·先进计算与数据处理· 文章编号:1000-3428(2019)09-0032-08 文献标志码:A 中图分类号:TP301.6

基于聚类的连续型数据缺失值充填方法

李国和[1a,1b],杨绍伟[1a,1b],吴卫江[1a,1b],郑艺峰[1a,1b,2a,2b]

(1. 中国石油大学(北京) a. 石油数据挖掘北京市重点实验室; b. 地球物理与信息工程学院,北京 102249;
2. 闽南师范大学 a. 数据科学与智能应用福建省高等学校重点实验室; b. 计算机学院,福建 漳州 363000)

摘 要: 在大数据应用中,多数建模方法是在完备数据集基础上进行的,但在数据采集过程或存储过程中容易出现数据缺失的现象,导致无法建模。为此,提出一种基于聚类的递归充填方法。使用同类簇的均值对不完备数据进行预填充,形成初始完备数据集,针对得到的完整数据进行聚类,并运用同类簇的均值修正初始充填值。根据充填效果误差判定充填稳定性,并进行多次递归聚类修正充填值,直到前后两次充填较为稳定或迭代次数超过阈值时停止迭代。实验结果表明,与均值充填、K 最近邻充填、聚类充填及粗糙集不完备数据分析等方法相比,该方法能够进行更为精准的充填,使得最终充填更加接近真实数据。

关键词: 缺失值; 预充填; 聚类; 递归充填; 平方误差

开放科学(资源服务)标志码(OSID):

中文引用格式: 李国和,杨绍伟,吴卫江,等. 基于聚类的连续型数据缺失值充填方法[J]. 计算机工程,2019,45(9):32-39.
英文引用格式: LI Guohe, YANG Shaowei, WU Weijiang, et al. Clustering-based missing value filling method for continuous data[J]. Computer Engineering,2019,45(9):32-39.

Clustering-based Missing Value Filling Method for Continuous Data

LI Guohe[1a,1b], YANG Shaowei[1a,1b], WU Weijiang[1a,1b], ZHENG Yifeng[1a,1b,2a,2b]

(1a. Beijing Key Lab of Petroleum Data Mining; 1b. College of Geophysics and Information Engineering, China University of Petroleum(Beijing), Beijing 102249, China; 2a. Key Laboratory of Data Science and Intelligence Application; 2b. School of Computer Sciences, Minnan Normal University, Zhangzhou, Fujian 363000, China)

【Abstract】 In big data applications, most modeling methods are based on a complete data set, but data missing in the data acquisition process or storing process tend to result in failure to modeling. Therefore, a clustering-based recursive filling method is proposed. The incomplete data is pre-filled using the mean of the same cluster to form an initial complete data set. The complete data obtained are clustered, and the initial filling is corrected using the mean of the same cluster. The filling stability is determined according to the deviation of filling results, and the filling value is corrected through multiple times of recursive clustering until the last two times of filling is stable or the number of iterations exceeds the threshold. Experimental results show that compared with the methods of mean filling, K nearest neighbor filling, cluster filling and incomplete data analysis for rough sets, the method can implement more precise filling, making the final filling more close to real data.

【Key words】 missing value; prefilling; clustering; recursive filling; square error
DOI: 10.19678/j.issn.1000-3428.0053331

0 概述

随着信息技术的发展与应用,当今社会已进入大数据时代,人们为获取潜在的有价值的数据信息,需要通过数据进行建模。真实数据偶尔存在缺失值,即数据不完备,将导致无法建模,因此,缺失值处理成为数据预处理的重要研究方向之一[1]。

针对连续型数据的缺失值,国内外学者提出许多处理方法,常见的充填方法是通过已知数据填补未知数据来达到完备数据。主流充填方法主要有均

基金项目: 国家自然科学基金(61701213);国家油气重点专项子课题(G-5800-08-ZS-WX);中国石油大学(北京)克拉玛依校区科研启动基金(RCYJ2016B-03-001);福建省教育厅中青年基金(JA15300)。
作者简介: 李国和(1965—),男,教授、博士、博士生导师,主研方向为人工智能、大数据、计算机图形技术;杨绍伟,硕士研究生;吴卫江,副教授、博士研究生;郑艺峰,博士研究生。
收稿日期: 2018-12-06 **修回日期:** 2019-02-22 **E-mail:** 787866446@qq.com

值充填算法、K 最近邻(K-Nearest Neighbor, KNN)充填算法、聚类充填算法[2]及粗糙集不完备数据分析(Rough Set Incomplete Data Analysis, ROUSTIDA)算法[3]。上述方法简单有效,应用比较广泛,但存在样本分布偏差较大,近邻点界定模糊以及求解过程较慢等问题。

针对上述方法存在的不足,本文提出一种基于聚类的递归充填方法,将均值充填与聚类相结合,以得到更加接近原始数据的完备数据集。

1　充填方法

目前主流充填方法主要有以下 4 种:

1)平均值充填算法。该算法数据属性分为连续型和离散型。对于离散型属性缺失值,在该属性的非缺失值样本中取频次最多的值进行缺失值充填,即选取其中的众数;对于连续型属性缺失值,用该属性所有非缺失值的均值进行缺失值充填。该算法用数据集的非缺失值推测缺失值,将出现次数最多或取值分布中心点作为充填值[4],充填过程快速高效。但所有充填值都集中在众数或均值,充填后数据集在分布上容易形成尖峰,导致样本之间的差异,这改变了数据原有分布,而且对小样本集的缺失充填会造成更大的分布偏差[5]。

2)K 最近邻充填算法。该算法通过暂时忽略缺失值的属性,形成完备数据集。根据最相似的 k 个完备样本,恢复所有属性,利用 k 个样本中的属性已知值充填缺失值。对离散型属性缺失值,用 k 个样本的众数充填。对连续型属性缺失值用已知样本均值充填。在寻找 k 个近邻的过程中,针对连续型数据,使用样本距离为相似度,而针对离散型数据,使用相同属性取值一致的数据作为相似度[6]。在此基础上给出可利用含有缺失值的相似样本进行充填的有序最近邻居算法(SKNN)[7]、基于马氏距离最近邻查找算法(MKNN)[8]。通过近邻点充填缺失值,相比于均值充填更为准确,使用合理的排序算法也可将充填时计算次数控制在可承受范围内。但该类方法难于确定 k 值,近邻点界定比较模糊。在缺失数据比例过高时,需要针对每一条不完备样本进行一次查找最近邻的操作,计算量随缺失数据比例呈线性增长,存在部分数据无法得到完备数据集的可能[9]。同时,对于有监督的分类问题,需要对样本进行按标签划分,加大了在同类数据中找不到完备近邻样本的概率。

3)聚类充填算法。与 KNN 算法类似,该算法可以使用同类簇点作为近邻点来填补缺失值,但是聚类算法通常无法确定缺失样本之间的距离,若缺失较多则常见的聚类方法不具备可用性。针对该问题,文献[10]提出了可以忽略不完整特征而衡量样本相似性的、基于部分距离的不完整数据 CFS 聚类(PDPCM)算法,该算法可无需舍弃缺失特征进行聚类,能够有效确定类簇并进行充填。但是这种方法针对缺失特征进行忽略处理,会形成样本近邻点选择完全依赖于完备特征的现象。

4)ROUSTIDA 算法通过扩充差异矩阵来判断对象之间的相似程度。建立一个维度为 $N \times N$ 的差异矩阵,其中,N 为数据集中的样本数,矩阵第 i、j 位置代表第 i 个样本和第 j 个样本之间的相似程度,对角线设置为 -1,以保证不被认为是近邻点。通过一次扫描,对缺失的样本使用近似样本的权值进行充填[11]。文献[3]针对差异矩阵,选择出样本的近似样本集合,使用该集合众数填补缺失值,确定了近邻点数目,进一步提高了精度。该类方法在处理缺失值时通过属性取值求出一个相似度值,选择近似样本为数据全体,无需单独挑选出完备数据部分。同时该方法对监督学习中有标签数据具备处理能力。但该方法只能处理离散型数据,在处理连续型数据时,需要对属性逐列进行离散化,若样本过多则求解过程较慢[12]。

2　基本概念

2.1　不完备连续型属性的决策信息系统

定义 1(信息系统)　$K = <U, R, V, f>$ 为信息系统,其中,$U = \{u_i \mid i = 1, 2, \cdots, |U|\}$ 为论域(即对象集合),$R = \{r_j \mid j = 1, 2, \cdots, |R|\}$ 为属性集(即属性集合)。对于 $\forall u_i \in U, \forall r_j \in R$,存在 $v_{ij} \in V_j$,即 $r_j: U \times \{r_j\} \rightarrow V_j$,$r_j(u_i) = v_{ij}$,称为对象 u_i 在 r_j 上的投影。决策信息系统记为 $K = <U, C \cup D, V, f>$。$V = \bigcup_{j=1}^{|R|} V_j$,存在 $f: U \times R \rightarrow V$ 为信息函数[13]。如果 $R = C \cup D$,$C \cap D = \varphi$,则 $K = <U, C \cup D, V, f>$ 为决策信息系统(或称决策表),其中,C 和 D 分别为条件属性集和决策属性集。

定义 2(连续型属性的信息系统)　对于 $K = <U, R, V, f>$,$\forall u_i \in U, \forall r_j \in R$,存在数值 ε,使得 $r_j(u_i) = v_{ij} - \varepsilon \in V_j$,$\varepsilon$ 为该属性取值的邻域,即 V_j 是连续区间,r_j 是实型。如果不存在 ε,则属性为离散型属性。

定义 3(不完备信息系统)　对于 $K = <U, R, V, f>$,$\forall u_i \in U, \forall r_j \in R$,存在数值 ε,使得 $r_j(u_i) = * \in V_j$,即知道值的存在,但不知具体值。其中,$*$ 为缺失值。如果 V 中含有缺失值 $*$,则 K 为不完备连续型属性的决策信息系统。

