

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2024 年 10 月 24 日 (24.10.2024)



(10) 国际公布号
WO 2024/217543 A1

(51) 国际专利分类号:
G06N 3/04 (2023.01) **G06N 3/08** (2023.01)
G06F 30/27 (2020.01) **G06F 111/10** (2020.01)
G06N 20/00 (2019.01)

(21) 国际申请号: PCT/CN2024/088822

(22) 国际申请日: 2024 年 4 月 19 日 (19.04.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202310431810.1 2023年4月20日 (20.04.2023) CN

(71) 申请人: 上海交通大学 (SHANGHAI JIAO TONG UNIVERSITY) [CN/CN]; 中国上海市闵行区东川路800号, Shanghai 200240 (CN)。

(72) 发明人: 张拳石 (ZHANG, Quanshi); 中国上海市闵行区东川路800号, Shanghai 200240 (CN)。

(74) 代理人: 上海一平知识产权代理有限公司 (XU & PARTNERS, LLC); 中国上海市普陀区真北路 958 号天地科技广场 1 号楼 106 室, Shanghai 200333 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD,

(54) Title: METHOD FOR DETERMINING SPARSE INTERACTION EFFECT OF BLACK-BOX ARTIFICIAL INTELLIGENCE MODEL

(54) 发明名称: 一种测定黑盒人工智能模型的稀疏交互效用的方法

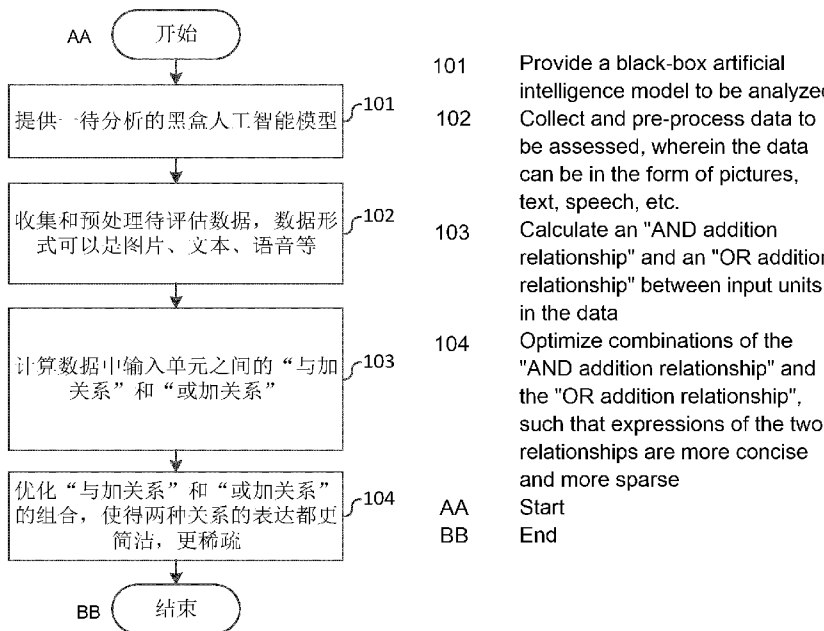


图 1

(57) Abstract: The present application relates to the technical field of machine learning. Disclosed are a method and system for interpreting a sparse interaction effect modeled by a black-box artificial intelligence model. The method and system can automatically analyze an interactive distribution modeled by a model. The implementation of the method and system comprises the following steps: providing data that needs to be assessed; using a black-box model to perform prediction on the data, so as to obtain a prediction result of the model; on the basis of an output of the black-box model, modeling the interaction effect between input units of samples, calculating the interaction intensity between combinations formed by the input units, and expressing the black-box model as an "AND addition

SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ,
UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚
(AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE,
BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR,
HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO,
PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF,
CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN,
TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

relationship" and an "OR addition relationship" between the combinations of the input units; and performing optimization, such that the "AND addition relationship" and the "OR addition relationship" are more sparse. The advantages of the present invention lie in that a quantification method for interpreting the interaction modeled by a black-box artificial intelligence model is provided, and in comparison with previous research, a more sparse and concise interactive interpretation can be obtained.

(57) 摘要: 本申请涉及机器学习技术领域, 公开了一种解释黑盒人工智能模型所建模的稀疏交互效用的方法和系统。该方法和系统可以自动地分析模型所建模的交互分布。该方法和系统的实施包括以下步骤: 提供需要评估的数据; 使用黑盒模型对数据进行预测, 获得模型的预测结果; 基于黑盒模型的输出, 对样本的输入单元间的交互作用进行建模, 计算输入单元间形成的组合的交互强度, 将黑盒模型表达成输入单元组合间的"与加关系"和"或加关系"; 通过优化, 使得"与加关系"和"或加关系"更加稀疏。本发明的优点在于, 提供了一种解释黑盒人工智能模型所建模的交互的量化方法, 相比于前人研究, 可以获得更稀疏、简洁的交互解释。

一种测定黑盒人工智能模型的稀疏交互效用的方法

技术领域

5 本发明涉及机器学习技术领域，具体涉及一种解释黑盒人工智能模型所建模的稀疏交互效用的方法和系统。

背景技术

10 目前，深度学习已在各个领域得到广泛应用，并展现出强大性能，但使用者与研究者很难分析人工智能模型的黑盒本质。在现有的技术中，基于交互的解释方法能够在保证可靠性的前提下，定量分析模型输入中各个单元、属性组合的产生的交互效用，但无法保证解释结果的稀疏性、简洁性，而稀疏、简洁的解释结果恰是人工智能可解释性方法的必要条件。

因此，如何可靠地解释黑盒模型所建模的交互效用，并同时保证其稀疏性、简洁性，是可解释性领域亟待解决的问题。

15

发明内容

本发明的目的在于发明一种解释黑盒人工智能模型所建模的稀疏交互效用的方法和系统，可以自动地分析模型所建模的交互分布。该方法包括以下步骤：

20

(1) 选取模型

选取一待分析的黑盒人工智能模型，该黑盒人工智能模型为在某一数据集上预先训练好的模型。

(2) 获取需要进行评估的数据

25 收集和预处理待评估数据，数据形式可以是表格数据、图片、文本、语音等，但需与步骤（1）中的模型兼容。所述待评估数据中包含一定数量的输入单元。

(3) 建模输入单元间的“与加关系”和“或加关系”

30 将步骤（2）中待评估的数据输入步骤（1）中的模型，并获取模型输出。基于黑盒模型的输出，对输入单元间的交互配合作用进行建模，获得由不同输入单元组成的集合，并计算每个集合的交互强度。将黑盒模型表达成输入单元组合间的“与加关系”和“或加关系”的组合。

(4) 优化“与加关系”和“或加关系”的组合

基于步骤（3）中所述的“与加关系”和“或加关系”，通过优化“与加关系”和“或加关系”的组合，使得两种关系的表达都更简洁，更稀疏。

35

在一个优选例中，所述步骤（1）进一步包括以下步骤：

基于某一数据集，训练一黑盒人工智能模型，作为待分析的模型。可选地，该模型可以为一个深度神经网络。

在一个优选例中，所述步骤（2）进一步包括以下子步骤：

（a）收集待分析数据。可选地，如果步骤（1）中的模型可以用于图像分类，则待分析数据可以为一张图片。

（b）预处理待分析数据。可选地，可将待分析数据作降采样或升采样、归一化等处理，使其适应上述模型输入。

在一个优选例中，所述步骤（3）进一步包括以下子步骤：

- （3a）计算上述待分析数据在任意遮挡的情况下，黑盒模型的输出值；
- （3b）计算黑盒模型所建模待分析数据中的“与交互”；
- （3c）计算黑盒模型所建模待分析数据中的“或交互”。

上述子步骤（3a）中，所述黑盒人工智能模型表示为 v ，所待分析数据样本 x 中包含 n 个输入单元，表示为集合 $N = \{1, 2, \dots, n\}$ ，其中，输入单元的个数一般为大于等于5的正整数。可选地，可将输入数据中的每一维视作一个输入单元，也可以将输入数据划分为多个区域，将每个区域视作一个输入单元。

将输入单元的任意子集 $S \subseteq N$ 称为输入单元间的组合， $v(S)$ 代表黑盒模型在给定某一输入单元间的组合 S ，并将在 $N \setminus S$ 中的输入单元遮挡时模型的输出。特别地，如果将待分析数据不加遮挡地输入模型，获得的输出表示为 $v(N)$ ，如果将待分析数据完全遮挡输入模型，获得的输出表示为 $v(\emptyset)$ 。

计算某一输入单元组合 S 上黑盒模型的输出 $v(S)$ 时，需要在 S 包含的输入单元上保持待分析样本原本的输入值，而将所有 S 的补集 $N \setminus S$ 中的输入单元替换为基准值。特别地，假设基准值为 b ，遮挡后的样本 x_S 的各个输入单元由下式定义：

$$(x_S)_i = \begin{cases} x_i & \text{如果输入的第 } i \text{ 个输入单元被输入单元组合 } S \text{ 所包含} \\ b_i & \text{如果输入的第 } i \text{ 个输入单元未被输入单元组合 } S \text{ 所包含} \end{cases}$$

