

一种数据处理方法以及装置

本申请要求于 2020 年 12 月 02 日提交中国专利局、申请号为 202011391497.6、发明名称为“一种数据处理方法以及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 **技术领域**

本申请涉及人工智能领域，尤其涉及一种数据处理方法以及装置。

背景技术

人工智能 (Artificial Intelligence, AI) 是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个分支，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。人工智能领域的研究包括机器人，自然语言处理，计算机视觉，决策与推理，人机交互，推荐与搜索，AI 基础理论等。

15 目前通常是通过神经网络实现 AI 的各种功能，例如自然语言处理、图像处理、语音处理等。随着 AI 技术的发展，AI 的功能逐渐多样化。但是这些功能的实现都是基于输入数据和标签，通过梯度下降等优化方法训练机器学习模型的参数，当模型参数收敛之后，可利用该模型来完成未知数据的预测。因此输入的数据和标签的处理可以看作是 AI 的基础。基于数据类型的差异，待输入的数据可以分为连续特征和离散特征。目前对于离散特征一般使用独热编码 (one-hot encoding)。而针对连续特征，常用的特征处理方法有类别化、归一化和离散化三类，其中离散化方法被工业界广泛常用，包括等频离散化、等距离散化、Logarithms 方法和基于树模型等方法。

目前的连续特征处理方法是通过对离散化策略 (如启发式规则或模型) 将连续特征值离散化到不同的区间 (如桶, bucket) 中，将区间的编号代替原本的特征值，然后使用和离散特征相同的方式获取向量化表示。但是该方法通常需要大量的人力和时间进行离散化策略尝试和优化，才能得到最优的离散化规则或模型，进而获取最终的 embedding 向量表示。同时在这些离散化策略中，均存在如下缺点：两阶段 (two-phase problem, TPP)、相近值却不同表示 (similar value but dis-similar embedding, SBD)、不同值却相同表达 (dissimilar value but same embedding, DBS)。例如：我们将年龄特征分为若干组，其中，18 至 40 为一组，40 至 60 为一组，年龄差距很大的 18 和 40 将使用同一个 embedding，无法体现二者的差异，年龄接近的 40 和 41 却分到了两个组，embedding 很可能差异巨大。因此现有方案中连续特征的向量表达值不够。

发明内容

35 本申请实施例提供了一种数据处理方法以及装置，用于更好地学习连续特征中每个特征值的向量表达值，使其具有更好的表达能力。

第一方面，本申请实施例提供一种数据处理方法，具体包括：该数据处理装置从样本数据中获取连续特征；然后通过离散化模型对该连续特征进行离散化处理得到该连续特征

对应的 N 个离散化概率值，其中，所述 N 个离散化概率值对应于预设的 N 个基向量表达值，该 N 为大于 1 的整数；最后，该数据处理装置根据该 N 个离散化概率值和该 N 个基向量表达值确定该连续特征的一个向量表达值。

本实施例中，该连续特征为该样本数据中具有连续统计特征值的特征数据。一种示例性方案中，该连续特征包括但不限于推荐模型中的年龄特征、点击量特征和分数特征。其中，点击量特征的取值可以是整个正整数的取值范围，且取值为连续的特征，比如取值为 1、2、3、4、……。而年龄特征的取值可以是 0 至现有已记录的最大年龄。而样本数据中的离散特征通常只可以按一定顺序一一列举，例如性别特征，其取值只有男或女。

向量表达是指用向量来表示具体的特征。本实施例中，该基向量表达值为可以为预设的初始化向量表达值，也可以是经过优化之后的向量表达值。而连续特征的向量表达值是根据该基向量表达值确定的用于表示该连续特征的具体特征的向量值。可以理解的是该连续特征的向量表达值与该基向量表达值的维度相同。一种示例性方案中，该基向量表达值可以为 5 维的向量值，比如 (01010)；该连续特征对应的向量表达值也同为 5 维的向量值，比如 (11010)。

本实施例中，该数据处理装置通过将离散化模型为每一个连续特征的特征值计算得到大于 1 维的离散化概率并为该连续特征中每个连续特征域预设大于 1 维的元向量表达值 meta-embedding，并利用聚合函数和该离散化概率从该 meta-embedding 中为特征值确定一个向量表达值。这样对于来说，与现有技术相比其学习表达的向量表达值具有更好的表达能力，有助于提升预测结果的精确性。

可选的，该数据处理装置在利用离散化模型对所述连续特征中进行离散化处理得到所述连续特征对应的 N 个离散化概率的具体方式可以如下：该数据处理装置预设该离散化模型中的初始变量；然后根据该初始变量确定该连续特征对应的 N 个映射值；最后根据该映射值计算得到该连续特征的 N 个离散化概率值。本实施例中，该初始变量可以为初始映射域，然后对于连续特征来说，通过该初始变量可以计算得到该连续特征对应的 N 个映射值；然后根据该 N 个映射值中的每一个映射值计算一个对应的概率值，从而得到 N 个概率值，此时该 N 个概率值作为该连续特征对应的 N 个离散化概率。

可选的，该离散化模型可以是多分类神经网络、注意力网络或线性映射和 softmax。本实施例中，该离散化模型只要可以实现该特征值对应多个离散化概率即可，具体离散化模型此处不做限定。同时该离散化模型也可以根据不同的应用场景进行相应的选择。比如在连续特征可以进行分类处理的系统中，该离散化模型可以是上述提供的离散化模型。

可选的，基于上述方式，该离散化模型为线性映射和 softmax 时，利用该离散化模型得到该连续特征的 N 个离散化概率值具体可以为：

预设初始化线性映射变量为： $W_{logit} = R^{1 \times h}$ ，所述初始化线性映射变量为所述初始变量；通过线性映射公式确定所述连续特征中的对应的 N 个映射值，

其中，所述线性映射公式为： $cont_{logit} = cont \cdot W_{logit}$ ；根据离散化公式计算所述 N 个映射值中每一个映射值对应的概率值得到 N 个概率值，其中所述离散化公式为：

-3-

$$cont_p_k = e^{\frac{1}{\tau} cont_{logit}^k} / \sum_{i=1}^h e^{\frac{1}{\tau} cont_{logit}^i}, \text{ 所述 } N \text{ 个概率值作为所述 } N \text{ 个离散化概率;}$$

其中, 所述 W_{logit} 用于指示线性映射变量, 所述 R 用于指示实数域, 所述 h 用于指示所述连续特征离散化后的桶数, 所述 h 等于所述 N , 所述 $cont_{logit}$ 用于指示所述连续特征线性映射后的表达, 所述 $cont_p_k$ 用于指示所述连续特征离散化到第 k 个桶的概率值, 所述

5 $cont_{logit}^k$ 指示所述连续特征线性映射后第 k 个神经元输出, 所述 τ 用于指示 softmax 的温度控制系数, 所述 $cont_{logit}^i$ 用于指示所述连续特征线性映射后第 i 个神经元输出。

