# 摘要

一种多源纵向联邦学习的无监督灵活就业人员公积金缴存推荐方法

一种无监督情况下面向灵活就业人员的公积金缴存推荐方法

一种联合多方数据的无监督灵活就业人员公积金缴存推荐方法

本发明涉及一种联合多方数据的无监督灵活就业人员公积金缴存推荐方法，S1：对灵活就业人员多方数据源进行预处理，建立基于多方的已缴存灵活就业人员特征数据集和未缴存灵活就业人员特征数据集；S2：构建一个面向多方纵向联邦学习的分类器，由基于各方数据源建立的神经网络编码器子模型组成，对已缴存人员的各方特征信息进行学习，形成其对应的特征编码子向量；S3：建立特征中心向量及各方中心子向量，对各方特征编码子向量采用向量距离损失函数和中心损失函数进行相似度度量计算，并加密传输给中心服务器；S4：在中心服务器中对各方损失进行聚合，然后将聚合结果解密返回给各方编码器子模型，进行参数更新；S5：将各方未缴存灵活就业人员的特征信息输入训练好的联邦学习编码器，得到各方特征编码子向量。再与已缴存灵活就业人员的各方特征编码向量和特征中心向量进行相似度计算并聚合，根据聚合结果进行分类推荐。本发明在保证多方数据安全隐私的情况下，解决基于多方数据的无监督灵活就业人员精准分类推荐问题。

灵活就业人员1： [xxxxx] [xxxxx] [xxxxxx] [xxxxxx]（加密后）

灵活就业人员2： [xxxxx] [xxxxx] [xxxxxx] [xxxxxx]（加密后）

流程图

1. 一种联合多方数据的无监督灵活就业人员公积金缴存推荐方法，其特征在于：包括以下步骤：

S1：对灵活就业人员多方数据源进行预处理，建立基于多方的已缴存灵活就业人员特征数据集和未缴存灵活就业人员特征数据集；

S2：构建一个面向多方纵向联邦学习的分类器，由基于各方数据源建立的神经网络编码器子模型组成，对已缴存人员的各方特征信息进行学习，形成其对应的特征编码子向量；

S3：建立特征中心向量及各方中心子向量，对各方特征编码子向量采用向量距离损失函数和中心损失函数进行相似度度量计算，并加密传输给中心服务器；

S4：在中心服务器中对各方损失进行聚合，然后将聚合结果解密返回给各方编码器子模型，进行参数更新；

S5：将各方未缴存灵活就业人员的特征信息输入训练好的联邦学习编码器，得到各方特征编码子向量。再与已缴存灵活就业人员的各方特征编码向量和特征中心向量进行相似度计算并聚合，根据聚合结果进行分类推荐。

**一种联合多方数据的无监督灵活就业人员公积金缴存推荐方法**

### 技术领域

本发明属于预测方法应用领域，涉及一种联合多方数据的无监督灵活就业人员公积金缴存推荐方法。

### 背景技术

### 发明内容

有鉴于此，本发明的目的在于提供一种联合多方数据的无监督灵活就业人员公积金缴存推荐方法。

为达到上述目的，本发明提供如下技术方案：

一种联合多方数据的无监督灵活就业人员公积金缴存推荐方法，包括以下步骤：

S1：对灵活就业人员多方数据源进行预处理，建立基于多方的已缴存灵活就业人员特征数据集和未缴存灵活就业人员特征数据集；

进一步，所述步骤S1中多方数据源指关于灵活就业人员的税务数据、房管数据、社保数据和个人基本信息数据。

使用灵活就业人员的税务数据源，获取税务数据，包括个人所得税。

使用房管数据源，获取房管数据，包括。

使用社保数据源，获取社保数据，包括。

使用个人基本信息数据源，获取个人基本信息数据，包括。（实例中说明，这里只需要说明使用哪些变量）

进一步地，所述S1中预处理操作包括离群值处理、缺失值处理和数据标准化。

所述离群值处理具体包括：采用箱型图的方式对数据进行可视化统计，剔除其中明显偏离大部分数据的异常值。

所述缺失值处理具体包括：对数据中的每个特征进行缺失率统计，对于缺失数据数目大于总体样本数据规模一半的数据视为无效数据进行剔除，其余数据根据数据分布采用特定方法进行填充，包括众数填充、近邻填充、均值填充和中位数填充。