这样， $v(S)$ 可以被计算为将样本 x_S 输入模型得到的输出值。可选地，基准值可以设置为样本集的均值，随机值，零值等。进一步，可将基于基准值 b 计算得到的黑盒模型输出 $v(S)$ 具体地记作 $v(S|b)$ ，应理解，在不产生歧义的情况下，可以将 $v(S|b)$ 简写为 $v(S)$ 。

上述子步骤（3b）中，通过计算“与交互”来评测任意输入单元组合之间的与关系对于黑盒模型输出的贡献值，“与交互”表示组合中的单元同时触发（即不被遮挡）时，对于黑盒模型输出产生的额外效用。应理解，在黑盒模型中，输入单元往往并不是单独地贡献于模型的输出，而是彼此交互配合影响模型的输出。

因此, 对于 $S \subseteq N$, 定义 $I^{AND}(S)$ 为输入单元间组合 S 对应的“与交互”, 使得当 S 中所有的输入变量都触发时, $I^{AND}(S)$ 才被触发, 并被累加到黑盒模型的输出中。具体地, $I^{AND}(S)$ 可以由如下公式计算得到。

$$I^{AND}(S) = \sum_{L \subseteq S} (-1)^{s-l} v(L)$$

其中, $s = |S|$ 且 $l = |L|$ 。可选地, 可以基于步骤 (2) 中的不同基准值计算得到不同的“与交互”值, 将基于基准值 b 计算得到的与交互记作 $I^{AND}(S|b)$ 。应理解, 在不产生歧义的情况下, 可以将 $I^{AND}(S|b)$ 简写为 $I^{AND}(S)$ 。

上述“与交互”满足, 对于任意给定的输入单元组合 $T \subseteq N$, 黑盒模型在这一输入单元组合下的输出 $v(T)$ 可以拆分为所有触发的与交互 $I^{AND}(S)$ 之和, 如下式所示。

$$\forall T \subseteq N, v(T) = \sum_{S \subseteq N} I^{AND}(S) \left[\bigwedge_{i \in S} (e_T)_i \right] \Leftrightarrow v(T) = I^{AND}(\emptyset) + \sum_{\emptyset \neq S \subseteq T} I^{AND}(S)$$

其中, e_T 满足, 当 $i \in T$ 时, $(e_T)_i = 1$, 否则, $(e_T)_i = 0$; $\bigwedge_{i \in S} (e_T)_i$ 表示输入单元组合 S 中的单元是否全部触发, 对 $\bigwedge$ 符号后的元素做“与”操作得到的结果, 即 $\bigwedge_{i \in S} (e_T)_i = \prod_{i \in S} (e_T)_i$ 。

进一步, 上述与交互的计算公式可等价如下形式, 假设 $\mathbf{v} = [v(S_1), v(S_2), \dots, v(S_{2^n})] \in \mathbb{R}^{2^n}$, $\mathbf{I}^{AND} = [I^{AND}(S_1), I^{AND}(S_2), \dots, I^{AND}(S_{2^n})] \in \mathbb{R}^{2^n}$, 则存在 $\mathbf{A} \in \{0, 1, -1\}^{2^n \times 2^n}$, 使得 $\mathbf{I}^{AND} = \mathbf{A}\mathbf{v}$ 。

上述子步骤 (3c) 中, 通过计算“或交互”来评测任意输入单元组合之间的或关系对于黑盒模型输出的贡献值, “或交互”表示组合中的单元至少有一个触发 (即不被遮挡) 时, 对于黑盒模型输出产生的额外效用。应理解, 在黑盒模型中, 输入单元往往并不是单独地贡献于模型的输出, 而是彼此交互配合影响模型的输出。

因此, 对于 $S \subseteq N$, 定义 $I^{OR}(S)$ 为输入单元间组合 S 对应的“或交互”, 使得当 S 中所有的输入变量至少有一个触发时, $I^{OR}(S)$ 则触发, 并被累加到黑盒模型的输出中。具体地, $I^{OR}(S)$ 可以由如下公式计算得到。

$$I^{OR}(S) = \begin{cases} v(\emptyset), & S = \emptyset \\ -\sum_{L \subseteq S} (-1)^{s-l} v(N \setminus L), & S \neq \emptyset \end{cases}$$

其中, $s = |S|$ 且 $l = |L|$ 。上述“或交互”满足, 对于任意给定的输入单元组合 $T \subseteq N$, 黑盒模型在这一输入单元组合下的输出 $v(T)$ 可以拆分为所有触发的或交互 $I^{OR}(S)$ 之和, 如下式所示。

$$\forall T \subseteq N, v(T) = \sum_{S \subseteq N} I^{OR}(S) \left[\bigvee_{i \in S} (e_T)_i \right] \Leftrightarrow v(T) = I^{OR}(\emptyset) + \sum_{S: S \cap T \neq \emptyset} I^{OR}(S)$$

其中, e_T 满足, 当 $i \in T$ 时, $(e_T)_i = 1$, 否则, $(e_T)_i = 0$; $\bigvee_{i \in S} (e_T)_i$ 表示输入单元组合 S 中的单元是否至少有一个被触发, 对 $\vee$ 符号后的元素做“或”操作得到的结果, 即 $\bigvee_{i \in S} (e_T)_i = 1 - \prod_{i \in S} [1 - (e_T)_i]$ 。

进一步, 上述或交互的计算公式可等价如下形式, 假设 $\mathbf{v} = [v(S_1), v(S_2), \dots, v(S_{2^n})] \in \mathbb{R}^{2^n}$, $\mathbf{I}^{OR} = [I^{OR}(S_1), I^{OR}(S_2), \dots, I^{OR}(S_{2^n})] \in \mathbb{R}^{2^n}$, 则存在 $\mathbf{B} \in \{0, 1, -1\}^{2^n \times 2^n}$, 使得 $\mathbf{I}^{OR} = \mathbf{B}\mathbf{v}$ 。

在一个优选例中, 所述步骤 (4) 进一步包括以下子步骤:

(4a) 基于步骤 (3) 中的“与交互”和“或交互”, 将黑盒模型输出解释成输入单元组合间的“与加关系”和“或加关系”的组合。

(4b) 基于子步骤 (4a) 中“与加关系”和“或加关系”的组合, 进一步优化“与加关系”和“或加关系”的组合, 使得两种关系的表达更简洁, 更稀疏。

在子步骤 (4a) 中, 基于步骤 (3) 中的“与交互”和“或交互”, 将黑盒模型在任意输入单元组合 T 下的输出 $v(T)$ 解释为“与加关系”和“或加关系”的组合。可选地, 基于步骤 (3) 中不同的基准值, 可将黑盒模型的输出 $v(T)$ 解释为“基于不同基准值的与加关系”和“或加关系”的组合。一般地, 假设有 m 种不同的基准值 $b^{(1)}, \dots, b^{(m)}$, 则黑盒模型在任意输入单元组合下的输出 $v(T)$ 可以解释为下式。

$$\begin{aligned} \forall T \subseteq N, v(T) = & \hat{I}^{AND}(\emptyset | b^{(1)}) + \sum_{\emptyset \neq S \subseteq N} \hat{I}^{AND}(S | b^{(1)}) \left[\bigwedge_{i \in S} (e_T)_i \right] \\ & + \dots \\ & + \hat{I}^{AND}(\emptyset | b^{(m)}) + \sum_{\emptyset \neq S \subseteq N} \hat{I}^{AND}(S | b^{(m)}) \left[\bigwedge_{i \in S} (e_T)_i \right] \\ & + \hat{I}^{OR}(\emptyset) + \sum_{\emptyset \neq S \subseteq N} \hat{I}^{OR}(S) \left[\bigvee_{i \in S} (e_T)_i \right] \end{aligned}$$

因此, 基于上述对黑盒模型的解释, 可将模型在完整样本上的输出 $v(N)$ 的拆分为 $v(N) = \sum_{S \subseteq N} \hat{I}^{AND}(S | b^{(1)}) + \dots + \sum_{S \subseteq N} \hat{I}^{AND}(S | b^{(m)}) + \sum_{S \subseteq N} \hat{I}^{OR}(S)$ 。其中, 黑盒模型输出 $v(N)$ 被解释为 m 组“与交互”的效用之和, 其中每组“与交互”包含 2^n 个输入单元间组合内的“与交互”效用 $\hat{I}^{AND}(S | b)$, $S \subseteq N$; 以及一组“或交互”的效用之和, 包含 2^n 个输入单元间组合内的“或交互”效用 $\hat{I}^{OR}(S)$, $S \subseteq N$ 。可选地, 上述 m 种不同的基准值可

以选取为均值、随机值、零值等。

在子步骤（4b）中，为了进一步得到稀疏的交互，本发明希望最小化所有交互的 L_1 -范数之和，即

$$\min_{\hat{\mathbf{I}}_b^{AND}, \dots, \hat{\mathbf{I}}_{b(m)}^{AND}, \hat{\mathbf{I}}^{OR}} \|\hat{\mathbf{I}}_{b(1)}^{AND}\|_1 + \dots + \|\hat{\mathbf{I}}_{b(m)}^{AND}\|_1 + \|\hat{\mathbf{I}}^{OR}\|_1$$