可选的, 本实施例中, 该数据处理装置可以利用聚合函数根据该 N 个离散化概率值和该 N 个基向量表达值确定该连续特征对应的一个向量表达值, 其中该聚合函数为 Max-Pooling、Top-K-Sum 或 Weighted-Average。

10 可以理解的是, 该聚合函数用于将与 N 个离散化概率值对应的 N 个 meta-embedding 聚合为该连续特征对应的一个 embedding。一种示例性方案中, 该离散化概率值为 (a1、a2、a3、a4), 该 meta-embedding 为 (b1、b2、b3、b4), 则该 a1 对应 b1, a2 对应 b2, a3 对应 b3, a4 对应 b4。该聚合函数用于从 (b1、b2、b3、b4) 聚合出一个 embedding, 且用这一个 embedding 表示该连续特征。比如, 可以从 (b1、b2、b3、b4) 聚合得到 b3 作为该

15 连续特征对应的向量表达值。

其中, 该 Max-Pooling 的计算方式是根据离散化概率值中的最大值对应的索引从 Meta Embedding 中获取对应的 embedding 作为该连续特征对应的向量表达值。其计算公式为:

$vx_cont = E^k, \text{ where } k = \arg \max_h \{cont_p_h\}$ 。例如, 假设该离散化概率值为 (a1、a2、a3、a4), 该 meta-embedding 为 (b1、b2、b3、b4), 则该 a1 对应 b1, a2 对应 b2, a3 对应 b3, a4 对应 b4。若该 a3 的值最大, 则该 b3 作为该连续特征的向量表达值。

20

该 Top-K-Sum 的计算方式是取离散化概率中最大的 k 个值对应的索引, 再从 Meta Embedding 中获取对应的 embedding 并求和, 作为该连续特征对应的向量表达值。其计算

公式为: $vx_cont = \sum_{k=1}^K E^k, \text{ where } k = \arg \text{top}k_h \{cont_p_h\}$ 。例如, 假设该

25 离散化概率为 (a1、a2、a3、a4), 该 meta-embedding 为 (b1、b2、b3、b4), 则该 a1 对应 b1, a2 对应 b2, a3 对应 b3, a4 对应 b4。若该 a2 和 a3 的值最大, 则该 b2 加上 b3 作为该连续特征的向量表达值。

该 Weighted-Average 的计算方式是将 N 个概率值和 Meta Embedding 进行加权求和, 作为该连续特征对应的向量表达值, 其计算公式为:

30 $vx_cont = \sum_{k=1}^h cont_p_k \times E^k$ 。可以理解的是, 该计算公式中的 h 等于该 N 。例

—4—

如,假设该离散化概率为(a1、a2、a3、a4),该 meta-embedding 为(b1、b2、b3、b4),则该 a1 对应 b1, a2 对应 b2, a3 对应 b3, a4 对应 b4。该连续特征的向量表达值等于(a1*b1+ a2*b2+ a3*b3+ a4*b4)。

5 可选的,该方法还包括将用户特征和对象特征输入推荐模型或搜索模型得到预测结果;所述用户特征或所述对象特征中包括所述向量表达值;所述用户特征或所述待推荐对象特征采用所述向量表达值表示所述连续特征。即本申请实施例提供的数据处理方法还可以应用于具体的应用场景。当该数据处理方法应用于推荐模型或搜索模型时,该推荐模型或搜索模型中需要输入用户特征和对象特征,若该用户特征或该对象特征中包括该连续特征,则该数据处理模型可以采用上述方法将输出该连续特征的向量表达值,而该向量表达值会包含于该用户特征或对象特征中输入至该推荐模型或搜索模型。

10 可选的,该向量表达值可以是直接包含于该用户特征或对象特征中,也可以是与其它连续特征进行拼接之后作为一个整体的输入特征表达或者是与其它连续特征和离散特征进行拼接之后作为一个整体的输入特征表达。只要该向量表达值可以作为模型应用或模型训练的输入即可,具体此处不做限定。这样由于该连续特征具有更好的向量表达值,使得该输入特征表达也更具有表达能力,因此在模型训练以及模型应用过程中,可以更精确的实现模型的功能。

可选的,在上述模型的应用或者训练过程中,该数据处理装置还可以获取到应用过程中的实际结果,然后根据该预测结果和该实际结果利用模型训练过程中的损失函数调整该离散化模型中的权重参数。该权重参数可以理解为不包括该 N 值的其它参数,比如该基向量表达值。这样该离散化模型与应用该离散化模型的机器学习模型相互影响,从而根据实时数据分布进行该离散化模型中的权重参数,实现离散化模型的优化。

20 可选的,该 N 的取值大于等于 20 且小于等于 100。在此范围内,该离散化模型可以具有较好的应用效果。

可选的,该连续特征可以是经过归一化处理的连续特征,这样可以更快速的实现连续特征的离散化。

25 第二方面,本申请提供一种数据处理装置,该装置具有实现上述第一方面中数据处理装置行为的功能。该功能可以通过硬件实现,也可以通过硬件执行相应的软件实现。该硬件或软件包括一个或多个与上述功能相对应的模块。

30 在一个可能的实现方式中,该装置包括用于执行以上第一方面各个步骤的单元或模块。例如,该装置包括:获取模块,用于获取连续特征;

处理模块,用于利用离散化模型对所述连续特征进行离散化处理得到所述连续特征对应的 N 个离散化概率值,所述 N 为大于 1 的整数;所述 N 个离散化概率值对应于预设的 N 个基向量表达值;根据所述 N 个离散化概率值和所述 N 个基向量表达值确定所述连续特征的一个向量表达值。

35 可选的,还包括存储模块,用于保存数据处理装置必要的程序指令和数据。

在一种可能的实现方式中,该装置包括:处理器和收发器,该处理器被配置为支持数据处理装置执行上述第一方面提供的方法中相应的功能。收发器用于指示数据处理装置和样本数据存储装置之间的通信,比如从该样本数据存储装置获取该连续特征。可选的,此装置还可以包括存储器,该存储器用于与处理器耦合,其保存数据处理装置必要的程序指

令和/数据。

在一种可能的实现方式中，当该装置为数据处理装置内的芯片时，该芯片包括：处理模块和收发模块，该收发模块例如可以是该芯片上的输入/输出接口、管脚或电路等，将连续特征传送给与此芯片耦合的其他芯片或模块中。该处理模块例如可以是处理器，此处理器用于利用离散化模型对所述连续特征进行离散化处理得到所述连续特征对应的 N 个离散化概率值，所述 N 为大于 1 的整数；所述 N 个离散化概率值对应于预设的 N 个基向量表达值；根据所述 N 个离散化概率值和所述 N 个基向量表达值确定所述连续特征的一个向量表达值。该处理模块可执行存储单元存储的计算机执行指令，以支持数据处理装置执行上述第一方面提供的方法。可选地，该存储单元可以为该芯片内的存储单元，如寄存器、缓存等，该存储单元还可以是位于该芯片外部的存储单元，如只读存储器（read-only memory, ROM）或可存储静态信息和指令的其他类型的静态存储设备，随机存取存储器（random access memory, RAM）等。

