

基于深度学习的不完整大数据填充算法

卜范玉^{1,2}, 陈志奎², 张清辰²

(1 内蒙古财经大学 职业学院, 内蒙古 呼和浩特 010010; 2. 大连理工大学 软件学院, 辽宁 大连 116620)

摘 要: 提出一种基于深度学习的完整大数据填充算法, 算法首先以自动编码器为基础建立填充自动编码器, 在此基础上, 构建深度填充网络模型, 分析不完整大数据的深度特征并根据逐层训练思想和反向传播算法计算网络参数, 最后利用深度填充网络来还原不完整大数据, 对缺失值进行填充. 实验表明, 提出的算法能够有效提高不完整大数据的填充精度.

关键词: 深度学习; 缺失数据填充; 自动编码器; 大数据

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2014)11-0173-04

Incomplete Big Data Imputation Algorithm Based on Deep Learning

BU Fan-yu^{1,2}, CHEN Zhi-kui², ZHANG Qing-chen²

(1 College of Vocation, Inner Mongolia University of Finance and Economics, Hohhot 010010, China;

2 School of Software Technology, Dalian University of Technology, Dalian 116620, China)

Abstract: This paper presents an imputation algorithm based on learning for incomplete big data. The proposed algorithm establishes a novel auto-encoder, called imputation auto-encoder, and then builds a deep imputation network model to analyze the deep features of incomplete big data and to calculate network parameters based on drill training ideas and back-propagation algorithm. Finally, the deep imputation network is used to impute the missing values. Experimental results show that the proposed algorithm can effectively improve the imputation accuracy for incomplete big data.

Key words: deep learning; missing data imputation; auto-encoder; big data

1 引言

随着物联网、社交网络及电子商务技术的兴起和发展, 数据正以前所未有的速度增长, 大数据的研究和应用时代已经到来^[1]. 在大数据的采集和传输过程中, 每个环节都可能出现故障, 导致很多大数据集中存在着大量的缺失. 大数据的不完整性为大数据的分析和处理带来巨大的挑战. 因此, 不完整数据的填充对大数据的分析和处理具有重要的意义.

近些年, 国内外研究人员提出了很多不完整数据填充方法. 包括基于最大期望的数据填充算法^[2]、基于模糊聚类的填充算法^[3]以及基于最近邻对象的填充算法等^[4-5]. 这些算法在填充小规模数据集方面

取得了显著的效果. 然而在填充不完整大数据方面, 精度急剧下降. 这是由于大数据存在着丰富的信息维度, 而传统的数据填充算法不能体现大数据的深度特征.

针对这个问题, 本文提出一种基于深度学习的不完整大数据填充算法.

2 填充自动编码器

设整个数据对象集合为 O , 且 $O = \{o_1, o_2, \dots, o_k\}$, 集合 O 中每个对象由 m 个属性 $\{a_1, a_2, \dots, a_m\}$, 即 $A = \{a_1, a_2, \dots, a_m\}$ 为属性集合. 算法首先将整个数据集划分为两个集合 C 和 U , 即 $O = C \cup I$, 其中 $C = \{c_1, c_2, \dots, c_i\}$, $I = \{i_1, i_2, \dots, i_p\}$, $C \cap I =$

收稿日期: 2014-02-03; 修回日期: 2014-03-29

基金项目: 国家重点自然科学基金(U1301253); 辽宁省自然科学基金(201202032)

Φ 且 $t+p=k$, C 为完整数据集, 即 C 中所有数据对象的属性值都是完整的; I 为非完整数据集, 即 I 中所有数据对象都存在缺失属性值.

本文构建的深度填充网络以填充自动编码器为基础模块, 从完整数据子集中随机采取一部分数据对象作为实例训练填充自动编码器的网络参数. 在构造填充自动编码器过程中, 使用选中的数据对象模拟缺失数据对象, 随机地将每个实例数据对象的部分属性值置 0, 模拟不完整对象作为填充自动编码器的输入, 通过最小化重构数据与实例原型来训练网络参数. 实例自动编码器如图 1 所示.