其中， $\hat{\mathbf{I}}_b^{AND} = [\hat{\mathbf{I}}_b^{AND}(S_1|b), \hat{\mathbf{I}}_b^{AND}(S_2|b), \dots, \hat{\mathbf{I}}_b^{AND}(S_{2^n}|b)] \in \mathbb{R}^{2^n}$ ，
5 $\hat{\mathbf{I}}^{OR} = [\hat{\mathbf{I}}^{OR}(S_1), \hat{\mathbf{I}}^{OR}(S_2), \dots, \hat{\mathbf{I}}^{OR}(S_{2^n})] \in \mathbb{R}^{2^n}$ ， $\|\cdot\|_1$ 表示向量的 L_1 -范数，上式中的 m 即前述基准值的种数。

可选地，在优化上述表达式时，可以定义 $\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(m)}, \mathbf{p}^{(or)} \in \mathbb{R}^{2^n}$ ，并满足 $\mathbf{p}^{(1)} + \dots + \mathbf{p}^{(m)} + \mathbf{p}^{(or)} = \mathbf{0}$ ，这里的 $\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(m)}, \mathbf{p}^{(or)}$ 表示对于神经网络输出 $\mathbf{v}$ 的一种划分；并定义 $\mathbf{q} \in \mathbb{R}^{2^n}$ ， $\mathbf{q}$ 为一个各维绝对值较小的向量，表示将模型输出解释为“与加关系”和“或加关系”所产生的偏差。使得
10 $\hat{\mathbf{I}}_{b(1)}^{AND} = A\left(\frac{1}{m+1}(\mathbf{v} + \mathbf{q}) + \mathbf{p}^{(1)}\right), \dots, \hat{\mathbf{I}}_{b(m)}^{AND} = A\left(\frac{1}{m+1}(\mathbf{v} + \mathbf{q}) - \mathbf{p}^{(m)}\right)$ ，
 $\hat{\mathbf{I}}^{OR} = B\left(\frac{1}{m+1}(\mathbf{v} + \mathbf{q}) + \mathbf{p}^{(or)}\right)$ 。则上述优化问题可以转换为如下形式。

$$\min_{\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(m)}, \mathbf{p}^{(or)}, \mathbf{q}} \|\hat{\mathbf{I}}_{b(1)}^{AND}\|_1 + \dots + \|\hat{\mathbf{I}}_{b(m)}^{AND}\|_1 + \|\hat{\mathbf{I}}^{OR}\|_1$$

s. t. $|\mathbf{q}_i| < \tau_i$

其中， τ_i 为预设的阈值。优选地， $\tau_i = 0.05 \times |\mathbf{v}(N) - \mathbf{v}(\emptyset)|$ ， m 的取值范围一般为1至10。

特别地，在一个实施例中，在上述子步骤（a）中，仅选取一种基准值，即 $m = 1$ 时，可将黑盒模型的输出解释为下式所示的“与加关系”和“或加关系”的组合。

$$\forall T \subseteq N, \mathbf{v}(T) = \hat{\mathbf{I}}^{AND}(\emptyset) + \sum_{\emptyset \neq S \subseteq N} \hat{\mathbf{I}}^{AND}(S) \left[\bigwedge_{i \in S} (e_T)_i \right] + \hat{\mathbf{I}}^{OR}(\emptyset) + \sum_{\emptyset \neq S \subseteq N} \hat{\mathbf{I}}^{OR}(S) \left[\bigvee_{i \in S} (e_T)_i \right]$$

其中， $\hat{\mathbf{I}}^{AND}(S)$ 表示输入单元组合 S 的“与交互”， $\hat{\mathbf{I}}^{OR}(S)$ 表示输入单元组合 S 的“或交互”。同样地，黑盒模型输出被解释为 2^n 个输入单元间组合“与交互”的效用之和，以及 2^n 个输入单元间组合“或交互”的效用之和。可选地，上述基准值可以选取为均值、随机值、零值等中的一种。

进一步，在子步骤（b）中，为了进一步得到稀疏的交互，本发明希望最小化所有交互的 L_1 -范数之和，即

$$\min_{\hat{\mathbf{I}}^{AND}, \hat{\mathbf{I}}^{OR}} \|\hat{\mathbf{I}}^{AND}\|_1 + \|\hat{\mathbf{I}}^{OR}\|_1$$

25 其中， $\hat{\mathbf{I}}^{AND} = [\hat{\mathbf{I}}^{AND}(S_1), \hat{\mathbf{I}}^{AND}(S_2), \dots, \hat{\mathbf{I}}^{AND}(S_{2^n})] \in \mathbb{R}^{2^n}$ ， $\hat{\mathbf{I}}^{OR} =$

$[\hat{I}^{OR}(S_1), \hat{I}^{OR}(S_2), \dots, \hat{I}^{OR}(S_{2^n})] \in \mathbb{R}^{2^n}$, $\|\cdot\|_1$ 表示向量的 L_1 -范数。

可选地, 在优化上述表达式时, 可以定义 $\mathbf{p} \in \mathbb{R}^{2^n}$, $\mathbf{q} \in \mathbb{R}^{2^n}$ 。使得 $\hat{\mathbf{I}}^{AND} = A\left(\frac{1}{2}(\mathbf{v} + \mathbf{q}) - \mathbf{p}\right)$, $\hat{\mathbf{I}}^{OR} = B\left(\frac{1}{2}(\mathbf{v} + \mathbf{q}) + \mathbf{p}\right)$ 。则上述优化问题可以转换为如下形式。

$$\begin{aligned} \min_{\mathbf{p}, \mathbf{q}} \quad & \|\hat{\mathbf{I}}^{AND}\|_1 + \|\hat{\mathbf{I}}^{OR}\|_1 \\ \text{s. t.} \quad & |\mathbf{q}_i| < \tau_i \end{aligned}$$

5 其中, τ_i 为预设的阈值。优选地, $\tau_S = 0.05 \times |v(N) - v(\emptyset)|$ 。

在一个优选例中, 上述步骤(3)和步骤(4)中对于输入单元间组合的标号可以进一步包括以下步骤:

10 输入样本 N 对应的输出 $v(N)$ 可以建模 2^n 个输入单元间组合内的与交互作用 I^{AND} 和 2^n 个输入单元间组合内的或交互作用 I^{OR} , 用如下方式对 2^n 个输入单元间组合进行标号。例如, 当 $n = 3$ 时, 设 $N = \{i, j, k\}$, 优选地, 可以用表一所示, 标记所有输入单元间组合。

表一

标号	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
输入单元组合	$\{\emptyset\}$	$\{i\}$	$\{j\}$	$\{k\}$	$\{i, j\}$	$\{i, k\}$	$\{j, k\}$	$\{i, j, k\}$

15 例如, 当 $n = 5$ 时, 设 $N = \{i, j, k, l, m\}$, 优选地, 可以用表二所示, 标记所有输入单元间组合。

表二

标号	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
输入单元组合	$\{\emptyset\}$	$\{i\}$	$\{j\}$	$\{k\}$	$\{l\}$	$\{m\}$	$\{i, j\}$	$\{i, k\}$
标号	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}
输入单元组合	$\{i, l\}$	$\{i, m\}$	$\{j, k\}$	$\{j, l\}$	$\{j, m\}$	$\{k, l\}$	$\{k, m\}$	$\{l, m\}$

标 号	S_{17}	S_{18}	S_{19}	S_{20}	S_{21}	S_{22}	S_{23}	S_{24}
输入单元组合	$\{i, j, k\}$	$\{i, j, l\}$	$\{i, j, m\}$	$\{i, k, l\}$	$\{i, k, m\}$	$\{i, l, m\}$	$\{j, k, l\}$	$\{j, k, m\}$
标 号	S_{25}	S_{26}	S_{27}	S_{28}	S_{29}	S_{30}	S_{31}	S_{32}
输入单元组合	$\{j, l, m\}$	$\{k, l, m\}$	$\{i, j, k, l\}$	$\{i, j, k, m\}$	$\{i, j, l, m\}$	$\{i, k, l, m\}$	$\{j, k, l, m\}$	$\{i, j, k, l, m\}$

本发明的第二方面，提供了一种解释黑盒人工智能模型所建模的稀疏交互效用的系统，其特征在于，包括：

5 (1) 输入模块，被配置为一个预训练好的黑盒人工智能模型和一组待分析数据。

(2) 计算模块，被配置为，基于输入模块中的模型和数据，计算模型所建模数据输入单元间的“与加关系”和“或加关系”，并优化“与加关系”和“或加关系”的组合，使得两种关系的表达都更简洁，更稀疏。

10 (3) 输出模块，被配置为，输出计算模块优化得到的“与加关系”和“或加关系”的组合。

应理解，在本发明范围内中，本发明的上述各技术特征和在下文(如实施例)中具体描述的各技术特征之间都可以互相组合，从而构成新的或优选的技术方案。限于篇幅，在此不再一一累述。

15 通过阅读参照以下附图对非限制性实施例所作的详细描述，本发明的其他特征、目的和优点将会变得更加明显。

本发明的主要优点在于：

(1) 本发明提出了一种能够自动分析黑盒模型内部“与加关系”与“或加关系”的方法和系统；

20 (2) 本发明所公开的方法和系统能够保证解释结果的严谨性，即，基于本发明中提出的“与交互”和“或交互”，能够实现将黑盒模型输出精确拆分为不同输入单元之间的交互效用之和；