在一种可能实现方式中，该装置包括通信接口和逻辑电路，该通信接口用于获取连续特征；该逻辑电路，用于利用离散化模型对所述连续特征进行离散化处理得到所述连续特征对应的 N 个离散化概率值，所述 N 为大于 1 的整数；所述 N 个离散化概率值对应于预设的 N 个基向量表达值；根据所述 N 个离散化概率值和所述 N 个基向量表达值确定所述连续特征的一个向量表达值。

其中，上述任一处提到的处理器，可以是一个通用中央处理器（Central Processing Unit, CPU），微处理器，特定应用集成电路（application-specific integrated circuit, ASIC），或一个或多个用于控制上述各方面数据传输方法的程序执行的集成电路。

第三方面，本申请实施例提供一种计算机可读存储介质，该计算机存储介质存储有计算机指令，该计算机指令用于执行上述各方面中任意一方面任意可能的实施方式的方法。

第四方面，本申请实施例提供一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机执行上述各方面中任意一方面该方法。

第五方面，本申请提供了一种芯片系统，该芯片系统包括处理器，用于支持数据处理装置实现上述方面中所涉及的功能，例如生成或处理上述方法中所涉及的数据和/或信息。在一种可能的设计中，该芯片系统还包括存储器，该存储器，用于保存数据处理装置必要的程序指令和数据，以实现上述各方面中任意一方面的功能。该芯片系统可以由芯片构成，也可以包含芯片和其他分立器件。

附图说明

- 图 1 为一种人工智能主体框架示意图；
- 图 2 为推荐系统的处理流程示意图；
- 图 3 为推荐系统的结构示意图；
- 图 4 为本申请实施例中离散化模型与深度学习模型结构的一个示例性模型架构图；
- 图 5 为本申请实施例中数据处理装置的一个实施例示意图；

图 6 为本申请实施例中数据处理装置的另一个实施例示意图；

图 7 为本申请实施例中数据处理方法的一个实施例示意图；

图 8 为本申请实施例应用程序推荐显示的一个应用场景示意图；

图 9 为本申请实施例中数据处理方法的另一个实施例示意图；

5 图 10 为本申请实施例中数据处理方法的另一个实施例示意图。

具体实施方式

为了使本申请的目的、技术方案及优点更加清楚明白，下面结合附图，对本申请的实施例进行描述，显然，所描述的实施例仅仅是本申请一部分的实施例，而不是全部的实施例。本领域普通技术人员可知，随着新应用场景的出现，本申请实施例提供的技术方案对于类似的技术问题，同样适用。

本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的实施例能够以除了在这里图示或描述的内容以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，
15 例如，包含了一系列步骤或模块的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或模块，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或模块。在本申请中出现的对步骤进行的命名或者编号，并不意味着必须按照命名或者编号所指示的时间/逻辑先后顺序执行方法流程中的步骤，已经命名或者编号的流程步骤可以根据要实现的技术目的变更执行次序，只要能达到相同或者相类似的技术效果
20 即可。本申请中所出现的单元的划分，是一种逻辑上的划分，实际应用中实现时可以有另外的划分方式，例如多个单元可以结合成或集成在另一个系统中，或一些特征可以忽略，或不执行，另外，所显示的或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，单元之间的间接耦合或通信连接可以是电性或其他类似的形式，本申请中均不作限定。并且，作为分离部件说明的单元或子单元可以是也可以不是物理上的分离，可以是
25 也可以不是物理单元，或者可以分布到多个电路单元中，可以根据实际的需要选择其中的部分或全部单元来实现本申请方案的目的。

为了更好地理解本发明实施例，在此首先释明下面实施例可能涉及到的概念：

机器学习系统：基于输入数据和标签，通过梯度下降等优化方法训练机器学习模型的参数，最终利用训练得到的模型来完成未知数据的预测。

30 个性化推荐系统：根据用户的历史数据，利用机器学习算法进行分析和建模，并以此对新的用户请求进行预测，给出个性化的推荐结果的系统。

连续特征：特征按特征值是否连续可以划分为连续特征与离散特征两种。在一定区间内可以任意取值的特征称为连续特征，其数值是连续不断的，相邻两个数值可作无限分割，即可取无限个数值。

35 离散特征：离散特征指特征值可以按一定顺序一一列举，通常以整数位取值的特征。如用户性别、国籍、物品类型等。有些性质上属于连续特征的场景也按整数取值，即可以把它们作为离散特征来看待。

特征域和特征值：特征域是特征值的集合，如性别是特征域；特征值是某个特征域中的取值，如性别特征域中，男和女都是对应的特征值。

连续特征离散化：离散化处理是一种常见的数据预处理方法，用于将连续的数值属性转化为离散的数值属性。

- 5 向量表达是指用向量来表示具体的特征。本实施例中，该基向量表达值为可以为预设的初始化向量表达值，也可以是经过优化之后的向量表达值。而连续特征的向量表达值是根据该基向量表达值确定的用于表示该连续特征的具体特征的向量值。可以理解的是该连续特征的向量表达值与该基向量表达值的维度相同。一种示例性方案中，该基向量表达值可以为 5 维的向量值，比如 (01010)；该连续特征对应的向量表达值也同样为 5 维的向量值，比如 (11010)。

10 图 1 示出一种人工智能主体框架示意图，该主体框架描述了人工智能系统总体工作流程，适用于通用的人工智能领域需求。

下面从“智能信息链”（水平轴）和“IT 价值链”（垂直轴）两个维度对上述人工智能主题框架进行阐述。

- 15 “智能信息链”反映从数据的获取到处理的一系列过程。举例来说，可以是智能信息感知、智能信息表示与形成、智能推理、智能决策、智能执行与输出的一般过程。在这个过程中，数据经历了“数据—信息—知识—智慧”的凝练过程。

“IT 价值链”从人工智能的底层基础设施、信息（提供和处理技术实现）到系统的产业生态过程，反映人工智能为信息技术产业带来的价值。

- 20 (1) 基础设施：

基础设施为人工智能系统提供计算能力支持，实现与外部世界的沟通，并通过基础平台实现支撑。通过传感器与外部沟通；计算能力由智能芯片（CPU、NPU、GPU、ASIC、FPGA 等硬件加速芯片）提供；基础平台包括分布式计算框架及网络等相关的平台保障和支持，可以包括云存储和计算、互联互通网络等。举例来说，传感器和外部沟通获取数据，这些

- 25 数据提供给基础平台提供的分布式计算系统中的智能芯片进行计算。

(2) 数据

基础设施的上一层的数据用于表示人工智能领域的数据来源。数据涉及到图形、图像、语音、文本，还涉及到传统设备的物联网数据，包括已有系统的业务数据以及力、位移、液位、温度、湿度等感知数据。

