } \bar{a}_i^{(t+1)} < -1 \\ \bar{a}_i^{(t+1)}, & \text{if } -1 \leq \bar{a}_i^{(t+1)} \leq 1 \\ 1, & \text{if } \bar{a}_i^{(t+1)} > 1 \end{cases} \quad (2)$$

such that $\bar{a}_i^{(t+1)}$ is the raw activation value of the ith concept in the iteration being processed.

The hyperbolic tangent function shown in Eq. (3) is quite popular in real-world FCM applications and produces values in the $(-1, 1)$ interval,

$$f_h\left(\bar{a}_i^{(t+1)}\right) = \frac{e^{\bar{a}_i^{(t+1)}} - e^{-\bar{a}_i^{(t+1)}}}{e^{\bar{a}_i^{(t+1)}} + e^{-\bar{a}_i^{(t+1)}}}. \tag{3}$$

The continuous activation functions usually pose some challenges that might affect the reliability of simulations. On the one hand, these functions often cause the neurons to saturate towards the endpoints of the activation intervals. If that happens, the neural concept will hardly capture any variations in its raw activation values. On the other hand, these functions are closely linked to the convergence properties of the network [13–15] while providing modelers little controllability options.

To tackle these issues, Nápoles et al. [11] introduced the re-scaled activation function,

$$f_r\left(\bar{a}_i^{(t+1)}\right) = \frac{\bar{a}_i^{(t+1)}}{\left\|\bar{\mathbf{A}}^{(t+1)}\right\|_2}, \left\|\bar{A}^{(t+1)}\right\|_2 \neq 0 \tag{4}$$

where $\bar{\mathbf{A}}^{(t)} = (\bar{a}_1^{(t)}, \ldots, \bar{a}_i^{(t)}, \ldots, \bar{a}_N^{(t)})$ is the raw activation vector given by $\bar{\mathbf{A}}^{(t)} = \mathbf{A}^{(t)}\mathbf{W}$. The authors in [11] derived analytical conditions for convergence, which are related to the ϕ parameter in Eq. (1) and the eigenvalues of $\mathbf{W}$. For example, if $0 \leq \phi < 1$, then we can ensure that the fixed point will not be unique, although cycles might appear under special circumstances. In contrast, if $\phi = 1$, then the fixed point will be unique provided that $\mathbf{W}$ is diagonalizable and contains an eigenvalue that is strictly greater in magnitude than other eigenvalues, and that the initial vector must have a nonzero component in the direction of the dominant eigenvector. Moreover, if the initial vector is orthogonal to the dominant eigenvector, then the FCM model will not converge.

3. Related work

This section will delve into two primary methodologies to determine relevant concepts in FCM models - network reduction and centrality-based approaches.

3.1. Network reduction approach

Next, we provide an overview of relevant network reduction approaches in previous research to determine the most important concepts in FCM models. By understanding the methods used in previous studies, decision-makers can leverage this approach to gain insights into complex systems and make informed decisions.

The authors in [16] explain how FCMs can quickly become unwieldy and difficult to interpret as the number of concepts and relationships increases, hindering their practical usefulness. They propose a network reduction method that analyzes the relationship strength between concepts and their neighbors while removing those with weak impact. The method reduces the number of concepts while maintaining accuracy for real-world decision-making. Another study [17] presents a reduction approach based on a combination of two-step learning and conceptual reduction techniques. First, a learning method creates an initial FCM model, and then a conceptual reduction technique reduces its dimensionality. Testing on various datasets demonstrated that this approach effectively reduces the dimensionality while maintaining accuracy by adapting the FCM model to the system's requirements.

To facilitate the analysis and interpretation of models, the study [18] is devoted to building hybrid FCM-based models with learning capabilities. The contributions comprise a scheme for constructing hybrid FCMs where experts are requested to identify the interaction among the input concepts. The weights connecting the inputs and outputs are computed from data using a rapid learning rule based on the Moore–Penrose inverse. A network reduction approach discards superfluous concepts and relationships, considering the expected activation values of concepts and the absolute values of weights. Furthermore, the authors presented two calibration methods to fine-tune the model after eliminating potentially redundant weights. While one of these methods focuses on recomputing the retained weights, the other modifies the non-linear parameters associated with the sigmoid function.

Developing a model that is both precise and straightforward is often a difficult task for experts. The authors in [19] compared established techniques with novel FCM reduction methods based on k-means and fuzzy c-means clustering. The goal was to create simpler models capable of mimicking the behavior of the original model better than existing methods. By reducing the complexity of the model while maintaining its accuracy, these new methods could provide a more efficient and effective approach to modeling complex systems. Ultimately, the proposed reduction techniques could significantly contribute to advancing our understanding of complex systems by providing more accurate and straightforward models.

3.2. Centrality-based approach

Next, we will elaborate on the centrality-based approach used for concept importance, firstly introduced in [1]. It is important to mention that contrary to the network reduction approach, the centrality-based approach allows quantifying concept importance.

The authors in [20] examined the effects of various policy scenarios on identified influential variables by using betweenness, closeness, and degree centrality measures. The study aims to suggest recommendations or solutions for decision-makers to address the problems identified and shift the system towards a more desirable outcome than the status quo. Closeness and betweenness centrality measurements consider the indirect links of a concept and highlight the shortest routes between that concept and others [21]. This approach reveals how a concept can have a global impact on the entire network rather than just a local impact. Overall, these measures are used during the condensation and aggregation of FCM models.

Several other studies [22–26] have applied in-degree, out-degree or degree centrality measures to obtain concept relevance scores in FCM-based models. These measures can be used to calculate either the overall influence of concepts on the whole model or the influence of individual concepts on others. For the entire model, the calculations consider both the positive and negative relationships, indicating the total influence of all components in the system. For individual concepts, the centrality scores measure their conceptual weight or importance by accounting for the magnitude of their positive and negative connections. To identify concepts that require special attention and determine the centrality of each factor in relation to others, the study [27] used absolute values for in-degree and out-degree measures.

The review paper in [7] elaborates on the comparative analysis of FCM models based on the structure of their underlying cognitive models. It aims to extract which neural concepts strongly influence the overall perspective of each FCM model. Within the metrics used to perform such a structural comparison, we can find network density, link-node ratio and centrality-based measures such as degree centrality, betweenness, and closeness.

Developing models with inherent explainability and self-generated justifications remains a persistent challenge. The model in [28] presents a novel classifier based on Long-term Cognitive Networks (LTCNs) with such characteristics. This model is rooted in the FCM formalism and uses a recurrence-aware decision model to avoid the issues associated with the unique fixed-point attractors. This classifier incorporates an intrinsic interpretability mechanism by assessing the relevance of each feature in the decision-making process. Such a mechanism scores the features according to the absolute centrality of neural concepts, which can be reasonably mapped to problem features.

Based on the revised literature, it has become evident that existing concept importance methods do not fully leverage the intrinsic interpretability of FCM models while also accounting for their dynamic properties.

4. Centrality-based measures

In this section, we formalize three centrality-based measures [29] to determine the relevance of each concept using the static information describing the relationships in the network. Such measures are not a contribution of this paper but will be employed as baselines.

