



(21) 申请号 202510131948.9

(22) 申请日 2025.02.06

(65) 同一申请的已公布的文献号

申请公布号 CN 119577402 A

(43) 申请公布日 2025.03.07

(73) 专利权人 北京科杰科技有限公司

地址 100096 北京市海淀区东升科技园北

街6号院10号楼11层101

(72) 发明人 高海玲 高经郡 董明慧

(74) 专利代理机构 北京智行阳光知识产权代理

事务所(普通合伙) 11738

专利代理师 张毅群

(51) Int.Cl.

G06F 18/2113 (2023.01)

G06F 18/25 (2023.01)

G06F 18/2323 (2023.01)

G06F 18/15 (2023.01)

G06F 18/213 (2023.01)

G06F 18/214 (2023.01)

G06F 18/22 (2023.01)

G06N 3/0455 (2023.01)

G06N 3/0464 (2023.01)

G06N 3/0442 (2023.01)

G06N 3/042 (2023.01)

G06N 3/084 (2023.01)

G06N 3/048 (2023.01)

(56) 对比文件

CN 117333770 A, 2024.01.02

CN 119097322 A, 2024.12.10

审查员 黄睿

权利要求书5页 说明书19页 附图2页

(54) 发明名称

基于Transformer的结构化数据建模分析方法

法

(57) 摘要

本发明提供一种基于Transformer的结构化数据建模分析方法,涉及数据分析技术领域,包括通过双流特征编码器对数值型特征和类别型特征分别执行分段线性变换和自适应量子编码,并通过门控机制融合得到融合特征;将增强训练样本输入包含层次注意力模块和自校准模块的特征提取网络,所述层次注意力模块计算样本特征的相似度矩阵,基于所述相似度矩阵构建特征亲和图,通过图分割算法将所述特征亲和图划分为多个特征子图;在每个所述特征子图内,采用因果注意力机制提取局部特征表示,并通过跨子图的注意力融合机制整合多个局部特征表示,得到全局特征表示;所述自校准模块接收所述全局特征表示,通过残差连接和动态归一化进行特征非线性变换,得到深层特征表示

构建信息瓶颈特征筛选器,通过最大化任务相关信息和最小化冗余信息,对所述深层特征表示进行筛选,得到关键特征子集;将所述关键特征子集输入包含主预测分支和辅助重建分支的双分支网络,所述主预测分支基于注意力机制对所述关键特征子集进行处理得到特征权重向量,并根据所述特征权重向量生成预测结果,所述辅助重建分支通过重建所述关键特征子集提供正则化约束;根据所述特征权重向量和所述预测结果,采用结构化梯度方法计算所述关键特征子集中各特征对所述预测结果的贡献度,生成分析结果

接收结构化原始数据,将所述结构化原始数据划分为数值型特征和类别型特征,通过双流特征编码器,对所述数值型特征执行分段线性变换,得到归一化主干特征,对所述类别型特征执行自适应量子编码,得到连续值映射特征;通过可学习的门控机制进行自适应权重分配,融合所述归一化主干特征和所述连续值映射特征,得到融合特征;基于所述融合特征的重要性评分进行动态采样,生成增强训练样本集

S101

将所述增强训练样本集输入包含层次注意力模块和自校准模块的特征提取网络,所述层次注意力模块计算样本特征的相似度矩阵,基于所述相似度矩阵构建特征亲和图,通过图分割算法将所述特征亲和图划分为多个特征子图;在每个所述特征子图内,采用因果注意力机制提取局部特征表示,并通过跨子图的注意力融合机制整合多个局部特征表示,得到全局特征表示;所述自校准模块接收所述全局特征表示,通过残差连接和动态归一化进行特征非线性变换,得到深层特征表示

S102

构建信息瓶颈特征筛选器,通过最大化任务相关信息和最小化冗余信息,对所述深层特征表示进行筛选,得到关键特征子集;将所述关键特征子集输入包含主预测分支和辅助重建分支的双分支网络,所述主预测分支基于注意力机制对所述关键特征子集进行处理得到特征权重向量,并根据所述特征权重向量生成预测结果,所述辅助重建分支通过重建所述关键特征子集提供正则化约束;根据所述特征权重向量和所述预测结果,采用结构化梯度方法计算所述关键特征子集中各特征对所述预测结果的贡献度,生成分析结果

S103

1. 基于Transformer的结构化数据建模分析方法,其特征在于,包括:

接收结构化原始数据,将所述结构化原始数据划分为数值型特征和类别型特征,其中所述数值型特征包括用户年龄、访问次数、消费总额、最近购买距今天数、收藏商品数、购物车商品数,所述类别型特征包括性别、会员等级、设备类型、偏好购物品类、收货地区;通过双流特征编码器,对所述数值型特征执行分段线性变换,得到归一化主干特征,对所述类别型特征执行自适应量子编码,得到连续值映射特征;通过可学习的门控机制进行自适应权重分配,融合所述归一化主干特征和所述连续值映射特征,得到融合特征;基于所述融合特征的重要性评分进行动态采样,生成增强训练样本集;

将所述增强训练样本集输入包含层次注意力模块和自校准模块的特征提取网络,所述层次注意力模块计算样本特征的相似度矩阵,基于所述相似度矩阵构建特征亲和图,通过图分割算法将所述特征亲和图划分为多个特征子图;在每个所述特征子图内,采用因果注意力机制提取局部特征表示,并通过跨子图的注意力融合机制整合多个局部特征表示,得到全局特征表示;所述自校准模块接收所述全局特征表示,通过残差连接和动态归一化进行特征非线性变换,得到深层特征表示;

构建信息瓶颈特征筛选器,通过最大化任务相关信息和最小化冗余信息,对所述深层特征表示进行筛选,得到关键特征子集;将所述关键特征子集输入包含主预测分支和辅助重建分支的双分支网络,所述主预测分支基于注意力机制对所述关键特征子集进行处理得到特征权重向量,并根据所述特征权重向量生成预测结果,所述辅助重建分支通过重建所述关键特征子集提供正则化约束;根据所述特征权重向量和所述预测结果,采用结构化梯度方法计算所述关键特征子集中各特征对所述预测结果的贡献度,生成用户购买意向预测的分析结果。

2. 根据权利要求1所述的方法,其特征在于,接收结构化原始数据,将所述结构化原始数据划分为数值型特征和类别型特征,通过双流特征编码器,对所述数值型特征执行分段线性变换,得到归一化主干特征,对所述类别型特征执行自适应量子编码,得到连续值映射特征;通过可学习的门控机制进行自适应权重分配,融合所述归一化主干特征和所述连续值映射特征,得到融合特征;基于所述融合特征的重要性评分进行动态采样,生成增强训练样本集包括:

接收结构化原始数据,将所述结构化原始数据划分为数值型特征和类别型特征;将数值型特征和类别型特征输入双流特征编码器进行特征处理,所述双流特征编码器包括特征主干流和特征映射流;

在特征主干流中,统计所述数值型特征的数据分布,得到数据密度,基于所述数据密度确定最优分段点的位置,根据所述最优分段点,将数值型特征划分为多个分段区间,在每个所述分段区间,构造线性变换参数矩阵,计算相邻的分段区间之间的特征差异值,通过最小化所述特征差异值,确定各分段区间的优化线性变换参数矩阵,使用优化线性变换参数矩阵对各分段区间的数值型特征执行线性变换,得到归一化主干特征;

在特征映射流中,将所述类别型特征中的每个类别值映射为对应的量子基态,为每个所述量子基态分配初始振幅参数和相位参数,构建包含旋转门和控制门的量子门序列,将所述量子门序列依次作用于所述量子基态,通过量子态变换得到变换量子态,计算变换量子态的熵值,基于所述熵值将变换量子态投影到连续空间,得到连续值映射特征;

分别计算所述归一化主干特征和所述连续值映射特征各维度之间的相关性系数,将所述相关性系数组合构建特征相关性系数矩阵,基于所述特征相关性系数矩阵训练预设的注意力网络,得到特征维度的重要度向量,使用所述重要度向量对所述归一化主干特征和所述连续值映射特征进行加权组合得到融合特征;

计算所述融合特征的各个维度对目标任务的贡献度,基于所述贡献度生成每个维度的特征重要性分数,根据所述特征重要性分数构建采样概率分布函数,使用所述采样概率分布函数对训练样本进行重要性采样,生成增强训练样本集。

3. 根据权利要求2所述的方法,其特征在于,统计所述数值型特征的数据分布,得到数据密度,基于所述数据密度确定最优分段点的位置包括:

基于所述数值型特征的四分位距和样本数量,计算基础区间宽度,所述数值型特征包括用户年龄;基于基础区间宽度,将所述数值型特征的值域范围划分为多个基础统计区间,对每个所述基础统计区间内的样本数量进行统计得到区间频数序列;

计算所述区间频数序列中相邻基础统计区间的频数比值,得到区间相似度,将所述区间相似度大于预设合并阈值的相邻的基础统计区间合并为新统计区间,得到合并频数分布序列;

根据所述合并频数分布序列计算累积频数分布,基于所述累积频数分布生成数据密度函数;在所述合并频数分布序列中选取候选分割位置,计算每个所述候选分割位置的两侧区间的样本数量和数值均值,利用所述样本数量和所述数值均值计算所述候选分割位置的区间方差值;

在当前待分割区间内选择具有最大区间方差值的位置作为待定分段点,计算所述待定分段点的区间方差值与所述数值型特征的总体方差的比值,确定所述待定分段点的方差贡献比;

对所述待定分段点的两侧区间计算局部分布偏度,当所述局部分布偏度的差值大于预设偏度阈值,且所述方差贡献比大于预设分割阈值时,将所述待定分段点确定为最优分段点,重复执行,确定所有最优分段点的位置。

4. 根据权利要求1所述的方法,其特征在于,将所述增强训练样本集输入包含层次注意力模块和自校准模块的特征提取网络,所述层次注意力模块计算样本特征的相似度矩阵,基于所述相似度矩阵构建特征亲和图,通过图分割算法将所述特征亲和图划分为多个特征子图;在每个所述特征子图内,采用因果注意力机制提取局部特征表示,并通过跨子图的注意力融合机制整合多个局部特征表示,得到全局特征表示;所述自校准模块接收所述全局特征表示,通过残差连接和动态归一化进行特征非线性变换,得到深层特征表示包括:

将增强训练样本集输入卷积神经网络得到初始特征映射,计算所述初始特征映射中任意两个样本特征的点积,并将所述点积除以对应特征向量的模长乘积,得到特征相似度矩阵;基于所述特征相似度矩阵构建特征亲和图,所述特征亲和图的顶点对应样本特征,所述特征亲和图的边权重对应所述特征相似度矩阵中的相似度值,通过谱聚类算法将所述特征亲和图划分为多个特征子图;

