Fuzzy Cognitive Maps and Explainable Artificial Intelligence: a critical perspective

Ifigeneia Athanasoula[1], Ioannis D. Apostolopoulos[2], and Peter P. Groumpos[1]

[1] Department of Electrical and Computer Technology Engineering, University of Patras, Patras, Greece; ece6932@upnet.gr, groumpos.ece.upatras.gr
[2] Department of Medical Physics, School of Medicine, University of Patras, Patras, Greece; ece7216@upnet.gr

Abstract. There is a lot of discussion regarding the interpretability and explainability of modern artificial intelligence methodologies, especially in applications such as medical imaging. Scientists argue that the most vital drawback of complex algorithms is their behaviour as black boxes. It is agreed that applying the newly developed methods in industry, medicine, agriculture, and other modern fields, such as the Internet of Things, requires the trustfulness of the systems from the users. Users are always entitled to know why and how each method made a decision and which factors played a role. Otherwise, they will always be wary of using new techniques. Fuzzy Cognitive Maps are an evolving computational method to model human knowledge, provide decisions handling uncertainty, and are the core of many modern intelligent systems. Numerous studies in various fields employ FCMs, which report top performance, sometimes proving superior to several Machine Learning models. In this work, we analyse the nature of FCMs in terms of their trust, transferability, causality, informativeness, and transparency, providing the reader with several success stories that reveal the suitability of FMCs in many domains.

Keywords: Fuzzy Cognitive Maps; Explainability; Explainable Artificial Intelligence; Interpretability

1 Introduction

Modern Artificial Intelligence, especially Machine Learning (ML) and Deep Learning (DL) [1], has become an established and dominant discipline in many activity sectors embracing new technologies. The feature development of human society lies in ML and DL to solve intricate problems and offer reliable solutions [2]. It is often discussed that the potential of ML and DL may transform human-oriented processes into automatic everyday tasks, wherein human intervention is no longer required.

In medicine, the decisions of ML and DL models could affect human lives directly. Human health differentiates from other human activities in many ways. In medicine, every decision must be justified based on golden, globally accepted standards, although, on many occasions, medical staff are required to improvise.

In this context, the act of DL as a black box [3] makes the medical community reluctant to adopt DL in assisting with everyday challenges. There is an increasing demand for transparency and interpretability of the new methods. Since 2018, a new discipline has been introduced by an increasing number of researchers. This discipline is called eXplainable Artificial Intelligence (XAI) [4]. XAI refers not only to technical aspects of the DL models that ensure some level of interpretability but also integrates the concepts of data privacy and accountability.

Fuzzy Cognitive Maps (FCM) are elements and methods union of fuzzy logic and Neural Networks. It is a calculating method capable of processing uncertain information. An FCM describes a system with a graphic display, which includes concepts and the relationships among them. They intend to model human knowledge, not discover it from raw data.

Our motivation for the structure of this work is the recent work of Arrieta et al. [4], which raises issues regarding the interpretability of modern AI methods. Moreover, in the prementioned work, the author attempts to divide the problem and the concept of interpretability into many pieces - aspects that each system must be evaluated. We intend to defend FCMs and their approach against the correctly raised issues in our work. Moreover, we attempt to fortify the theory of FCM with contemporary answers to vital issues, such as accuracy metrics, experts' collaboration, causality and transparency.

2 Methods

From a technical point of view, considering the interpretability of a newly developed ML or DL model can improve its implementability. Firstly, designing an interpretable model ensures impartiality in the decision-making process. Secondly, interpretability can point out potential adversarial perturbations that affect the prediction. This enables specific improvements to the core of the model itself. Thirdly, interpretability can ensure that only the meaningful features infer the desired output, thereby highlighting that an underlying causality exists in the given data and the model reasoning.

Several research papers mention crucial XAI aspects of the proposed models to facilitate interpretability. Arrieta et al. [4] classified those aspects as follows:

a) Trustworthiness: We trust the model to operate as usual without supervision, even in unknown conditions.

b) Causality: Causality requires a wide frame of prior knowledge. Simply discovering data variables correlated with the desired effect does not ensure that the model can explain the deep cause and effect relationships between the features.

c) Transferability: The ability of the model to operate as usual when facing a different problem by transferring its knowledge, which is initially obtained by training in a specific domain.

d) Informativeness: an XAI model should be able to be informative as to its inner structure and the decision-making problem it is solving.

