Explainable artificial intelligence: problems, questions and applications

Marchenko M.A.
Institute of Computational Mathematics and Mathematical Geophysics of SB RAS,
Novosibirsk, Russia
marchenko@sscc.ru

The explained artificial intelligence (Explainable AI) is a mathematical model that can explain the mechanisms underlying artificial intelligence algorithms (machine learning models). Many solutions that use AI algorithms are like a “black box”. Therefore, often not only end users, but the developers themselves cannot determine exactly how exactly the machine learning model came to one or another conclusions during the processing of the source data. Understanding the work of artificial intelligence algorithms will allow developers to accurately evaluate the influence of the input features on the output result of the model, to identify bias and disadvantages associated with the work of the model, as well as to carry out subtle tuning and optimizing this kind of method. For users, the explanation of the result of the work of artificial intelligence methods is important regarding the understanding of the causes of the conclusions made by the model, and for experts, to explain those conclusions that at first glance have no grounds, for example, when looking for hidden laws in the data.

The report makes an overview of algorithmically sound approaches to the development of artificial intelligence methods, and methods of interpreting the results of their application are described to solve specific applications.

可解釋的人工智能：問題、問題和應用

馬爾琴科・米哈伊爾・亞歷山德羅維奇
Marchenko M.A.
ICM&MG SB RAS 總監
Marchenko@sscc.ru

Explainable AI
是一種數學模型，能夠解釋人工智能算法（機器學習模型）背後的機制。
許多使用人工智能算法的解決方案就像一個“黑匣子”。
因此，不僅最終用戶，而且開發人員本身往往無法準確確定機器學習模型在處理初始數據期間如何得出某些結論。
了解人工智能算法的運行將使開發人員能夠準確評估輸入特徵對模型輸出的影響，識別與模型運行相關的偏差和缺點，以及微調和優化此類方法。
對於用戶來說，人工智能方法工作結果的可解釋性對於理解模型得出結論的原因很重要，對於專家來說，對於解釋那些乍一看沒有根據的結論，例如，在搜索數據中的隱藏模式時。

該報告概述了基於算法的人工智能方法開發方法，描述了解釋其應用結果以解決特定問題的方法。

DOI 10.18699/sblai2023-33

Три определения меры знания

Михайловский Н.
Высшая ИТ-школа ТГУ и ООО «НТР», Томск, Россия
nickm@ntr.ai

Познающие агенты, включая людей, решают проблемы, строя модели. Модель – это объяснение явления, то есть предположение механизма того, как оно может происходить.

При построении модели принимают точку зрения некоторой теории на явление, отбрасывая несущественные детали. Каждая модель имеет ограниченную область применимости. В рамках одной теории можно строить модели множества явлений. Назовем такой тип знания «знанием научного типа» [1].

Определим содержание знаний модели M, объясняющей явление Ф и относящейся к теории $\mathcal{T}$. Ф проявляется как измерения конечной точности $P = \{p_1, p_2, \dots\}$. Будем считать точку зрения $\mathcal{T}$ оператором проецирования $T: p_k \to p_k^t$ и положим, что $p_k^t \in R^l$ и $p_k \in R^n$, $l \le n$. Если ядро T нетривиально и/или $l < n$, в результате часть информации будет утеряна. Модель M предсказывает некоторые измерения p_k^t, используя остальные как входные данные. Для набора индексов входных измерений $I \subset N$ модель M выдает набор приближений $p_o^{M(I)} \cong p_o^t$. Перенормируем измерения и прогнозы и используем безразмерные $\tilde{p}_k^t = p_k^t / p_a$, где p_a – среднее измерение, и обозначим вводимую $\mathcal{T}$ метрику $d(\tilde{p}_1, \tilde{p}_2)$.

Определение 1. Содержание знаний C в модели M явления Ф, наблюдаемого через измерения P с точки зрения теории $\mathcal{T}$, определяется как

$$C = \min_{\varepsilon} \mathbb{E}\left[d\left(\tilde{p}_o^{M(I)}, \tilde{p}_o^t + \varepsilon\right)^{-1}\right], \quad (1)$$

где $\mathbb{E}$ – матожидание и $\varepsilon \in R^l$.

Примем во внимание бритву Оккама в отношении размера модели [2] и дадим

Определение 2. Плотность знаний D в модели M явления Ф определяется как

$$D = \frac{C}{J(M)}, \quad (2)$$

где C – содержание знаний модели M и $J(M)$ – колмогоровская сложность [3] модели M.