在每个所述特征子图内,将所述特征子图的特征序列划分为查询特征序列和键值对特征序列,构建时序因果掩码矩阵,利用所述时序因果掩码矩阵对所述查询特征序列与所述键值对特征序列的交互结果进行屏蔽,得到仅包含历史特征信息的初始注意力权重;计算

每个所述特征子图内样本特征的空间距离矩阵,基于所述空间距离矩阵计算所述特征子图的空间分布熵,根据所述空间分布熵对所述初始注意力权重进行自适应调节,得到自适应注意力权重;基于所述自适应注意力权重对所述键值对特征序列进行加权,并叠加位置编码信息,得到每个所述特征子图的局部特征表示;

基于预先构建的多层感知机网络,输入相邻特征子图的局部特征表示的拼接向量,经所述多层感知机网络前向传播,得到特征子图间的关联权重,基于所述关联权重对所述局部特征表示进行加权求和,得到全局特征表示;

对所述全局特征表示在特征维度上分别计算均值和标准差;基于预先构建的动态归一化网络,将所述全局特征表示输入所述动态归一化网络,经前向传播得到特征变换系数和特征偏移量;将所述全局特征表示减去所述均值后除以所述标准差得到标准化特征,将所述标准化特征与所述特征变换系数相乘,并加上所述特征偏移量得到变换特征,将所述变换特征与所述全局特征表示相加,得到自校准特征;

对所述自校准特征进行层归一化处理,并通过ReLU激活函数进行非线性变换,得到深层特征表示。

5. 根据权利要求4所述的方法,其特征在于,将所述特征子图的特征序列划分为查询特征序列和键值对特征序列,构建时序因果掩码矩阵,利用所述时序因果掩码矩阵对所述查询特征序列与所述键值对特征序列的交互结果进行屏蔽,得到仅包含历史特征信息的初始注意力权重包括:

所述特征子图包括用户连续30天的电商用户行为,构成用户行为的特征序列,计算所述特征序列中相邻时间步特征的相似度,对所述特征序列进行分组,得到查询特征序列和键值对特征序列,其中,所述相邻时间步特征的相似度大于基于所述特征序列的全局统计特性确定的动态阈值的时间步特征划分为所述查询特征序列,其余特征划分为所述键值对特征序列,得到所述查询特征序列和所述键值对特征序列的特征划分结果;

基于所述特征划分结果,通过递归神经网络分别计算所述查询特征序列和所述键值对特征序列的时序关联强度,所述递归神经网络的隐状态包含所述查询特征序列和所述键值对特征序列的长短期依赖信息,得到所述查询特征序列的时序关联强度和所述键值对特征序列的时序关联强度;

基于所述特征序列的时序位置关系构建第一掩码矩阵作为时序因果掩码矩阵,将所述查询特征序列的时序关联强度和所述键值对特征序列的时序关联强度转换为软掩码权重,得到第二掩码矩阵;

将所述查询特征序列与所述键值对特征序列进行矩阵乘法运算,得到原始注意力分数;将所述原始注意力分数与所述第一掩码矩阵进行加法运算,得到第一次掩码结果,将所述第一次掩码结果与所述第二掩码矩阵进行元素级乘法运算,得到双重掩码后的注意力分数;对双重掩码后的注意力分数进行归一化处理,得到仅包含历史特征信息的初始注意力权重。

6. 根据权利要求1所述的方法,其特征在于,构建信息瓶颈特征筛选器,通过最大化任务相关信息和最小化冗余信息,对所述深层特征表示进行筛选,得到关键特征子集包括:

构建信息瓶颈编码器,所述信息瓶颈编码器包括三层感知机结构,每层感知机结构之后连接归一化层,将所述深层特征表示输入所述信息瓶颈编码器,所述信息瓶颈编码器的

输出层生成压缩特征空间的概率分布参数;基于所述概率分布参数对所述深层特征表示进行编码映射,得到压缩特征表示;

构建互信息评估模块,所述互信息评估模块包含三层感知机结构,将所述压缩特征表示与目标任务的标签的拼接向量输入所述互信息评估模块,输出所述压缩特征表示与所述目标任务之间的互信息量;

计算所述压缩特征表示的自相关矩阵,所述自相关矩阵中每个元素为对应特征向量减去相应的均值后的乘积,除以对应标准差的乘积;基于所述自相关矩阵与单位矩阵的差值,计算特征冗余度;

构建特征选择网络,所述特征选择网络包括第一全连接层和第二全连接层,将所述压缩特征表示输入所述第一全连接层,所述第一全连接层的输出经过激活函数后输入第二全连接层,所述第二全连接层的输出经过门控函数得到门控值,将所述门控值与所述压缩特征表示进行逐元素相乘,得到初始筛选特征;

基于温度参数对所述门控函数的输出进行调节,所述温度参数基于预设的递减量随训练轮数递减,将调节后的门控值与所述压缩特征表示进行逐元素相乘,得到调节后筛选特征;

将所述互信息量的相反数、所述特征冗余度以及所述调节后筛选特征的范数线性组合构建目标函数;

通过迭代优化所述目标函数,得到所述信息瓶颈编码器的优化参数和所述特征选择网络的优化参数,获得优化信息瓶颈编码器和优化特征选择网络;

将所述深层特征表示输入优化信息瓶颈编码器得到优化压缩特征表示,将所述优化压缩特征表示输入优化特征选择网络,得到关键特征子集。

7.根据权利要求1所述的方法,其特征在于,根据所述特征权重向量和所述预测结果,采用结构化梯度方法计算所述关键特征子集中各特征对所述预测结果的贡献度,生成用户购买意向预测的分析结果包括:

基于特征权重向量和预测结果,计算所述特征权重向量中每对特征之间的相关系数,生成特征相关性矩阵;基于所述特征相关性矩阵构建特征依赖图,将所述关键特征子集中的每个特征作为所述特征依赖图中的节点,将所述特征相关性矩阵中的相关系数作为对应节点之间的边权重;

根据所述特征依赖图构建图拉普拉斯矩阵,将各节点与对应相连节点的边权重之和作为所述图拉普拉斯矩阵的对角元素值,将节点间的边权重的负值作为所述图拉普拉斯矩阵的非对角元素值;

计算所述预测结果关于所述关键特征子集的雅可比矩阵,得到初始梯度;将所述图拉普拉斯矩阵作为结构化正则项,将所述结构化正则项与所述初始梯度进行加权组合,得到考虑特征依赖关系的结构化梯度;

计算所述关键特征子集中每个特征的结构化梯度幅值、每个特征与其他特征的交互梯度值、以及在所述特征依赖图中每个特征的中心性度量;

将所述结构化梯度幅值、所述交互梯度值和所述中心性度量进行加权组合,得到特征贡献度;对所述特征贡献度进行归一化处理,得到所述关键特征子集中各特征对所述预测结果的标准化贡献度;基于所述标准化贡献度对所述关键特征子集中的特征进行重要性排

序,生成用户购买意向预测的分析结果。

8. 基于Transformer的结构化数据建模分析系统,用于实现前述权利要求1-7中任一项所述的方法,其特征在于,包括:

第一单元,用于接收结构化原始数据,将所述结构化原始数据划分为数值型特征和类别型特征,通过双流特征编码器,对所述数值型特征执行分段线性变换,得到归一化主干特征,对所述类别型特征执行自适应量子编码,得到连续值映射特征;通过可学习的门控机制进行自适应权重分配,融合所述归一化主干特征和所述连续值映射特征,得到融合特征;基于所述融合特征的重要性评分进行动态采样,生成增强训练样本集;

第二单元,用于将所述增强训练样本集输入包含层次注意力模块和自校准模块的特征提取网络,所述层次注意力模块计算样本特征的相似度矩阵,基于所述相似度矩阵构建特征亲和图,通过图分割算法将所述特征亲和图划分为多个特征子图;在每个所述特征子图内,采用因果注意力机制提取局部特征表示,并通过跨子图的注意力融合机制整合多个局部特征表示,得到全局特征表示;所述自校准模块接收所述全局特征表示,通过残差连接和动态归一化进行特征非线性变换,得到深层特征表示;

第三单元,用于构建信息瓶颈特征筛选器,通过最大化任务相关信息和最小化冗余信息,对所述深层特征表示进行筛选,得到关键特征子集;将所述关键特征子集输入包含主预测分支和辅助重建分支的双分支网络,所述主预测分支基于注意力机制对所述关键特征子集进行处理得到特征权重向量,并根据所述特征权重向量生成预测结果,所述辅助重建分支通过重建所述关键特征子集提供正则化约束;根据所述特征权重向量和所述预测结果,采用结构化梯度方法计算所述关键特征子集中各特征对所述预测结果的贡献度,生成分析结果。

9. 一种电子设备,其特征在于,包括:

处理器;

用于存储处理器可执行指令的存储器;

其中,所述处理器被配置为调用所述存储器存储的指令,以执行权利要求1至7中任意一项所述的方法。

10. 一种计算机可读存储介质,其上存储有计算机程序指令,其特征在于,所述计算机程序指令被处理器执行时实现权利要求1至7中任意一项所述的方法。

基于Transformer的结构化数据建模分析方法

技术领域

[0001] 本发明涉及数据分析技术领域,尤其涉及一种基于Transformer的结构化数据建模分析方法。

背景技术

[0002] 随着大数据时代的到来,结构化数据的分析和建模已成为人工智能领域的重要研究方向。传统的结构化数据分析方法主要依赖特征工程和浅层机器学习模型,如决策树、逻辑回归等。近年来,Transformer架构在自然语言处理领域取得了突破性进展,其自注意力机制能够有效捕获序列数据中的长程依赖关系。然而,将Transformer应用于结构化数据分析时,需要解决数值型特征和类别型特征的异构性处理、特征交互建模以及模型可解释性等关键问题。

[0003] 然而,现有技术在结构化数据建模分析领域仍然存在问题,传统方法难以有效处理数值型特征和类别型特征的异构性,特征编码方式过于简单,无法充分利用特征间的关联信息;现有的特征提取方法主要关注局部特征交互,缺乏对全局特征依赖关系的建模能力,导致特征表示能力受限;特征选择过程往往采用静态的评分机制,忽视了特征间的动态关联性,且缺乏可解释性支持;模型预测结果缺乏细粒度的解释机制,难以量化各个特征对预测结果的贡献程度。

[0004] 综上所述,亟需一种基于Transformer的结构化数据建模分析方法,通过设计双流特征编码器,分别对数值型特征和类别型特征进行适应性编码;利用层次注意力机制和特征亲和图实现多尺度特征交互建模;构建信息瓶颈特征筛选器和双分支预测网络,在保证模型性能的同时提供可解释的特征重要性评估机制;最后通过结构化梯度方法定量分析特征贡献度,为模型预测结果提供细粒度的解释支持。本发明不仅提高了结构化数据的建模效果,还增强了模型的可解释性和可信性,能够解决现有技术中的问题。