- 30 (3) 数据处理

数据处理通常包括数据训练，机器学习，深度学习，搜索，推理，决策等方式。

其中，机器学习和深度学习可以对数据进行符号化和形式化的智能信息建模、抽取、预处理、训练等。

- 35 推理是指在计算机或智能系统中，模拟人类的智能推理方式，依据推理控制策略，利用形式化的信息进行机器思维和求解问题的过程，典型的功能是搜索与匹配。

决策是指智能信息经过推理后进行决策的过程，通常提供分类、排序、预测等功能。

(4) 通用能力

对数据经过上面提到的数据处理后,进一步基于数据处理的结果可以形成一些通用的能力,比如可以是算法或者一个通用系统,例如,翻译,文本的分析,计算机视觉的处理,语音识别,图像的识别等等。

(5) 智能产品及行业应用

5 智能产品及行业应用指人工智能系统在各领域的产品和应用,是对人工智能整体解决方案的封装,将智能信息决策产品化、实现落地应用,其应用领域主要包括:智能制造、智能交通、智能家居、智能医疗、智能安防、自动驾驶,平安城市,智能终端等。

本申请实施例提供的连续特征的处理方法可以应用于各种机器学习系统的应用场景。一种示例性方案中,该机器学习系统可以是个性化推荐系统。在该个性化推荐系统中,其具体处理流程可以如下图 2 所示:首先进行原始数据的采集,然后再通过特征处理得到可以输入到模型参与训练的特征数据;再将特征数据输入初始化模型,利用模型训练得到最终的推荐模型;最后将推荐模型线上服务模块,为用户产生推荐列表。而其主要结构可以如图 3 所示,该结构包括用户数据采集、日志存储模块、离线训练模块、预测模型以及线上预测模块。其基本运行逻辑为:用户在前端展示列表中进行一系列如浏览、点击、评论、
10 下载等行为产生行为数据(即前端进行用户数据采集);然后这些行为数据存储于日志模块中;该个性化推荐系统利用包括用户行为日志在内的数据进行离线的模型训练,在训练收敛后产生预测模型,再将模型部署在线上服务环境并基于用户的请求访问、物品特征和上下文信息给出推荐结果,然后用户对该推荐结果产生反馈形成新一轮的用户数据。在图 2 所示的流程图或者图 3 所示的结构图中,原始数据或用户数据(即日志数据)的特征处理
15 是模型训练、线上服务的基础,在机器学习系统中占有十分重要的地位。

下面结合模型架构图详细说明特征处理与机器学习模型或者深度学习模型的关系和作用。一种示例性方案如图 4 所示,该模型架构图以深度学习为例。模型架构图包括输入层,向量化表示层(Embedding),多层神经网络(Multi-layer Perceptron, MLP)与特征交互层和输出层。其中,输入层传入数据特征,包括连续特征和离散特征。在向量化表示层,
25 对该数据特征进行处理。具体如下:对于离散特征,使用 one-hot 进行稀疏编码,然后通过稀疏编码的 id 从 embedding table 中取出对应的 embedding 向量表达,并最终将所有离散特征的 embedding 向量表达按顺序拼接;对于连续特征通过离散化之后进行向量化表达;在获取每个特征的向量化表示之后,将所有连续特征和离散特征的 embedding 进行拼接,作为神经网络的输入数据;该输入数据通过 MLP 与特征交互层和输出层得到预测值。可以发现,连续特征的离散处理不仅用于连续特征的表示建模,而且直接影响上层 MLP 与特征交互层的参数学习,同时通过梯度的反向传播也间接影响了离散特征相关参数的学习,对
30 最终模型预测效果起到至关重要的作用。

如图 5 所示,本申请实施例提供一种数据处理装置 500,包括日志数据存储模块 501、连续特征离散化模块 502、向量表达聚合模块 503、向量表达拼接模块 504 以及输出模块
35 505。其中,该日志数据存储模块 501 用于采集前端用户反馈的行为数据并存储;该连续特征离散化模块 502 通过线性变换和 softmax、多分类神经网络或者注意力网络为该连续特征输出 N 个离散化概率值;该向量表达聚合模块 503 用于根据该 N 个离散化概率以及预置

的 N 个基向量表达值确定连续特征对应的一个向量表达值；最后通过向量表达拼接模块 504 将各个特征的向量表达值按序拼接作为机器学习或深度学习模型的输入数据；最后该输出模块 505 将该输入数据输出至机器学习或者深度学习模型。

在特征处理的过程中，日志数据存储模块 501 中还包括离散特征，因此该数据处理装置 500 还需要处理离散特征。同时该日志数据中的连续特征也可以先进行归一化处理。因此，在一种示例性方案中，与机器学习或者深度学习相结合的数据处理装置 600 可以如图 6 所示，其包括：日志数据存储模块 601、归一化模块 602、连续特征离散化模块 603、向量表达聚合模块 604，离散特征处理模块 605、向量表达拼接模块 606 以及输出模块 607。其中，该日志数据存储模块 601 用于采集前端用户反馈的行为数据并存储；该归一化模块 602 用于将该行为数据中的连续特征进行归一化处理；然后将归一化后的连续特征输入至该连续特征离散化模块 603；该连续特征离散化模块 603 通过线性映射和 softmax、多分类神经网络或者注意力网络为该连续特征输出 N 个离散化概率值；该向量表达聚合模块 604 用于根据该 N 个离散化概率以及预置的 N 个基向量表达值确定连续特征对应的一个向量表达值；而离散特征处理模块 605 用于将该行为数据中的离散特征使用 one-hot 进行稀疏编码，然后通过稀疏编码的 id 从 embedding table 中取出对应的 embedding 向量表达值。最后该向量表达拼接模块 606 将该离散特征的向量表达值与该连续特征的向量表达值按序拼接作为机器学习或深度学习模型的输入数据；最后该输出模块 607 将该输入数据输出至机器学习或者深度学习模型。

本申请实施例提供一种连续特征的处理方法，具体请参阅图 7 所示，该连续特征的处理方法具体步骤如下：

701、获取连续特征。

在模型训练过程或者模型应用过程中，将前端用户的行为数据进行数据存储，并作为模型训练过程或者模型应用过程中的样本数据。该样本数据包括连续特征和/或离散特征，本实施例中主要说明连续特征的处理方法。该数据处理装置从该样本数据中获取该连续特征。

可选的，该连续特征可以是由该数据处理装置进行归一化处理后的连续特征。一种示例性方案中，该连续特征可以根据公式 $\bar{X} = (x - \min) / (\max - \min)$ 进行归一化处理。

702、利用离散化模型对该连续特征进行离散化处理得到该连续特征中对应的 N 个离散化概率值，其中，该 N 为大于 1 的整数，该 N 个离散化概率值对应于预设的 N 个基向量表达值。

该数据处理装置在获取到该连续特征之后，将该连续特征输入到离散化模型中，将该连续特征中的进行离散化处理得到该连续特征中对应的 N 个离散化概率值。

本实施例中，该数据处理装置在利用离散化模型对所述连续特征中进行离散化处理得到所述连续特征对应的 N 个离散化概率值的具体方式可以如下：该数据处理装置预设该离散化模型中的初始变量；然后根据该初始变量确定该连续特征对应的 N 个映射值；最后根据该映射值计算得到该连续特征的 N 个离散化概率值。本实施例中，该初始变量可以为初

