文章编号：2095-2163(2021)11-0031-04 中图分类号：O212.1；TP274 文献标志码：A

基于加权 K 近邻算法的缺失数据填补研究

郑智泉，王孟孟，田维琦
（贵州民族大学 数据科学与信息工程学院，贵阳 550025）

摘 要：针对数据缺失问题，本文在完全随机缺失的前提下，对完整数据集进行不同比例的挖空处理，并使用 K 近邻算法进行缺失值填补；采用交叉验证法优化 K 值；最后借用高斯函数，对传统 K 近邻算法进行加权处理，提出加权 K 近邻算法。实验结果表明，不论 K 取值多大，加权 K 近邻算法填补效果均优于传统 K 近邻算法；且 $K = 2$ 时，两种算法填补效果达到最佳。

关键词：数据缺失；K 近邻；交叉验证；高斯函数；加权 K 近邻

Research on missing data filling based on Weighted K-Nearest Neighbor algorithm

ZHENG Zhiquan, WANG Mengmeng, TIAN Weiqi
(College of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China)

【Abstract】 In order to solve the problem of missing data, this paper carries out different proportion of hollowed-out processing on the complete data set under the premise of completely random missing and uses the K-Nearest Neighbor algorithm to fill in the missing values. Then cross validation method was used to optimize K value. Finally, the traditional K-Nearest Neighbor algorithm is weighted by Gaussian function, and the Weighted K-Nearest Neighbor algorithm is proposed. The experimental results show that the Weighted K-Nearest Neighbor algorithm is better than the traditional K-Nearest Neighbor algorithm no matter what the value of K is, and the two algorithms achieve the best filling effect when $K = 2$.

【Key words】 missing data; K-Nearest Neighbor; cross validation; Gaussian function; Weighted K-Nearest Neighbor

0 引 言

样本数据存在缺失值的问题，一直是科学研究领域和社会生产活动中广泛关注的问题之一。造成数据缺失的原因很多，若按照收集过程来分，数据的缺失存在于收集过程中和收集完成后。在数据收集的过程中，由于技术上无法获取[1]、数据获取代价过高、数据采集设备故障、因隐私问题而造成的单元无回答等情况，都会影响样本集的完整性[2]；样本数据收集完成后，因工作人员的失误而导致的数据丢失[3]，或是收集上来的部分数据有误、数据不可用也会间接造成样本集的不完整。

针对数据缺失原因的多样化，文献[4]中提出了 3 种缺失机制，定义了数据缺失的类型。即完全随机缺失（MCAR，Missing Completely at Random）、随机缺失（MAR，Missing at Random）、非随机缺失（NMAR，Not Missing at Random）。其中，MCAR 表示数据的缺失与完全变量和不完全变量均是无关的；MAR 表示数据的缺失只与完全变量有关；NMAR 表示数据的缺失与不完全变量有关。不同的缺失机制需要选择不同的处理办法，不处理、直接删除含缺失值的样本点和缺失值填补是处理数据缺失问题的主要方式。对大部分数据而言，不处理会导致统计推断出现较大偏差，而直接删除含缺失值的样本点又会对数据造成浪费，因此，对缺失值选用适当的方法进行填补处理成为人们关注的焦点。

1 数据填补算法

传统的数据填补算法有均值填补、众数填补、中位数填补、热卡填补、冷卡填补、回归填补、多重插补等方法。每种数据填补算法都有其特定的使用条件和应用场景，对于不同的数据缺失机制和具体问题，选择合适的填补方法尤为重要。近年来，随着数据的爆发式增长，机器学习算法有了很大的发展。针对数据缺失问题，国内外很多学者开始尝试使用机器学习算法进行填补处理，填补效果优良。本文使用 K 近邻算法进行缺失数据的填补处理，使用交叉验证法优化 K 值，并在此基础上对 K 近邻进行改进，提出加权 K 近邻算法对缺失值进行填补。

基金项目：贵州民族大学“部校共建”专项项目（GZMDBXSZM1908）。

作者简介：郑智泉（1990-），男，硕士研究生，主要研究方向：统计模型与统计计算、机器学习；王孟孟（1995-），女，硕士研究生，主要研究方向：统计模型与统计计算；田维琦（1996-），男，硕士研究生，主要研究方向：统计模型与统计计算。

收稿日期：2021-08-13

1.1 K 近邻填补(K- Nearest Neighbor,KNN)

KNN 算法被广泛应用于图像分类等领域,是机器学习算法的一种。其核心思想,是根据已有训练集样本进行模型训练,训练集中样本点的类别需要人为进行标注。模型训练完成后,根据测试集中已有的样本信息为待分类样本点在训练集中寻找 $K(K \geqslant 1)$ 个近邻,然后根据被选出的 K 个近邻来判断待分类样本点的类别。根据 KNN 算法的核心思想,如何选取 K 个近邻和选取几个近邻成了该算法的核心问题。