发明内容

[0005] 本发明实施例提供一种基于Transformer的结构化数据建模分析方法,能够解决现有技术中的问题。

[0006] 本发明实施例的第一方面,

[0007] 提供基于Transformer的结构化数据建模分析方法,包括:

[0008] 接收结构化原始数据,将所述结构化原始数据划分为数值型特征和类别型特征,通过双流特征编码器,对所述数值型特征执行分段线性变换,得到归一化主干特征,对所述类别型特征执行自适应量子编码,得到连续值映射特征;通过可学习的门控机制进行自适应权重分配,融合所述归一化主干特征和所述连续值映射特征,得到融合特征;基于所述融合特征的重要性评分进行动态采样,生成增强训练样本集;

[0009] 将所述增强训练样本集输入包含层次注意力模块和自校准模块的特征提取网络,所述层次注意力模块计算样本特征的相似度矩阵,基于所述相似度矩阵构建特征亲和图,

通过图分割算法将所述特征亲和图划分为多个特征子图；在每个所述特征子图内，采用因果注意力机制提取局部特征表示，并通过跨子图的注意力融合机制整合多个局部特征表示，得到全局特征表示；所述自校准模块接收所述全局特征表示，通过残差连接和动态归一化进行特征非线性变换，得到深层特征表示；

[0010] 构建信息瓶颈特征筛选器，通过最大化任务相关信息和最小化冗余信息，对所述深层特征表示进行筛选，得到关键特征子集；将所述关键特征子集输入包含主预测分支和辅助重建分支的双分支网络，所述主预测分支基于注意力机制对所述关键特征子集进行处理得到特征权重向量，并根据所述特征权重向量生成预测结果，所述辅助重建分支通过重建所述关键特征子集提供正则化约束；根据所述特征权重向量和所述预测结果，采用结构化梯度方法计算所述关键特征子集中各特征对所述预测结果的贡献度，生成分析结果。

[0011] 在一种可选的实施例中，

[0012] 接收结构化原始数据，将所述结构化原始数据划分为数值型特征和类别型特征，通过双流特征编码器，对所述数值型特征执行分段线性变换，得到归一化主干特征，对所述类别型特征执行自适应量子编码，得到连续值映射特征；通过可学习的门控机制进行自适应权重分配，融合所述归一化主干特征和所述连续值映射特征，得到融合特征；基于所述融合特征的重要性评分进行动态采样，生成增强训练样本集包括：

[0013] 接收结构化原始数据，将所述结构化原始数据划分为数值型特征和类别型特征；将数值型特征和类别型特征输入双流特征编码器进行特征处理，所述双流特征编码器包括特征主干流和特征映射流；

[0014] 在特征主干流中，统计所述数值型特征的数据分布，得到数据密度，基于所述数据密度确定最优分段点的位置，根据所述最优分段点，将数值型特征划分为多个分段区间，在每个所述分段区间，构造线性变换参数矩阵，计算相邻的分段区间之间的特征差异值，通过最小化所述特征差异值，确定各分段区间的优化线性变换参数矩阵，使用优化线性变换参数矩阵对各分段区间的数值型特征执行线性变换，得到归一化主干特征；

[0015] 在特征映射流中，将所述类别型特征中的每个类别值映射为对应的量子基态，为每个所述量子基态分配初始振幅参数和相位参数，构建包含旋转门和控制门的量子门序列，将所述量子门序列依次作用于所述量子基态，通过量子态变换得到变换量子态，计算变换量子态的熵值，基于所述熵值将变换量子态投影到连续空间，得到连续值映射特征；

[0016] 分别计算所述归一化主干特征和所述连续值映射特征各维度之间的相关性系数，将所述相关性系数组合构建特征相关性系数矩阵，基于所述特征相关性系数矩阵训练预设的注意力网络，得到特征维度的重要度向量，使用所述重要度向量对所述归一化主干特征和所述连续值映射特征进行加权组合得到融合特征；

[0017] 计算所述融合特征的各个维度对目标任务的贡献度，基于所述贡献度生成每个维度的特征重要性分数，根据所述特征重要性分数构建采样概率分布函数，使用所述采样概率分布函数对训练样本进行重要性采样，生成增强训练样本集。

[0018] 在一种可选的实施例中，

[0019] 统计所述数值型特征的数据分布，得到数据密度，基于所述数据密度确定最优分段点的位置包括：

[0020] 基于所述数值型特征的四分位距和样本数量，计算基础区间宽度，基于基础区间

宽度,将所述数值型特征的值域范围划分为多个基础统计区间,对每个所述基础统计区间内的样本数量进行统计得到区间频数序列;

[0021] 计算所述区间频数序列中相邻基础统计区间的频数比值,得到区间相似度,将所述区间相似度大于预设合并阈值的相邻的基础统计区间合并为新统计区间,得到合并频数分布序列;

[0022] 根据所述合并频数分布序列计算累积频数分布,基于所述累积频数分布生成数据密度函数;在所述合并频数分布序列中选取候选分割位置,计算每个所述候选分割位置的两侧区间的样本数量和数值均值,利用所述样本数量和所述数值均值计算所述候选分割位置的区间方差值;

[0023] 在当前待分割区间内选择具有最大区间方差值的位置作为待定分段点,计算所述待定分段点的区间方差值与所述数值型特征的总体方差的比值,确定所述待定分段点的方差贡献比;

[0024] 对所述待定分段点的两侧区间计算局部分布偏度,当所述局部分布偏度的差值大于预设偏度阈值,且所述方差贡献比大于预设分割阈值时,将所述待定分段点确定为最优分段点,重复执行,确定所有最优分段点的位置。

[0025] 在一种可选的实施例中,

[0026] 将所述增强训练样本集输入包含层次注意力模块和自校准模块的特征提取网络,所述层次注意力模块计算样本特征的相似度矩阵,基于所述相似度矩阵构建特征亲和图,通过图分割算法将所述特征亲和图划分为多个特征子图;在每个所述特征子图内,采用因果注意力机制提取局部特征表示,并通过跨子图的注意力融合机制整合多个局部特征表示,得到全局特征表示;所述自校准模块接收所述全局特征表示,通过残差连接和动态归一化进行特征非线性变换,得到深层特征表示包括:

[0027] 将增强训练样本集输入卷积神经网络得到初始特征映射,计算所述初始特征映射中任意两个样本特征的点积,并将所述点积除以对应特征向量的模长乘积,得到特征相似度矩阵;基于所述特征相似度矩阵构建特征亲和图,所述特征亲和图的顶点对应样本特征,所述特征亲和图的边权重对应所述特征相似度矩阵中的相似度值,通过谱聚类算法将所述特征亲和图划分为多个特征子图;

[0028] 在每个所述特征子图内,将所述特征子图的特征序列划分为查询特征序列和键值对特征序列,构建时序因果掩码矩阵,利用所述时序因果掩码矩阵对所述查询特征序列与所述键值对特征序列的交互结果进行屏蔽,得到仅包含历史特征信息的初始注意力权重;计算每个所述特征子图内样本特征的空间距离矩阵,基于所述空间距离矩阵计算所述特征子图的空间分布熵,根据所述空间分布熵对所述初始注意力权重进行自适应调节,得到自适应注意力权重;基于所述自适应注意力权重对所述键值对特征序列进行加权,并叠加位置编码信息,得到每个所述特征子图的局部特征表示;

[0029] 基于预先构建的多层感知机网络,输入相邻特征子图的局部特征表示的拼接向量,经所述多层感知机网络前向传播,得到特征子图间的关联权重,基于所述关联权重对所述局部特征表示进行加权求和,得到全局特征表示;

[0030] 对所述全局特征表示在特征维度上分别计算均值和标准差;基于预先构建的动态归一化网络,将所述全局特征表示输入所述动态归一化网络,经前向传播得到特征变换系

数和特征偏移量;将所述全局特征表示减去所述均值后除以所述标准差得到标准化特征,将所述标准化特征与所述特征变换系数相乘,并加上所述特征偏移量得到变换特征,将所述变换特征与所述全局特征表示相加,得到自校准特征;

[0031] 对所述自校准特征进行层归一化处理,并通过ReLU激活函数进行非线性变换,得到深层特征表示。

[0032] 在一种可选的实施例中,

[0033] 将所述特征子图的特征序列划分为查询特征序列和键值对特征序列,构建时序因果掩码矩阵,利用所述时序因果掩码矩阵对所述查询特征序列与所述键值对特征序列的交互结果进行屏蔽,得到仅包含历史特征信息的初始注意力权重包括:

[0034] 基于特征子图中相邻时间步特征的相似度,对所述特征子图的特征序列进行分组,得到查询特征序列和键值对特征序列,其中,所述相邻时间步特征的相似度大于基于所述特征序列的全局统计特性确定的动态阈值的时间步特征划分为所述查询特征序列,其余特征划分为所述键值对特征序列,得到所述查询特征序列和所述键值对特征序列的特征划分结果;

[0035] 基于所述特征划分结果,通过递归神经网络分别计算所述查询特征序列和所述键值对特征序列的时序关联强度,所述递归神经网络的隐状态包含所述查询特征序列和所述键值对特征序列的长短期依赖信息,得到所述查询特征序列的时序关联强度和所述键值对特征序列的时序关联强度;

[0036] 基于所述特征序列的时序位置关系构建第一掩码矩阵作为时序因果掩码矩阵,将所述查询特征序列的时序关联强度和所述键值对特征序列的时序关联强度转换为软掩码权重,得到第二掩码矩阵;

[0037] 将所述查询特征序列与所述键值对特征序列进行矩阵乘法运算,得到原始注意力分数;将所述原始注意力分数与所述第一掩码矩阵进行加法运算,得到第一次掩码结果,将所述第一次掩码结果与所述第二掩码矩阵进行元素级乘法运算,得到双重掩码后的注意力分数;对双重掩码后的注意力分数进行归一化处理,得到仅包含历史特征信息的初始注意力权重。

[0038] 在一种可选的实施例中,

[0039] 构建信息瓶颈特征筛选器,通过最大化任务相关信息和最小化冗余信息,对所述深层特征表示进行筛选,得到关键特征子集包括:

[0040] 构建信息瓶颈编码器,所述信息瓶颈编码器包括三层感知机结构,每层感知机结构之后连接归一化层,将所述深层特征表示输入所述信息瓶颈编码器,所述信息瓶颈编码器的输出层生成压缩特征空间的概率分布参数;基于所述概率分布参数对所述深层特征表示进行编码映射,得到压缩特征表示;