定义 4(缺失属性集、缺失对象集)　对于不完备连续型属性的决策信息系统 $K = <U, R, V, f>$:

$$MAS = \{r_j \mid r_j \in R, \exists u_i \in U, r_j(u_i) = *\} \quad (1)$$

$$MOS = \{u_i \mid \forall r_j \in MAS, \exists u_i \in U, r_j(u_i) = *\} \quad (2)$$

其中,MAS 为缺失属性集,MOS 为缺失对象集,$U - MOS$ 为完备对象集,即 $K' = <U - MOS, R - MAS, V, f>$ 为 K 的完备决策信息子系统。

定义 5(信息系统的平方误差) 完备信息系统 $K_1 = <U,R,V_1,f_1>$, $K_2 = <U,R,V_2,f_2>$ 的平方误差为:

$$SQ(K_1,K_2) = \sum_{\forall v_{1ij}\in V_1, v_{2ij}\in V_2} (v_{1ij} - v_{2ij})^2 \tag{3}$$

2.2 k-means 聚类算法

定义 6(对象距离) 根据欧氏距离定义,对于完备信息系统 $K = <U,R,V,f>$, $\forall u_i,u_j \in U$ 的欧式距离为:

$$D(u_i,u_j) = \sqrt{\sum_{k\in R}(r_k(u_i) - r_k(u_j))^2} \tag{4}$$

定义 7(属性归一化) 对于完备信息系统 $K = <U,R,V,f>$, $u_i \in U$, $r_j \in R$,有:

$$normal(r_j(u_i)) = \frac{r_j(u_i) - \min\limits_{\forall u_k\in U}(r_j(u_k))}{\max\limits_{\forall u_k\in U}(r_j(u_k)) - \min\limits_{\forall u_k\in U}(r_j(u_k))} \tag{5}$$

其中, $normal(r_j(u_i))$ 为对象 u_i 的属性 r_j 的标准化值[14]。

通过属性成归一化,消除不同属性量纲和取值范围不同的影响。后续决策信息系统均指归一化后的决策信息系统,对于离散属性,无需对其进行归一化。

定义 8(线性回归的属性权重) 对于完备决策信息系统 $K = <U,C\cup\{d\},V,f>$,有:

$$\omega_c = \arg\min \sum_{\forall u\in U}\sum_{c\in C} \omega_c c(u) - \lambda \| \omega_c \|^2 \tag{6}$$

其中, d 为连续型决策值, ω_c 为线性回归模型下条件属性 C 的权重, λ 为正则化系数。

定义 9(独热编码) 独热编码(One-Hot 编码),即一位有效编码,定义为用 N 位状态寄存器编码 N 个状态,每个状态都有独立的寄存器位,且在任意时候只有一位有效。

对于完备信息系统 $K = <U,R,V,f>$,对 U 进行划分,记为 $DS_R(U)$,满足: $\forall U_i,U_j\subseteq U$, $U_i \neq U_j$, $U_i \cap U_j = \phi$, $U = \bigcup\limits_{\forall U_k\in DS_R(U)} U_k$。

对完备信息系统 $K = <U,R,V,f>$ 进行聚类,形成聚类簇 $Clus$, $Clus$ 为 U 的划分 $DS_R(U)$。对于 $clu\subseteq Clus$, $u_i \in clu$,有:

$$a(u_i) = \frac{\sum\limits_{\forall u_j\in clu - \{u_i\}} D(u_i,u_j)}{|clu| - 1} \tag{7}$$

其中, $a(u_i)$ 为对象 u_i 与同聚类 clu 中其他对象的平均距离。

$$b(u_i) = \min_{\substack{u_i\in clu,\\ \forall u_j\in U - clu - \{u_i\}}} D(u_i,u_j) \tag{8}$$

其中, $b(u_i)$ 为对象 u_i 与所有不同聚类 clu 中对象的最短距离。

$$S(u_i) = \frac{b(u_i) - a(u_i)}{\max\{b(u_i),a(u_i)\}} \tag{9}$$

其中, $S(u_i)$ 为对象 u_i 的轮廓系数,其综合了聚类的凝聚度和分离度。取值范围为 $[-1,1]$,值越大,表示聚类效果越好[15],可确定 k-means 算法中 k 的取值。

定义 10(轮廓系数) 轮廓系数为评估聚类效果的参数。

对于完备决策信息系统 $K = <U,R,V,f>$,存在划分 $DS_R(U)$,概率分布为:

$$P_R(U) = \left\{ p(U_i) = \frac{|U_i|}{|U|} \,\middle|\, \forall U_i \in DS_R(U) \right\} \tag{10}$$

对于完备决策信息系统 $K = <U,R,V,f>$,可分解为 2 个完备信息子系统,即 $K_{R1} = <U,R_1,V_{R1},f_{R1}>$ 和 $K_{R_2} < U,R_2,V_{R_2},f_{R_2}>$,且 $R_1,R_2\subseteq R$(如决策表, $R_1 = C$, $R_2 = D$),分别形成 2 个划分 $DS(R_1)$、$DS(R_2)$,对应 2 个概率分布 $P_{R_1}(U)$、$P_{R_2}(U)$。根据互信息熵的定义,可给出完备信息系统的互信息熵定义。