度量不同样本点之间的距离公式有很多,常用的距离公式有欧式距离、马氏距离、曼哈顿距离等。针对如何选取 K 个近邻的问题,本文选取曼哈顿距离公式进行度量,公式如式(1):

$$d_{(x,y)} = \sum_{i=1}^{m} |x_i - y_i| \tag{1}$$

其中,x 代表测试集样本点;y 代表训练集样本点;$d_{(x,y)}$ 代表 x 和 y 两个样本点之间的距离大小(值越小代表两个样本点越相似);m 代表不包含缺失值的变量个数。

KNN 算法具有简单容易实现的优点,但缺点也十分明显。由于需要为每个测试集中的样本点寻找 K 个近邻,在算法执行过程中,为了计算当前待分类样本与训练集中每个样本点的距离,算法需要遍历整个训练集,时间开销巨大,尤其是在训练集样本点或测试集样本点数量庞大时。

K 的取值会影响最终的分类结果。K 取值过大,会导致分类不准确,结果变得模糊;K 取值过小,会影响算法的鲁棒性,导致分类结果课程出错。选择合适的 K 值也是 KNN 算法需要解决的问题之一。在 K 值确定之后,算法会为待分类样本在训练集中点寻找 K 个近邻,然后对选出的 K 个近邻进行分析,进而判断待分类样本点的类别。

1.2 交叉验证法

针对 K 值优化问题,本文采用交叉验证法对 K 近邻算法核心参数进行选取。交叉验证法的基本思想,是使用训练集数据,通过合适的方法寻找最优 K 值。本文交叉验证法的具体实现步骤如下:

Step 1 将训练集样本进行随机排序得到 X,将 X 等分为 n 个子样本集 $X_1, X_2, X_3, \cdots, X_n$。

其中,$n = floor(\frac{N}{N \times p})$;$N$ 为训练集样本数量;p 为缺失率;$floor(\cdot)$ 为取整函数。

Step 2 令 $X_i, i \in (1, n-1)$ 作为新的测试集样本,并将 X_i, X_{i+1} 中对应的缺失变量 x_{mis} 的真实值 $y^i_{obs}, i \in (1, n-1)$ 进行删除处理,X 中的其余子样本集作为新的训练集样本。使用 KNN 算法对 x_{mis} 的值进行填补,得到填补值 $\hat{y}^i_{obs}, i \in (1, n-1)$。

Step 3 令 $N^* = \frac{N}{n}$。

其中,N^* 为子样本集中样本点数量;N 为训练集样本数量;n 为子样本集数量。

Step 4 令 $MAE_{CV} = \{MAE^1_{CV}, MAE^2_{CV}, \cdots, MAE^i_{CV}, \cdots, MAE^{n-1}_{CV}\}$。

其中,MAE^i_{CV} 代表第 i 次运算得到的均方误差,$MAE^i_{CV} = \frac{1}{N^*} \sum_{j}^{N^*} |\hat{y}^i_{obs} - y^i_{obs}|, i \in (1, n-1)$。

Step 5 令 $\overline{MAE_{CV}} = mean(MAE_{CV})$、$MAE\_SD_{CV} = sd(MAE_{CV})$。

其中,$mean(\cdot)$ 为求均值函数,$sd(\cdot)$ 为求标准差函数。

Step 6 将 K 的取值设为 1~20,重复 Step2-5,寻找最小的 $\overline{MAE_{CV}}$,MAE_SD_{CV} 即可。

根据以上步骤进行实验,得到结果如图 1 所示:

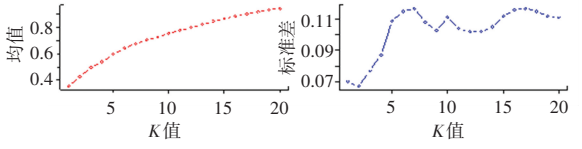

图 1 不同 K 值下 MAE 结果的均值与标准差

Fig. 1 Mean and standard deviation of *MAE* results with different *K* values

根据图 1 实验结果可得,随着 K 值的不断增大,$\overline{MAE_{CV}}$ 结果也逐步变大;当 $2 \leqslant K \leqslant 5$ 时,MAE_SD_{CV} 结果呈现逐步上升趋势,算法稳定性随 K 值的增大而降低;当 $K > 7$ 时,MAE_SD_{CV} 结果开始呈现震荡状态,算法稳定性与 K 值的变化趋势呈现出不一致性;当 $K = 2$ 时,模型结果相对较好,且稳定性达到最佳。

1.3 加权 k 近邻(Weighted K-Nearest Neighbor,WKNN)

WKNN 算法是对 KNN 算法的一种改进。传统的 KNN 算法在对选出的 K 个近邻进行分析时,一般采用少数服从多数的原则来判断待分类样本点的类别,这可能导致分类错误。为了弥补这一缺陷,本文采用加权的方式对传统 KNN 算法进行改进,核心思想是为最邻近待分类样本点的近邻赋予更高权重,依此类推。本文使用高斯函数为基础进行加权处

理，高斯函数形式如下：

$$f(x) = ae^{-(x-b)^2/2c^2}$$

其中，a,b,c 为实数常数，且 $a > 0$。权重函数如式(2)：

$$\begin{cases} w_l = \dfrac{ae^{-(x^l-b)^2/2c^2}}{\sum_{l=1}^{k} ae^{-(x^l-b)^2/2c^2}} \\ \sum_{l=1}^{k} w_l = 1 \end{cases} \tag{2}$$