The degree centrality of a node in a weighted digraph can be calculated by summing the absolute weights of the edges that arrive at or depart from the node. Hence, the degree centrality of a node indicates its overall connectivity within the causal network. In an FCM model, this measure can easily be computed as follows:

$$\Gamma(C_i) = \sum_j |w_{ji}| + \sum_l |w_{il}|. \tag{5}$$

The degree of a node in a digraph can be classified into two types: in-degree and out-degree. In a weighted directed digraph, the in-degree of a node is defined as the sum of the (absolute) weights of the incoming edges to the node being analyzed. This measure is useful in identifying important nodes in terms of receiving information from other nodes in the network. In an FCM model with signed weights, this measure can be computed as follows:

$$\Gamma^{-}(C_i) = \sum_j |w_{ji}|. \tag{6}$$

The out-degree centrality is defined as the sum of the (absolute) weights of the outgoing edges from the node. It provides a measure of the total weight of outgoing edges from the node, so it is useful in identifying important nodes in terms of providing information to other nodes in the network. Similarly to the in-degree centrality, we can adapt this measure to FCM models where the network involves signed weights as shown below,

$$\Gamma^{+}(C_i) = \sum_j w_{ij}. \tag{7}$$

Although these measures provide insights into the relevance of concepts, they neither consider their activation values nor the activation function used in the reasoning process. As such, centrality-based measures fail to capture the network's dynamic behavior, which is probably the most distinctive feature of FCM models.

5. SHAP-based feature importance

Aiming to tackle the limitations of centrality-based concept importance measures, we will modify the traditional SHAP method to deal with FCM models devoted to scenario simulation. More explicitly, the method is aimed to derive concept importance from the model's dynamic properties, which are materialized through its temporal states. To do that, we will rely on two main assumptions. Firstly, all concepts could be deemed as inputs and outputs, so there is no distinction between them in that regard. Secondly, the method should be robust to the activation function, the nonlinearity parameter and the network's convergence status. The latter is particularly problematic when having cyclic or chaotic behaviors since the method's output will depend on the number of iterations T, affecting the method's reliability. The convergence to a unique fixed point is even more problematic since the SHAP method could not derive any relevance scores from invariant outputs that are independent from the inputs.

To overcome these convergence issues, our method attributes the prediction of the FCM model to its initial activation values of neural concepts, while the outputs are regarded as the hidden states produced by the model during the inference process. This means that an FCM model with N concepts will have the same number of inputs given by $\mathbf{A}^{(0)} = (a_1^{(0)}, \ldots, a_i^{(0)}, \ldots, a_N^{(0)})$ whereas the outputs associated is given by the temporal state vector $\mathbf{H}_i^{(T)}$ obtained with the recursive relation $\mathbf{H}^{(t)} = (\mathbf{H}^{(t-1)}|\mathbf{A}^{(t)})$ where $\mathbf{H}^{(0)} = \mathbf{A}^{(0)}$. In this formulation, the $(\cdot|\cdot)$ operation stands for the horizontal concatenation operator of vectors. Therefore, it holds that

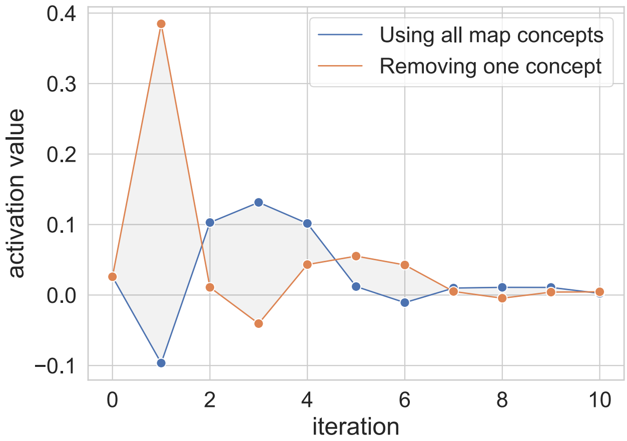


Fig. 2. Perturbation in the trajectory of a concept to a fixed point after inducing a perturbation in the network.

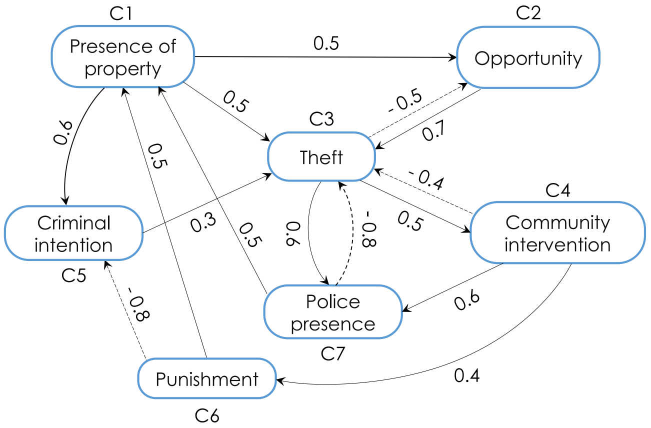


Fig. 3. FCM model concerning the "crime and punishment" case study described by seven neural concepts.

$$\mathbf{H}^{(T)} = \left(\mathbf{A}^{(0)}|\mathbf{A}^{(1)}|\mathbf{A}^{(2)}| \ldots |\mathbf{A}^{(T-2)}|\mathbf{A}^{(T-1)}|\mathbf{A}^{(T)}\right). \tag{8}$$

Eq. (9) shows how to compute the Shapley value denoting the importance of the neural concept C_i in an FCM model for scenario analysis,

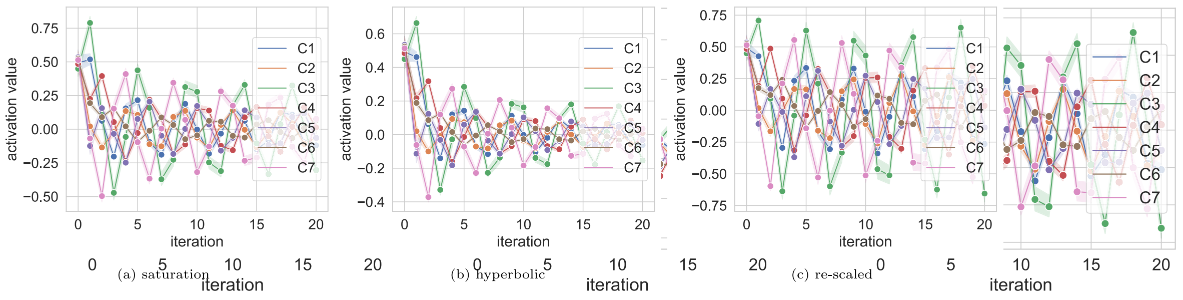
$$\Psi(C_i) = \sum_{\tilde{C} \subseteq C \setminus \{C_i\}} \Omega \cdot \left[\mathbf{H}_j^{(T)}(\tilde{C} \cup \{C_i\}) - \mathbf{H}_j^{(T)}(\tilde{C})\right] \tag{9}$$

where

$$\Omega = \frac{|\tilde{C}|!(|C| - |\tilde{C}| - 1)!}{|C|!} \tag{10}$$

such that C is the concept set, $\tilde{C}$ is a subset of $C \setminus \{C_i\}$, while $\mathbf{H}_j(\tilde{C} \cup \{C_i\})$ gives a concept's activation value assuming that all concepts are used in the calculation. Similarly, $\mathbf{H}_j(\tilde{C})$ gives a concept's activation value after excluding C_i from the reasoning process. While marginalizing a feature in pattern classification is obtained by replacing its values with the mean, in the FCM context, suppressing a concept means that its incoming and outgoing connections are zeroed. In that way, the removed concept will not influence others during the reasoning process since the activation functions used in this paper do not arbitrarily activate a concept when its raw activation value is zero, as happens with the sigmoid activation function.