定义 11(完备信息系统的互信息) 对于完备决策信息系统 $K = <U,R,V,f>$,存在 $R_1,R_2\subseteq R$ 及其概率分布 $P_{R_1}(U)$, $P_{R_2}(U)$, $P_{R_1\cap R_2}(U)$, $R_1,R_2\subseteq R$ 及的互信息定义为:

$$I(R_1,R_2) = \sum_{\forall U_i\in DS_{R1}(U)}\sum_{\forall U_i\in DS_{R1}(U)} p(U_1\cap U_2)\times \mathrm{lb}\frac{p(U_1\cap U_2)}{p(U_1)p(U_2)} \tag{11}$$

其中, $I(R_1、R_2)$ 表达了 2 个属性集 R_1、R_2 间相互依赖性程度。特别对于完备决策信息系统 $K = <U,C\cup D,V,f>$,可分解为 2 个完备信息子系统,即 $K_C = <U,C,V_C,f_C>$ 和 $K_D = <U,D,V_D,f_D>$,其中 $I(C,D)$ 反映了条件属性 C 对决策属性 D 的相互关联紧密程度。

3 基于聚类的递归均值充填

本文提出基于聚类的递归均值充填(Recursive Mean Filling Based on Clustering,RMFBC)算法主要包括均值预充填、完备数据聚类与更新、稳定检测 3 个阶段,如图 1 所示,输入为不完备连续型属性信息系统 K,输出为完备连续型属性信息系统 K_1,整个过程为记录原始缺失样本,首先使用均值进行预充填,对完备信息系统进行聚类,使用同类簇中样本均值更新缺失样本,然后对新的完备系统进行聚类,再使用同类簇中样本均值更新缺失样本,反复进行,直至聚类结果较为稳定后停止。

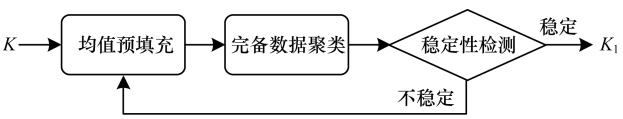

图 1 缺失值充填流程

3.1 均值预充填

均值预充填过程使用均值充填缺失样本,具体过程如算法 1 所示。

算法 1　MeanFilling 算法

```
MeanFill(K)//K = <U,R,V,f> 不完备连续型属性的
//决策信息系统
UC = U - MOS //记录完备样本
For(i = 0;i < |U|;i + + )//遍历所有样本
If(u_i ∈ MOS) //针对有缺失值样本进行处理
For(j:MAS_i) //对于该样本缺失属性
If(r_j ∈ C) //如果缺失条件属性
v_ij = ave(r_j(UC)) //完备部分均值充填
End If
If(r_j ∈ D) //如果缺失决策属性
v_ij = mod(r_j(UC)) //完备部分众数充填
End If
End For
End If
End For
ω_c = argmin Σ_{∀u∈U} Σ_{∀c∈C} ω_c c(u) - λ ‖ ω_c ‖^2 //通过完备数据
//得到每个属性的权重系数
```

在算法 1 中,ω_c 为通过线性回归模型得到的权重。为了补全数据,不考虑分类,则其可置为全 1,对决策属性重要性不造成影响。

3.2　完备数据集聚类

针对上述步骤得到的完备信息系统(数据集),采用 k-means 聚类寻找缺失样本的近邻点,具体如算法 2 所示。

算法 2　k-means 算法

```
List k-means (K,k,time)
//time 为最大迭代次数,返回所有样本的类别列表,k 为
//类别数
//K = <U,R,V,f> 完备连续型属性的决策信息系统
mk = Random(U,k) //随机选择 k 个初始质心的 list
mk_last = mk //记录上一次聚类的质心
while(true)
For(i = 0;i < |U|;i + + )
Get min(D(u_i,mk)) //计算每点最近距离的质心,归为
//该类
End For
For(i = 0;i < k;i + + )
cacul(mk)  //重新计算每个类的质心
End For
time
if(mk = = mk_last ||time = = 0) //2 次聚类结果相同或
//达到最大迭代次数
break
End If
mk_last = mk
End While
UK = {U1,U2,…,Uk} = DS_mk(U) //用以记录类别对
//样本的划分
Initial result[|U|] //用以记录样本属于的类别
For(i = 0;i < |U|;i + + )
result[i] = getindex(UK,ui) //得到样本所属的类别
//索引
End For
return result
```

