doi: 10.3969/j.issn.0490-6756.2020.05.008

基于灰色关联分析的类中心缺失值填补方法

刘 莎, 杨有龙
(西安电子科技大学数学与统计学院, 西安 710126)

摘 要: 真实数据集中含有缺失值,许多数据分析技术不能直接应用到不完整数据上,且缺失值的存在会明显地降低算法的有效性,缺失数据处理是一个不可缺少的数据预处理过程,因此提出了一个基于统计度量的缺失值填补算法,名为灰色类中心缺失值填补(GCCMVI)方法,利用数据点的类中心和标准差来填补缺失值,此外,通过比较阈值和实例与类中心间相关性的大小关系,决定是否加上(减去)标准差,灰色关联分析用来计算相关性,在缺失值被填补后,得到的完整的数据集用来训练支持向量机(SVM)分类器. 在三种类型不同的数据集上进行比较,以分类精度,填补效果,填补时间作为评估准则来衡量算法的有效性. 实验结果表明,所提出的算法显著地提高了分类精度和填补效果.

关键词: 数据分析; 不完整数据; 缺失值填补; 类中心; 灰色关联分析
中图分类号: TP391 **文献标识码**: A **文章编号**: 0490-6756(2020)05-0871-08

Imputing missing value by class center based on grey relational analysis

LIU Sha, *YANG You-Long*
(School of Mathematics and Statistics, Xidian University, Xi'an 710126, China)

Abstract: Many data mining techniques cannot be applied directly to incomplete dataset which contains missing values. Furthermore, missing values will significantly reduce the effectiveness of the algorithm. So missing data management is an indispensable data preprocessing process. The proposed imputation method is based on statistical measurements named as grey class center missing value imputation (GCCMVI) approach. The missing values are imputed based on class center and standard deviation. Besides, the standard deviation is added (subtracted) or not determined by comparing the threshold and the relevance between class center and instance. Grey relational analysis is used to compute relevance. After the missing values are filled, the complete dataset is used to train the support vector machine (SVM) classifier. The comparative experiments are carried out on three datasets in different types. The classification accuracy, imputation performance and imputation time are used as criteria to evaluate the effectiveness of the proposed algorithm, experimental results show that it significantly improves the classification accuracy and imputation performance.

Keywords: Data mining; Incomplete data; Missing value imputation; Class center; Grey relational analysis

收稿日期: 2019-06-04
基金项目: 国家自然科学基金(61573266)
作者简介: 刘莎(1995—), 女, 陕西汉中人, 硕士研究生, 研究方向为数据分析理论及应用. E-mail: liushasha418@163.com
通讯作者: 杨有龙. E-mail: ylyang@mail.xidian.edu.cn

1 引 言

缺失数据是数据分析中一个不可避免的问题,且缺失值的存在会严重地降低算法的有效性. 因此,缺失数据的处理是一个不可缺少的数据预处理过程,一般分为两类,一类是直接删除含有缺失值的数据点,这种方法简单易操作,但缺点是在缺失比例较高时,该方法会造成信息的大量流失从而降低有效性. 另一类是缺失值填补方法,用估计值来代替缺失值. 一般地,缺失值填补分为基于统计技术和基于机器学习技术的,机器学习技术包括:k 近邻,人工神经网络,支持向量机,决策树和随机森林等.

广泛应用的统计技术包括:均值或众数填补和回归法. 均值或众数填补用相同属性的平均值或众数来代替缺失值. 最近一个名为基于类中心的缺失值填补(Class Center Missing Value Imputation, CCMVI)方法在文献[1]被提出,其主要思想与聚类中心应用到 k 均值算法中的想法相似,基于类中心,标准差和欧氏距离来填补缺失值,该算法的主要缺点是不能适用于缺失比例较高的情况. 因此,我们提出了一个改进的类中心缺失值填补方法,名为灰色类中心缺失值填补(Grey Class Center Missing Value Imputation, GCCMVI),我们改进了类中心,标准差和阈值的计算,且利用灰色关联度代替欧式距离来计算实例间的相关性. 在 17 个 UCI 数据集上进行了对比实验,最终实验结果表明,我们所提出的方法显著提高了分类精度和填补效果.