The Shapley value approach satisfies the efficiency property, which states that the sum of the individual contributions should be equivalent

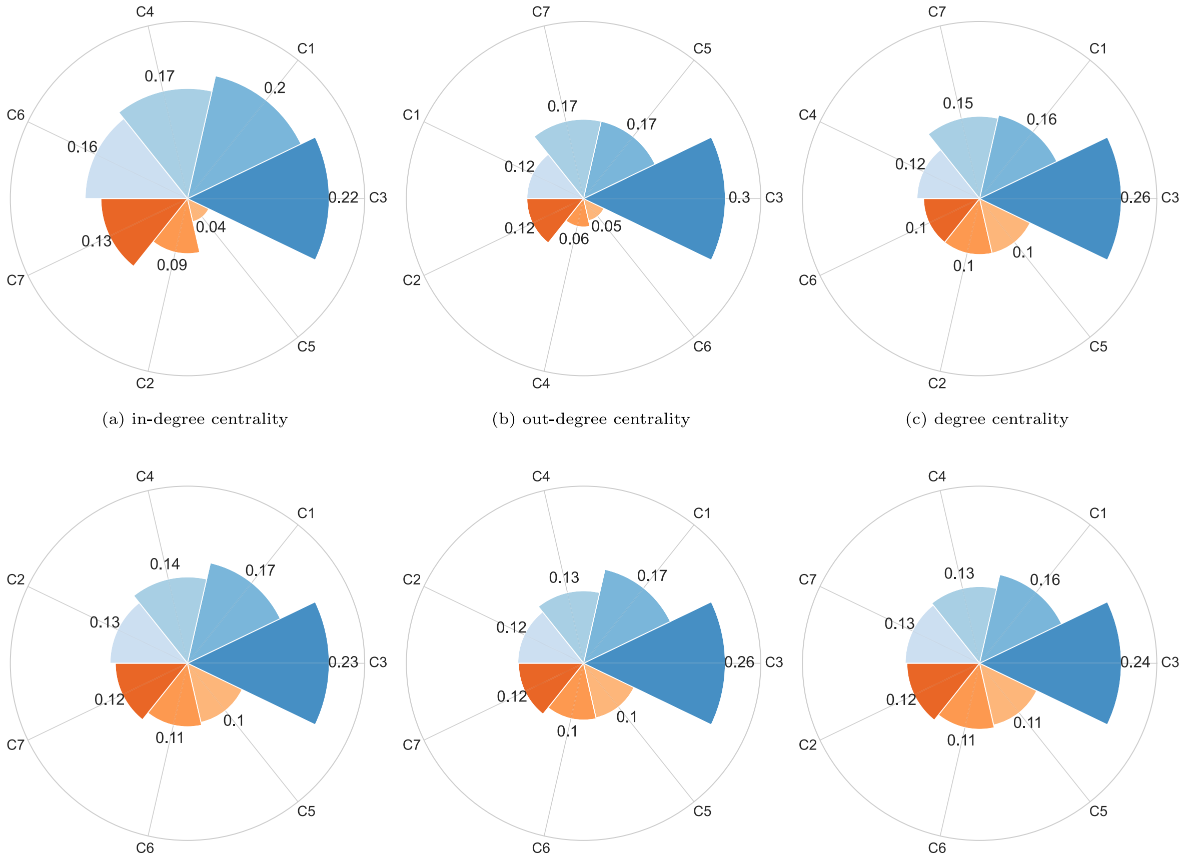


(a) saturation

(b) hyperbolic

(c) re-scaled

Fig. 4. Reasoning process of the FCM model concerning the "crime and punishment" case study for 100 randomly generated initial activation vectors and $T = 20$ iterations. The network does not converge to a fixed-point attractor for any of the activation functions used to perform the reasoning process. The re-scaled model does not converge since the transposed weight matrix does not have an eigenvalue that is strictly greater in magnitude than other eigenvalues.



(a) in-degree centrality

(b) out-degree centrality

(c) degree centrality

(d) shap saturation

(e) shap hyperbolic

(f) shap re-scaled

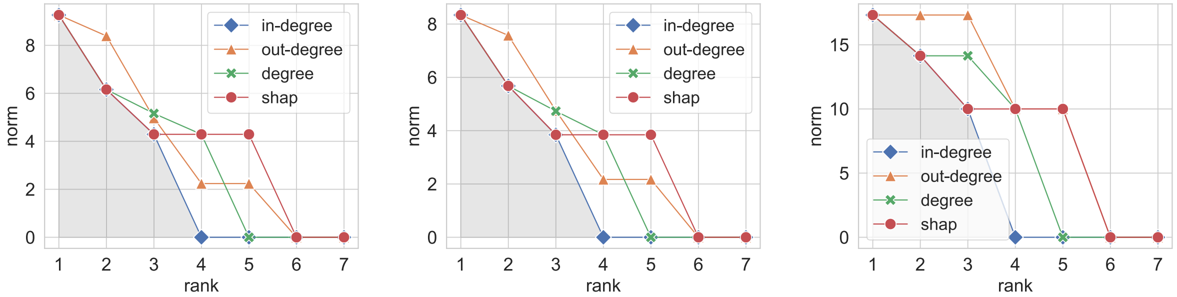
Fig. 5. Concept relevance scores produced by the degree centrality measures and the proposed SHAP method for the "crime and punishment" case study. In the case of the SHAP method, we also report the relevance scores for the activation functions adopted in the study (saturation, hyperbolic tangent and re-scaled).

to the discrepancy between the concepts' activation values across iterations and the average activation value. This means that the attribution values can be aggregated in order to obtain a single ranking about concept relevance in FCM models.

Overall, our method uses a holistic concept importance approach since the model's dynamic properties depend on the (a) initial activation values, (b) the activation function, (3) the reasoning rule, and (4) the weight matrix. Therefore, it also uses the network's static properties to derive the concept relevance scores. For example, marginalizing a concept with no outgoing connections will not bring any variability in the activation values of retained concepts. As such, it will pushed to the bottom of the concept relevance ranking. The same reasoning applies to concepts with near-zero incoming connections.

Fig. 2 portrays the algorithm's intuition using a functional example. Firstly, we assume that we have several neural concepts in the network. Secondly, we assume that one of these concepts will be used to monitor the alterations in the reasoning process after removing one or several neurons. Thirdly, we assume that all neurons converge to a unique fixed-point attractor. The example shows that removing a neuronal concept causes a significant disruption (visualized with the gray area) in the trajectory of the sensing concept to the fixed point. Therefore, the bigger the area, the more relevant the concept being analyzed. Note that the SHAP algorithm would fail if it focused on the sensing concept's outcome only rather than the concept's trajectory to the fixed point.