- 10 -

始映射域, 然后对于连续特征来说, 通过该初始变量可以计算得到该连续特征对应的 N 个映射值; 然后根据该 N 个映射值中的每一个映射值计算一个对应的概率值, 从而得到 N 个概率值, 此时该 N 个概率值作为该连续特征对应的 N 个离散化概率。

5 可选的, 该 N 的取值大于等于 20 且小于等于 100。在此范围内, 该离散化模型可以具有较好的应用效果。

在本申请实施提供的推荐系统中, 该离散化模型可以是多分类神经网络、注意力网络或者线性映射和 softmax。

一种示例性方案中, 以线性映射和 softmax 为例进行说明。

10 预设该连续特征的离散化后的桶数为 h (即相当于预设该连续特征的离散化后的分布区域为 h 个, 可以理解的是, 该 h 等于该 N), temperature 为 τ , 初始化线性映射变量 $W_{logit} = R^{1 \times h}$;

然后第一步骤将, 该连续特征根据线性映射公式进行线性映射得到 N 个映射值, 其中, 该线性映射公式为: $cont_{logit} = cont \cdot W_{logit}$;

第二步骤, 根据离散化公式和该 N 个映射值计算该连续特征对应的 N 个离散化概率值,

15 其中该离散化公式为: $cont_p_k = e^{\frac{1}{\tau} cont_{logit}^k} / \sum_{i=1}^h e^{\frac{1}{\tau} cont_{logit}^i}$; 其中, 所述 W_{logit} 用于指示线性映射变量, 所述 R 用于指示实数域, 所述 h 用于指示所述连续特征离散化后的桶数, 所述 $cont_{logit}$ 用于指示所述连续特征线性映射后的表达, 所述 $cont_p_k$ 用于指示所述连续特征离散化到第 k 个桶的概率值, 所述 $cont_{logit}^k$ 指示所述连续特征线性映射后第 k 个神经元输出, 所述 τ 用于指示 softmax 的温度控制系数, 所述 $cont_{logit}^i$ 用于指示所述连续特征线性映射后第 i 个神经元输出。

20 经过上述第一步骤和第二步骤之后, 该数据处理装置得到大小为 $1 \times h$ 的概率分布 $cont_p$, 表示该连续特征离散化到不同桶的概率。一种示例性方案中, 以年龄为例, 假设某个年龄值为 20, 对年龄这个域分配了 4 个桶, h_1, h_2, h_3, h_4 , 通过上述步骤, 会得到一个 1×4 的概率分布, 分别是 0.1、0.15、0.7 以及 0.05。即该年龄值为 20 在 4 个桶的概率分布为 (0.1、0.15、0.7 以及 0.05)。同时根据上述结果可以看出该年龄值为 20 分布在第三个桶的概率是最高的。

另一种示例性方案中, 以多分类神经网络为例进行说明。

30 预设该连续特征的离散化后的桶数为 h (即相当于预设该连续特征的离散化后的分布区域为 h 个), 然后第一步骤, 将该连续特征中的各个特征值 $cont$ 输入到多层感知机 MLP 中, MLP 的第 l 层公式为: $O^l = \sigma(O^{l-1} \cdot W^l + b^l)$, 其中 σ 是激活函数, 可以是 sigmoid 函数或 tanh 函数, W^l 是第 l 层的权重参数, b^l 是第 l 层的偏差参数, O^{l-1} 是上一层的输出, $O^0 = cont$ 即原始特征值是第一层的输入。

-11-

第二步骤，MLP 最后一层的激活函数设置为 softmax，公式为：
 $cont_p = O^L = softmax(O^{L-1} \cdot W^L + b^L)$ ，softmax 的公式为 $S_k = e^k / \sum_{i=1}^h e^i$ ，其中， $cont_p$ 用于指示所述连续特征离散化到每个桶的概率值。

703、根据所述N个离散化概率值和所述N个基向量表达值确定所述连续特征的一个向量表达值。

该数据处理装置为该连续特征中的每一个连续特征域初始化 N 个向量表达值（也称为 Meta Embedding），即 $V = R^{h \times e}$ ，其中该 e 用于表示该向量表达值的描述维度（也称为 Embedding size）。上述向量表达值将用于表示一个连续特征域的所有特征值。然后，该数据处理装置根据步骤 702 中获取的离散化概率，通过聚合函数对该 Meta Embedding 进行聚合得到该连续特征中对应的向量表达值（也称为 embedding vx_cont ）。其中，该聚合函数用于将连续特征域中的对应的 N 个 meta-embedding 聚合为该特征值对应的一个 embedding。

一种示例性方案中，该连续特征域对应的 Meta Embedding 中的向量表达与特征值的离散化概率是一一对应的关系。假设该离散化概率为（a1、a2、a3、a4），该向量表达值为（b1、b2、b3、b4），则该 a1 对应 b1，a2 对应 b2，a3 对应 b3，a4 对应 b4。该聚合函数用于将该特征值对应的 meta-embedding（b1、b2、b3、b4）聚合为一个 embedding。一种示例性方案中，可以从（b1、b2、b3、b4）聚合得到 b3 作为该特征值对应的向量表达值。

可选的，该聚合函数可以有多种，具体情况可以如下：

一种可能实现方式中，该聚合函数为 Max-Pooling，从步骤 702 中获取得到的离散化概率中确定最大值，然后从 Meta Embedding 中获取与该最大值对应的 embedding 作为该特征值对应的向量表达值，其计算公式为：

$vx_cont = E^k, where k = \arg \max_h \{cont - p_h\}$ 。例如，假设该离散化概率为

（a1、a2、a3、a4），该 meta-embedding 为（b1、b2、b3、b4），则该 a1 对应 b1，a2 对应 b2，a3 对应 b3，a4 对应 b4。若该 a3 的值最大，则该 b3 作为该特征值的向量表达值。

另一种可能实现方式中，该聚合函数为 Top-K-Sum，从步骤 702 中获取得到的离散化概率中选择最大的 k 个概率值，然后再从 Meta Embedding 中获取与该概率值对应的 embedding 并求和，作为该特征值对应的向量表达值，其计算公式为：

$vx_cont = \sum_{k=1}^K E^k, where k = \arg \text{top}k_h \{cont - p_h\}$ 。例如，假设该离散化概

率为（a1、a2、a3、a4），该 meta-embedding 为（b1、b2、b3、b4），则该 a1 对应 b1，a2 对应 b2，a3 对应 b3，a4 对应 b4。若该 k 取值为 2，且该 a2 和 a3 的值分别为前两位，则该特征值的向量表达值为 b2+b3。

另一种可能实现方式中，该聚合函数为 Weighted-Average，从步骤 702 中获取得到离散化概率，然后将 N 个概率值和 Meta Embedding 进行加权求和，作为该特征值对应的向量

表达值，其计算公式为： $vx_cont = \sum_{k=1}^h cont - p_k \times E^k$ 。例如，假设该离散化概

率为 (a1、a2、a3、a4)，该 meta-embedding 为 (b1、b2、b3、b4)，则该 a1 对应 b1，a2 对应 b2，a3 对应 b3，a4 对应 b4。该特征值的向量表达值等于 (a1*b1+ a2*b2+ a3*b3+ a4*b4)。