其中，a,b,c 为实数常数，且 $a > 0$；x^l 为第 l 个近邻与待分类样本点之间的距离；k 为选取的近邻数量。

本文中，令 $a = 1, b = 0, c = 10$。对上述权重函数整理后，得到本文所使用的权重函数如式(3)：

$$\begin{cases} w_l = \dfrac{e^{-d_{(x,y)}^{l\,2}/200}}{\sum_{l=1}^{k} e^{-d_{(x,y)}^{l\,2}/200}} \\ \sum_{l=1}^{k} w_l = 1 \end{cases} \tag{3}$$

其中，$d_{(x,y)}^{l}$ 代表第 l 个近邻与待分类样本点间的距离，k 为选取的近邻数量。

2　实验方法

本文实验运行系统环境为：Windows10、软件版本为 RStudio Version 1.4.1106；所用数据集来自 R 语言 MASS 包中的经典数据集 Boston。根据数据缺失机制的定义，本文使用计算机模拟数据缺失过程，采用随机数生成法模拟完全随机缺失[5-6]。将 Boston 数据集中“rad(径向公路可达性指数)”指标列和“ptratio(城镇的师生比例)”指标列进行 1%、5%、10%的随机挖空处理；将含有缺失值的数据集使用 KNN 算法和 WKNN 算法进行缺失值填补，在填补过程中，令两种算法中的 K 值分别取为 2、5、10。最后，在平均绝对误差(Mean Absolute Error，*MAE*)、均方根误差(Root Mean Square Error，*RMSE*)、平均绝对误差百分比(Mean Absolute Percentage Error，*MAPE*) 3 个评价准则下，分别比较 KNN 算法、WKNN 算法在不同 K 值下的填补效果。本文实验过程有两点需要特别强调，一是在计算过程中，本文不对样本集进行任何标准化处理；二是在确定缺失率和 K 值以后，本文再对原始数据采用随机数生成法进行数据缺失模拟，进而在相同的缺失率和 K 值下比较 KNN、WKNN 两种算法的填补效果。

3　实验分析

针对不同的缺失率所生成的不完整数据集，确定 K 值之后，分别使用 KNN 填补法和 WKNN 填补法进行缺失值填补，进而计算不同缺失率、不同 K 值下 *MAE*、*RMSE*、*MAPE* 的值。实验结果见表 1。

表 1　不同缺失率、不同 *K* 值下两类方法的 3 类评价结果

Tab. 1　Three types of evaluation results of two types of methods under different miss rates and different *K* values

缺失率	*K* 值	填补算法	*MAE*	*RMSE*	*MAPE*
10%	*K* = 10	KNN	0.997 8	1.223 3	0.120 1
		WKNN	0.527 2	0.938 5	0.064 4
	K = 5	KNN	0.910 0	1.331 9	0.110 5
		WKNN	0.445 7	0.906 3	0.052 2
	K = 2	KNN	0.632 0	1.185 9	0.057 8
		WKNN	0.351 9	0.850 9	0.030 4
5%	*K* = 10	KNN	0.983 2	1.110 0	0.096 3
		WKNN	0.430 4	0.774 4	0.048 0
	K = 5	KNN	0.472 0	0.641 5	0.044 4
		WKNN	0.183 6	0.311 3	0.026 0
	K = 2	KNN	0.392 0	0.777 9	0.036 9
		WKNN	0.254 9	0.690 3	0.019 6
1%	*K* = 10	KNN	0.544 0	0.718 6	0.036 0
		WKNN	0.069 4	0.146 5	0.013 3
	K = 5	KNN	0.448 0	0.817 0	0.033 5
		WKNN	0.055 2	0.121 6	0.003 3
	K = 2	KNN	0.030 0	0.067 1	0.001 8
		WKNN	0.001 0	0.002 3	0.000 1

由实验结果可知：当缺失率为 1%、5%、10%时，不论 K 取值如何，WKNN 算法填补效果均优于 KNN 算法；当 $K = 2$ 时，KNN 算法和 WKNN 算法填补效果均达到最优，印证了前文所述交叉验证法的有效性；WKNN 算法的 *MAPE* 值均小于 KNN 算法，证明 WKNN 算法填补缺失值的可靠性也优于 KNN 算法；随着缺失率的增大，KNN 算法和 WKNN 算法的填补效果均有所下降。

4　结束语

针对数据缺失问题，本文提出一种新的算法对缺失值进行填补处理，通过计算机模拟不同缺失率的数据集，使用 KNN、WKNN 算法对其进行填补，并用交叉验证法对 K 值进行优化处理。大量的实验证明，在不同缺失率下，WKNN 算法的填补效果和可靠性均优于 KNN 算法，这也表明，机器学习算法在缺失数据处理方面有很好的应用前景。

参考文献

[1] 邓建新，单路宝，贺德强，等. 缺失数据的处理方法及其发展趋势[J]. 统计与决策，2019，35(23)：28-34.

(下转第 42 页)