It is worth mentioning that our SHAP method works in two different settings that can be defined by the user. In the first setting, all concepts



(a) saturation (b) hyperbolic (c) re-scaled

Fig. 6. Average norm of the activation vector across iterations when suppressing the concepts in the same order as determined by each post-hoc method for the "crime and punishment" case study. The gray area is defined by the best-performing value reported by these methods after removing the concept ranked in the ith position of the relevance ranking.

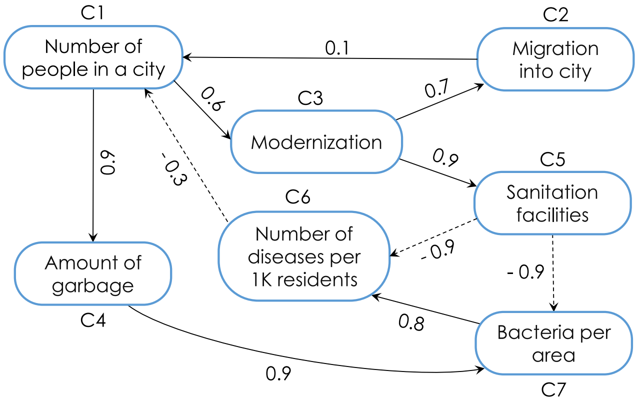


Fig. 7. FCM model concerning the "public health" case study described by seven neural concepts.

are regarded as inputs and outputs, which means that all concepts are used to quantify perturbations in the network. In the second setting, we can distinguish between inputs and outputs such that only sensing concepts are used to determine the relevance of those concepts deemed as inputs.

The last aspect to be discussed is the method's computational complexity. Predictably, the proposed SHAP method will be slower than any centrality-based measure since it needs to explore several coalitions (concept combinations) across several initial conditions. For each combination, the FCM's recurrent reasoning process needs to be executed using a pre-defined set of initial conditions. A recommendation to improve the method's efficiency includes stopping the reasoning process as soon a fixed point (not necessarily unique) is found or a clear cycle is observed. In addition, we can sample the temporal states produced by concepts in odd and even cycles since a few states could capture well the changes induced by concept marginalization. Concerning the number of features and samples, FCM models used for simulation purposes typically involve a low number of concepts and representative samples defined by domain experts.

6. Numerical simulations

To evaluate the performance of the propose method, we will rely on three real-world case studies for scenario analysis with FCM models. The experimental methodology can be briefly described as follows. Firstly, we build the FCM model for each problem and randomly generate 100 initial activation vectors. Secondly, we perform the reasoning rule depicted in Eq. (1) using the saturation, hyperbolic tangent, and re-scaled activation functions. In these simulations, the nonlinearity parameter is arbitrarily set to $\phi = 1.0$ while the number of iterations is set to $T = 20$. Thirdly, we compute the centrality-based metrics and the proposed SHAP method for each parametric setting and assess the agreement among these concept relevance approaches. Finally, we conduct a "pixel-flipping" experiment to determine which method accurately determines the importance of concepts in the investigated FCM models. The intuition of this experiment is that removing the most relevant concepts causes a significant drop in the activation values of retained concepts.

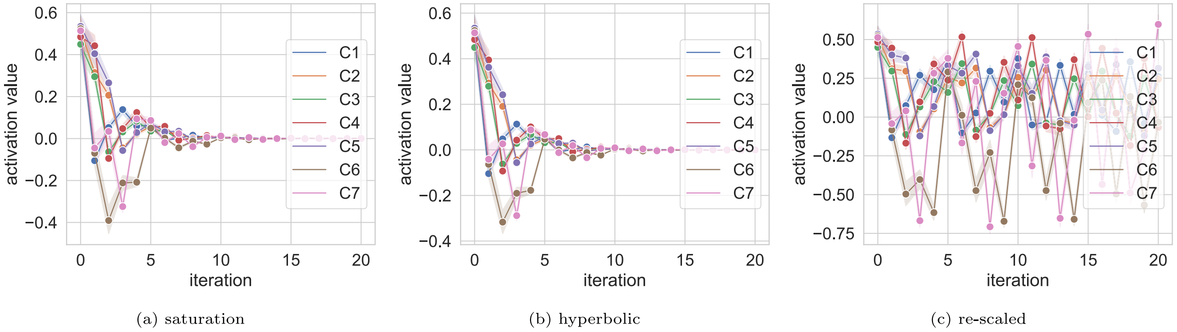
6.1. Case study 1: Crime and punishment

The first case study concerns the "crime and punishment" cognitive network, as described in [30], which is used to model the relationship between various factors that influence crime and the criminal justice system. The concepts in this model are presence of property ($C1$), opportunity ($C2$), theft ($C3$), community intervention ($C4$), criminal intention ($C5$), punishment ($C6$), and police presence ($C7$). Fig. 3 shows the causal network such that positive relationships are depicted as solid lines while negative ones are represented with dashed lines.

Firstly, it seems convenient to inspect the convergence behavior of the FCM model for the three activation functions (see Fig. 4). The reader can notice that the network fails to converge for the specified number of iterations, which means that the network produces state vectors that continue to change from one iteration to another. In the case of the model using the re-scaled activation function, this result brings no surprise since the transposed weight matrix does not have an eigenvalue that is strictly greater in magnitude than other eigenvalues.