[0041] 构建互信息评估模块,所述互信息评估模块包含三层感知机结构,将所述压缩特征表示与目标任务的标签的拼接向量输入所述互信息评估模块,输出所述压缩特征表示与所述目标任务之间的互信息量;

[0042] 计算所述压缩特征表示的自相关矩阵,所述自相关矩阵中每个元素为对应特征向量减去相应的均值后的乘积,除以对应标准差的乘积;基于所述自相关矩阵与单位矩阵的差值,计算特征冗余度;

[0043] 构建特征选择网络,所述特征选择网络包括第一全连接层和第二全连接层,将所述压缩特征表示输入所述第一全连接层,所述第一全连接层的输出经过激活函数后输入第二全连接层,所述第二全连接层的输出经过门控函数得到门控值,将所述门控值与所述压缩特征表示进行逐元素相乘,得到初始筛选特征;

[0044] 基于温度参数对所述门控函数的输出进行调节,所述温度参数基于预设的递减量随训练轮数递减,将调节后的门控值与所述压缩特征表示进行逐元素相乘,得到调节后筛选特征;

[0045] 将所述互信息量的相反数、所述特征冗余度以及所述调节后筛选特征的范数线性组合构建目标函数;

[0046] 通过迭代优化所述目标函数,得到所述信息瓶颈编码器的优化参数和所述特征选择网络的优化参数,获得优化信息瓶颈编码器和优化特征选择网络;

[0047] 将所述深层特征表示输入优化信息瓶颈编码器得到优化压缩特征表示,将所述优化压缩特征表示输入优化特征选择网络,得到关键特征子集。

[0048] 在一种可选的实施例中,

[0049] 根据所述特征权重向量和所述预测结果,采用结构化梯度方法计算所述关键特征子集中各特征对所述预测结果的贡献度,生成分析结果包括:

[0050] 基于特征权重向量和预测结果,计算所述特征权重向量中每对特征之间的相关系数,生成特征相关性矩阵;基于所述特征相关性矩阵构建特征依赖图,将所述关键特征子集中的每个特征作为所述特征依赖图中的节点,将所述特征相关性矩阵中的相关系数作为对应节点之间的边权重;

[0051] 根据所述特征依赖图构建图拉普拉斯矩阵,将各节点与对应相连节点的边权重之和作为所述图拉普拉斯矩阵的对角元素值,将节点间的边权重的负值作为所述图拉普拉斯矩阵的非对角元素值;

[0052] 计算所述预测结果关于所述关键特征子集的雅可比矩阵,得到初始梯度;将所述图拉普拉斯矩阵作为结构化正则项,将所述结构化正则项与所述初始梯度进行加权组合,得到考虑特征依赖关系的结构化梯度;

[0053] 计算所述关键特征子集中每个特征的结构化梯度幅值、每个特征与其他特征的交互梯度值、以及在所述特征依赖图中每个特征的中心性度量;

[0054] 将所述结构化梯度幅值、所述交互梯度值和所述中心性度量进行加权组合,得到特征贡献度;对所述特征贡献度进行归一化处理,得到所述关键特征子集中各特征对所述预测结果的标准化贡献度;基于所述标准化贡献度对所述关键特征子集中的特征进行重要性排序,生成分析结果。

[0055] 本发明实施例的第二方面,

[0056] 提供基于Transformer的结构化数据建模分析系统,包括:

[0057] 第一单元,用于接收结构化原始数据,将所述结构化原始数据划分为数值型特征和类别型特征,通过双流特征编码器,对所述数值型特征执行分段线性变换,得到归一化主干特征,对所述类别型特征执行自适应量子编码,得到连续值映射特征;通过可学习的门控机制进行自适应权重分配,融合所述归一化主干特征和所述连续值映射特征,得到融合特征;基于所述融合特征的重要性评分进行动态采样,生成增强训练样本集;

[0058] 第二单元,用于将所述增强训练样本集输入包含层次注意力模块和自校准模块的特征提取网络,所述层次注意力模块计算样本特征的相似度矩阵,基于所述相似度矩阵构建特征亲和图,通过图分割算法将所述特征亲和图划分为多个特征子图;在每个所述特征子图内,采用因果注意力机制提取局部特征表示,并通过跨子图的注意力融合机制整合多个局部特征表示,得到全局特征表示;所述自校准模块接收所述全局特征表示,通过残差连接和动态归一化进行特征非线性变换,得到深层特征表示;

[0059] 第三单元,用于构建信息瓶颈特征筛选器,通过最大化任务相关信息和最小化冗余信息,对所述深层特征表示进行筛选,得到关键特征子集;将所述关键特征子集输入包含主预测分支和辅助重建分支的双分支网络,所述主预测分支基于注意力机制对所述关键特征子集进行处理得到特征权重向量,并根据所述特征权重向量生成预测结果,所述辅助重建分支通过重建所述关键特征子集提供正则化约束;根据所述特征权重向量和所述预测结果,采用结构化梯度方法计算所述关键特征子集中各特征对所述预测结果的贡献度,生成分析结果。

[0060] 本发明实施例的第三方面,

[0061] 提供一种电子设备,包括:

[0062] 处理器;

[0063] 用于存储处理器可执行指令的存储器;

[0064] 其中,所述处理器被配置为调用所述存储器存储的指令,以执行前述所述的方法。

[0065] 本发明实施例的第四方面,

[0066] 提供一种计算机可读存储介质,其上存储有计算机程序指令,所述计算机程序指令被处理器执行时实现前述所述的方法。

[0067] 在本发明实施例中,通过双流特征编码器对数值型和类别型特征进行分别处理,并使用可学习的门控机制进行自适应融合,有效提高了对不同类型数据的处理能力;基于融合特征的重要性评分进行动态采样,生成增强训练样本集,显著提升了模型的泛化能力和鲁棒性;采用层次注意力模块和自校准模块的特征提取网络,通过特征亲和图和图分割算法捕捉特征间的复杂关系,并利用因果注意力机制和跨子图注意力融合提取全局特征表示;自校准模块通过残差连接和动态归一化进一步优化特征表示,有效提高了特征提取的准确性和模型的表达能力;引入信息瓶颈特征筛选器,在保留任务相关信息的同时最小化冗余信息,提高了特征选择的效率;双分支网络的设计既保证了预测的准确性,又通过辅助重建分支提供了正则化约束,增强了模型的稳定性;采用结构化梯度方法计算特征贡献度,为模型决策提供了可解释性,有助于用户理解和信任模型的预测结果。

附图说明

[0068] 图1为本发明实施例基于Transformer的结构化数据建模分析方法的流程示意图;

[0069] 图2为本发明实施例基于Transformer的结构化数据建模分析系统的结构示意图。

具体实施方式

[0070] 为使本发明实施例的目的、技术方案和优点更加清楚,下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅

仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0071] 下面以具体地实施例对本发明的技术方案进行详细说明。下面这几个具体的实施例可以相互结合,对于相同或相似的概念或过程可能在某些实施例不再赘述。

[0072] 图1为本发明实施例基于Transformer的结构化数据建模分析方法的流程示意图,如图1所示,所述方法包括:

[0073] S101.接收结构化原始数据,将所述结构化原始数据划分为数值型特征和类别型特征,通过双流特征编码器,对所述数值型特征执行分段线性变换,得到归一化主干特征,对所述类别型特征执行自适应量子编码,得到连续值映射特征;通过可学习的门控机制进行自适应权重分配,融合所述归一化主干特征和所述连续值映射特征,得到融合特征;基于所述融合特征的重要性评分进行动态采样,生成增强训练样本集;

[0074] 在本实施例中,通过双流特征编码器分别处理数值型特征和类别型特征,能够针对不同特征类型采用适配的编码方式,提高特征表达的精确性;通过可学习的门控机制对不同特征类型分配自适应权重,实现数值型特征和类别型特征的有效融合,充分利用多源信息,提高模型的表现力;动态样本优化:基于融合特征的重要性评分进行动态采样,有助于平衡样本分布,突出关键样本,从而生成更具代表性的增强训练样本集,提升模型的训练效果和泛化能力;整体方法在数据预处理和特征表示阶段进行优化,为下游任务提供更加精准且高效的特征输入,进一步提升模型的学习效率和性能表现。

[0075] S102.将所述增强训练样本集输入包含层次注意力模块和自校准模块的特征提取网络,所述层次注意力模块计算样本特征的相似度矩阵,基于所述相似度矩阵构建特征亲和图,通过图分割算法将所述特征亲和图划分为多个特征子图;在每个所述特征子图内,采用因果注意力机制提取局部特征表示,并通过跨子图的注意力融合机制整合多个局部特征表示,得到全局特征表示;所述自校准模块接收所述全局特征表示,通过残差连接和动态归一化进行特征非线性变换,得到深层特征表示;

[0076] 在本实施例中,通过层次注意力模块,构建特征亲和图并分割为多个特征子图,能够捕捉特征之间的结构化关联,提升对样本内部关系的建模能力;利用因果注意力机制提取局部特征表示,并通过跨子图的注意力融合机制整合局部特征,能够有效结合局部细节和全局模式,提升特征表达的全面性;校准模块通过残差连接和动态归一化,实现对全局特征表示的非线性变换和自适应调整,有助于强化关键特征并抑制冗余信息,提高深层特征的表示能力;整体网络架构结合层次注意力和自校准模块,能够更加高效地提取复杂样本的多维特征,为下游任务提供具有高辨识度的深层特征表示;特征提取网络通过图分割、注意力机制和动态校准,增强了对数据复杂关系的建模能力,提升了模型对多样化输入的适应性和泛化能力。

[0077] S103.构建信息瓶颈特征筛选器,通过最大化任务相关信息和最小化冗余信息,对所述深层特征表示进行筛选,得到关键特征子集;将所述关键特征子集输入包含主预测分支和辅助重建分支的双分支网络,所述主预测分支基于注意力机制对所述关键特征子集进行处理得到特征权重向量,并根据所述特征权重向量生成预测结果,所述辅助重建分支通过重建所述关键特征子集提供正则化约束;根据所述特征权重向量和所述预测结果,采用结构化梯度方法计算所述关键特征子集中各特征对所述预测结果的贡献度,生成分析结

果。

[0078] 首先构建信息瓶颈特征筛选器,对深层特征表示进行分析和筛选。计算每个特征维度与预测目标之间的互信息值,互信息值越高表示该维度特征与预测任务的相关性越强。同时计算特征维度之间的冗余信息,用于发现和去除信息重叠的冗余特征。通过迭代优化过程,在最大化任务相关信息和最小化特征冗余信息的约束下,从完整的深层特征表示中筛选出关键特征子集。