Сложность разных явлений естественно различается. Чтобы сравнить эффективность моделей разных явлений, надо оценивать сложность явления, не моделируя его. Для этого применим универсальный алгоритм сжатия информации без потерь к измерениям P и сравним сжатие с потерями с помощью модели M со сжатием тех же данных без потерь. Чем лучше модель M, тем меньше будет отношение между размером данных после сжатия M и алгоритмом сжатия общего назначения. Обозначим $\tilde{I}$ минимальный набор индексов входных измерений $P(\tilde{I})$, который позволяет M предсказать все остальные измерения. Обозначим $G(X)$ размер объекта X, сжатого алгоритмом сжатия без потерь G, например, LZW [4].

Определение 3. Знания K модели M явления Ф определяются как:

$$K = \frac{G(P)}{G\left(P(\tilde{I})\right) + G(M)}. \quad (3)$$

Значение K независимо от D и C.

Литература

1. Mikhaylovskiy N.: How Do You Test the Strength of AI? In: 13th International Conference AGI 2020. pp. 257–266.

2. Manin D.Yu. Science in a False Mirror: Lakatos, Feyerabend, Kuhn. In defense of science. (2008) No. 3. (In Russian)

3. Kolmogorov A.N. Three Approaches to the Quantitative Definition of Information. Problems Inform. Transmission. 1965. vol. 1 issue 1, pp. 1–7.

4. Ziv J., Lempel A. Compression of individual sequences via variable-rate coding. IEEE Transactions on Information Theory. 1978. vol. 24, issue 5, pp. 530.

Three Metrics of Knowledge

Mikhaylovskiy N.

Higher IT School of Tomsk State University and NTR Labs, Tomsk, Russia

nickm@ntr.ai

Sentient agents, including humans, solve problems by building models. A model is an explanation of a phenomenon, that is, an assumption of the mechanism of how can it occur. When

building a model, we take a point of view of some theory on the phenomenon, discarding irrelevant details. Each model has a limited domain of applicability. Within one theory, it is possible to build models of multiple phenomena. We call such type of knowledge a "Scientific-type knowledge" [1].

We want to measure the knowledge content of a model M explaining a phenomenon Φ and pertaining to a theory $\mathcal{T}$. Φ manifests itself as finite precision measurements $P = \{p_1, p_2, \ldots\}$. We consider the viewpoint of $\mathcal{T}$ to be a projection operator $T: p_k \to p_k^t$ and assume that $p_k^t \in R^l$ and $p_k \in R^n$, $l \leq n$. Some information is lost as a result of this projection if the kernel of T is non-trivial and/or $l < n$. The model M can predict some of the measurements p_k^t, using some others as an input. For a set of indices of input measurements $I \subset N$ the model M produces a set of output measurement approximations $p_o^{M(I)} \cong p_o^t$. We rescale the measurements and predictions and use dimensionless $\tilde{p}_k^t = p_k^t / p_a$, where p_a is an average measurement and assume $\mathcal{T}$ implies a distance $d(\tilde{p}_1, \tilde{p}_2)$.

Definition 1. Knowledge content C of a model M of the phenomenon Φ observed through measurements p_k from a viewpoint of a theory $\mathcal{T}$ is:

$$C = \min_{\varepsilon} \mathbb{E}\left[d\left(\tilde{p}_o^{M(I)}, \tilde{p}_o^t + \varepsilon\right)^{-1}\right], \tag{1}$$

where $\mathbb{E}$ is the expectation and $\varepsilon \in R^l$.

We take into account Occam's razor, especially as applied to the model size [2] and give

Definition 2. Knowledge density D of a model M of the phenomenon Φ is:

$$D = \frac{C}{J(M)}, \tag{2}$$

where C is the knowledge content of M and $J(M)$ is the Kolmogorov complexity [3] of M.

Some phenomena are naturally more complex than others are. To compare the efficiency of models of different phenomena, we want to measure the complexity of a phenomenon without specifically modelling it. To do so one can apply a general-purpose, lossless information compression algorithm to the measurements P and compare the lossy compression by model M with lossless compression of the same data. The better the model M is, the smaller will be the ratio between the sizes of data compressed by M and a general-purpose compression algorithm. Let's denote $\tilde{I}$ a minimal set of indices of input measurements that allows M to produce outputs for all the other measurements, and $P(\tilde{I})$ the respective measurements. Let's denote $G(X)$ the size of object X compressed by a general-purpose lossless compression algorithm G, such as LZW [4].

Definition 3. Knowledge K of a model M of the phenomenon Φ is:

$$K = \frac{G(P)}{G\left(P(\tilde{I})\right) + G(M)}. \tag{3}$$

K defines a quantity that is independent from D and C.