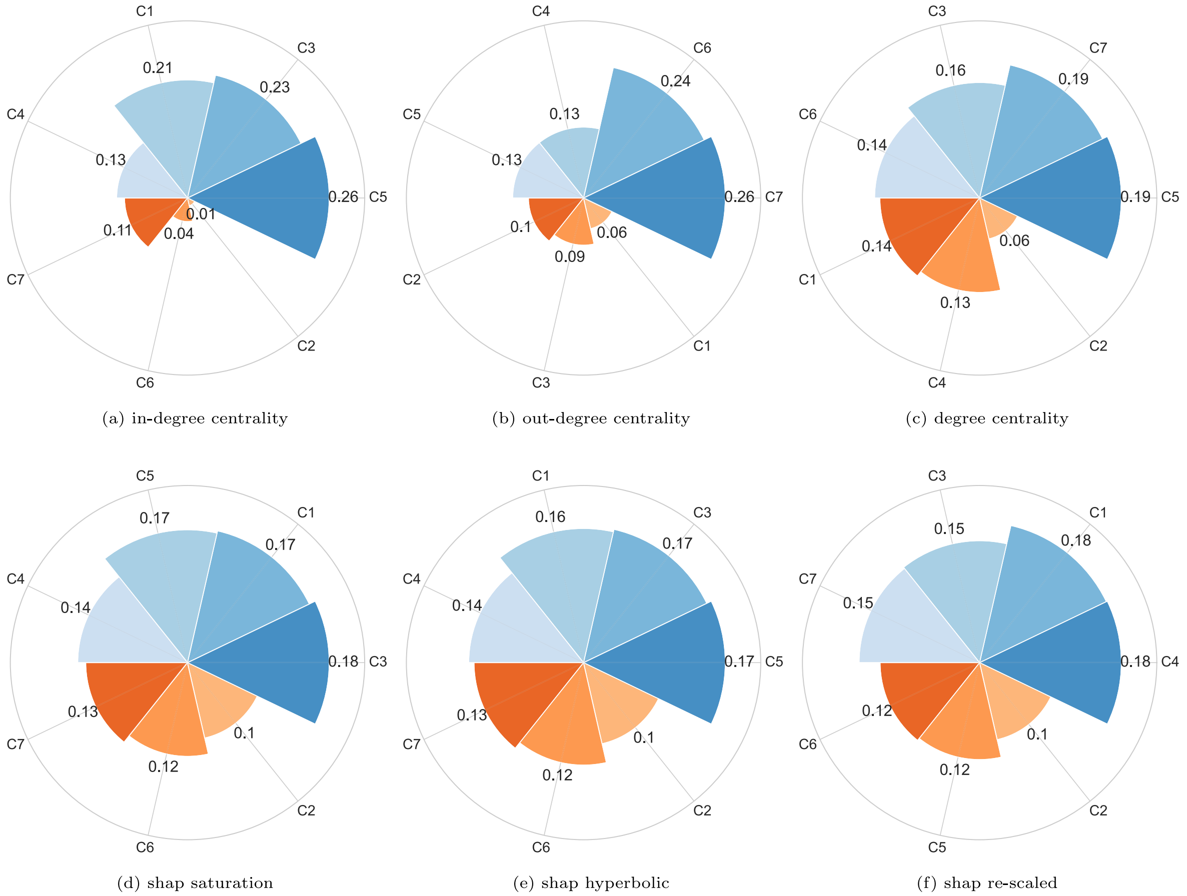
Fig. 5 reports the normalized concept relevance scores as computed by the centrality-based measures and the proposed SHAP method. It can be noted that $C3$ reports the largest centrality values overall and that the concept rankings differ in the remaining positions. These rankings will not change if we use other initial activation vectors, activation functions or reasoning rules, as they depend on the weights only. The SHAP method also detected $C3$ as the most important concept in all cases. However, the rankings barely change despite using the model's dynamic components to elicit concept importance.

In order to determine which method computed the most accurate concept relevance results, we will resort to a version of the "pixel flipping" experiment [31,32]. This procedure is widely used to understand the impact of individual pixels on the predictions made by a machine learning model. While the term "pixel flipping" refers specifically to image data, a similar reasoning can be applied to other types of data with feature-based representations. The adapted variant implemented in this study consists of systematically removing the concepts from the cognitive network in the same order they were ranked by a given post-hoc method. The intuition of this experiment is that removing relevant concepts causes a significant drop in the activation values of retained concepts.



(a) saturation (b) hyperbolic (c) re-scaled

Fig. 8. Reasoning process of the FCM model concerning the "public health" case study for 100 randomly generated initial activation vectors and $T = 20$ iterations. This model converges to a fixed point when using the saturation and the hyperbolic tangent function. The fixed point seems to be unique since all final activation vectors are the same, which is visualized with the absence of a shadow (standard deviation). In the case of the re-scaled transfer function, the network fails to converge since the transposed weight matrix does not have an eigenvalue that is strictly greater in magnitude than other eigenvalues.



(a) in-degree centrality (b) out-degree centrality (c) degree centrality

(d) shap saturation (e) shap hyperbolic (f) shap re-scaled

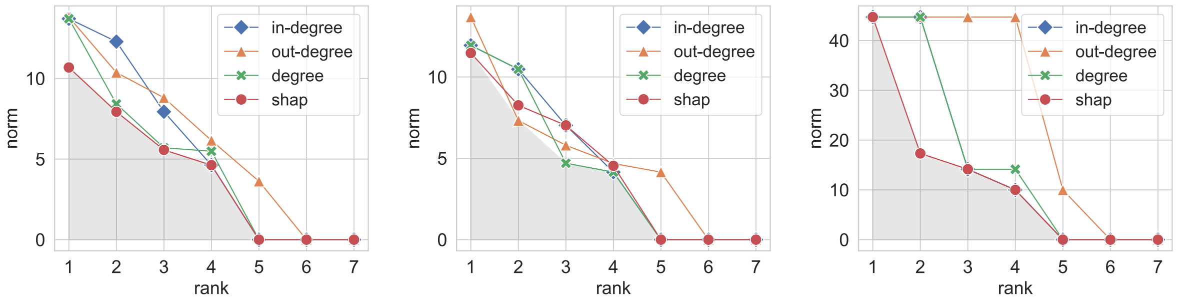
Fig. 9. Concept relevance scores produced by the degree centrality measures and the proposed SHAP method for the "public health" case study. In the case of the SHAP method, we also report the relevance scores for the activation functions adopted in the study (saturation, hyperbolic tangent and re-scaled).

Fig. 6 shows the average norm of the activation vector across iterations when suppressing the concepts in the same order as ranked by the post-hoc method. Therefore, the x-axis gives the concept ranking while the y-axis gives the average norm value. The gray area is defined by the best-performing value after removing the concept ranked in the ith position of the relevance ranking. Notice that a zero-norm activation vector (after performing a fixed number of iterations) means that concepts are no longer connected with each other. As such, the faster we approach this extreme case, the more reliable the relevance scores used to build the concept ranking.

The simulation results indicate that the in-degree centrality measure is the best-performing post-hoc method. The proposed method performs similarly to the in-degree centrality measure in the most and least important concepts, which is a positive outcome. In contrast, the out-degree measure reports the worst results since it fails to quantify a large drop when removing the most relevant concept from the network.

6.2. Case study 2: Public health

The second case study concerns civil engineering and studies the consequences of the increase of a city's population and modernization to the city's public health (see Fig. 7). This FCM was used in [14] to compare the inference capabilities of binary, trivalent and sigmoid FCM models. The concepts in this model are people in a city ($C1$), migration



(a) saturation

(b) hyperbolic

(c) re-scaled

Fig. 10. Average norm of the activation vector across iterations when suppressing the concepts in the same order as determined by each post-hoc method for the "public health" case study. The gray area is defined by the best-performing value reported by these methods after removing the concept ranked in the ith position of the relevance ranking.

into city ($C2$), modernization ($C3$), amount of garbage ($C4$), sanitation facilities ($C5$), diseases per 1000 residents ($C6$), and bacteria per area ($C7$).

A closer inspection of the dynamic behavior of this FCM model (see Fig. 8) shows that it converges to a fixed point when using the saturation and hyperbolic tangent functions. The fixed point seems to be unique since the FCM model produces the same state vector for all initial activation vectors after performing $T = 20$ iterations. In the case of the re-scaled transfer function, the network unsurprisingly fails to converge since the transposed weight matrix does not have an eigenvalue that is strictly greater in magnitude than other eigenvalues.

