



(12)发明专利

(10)授权公告号 CN 109165664 B

(45)授权公告日 2020.09.22

(21)申请号 201810722774.3

G06N 3/08(2006.01)

(22)申请日 2018.07.04

(56)对比文件

(65)同一申请的已公布的文献号

申请公布号 CN 109165664 A

CN 107945118 A,2018.04.20

CN 107133934 A,2017.09.05

AU 2017101166 A4,2017.11.02

(43)申请公布日 2019.01.08

CN 107945140 A,2018.04.20

(73)专利权人 华南理工大学

CN 106952239 A,2017.07.14

地址 510640 广东省广州市天河区五山路
381号

KR 20170137350 A,2017.12.13

Jinsung Yoon et al..GAIN: Missing

(72)发明人 赵跃龙 王禹

Data Imputation using Generative

(74)专利代理机构 广州市华学知识产权代理有
限公司 44245

Adversarial Nets.《Proceedings of the 35
th International Conference on Machine》
.2018,

代理人 冯炳辉

审查员 张雅琳

(51)Int.Cl.

G06K 9/62(2006.01)

G06N 3/04(2006.01)

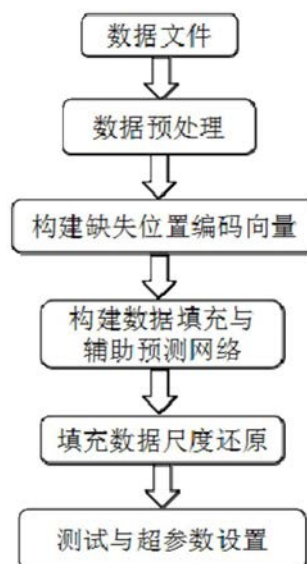
权利要求书2页 说明书5页 附图2页

(54)发明名称

一种基于生成对抗网络的属性缺失数据集
补全与预测方法

(57)摘要

本发明公开了一种基于生成对抗网络的属性缺失数据集补全与预测方法,包括步骤:1)对数据minmax归一化,同时对离散类型的属性使用one hot编码,缺失值标记为0;2)使用数据集建立关于样本的缺失位置编码向量;3)构建生成式对抗网络与辅助预测网络进行数据填充与标签的预测;4)根据属性中最大最小值还原为minmax归一化前的结果;5)通过测试选取合适的超参数;本发明充分利用数据集中数据分布信息与标签信息,能够对高维度缺失数据集进行有效的数据填充,同时在训练完成之后,该方法中包含的另一辅助预测网络能够直接队输入的属性缺失数据给出标签的预测结果,流程简捷、具有更高的预测准确率。



1. 一种基于生成对抗网络的属性缺失数据集补全与预测方法,其特征在於:首先,针对属性缺失的数据集进行数据预处理,包括minmax归一化和离散的数值变量的one hot编码转换;然后针对具有属性缺失的样本,构建缺失位置的编码向量,从而表达缺失的位置信息;接着构建缺失数据的填充网络与辅助预测网络同步完成缺失数据的填充与标签预测;在网络训练完成之后,以填充网络中生成网络的输出结果为填充的结果,根据minmax归一化时记录的列最大最小值进行尺度还原;最后,通过不断修改超参数观测其在验证集的预测结果的损失来完成超参数的设置;其包括以下步骤

1) 数据预处理;

2) 构建缺失位置编码向量,其情况是:在数据填充时,样本缺失的属性位置是一种重要的信息,在使用神经网络进行填充时,只需要对这些缺失的位置进行填充,在构建缺失位置编码向量时,对所有样本的每一列进行遍历,如果该属性缺失,记为“1”,否则记为“0”,按此流程执行,每个样本都会有一个缺失位置编码向量对应;

3) 构建缺失数据填充网络与辅助预测网络,其情况是:该网络在原始的生成式对抗网络做了如下改进:①在生成网络的输入中移除了噪声;②使用生成的数据与缺失位置向量编码来组成填充的数据;此外,辅助预测网络的引入充分的考虑了属性与标签之间的联系,在使用属性缺失数据进行预测同时,使用辅助预测网络预测标签与真实标签之间损失通过BP算法进行反馈计算更新了生成网络;联合生成式对抗网络中的损失函数与辅助预测网络中的损失函数,通过超参数控制其权重比,来决定生成的填充数据分布贴近完整数据的分布;其中,数据填充网络与辅助预测网络的结构包含生成网络、判别网络、辅助预测网络;下面对这三个网络的结构进行详细的介绍:

生成网络:输入部分由具有属性缺失的数据与其对应的缺失位置编码向量拼接构成;输入的数据是图片类型数据,使用反卷积操作得到生成的填充数据;

判别网络:输入的数据有两部分,第一部分是基于生成网络的输出得到的填充数据结果,第二部分是属性未缺失的样本数据,输出结果为0~1之间的小数,代表判别网络认为接收的输入数据是否来自属性未缺失的数据的概率;输入数据为图像类型数据,判别网络的网络结构由卷积神经网络构建;

辅助预测网络:输入与判别网络完全一致,输出则是对输入样本关于标签的预测值,当预测问题是分类问题时,采用交叉熵作为损失函数,当预测问题是回归问题时,采用L2范数或者L1范数作为损失函数;网络结构与判别网络的设置方法相同;

4) 填充数据尺度还原;

5) 测试与超参数设置。

2. 根据权利要求1所述的一种基于生成对抗网络的属性缺失数据集补全与预测方法,其特征在於:在步骤1)中,对不同数据类型进行不同的预处理,涉及的数据类型分为连续型数值与离散型数值,对于连续型数值,直接使用minmax进行归一化;对于离散型数值,转化为one hot编码之后,使用minmax归一化,对于缺失位置统一补上0;此外,将数据集根据是否有属性缺失分为两部分:具有属性缺失的数据与属性未缺失的数据。

3. 根据权利要求1所述的一种基于生成对抗网络的属性缺失数据集补全与预测方法,其特征在於:在步骤4)中,对生成的填充数据进行尺度还原,由于预处理阶段使用了minmax进行了数据归一化,根据记录的每个属性的最大值与最小值,能够还原得到最终的填充的

结果。

4.根据权利要求1所述的一种基于生成对抗网络的属性缺失数据集补全与预测方法,其特征在于:在步骤5)中,测试与超参数设置,其情况是:网络在训练的过程中,损失来源于两部分:生成式对抗网络中的损失与辅助预测网络的预测损失;这两部分损失以不同的比例 λ 组合得到综合的损失;不同的 λ 会影响模型的训练;在操作过程中,切分数据集为训练集和测试集,在训练集上选取不同尺度的 λ ,分别为0.1,0.3,0.5,0.7,0.9进行训练,同时,使用测试集进行测试,以测试集上辅助预测网络的损失最小作为超参数的选取标准。

一种基于生成对抗网络的属性缺失数据集补全与预测方法

技术领域

[0001] 本发明涉及数据预处理的技术领域,尤其是指一种基于生成对抗网络的属性缺失数据集补全与预测方法。

背景技术

[0002] 数据集属性缺失这一现象在各类数据集中广泛存在,通常是在数据采集或者传输的过程中信息丢失造成的。数据集中的样本丢失一个与多个属性会使后续建立预测、分类的模型预测精度下降。如何对这些缺失数据进行补全,并利用具有属性缺失的样本蕴含的信息来构建高精度的预测模型,是数据预处理面临的一个关键问题。

[0003] 多数统计工具采取删除缺失样本对应行、列的方式处理属性缺失的问题,或者使用该列中位数、平均数对缺失位置进行填充;这类方式虽然高效、便捷,但未能完全利用样本数据分布信息,造成计算结果的不准确。在多维数据处理的过程中,数据不同属性之间往往存在很多关联性,这些属性之间的关联性可以为数据的填充提供更多的信息,考虑到这类关联性的数据填充方法,在对缺失值进行估计时会有更小的偏差,从而能够深度挖掘缺失样本蕴含的信息。