(3) 本发明所公开的方法和系统能够保证解释结果的简洁性，即，

基于本发明中提出的对“与加关系”和“或加关系”的优化，能够得到更稀疏、简洁的交互解释；

(4) 本发明所公开的方法和系统具有广泛的普适性和一般性，即，适用于大部分黑盒模型，包括但不限于深度神经网络等。

5 本申请的说明书中记载了大量的技术特征，分布在各个技术方案中，如果要罗列出本申请所有可能的技术特征的组合（即技术方案）的话，会使得说明书过于冗长。为了避免这个问题，本申请上述发明内容中公开的各个技术特征、在下文各个实施方式和例子中公开的各技术特征、以及附图中公开的各个技术特征，都可以自由地互相组合，从而构成各种新的技术方案（这些技术方案均因视为在本说明书中已经记载），除非这种技术特征的组合在技术上是不可行的。例如，在一个例子中公开了特征 $A+B+C$ ，在另一个例子中公开了特征 $A+B+D+E$ ，而特征 C 和 D 是起到相同作用的等同技术手段，技术上只要择一使用即可，不可能同时采用，特征 E 技术上可以与特征 C 相组合，则， $A+B+C+D$ 的方案因技术不可行而应当不被视为已经记载，而 $A+B+C+E$ 的方案应当视为已经被记载。

10

15

附图说明

下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

20

图 1 是根据本发明第一实施方式的解释黑盒人工智能模型所建模的稀疏交互效用方法的流程示意图；

图 2 中的 (a) 是本发明所述的待分析样本和输入变量的示意图，图 2 的 (b) 是根据本发明得到的“与加关系”和“或加关系”的初始效用分布，图 2 的 (c) 是根据本发明优化得到的“与加关系”和“或加关系”的最终效用分布，该结果基于表格数据；

25

图 3 是根据本发明优化得到的部分“与交互”和“或交互”的效用值，该结果基于表格数据。

图 4 的 (a) 是本发明所述的待分析样本和输入变量的示意图，图 4 的 (b) 是根据本发明得到的“与加关系”和“或加关系”的初始效用分布，图 4 的 (c) 是根据本发明优化得到的“与加关系”和“或加关系”的最终效用分布，该结果基于图像数据；

30

图 5 是根据本发明优化得到的部分“与交互”和“或交互”的效用值，该结果基于图像数据。

35

图 6 是根据本发明第二实施方式的解释黑盒人工智能模型所建模的稀疏交互效用系统的结构示意图。

5 具体实施方式

本发明人经过细致深入的研究，首次开发了一种解释黑盒人工智能模型所建模的稀疏交互效用的方法和系统。

通用方法

10 典型地，本发明包括以下步骤：

(1) 选取模型。选取一待分析的黑盒人工智能模型，该黑盒人工智能模型为在某一数据集上预先训练好的模型。

(2) 获取需要进行评估的数据。收集和预处理待评估数据，数据形式可以是表格数据、图片、文本、语音等，但需与步骤 (1) 中的模型兼容。
15 所述待评估数据中包含一定数量的输入单元。

(3) 建模输入单元间的“与加关系”和“或加关系”。将步骤 (2) 中待评估的数据输入步骤 (1) 中的模型，并获取模型输出。基于黑盒模型的输出，对输入单元间的交互配合作用进行建模，获得由不同输入单元组成的集合，并计算每个集合的交互强度。将黑盒模型表达成输入单元组合间的“与加关系”和“或加关系”的组合。
20

(4) 优化“与加关系”和“或加关系”的组合。基于步骤 (3) 中所述的“与加关系”和“或加关系”，通过优化“与加关系”和“或加关系”的组合，使得两种关系的表达都更简洁，更稀疏。

实施例

25 为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请的实施方式作进一步地详细描述。

本发明的第一实施方式涉及一种解释黑盒人工智能模型所建模的稀疏交互效用的方法，其流程如图 1 所示，该方法包括以下步骤：

在步骤 101 中：基于某一数据集，训练一黑盒人工智能模型，作为待分析的模型。可选地，该模型可以为一个深度神经网络。在一个实施例中，该模型选定为在 UCI Wireless Indoor Localization 数据集上预训练的五层感知机模型 (MLP-5)，为一个深度神经网络；在另一实施例中，该模型选定为在 CelebA 数据集上预训练的 ResNet-18 模型，也为一个深度神经网络。
30

35 之后，进入步骤 102，该步骤可以进一步分为如下两个子步骤：

(a) 收集待分析数据。可选地，如果步骤 (1) 中的模型可以用于图像分类，则待分析数据可以为一张图片。

(b) 预处理待分析数据。可选地，可将待分析数据作降采样或升采样、归一化等处理，使其适应上述模型输入。

5 在一个实施例中，如图 2(左)所示，该数据选定为 UCI Wireless Indoor Localization 数据集中的任一数据，包含 7 个输入变量，分别表示第 1 至 7 个 wifi 信号强度值，记做信号 1，信号 2，…，信号 7；在另一实施例中，如图 4(左)所示，该模型选定为在 CelebA 数据集中的任一数据，划分为 9 个输入变量，即头发、额头、左眼、右眼、鼻子、左脸颊、右脸颊、嘴、
10 下巴。

之后，进入步骤 103，该步骤可以进一步分为如下三个子步骤：

(a) 计算上述待分析数据在任意遮挡的情况下，黑盒模型的输出值；

(b) 计算黑盒模型所建模待分析数据中的“与交互”；

(c) 计算黑盒模型所建模待分析数据中的“或交互”。

15 上述子步骤 (a) 中，所述黑盒人工智能模型表示为 v ，所待分析数据样本 x 中包含 n 个输入单元，表示为集合 $N = \{1, 2, \dots, n\}$ ，其中，输入单元的个数一般为 5 至 20。可选地，可将输入数据中的每一维视作一个输入单元，也可以将输入数据划分为多个区域，将每个区域视作一个输入单元。

将输入单元的任意子集 $S \subseteq N$ 称为输入单元间的组合， $v(S)$ 代表黑盒
20 模型在给定某一输入单元间的组合 S ，并将在 $N \setminus S$ 中的输入单元遮挡时模型的输出。特别地，如果将待分析数据不加遮挡地输入模型，获得的输出表示为 $v(N)$ ，如果将待分析数据完全遮挡输入模型，获得的输出表示为 $v(\emptyset)$ 。

计算某一输入单元组合 S 上黑盒模型的输出 $v(S)$ 时，需要在 S 包含的输入单元上保持待分析样本原本的输入值，而将所有 S 的补集 $N \setminus S$ 中的输入单元
25 替换为基准值。特别地，假设基准值为 b ，遮挡后的样本 x_S 的各维取值由下式定义：

$$(x_S)_i = \begin{cases} x_i & \text{如果输入的第 } i \text{ 维被输入单元组合 } S \text{ 所包含} \\ b_i & \text{如果输入的第 } i \text{ 维未被输入单元组合 } S \text{ 所包含} \end{cases}$$

这样， $v(S)$ 可以被计算为将样本 x_S 输入模型得到的输出值。可选地，基准值可以设置为样本集的均值，随机值，零值等。进一步，可将基于基准值 b 计算得到的黑盒模型输出 $v(S)$ 具体地记作 $v(S|b)$ ，应理解，在不产生歧义的情况下，可以将 $v(S|b)$ 简写为 $v(S)$ 。在一个实施例中，基准值 b 设置
30 为样本集不同样本的均值。

上述子步骤 (b) 中，通过计算“与交互”来评测任意输入单元组合之间的与关系对于黑盒模型输出的贡献值，“与交互”表示组合中的单元

同时触发（即不被遮挡）时，对于黑盒模型输出产生的额外效用。应理解，在黑盒模型中，输入单元往往并不是单独地贡献于模型的输出，而是彼此交互配合影响模型的输出。

因此，对于 $S \subseteq N$ ，定义 $I^{AND}(S)$ 为输入单元间组合 S 对应的“与交互”，使得当 S 中所有的输入变量都触发时， $I^{AND}(S)$ 才被触发，并被累加到黑盒模型的输出中。具体地， $I^{AND}(S)$ 可以由如下公式计算得到。

$$I^{AND}(S) = \sum_{L \subseteq S} (-1)^{s-l} v(L)$$

其中， $s = |S|$ 且 $l = |L|$ 。可选地，可以基于步骤（2）中的不同基准值计算得到不同的“与交互”值，将基于基准值 b 计算得到的与交互记作 $I^{AND}(S|b)$ 。应理解，在不产生歧义的情况下，可以将 $I^{AND}(S|b)$ 简写为 $I^{AND}(S)$ 。

上述“与交互”满足，对于任意给定的输入单元组合 $T \subseteq N$ ，黑盒模型在这一输入单元组合下的输出 $v(T)$ 可以拆分为所有触发的与交互 $I^{AND}(S)$ 之和，如下式所示。

$$\forall T \subseteq N, v(T) = \sum_{S \subseteq N} I^{AND}(S) \bigwedge_{i \in S} (e_T)_i \Leftrightarrow v(T) = I^{AND}(\emptyset) + \sum_{\emptyset \neq S \subseteq T} I^{AND}(S)$$

其中， e_S 满足，当 $i \in S$ 时， $(e_S)_i = 1$ ，否则， $(e_S)_i = 0$ ； $\bigwedge_{i \in S} (e_T)_i$ 表示输入单元组合 S 中的单元是否全部触发，对 $\bigwedge$ 符号后的元素做“与”操作得到的结果。