Fig. 9 reports the normalized concept relevance scores computed by the centrality-based measures and the proposed SHAP post-hoc method. The simulation results show that $C5$ reports the largest in-degree centrality and degree centrality scores, while $C7$ is the concept with the largest out-degree centrality. The SHAP method indicates that $C3$, $C5$ and $C4$ are the most relevant concepts when using the saturation, hyperbolic tangent, and re-scaled activation function, respectively. Note how differences in the activation values (what is the information captured by the proposed SHAP method to derive concept importance) change from one setting to another.

To determine which method computed the most accurate concept relevance results, we will again rely on the "pixel flipping" experiment. Fig. 10 shows the average norm of the activation vector across iterations when suppressing the concepts in the same order as ranked by the post-hoc method. As before, the gray area is defined by the best-performing value after removing the concept ranked in the ith position of the relevance ranking. It should be recalled that the rankings associated with each method are visualized in Fig. 9. The results indicate the proposed SHAP method is clearly superior to the centrality-based measures, which becomes more evident for the saturation and re-scaled activation functions. In other words, it reports the smallest area under the curve while approaching zero faster than the other methods.

6.3. Case study 3: Car industry

The third case study refers to a complex system representing a car industry, which was taken from [33]. The neural concepts describing this system are high profits ($C1$), customer satisfaction ($C2$), high sales ($C3$), union raises ($C4$), safer cars ($C5$), foreign competition ($C6$), and lower prices ($C7$). Fig. 11 visualizes the network such that positive relationships are shown as solid lines while negative ones are depicted with dashed lines.

By inspecting the model's convergence, we notice that the network does not converge for any of the activation functions (see Fig. 12). As in the previous case studies, the model using the re-scaled transfer function was not expected to converge since the transposed weight matrix does not have an eigenvalue that is strictly greater in magnitude than other eigenvalues.

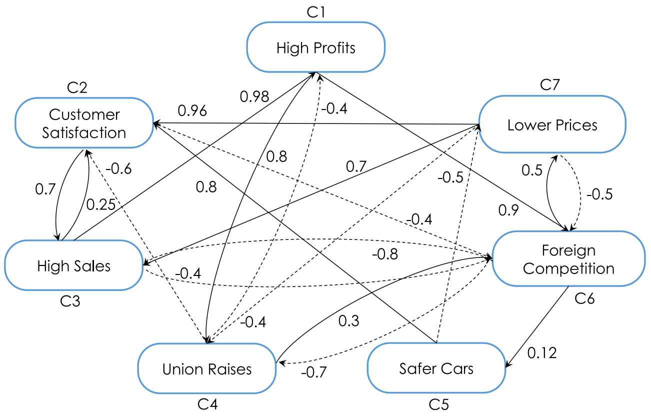


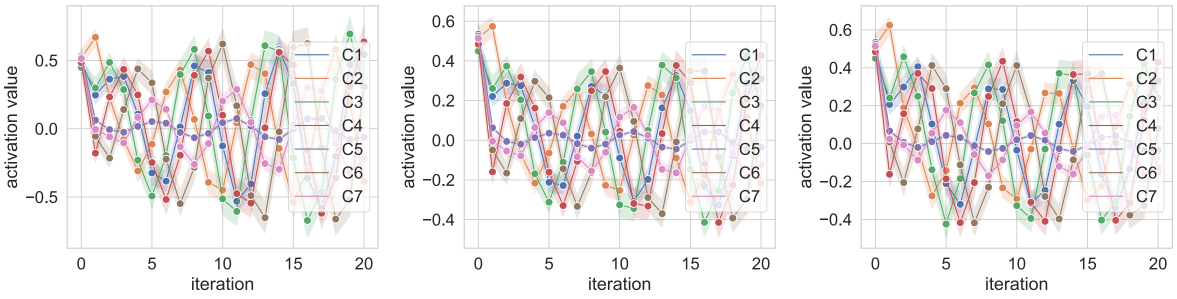
Fig. 11. FCM model concerning the "car industry" case study described by seven neural concepts.

Fig. 13 shows the normalized concept relevance scores computed with the centrality-based measures and the proposed SHAP post-hoc method. While $C7$ has the largest in-degree centrality, $C2$ and $C6$ report the largest out-degree and degree centrality scores, respectively. When considering the initial activation values, neurons' activation values and the activation function besides the edge weights, the concept rankings further change. For example, it can be noticed that $C2$ is no longer an important concept as indicated by the out-degree centrality measure, but $C6$ and $C7$. This result further supports the superiority of the SHAP method for determining concept importance in FCM models used for scenario analysis.

Fig. 14 displays the average norm of the activation vector across iterations when suppressing the concepts in the same order as ranked by the post-hoc method. The rankings for each post-hoc method are visualized in Fig. 13. The curves obtained using the saturation and hyperbolic tangent activation functions indicate that all methods perform comparably, with SHAP being slightly better since it does not report any peaks. The results concerning the re-scaled activation function indicate that the SHAP method is the winner. Note that all curves have a plateau that involves the top-3 relevant concepts in the network, meaning they are equally relevant.

7. Concluding remarks

The lack of approaches to exploit the interpretability of FCM models while considering their dynamic properties inspired the proposal of a new method. Although centrality-based feature importance approaches provide insights into the relevance of concepts, they fail to capture the network's dynamic behavior. As such, the paper proposed a SHAP-based approach for computing concept attribution in FCM models for scenario analysis.