References

1. Mikhaylovskiy N.: How Do You Test the Strength of AI? In: 13th International Conference AGI 2020. pp. 257–266.

2. Manin D.Yu. Science in a False Mirror: Lakatos, Feyerabend, Kuhn. In defense of science. (2008) No. 3. (In Russian)

3. Kolmogorov A.N. Three Approaches to the Quantitative Definition of Information. Problems Inform. Transmission. 1965. vol. 1 issue 1, pp. 1–7.

4. Ziv J., Lempel A. Compression of individual sequences via variable-rate coding. IEEE Transactions on Information Theory. 1978. vol. 24, issue 5, pp. 530.

知識的三個指標

Mikhaylovskiy N.

托木斯克國立大學高等 IT 學院和 NTR 實驗室，俄羅斯托木斯克

nickm@ntr.ai

包括人類在內的有感知能力的主體通過建立模型來解決問題。
模型是對現象的解釋，即對現像如何發生的機制的假設。
在建立模型時，我們從現象的某些理論角度出發，丟棄不相關的細節。
每個模型都有一個有限的適用範圍。 在一個理論中，可以建立多種現象的模型。
我們稱這類知識為“科學型知識”[1]。
我們想要測量解釋現象 M 並與理論 Φ 相關的模型 $\mathcal{T}$ 的知識內容。 Φ 表現為有限精度測量 $P = \{p_1, p_2, \ldots\}$ 我們考慮的觀點 $\mathcal{T}$ 成為一名投影操作員 $T: p_k \to p_k^t$ 並假設 $p_k^t \in R^l$ 和 $p_k \in R^n, l \leq n$. 如果內核 of T 是不平凡的和/或 $l < n$. 該模型 M 可以預測一些測量值 p_k^t, 使用其他一些作為輸入。 對於一組輸入測量指標 $I \subset N$ 該模型 M 產生一組輸出測量近似值 $p_o^{M(I)} \cong p_o^t$. 我們重新調整測量和預測並使用無量綱 $\tilde{p}_k^t = p_k^t / p_a$, 在哪裡 p_a 是平均測量值並假設 $\mathcal{T}$ 意味著距離 $d(\tilde{p}_1, \tilde{p}_2)$.

定義 1. 知識內容 C 模型的 M 現象的 Φ 通過測量觀察 p_k 從理論的角度 $\mathcal{T}$ 是：

$$C = \min_{\varepsilon} \mathbb{E}\left[d\left(\tilde{p}_o^{M(I)}, \tilde{p}_o^t + \varepsilon\right)^{-1}\right] \tag{1}$$

其中 $\mathbb{E}$ 是期望，$\varepsilon \in R^l$.
我們考慮了奧卡姆剃刀原則，尤其是應用於模型尺寸 [2] 時，給出
定義 2. 知識密度 D 模型的 M 現象的 Φ 是：

$$D = \frac{C}{J(M)} \tag{2}$$

其中 C 是 M的知識內容，$J(M$ 是 M 的 Kolmogorov 的複雜性 [3] 的 M.
有些現象自然比其他現象更複雜。
為了比較不同現像模型的效率，我們希望在不專門建模的情況下測量現象的複雜性。
為此，可以對測量值 P應用通用的無損信息壓縮算法，並將模型 M 的有損壓縮與相同數據的無損壓縮進行比較。 模型 M 越好，M 壓縮後的數據大小與通用壓縮算法的比值就越小。 讓我們表示 $\tilde{I}$ 一組最小的輸入測量指標，它允許 M 為所有其他測量產生輸出，而 $P(\tilde{I})$ 表示相應的測量。 讓我們用通用無損壓縮算法 G 壓縮對象 X 的大小來表示 $G(X)$，例如 LZW [4]。
定義 3. 現象 K 的模型 M 的知識 Φ 是：

$$K = \frac{G(P)}{G\left(P(\tilde{I})\right) + G(M)} \tag{3}$$

K定義了一個獨立於 D 和 C 的量。

文學

1. Mikhaylovskiy N.: How Do You Test the Strength of AI? In: 13th International Conference AGI 2020. pp. 257–266.
2. Manin D.Yu. Science in a False Mirror: Lakatos, Feyerabend, Kuhn. In defense of science. (2008) No. 3. (In Russian)
3. Kolmogorov A.N. Three Approaches to the Quantitative Definition of Information. Problems Inform. Transmission. 1965. vol. 1 issue 1, pp. 1–7.
4. Ziv J., Lempel A. Compression of individual sequences via variable-rate coding. IEEE Transactions on Information Theory. 1978. vol. 24, issue 5, pp. 530.