所述数据标准化具体包括：将已有特征按照数据类型分为连续型特征和离散型特征，对连续型特征采用数据最大-最小归一化处理，对离散型特征采用独热编码处理。

再根据灵活就业人员的公积金缴存情况，将多方源数据分为已缴存灵活就业人员多方源特征数据集和未缴存灵活就业人员多方源特征数据集。

S2：构建一个面向多方纵向联邦学习的分类器，对已缴存人员的各方特征信息进行学习，形成其对应的特征编码子向量；

所述S2中的多方纵向联邦学习的分类器由多个基于神经网络的编码器构成，每一个编码器都对应一方特征数据，通过网络学习找到特征数据存在的最佳特征表达形式，即特征编码子向量。

所述S2中的特征编码子向量为神经网络编码器的输出，其维度可作为超参数进行调节。对于每一方数据源都建立对应的编码器，编码器的公式表示如下：（了解多种自编码器）

其中*NN*表示神经网络，*W*代表可学习参数，*X*可以代表输入的各方特征数据，*F*代表输出的特征编码子向量。*F*可表示为：

其中*n*代表样本个数，*m*代表特征编码向量的最大维度，代表第*n*个样本的经过编码器后的特征向量表示，代表第*n*个样本第*m*个特征。（多方如何表示）

S3：建立特征中心向量及各方中心子向量，对各方特征编码子向量采用向量距离损失函数和中心损失函数进行相似度度量计算，并加密传输给中心服务器；

S31：根据S2中编码器输出的各方特征数据的特征编码子向量，建立其对应的特征中心向量*C*；

特别地，特征中心向量初始值获取方式为自定义设定，后续将通过参数更新来进行修正。*C*的表示方式：

其中，表示*m*特征的中心值。

S32：采用向量距离函数和中心损失函数对编码器输出的特征编码子向量进行距离计算，两个向量之间的距离函数采用欧式距离，具体表示公式如下：

根据欧式距离进行计算的向量距离损失函数为：

所述S32中的中心损失函数是深度学习中常用的一种损失函数，它可以降低类内之间距离，中心损失函数可表示为：

（前面的1/2可以换成）

其中代表*j*特征的特征中心。

中心损失函数的梯度为：

最终总的损失函数为：

（纵向梯度计算）(收敛问题需要考虑)

S33：将各方特征数据对应的编码器的计算的损失结果进行同态加密操作，将加密结果传递给中心服务器。

其中同态加密操作使用同态加密算法

S4：在中心服务器中对各方损失进行聚合，然后将聚合结果解密返回给各方编码器子模型，进行参数更新；

S5：将各方未缴存灵活就业人员的特征信息输入训练好的联邦学习编码器，得到各方特征编码子向量。再与已缴存灵活就业人员的各方特征编码向量和特征中心向量进行相似度计算并聚合，根据聚合结果进行分类推荐。

另一方面，本发明提供一种处理装置，包括处理器和存储装置，所述处理器，适用于执行各条程序，所述存储装置，适用于存储多条程序，所述程序适用于由处理器加载并执行以实现如上任一所述的方法。

本发明的有益效果在于：

本发明的其他优点、目标和特征在某种程度上将在随后的说明书中进行阐述，并且在某种程度上，基于对下文的考察研究对本领域技术人员而言将是显而易见的，或者可以从本发明的实践中得到教导。本发明的目标和其他优点可以通过下面的说明书来实现和获得。

### 附图说明

为了使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明作优选的详细描述，其中：

### 具体实施方式

以下通过特定的具体实例说明本发明的实施方式，本领域技术人员可由本说明书所揭露的内容轻易地了解本发明的其他优点与功效。本发明还可以通过另外不同的具体实施方式加以实施或应用，本说明书中的各项细节也可以基于不同观点与应用，在没有背离本发明的精神下进行各种修饰或改变。需要说明的是，以下实施例中所提供的图示仅以示意方式说明本发明的基本构想，在不冲突的情况下，以下实施例及实施例中的特征可以相互组合。

其中，附图仅用于示例性说明，表示的仅是示意图，而非实物图，不能理解为对本发明的限制；为了更好地说明本发明的实施例，附图某些部件会有省略、放大或缩小，并不代表实际产品的尺寸；对本领域技术人员来说，附图中某些公知结构及其说明可能省略是可以理解的。

本发明实施例的附图中相同或相似的标号对应相同或相似的部件；在本发明的描述中，需要理解的是，若有术语“上”、“下”、“左”、“右”、“前”、“后”等指示的方位或位置关系为基于附图所示的方位或位置关系，仅是为了便于描述本发明和简化描述，而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作，因此附图中描述位置关系的用语仅用于示例性说明，不能理解为对本发明的限制，对于本领域的普通技术人员而言，可以根据具体情况理解上述术语的具体含义。

一种联合多方数据的无监督灵活就业人员公积金缴存推荐方法包括以下步骤：