一种示例性方案中，若以年龄为例，假设某个年龄值为 20，对年龄这个域分配了 4 个桶，h1, h2, h3, h4，通过上述步骤，会得到一个 1*4 的概率分布，分别是 0.1、0.15、0.7 以及 0.05。即该年龄值为 20 在 4 个桶的概率分布可以看出分布在第三个桶的概率是最高的。若该数据处理装置选择 Max-Pooling 聚合函数进行计算，则该数据处理装置将会选择概率为 0.7 的桶，并将其对应的 embedding 做为该特征值的向量表达值。

本实施例中，该数据处理装置通过将离散化模型为每一个连续特征的特征值计算得到大于 1 维的离散化概率并为该连续特征中每个连续特征域预设大于 1 维的元向量表达值 meta-embedding，并利用聚合函数和该离散化概率从该 meta-embedding 中为特征值确定一个向量表达值。这样对于连续特征来说，与现有技术相比其学习表达的向量表达值具有更好的表达能力，有助于提升预测结果的精确性。

本实施例中，上述图 7 所示的数据处理方法可以应用于多种应用场景，比如推荐模型或搜索模型。下面以图 8 所示的手机应用市场推荐系统中的点击率预测场景为例说明本申请实施例提供的数据处理方法的应用场景。在该图 8 所示的应用场景下，其具体的数据模型为点击率预测模型（或者说推荐模型），其主要用于在图 8 所示的“精品应用”一栏根据用户特征（如用户年龄或者用户性别）和对象特征（应用程序）为用户推荐相应的应用程序（如图 8 所示，在精品应用中显示“拼多多”、“流利说-阅读”等应用程序）。其具体处理流程可以如下：获取用户特征和对象特征；然后将该用户特征和对象特征中的离散特征采用传统处理进行处理。即先进行 one-hot 编码，然后通过 embedding lookup 操作获取 embedding 表达。而对于用户特征和对像特征中的连续特征采用图 7 所示方法进行处理得到对应的向量表达值；然后该用户特征和对象特征中的离散特征的向量表达值和连续特征的向量表达值作为模型的输入特征表达输入该图 8 所示的应用场景对应的推荐模型得到推荐结果。

本实施例中，该推荐模型还可以利用预测结果和实际结果利用损失函数计算损失值 (loss)，基于 loss 完成推荐模型和离散化模型的参数更新。在线服务时，该该数据处理装置可作为推荐模型的一部分使用，在线完成连续特征离散化，学习每个连续特征的 embedding。相比于人工特征工程、分桶离散化预处理等技术可以节约处理时间。当使用增量训练时，离散化模型的权重参数可随着最新的数据分布进行调整，从而利用数据的效率更高。

下面可以具体的实验的数据对本申请提供的连续特征的处理方法进行说明，本实施例中提供了三个数据集，分别为 Criteo 数据集、AutoML 数据集和华为工业数据集。一种示例性方案中，各个数据集的统计信息如表 1 所示。

表 1

数据集名称	数据集大小	全部特征量	离散特征量	连续特征量
Criteo	45.8M	39	26	13

AutoML	4.69M	74	51	23
华为工业	8.75M	85	44	41

其中，该 M 等于 10 的 6 次方。

本实施例中，实验评价指示为 AUC（即为 area under curve），连续特征处理技术分别为归一化方法、等距、Logarithms、DeepGBM 以及本申请实施例提供的连续特征处理技术。在上述三个数据集上进行实验，以 DeepFM 作为顶层深度模型为例，实验结果如表 2 所示：

表 2

	Criteo	AutoML	Industrial
DeepFM-Norm	0.8107	0.7523	0.7248
DeepFM-EDD	0.8125	0.7545	0.7251
DeepFM-LD	0.8138	0.7527	0.7265
DeepFM-TD	0.8130	0.7531	0.7262
DeepFM-AutoDis	0.8149	0.7556	0.7277
% Impr.	0.14%	0.15%	0.17%

其中，该 AutoDis 用于指示本申请实施例中执行数据处理方法的框架或装置。由上述结果可知，本实施例提供的技术方案可以达到更好的结果。

同时，本实施例提供的技术方案可以应用于不同模型，同时也具有提升效果。本实施例选取了几种工业界常用的深度模型，用于进行点击通过率（click-through-rate, CTR）预测，包括了前馈神经网络（factorisation-machine supported neural networks, FNN）、Wide&Deep（即具有具有稀疏特征和变换的逻辑回归的 wide 模型与具有 Embedding 层和多个隐藏层的 feed-forward 神经网络的 deep 模型的联合训练模型）、DeepFM、DCN 和 IPNN 等，实验结果如表 3 所示：

表 3

	Criteo		AutoML		Industrial	
	基础模型	+AutoDis	基础模型	+AutoDis	基础模型	+AutoDis
FNN	0.8059	0.8091	0.7383	0.7448	0.7271	0.7286
Wide&Deep	0.8097	0.8121	0.7407	0.7442	0.7275	0.7287
DeepFM	0.8108	0.8149	0.7525	0.7556	0.7262	0.7277
DCN	0.8091	0.8128	0.7489	0.7508	0.7262	0.7281
IPNN	0.8101	0.8135	0.7519	0.7541	0.7269	0.7283

根据上述表 3 所示的结果可知, 在这些常用的深度模型中, 加入本实施例提供的连续特征的处理方法, 可以显著提升模型性能, 显示出连续特征的处理方法具有良好的兼容性。

图 9 示出了上述实施例中一种数据处理装置 900 可能的结构示意图, 该数据处理装置 900 可以配置成是前述数据处理装置。该数据处理装置 900 可以包括: 处理器 902、计算机可读存储介质/存储器 903、收发器 904、输入设备 905 和输出设备 906, 以及总线 901。其中, 处理器, 收发器, 计算机可读存储介质等通过总线连接。本申请实施例不限定上述部件之间的具体连接介质。

一个示例中, 该收发器 904 获取连续特征;

该处理器 902 利用离散化模型对所述连续特征进行离散化处理得到所述连续特征对应的 N 个离散化概率值, 所述 N 为大于 1 的整数; 所述 N 个离散化概率值对应于预设的 N 个基向量表达值; 根据所述 N 个离散化概率值和所述 N 个基向量表达值确定所述连续特征的一个向量表达值。

又一个示例中, 处理器 902 可以运行操作系统, 控制各个设备和器件之间的功能。收发器 904 可以包括基带电路和射频电路, 例如, 可以对向量表达值经由基带电路, 射频电路进行处理后发送给推荐系统或搜索系统。

该收发器 904 与该处理器 902 可以实现上述图 7 至 8 中任一实施例中相应的步骤, 具体此处不做赘述。

可以理解的是, 图 9 仅仅示出了数据处理装置的简化设计, 在实际应用中, 数据处理装置可以包含任意数量的收发器, 处理器, 存储器等, 而所有的可以实现本申请的数据处理装置都在本申请的保护范围之内。