[0079] 将筛选得到的关键特征子集输入双分支网络进行处理。在主预测分支中,采用多头注意力机制对特征进行加权。多头注意力层包含多个并行的注意力计算单元,每个注意力头独立计算特征权重。通过这种方式,模型能够从不同角度捕获特征之间的关联模式。将所有注意力头的输出进行合并,得到综合的特征权重向量。基于该权重向量对关键特征进行加权组合,通过全连接层生成预测结果。

[0080] 在辅助重建分支中,构建对称的编码器-解码器结构。编码器通过多层全连接网络将关键特征映射到低维隐空间,解码器则将隐空间表示重建回原始特征空间。这种重建任务作为辅助监督信号,通过最小化重建误差来约束模型的学习过程。重建损失与预测损失共同构成模型的优化目标,使得学习得到的特征表示既有利于预测任务,又保持了原始数据的结构信息。

[0081] 最后采用结构化梯度方法对预测结果进行解释分析。结合主分支得到的特征权重向量,计算每个输入特征对预测结果的贡献度。特征贡献度反映了不同特征在决策过程中的重要程度,可用于理解模型的预测依据。基于贡献度分析生成可解释的结果报告,帮助用户理解预测背后的关键影响因素。

[0082] 在本实施例中,通过信息瓶颈特征筛选器最大化任务相关信息和最小化冗余信息,从深层特征表示中提取关键特征子集,显著提高特征选择的效率和精度;主预测分支基于注意力机制对关键特征子集赋予权重,通过重点关注对预测任务影响较大的特征,生成更准确的预测结果;辅助重建分支通过重建关键特征子集引入正则化约束,防止模型过拟合,同时保证关键特征子集的完整性和表达能力;利用结构化梯度方法量化关键特征对子集中的每个特征对预测结果的贡献度,为模型解释性和可解释性提供支持,帮助理解模型的决策依据;通过筛选和优化,与任务目标高度相关的特征得以保留和强化,从而提升模型的任务适配能力和实际应用价值。

[0083] 在一种可选的实施方式中,接收结构化原始数据,将所述结构化原始数据划分为数值型特征和类别型特征,通过双流特征编码器,对所述数值型特征执行分段线性变换,得到归一化主干特征,对所述类别型特征执行自适应量子编码,得到连续值映射特征;通过可学习的门控机制进行自适应权重分配,融合所述归一化主干特征和所述连续值映射特征,得到融合特征;基于所述融合特征的重要性评分进行动态采样,生成增强训练样本集包括:

[0084] 接收结构化原始数据,将所述结构化原始数据划分为数值型特征和类别型特征;将数值型特征和类别型特征输入双流特征编码器进行特征处理,所述双流特征编码器包括特征主干流和特征映射流;

[0085] 在特征主干流中,统计所述数值型特征的数据分布,得到数据密度,基于所述数据密度确定最优分段点的位置,根据所述最优分段点,将数值型特征划分为多个分段区间,在每个所述分段区间,构造线性变换参数矩阵,计算相邻的分段区间之间的特征差异值,通过

最小化所述特征差异值,确定各分段区间的优化线性变换参数矩阵,使用优化线性变换参数矩阵对各分段区间的数值型特征执行线性变换,得到归一化主干特征;

[0086] 在特征映射流中,将所述类别型特征中的每个类别值映射为对应的量子基态,为每个所述量子基态分配初始振幅参数和相位参数,构建包含旋转门和控制门的量子门序列,将所述量子门序列依次作用于所述量子基态,通过量子态变换得到变换量子态,计算变换量子态的熵值,基于所述熵值将变换量子态投影到连续空间,得到连续值映射特征;

[0087] 分别计算所述归一化主干特征和所述连续值映射特征各维度之间的相关性系数,将所述相关性系数组合构建特征相关性系数矩阵,基于所述特征相关性系数矩阵训练预设的注意力网络,得到特征维度的重要度向量,使用所述重要度向量对所述归一化主干特征和所述连续值映射特征进行加权组合得到融合特征;

[0088] 计算所述融合特征的各个维度对目标任务的贡献度,基于所述贡献度生成每个维度的特征重要性分数,根据所述特征重要性分数构建采样概率分布函数,使用所述采样概率分布函数对训练样本进行重要性采样,生成增强训练样本集。

[0089] 在一种具体实施方式中,首先接收结构化原始数据,根据数据类型进行特征划分。以某电商平台的用户购买意向预测任务为例,原始数据包含用户基础信息、行为数据和交易记录等。数值型特征包括:年龄(18-80)、近30天访问次数(0-120)、近90天消费总额(0-50000)、最近一次购买距今天数(0-365)、收藏商品数(0-100)、购物车商品数(0-50)等;类别型特征包括:性别(男/女)、会员等级(普通/白银/黄金/钻石)、常用设备类型(PC/Android/iOS)、偏好购物品类(服装/数码/美妆/图书等)、收货地区等。

[0090] 在特征主干流中,首先对数值型特征进行分布统计和分段处理:以用户年龄特征为例,基于50万用户样本,统计得到年龄的数据分布密度。通过设定基础统计区间宽度为2岁,初步划分为31个基础区间(18-20,20-22,...,78-80)。计算相邻区间的样本频数比值,当比值大于0.85时进行区间合并。例如20-22、22-24和24-26这三个区间的频数分别为15000、14500、14800,两组相邻比值分别为0.97和0.98,满足合并条件,将其合并为20-26区间。经过合并后最终得到18-20、20-26、26-35、35-45、45-55、55-65、65-80这七个分段区间。

[0091] 对每个分段区间构造线性变换参数矩阵:以20-26岁区间为例,构造 2×2 的线性变换参数矩阵,初始值为单位矩阵。计算与相邻区间26-35的特征差异值,通过迭代优化使差异值低于阈值0.01。最终得到的变换矩阵应用于该区间内的年龄值,实现特征的归一化转换。例如原始年龄24岁经过变换后得到归一化值0.37。对其他数值型特征采用相同方法处理,得到完整的归一化主干特征集。

[0092] 在特征映射流中,对类别型特征进行量子编码转换:以会员等级特征为例,首先将四个等级映射为对应的量子基态:普通会员映射为 $|00\rangle$,白银会员映射为 $|01\rangle$,黄金会员映射为 $|10\rangle$,钻石会员映射为 $|11\rangle$ 。为每个基态分配初始参数:振幅均设为0.5,相位角设为0。构建包含8个量子门的序列,依次为:X门、CNOT门、Z门、X门、CNOT门、Z门、X门、CNOT门。将这个门序列依次作用于基态,得到变换后的量子态。计算量子态的熵值,普通会员得到熵值0.25,白银会员0.45,黄金会员0.65,钻石会员0.85。基于熵值将量子态投影到二维连续空间,得到会员等级的连续值表示,例如普通会员被映射为(0.2,0.3),白银会员被映射为(0.4,0.5)等。

[0093] 进行特征相关性分析和融合:计算所有特征两两之间的相关系数。例如,年龄与消

费总额的相关系数为0.35,会员等级与访问次数的相关系数为0.62等。将这些相关系数组织成相关性矩阵,矩阵维度为特征总数 $\times$ 特征总数。基于该矩阵训练一个包含三层结构的注意力网络,网络的隐层单元数分别为128、64、32,使用ReLU激活函数。通过网络训练得到特征维度的重要度向量,例如年龄特征的重要度为0.28,消费总额的重要度为0.35,会员等级的重要度为0.31等。使用这些重要度值对归一化主干特征和连续值映射特征进行加权组合,得到最终的融合特征。

[0094] 最后进行样本增强处理:计算融合特征中各维度对购买意向预测的贡献度。基于50万训练样本,通过计算得到各特征维度的贡献度值:年龄特征0.25、消费总额0.38、会员等级0.32、访问次数0.28等。将这些贡献度值归一化后作为采样概率,构建采样概率分布函数。对原始训练样本集进行重要性采样,采样概率超过0.8的样本复制3份,概率在0.5-0.8之间的复制2份,概率在0.3-0.5之间的复制1份。通过这种方式生成增强训练样本集,最终样本量达到原始数据集的1.8倍,约90万条样本记录。

[0095] 在本实施例中,通过双流特征编码器分别处理数值型特征和类别型特征,实现了不同类型特征的高效编码。分段线性变换保持了数值特征的连续性和局部特征,自适应量子编码则充分利用了量子计算的优势,提高了类别特征的表达能力;采用可学习的门控机制进行特征融合,实现了特征间的动态权重分配。通过注意力网络学习特征相关性,自适应调整不同特征的重要程度,提高了模型对不同特征组合的适应能力;基于特征重要性的动态采样策略,有效提升了训练样本的质量。通过计算特征对目标任务的贡献度,构建合理的采样概率分布,生成更具代表性的训练样本,提高了模型的泛化能力和鲁棒性。

[0096] 在一种可选的实施方式中,统计所述数值型特征的数据分布,得到数据密度,基于所述数据密度确定最优分段点的位置包括:

[0097] 基于所述数值型特征的四分位距和样本数量,计算基础区间宽度,基于基础区间宽度,将所述数值型特征的值域范围划分为多个基础统计区间,对每个所述基础统计区间内的样本数量进行统计得到区间频数序列;

[0098] 计算所述区间频数序列中相邻基础统计区间的频数比值,得到区间相似度,将所述区间相似度大于预设合并阈值的相邻的基础统计区间合并为新统计区间,得到合并频数分布序列;

[0099] 根据所述合并频数分布序列计算累积频数分布,基于所述累积频数分布生成数据密度函数;在所述合并频数分布序列中选取候选分割位置,计算每个所述候选分割位置的两侧区间的样本数量和数值均值,利用所述样本数量和所述数值均值计算所述候选分割位置的区间方差值;

[0100] 在当前待分割区间内选择具有最大区间方差值的位置作为待定分段点,计算所述待定分段点的区间方差值与所述数值型特征的总体方差的比值,确定所述待定分段点的方差贡献比;

[0101] 对所述待定分段点的两侧区间计算局部分布偏度,当所述局部分布偏度的差值大于预设偏度阈值,且所述方差贡献比大于预设分割阈值时,将所述待定分段点确定为最优分段点,重复执行,确定所有最优分段点的位置。

[0102] 所述局部分布偏度具体是指衡量某一区间内数据分布是否对称的一种指标,用来描述该区间数据分布的偏斜方向和程度。如果数据分布左右对称,偏度为零;如果数据分布

向右偏,说明较大的数值出现得更少但拉长了尾部,偏度为正;如果数据分布向左偏,说明较小的数值更稀少但形成了较长的尾部,偏度为负。局部分布偏度专注于某一区间内的分布特性,而不是整体数据的分布情况,因此能帮助更细致地分析数据在不同区间内的分布特征。