进一步，上述与交互的计算公式可等价如下形式，假设 $\mathbf{v} = [v(S_1), v(S_2), \dots, v(S_{2^n})] \in \mathbb{R}^{2^n}$ ， $\mathbf{I}^{AND} = [I^{AND}(S_1), I^{AND}(S_2), \dots, I^{AND}(S_{2^n})] \in \mathbb{R}^{2^n}$ ，则存在 $A \in \{0, 1, -1\}^{2^n \times 2^n}$ ，使得 $\mathbf{I}^{AND} = A\mathbf{v}$ 。

上述子步骤（c）中，通过计算“或交互”来评测任意输入单元组合之间的或关系对于黑盒模型输出的贡献值，“或交互”表示组合中的单元至少有一个触发（即不被遮挡）时，对于黑盒模型输出产生的额外效用。应理解，在黑盒模型中，输入单元往往并不是单独地贡献于模型的输出，而是彼此交互配合影响模型的输出。

因此，对于 $S \subseteq N$ ，定义 $I^{OR}(S)$ 为输入单元间组合 S 对应的“或交互”，使得当 S 中所有的输入变量至少有一个触发时， $I^{OR}(S)$ 则触发，并被累加到黑盒模型的输出中。具体地， $I^{OR}(S)$ 可以由如下公式计算得到。

$$I^{OR}(S) = \begin{cases} v(\emptyset), & S = \emptyset \\ -\sum_{L \subseteq S} (-1)^{s-l} v(N \setminus L), & S \neq \emptyset \end{cases}$$

其中， $s = |S|$ 且 $l = |L|$ 。上述“或交互”满足，对于任意给定的输入单元组合 $T \subseteq N$ ，黑盒模型在这一输入单元组合下的输出 $v(T)$ 可以拆分

为所有触发的或交互 $I^{OR}(S)$ 之和，如下式所示。

$$\forall T \subseteq N, v(T) = \sum_{S \subseteq N} I^{OR}(S) \bigvee_{i \in S} (e_T)_i \Leftrightarrow v(T) = I^{OR}(\emptyset) + \sum_{S: S \cap T \neq \emptyset} I^{OR}(S)$$

其中， e_S 满足，当 $i \in S$ 时， $(e_S)_i = 1$ ，否则， $(e_S)_i = 0$ ； $\bigvee_{i \in S} (e_T)_i$ 表示输入单元组合 S 中的单元是否至少有一个被触发，对 $\vee$ 符号后的元素做“或”操作得到的结果。

5 进一步，上述或交互的计算公式可等价为如下形式，假设 $\mathbf{v} = [v(S_1), v(S_2), \dots, v(S_{2^n})] \in \mathbb{R}^{2^n}$ ， $\mathbf{I}^{OR} = [I^{OR}(S_1), I^{OR}(S_2), \dots, I^{OR}(S_{2^n})] \in \mathbb{R}^{2^n}$ ，则存在 $\mathbf{B} \in \{0, 1, -1\}^{2^n \times 2^n}$ ，使得 $\mathbf{I}^{OR} = \mathbf{B}\mathbf{v}$ 。

在本发明的一个实施例中，基于步骤 101 中的 MLP-5 模型，利用所述“与交互”和“或交互”计算公式，分别计算“与交互”和“或交互”的值。如图 2（中）所示，显示了不同输入单元组合的“与交互”效用和“或交互”效用按降序排列的分布图。图 3（左）显示了部分“与交互”和“或交互”的效用值，图中按效用值的绝对值由大到小排列。（这个实施例是基于表格数据）

10 在本发明的另一实施例中，基于步骤 101 中的 ResNet-34 模型，利用所述“与交互”和“或交互”计算公式，分别计算“与交互”和“或交互”的值。如图 4（中）所示，显示了不同输入单元组合的“与交互”效用和“或交互”效用按降序排列的分布图。图 5（左）显示了部分“与交互”和“或交互”的效用值，图中按效用值的绝对值由大到小排列。（这个实施例是基于图像数据）

20

之后，进入步骤 104，该步骤可以进一步分为如下子步骤：

（a）基于步骤（3）中的“与交互”和“或交互”，将黑盒模型输出解释成输入单元组合间的“与加关系”和“或加关系”的组合。

25 （b）基于子步骤（a）中“与加关系”和“或加关系”的组合，进一步优化“与加关系”和“或加关系”的组合，使得两种关系的表达更简洁，更稀疏。

在子步骤（a）中，基于步骤（3）中的“与交互”和“或交互”，将黑盒模型在任意输入单元组合 T 下的输出 $v(T)$ 解释为“与加关系”和“或加关系”的组合。在本实施例中，在上述子步骤（a）中，仅选取一种基准值，将黑盒模型的输出解释为下式所示的“与加关系”和“或加关系”的组合。

$$\begin{aligned} \forall T \subseteq N, v(T) = & \hat{I}^{AND}(\emptyset) + \sum_{\emptyset \neq S \subseteq N} \hat{I}^{AND}(S) \bigwedge_{i \in S} (e_T)_i \\ & + \hat{I}^{OR}(\emptyset) + \sum_{\emptyset \neq S \subseteq N} \hat{I}^{OR}(S) \bigvee_{i \in S} (e_T)_i \end{aligned}$$

其中, $\hat{I}^{AND}(S)$ 表示输入单元组合 S 的“与交互”, $\hat{I}^{OR}(S)$ 表示输入单元组合 S 的“或交互”。同样地, 黑盒模型输出被解释为 2^n 个输入单元间组合“与交互”的效用之和, 以及 2^n 个输入单元间组合“或交互”的效用之和。可选地, 上述基准值可以选取为均值、随机值、零值等中的一种。

进一步, 在子步骤 (b) 中, 为了进一步得到稀疏的交互, 本发明希望最小化所有交互的 L_1 -范数之和, 即

$$\min_{\hat{I}^{AND}, \hat{I}^{OR}} \|\hat{I}^{AND}\|_1 + \|\hat{I}^{OR}\|_1$$

其中, $\hat{I}^{AND} = [\hat{I}^{AND}(S_1), \hat{I}^{AND}(S_2), \dots, \hat{I}^{AND}(S_{2^n})] \in \mathbb{R}^{2^n}$, $\hat{I}^{OR} = [\hat{I}^{OR}(S_1), \hat{I}^{OR}(S_2), \dots, \hat{I}^{OR}(S_{2^n})] \in \mathbb{R}^{2^n}$, $\|\cdot\|_1$ 表示向量的 L_1 -范数。

可选地, 在优化上述表达式时, 可以定义 $\mathbf{p} \in \mathbb{R}^{2^n}$, $\mathbf{q} \in \mathbb{R}^{2^n}$ 。使得 $\hat{I}^{AND} = A\left(\frac{1}{2}(\mathbf{v} + \mathbf{q}) - \mathbf{p}\right)$, $\hat{I}^{OR} = B\left(\frac{1}{2}(\mathbf{v} + \mathbf{q}) + \mathbf{p}\right)$ 。则上述优化问题可以转换为如下形式。

$$\begin{aligned} \min_{\mathbf{p}, \mathbf{q}, \alpha^{(1)}, \dots, \alpha^{(m)}} \quad & \|\hat{I}^{AND}\|_1 + \|\hat{I}^{OR}\|_1 \\ \text{s. t.} \quad & |\mathbf{q}_i| < \tau_i \end{aligned}$$

其中, τ_i 为预设的阈值。在本实施例中, $\tau_s = 0.05 \times |\mathbf{v}(N)|$ 。

在本发明的一个实施例中, 基于前述方法优化图 2 (中) 中计算得到的“与加关系”和“或加关系”, 得到更稀疏的交互效用。如图 2 (右) 所示, 图中显示了不同输入单元组合的“与交互”效用和“或交互”效用按降序排列的分布图, 相比于图 2 (中) 中的结果, “与交互”和“或交互”的效用分布更加稀疏。图 3 (右) 显示了部分“与交互”和“或交互”的效用值, 图中按效用值的绝对值由大到小排列, 相比于图 3 (左), 同样提升了“与交互”和“或交互”的稀疏性。

在本发明的另一个实施例中, 基于前述方法优化图 4 (中) 中计算得到的“与加关系”和“或加关系”, 得到更稀疏的交互效用。如图 4 (右) 所示, 图中显示了不同输入单元组合的“与交互”效用和“或交互”效用按降序排列的分布图, 相比于图 4 (中) 中的结果, “与交互”和“或交互”的效用分布更加稀疏。图 5 (右) 显示了部分“与交互”和“或交互”的效用值, 图中按效用值的绝对值由大到小排列, 相比于图 5 (左), 同样提升了“与交互”和“或交互”的稀疏性。

本发明的第二实施方式涉及一种解释黑盒人工智能模型所建模的稀疏交互效用的系统, 其结构如图 4 所示, 该系统包括:

(1) 输入模块, 被配置为一个预训练好的黑盒人工智能模型和一组待分析数据。

(2) 计算模块, 被配置为, 基于输入模块中的模型和数据, 计算模型所建模数据输入单元间的“与加关系”和“或加关系”, 并优化“与加关系”和“或加关系”的组合, 使得两种关系的表达都更简洁, 更稀疏。