[0004] 在此基础上,更进一步的数据填充方法通过建模来填补缺失值。如回归填补法将缺失属性作为因变量建立回归方程实现预测,EM算法先初始化缺失值,通过E步与M步迭代来得到最终的填补结果,k邻近算法(KNN)则根据未缺失的属性计算欧式距离匹配样本集中最相似的k个样本,通过加权平均得到填补结果。这些算法往往在数据量足够的情况下,取得比均值、中位数更准确的填补结果,然后通常也存在一些问题:回归填补法中,需要各属性之间有显著地线性关系,而基于EM算法的填充方法,计算复杂度高,并且容易陷入局部最优;基于k近邻的填充方法实现简单,但在面对大数据量时,计算量大复杂度极高导致计算困难。

[0005] 此外,数据填充的主要目的是为了提供更多完整的数据以供后续的建模预测。以上方法中未涉及到建模的过程,填充的数据往往和预测的标签往往会存在一些关联,将预测模型与填充方法结合起来能够使得填充得到的数据能起到更好的预测效果。针对传统数据填充方法处理高维度数据时存在计算复杂度高,未能充分挖掘标签信息以修正填充结果这两个问题;本发明将基于生成式对抗网络学习数据分布进行数据填充,同时建立一个辅助预测网络充分挖掘数据与标签之间的关联,使得其互信息达到最大。

发明内容

[0006] 本发明的目的在于克服现有技术的不足,提出了一种基于生成对抗网络的属性缺失数据集补全与预测方法,充分利用数据集中数据分布信息与标签信息,能够对高维度缺失数据集进行有效的数据填充,同时在训练完成之后,该方法中包含的另一辅助预测网络能够直接对输入的属性缺失数据给出标签的预测结果,流程简捷、具有更高的预测准确率。

[0007] 为实现以上目的,本发明所提供的技术方案为:一种基于生成对抗网络的属性缺

失数据集补全与预测方法,首先,针对属性缺失的数据集进行数据预处理,主要包括minmax归一化和离散的数值变量的one hot编码转换;然后针对具有属性缺失的样本,构建缺失位置的编码向量,从而表达缺失的位置信息;接着构建缺失数据的填充网络与辅助预测网络同步完成缺失数据的填充与标签预测;在网络训练完成之后,以填充网络中生成网络的输出结果为填充的结果,根据minmax归一化时记录的列最大最小值进行尺度还原;最后,通过不断修改超参数观测其在验证集的预测结果的损失来完成超参数的设置;其包括以下步骤

[0008] 1) 数据预处理;

[0009] 2) 构建缺失位置编码向量;

[0010] 3) 构建缺失数据填充网络与辅助预测网络;

[0011] 4) 填充数据尺度还原;

[0012] 5) 测试与超参数设置。

[0013] 在步骤1)中,对不同数据类型进行不同的预处理,涉及的主要数据类型分为连续型数值与离散型数值,对于连续型数值,直接使用minmax进行归一化;对于离散型数值,转化为one hot编码之后,使用minmax归一化,对于缺失位置统一补上0;此外,将数据集是否划分为两部分:具有属性缺失的数据与属性未缺失的数据。

[0014] 在步骤2)中,构建缺失位置编码向量,其情况是:在数据填充时,样本缺失的属性位置也是一种重要的信息,在使用神经网络进行填充时,只需要对这些缺失的位置进行填充,在构建缺失位置编码向量时,对所有样本的每一列进行遍历,如果该属性缺失,记为“1”,否则记为“0”,按此流程执行,每个样本都会有一个缺失位置编码向量对应。

[0015] 在步骤3)中,构建缺失数据填充网络与辅助预测网络,其情况是:该网络在原始的生成式对抗网络做了如下改进:①在生成网络的输入中移除了随机采样得到的噪声;②使用生成的数据与缺失位置向量编码来组成填充的数据;此外,辅助预测网络的引入更充分的考虑了属性与标签之间的联系,在使用属性缺失数据进行预测同时,使用辅助预测网络预测标签与真实标签之间损失通过BP算法进行反馈计算更新了生成网络,从而使得生成的填充数据在构建预测模型时具有更好的效果;联合生成式对抗网络中的损失函数与辅助预测网络中的损失函数,通过超参数控制其权重比,来决定生成的填充数据分布更贴近完整数据的分布或者是能够使得预测模型预测更准确;其中,数据填充网络与辅助预测网络的结构包含生成网络、判别网络、辅助预测网络;下面对这三个网络的结构进行详细的介绍:

[0016] 生成网络:输入部分由具有属性缺失的数据与其对应的缺失位置编码向量拼接构成;根据数据的结构不同,隐藏层能够使用全连接层或者反卷积层来构成,尤其在输入的数据是图片类型数据时,使用反卷积操作得到生成的填充数据;这里假定输入的数据记为I,是100维的向量,因而对应的缺失位置编码向量记为E,也是100维的,经拼接得到的输入向量维度为200;隐藏层由全连接层构成,激活函数使用relu;最终的输出层,具有100个输出单元,记为O,输出层的激活函数采用sigmoid;填充的数据最终由由 $I \cdot (1-E) + O \cdot E$ 构成;

[0017] 判别网络:输入的数据有两部分,第一部分是基于生成网络的输出得到的填充数据结果,第二部分是属性未缺失的样本数据,输出结果为0~1之间的小数,代表判别网络认为接收的输入数据是否来自属性未缺失的数据的概率;根据输入数据类型的不同,网络结构的设置也不同,在输入数据为图像类型数据时,由卷积神经网络构建;这里假定输入数据是100维向量,那么隐藏层能够选择由全连接层构成,激活函数设置为relu;输出层仅包含

一个单元,激活函数选择为sigmoid,表征概率;

[0018] 辅助预测网络:输入与判别网络完全一致,输出则是对输入样本关于标签的预测值,当预测问题是分类问题时,采用交叉熵作为损失函数,当预测问题是回归问题时,采用L2范数或者L1范数作为损失函数;网络结构与判别网络的设置方法相同;这里假定输入数据是100维向量,那么隐藏层能够选择由全连接层构成,激活函数设置为relu;输出层仅包含一个单元,激活函数按上述方式设置。

[0019] 在步骤4)中,对生成的填充数据进行尺度还原,由于预处理阶段使用了minmax进行了数据归一化,根据记录的每个属性的最大值与最小值,能够还原得到最终的填充的结果。

[0020] 在步骤5)中,测试与超参数设置,其情况是:网络在训练的过程中,损失来源于两部分:生成式对抗网络中的损失与辅助预测网络的预测损失;这两部分损失以不同的比例 λ 组合得到综合的损失;不同的 λ 会影响模型的训练;在操作过程中,切分数据集为训练集和测试集,在训练集上选取不同尺度的 λ ,分别为0.1,0.3,0.5,0.7,0.9进行训练,同时,使用测试集进行测试,以测试集上辅助预测网络的损失最小作为超参数的选取标准。

[0021] 本发明与现有技术相比,具有如下优点与有益效果:

[0022] 1、传统的填充方法如中位数、均值填充等,方法简单,填充效果不够好,而基于KNN,EM的方法往往时间复杂度大,在处理高维度数据集时,时间复杂度极大,甚至出现无法处理的情况。而生成式对抗网络在高维度数据的分布学习上有着极好的效果,因而可以解决高维度数据集带来的麻烦;另外通常未有属性缺失的样本与具有属性的缺失样本是服从同一分布的,让填充后的数据从分布上逼近没有属性缺失的数据集能够使得填充的结果不会偏离数据分布,给预测模型带来负面影响。

[0023] 2、传统的填充方法并未考虑到填充后的数据对后续建立预测模型的预测结果的影响,其步骤通常是先对缺失数据进行填充得到完成的数据,再利用填充后的数据建立预测模型,因而在不能使用预测的效果去指导数据的填充。本发明通过引入辅助预测网络计算每次填充的数据预测的值与真实标签之间的损失进行反向传播指导生成网络的数据填充,从而能够观测到填充的数据在预测模型上表现好坏选择出预测效果,结合判别网络的损失限制填充的数据与真实数据分布的差异,达到在有较好填充效果的同时具有好的预测结果。此外在完成训练之后,得到的是一个端到端的网络,在输入数据之后,能够直接得到辅助预测网络的预测结果。

附图说明

[0024] 图1为缺失数据填充与预测的流程图。

[0025] 图2为填充数据的生成式对抗网络与预测网络数据流图。

具体实施方式

[0026] 下面结合具体实施例对本发明作进一步说明。

[0027] 如图1所示,本实例所提供的基于生成对抗网络的属性缺失数据集补全与预测方法,其具体情况如下:

[0028] 1) 数据预处理:不同属性的数据类型不同,对应的处理方式也不同。涉及的主要数

据类型分为连续型数值与离散型数值,对于连续型数值,直接使用minmax进行归一化;对于离散型数值,转化为one hot编码之后,使用minmax归一化,对于缺失位置统一补上0。此外将数据集分为两部分:具有属性缺失的数据与属性未缺失的数据。

[0029] 2) 构建缺失位置编码向量:在数据填充时,样本缺失的属性位置也是一种重要的信息,在使用神经网络进行填充时,只需要对这些缺失的位置进行填充。在构建缺失位置编码向量时,对所有样本的每一列进行遍历,如果该属性缺失,记为“1”,否则记为“0”。按此流程执行,每个样本都会有一个缺失位置编码向量对应。

[0030] 3) 构建缺失数据填充网络与辅助预测网络:本发明提出了一种基于生成式对抗网络并联合辅助预测网络来进行数据填充同时能够进行预测的综合网络。该网络在原始的生成式对抗网络做了如下改进:①在生成网络的输入中移除了采样得到的噪声;②使用生成的数据与缺失位置向量编码来组成填充的数据。此外辅助预测网络的引入更充分的考虑了属性与标签之间的联系,在使用属性缺失数据进行预测同时,使用辅助预测网络预测标签与真实标签之间损失通过BP算法进行反馈计算更新了生成网络,从而使得生成的填充数据在构建预测模型时具有更好的效果。联合生成式对抗网络中的损失函数与辅助预测网络中的损失函数,通过超参数控制其权重比,来决定生成的填充数据分布更贴近完整数据的分布或者是能够使得预测模型预测更准确。图2是本发明中最重要的数据填充网络与辅助预测网络的结构图,包含生成网络、判别网络、辅助预测网络;下面对这三个网络的结构进行详细的介绍:

[0031] 生成网络:输入部分由具有属性缺失的数据与其对应的缺失位置编码向量拼接构成。根据数据的结构不同,隐藏层层可以使用全连接层或者反卷积层来构成,尤其在输入的数据是图片类型数据时,使用反卷积层得到生成的填充数据。这里假定输入的数据(记为I)是100维的向量,因而对应的缺失位置编码向量(记为E)也是100维的,经拼接得到的输入向量维度为200;隐藏层由全连接层构成,激活函数使用relu;最终的输出层,具有100个输出单元(记为O),输出层的激活函数采用sigmoid。填充的数据最终由由 $I \cdot (1-E) + O \cdot E$ 构成。

[0032] 判别网络:输入的数据有两部分,第一部分是基于生成网络的输出得到的填充数据结果,第二部分是属性未缺失的样本数据,输出结果为0~1之间的小数,代表判别网络认为接收的输入数据是否来自属性未缺失的数据的概率。根据输入数据类型不同,网络结构的设置也不同,在输入数据为图像类型数据时,可由卷积神经网络构建。这里假定输入数据是100维向量,那么隐藏层可选择为全连接层构成,激活函数设置为relu;输出层仅包含一个单元,激活函数选择为sigmoid,表征概率。

[0033] 辅助预测网络:输入与判别网络完全一致,输出则是对输入样本关于标签的预测值,当预测问题是分类问题时,采用交叉熵作为损失函数,当预测问题是回归问题时,采用L2范数或者L1范数作为损失函数。网络结构与判别网络的设置方法相同。这里假定输入数据是100维向量,那么隐藏层可选择为全连接层构成,激活函数设置为relu;输出层仅包含一个单元,激活函数按上述方式设置。

[0034] 4) 填充数据尺度还原:由于预处理阶段使用了minmax进行了数据归一化,根据记录的每个属性的最大值与最小值,可还原得到最终的填充的结果。

[0035] 5) 测试与超参数设置:网络在训练的过程中,损失来源于两部分由生成式对抗网络中的损失与辅助预测网络的预测损失;这两部分损失以不同的比例 λ 组合得到综合的损