之后使用同一类簇中所有点的均值或众数更新对应位置,具体如算法 3 所示。

算法 3　ClusterFilling 算法

```
ClusterFilling (K,a)
//K = <U,R,V,f> 完备连续型属性的决策信息系统
//a 为决策属性的权重系数
K1 = K //拷贝副本
For(j = 0;j < |C|;j + + ) //遍历所有条件属性
For(i = 0;i < |U|;i + + ) //遍历所有样本属性
r_j(u_i) = normal(r_j(u_i)) * ω_ci//对属性向量进行归一
//化,并乘以权重
End For
End For
onehot(D) → d //对决策属性进行独热编码化为连续
//属性
For(i = 0;i < |U|;i + + ) //遍历所有样本属性
r_d(u_i) = r_d(u_i) * a  //乘系数,增加(减少)决策属性的
//重要程度
End For
best_k = 1,S = -1,best_result = null
For(k = 2;k < 9;k + + ) //根据轮廓系数寻找最优 k 值
result = k-means (K,k,10)
S(u_0) > S? (best_k = k,S = S(u_0),best_result =
result):continue
End For
K = K1 //恢复原数据,使用聚类结果对原数据进行充填
For(i = 0;i < |U|;i + + )//遍历所有样本
For(j = 0;j < |R|;j + + )//遍历所有属性
If(r_j ∈ MAS && u_i ∈ MOS)//如果原数据该位置缺失
If(r_j ∈ C) //如果缺失条件属性
v_ij = ave(r_j(UK[result[i]]))//同类簇完备部分均值充填
End If
If(r_j ∈ D)  //如果缺失决策属性
v_ij = mod(r_j(UK[result[i]]))//同类簇完备部分众数
//充填
End If
End If
End For
End For//得到完备连续型属性的决策信息系统
```

在算法 2 中,a 为决策属性重要性的参数。为更好地寻找近邻点,将有监督的分类问题转化为无监督的聚类问题,相当于弱化了决策属性作用,因此在处理决策属性时需针对数据缺失情况分别讨论。若决策属性缺失数量较少,则应当充分利用该属性,将 a 的数值设置为较大;若决策属性缺失数量过多,则将 a 值设置为较小,避免对聚类影响过大。本文算法得到了完备的决策信息系统,同时保留本次聚类的结果,即 best_result 列表。

3.3 递归充填

在均值充填过程中,聚类 k 值并不固定,能够形成具有层次结构的聚类树。对于 k 值变化不大,分布合理的数据集,其聚类结果偏差应较小,即当前、后 2 次聚类结果稳定时缺失充填结果是稳定的。为了实现稳定的充填效果,本文算法选择多次聚类充填后,直到前后 2 次聚类结果偏差在容忍范围内或迭代次数超出预期时的数据集作为最终充填结果。使用定义 11 作为 2 次聚类充填的相似性评价,并给定阈值和迭代次数控制递归充填,具体如算法 4 所示。

算法 4 RecursiveFilling 算法

```
RecursiveFilling (K,threshold_i,time)
//time 为最大迭代次数
//threshold_i 为 2 次聚类结果相关系数阈值
//K = <U,R,V,f>完备连续型属性的决策信息系统
Initialresult //初始化全局变量 result
while(time >0)
last_result = result
ClusterFilling (K,a) //通过聚类更新 result
time
If(I(result,last_result) ≤ threshold_i) //2 次聚类结果
//相关性小于阈值
break
End If
End while  //得到最终完备连续型属性的决策信息系统
```

4 实验结果与分析

实验数据为 UCI 数据集,如表 1 所示。实验环境为 Windows10(64 位),8 GB 内存,Intel(R) Core(TM)2 CPU E7300 @2.66 GHz。

表 1 数据集信息

数据集	样本数	属性数	类别数
Hill-Valley	606	101	2
LSVT_ voice_rehabilitation	126	309	2
breast_cancer_wisconsin	198	34	2
BlogFeedback(4)	60 021	281	8
gesture_phase_dataset	9 900	50	5
Dataset for Sensorless Drive Diagnosis	58 509	49	11
Connectionist Bench	208	60	2

4.1 评价方式定义

实验采用比较平方误差方式,即将原始数据按比例随机缺失,然后使用缺失充填法进行充填,并计算所有对应的充填数值与原始数值差的平方和。若完备决策信息系统为 $K=<U,R,V,f>$,随机缺失并充填后的完备决策信息系统为 $K_1=<U,R,V_1,f_1>$,则平方误差定义如下:

$$Err_1(K_1,K)=\sum_{\substack{\forall r\in R\\ \forall u\in U}}(r(u)'-r(u))^2 \tag{12}$$

其中,$r(u)'$为 K_1 的函数映射,$r(u)$为 K 的函数映射。

为便于归一化比较,对于 K 的多种不同充填 $K_1,K_2,\cdots,K_l,K_i$ 相对误差为:

$$Err(K_i,K)=\frac{Err(K_i,K)}{\max\{Err(K_j,K)\mid j=1,2,\cdots,l\}} \tag{13}$$

由于均值充填误差较大,令 K_1 为均值充填,显然当 $Err(K_i,K)\in[0,1]\to 0$ 时,充填效果相对较好,而当 $Err(K_i,K)\in[0,1]\to 1$ 时,充填效果相对较差[18]。