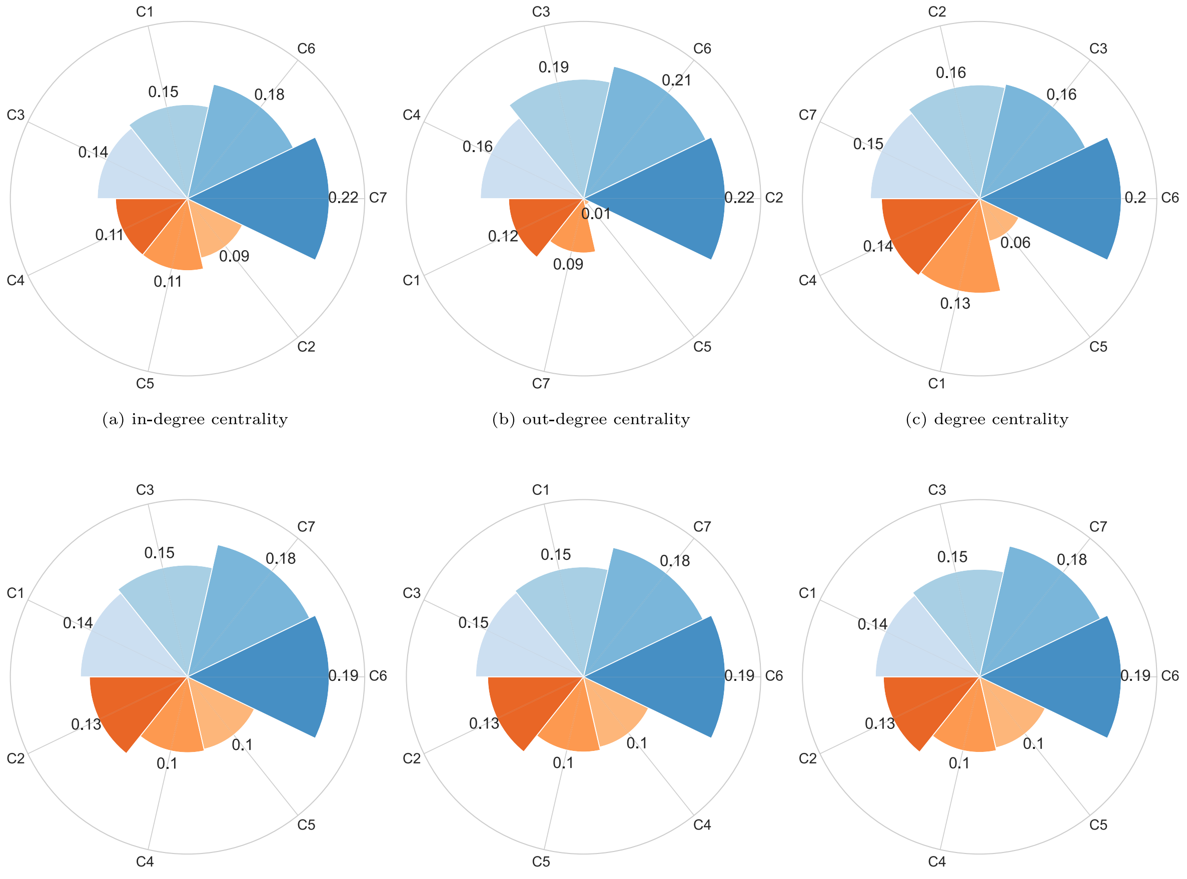


(a) saturation

(b) hyperbolic

(c) re-scaled

Fig. 12. Reasoning process of the FCM model concerning the "car industry" case study for 100 randomly generated initial activation vectors and $T = 20$ iterations. This model fails to converge to a fixed point for all activation functions used in the study since the concepts' activation values continue to change from one iteration to another. In the case of the re-scaled transfer function, this convergence behavior was predicted since the transposed weight matrix does not have an eigenvalue that is strictly greater in magnitude than other eigenvalues, thus indicating the presence of cycles or a chaotic behavior.



(a) in-degree centrality

(b) out-degree centrality

(c) degree centrality

(d) shap saturation

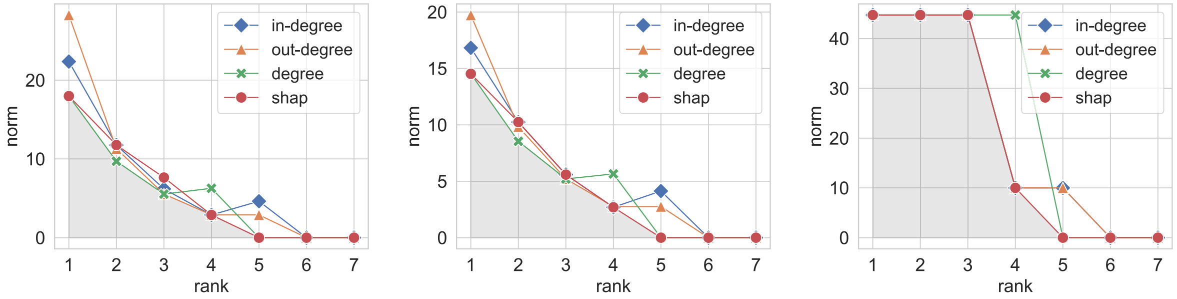
(e) shap hyperbolic

(f) shap re-scaled

Fig. 13. Concept relevance scores produced by the degree centrality measures and the proposed SHAP method for the "car industry" case study. In the case of the SHAP method, we also report the relevance scores for the activation functions adopted in the study (saturation, hyperbolic tangent and re-scaled).

The proposed method operates under the assumption that all concepts are inputs and outputs and should be robust to factors such as the activation function, nonlinearity parameters, and network convergence status. As stated before, convergence poses difficulties when dealing with cyclic or chaotic behavior, which can impact the method's reliability. Additionally, invariant outputs independent of inputs pose a bigger challenge since existing feature importance methods cannot derive relevance scores from them. To overcome these issues, the proposed method attributes the FCM model's prediction to its initial activation values of concepts while considering the hidden states produced by the model during inference as outputs. Using the Shapley value approach satisfies the efficiency property, allowing attribution values to be aggregated to obtain a single ranking of concept relevance in FCM models.

It is important to mention that existing centrality-based measures and the proposed SHAP method induce different types of knowledge characterizing concept importance. Centrality-based measures give information about the strength of incoming and outgoing weights associated with each neural concept, however, such static knowledge is insufficient to conclude the concepts' relevance. In contrast, the proposed SHAP method effectively captures situations that are missed by centrality-based measures since it uses all pieces of information produced by the network when performing the reasoning process. More explicitly, the numerical simulations showed that concept attribution



(a) saturation (b) hyperbolic (c) re-scaled

Fig. 14. Average norm of the activation vector across iterations when suppressing the concepts in the same order as determined by each post-hoc method for the "car industry" case study. The gray area is defined by the best-performing value reported by these methods after removing the concept ranked in the *i*th position of the relevance ranking.

scores often vary when changing the activation function, the initial conditions and the inference rule used to update the concept's activation values.

CRediT authorship contribution statement

Gonzalo Nápoles: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Nevena Ranković:** Writing – original draft, Resources, Investigation. **Yamisleydi Salgueiro:** Writing – review & editing, Resources, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Y. Salgueiro would like to acknowledge the support provided by the National Center for Artificial Intelligence CENIA FB210017, Basal ANID and the Chilean National Agency for Research and Development (ANID), InES-Género project INGE210025.

References

[1] B. Kosko, Fuzzy cognitive maps, Int. J. Man-Mach. Stud. 24 (1986) 65–75, http://dx.doi.org/10.1016/s0020-7373(86)80040-2.

[2] A. Jastrzębska, A. Cisłak, Interpretation-aware cognitive map construction for time series modeling, Fuzzy Sets and Systems 361 (2019) 33–55, http://dx.doi.org/10.1016/j.fss.2018.05.013.

[3] G. Nápoles, J.L. Salmeron, W. Froelich, R. Falcon, M.L. Espinosa, F. Vanhoenshoven, R. Bello, K. Vanhoof, Fuzzy cognitive modeling: Theoretical and practical considerations. volume 142, 2019, pp. 77–87. http://dx.doi.org/10.1007/978-981-13-8311-3_7.

[4] I.D. Apostolopoulos, P.P. Groumpos, Fuzzy cognitive maps: Their role in explainable artificial intelligence, Appl. Sci. (2023) 13, http://dx.doi.org/10.3390/app13063412.

[5] M. Tyrovolas, X.S. Liang, C. Stylios, A novel framework for enhanced interpretability in fuzzy cognitive maps, 2023, http://dx.doi.org/10.36227/techrxiv.22718032.v1.

[6] A. Garzón Casado, P. Cano Marchal, J. Gómez Ortega, J. Gámez García, Visualization and interpretation tool for expert systems based on fuzzy cognitive maps, IEEE Access 7 (2019) 6140–6150, http://dx.doi.org/10.1109/access.2018.2887355.