[0103] 在一种具体实施方式中,首先对数值型特征的数据分布进行统计分析。通过计算数值型特征的最大值、最小值确定整体值域范围。基于数据的四分位距和样本总量,设定基础统计区间的宽度。比如对于样本量为一万条的数据集,其中年龄特征的四分位距为二十岁,可以将基础区间宽度设置为五岁。按照这个基础区间宽度,将值域范围划分为若干个基础统计区间,统计每个区间内的样本数量,得到初始的区间频数序列。

[0104] 接着分析相邻基础统计区间的数据分布相似性。计算相邻区间的频数比值,设定合并阈值为零点八。以年龄特征为例,如果二十岁到二十五岁区间的样本数为一千,二十五岁到三十岁区间的样本数为九百,则这两个区间的相似度为零点九,大于合并阈值,可以将这两个区间合并为一个新的统计区间。依次处理所有相邻区间,得到合并后的频数分布序列。

[0105] 基于合并后的频数分布序列,计算各区间的累积频数,并构建数据密度函数。在频数分布序列中选取频数变化明显的位置作为候选分割位置。对每个候选位置,分别计算其左右两侧区间的样本数量和数值均值。以此计算该位置的区间方差值,表征该位置作为分段点的区分效果。

[0106] 在当前需要分割的区间内,选择具有最大区间方差值的位置作为待定分段点。计算该待定分段点对应的区间方差值与特征总体方差的比值,得到方差贡献比。同时计算待定分段点两侧区间的数据分布偏度。设定偏度阈值为零点五,分割阈值为零点三。当两侧区间的偏度差值大于偏度阈值,且方差贡献比大于分割阈值时,确认该待定分段点为最优分段点。

[0107] 重复执行上述分析过程,直到所有区间的分割效果都不满足阈值要求,即可得到所有最优分段点的位置。以年龄特征为例,可能的最优分段点包括十八岁、三十五岁、五十岁等,这些分段点将样本数据划分为具有显著统计特征差异的多个区间。

[0108] 在本实施例中,通过计算相邻区间的频数比值进行区间合并,避免了数据稀疏区间对分段结果的干扰,提高了分段结果的稳定性和可靠性;采用区间方差值和方差贡献比作为分段点评价指标,结合区间分布偏度的差异性约束,确保了分段点位置的统计显著性,使得分段结果具有良好的区分度;基于数据分布特征自适应确定最优分段点,无需人工指定分段数量,避免了主观因素的影响,提升了分段方法在不同场景下的适用性。

[0109] 在一种可选的实施方式中,将所述增强训练样本集输入包含层次注意力模块和自校准模块的特征提取网络,所述层次注意力模块计算样本特征的相似度矩阵,基于所述相似度矩阵构建特征亲和图,通过图分割算法将所述特征亲和图划分为多个特征子图;在每个所述特征子图内,采用因果注意力机制提取局部特征表示,并通过跨子图的注意力融合机制整合多个局部特征表示,得到全局特征表示;所述自校准模块接收所述全局特征表示,通过残差连接和动态归一化进行特征非线性变换,得到深层特征表示包括:

[0110] 将增强训练样本集输入卷积神经网络得到初始特征映射,计算所述初始特征映射中任意两个样本特征的点积,并将所述点积除以对应特征向量的模长乘积,得到特征相似

度矩阵;基于所述特征相似度矩阵构建特征亲和图,所述特征亲和图的顶点对应样本特征,所述特征亲和图的边权重对应所述特征相似度矩阵中的相似度值,通过谱聚类算法将所述特征亲和图划分为多个特征子图;

[0111] 在每个所述特征子图内,将所述特征子图的特征序列划分为查询特征序列和键值对特征序列,构建时序因果掩码矩阵,利用所述时序因果掩码矩阵对所述查询特征序列与所述键值对特征序列的交互结果进行屏蔽,得到仅包含历史特征信息的初始注意力权重;计算每个所述特征子图内样本特征的空间距离矩阵,基于所述空间距离矩阵计算所述特征子图的空间分布熵,根据所述空间分布熵对所述初始注意力权重进行自适应调节,得到自适应注意力权重;基于所述自适应注意力权重对所述键值对特征序列进行加权,并叠加位置编码信息,得到每个所述特征子图的局部特征表示;

[0112] 基于预先构建的多层感知机网络,输入相邻特征子图的局部特征表示的拼接向量,经所述多层感知机网络前向传播,得到特征子图间的关联权重,基于所述关联权重对所述局部特征表示进行加权求和,得到全局特征表示;

[0113] 对所述全局特征表示在特征维度上分别计算均值和标准差;基于预先构建的动态归一化网络,将所述全局特征表示输入所述动态归一化网络,经前向传播得到特征变换系数和特征偏移量;将所述全局特征表示减去所述均值后除以所述标准差得到标准化特征,将所述标准化特征与所述特征变换系数相乘,并加上所述特征偏移量得到变换特征,将所述变换特征与所述全局特征表示相加,得到自校准特征;

[0114] 对所述自校准特征进行层归一化处理,并通过ReLU激活函数进行非线性变换,得到深层特征表示。

[0115] 在一种具体实施方式中,首先,对输入的训练样本集进行数据增强处理。数据增强包括随机裁剪、水平翻转、亮度和对比度调整等操作,生成增强训练样本集。增强后的样本数量是原始样本的4倍,每个原始样本通过不同的增强方式产生3个新样本。

[0116] 将增强后的训练样本输入卷积神经网络进行特征提取。采用ResNet50作为基础网络,提取多尺度特征。在最后一个卷积层输出特征图的大小为14x14,通道数为2048。对于每个样本,将特征图展平为一维向量,得到初始特征映射。

[0117] 计算特征相似度矩阵时,对任意两个样本特征向量进行点积运算,并除以各自的模长乘积进行归一化。例如,对于两个2048维的特征向量,计算它们的余弦相似度,得到的相似度值范围在-1到1之间。基于相似度矩阵构建特征亲和图,图中的顶点表示样本特征,边的权重为对应的相似度值。

[0118] 使用谱聚类算法对特征亲和图进行划分。首先计算图的拉普拉斯矩阵,选取其最小的k个特征值对应的特征向量,其中k为预设的子图数量,通常设置为4到8之间。对特征向量进行K-means聚类,将样本划分到不同的特征子图中。

[0119] 在每个特征子图内实现因果注意力机制。将特征序列等分为查询序列和键值对序列。构建上三角形因果掩码矩阵,确保当前时刻只能访问历史信息。计算特征间的欧氏距离构建空间距离矩阵,并计算归一化的空间分布熵。根据熵值自适应调节注意力权重,熵值越大表示特征分布越分散,对应的注意力权重越小。

[0120] 对于特征子图间的注意力融合,构建三层的多层感知机网络。隐层维度分别为1024、512和256。输入相邻子图的局部特征拼接向量,输出它们之间的关联权重。根据关联

权重对所有局部特征进行加权求和,得到全局特征表示。

[0121] 在自校准模块中,首先计算全局特征在特征维度上的统计量。构建双层的动态归一化网络,隐层维度为1024。网络输出特征变换系数和偏移量,用于对标准化特征进行自适应调整。将调整后的特征与原特征相加,实现残差连接。

[0122] 最后对自校准特征进行层归一化,并通过ReLU激活函数得到最终的深层特征表示。特征维度保持为2048,便于后续任务使用。

[0123] 示例性地,以电商平台用户购买意向预测任务为例,首先将90万条增强训练样本输入卷积神经网络。该卷积网络包含3个卷积层,卷积核大小分别为 3×3 、 3×3 、 5×5 ,输出通道数分别为64、128、256。经过卷积网络处理后,得到维度为256的初始特征映射。对任意两个样本特征计算点积相似度,并进行归一化处理,得到大小为90万 $\times$ 90万的特征相似度矩阵。

[0124] 基于特征相似度矩阵构建特征亲和图。每个样本特征作为图的一个顶点,样本间的相似度值作为边的权重。例如,用户A的特征向量与用户B的特征向量相似度为0.85,则在亲和图中这两个顶点之间的边权重为0.85。使用谱聚类算法将特征亲和图划分为8个特征子图,每个子图包含具有相似行为模式的用户群体。

[0125] 在特征子图内部进行特征序列处理。以包含10万用户的子图为例,将特征序列按时间顺序排列,并基于特征相似度阈值0.7将序列划分为查询特征序列和键值对特征序列。构建大小为10万 $\times$ 10万的时序因果掩码矩阵,矩阵中只保留当前时刻之前的历史信息,对未来时刻的信息进行屏蔽。将掩码矩阵应用于查询序列和键值对序列的交互结果,得到初始注意力权重。

[0126] 计算子图内样本特征的空间距离矩阵。对于每对样本,基于其特征向量计算欧氏距离,构建大小为10万 $\times$ 10万的距离矩阵。基于距离矩阵计算空间分布熵,熵值反映了特征分布的均匀程度。例如,某子图的空间分布熵为0.75,表明特征分布较为均匀;另一子图的熵值为0.45,表明特征分布较为集中。根据熵值对初始注意力权重进行调节:对于分布均匀的子图,保持初始权重;对于分布集中的子图,增大局部区域的注意力权重。得到的自适应注意力权重用于对键值对特征序列进行加权。

[0127] 为了保留特征的位置信息,构造位置编码向量。位置编码的维度与特征维度相同,均为256。将位置编码与加权后的特征序列相加,得到每个子图的局部特征表示。例如,某子图的局部特征表示捕获了该用户群体在近期购物行为上的共同特征。

[0128] 构建四层的多层感知机网络,各层的神经元数量为512、256、128、64。输入相邻两个特征子图的局部特征表示的拼接向量,经网络前向传播得到子图间的关联权重。例如,表示年轻用户群体的子图与表示中年用户群体的子图之间的关联权重为0.6,表明这两个群体的购物行为具有一定相关性。基于关联权重对所有子图的局部特征表示进行加权求和,得到维度为256的全局特征表示。

[0129] 对全局特征表示进行自校准处理。首先在特征维度上计算均值和标准差。构建动态归一化网络,包含两个全连接层,神经元数量分别为128和256。将全局特征表示输入该网络,得到特征变换系数和特征偏移量。对全局特征表示进行标准化处理后,与变换系数相乘并加上偏移量,得到变换特征。将变换特征与原始全局特征表示相加,得到自校准特征。

[0130] 最后,对自校准特征执行层归一化操作,并通过ReLU激活函数进行非线性变换,得

到最终的深层特征表示。深层特征表示维度保持为256,其中每个维度都包含了原始特征经过多层次转换和融合后的信息。这些特征既保留了局部用户群体的行为特征,又包含了不同群体之间的关联信息,为后续的特征选择和预测任务提供了高质量的特征表示。