失。不同的 λ 会影响模型的训练。在操作过程中,切分数据集为训练集和测试集,在训练集上选取不同尺度的 λ ,分别为0.1,0.3,0.5,0.7,0.9进行训练,同时,使用测试集进行测试,以测试集上辅助预测网络的损失最小作为超参数的选取标准。

[0036] 以上所述实施例子只为本发明之较佳实施例子,并非以此限制本发明的实施范围,故凡依本发明之形状、原理所作的变化,均应涵盖在本发明的保护范围内。

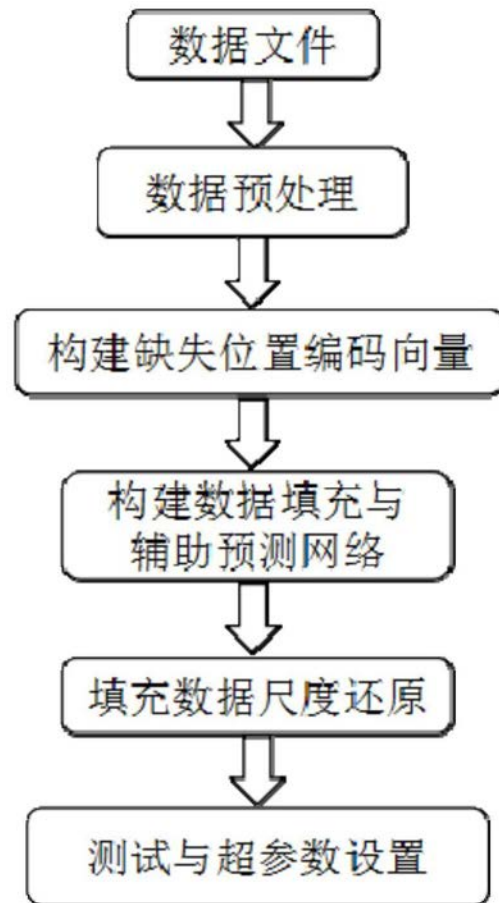


图1

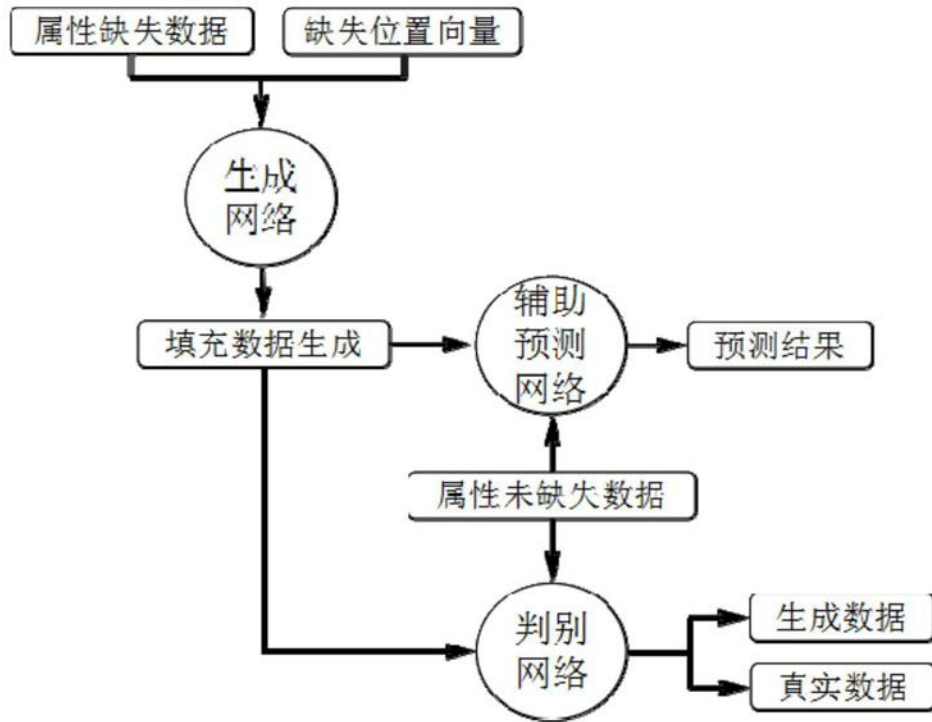


图2