上述装置 900 中涉及的处理器 902 可以是通用处理器, 例如 CPU、网络处理器(network processor, NP)、微处理器等, 也可以是 ASIC, 或一个或多个用于控制本申请方案程序执行的集成电路。还可以是数字信号处理器(digital signal processor, DSP)、现场可编程门阵列(field-programmable gate array, FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。控制器/处理器也可以是实现计算功能的组合, 例如包含一个或多个微处理器组合, DSP 和微处理器的组合等等。处理器通常是基于存储器内存储的程序指令来执行逻辑和算术运算。

上述涉及的总线 901 可以是外设部件互连标准(peripheral component interconnect, 简称 PCI)总线或扩展工业标准结构(extended industry standard architecture, 简称 EISA)总线等。该总线可以分为地址总线、数据总线、控制总线等。为便于表示, 图 8 中仅用一条粗线表示, 但并不表示仅有一根总线或一种类型的总线。

上述涉及的计算机可读存储介质/存储器 903 还可以保存有操作系统和其他应用程序。具体地, 程序可以包括程序代码, 程序代码包括计算机操作指令。更具体的, 上述存储器可以是 ROM、可存储静态信息和指令的其他类型的静态存储设备、RAM、可存储信息和指令的其他类型的动态存储设备、磁盘存储器等等。存储器 903 可以是上述存储类型的组合。并且上述计算机可读存储介质/存储器可以在处理器中, 还可以在处理器的外部, 或

在包括处理器或处理电路的多个实体上分布。上述计算机可读存储介质/存储器可以具体体现在计算机程序产品中。举例而言，计算机程序产品可以包括封装材料中的计算机可读介质。

可以替换的，本申请实施例还提供一种通用处理系统，例如通称为芯片，该通用处理系统包括：提供处理器功能的一个或多个微处理器；以及提供存储介质的至少一部分的外部存储器，所有这些都通过外部总线体系结构与其它支持电路连接在一起。当存储器存储的指令被处理器执行时，使得处理器执行数据处理装置在图 7 至图 8 该实施例中的数据传输方法中的部分或全部步骤，和/或用于本申请所描述的技术的其它过程。

结合本申请公开内容所描述的方法或者算法的步骤可以硬件的方式来实现，也可以是由处理器执行软件指令的方式来实现。软件指令可以由相应的软件模块组成，软件模块可以被存放于 RAM 存储器、闪存、ROM 存储器、EPROM 存储器、EEPROM 存储器、寄存器、硬盘、移动硬盘、CD-ROM 或者本领域熟知的任何其它形式的存储介质中。一种示例性的存储介质耦合至处理器，从而使处理器能够从该存储介质读取信息，且可向该存储介质写入信息。当然，存储介质也可以是处理器的组成部分。处理器和存储介质可以位于 ASIC 中。另外，该 ASIC 可以位于终端中。当然，处理器和存储介质也可以作为分立组件存在于数据处理装置中。

具体请参阅图 10 所示，本申请实施例示出了上述实施例中一种数据处理装置 1000 可能的结构示意图，该数据处理装置 1000 包括获取模块 1001 和处理模块 1002，其中该获取模块 1001 与该处理模块 1002 通过总线相连。数据处理装置 1000 可以是上述方法实施例中的数据处理的装置，也可以配置为数据处理装置内的一个或多个芯片。数据处理装置 1000 可以用于执行上述方法实施例中的数据处理的装置的部分或全部功能。同时，图 10 仅示出了该数据处理装置涉及到本申请实施例的部分模块。

该获取模块 1001 用于获取连续特征；

该处理模块 1002 用于利用离散化模型对所述连续特征进行离散化处理得到所述连续特征对应的 N 个离散化概率值，所述 N 为大于 1 的整数；所述 N 个离散化概率值对应于预设的 N 个基向量表达值；根据所述 N 个离散化概率值和所述 N 个基向量表达值确定所述连续特征的一个向量表达值。该处理模块 1002 还可以执行上述图 5 中连续特征离散化模块 502 和向量表达聚合模块 503 所执行的方法或者，该处理模块 1002 还可以执行上述图 6 中连续特征离散化模块 603 和向量表达聚合模块 604 所执行的方法。

可选的，该数据处理装置 1000 还包括存储模块，该存储模块可以存储执行指令，此时该存储模块与处理模块耦合，使得处理模块可执行存储模块中存储的计算机执行指令以实现上述方法实施例中数据处理装置的功能。在一个示例中，数据处理装置 1000 中可选的包括的存储模块可以为芯片内的存储单元，如寄存器、缓存等，该存储单元还可以是位于芯片外部的存储单元，如 ROM 或可存储静态信息和指令的其他类型的静态存储设备，RAM 等。

应理解，上述图 10 对应实施例中数据处理装置各模块之间所执行的流程与前述图 7 中对应方法实施例中的数据处理的装置执行的流程类似，具体此处不再赘述。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或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）执行本申请各个实施例所述方法的全部或部分步骤。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）、随机存取存储器（RAM，Random Access Memory）、磁碟或者光盘等各种可以存储程序代码的介质。

-17-

权 利 要 求

1. 一种数据处理方法，其特征在于，包括：
 - 获取连续特征；
 - 利用离散化模型对所述连续特征进行离散化处理得到所述连续特征对应的 N 个离散化
- 5 概率值，所述 N 为大于 1 的整数；
 - 所述 N 个离散化概率值对应于预设的 N 个基向量表达值；
 - 根据所述 N 个离散化概率值和所述 N 个基向量表达值确定所述连续特征的一个向量表
- 达值。
2. 根据权利要求 1 所述的方法，其特征在于，所述利用离散化模型对所述连续特征进
- 10 行离散化处理得到所述连续特征对应的 N 个离散化概率值包括：
 - 预设所述离散化模型中初始变量；
 - 根据所述初始变量确定所述连续特征对应的 N 个映射值；
 - 根据所述映射值计算得到所述连续特征的 N 个离散化概率。
3. 根据权利要求 1 或 2 所述的方法，其特征在于，所述离散化模型为多分类神经网络、
- 15 注意力网络或线性映射和 softmax。
4. 根据权利要求 3 所述的方法，其特征在于，在所述离散化模型为线性映射和 softmax
- 时，所述预设所述离散化模型中初始变量包括：包括：
 - 预设初始化线性映射变量为： $W_{logit} = R^{1 \times h}$ ，所述初始化线性映射变量为所述初始变
 - 量；
- 20 所述根据所述 N 确定所述连续特征对应的 N 个映射值包括：
 - 通过线性映射公式确定所述连续特征对应的 N 个映射值，
 - 其中，所述线性映射公式为： $cont_{logit} = cont \cdot W_{logit}$ ；
 - 根据所述映射值计算得到所述连续特征的 N 个离散化概率值包括：
 - 根据离散化公式计算所述 N 个映射值中每一个映射值对应的概率值得到 N 个概率值，
- 25 其中所述离散化公式为： $cont - p_k = e^{\frac{1}{\tau} cont_{logit}^k} / \sum_{i=1}^h e^{\frac{1}{\tau} cont_{logit}^i}$ ，所述 N 个概率值
- 作为所述 N 个离散化概率；
 - 其中，所述 W_{logit} 用于指示线性映射变量，所述 R 用于指示实数域，所述 h 用于指示所
 - 述连续特征离散化后的桶数，所述 h 等于所述 N，所述 $cont_{logit}$ 用于指示所述连续特征线性
 - 映射后的表达，所述 $cont - p_k$ 用于指示所述连续特征离散化到第 k 个桶的概率值，所述
 - 30 $cont_{logit}^k$ 指示所述连续特征线性映射后第 k 个神经元输出，所述 τ 用于指示 softmax 的温度
 - 控制系数，所述 $cont_{logit}^i$ 用于指示所述连续特征线性映射后第 i 个神经元输出。
5. 根据权利要求 1 至 4 中任一项所述的方法，其特征在于，所述根据所述 N 个离散化
- 概率值和 N 个基向量表达值确定所述连续特征的一个向量表达值包括：
 - 利用所述聚合函数根据所述 N 个离散化概率值和 N 个基向量表达值确定所述连续特征