[0131] 在本实施例中,通过层次化的注意力机制和特征子图划分,能够更好地捕获样本间的局部相关性,提高特征表示的判别能力。子图内的因果注意力确保了时序信息的有效利用,避免信息泄露;引入空间分布熵对注意力权重进行自适应调节,使得注意力分配更加合理,能够突出重要特征的贡献。特征子图间的关联权重学习实现了多尺度特征的有效融合;自校准模块通过动态归一化和残差连接,增强了特征的非线性表达能力,同时保持了原始信息。这种设计提高了模型的泛化性能,使得提取的特征更加鲁棒。

[0132] 在一种可选的实施方式中,将所述特征子图的特征序列划分为查询特征序列和键值对特征序列,构建时序因果掩码矩阵,利用所述时序因果掩码矩阵对所述查询特征序列与所述键值对特征序列的交互结果进行屏蔽,得到仅包含历史特征信息的初始注意力权重包括:

[0133] 基于特征子图中相邻时间步特征的相似度,对所述特征子图的特征序列进行分组,得到查询特征序列和键值对特征序列,其中,所述相邻时间步特征的相似度大于基于所述特征序列的全局统计特性确定的动态阈值的时间步特征划分为所述查询特征序列,其余特征划分为所述键值对特征序列,得到所述查询特征序列和所述键值对特征序列的特征划分结果;

[0134] 基于所述特征划分结果,通过递归神经网络分别计算所述查询特征序列和所述键值对特征序列的时序关联强度,所述递归神经网络的隐状态包含所述查询特征序列和所述键值对特征序列的长短期依赖信息,得到所述查询特征序列的时序关联强度和所述键值对特征序列的时序关联强度;

[0135] 基于所述特征序列的时序位置关系构建第一掩码矩阵作为时序因果掩码矩阵,将所述查询特征序列的时序关联强度和所述键值对特征序列的时序关联强度转换为软掩码权重,得到第二掩码矩阵;

[0136] 将所述查询特征序列与所述键值对特征序列进行矩阵乘法运算,得到原始注意力分数;将所述原始注意力分数与所述第一掩码矩阵进行加法运算,得到第一次掩码结果,将所述第一次掩码结果与所述第二掩码矩阵进行元素级乘法运算,得到双重掩码后的注意力分数;对双重掩码后的注意力分数进行归一化处理,得到仅包含历史特征信息的初始注意力权重。

[0137] 所述软掩码权重具体是指一种基于特征的重要性动态调整的权重分配方式,用于在处理特征序列时,强调不同特征对整体计算的贡献程度。它通过结合特征的时序关联强度和其他上下文信息,生成介于0和1之间的权重,用来表示某个特征在当前计算中应被关注的程度。

[0138] 软掩码权重的特点是连续可调,相较于硬掩码(完全屏蔽或完全通过),它能够以柔性的方式部分减少或增强特定特征的影响。这种动态权重能够根据数据的特性,在模型的运算过程中自适应地分配注意力资源,从而更准确地捕捉特征之间的时序关系和依赖性。

[0139] 在一种具体实施方式中,首先对特征子图中的特征序列进行动态划分。通过计算

相邻时间步特征之间的余弦相似度来衡量特征的相似程度。具体地,对于特征序列中任意相邻的两个时间步特征,计算它们的余弦相似度值。同时基于特征序列的均值和标准差统计特性,设定动态阈值。当相邻时间步特征的相似度大于该动态阈值时,将这些特征划分到查询特征序列中,其余特征划分为键值对特征序列。例如,对于长度为10的特征序列,若第2-4、第7-8时间步的特征相似度高于阈值0.8,则这些时间步的特征划分为查询特征序列,其余时间步的特征划分为键值对特征序列。

[0140] 接着利用长短时记忆网络分别处理查询特征序列和键值对特征序列,提取它们的时序关联信息。长短时记忆网络通过门控机制选择性地保留和更新特征信息,其隐状态包含了特征序列的长期依赖关系和短期相关性。对于查询特征序列,网络的隐状态输出表示其时序关联强度;对于键值对特征序列,网络的隐状态输出表示其时序关联强度。

[0141] 然后构建双重掩码机制。第一重掩码基于特征的时序位置关系构建因果掩码矩阵,确保当前时间步只能访问历史信息。具体实现时,将上三角区域的值设为负无穷,下三角区域的值设为0。第二重掩码将查询序列和键值对序列的时序关联强度转换为0到1之间的软掩码权重,形成软掩码矩阵。

[0142] 最后计算注意力权重。将查询特征序列与键值对特征序列进行矩阵乘法得到原始注意力分数,与因果掩码矩阵相加得到第一次掩码结果。将第一次掩码结果与软掩码矩阵进行逐元素相乘得到双重掩码后的注意力分数。对该分数进行softmax归一化,得到最终的注意力权重。这个注意力权重仅包含历史时序信息,可用于后续的特征融合。

[0143] 示例性地,以电商用户行为序列分析为例,对特征子图中的用户行为序列进行处理。首先计算特征序列中相邻时间步特征的相似度。假设某用户在连续30天内的行为序列,每天的行为特征向量维度为256。对于任意相邻两天的行为特征,计算它们的余弦相似度值。

[0144] 基于特征序列的全局统计特性确定动态阈值。统计30天内所有相邻时间步特征相似度的均值和标准差,例如均值为0.65,标准差为0.15。将均值加上0.5倍标准差作为动态阈值,在本例中阈值值为0.725。对相邻时间步的相似度值与该阈值进行比较,将相似度大于阈值的时间步特征划分为查询特征序列,其余特征划分为键值对特征序列。

[0145] 例如,第3天和第4天的行为特征相似度为0.78,大于阈值0.725,则第4天的特征划入查询特征序列;第15天和第16天的特征相似度为0.68,小于阈值,则第16天的特征划入键值对特征序列。经过划分后,得到包含12个时间步的查询特征序列和包含18个时间步的键值对特征序列。

[0146] 构建双向长短期记忆网络(Bi-LSTM)作为递归神经网络,用于计算时序关联强度。网络包含两层Bi-LSTM结构,隐层单元数为128。将查询特征序列输入该网络,网络的隐状态包含了特征序列的长短期依赖信息。对于12个时间步的查询特征序列,得到维度为 12×128 的时序关联强度矩阵。同样将18个时间步的键值对特征序列输入网络,得到维度为 18×128 的时序关联强度矩阵。

[0147] 基于特征序列的时序位置关系构建时序因果掩码矩阵。矩阵大小为 30×30 ,对应完整的30天序列。在矩阵中,只保留当前时间步及之前时间步的信息,将之后时间步的位置标记为负无穷大,作为第一掩码矩阵。例如,对于第10天的预测,只考虑第1天到第10天的特征信息,第11天到第30天的信息被掩码屏蔽。

[0148] 将查询特征序列和键值对特征序列的时序关联强度转换为软掩码权重。对关联强度进行Softmax归一化处理,得到第二掩码矩阵。该矩阵反映了不同时间步特征之间的关联程度,关联强度越大的时间步对应的掩码权重越大。

[0149] 将查询特征序列和键值对特征序列进行矩阵乘法运算,得到原始注意力分数矩阵。矩阵维度为查询序列长度乘以键值对序列长度,即 12×18 。将该矩阵与第一掩码矩阵相加,实现时序因果掩码,确保预测时只使用历史信息。然后将掩码结果与第二掩码矩阵进行元素级乘法运算,得到同时考虑时序因果关系和特征关联强度的双重掩码注意力分数。

[0150] 最后对双重掩码后的注意力分数进行Softmax归一化处理,得到初始注意力权重。这些权重体现了查询序列中每个时间步与键值对序列中历史时间步的关联程度。例如,第10天的查询特征与第8天的键值对特征的注意力权重为0.15,表明这两个时间步的特征具有较强的相关性。

[0151] 在本实施例中,通过动态特征序列划分机制,根据特征相似度自适应地划分查询序列和键值对序列,提高了特征表示的区分度,增强了模型对重要特征的感知能力;采用双重掩码机制,既保证了严格的时序因果性,又通过软掩码权重引入了灵活的特征选择机制,有效抑制了无关特征的干扰,提升了模型的鲁棒性;基于递归神经网络提取时序关联强度,充分利用了特征序列的长短期依赖信息,增强了模型对时序模式的建模能力,提高了特征表示的时序一致性。

[0152] 在一种可选的实施方式中,构建信息瓶颈特征筛选器,通过最大化任务相关信息和最小化冗余信息,对所述深层特征表示进行筛选,得到关键特征子集包括:

[0153] 构建信息瓶颈编码器,所述信息瓶颈编码器包括三层感知机结构,每层感知机结构之后连接归一化层,将所述深层特征表示输入所述信息瓶颈编码器,所述信息瓶颈编码器的输出层生成压缩特征空间的概率分布参数;基于所述概率分布参数对所述深层特征表示进行编码映射,得到压缩特征表示;

[0154] 构建互信息评估模块,所述互信息评估模块包含三层感知机结构,将所述压缩特征表示与目标任务的标签的拼接向量输入所述互信息评估模块,输出所述压缩特征表示与所述目标任务之间的互信息量;

[0155] 计算所述压缩特征表示的自相关矩阵,所述自相关矩阵中每个元素为对应特征向量减去相应的均值后的乘积,除以对应标准差的乘积;基于所述自相关矩阵与单位矩阵的差值,计算特征冗余度;

[0156] 构建特征选择网络,所述特征选择网络包括第一全连接层和第二全连接层,将所述压缩特征表示输入所述第一全连接层,所述第一全连接层的输出经过激活函数后输入第二全连接层,所述第二全连接层的输出经过门控函数得到门控值,将所述门控值与所述压缩特征表示进行逐元素相乘,得到初始筛选特征;

[0157] 基于温度参数对所述门控函数的输出进行调节,所述温度参数基于预设的递减量随训练轮数递减,将调节后的门控值与所述压缩特征表示进行逐元素相乘,得到调节后筛选特征;

[0158] 将所述互信息量的相反数、所述特征冗余度以及所述调节后筛选特征的范数线性组合构建目标函数;

[0159] 通过迭代优化所述目标函数,得到所述信息瓶颈编码器的优化参数和所述特征选

择网络的优化参数,获得优化信息瓶颈编码器和优化特征选择网络;

[0160] 将所述深层特征表示输入优化信息瓶颈编码器得到优化压缩特征表示,将所述优化压缩特征表示输入优化特征选择网络,得到关键特征子集。

[0161] 在一种具体实施方式中,信息瓶颈特征筛选器的构建过程首先需要搭建信息瓶颈编码器。该编码器采用三层感知机结构,每层感知机的神经元数量分别为1024、512和256,使用ReLU作为激活函数。每层感知机后接归一化层,用于加速训练收敛。输入的深层特征表示维度为2048,经过编码器处理后输出两组参数:均值向量和方差向量,分别用于构建压缩特征空间的概率分布。基于重参数化技术,从该概率分布中采样得到维度为256的压缩特征表示。