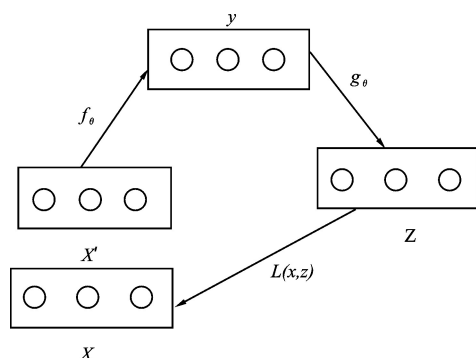


图 1 填充自动编码器结构

选定 n 个完整数据对象构成实例集合 $I = \{x_i | i = 1, 2, \dots, n\}$; 随即将每个实例对象的部分属性值置 0, 构成训练数据集 $D = \{x_i'\}$; 重构数据集为 $Z = \{z_i | i = 1, 2, \dots, n\}$. 使用训练集合 D 中的每个实例对象作为输入, 隐藏层表达式为 $y = f_\theta(x') = s(Wx' + b)$, 重构 $z = g_\theta(y) = s(W'y + b')$. 通过最小化平均重构误差来训练模型参数, 使得重构数据对象尽力逼近实例对象.

$$\begin{aligned} \theta^*, \theta'^* &= \operatorname{argmin}_{\theta, \theta'} \frac{1}{n} \sum_{i=1}^n L(x_i, z_i) \\ &= \operatorname{argmin}_{\theta, \theta'} \frac{1}{n} \sum_{i=1}^n L(x_i, g_\theta(f_\theta(x_i'))) \end{aligned} \quad (1)$$

与自动编码器一致, $\theta = \{W, b\}$ 和 $\theta' = \{W', b'\}$ 为网络参数, W 和 W' 分别是大小为 $d' \times d$ 和 $d \times d'$ 的权值矩阵并限定 $W' = W^T$, b 和 b' 是偏置向量. L 为损失函数, 本文采用传统的平方差函数 $L(x, z) = \|x - z\|^2$ 作为损失函数.

根据随机梯度下降算法, 每当从数据集 I 中选择一个实例进行训练, 填充自动编码器首先随机地选择该实例的部分属性, 将其属性值置 0, 得到一对数据 x 和 x' , 然后通过如下公式对自动编码器的权

值进行一次更新. 如此更新网络参数, 直到整个网络趋于稳定.

$$W = W - \eta \left(\frac{\partial L(X, Z)}{\partial W} + \lambda W \right) \quad (2)$$

$$b = b - \eta \left(\frac{\partial L(X, Z)}{\partial b} + \lambda b \right) \quad (3)$$

式中, η 为学习速率, λ 是权重衰减因子, 防止过度拟合

3 深度填充网络与数据填充

本文以填充自动编码器为基础模块, 构建三层网络模型. 每一层网络输出都将作为上一层网络的输入, 最上层作为提取的特征输出. 训练过程分为预训练和微调两个阶段. 首先自下而上地进行逐层训练获得网络初始化参数, 最终通过反向传播算法对全局参数进行微调.

为了获取网络逐层训练监督对象, 首先利用实例数据作为输入构建叠加自动编码器, 获得实例数据的两层特征. 叠加自动编码器结构如图 2 所示.

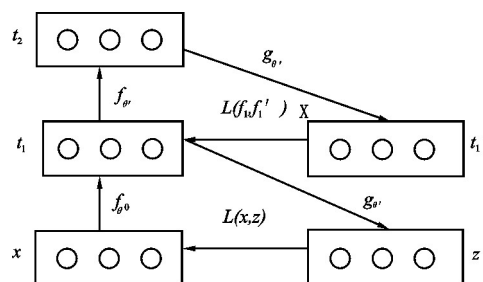


图 2 两层叠加自动编码器结构

本文以未经处理的原始实例数据 x 作为网络输入, 在最下层可获得第一层特征 $t_1 = f_\theta(x)$, 把特征 t_1 作为上一层网络的输入, 获得第二层特征 $t_2 = f_{\theta'}(t_1)$, 该训练过程是局部的, 即第二层网络更新本层的网络权重, 对下层网络没有影响. 通过这种方式可以初始化叠层网络参数, 最后通过反向传播算法对网络全局参数进行微调. 如此能够获得对应于原始数据实例的两层特征 t_1 和 t_2 .

本文构建的三层深度填充网络结构如图 3 所示.

深度实例网络以 x 和 t_1 和 t_2 将作为监督数据, 采用逐层向上的训练方法来初始化每层网络参数. 首先对数据实例 x 进行加噪处理, 随即选取部分属性, 将其属性值置 0, 获得不完整数据模拟实例 x' ; 输入 x' 获得不完整数据的第一层特征 $t_1' = f_\theta(x')$ 及 $z = g_\theta(t_1')$; 然后根据叠加自动编码器在 f_θ 的基础上学习实例 x 的第一层特征 t_1 , 并以 t_1 作为监