e) Confidence: Confidence should always be assessed on a model in which reliability is expected. One of the most crucial aspects of the model to ensure confidence is its stability to produce reliable results.

f) Fairness: an XAI model should suggest a clear visualisation of any relations affecting the desired task, allowing for fairness of ethical analysis

g) Accessibility: Accessibility refers to the property that allows end users to understand how the model works.

h) Interactivity: In fields where the end users are an important part of the process, the ability of a model to offer interaction with the user is often considered a vital aspect.

i) Privacy Awareness: Privacy breaches may be entailed when a model is not transparent enough about what information has captured and is used to predict the desired outcome. On the contrary, total transparency regarding inner relations may also raise privacy concerns, especially when non-authorised users gain access to the model's mechanism.

2.1 Fuzzy Cognitive Maps in a nutshell

A Fuzzy Cognitive Map shows a graphic display to present the model and the system's behaviour. The notions of an FCM interact according to non-precise rules, so the procedures of multi-complex systems are simulated. The FCMs constitute a modelling method consisting of a grid of interconnected and interdependent concepts Ci (variables), as well as of the existing relations among them, W (weights). Fuzzy Cognitive Maps operate through a knowledge integration of a group of experts, who examine and describe the system. A very descriptive overview can be found in the work of Groumpos et al. [5].

3 Results and Discussion

Considering that the is no specific formal technical meaning for the node of interpretability, we analyse the aspects of Fuzzy Cognitive Maps in the following areas – fields: (a) trust, (b) transferability, (c) causality, (d) informativeness, (e) transparency, (f) post – hoc interpretability.

3.1 FCM and Trust

Trust is directly related to the model's accuracy. However, trust as a notion itself has not a specific meaning when it comes to machine learning. It is wrong to identify the algorithm's accuracy with our confidence. We trust the FCM models as much as we trust human knowledge. For example, in [6], the authors presented an FCM that achieves 90.26% accuracy in identifying brain tumours of 100 patients. Trust is also related to the amount of evaluation data and the origin of those data.

If we are sure of the authenticity of the scientific knowledge of those involved in creating the model, we are sure of the model itself, as long as the model does indeed work. A second factor is our confidence in the model's technical creators. But this topic applies to everything that man creates, and it makes no sense to specialise it further in FCMs. Fuzzy Cognitive Maps are trustworthy because they incorporate human knowledge. The experts only define their weights and mechanism. The model can be

trusted as long as the experts' way is trustworthy and based on proven assumptions. Concluding, FCMs give what they take.

3.2 FCM and Transferability

Transferability usually applies to machine learning methods, such as Neural Networks or Convolutional Neural Networks. Typically, we choose training and test data by randomly partitioning examples from the same distribution. We then judge a model's generalisation error by the gap between its training and test data performance. However, humans exhibit a far richer capacity to generalise, transferring learned skills to unfamiliar situations.

FCMs may or may not use data [5]. The ability to generalise depends strongly on their architecture, complexity, and overall simplicity [5]. FCM's ability to generalise at a higher level than human beings is to be investigated. So far, the proposed FCM models try to mimic the generalisation ability of the developers and the experts. That ability is reflected in the weight values, the concept handling, and the ways the concepts interact. A very intuitive example of FCM's transferability is presented in [7], where the authors present an FCM model to predict the spread of COVID-19 diseases across different countries. It is found that their model is applicable in many situations without dropping its accuracy.

3.3 FCMs and Causality

Fuzzy Cognitive Maps provide a distinguishable way to express the cause-effect relationship between phenomena, between numeric, nominal, binary, or categorical parameters. A complex system may include all the factors mentioned above. Relationships between them, provided that they are discovered, can be represented visually and mathematically through FCMs. This way, not only the developers but also the experts in the field may observe and understand the FCM representations. Inspecting the visual side of the FCM, one can understand every parameter and connection and have a first sight of how each concept affects the other.

The interconnections between concepts learned by supervised learning algorithms are not ensured to reflect causal connections. That does not mean that unobserved causes may not exist; one target of supervised learning is to make assumptions regarding the relations between mutually affected concepts during training and testing. Assumptions that can be confirmed or denied experimentally later. Compared to trainable artificial intelligence algorithms, FCM does not intend to discover associations that may or may not exist.

In the work of Morone et al. [8], the authors demonstrated some policy drivers, such as "Public food waste rules", "Investments and infrastructure", and "Small-scale farming", that are particularly effective in supporting a new and sustainable food consumption model. Their FCM model modelled the causality of the involved attributes successfully.