[7] B.S. Yoon, A.J. Jetter, Comparative analysis for fuzzy cognitive mapping, in: 2016 Portland International Conference on Management of Engineering and Technology (PICMET), 2016, pp. 1897–1908, http://dx.doi.org/10.1109/picmet.2016.7806755.

[8] M.T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you? explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, http://dx.doi.org/10.1145/2939672.2939778.

[9] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. (2017) 30.

[10] M. León, G. Nápoles, C. Rodriguez, M.M. García, R. Bello, K. Vanhoof, A fuzzy cognitive maps modeling, learning and simulation framework for studying complex system, in: New Challenges on Bioinspired Applications: 4th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2011, Springer, 2011, pp. 243–256, http://dx.doi.org/10.1007/978-3-642-21326-7_27.

[11] G. Nápoles, I. Grau, L. Concepción, L.K. Koumeri, J.P. Papa, Modeling implicit bias with fuzzy cognitive maps, Neurocomputing 481 (2022) 33–45, http://dx.doi.org/10.1016/j.neucom.2022.01.070.

[12] S. Bueno, J.L. Salmeron, Benchmarking main activation functions in fuzzy cognitive maps, Expert Syst. Appl. 36 (2009) 5221–5229, http://dx.doi.org/10.1016/j.eswa.2008.06.072.

[13] I.Á. Harmati, M.F. Hatwágner, L.T. Kóczy, Global stability of fuzzy cognitive maps, Neural Comput. Appl. 35 (2023) 7283–7295, http://dx.doi.org/10.1007/s00521-021-06742-9.

[14] A.K. Tsadiras, Comparing the inference capabilities of binary, trivalent and sigmoid fuzzy cognitive maps, Inform. Sci. 178 (2008) 3880–3894, http://dx.doi.org/10.1016/j.ins.2008.05.015.

[15] L. Concepción, G. Nápoles, R. Falcon, K. Vanhoof, R. Bello, Unveiling the dynamic behavior of fuzzy cognitive maps, IEEE Trans. Fuzzy Syst. 29 (2021) 1252–1261, http://dx.doi.org/10.1109/tfuzz.2020.2973853.

[16] E.I. Papageorgiou, M.F. Hatwágner, A. Buruzs, L.T. Kóczy, A concept reduction approach for fuzzy cognitive map models in decision making and management, Neurocomputing 232 (2017) 16–33, http://dx.doi.org/10.1016/j.neucom.2016.11.060.

[17] M.F. Hatwágner, E. Yesil, M.F. Dodurka, E. Papageorgiou, L. Urbas, L.T. Kóczy, Two-stage learning based fuzzy cognitive maps reduction approach, IEEE Trans. Fuzzy Syst. 26 (2018) 2938–2952, http://dx.doi.org/10.1109/tfuzz.2018.2793904.

[18] A. Nápoles, C. Mosquera, K. Vanhoof, W. Homenda, Deterministic learning of hybrid fuzzy cognitive maps and network reduction approaches, Neural Netw. 124 (2020) 258–268, http://dx.doi.org/10.1016/j.neunet.2020.01.019.

[19] M.F. Hatwágner, L.T. Kóczy, Novel methods of FCM model reduction, 2022, pp. 101–112, http://dx.doi.org/10.1007/978-3-030-88817-6_12.

[20] M. Obiedat, S. Samarasinghe, A novel semi-quantitative fuzzy cognitive map model for complex systems for addressing challenging participatory real life problems, Appl. Soft Comput. 48 (2016) 91–110, http://dx.doi.org/10.1016/j.asoc.2016.06.001.

[21] E.A. Lavin, P.J. Giabbanelli, A.T. Stefanik, S.A. Gray, R. Arlinghaus, Should we simulate mental models to assess whether they agree? in: Proceedings of the Annual Simulation Symposium, 2018.

[22] K. Kokkinos, E. Lakioti, E. Papageorgiou, K. Moustakas, V. Karayannis, Fuzzy cognitive map-based modeling of social acceptance to overcome uncertainties in establishing waste biorefinery facilities, Front. Energy Res. 6 (112) (2018) http://dx.doi.org/10.3389/fenrg.2018.00112.

[23] K. Papageorgiou, P.K. Singh, E.I. Papageorgiou, H. Chudasama, D. Bochtis, G. Stamoulis, Participatory modelling for poverty alleviation using fuzzy cognitive maps and owa learning aggregation, PLoS One 15 (2020) e0233984, http://dx.doi.org/10.1371/journal.pone.0233984.

[24] B.G. Giles, C.S. Findlay, G. Haas, B. LaFrance, W. Laughing, S. Pembleton, Integrating conventional science and aboriginal perspectives on diabetes using fuzzy cognitive maps, Soc. Sci. Med. 64 (2007) 562–576, http://dx.doi.org/10.1016/j.socscimed.2006.09.007, URL: https://www.sciencedirect.com/science/article/pii/S0277953606004758.

[25] V. Mago, H.K. Morden, C. Fritz, T. Wu, S. Namazi, P. Geranmayeh, R. Chattopadhyay, V. Dabbaghian, Analyzing the impact of social factors on homelessness: a fuzzy cognitive map approach, BMC Med. Inform. Decis. Mak. 13 (2013) 94.

[26] K. Fonseca, E. Espitia, L. Breuer, A. Correa, Using fuzzy cognitive maps to promote nature-based solutions for water quality improvement in developing-country communities, J. Clean. Prod. 377 (2022) 134246, http://dx.doi.org/10.1016/j.jclepro.2022.134246.

[27] M. Malakoutikhah, M. Alimohammadlou, M. Jahangiri, H. Rabiei, S.A. Faghihi, M. Kamalinia, Modeling the factors affecting unsafe behaviors using the fuzzy best–worst method and fuzzy cognitive map, Appl. Soft Comput. 114 (2022) 108119, http://dx.doi.org/10.1016/j.asoc.2021.108119.

[28] Y. Nápoles, I. Grau, M.L. Espinosa, Recurrence-aware long-term cognitive network for explainable pattern classification, IEEE Trans. Cybern. (2022) 1–12, http://dx.doi.org/10.1109/tcyb.2022.3165104.

[29] V. Latora, V. Nicosia, G. Russo, Centrality measures, 2017, pp. 31–68, http://dx.doi.org/10.1017/9781316216002.004.

[30] J. Carvalho, J.A. Tomè, Rule based fuzzy cognitive maps and fuzzy cognitive maps-a comparative study, in: 18th International Conference of the North American Fuzzy Information Processing Society-NAFIPS, 1999, pp. 115–119, http://dx.doi.org/10.1109/nafips.1999.781665.

[31] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.R. Müller, Evaluating the visualization of what a deep neural network has learned, IEEE Trans. Neural Netw. Learn. Syst. 28 (2016) 2660–2673, http://dx.doi.org/10.1109/tnnls.2016.2599820.

[32] L. Tětková, L.K. Hansen, Robustness of visual explanations to common data augmentation methods, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3714–3719, http://dx.doi.org/10.48550/arXiv.2304.08984.

[33] A. Tsadiras, N. Bassiliades, Ruleml representation and simulation of fuzzy cognitive maps, Expert Syst. Appl. 40 (2013) 1413–1426, http://dx.doi.org/10.1016/j.eswa.2012.08.035.