(3) 输出模块, 被配置为, 输出计算模块优化得到的“与加关系”和“或加关系”的组合。

需要说明的是, 本领域技术人员应当理解, 上述解释黑盒人工智能模型所建模的稀疏交互效用的系统的实施方式中所示的各模块的实现功能, 可参照前述对解释黑盒人工智能模型所建模的稀疏交互效用的方法的相关描述而理解。上述解释黑盒人工智能模型所建模的稀疏交互效用的系统的实施方式中所示的各模块的功能可通过运行于处理器上的程序(可执行指令)而实现, 也可通过具体的逻辑电路而实现。本发明实施例上述解释黑盒人工智能模型所建模的稀疏交互效用的系统如果以软件功能模块的形式实现并作为独立的产品销售或使用, 也可以存储在一个计算机可读存储介质中。基于这样的理解, 本发明实施例的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来, 该计算机软件产品存储在一个存储介质中, 包括若干指令用以使得一台计算机设备(可以是个人计算机、服务器、或者网络设备等)执行本发明各个实施例该方法的全部或部分。而前述的存储介质包括: U 盘、移动硬盘、只读存储器(ROM, Read Only Memory)、磁碟或者光盘等各种可以存储程序代码的介质。这样, 本发明实施例不限制于任何特定的硬件和软件结合。

相应地, 本发明实施方式还提供一种计算机可读存储介质, 其中存储有计算机可执行指令, 该计算机可执行指令被处理器执行时实现本发明的各方法实施方式。计算机可读存储介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括但不限于, 相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带, 磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质, 可用于存储可以被计算设备访问的信息。按照本文中的界定, 计算机可读存储介质不包括暂存电脑可读媒体(transitory media), 如调制的数据信号和载波。

此外, 本发明实施方式还提供一种解释黑盒人工智能模型所建模的稀

疏交互效用的系统，其中包括用于存储计算机可执行指令的存储器，以及，
处理器；该处理器用于在执行该存储器中的计算机可执行指令时实现上述
各方法实施方式中的步骤。其中，该处理器可以是中央处理单元(Central
Processing Unit，简称“CPU”)，还可以是其他通用处理器、数字信号处
理器(Digital Signal Processor，简称“DSP”)、专用集成电路(Application
Specific Integrated Circuit，简称“ASIC”)等。前述的存储器可以是
只读存储器(read-only memory，简称“ROM”)、随机存取存储器(random
access memory，简称“RAM”)、快闪存储器(Flash)、硬盘或者固态硬盘
等。本发明各实施方式所公开的方法的步骤可以直接体现为硬件处理器
执行完成，或者用处理器中的硬件及软件模块组合执行完成。

需要说明的是，在本专利的发明文件中，诸如第一和第二等之类的关
系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不
一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺
序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他
性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包
括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种
过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由
语句“包括一个”限定的要素，并不排除在包括该要素的过程、方法、物
品或者设备中还存在另外的相同要素。本专利的发明文件中，如果提到根
据某要素执行某行为，则是指至少根据该要素执行该行为的意思，其中包
括了两种情况：仅根据该要素执行该行为、和根据该要素和其它要素执行
该行为。多个、多次、多种等表达包括 2 个、2 次、2 种以及 2 个以上、2
次以上、2 种以上。

在本发明提及的所有文献都被认为是整体性地包括在本发明的公开
内容中，以便在必要时可以作为修改的依据。此外应理解，以上该仅为本
说明书的较佳实施例而已，并非用于限定本说明书的保护范围。凡在本说
明书一个或多个实施例的精神和原则之内，所作的任何修改、等同替换、
改进等，均应包含在本说明书一个或多个实施例的保护范围之内。

权利要求

1. 一种测定黑盒人工智能模型的稀疏交互效用的方法，其特征在于，包括以下步骤：

5 (1) 提供模型

提供一待分析的黑盒人工智能模型，所述黑盒人工智能模型包括基于某一数据集上预先训练好的人工智能模型；

(2) 获取待评估数据

10 基于所述黑盒人工智能模型，获取与所述人工智能模型兼容的待评估数据，其中，所述待评估数据包含 n 个输入单元， n 为 ≥ 4 的正整数；其中，步骤 (1) 和步骤 (2) 可互换或同时进行；

(3) 计算所述输入单元间的“与交互”和“或交互”

15 将所述待评估数据输入所述黑盒人工智能模型，并获取所述黑盒人工智能模型的输出；基于所述黑盒人工智能模型的输出，对所述输入单元间的交互配合作用进行建模，从而获得由不同所述输入单元组成的集合，记为输入单元组合 S （或评估集合 S ）；并计算每个输入单元组合 S 的“与交互”和“或交互”；

(4) 测定稀疏交互效用

20 基于步骤 (3) 中获得的输入单元组合 S 间的“与交互”和“或交互”，计算所有输入单元组合 S 间的“与加关系”和“或加关系”，并将黑盒模型输出表示成输入单元组合间的“与加关系”和“或加关系”的组合，并用于解释黑盒人工智能模型的稀疏交互效用。

2. 根据权利要求 1 所述的方法，其特征在于，所述方法还包括：

25 (5) 对所述的“与加关系”和“或加关系”进行优化，使得经优化后的“与加关系”和“或加关系”组合的稀疏程度升高，并用经优化后“与加关系”和“或加关系”来解释黑盒人工智能模型的稀疏交互效用。

3. 根据权利要求 1 所述的方法，其特征在于，所述步骤 (3) 中的“与交互” $I^{AND}(S)$ 通过以下公式得出：

$$I^{AND}(S) = \sum_{L \subseteq S} (-1)^{s-l} v(L)$$

其中， $s = |S|$ 且 $l = |L|$ ； s 为某一所述输入单元组合 S 中所述输入单元的个数，记为 $s = |S|$ ； l 为输入单元间组合 L 中所述输入单元的个数，记为 $l = |L|$ ，所述输入单元间组合 L 为所述输入单元组合 S 的任一子集，

公式中的 $v(L)$ 的计算方式如下，将 n 个所述输入变量记作 $N = \{1, 2, \dots, n\}$ ， $v(L)$ 表示在所述待评估数据中仅保留所述输入单元间组合 L 中的输入单元，而将其他输入单元遮挡为基准值 b ，所述基准值 b 是一个与待评估数据维数相同的量，所述基准值 b 可以选取为所述输入样本集的均值、随机值、零值等；将遮挡后的所述待评估数据 x_L 输入黑盒模型得到的输出值记为 $v(L)$ ，具体地，遮挡后的所述待评估数据 x_L 的各个输入单元由下式定义：

$$(x_S)_i = \begin{cases} x_i & \text{如果输入的第 } i \text{ 个输入单元被输入单元组合 } S \text{ 所包含} \\ b_i & \text{如果输入的第 } i \text{ 个输入单元未被输入单元组合 } S \text{ 所包含} \end{cases}。$$

4. 根据权利要求 1 所述的方法，其特征在于，所述步骤（3）中的“或交互” $I^{OR}(S)$ 通过以下公式得出：

$$I^{OR}(S) = \begin{cases} v(\emptyset), & S = \emptyset \\ -\sum_{L \subseteq S} (-1)^{s-l} v(N \setminus L), & S \neq \emptyset \end{cases}$$

其中，其中， $s = |S|$ 且 $l = |L|$ ； s 为某一所述输入单元组合 S 中所述输入单元的个数，计为 $s = |S|$ ； l 为输入单元间组合 L 中输入单元的个数，记为 $l = |L|$ ，所述输入单元间组合 L 为所述输入单元组合 S 的任一子集。

5. 根据权利要求 1 所述的方法，其特征在于，所述步骤（4）中通过如下公式将所述黑盒模型的输出 $v(T)$ 解释为“与加关系”和“或加关系”的组合：

$$\begin{aligned} \forall T \subseteq N, v(T) = & \hat{I}^{AND}(\emptyset | b^{(1)}) + \sum_{\emptyset \neq S \subseteq N} \hat{I}^{AND}(S | b^{(1)}) \left[\bigwedge_{i \in S} (e_T)_i \right] \\ & + \dots \\ & + \hat{I}^{AND}(\emptyset | b^{(m)}) + \sum_{\emptyset \neq S \subseteq N} \hat{I}^{AND}(S | b^{(m)}) \left[\bigwedge_{i \in S} (e_T)_i \right] \\ & + \hat{I}^{OR}(\emptyset) + \sum_{\emptyset \neq S \subseteq N} \hat{I}^{OR}(S) \left[\bigvee_{i \in S} (e_T)_i \right] \end{aligned}$$

其中，黑盒模型输出 $v(N)$ 被解释为 m 组“与交互”的效用以及一组“或交互”的效用之和， $b^{(1)}, \dots, b^{(m)}$ 表示 m 个基准值， m 为基准值的个数，所述基准值 m 可以选取为均值、随机值、零值等，所述基准值包括其中每组“与交互”包含 2^n 个输入单元间组合内的“与交互”效用 $\hat{I}^{AND}(S | b), S \subseteq N$ ；其中该一组“或交互”的效用之和了包含 2^n 个输入单元间组合内的“或交互”效用 $\hat{I}^{OR}(S), S \subseteq N$ ；所述 m 种不同的基准值可以选取