的一个向量表达值, 其中, 所述聚合函数为 Max-Pooling、Top-K-Sum 或 Weighted-Average。

6. 根据权利要求 1 至 5 中任一项所述的方法, 其特征在于, 所述方法还包括:
将用户特征和对象特征输入推荐模型或搜索模型得到预测结果;
所述用户特征或所述对象特征中包括所述向量表达值;
- 5 所述用户特征或所述待推荐对象特征采用所述向量表达值表示所述连续特征。
7. 根据权利要求 6 所述的方法, 其特征在于, 所述方法还包括:
获取实际结果;
根据所述预测结果和所述实际结果利用损失函数调整所述离散化模型的权重参数。
8. 根据权利要求 1 至 7 中任一项所述的方法, 其特征在于, 所述 N 的取值大于等于 20
- 10 且小于等于 100。
9. 根据权利要求 1 至 8 中任一项所述的方法, 所述连续特征为样本数据中具有连续统计特征值的特征数据。
10. 根据权利要求 9 所述的方法, 其特征在于, 所述连续特征包括但不限于推荐系统中年龄特征、点击量特征和分数特征。
- 15 11. 根据权利要求 1 至 10 中任一项所述的方法, 其特征在于, 所述连续特征为经过归一化处理的连续特征。
12. 一种数据处理装置, 其特征在于, 包括:
获取模块, 用于获取连续特征;
处理模块, 用于利用离散化模型对所述连续特征进行离散化处理得到所述连续特征对应的 N 个离散化概率值, 所述 N 为大于 1 的整数; 所述 N 个离散化概率值对应于预设的 N 个基向量表达值; 根据所述 N 个离散化概率值和所述 N 个基向量表达值确定所述连续特征的一个向量表达值。
- 20 13. 根据权利要求 12 所述的装置, 其特征在于, 所述处理模块, 具体用于预设所述离散化模型中初始变量; 根据所述初始变量确定所述连续特征中对应的 N 个映射值; 根据所述映射值计算得到所述连续特征的 N 个离散化概率值。
- 25 14. 根据权利要求 12 或 13 所述的装置, 其特征在于, 所述离散化模型为多分类神经网络、注意力网络或线性映射和 softmax。
15. 根据权利要求 14 所述的装置, 其特征在于, 在所述离散化模型为线性映射和 softmax 时, 所述处理模块, 具体用于预设初始化线性映射变量为: $W_{logit} = R^{1 \times h}$, 所述初始化线性映射变量为所述初始变量;
- 30 通过线性映射公式确定所述连续特征对应的 N 个映射值,
其中, 所述线性映射公式为: $cont_{logit} = cont \cdot W_{logit}$;
根据离散化公式计算所述 N 个映射值中每一个映射值对应的概率值得到 N 个概率值,
其中所述离散化公式为: $cont_p_k = e^{\frac{1}{\tau} cont_{logit}^k} / \sum_{i=1}^h e^{\frac{1}{\tau} cont_{logit}^i}$, 所述 N 个概率值
- 35 作为所述 N 个离散化概率值;

其中, 所述 W_{logit} 用于指示线性映射变量, 所述 R 用于指示实数域, 所述 h 用于指示所述连续特征离散化后的桶数, 所述 h 等于所述 N , 所述 $cont_{logit}$ 用于指示所述连续特征线性映射后的表达, 所述 $cont_p_k$ 用于指示所述连续特征离散化到第 k 个桶的概率值, 所述 $cont_{logit}^k$ 指示所述连续特征线性映射后第 k 个神经元输出, 所述 τ 用于指示 softmax 的温度

5 控制系数, 所述 $cont_{logit}^i$ 用于指示所述连续特征线性映射后第 i 个神经元输出。

16. 根据权利要求 12 至 15 中任一项所述的装置, 其特征在于, 所述处理模块, 具体用于利用聚合函数根据所述 N 个离散化概率和 N 个基向量表达值确定所述连续特征对应的一个向量表达值, 其中, 所述聚合函数为 Max-Pooling、Top-K-Sum 或 Weighted-Average。

10 17. 根据权利要求 12 至 16 中任一项所述的装置, 其特征在于, 所述处理模块, 还用于将用户特征和对象特征输入推荐模型或搜索模型得到预测结果; 所述用户特征或所述对象特征中包括所述向量表达值; 所述用户特征或所述待推荐对象特征采用所述向量表达值表示所述连续特征。

15 18. 根据权利要求 17 所述的装置, 其特征在于, 所述获取模块, 还用于获取实际结果; 所述处理模块, 还用于根据所述预测结果和所述实际结果利用损失函数调整所述离散化模型的权重参数。

19. 根据权利要求 12 至 18 中任一项所述的装置, 其特征在于, 所述 N 的取值大于等于 20 且小于等于 100

20. 根据权利要求 12 至 19 中任一项所述的装置, 所述连续特征为样本数据中具有连续统计特征值的特征数据。

20 21. 根据权利要求 20 所述的装置, 其特征在于, 所述连续特征包括但不限于推荐系统中的年龄特征、点击量特征和分数特征。

22. 根据权利要求 12 至 20 中任一项所述的装置, 其特征在于, 所述连续特征为经过归一化处理的连续特征。

25 23. 一种数据处理装置, 其特征在于, 包括至少一个处理器和存储器, 所述处理器用于与所述存储器耦合, 所述处理器调用所述存储器中存储的指令以控制所述数据处理装置执行权利要求 1 至 11 中任一项所述的方法。

24. 一种计算机存储介质, 所述计算机存储介质存储有计算机指令, 所述计算机指令用于执行上述权利要求 1 至权利要求 11 中任意任一项所述的方法。

30 25. 一种包含指令的计算机程序产品, 当其在计算机上运行时, 使得计算机执行上述权利要求 1 至权利要求 11 中任意任一项所述的方法。