4.2 方法对比及参数设置

本文分别采用均值充填、KNN 充填、PDPCM 充填和 ROUSTIDA 充填,均值充填无需参数。对于 KNN 充填,若簇内完备数据不足 k 个,则用簇完备数据充填。需要选定参数 K 值,即视为近邻点的数目。PDPCM 充填需指定聚类类别数目 Cn。ROUSTIDA 充填是针对离散化属性充填的,对连续型属性则需离散化,本文实验采用等距离分桶方式离散化[17],分桶值为 N,计算差异矩阵后用最大相似对象集的均值进行充填,最后仍然缺失的用全部数据进行均值充填。具体参数设置如表 2 所示。

表 2 具体参数说明

方法	参数名称	参数取值
均值充填方法	null	null
KNN 充填方法	k	5
PDPCM 充填方法	C_n	10
ROUSTIDA 充填方法	N	10

表 1 中参数的选取,能够确保充填具有较好的效率和效果。

4.3 超参数设置

实验涉及到 5 个超参数,分别为正则化系数 λ、决策属性权重 a 值、k-means 聚类迭代次数、递归充填相似性阈值 *threshold_i* 和递归次数。

l2 正则化系数 λ 本文选取为经验值 1,k-means 聚类迭代次数选择常用经验值 10[19],最终充填时迭代次数是反复执行聚类的次数,与最终充填的精度正相关,为保证执行时间,将其设置为 20。

针对决策属性的权重 a 值,缺失数据比例越高,该值越小,而缺失数据比例越低,该值较大。在数据集 1 上随机缺失 10% ~60% ,a 取值分别为 0.5、1.0、2.0(相似性阈值取为 1,通过迭代次数控制停止时机)下充填误差如图 2 所示。

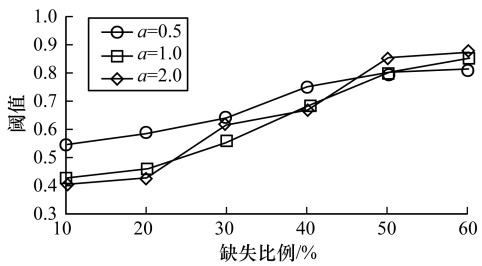

图 2　不同 *a* 值下充填误差

从图 2 可以看出,缺失比例越大,*a* 值应当越小。在本文实验过程中,由于需要针对多个数据集的不同缺失比例进行实验,因此对 *a* 值不进行调节,均设置为 1。

相似性阈值 *threshold_i* 表示 2 次聚类结果列表的互信息,k-means 随机选择初始质心将产生局部点的误分[20]。对第 1 次聚类结果列表 best_result 修改 5 次,每次随机修改(在可能类别中选择)其中的 20%,以修改前后两列表的 5 个互信息均值作为 *threshold_i*。这样取值意味着对最后的结果要求变化不超过 20%。不同数据集聚类结果随机更改 20% 的互信息如表 3 所示。

表 3　各数据集聚类结果随机更改 20% 的互信息

数据集	第 1 次	第 2 次	第 3 次	第 4 次	第 5 次
ill-Valley	0.478	0.403	0.473	0.503	0.420
LSVT_ voice_rehabilitation	0.418	0.423	0.511	0.478	0.456
breast_cancer_wisconsin	0.509	0.425	0.485	0.471	0.416
BlogFeedback	0.933	0.921	0.952	0.931	0.944
gesture_phase_dataset	0.851	0.905	0.920	0.881	0.902
Dataset for Sensorless Drive Diagnosis	0.967	0.961	0.951	0.970	0.939
Connectionist Bench	0.424	0.431	0.479	0.482	0.422

4.4　充填效果对比

将 7 个数据集分别随机缺失 10% ~60% 进行充填实验,并以均值充填为标准,根据式(13)进行评价对比。实验结果如表 4 ~ 表 10 所示,其中,设缺失比例为 *X*。

表 4　1 号数据集充填误差

方法	*X* 为 10%	*X* 为 20%	*X* 为 30%	*X* 为 40%	*X* 为 50%	*X* 为 60%
KNN	0.367	0.500	0.623	0.668	0.681	0.751
PDPCM	0.277	0.343	0.523	0.688	0.701	0.731
ROUSTIDA	0.275	0.577	0.594	0.724	0.759	0.772
RMFBC	0.207	0.377	0.392	0.534	0.576	0.682

表 5　2 号数据集充填误差

方法	*X* 为 10%	*X* 为 20%	*X* 为 30%	*X* 为 40%	*X* 为 50%	*X* 为 60%
KNN	0.413	0.477	0.613	0.768	0.681	0.733
PDPCM	0.342	0.414	0.474	0.789	0.702	0.777
ROUSTIDA	0.442	0.514	0.574	0.689	0.714	0.791
RMFBC	0.279	0.361	0.433	0.538	0.676	0.752