为均值、随机值、零值等；公式中的 $\forall$ 表示“任意”；公式中的 e_S 满足，当 $i \in S$ 时， $(e_T)_i = 1$ ，否则， $(e_T)_i = 0$ ； $\bigwedge_{i \in S}(e_T)_i$ 表示输入单元组合 S 中的单元是否全部触发，对 $\bigwedge$ 符号后的元素做“与”操作得到的结果，即 $\bigwedge_{i \in S}(e_T)_i = \prod_{i \in S}(e_T)_i$ ； $\bigvee_{i \in S}(e_T)_i$ 表示输入单元组合 S 中的单元是否至少有一个被触发，对 $\bigvee$ 符号后的元素做“或”操作得到的结果，即 $\bigvee_{i \in S}(e_T)_i = 1 - \prod_{i \in S}[1 - (e_T)_i]$ 。

6. 根据权利要求 2 所述的方法，其特征在于，在所述步骤（5）中，通过如下公式进一步得到更加稀疏的交互关系：

$$\min_{\hat{\mathbf{I}}_{b(1)}^{AND}, \dots, \hat{\mathbf{I}}_{b(m)}^{AND}, \hat{\mathbf{I}}^{OR}} \|\hat{\mathbf{I}}_{b(1)}^{AND}\|_1 + \dots + \|\hat{\mathbf{I}}_{b(m)}^{AND}\|_1 + \|\hat{\mathbf{I}}^{OR}\|_1$$

10 其中， $\hat{\mathbf{I}}_b^{AND} = [\hat{I}^{AND}(S_1|b), \hat{I}^{AND}(S_2|b), \dots, \hat{I}^{AND}(S_{2^n}|b)] \in R^{2^n}$ ， $\hat{\mathbf{I}}^{OR} = [\hat{I}^{OR}(S_1), \hat{I}^{OR}(S_2), \dots, \hat{I}^{OR}(S_{2^n})] \in R^{2^n}$ ，其中 $\|\cdot\|_1$ 表示向量的 L_1 -范数，上式中的 m 为所选取的基准值的个数，所述基准值包括为所述输入样本集的均值、随机值，零值。

15 7. 根据权利要求 6 所述的方法，其特征在于，可以通过如下方式简化如权利要求 6 所述的公式：

可以定义 $\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(m)}, \mathbf{p}^{(or)} \in \mathbb{R}^{2^n}$ ，并满足 $\mathbf{p}^{(1)} + \dots + \mathbf{p}^{(m)} + \mathbf{p}^{(or)} = \mathbf{0}$ ，所述 $\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(m)}, \mathbf{p}^{(or)}$ 表示对于神经网络输出 $\mathbf{v}$ 的一种划分；并定义 $\mathbf{q} \in \mathbb{R}^{2^n}$ ，所述 $\mathbf{q}$ 为一个各维绝对值较小的向量，表示将模型输出解释为“与加关系”和“或加关系”所产生的偏差；使得 $\hat{\mathbf{I}}_{b(1)}^{AND} = A\left(\frac{1}{m+1}(\mathbf{v} + \mathbf{q}) + \mathbf{p}^{(1)}\right), \dots, \hat{\mathbf{I}}_{b(m)}^{AND} = A\left(\frac{1}{m+1}(\mathbf{v} + \mathbf{q}) - \mathbf{p}^{(m)}\right)$ ， $\hat{\mathbf{I}}^{OR} = B\left(\frac{1}{m+1}(\mathbf{v} + \mathbf{q}) + \mathbf{p}^{(or)}\right)$ ；则上述优化问题可以转换为如下形式：

$$\min_{\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(m)}, \mathbf{p}^{(or)}, \mathbf{q}} \|\hat{\mathbf{I}}_{b(1)}^{AND}\|_1 + \dots + \|\hat{\mathbf{I}}_{b(m)}^{AND}\|_1 + \|\hat{\mathbf{I}}^{OR}\|_1$$

s. t. $|\mathbf{q}_i| < \tau_i$

其中， τ_i 为预设的阈值；优选地， $\tau_i = 0.05 \times |\mathbf{v}(N) - \mathbf{v}(\emptyset)|$ ， m 的取值范围为 1 至 10。

25 8. 根据权利要求 1 所述的方法，其特征在于，所述步骤（3）还包括以下步骤：

计算所述待分析数据在任意所述输入单元被遮挡的情况下，所述黑盒模型的输出值。

30 9. 根据权利要求 1 所述的方法，其特征在于，所述待评估数据选自下

组：格数据、图片、文本、语音或其组合。

10. 一种解释黑盒人工智能模型所建模的稀疏交互效用的系统，其特征在于，包括如下模块：

- 5 (1) 输入模块，被配置为一个预训练好的黑盒人工智能模型和一组待分析数据；
- (2) 计算模块，被配置为，基于输入模块中的模型和数据，计算模型所建模数据输入单元间的“与加关系”和“或加关系”，并优化“与加关系”和“或加关系”的组合，使得两种关系的表达都更简洁，更稀疏；
- 10 (3) 输出模块，被配置为，输出计算模块优化得到的“与加关系”和“或加关系”的组合。

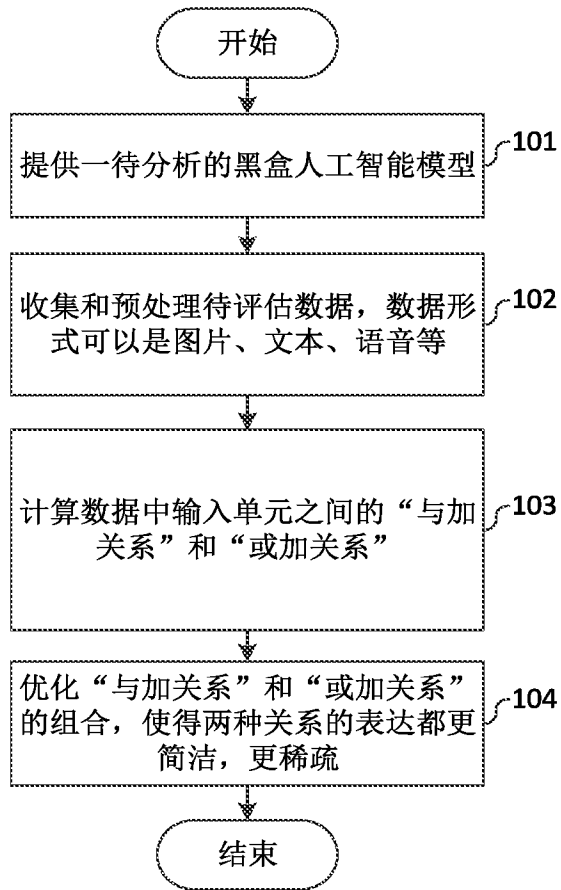


图 1

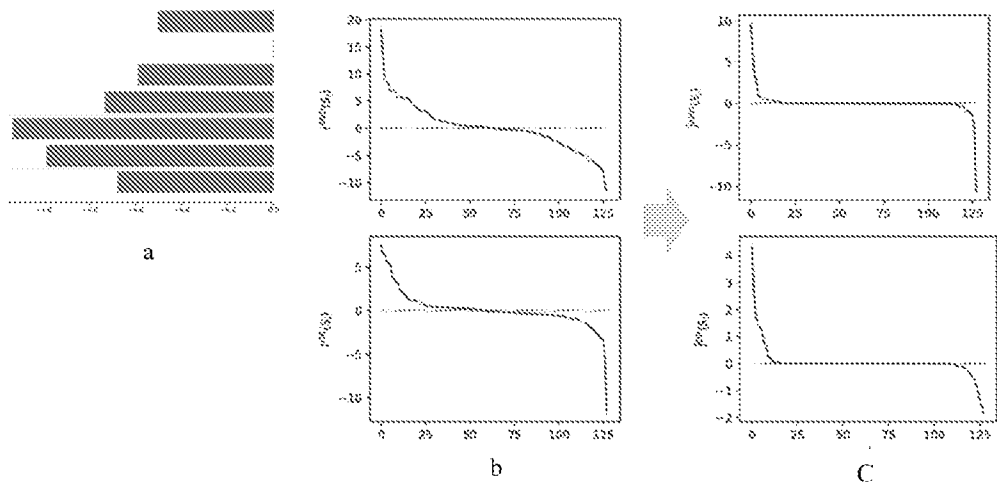


图 2

序号	交互项	交互类型	交互值
1	信号1	与交互	18.5947
2	信号4	与交互	13.9662
3	信号1,信号4	与交互	-10.1763
4	信号2,信号3,信号4	与交互	0.0897
5	信号3	与交互	8.9238
...			
21	信号1,信号4	或交互	6.3960
22	信号1,信号3	与交互	-6.0364
23	信号3,信号6	与交互	-5.9252
24	信号2,信号3	与交互	-5.9153
25	信号1,信号2,信号4,信号5,信号7	与交互	5.8323
...			
41	信号1,信号4,信号5,信号6	与交互	-4.7161
42	信号1,信号3,信号4,信号5	与交互	-4.5664
43	信号1,信号2,信号4,信号7	与交互	-4.5381
44	信号1,信号4,信号5	与交互	4.2696
45	信号1,信号3,信号4	与交互	4.1863
...			

序号	交互项	交互类型	交互值
1	信号1	与交互	9.6536
2	信号4	与交互	6.1205
3	信号1,信号4	或交互	4.3571
4	信号5	与交互	3.3949
5	信号1,信号5	或交互	2.7993
...			
21	信号6	与交互	-0.9404
22	信号2,信号3,信号7	与交互	-0.9346
23	信号1,信号3,信号5	与交互	-0.9150
24	信号4,信号6	与交互	0.8704
25	信号2,信号3,信号4	与交互	0.8560
...			
41	信号1,信号2,信号3,信号7	与交互	0.3026
42	信号2,信号7	或交互	-0.2738
43	信号3,信号4,信号5,信号7	或交互	0.2490
44	信号3,信号4,信号7	或交互	0.2317
45	信号3,信号5,信号7	与交互	0.2236
...			