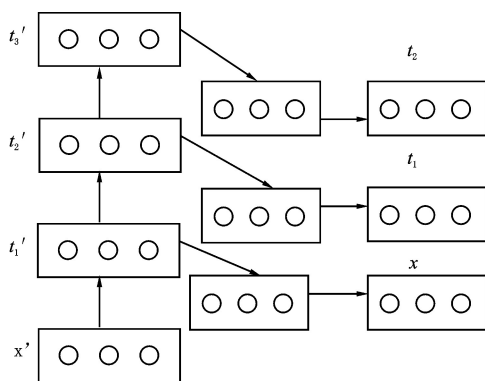


图3 深度填充网络结构

督数据,以 t_1' 为输入,进一步训练深度实例网络上一层的自动编码器,获得 x' 的第二层特征 $t_2' = f_{\theta^1}(t_1')$ 及 $z_1 = g_{\theta^1}(t_2')$,同时根据叠加自动编码器在 f_{θ^1} 的基础上学习实例 x 的第一层特征 t_2 。最后,以 t_2 作为监督数据,以 t_2' 为输入,进一步训练深度实例网络上一层的自动编码器,获得 x' 的第三层特征 $t_3' = f_{\theta^2}(t_2')$ 及 $z_2 = g_{\theta^2}(t_3')$ 。深度实例网络逐层对实例特征进行逼近,在每一层都会降低缺失数据的影响,达到网络最顶层时即可对大数据本质进行深度抽象和特征提取。

从数据集 I 中逐一取出实例对深度实例网络进行训练,每训练一次,对网络参数进行一次更新,直到整个网络趋于稳定,获得最终的网路参数 θ^1, θ^1 和 θ^2 。

在获得网络参数之后,本文首先抽取不完整数据集中每个数据对象的深度特征。对于不完整数据对象 x_c 而言,首先将其缺失属性的属性值置0,构成 x_c' ,以 x_c' 为输入,利用式(4)提取其深度特征:

$$t = f_{\theta^2}(f_{\theta^1}(f_{\theta^1}(x_c))) \quad (4)$$

最后利用式(5)对不完整数据对象进行还原,获得缺失数据的填充值。

$$\tilde{x} = g_{\theta^2}(g_{\theta^1}(g_{\theta^1}(t))) \quad (5)$$

4 实验分析

为了验证本文提出的算法(DLDBI)的有效性,将本文提出的算法和两种填充算法 FIMUS 和 DMI 进行对比。本文采用的数据集采自数字家庭与无线传感网络实验室,数据集总量达到 10 G 每个数据对象包含 650 个数值属性。我们首先人为地从数据集中删除一部分数据,模拟不完整数据集,在填充完成之后,将填充值与真实值进行比较,得到算法的填充精度。

本文人工制造两种缺失值,单模式缺失和多模

式缺失。在单模式缺失中,每个数据对象只允许含有一个缺失值,多模式缺失则允许每个数据对象含有多个缺失值。本文分别从数据集中选择 1%, 3%, 5% 和 10% 的数据对象并删除这些数据对象的部分属性值,模拟缺失数据。

本文使用两个标准来衡量算法的填充精度^[7]。第一个标准被称 d_2 标准,该标准用于衡量填充值与真实值的匹配程度,计算公式如下:

$$d_2 = 1 - \left[\frac{\sum_{i=1}^n (e_i - r_i)^2}{\sum_{i=1}^n (|e_i - E| + |r_i - R|)^2} \right] \quad (6)$$

第二个标准是 RMSE,用于衡量填充值与真实值之间的平均误差,计算公式如下:

$$\text{RMSE} = \left(\frac{1}{n} \sum_{i=1}^n |r_i - e_i|^2 \right)^{1/2} \quad (7)$$

在式(6)和式(7)中, n 代表缺失数值的数目, r_i 代表第 i 个缺失值的真实值, e_i 代表第 i 个缺失值的填充值, R 代表 r_i 的平均值, E 代表 e_i 的平均值, $i = 1, 2, \dots, n$ 。更具两个标准的定义,对于一个算法来讲, d_2 的值越大,说明该算法的填充精度越高。相反, RMSE 的值越小,则算法的填充精度越高。填充结果如表 1 和表 2 所示。

表1 填充精度指标: d_2

组合		算法		
缺失率/%	缺失模式	DLDBI	FIMUS	DMI
1	single	0.842	0.741	0.732
	multiple	0.819	0.727	0.721
3	single	0.891	0.729	0.712
	multiple	0.849	0.708	0.697
5	single	0.855	0.692	0.684
	multiple	0.840	0.681	0.672
10	single	0.865	0.657	0.643
	multiple	0.842	0.635	0.616

表2 填充精度指标: RMSE

组合		算法		
缺失率/%	缺失模式	DLDBI	FIMUS	DMI
1	single	0.151	0.261	0.289
	multiple	0.174	0.267	0.293
3	single	0.118	0.272	0.302
	multiple	0.143	0.295	0.319
5	single	0.136	0.293	0.309
	multiple	0.156	0.317	0.328
10	single	0.161	0.316	0.336
	multiple	0.176	0.328	0.362

从表 1 可以看出,对于任何一种缺失组合,本文提出的算法所得到的 d_2 都明显高于其他两种算法.除此之外,随着数据缺失率的增大,算法 FIMUS 和 DMI 所得到的 d_2 都在下降,即这两种算法的填充精度随着数据缺失率的增大而降低.而本文提出的算法的填充精度一直保持在一个很高的水平之上.因此,就 d_2 而言,本文提出的算法的填充精度明显高于 FIMUS 和 DMI.