表 6　3 号数据集充填误差

方法	*X* 为 10%	*X* 为 20%	*X* 为 30%	*X* 为 40%	*X* 为 50%	*X* 为 60%
KNN	0.418	0.587	0.691	0.772	0.691	0.781
PDPCM	0.298	0.490	0.712	0.773	0.821	0.830
ROUSTIDA	0.388	0.477	0.694	0.785	0.791	0.799
RMFBC	0.319	0.417	0.602	0.734	0.576	0.812

表 7　4 号数据集充填误差

方法	*X* 为 10%	*X* 为 20%	*X* 为 30%	*X* 为 40%	*X* 为 50%	*X* 为 60%
KNN	0.237	0.348	0.471	0.643	0.731	0.841
PDPCM	0.232	0.316	0.369	0.528	0.635	0.618
ROUSTIDA	0.288	0.572	0.516	0.614	0.659	0.753
RMFBC	0.212	0.299	0.389	0.588	0.535	0.698

表 8　5 号数据集充填误差

方法	*X* 为 10%	*X* 为 20%	*X* 为 30%	*X* 为 40%	*X* 为 50%	*X* 为 60%
KNN	0.412	0.477	0.674	0.668	0.783	0.824
PDPCM	0.361	0.447	0.593	0.671	0.788	0.833
ROUSTIDA	0.477	0.541	0.619	0.637	0.793	0.764
RMFBC	0.374	0.451	0.683	0.604	0.776	0.802

表 9　6 号数据集充填误差

方法	*X* 为 10%	*X* 为 20%	*X* 为 30%	*X* 为 40%	*X* 为 50%	*X* 为 60%
KNN	0.218	0.315	0.457	0.518	0.752	0.702
PDPCM	0.194	0.272	0.383	0.613	0.634	0.711
ROUSTIDA	0.275	0.309	0.494	0.684	0.699	0.731
RMFBC	0.164	0.245	0.392	0.573	0.574	0.608

表 10　7 号数据集充填误差

方法	*X* 为 10%	*X* 为 20%	*X* 为 30%	*X* 为 40%	*X* 为 50%	*X* 为 60%
KNN	0.394	0.511	0.553	0.686	0.731	0.788
PDPCM	0.303	0.422	0.549	0.671	0.666	0.793
ROUSTIDA	0.405	0.592	0.638	0.624	0.667	0.772
RMFBC	0.288	0.413	0.492	0.655	0.673	0.782

由表 4 ~ 表 10 结果可看出,在缺失比例逐渐升高时,4 种充填方法都接近均值充填的效果。RMFBC 充填通过聚类确定缺失样本所处的类簇,可较明确地选择近邻点进行充填,因此在数据缺失量较小时,充填优势更为明显。在缺失样本较多时,尤其是同一特征大量缺失时,则 RMFBC 相当于均值充填,可能存在充填效果逊于选择近邻的其他算法。由于决策属性权重 a 并未设置为较大数值,决策属性重要程度体现不明显,因此在多分类数据集 3、4、5 和二分类数据集 1、6、7 中,充填效果区别不大。

4.5 应用结果分析

在油气勘探领域,人工地震方法采集的声波是在不同地层传播的声波信号(数据),通过声波信号与地层地质性质关系,实现对地层的认识。地层的岩性(如砂岩、泥岩、碳酸盐岩等)识别对油气勘探生产至关重要。地层复杂性导致在收集地层数据时存在局部区域无法采样及采样不均匀,若直接丢弃将损失大量数据,因此,在岩性识别建模前需对不完整岩性样本进行充填预处理。

为证明本文方法在地震数据充填上的有效性,采用的实验数据来自华北油田某地区,包括 16 口测井的岩性数据,共 1 956 道、501 个采样点,总样本数为 2 003 个。低频振幅数据能有效地反映地层结构,而高频振幅数则能有效地反映地层细节,组成了由 1 个全频和 8 个分频地震数据并添加包括峰度、均值、方差等 5 种属性以进行砂岩、泥岩、碳酸盐岩3 类岩性的识别工作(即共14 个条件属性及 1 个决策属性的数据为输入),对其中完备部分进行随机缺失和充填实验,实验结果如图 3 所示。从图 3 可以看出,与其他方法相比,RMFBC 方法充填误差较小。

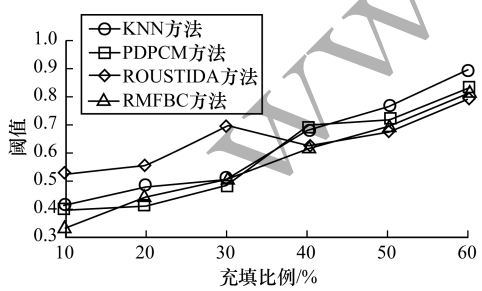

图 3 地震数据集充填误差

4.6 时间复杂度分析

针对样本容量为 N 的数据集,KNN 寻找近邻样本的时间复杂度为 $O(N)$[21],在对每一个缺失样本进行近邻点查找时需要进行一次完整的扫描,因此,KNN 充填方法的时间复杂度为 $O(N^2)$。在 RMFBC 方法充填中,均值充填部分时间复杂度为 $O(N)$,使用一次 k-means 聚类时间复杂度为 $O(|R| \times I \times k \times N)$,其中,$|R|$为属性数,$I$ 为迭代次数,k 为目标类别数,因此本文算法综合时间复杂度为 $O(time \times (|R| \times I \times k \times N + N))$,其中,$time$ 为算法最大迭代次数。上述条件除 $|R|$外皆为常数,总体计算时间与 KNN 充填算法相同,处理速度在可接受范围之内。