优化前的交互值（按绝对值由大到小排序）
交互值都在同一量级，交互值分布不稀疏

优化后的交互值（按绝对值由大到小排序）
相比于优化前，交互值分布更稀疏

图 3

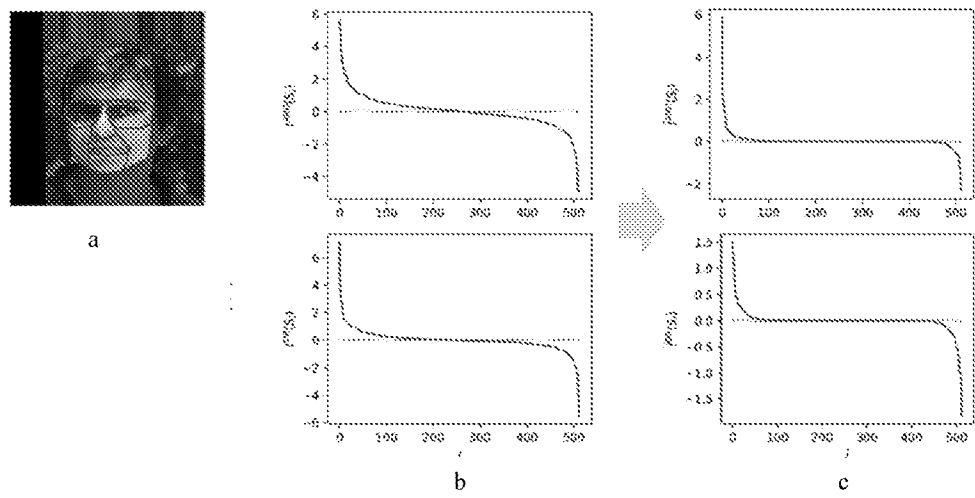


图 4

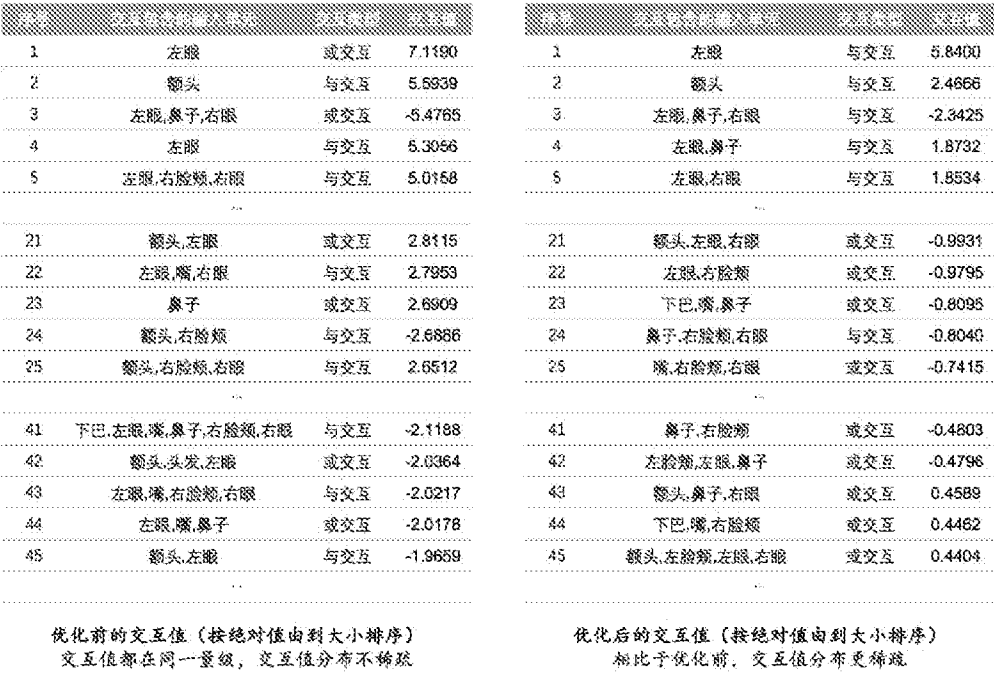


图 5

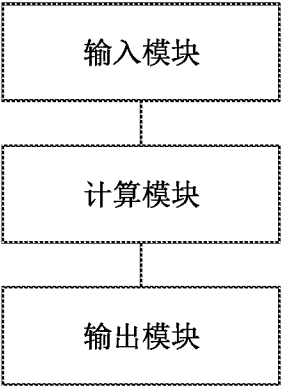


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2024/088822

A. CLASSIFICATION OF SUBJECT MATTER G06N3/04(2023.01)i; G06F30/27(2020.01)i; G06N20/00(2019.01)i; G06N3/08(2023.01)i; G06F111/10(2020.01)i According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) IPC:G06N G06F Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 黑盒, 模型, 人工智能, 神经网络, 稀疏, 解释, 与, 或, 与加, 或加, 关系; black box, model, artificial intelligence, neural network, sparse, interpretation, AND, OR, AND 1W Add, OR 1W Add, relationship		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 114462291 A (SHANGHAI JIAO TONG UNIVERSITY) 10 May 2022 (2022-05-10) claims 1-2	1-10
A	CN 113191441 A (BEIJING INSTITUTE OF TECHNOLOGY) 30 July 2021 (2021-07-30) entire document	1-10
A	CN 113518999 A (SIEMENS AG) 19 October 2021 (2021-10-19) entire document	1-10
A	CN 115457365 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 09 December 2022 (2022-12-09) entire document	1-10
A	US 2022172050 A1 (UMNAI LIMITED) 02 June 2022 (2022-06-02) entire document	1-10
A	US 2022405531 A1 (ETSY, INC.) 22 December 2022 (2022-12-22) entire document	1-10
☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents: “A” document defining the general state of the art which is not considered to be of particular relevance “D” document cited by the applicant in the international application “E” earlier application or patent but published on or after the international filing date “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) “O” document referring to an oral disclosure, use, exhibition or other means “P” document published prior to the international filing date but later than the priority date claimed “T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art “&” document member of the same patent family		
Date of the actual completion of the international search 21 June 2024		Date of mailing of the international search report 27 June 2024
Name and mailing address of the ISA/CN China National Intellectual Property Administration (ISA/ CN) China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088		Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2024/088822

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	114462291	A	10 May 2022	None			
CN	113191441	A	30 July 2021	None			
CN	113518999	A	19 October 2021	EP	3918524	A1	08 December 2021
				US	2022172146	A1	02 June 2022
				EP	3709222	A1	16 September 2020
				WO	2020182706	A1	17 September 2020
CN	115457365	A	09 December 2022	US	2024104906	A1	28 March 2024
US	2022172050	A1	02 June 2022	WO	2022101515	A1	19 May 2022
				EP	4244767	A1	20 September 2023
US	2022405531	A1	22 December 2022	EP	4356317	A1	24 April 2024
				WO	2022265782	A1	22 December 2022

A. 主题的分类

G06N3/04(2023.01)i; G06F30/27(2020.01)i; G06N20/00(2019.01)i; G06N3/08(2023.01)i; G06F111/10(2020.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC:G06N G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 黑盒, 模型, 人工智能, 神经网络, 稀疏, 解释, 与, 或, 与加, 或加, 关系; black box, model, artificial intelligence, neural network, sparse, interpretation, AND, OR, AND 1W Add, OR 1W Add, relationship

C. 相关文件

类型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 114462291 A (上海交通大学) 2022年5月10日 (2022 - 05 - 10) 权利要求1-2	1-10
A	CN 113191441 A (北京理工大学) 2021年7月30日 (2021 - 07 - 30) 全文	1-10
A	CN 113518999 A (西门子股份公司) 2021年10月19日 (2021 - 10 - 19) 全文	1-10
A	CN 115457365 A (北京百度网讯科技有限公司) 2022年12月9日 (2022 - 12 - 09) 全文	1-10
A	US 2022172050 A1 (UMNAI LTD.) 2022年6月2日 (2022 - 06 - 02) 全文	1-10
A	US 2022405531 A1 (ETSY INC.) 2022年12月22日 (2022 - 12 - 22) 全文	1-10

☐ 其余文件在C栏的续页中列出。

☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2024年6月21日

国际检索报告邮寄日期

2024年6月27日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

邱爽

电话号码 (+86) 010-53961362

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2024/088822

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	114462291	A	2022年5月10日	无			
CN	113191441	A	2021年7月30日	无			
CN	113518999	A	2021年10月19日	EP	3918524	A1	2021年12月8日
				US	2022172146	A1	2022年6月2日
				EP	3709222	A1	2020年9月16日
				WO	2020182706	A1	2020年9月17日
CN	115457365	A	2022年12月9日	US	2024104906	A1	2024年3月28日
US	2022172050	A1	2022年6月2日	WO	2022101515	A1	2022年5月19日
				EP	4244767	A1	2023年9月20日
US	2022405531	A1	2022年12月22日	EP	4356317	A1	2024年4月24日
				WO	2022265782	A1	2022年12月22日