[0162] 互信息评估模块同样采用三层感知机结构,神经元数量为512、256和1。将压缩特征表示与任务标签进行拼接作为输入,标签采用one-hot编码形式。通过非线性变换,最终输出一个标量值,表示压缩特征与任务之间的互信息估计。

[0163] 特征冗余度的计算首先需要对压缩特征进行标准化处理。计算每个特征维度的均值和标准差,将特征减去均值后除以标准差。然后计算标准化特征之间的内积,得到自相关矩阵。该矩阵与单位矩阵的差值的平方和即为特征冗余度量。

[0164] 特征选择网络包含两个全连接层,第一层神经元数量为512,使用ReLU激活函数;第二层神经元数量与压缩特征维度相同,即256,使用Sigmoid函数作为门控函数。网络输入压缩特征表示,输出门控值介于0到1之间。门控值与压缩特征逐元素相乘,实现特征的软选择。

[0165] 为了使特征选择更加稀疏,引入温度参数对门控函数进行调节。温度参数初始值设为1.0,每训练100轮递减0.1,直至最小值0.1。温度参数越小,门控值越趋近于0或1,特征选择越稀疏。

[0166] 目标函数由三部分组成:互信息损失、冗余度损失和稀疏性损失。互信息损失取互信息估计值的相反数;冗余度损失为特征冗余度;稀疏性损失为调节后筛选特征的L1范数。三项损失的权重比例为1:0.5:0.1。

[0167] 在训练过程中,采用Adam优化器进行参数更新,学习率设为0.001。每个mini-batch包含128个样本,总共训练50个epoch。训练完成后,使用优化后的编码器和特征选择网络对新的深层特征进行处理,得到最终的关键特征子集。

[0168] 在本实施例中,通过信息瓶颈编码器实现特征压缩和信息提取,有效降低特征维度,提高特征表示的紧凑性和有效性,同时保留任务相关的关键信息;基于互信息评估和冗余度度量的双重约束,确保筛选得到的特征既与任务高度相关,又具有较低的信息冗余,提高特征表示的效率和区分性;采用可学习的特征选择网络结合温度调节机制,实现特征的自适应软选择,既保持了选择过程的可微性,又能得到稀疏的特征表示,提高模型的可解释性和泛化能力。

[0169] 在一种可选的实施方式中,根据所述特征权重向量和所述预测结果,采用结构化梯度方法计算所述关键特征子集中各特征对所述预测结果的贡献度,生成分析结果包括:

[0170] 基于特征权重向量和预测结果,计算所述特征权重向量中每对特征之间的相关系数,生成特征相关性矩阵;基于所述特征相关性矩阵构建特征依赖图,将所述关键特征子集中的每个特征作为所述特征依赖图中的节点,将所述特征相关性矩阵中的相关系数作为对

应节点之间的边权重；

[0171] 根据所述特征依赖图构建图拉普拉斯矩阵,将各节点与对应相连节点的边权重之和作为所述图拉普拉斯矩阵的对角元素值,将节点间的边权重的负值作为所述图拉普拉斯矩阵的非对角元素值；

[0172] 计算所述预测结果关于所述关键特征子集的雅可比矩阵,得到初始梯度;将所述图拉普拉斯矩阵作为结构化正则项,将所述结构化正则项与所述初始梯度进行加权组合,得到考虑特征依赖关系的结构化梯度；

[0173] 计算所述关键特征子集中每个特征的结构化梯度幅值、每个特征与其他特征的交互梯度值、以及在所述特征依赖图中每个特征的中心性度量；

[0174] 将所述结构化梯度幅值、所述交互梯度值和所述中心性度量进行加权组合,得到特征贡献度;对所述特征贡献度进行归一化处理,得到所述关键特征子集中各特征对所述预测结果的标准化贡献度;基于所述标准化贡献度对所述关键特征子集中的特征进行重要性排序,生成分析结果。

[0175] 在一种具体实施方式中,首先获取特征权重向量和预测结果数据。以一个包含十个关键特征的数据集为例,特征包括温度、湿度、压力等环境参数。预测结果为某种产品的质量指标。

[0176] 对特征权重向量中的特征对进行相关性分析。通过计算皮尔逊相关系数得到特征相关性矩阵。例如温度与湿度的相关系数为零点七,表示较强的正相关关系。将所有特征对的相关系数计算完成后,得到完整的特征相关性矩阵。

[0177] 基于特征相关性矩阵构建特征依赖图。将每个特征作为图中的节点,相关系数作为连接节点的边的权重。例如温度节点和湿度节点之间的边权重为零点七。对相关系数低于阈值的特征对,可以不在图中连接边,以突出重要的特征依赖关系。

[0178] 构建图拉普拉斯矩阵时,计算每个节点与其相连节点的边权重之和作为对角元素。例如温度节点与其他节点的边权重之和为二点五,则作为对角元素。非对角元素取对应节点间边权重的负值,没有边相连的节点对应位置填零。

[0179] 计算预测结果对每个特征的初始梯度。采用有限差分方法,对每个特征进行微小扰动,观察预测结果的变化。将图拉普拉斯矩阵与初始梯度进行加权组合,得到考虑特征依赖关系的结构化梯度。加权系数可以根据实际需求调整,用于平衡特征独立性与依赖性的影响。

[0180] 计算每个特征的结构化梯度幅值,表示特征对预测结果的直接影响程度。计算特征间的交互梯度,表示特征协同作用的强度。计算特征在依赖图中的中心性,包括度中心性、接近中心性等指标,表示特征在整体特征网络中的重要程度。

[0181] 将上述三个指标进行加权组合得到特征贡献度。例如可以设置结构化梯度幅值权重为零点五,交互梯度权重为零点三,中心性权重为零点二。对特征贡献度进行归一化处理,使所有特征的贡献度之和为一。最后根据标准化贡献度对特征进行排序,生成分析结果报告。

[0182] 在本实施例中,通过构建特征依赖图和图拉普拉斯矩阵,有效捕捉了特征之间的复杂依赖关系,避免了传统特征重要性分析方法忽视特征交互作用的问题,提高了特征贡献度评估的准确性;采用结构化梯度方法,将特征的直接影响、交互作用和网络中心性进行

综合考虑,实现了多维度的特征重要性评估,使分析结果更加全面和可靠,能够更好地指导实际应用中的特征选择和优化;通过标准化处理和排序,得到直观清晰的特征重要性排序结果,便于理解和应用,同时保证了不同场景下分析结果的可比性,为后续的模型优化和决策提供了可靠的依据。

[0183] 图2为本发明实施例基于Transformer的结构化数据建模分析系统的结构示意图,如图2所示,所述系统包括:

[0184] 第一单元,用于接收结构化原始数据,将所述结构化原始数据划分为数值型特征和类别型特征,通过双流特征编码器,对所述数值型特征执行分段线性变换,得到归一化主干特征,对所述类别型特征执行自适应量子编码,得到连续值映射特征;通过可学习的门控机制进行自适应权重分配,融合所述归一化主干特征和所述连续值映射特征,得到融合特征;基于所述融合特征的重要性评分进行动态采样,生成增强训练样本集;

[0185] 第二单元,用于将所述增强训练样本集输入包含层次注意力模块和自校准模块的特征提取网络,所述层次注意力模块计算样本特征的相似度矩阵,基于所述相似度矩阵构建特征亲和图,通过图分割算法将所述特征亲和图划分为多个特征子图;在每个所述特征子图内,采用因果注意力机制提取局部特征表示,并通过跨子图的注意力融合机制整合多个局部特征表示,得到全局特征表示;所述自校准模块接收所述全局特征表示,通过残差连接和动态归一化进行特征非线性变换,得到深层特征表示;

[0186] 第三单元,用于构建信息瓶颈特征筛选器,通过最大化任务相关信息和最小化冗余信息,对所述深层特征表示进行筛选,得到关键特征子集;将所述关键特征子集输入包含主预测分支和辅助重建分支的双分支网络,所述主预测分支基于注意力机制对所述关键特征子集进行处理得到特征权重向量,并根据所述特征权重向量生成预测结果,所述辅助重建分支通过重建所述关键特征子集提供正则化约束;根据所述特征权重向量和所述预测结果,采用结构化梯度方法计算所述关键特征子集中各特征对所述预测结果的贡献度,生成分析结果。

[0187] 本发明实施例的第三方面,

[0188] 提供一种电子设备,包括:

[0189] 处理器;

[0190] 用于存储处理器可执行指令的存储器;

[0191] 其中,所述处理器被配置为调用所述存储器存储的指令,以执行前述所述的方法。

[0192] 本发明实施例的第四方面,

[0193] 提供一种计算机可读存储介质,其上存储有计算机程序指令,所述计算机程序指令被处理器执行时实现前述所述的方法。

[0194] 本发明可以是方法、装置、系统和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质,其上载有用于执行本发明的各个方面的计算机可读程序指令。

[0195] 最后应说明的是:以上各实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述各实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分或者全部技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的范围。

接收结构化原始数据，将所述结构化原始数据划分为数值型特征和类别型特征，通过双流特征编码器，对所述数值型特征执行分段线性变换，得到归一化主干特征，对所述类别型特征执行自适应量子编码，得到连续值映射特征；通过可学习的门控机制进行自适应权重分配，融合所述归一化主干特征和所述连续值映射特征，得到融合特征；基于所述融合特征的重要性评分进行动态采样，生成增强训练样本集

S101

将所述增强训练样本集输入包含层次注意力模块和自校准模块的特征提取网络，所述层次注意力模块计算样本特征的相似度矩阵，基于所述相似度矩阵构建特征亲和图，通过图分割算法将所述特征亲和图划分为多个特征子图；在每个所述特征子图内，采用因果注意力机制提取局部特征表示，并通过跨子图的注意力融合机制整合多个局部特征表示，得到全局特征表示；所述自校准模块接收所述全局特征表示，通过残差连接和动态归一化进行特征非线性变换，得到深层特征表示

S102

构建信息瓶颈特征筛选器，通过最大化任务相关信息和最小化冗余信息，对所述深层特征表示进行筛选，得到关键特征子集；将所述关键特征子集输入包含主预测分支和辅助重建分支的双分支网络，所述主预测分支基于注意力机制对所述关键特征子集进行处理得到特征权重向量，并根据所述特征权重向量生成预测结果，所述辅助重建分支通过重建所述关键特征子集提供正则化约束；根据所述特征权重向量和所述预测结果，采用结构化梯度方法计算所述关键特征子集中各特征对所述预测结果的贡献度，生成分析结果

S103

图1



图2