5 结束语

目前进行数据分析的多种模型对于存在缺失值的数据集缺乏有效的处理,因此,需要对缺失数据进行充填补齐。本文提出一种 RMFBC 数据充填方法,通过均值预充填,采用聚类修正原来缺失的样本,根据相对误差评价最终充填效果。实验结果表明,该方法通过反复聚类修正,克服聚类依赖于质心的随机性,使得最终充填更加接近真实数据。

参考文献

[1] LIU Zhunga, PAN Quan, DEZERT J, et al. Adaptive imputation of missing values for incomplete pattern classification[J]. Pattern Recognition, 2016, 52(C): 85-95.

[2] 高科,刁兴春,曹建军. 含缺失属性值的问题数据检测与修复[J]. 计算机工程与设计,2016,37(3):643-649.

[3] 韩飞,沈镇林. 基于不完备集双聚类的缺失数据填补算法[J]. 计算机工程,2016,42(4):20-26.

[4] ARMINA R, ZAIN A M, ALI N A, et al. A review on missing value estimation using imputation algorithm[C]// Proceedings of JPCS'17. Washington D. C., USA: IEEE Press, 2017: 125-136.

[5] YAN Xiaobo, XIONG Weiqing, HU Liang, et al. Missing value imputation based on gaussian mixture model for the internet of things[J]. Mathematical Problems in Engineering, 2015(3): 1-8.

[6] PURWAR A, SINGH S K. Hybrid prediction model with missing value imputation for medical data[J]. Expert Systems with Applications, 2015, 42(13): 5621-5631.

[7] XUE Wenzhuo, LI Hui, PENG Yanguo, et al. Secure k nearest neighbors query for high-dimensional vectors in outsourced environments[J]. IEEE Transactions on Big Data, 2017(99): 1.

[8] 杨涛,骆嘉伟,王艳. 基于马氏距离的缺失值充填算法[J]. 计算机应用,2005,25(12):2868-2871.

[9] LIEW W C, LAW N F. Missing value imputation for gene expression data: computational techniques to recover missing data from available information[J]. Briefings in Bioinformatics, 2011, 12(5): 498-513.

[10] 卜范玉,陈志奎. 基于聚类和自动编码机的缺失数据充填算法[J]. 计算机工程与应用, 2015, 51(18): 13-17.

[11] SIMINSKI K. Neuro-rough-fuzzy approach for regression modelling from missing data[J]. International Journal of Applied Mathematics and Computer Science, 2012, 22(2): 461-476.

[12] 焦媛. 云计算下多维数据缺失特征填补仿真研究[J]. 计算机仿真, 2018, 35(2): 262-265.

[13] WANG Changzhong, QI Yali, SHAO Mingwen, et al. A fitting model for feature selection with fuzzy rough sets[J]. IEEE Transactions on Fuzzy Systems, 2017, 25(4): 741-753.

[14] 顾爱华. 云计算网络中高维数据标准化处理优化仿真[J]. 计算机仿真, 2017, 34(3): 317-320.

[15] KISORE N R, KOTESWARAIAH C B. Improving ATM coverage area using density based clustering algorithm and voronoi diagrams[J]. Information Sciences, 2017, 376: 1-20.

[16] VINH N X, ZHOU Shuo, CHAN J, et al. Can high-order dependencies improve mutual information based feature selection? [J]. Pattern Recognition, 2016, 53(C): 46-58.

[17] 牛咏梅. 基于粗糙集的海量数据挖掘算法研究[J]. 现代电子技术, 2016, 39(7): 115-119.

[18] YUAN Jingling, CHEN Mincheng, JIANG Tao, et al. Complete tolerance relation based parallel filling for incomplete energy big data[J]. Knowledge-Based Systems, 2017, 132: 1-23.

[19] YANG Jie, MA Yan, ZHANG Xiangfen, et al. An initialization method based on hybrid distance for k-means algorithm[J]. Neural Computation, 2017, 29(11): 1-24.

[20] AHMADIAN S, NOROUZI-FARD A, SVENSSON O, et al. Better guarantees for k-means and euclidean k-median by primal-dual algorithms[C]//Proceedings of the 58th IEEE Annual Symposium on Foundations of Computer Science. Washinton D. C., USA: IEEE Press, 2017: 61-72.

[21] MAILLO J, RAMIREZ S, TRIGUERO I, et al. kNN-IS: an iterative spark-based design of the k-nearest neighbors classifier for big data[J]. Knowledge-Based Systems, 2016, 99: 1-21.

编辑　索书志