3.4 FCMs and Informativeness

Developing an FCM that expresses the reality may conflict with the development of an FCM based on a specific dataset, which does not explain the reality accurately. A specific weight between two concepts suggested by an expert on the matter may not be confirmed by a sample of data. This suggests that the dataset is bad. Designing from scratch, or employing a state-of-the-art algorithm to make predictions on this dataset, is pointless. One can obtain high accuracy; however, the trained system's weights will not explain the real relationships; thus, the dataset's flaws are transferred to the model.

One can blame FCMs for non-dynamism. From one aspect, that is correct. FCMs' way of treating the concepts and the relationships are static, although several approaches argue that FCM weights can become trainable.

Nevertheless, the nature of the FCM enables them to present the user with the entire reasoning in both quantitative and qualitative manners. A study by Papageorgiou et al. [9] presented a very informative graph that reveals all the relationships among the concepts and informs the user about the assigned weights.

3.5 FCMs and Transparency

New regulations in the European Union proposed that people affected by algorithmic decisions have a right to an explanation [10]. Exactly what structure such clarification may take or how such a clarification could be demonstrated right remain open questions. Moreover, the same regulations suggest that algorithmic decisions should be contestable. The comprehensiveness of FCM, its ability to visualise the procedure and its transparency in every step suggests that the user always has complete control of the decisions the model suggests. We will discuss more on this matter in post hoc interpretability.

3.6 FCMs and Post-hoc Interpretability

Post-hoc interpretability introduces a particular way to deal with extricating data from learned models. While posthoc understandings regularly don't clarify decisively how a model functions, they may, in any case, give valuable data for specialists and end clients of AI. Some normal ways to deal with posthoc elucidations incorporate regular language clarifications, perceptions of scholarly portrayals or models, and clarifications by model.

Humans often justify decisions verbally. Providing the user information regarding the reason the FCM predicts the specific class or suggests certain actions are methods to enhance the trust between the model and the user. Utilising common algorithms, FCMs can translate their predictions to any format. A major advantage of FCMs, which aids in the post hoc interpretability of a whole system, is their ability to play a unifying role. In other words, machine learning predictions, algorithms and rules can be parts of a bigger system. Those parts shall be united under the FCM, which makes the final decision based on the accuracies of the classifiers and perhaps some other non-trainable

parameters. This way, a non-interpretable classification method is embedded into an interpretable system that reduces ambiguity and is user-friendly.

References

1. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016).
2. Asan, O., Choudhury, A.: Research Trends in Artificial Intelligence Applications in Human Factors Health Care: Mapping Review. JMIR Hum Factors. 8, e28236 (2021). https://doi.org/10.2196/28236.
3. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy. 23, 18 (2020). https://doi.org/10.3390/e23010018.
4. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 58, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012.
5. Groumpos, P.P.: Intelligence and fuzzy cognitive maps: scientific issues, challenges and opportunities. Studies in Informatics and Control. 27, 247–264 (2018).
6. Papageorgiou, E., Spyridonos, P., Glotsos, D.T., Stylios, C.D., Ravazoula, P., Nikiforidis, G., Groumpos, P.P.: Brain tumor characterisation using the soft computing technique of fuzzy cognitive maps. Applied Soft Computing. 8, 820–828 (2008).
7. Groumpos, P.P., Apostolopoulos, I.D.: Modeling the spread of dangerous pandemics with the utilisation of a hybrid-statistical–Advanced-Fuzzy-Cognitive-Map algorithm: the example of COVID-19. Res. Biomed. Eng. 37, 749–764 (2021). https://doi.org/10.1007/s42600-021-00182-z.
8. Morone, P., Falcone, P.M., Lopolito, A.: How to promote a new and sustainable food consumption model: A fuzzy cognitive map study. Journal of Cleaner Production. 208, 563–574 (2019). https://doi.org/10.1016/j.jclepro.2018.10.075.
9. Papageorgiou, E.I., Papandrianos, N.I., Karagianni, G., Kyriazopoulos, G.C., Sfyras, D.: A fuzzy cognitive map based tool for prediction of infectious diseases. In: 2009 IEEE International Conference on Fuzzy Systems. pp. 2094–2099. IEEE (2009).
10. Goodman, B., Flaxman, S.: European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." AIMag. 38, 50–57 (2017). https://doi.org/10.1609/aimag.v38i3.2741.