从表 2 可以看出,对于任何一种缺失组合,本文提出的算法所得到的 RMSE 都明显低于其他两种算法.随着数据缺失率的增大,算法 FIMUS 和 DMI 所得到的 RMSE 不断升高,即这两种算法的填充精度随着数据缺失率的增大而降低.而本文提出的算法的得到的 RMSE 值一直低于 0.2.因此,就 RMSE 而言,本文提出的算法的填充精度明显高于 FIMUS 和 DMI.

对于任意的缺失组合,本文随即的选取不同的对象作为训练数据,运行算法 20 次,统计算法得到 d_2 值和 RMSE 值的平均值,如图 4 和图 5 所示.

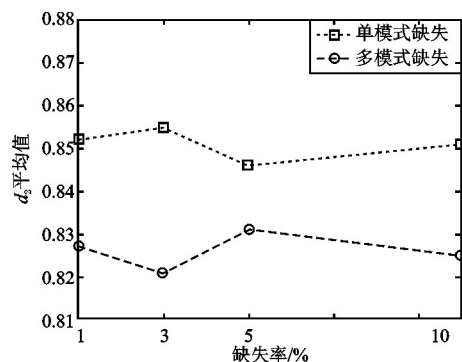


图 4 d_2 平均值

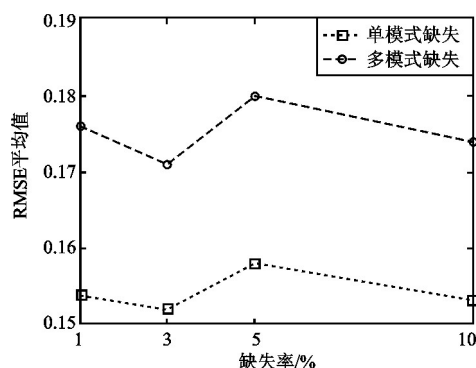


图 5 RMSE 平均值

从图 3 和图 4 可以看出,本文提出的算法填充精度相对比较稳定.具体的说,当数据缺失率在 1%到 10%之间, d_2 值能够稳定的保持在 0.8 以上, RMSE 值稳定在 0.15 到 0.2 之间.此外,对于任意一种缺失率而

言,单缺失模式的填充精度明显高于多缺失模型的填充精度,这是因为多填充模式缺失数据大,对特征提取和还原造成的干扰高于单缺失模式.

5 结束语

本文提出一种基于深度学习的不完整大数据填充算法,算法针对大数据具有丰富的信息维度,构建深度填充网络类提取大数据的深度特征,进而对缺失值进行还原.实验结果表明本文提出的算法能够有效的提高数据填充精度.在下一步工作中,探索如何提高多缺失模式下的数据填充精度.

参考文献:

- [1] Wu X, Zhu X, Wu G Q, et al. Data mining with big data[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(1): 97-107.
- [2] Schafer J L, Graham J W. Missing data: our view of the state of the art [J]. Psychological methods, 2002, 7(2): 147.
- [3] Wang X, Ao Li, Zhao hui, Jiang, et al. Missing value estimation for dna microarray gene expression data by support vector regression imputation and orthogonal coding scheme[J]. BMC Bioinformatics, 2006, 7 (1): 32-37.
- [4] Li D, Jitender Deogun, Bill Shuarrrt, et al. Towards missing data imputation: A study of fuzzy k-means clustering method [C] // Rough Sets and Current Trends in Computing. Berlin Heidelberg: Springer, 2004: 573-579.
- [5] 武森, 冯小东, 单志广. 基于不完备数据聚类的缺失数据填补方法 [J]. 计算机学报, 2012, 35 (8): 2053-2053.
- [6] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.
- [7] Rahman M G, Islam M Z. FIMUS: A framework for imputing missing values using co-appearance, correlation and similarity analysis[J]. Knowledge-Based Systems, 2014(56): 311-327.

作者简介:

卜范玉 男, (1981-), 博士研究生, 讲师. 研究方向为大数据、无线传感网络. E-mial: 623759909@qq.com.

陈志奎 男, (1968-), 博士, 博导教授. 研究方向为物联网、大数据.

张清辰 男, (1988-), 博士研究生. 研究方向为大数据、云计算.