2 相关工作

2.1 缺失机制

三种不同的缺失机制,分别为:完全随机缺失(Missing Completely at Random,MCAR),随机缺失(Missing at Random,MAR)和非随机缺失(Not Missing at Random,NMAR)[1].

给出如下条件. D 是一个不完整数据集,有 r 个特征,$D=\{A_1,A_2,\cdots,A_r\}$,含有 n 个实例,则整个数据集可以分为两个部分,$D=\{D^{obs},D^{mis}\}$,其中 D^{obs} 是所有观测到的实例的集合,D^{mis} 是含有缺失值的实例的集合,用 R 表示一个回应指标矩阵,与 D 大小相同来表述 D 的缺失度,R 的每一项定义如下.

$$R_{ij}=\begin{cases}0, & \text{if } v_{ij} \text{ is missing.}\\ 1, & \text{if } v_{ij} \text{ is observed.}\end{cases} \tag{1}$$

其中,v_{ij} 是第 i 个实例在特征 A_j 处的值,$i=1,2,\cdots,n$, $j=1,2,\cdots,r$.

(1) 完全随机缺失(MCAR). 若缺失值在整个数据集中的分布都是完全随机的,换句话说,一个实例中的缺失值独立于任何其他的实例,不管该实例是缺失的还是非缺失的,则称为完全随机缺失. 以概率公式表示有:$Pr(R|D^{mis},D^{obs})=Pr(R)$.

(2) 随机缺失(MAR). 如果缺失值可以根据不完整数据集中的其他非缺失的特征值猜测得到,它意味着缺失值独立于任何缺失值但与观测值有关,则称为随机缺失. 以概率公式表示如下:$Pr(R|D^{mis},D^{obs})=Pr(R|D^{obs})$.

(3) 非随机缺失(NMAR). 如果不完整实例中的一个缺失值依赖于该实例中的至少一个其他的缺失值,即 $\Pr(R|D^{mis},D^{obs})$ 不等于 $\Pr(R|D^{obs})$,换句话说,它依赖于 D^{mis}.

2.2 灰色关联分析(Grey Ralational Analysis, GRA)

灰色系统理论被提出用来处理不确定系统(有部分知道的信息和部分不知道的信息),且可以从知道的信息中提取有价值的信息. 灰色关联系数和灰色关联度是灰色系统理论中两个重要的参数,它们被用来衡量两个随机实例间的相关性. 例如,Pan 等人[2]和 Huang 等人[3]展示了用灰色关联分析代替欧式距离或其变体来衡量两个实例间的相似度或相关性时的有效性. 此外,Sefidian 等人[4]提出了一个新的缺失值填补算法,用一个新的基于灰色的模糊 c 均值,基于互信息的特征选择和回归模型. Tian 等人[5]提出了一个缺失数据分析,即一个利用灰色系统理论和基于熵的聚类的综合的多重填补算法. 这些文献进一步反映了用灰色关联分析作为相关性度量的有效性,这鼓励我们在本文中使用灰色关联分析,相关的细节和公式如下.

考虑数据集 $D=\{x_0,x_1,x_2,\cdots,x_n\}$,这里 x_0 是参考实例,其他的是比较实例,每个实例 x_i 有 m 个特征,表示为:$x_i=(x_i(1),x_i(2),\cdots,x_i(m))$, $i=0,1,2,\cdots,n$.

由文献[2]两个实例间的灰色关联系数(GRC)的定义如下.

$$GRC(x_0(p),x_i(p))=\frac{\min_{\forall j}\min_{\forall k}|x_0(k)-x_j(k)|+\rho\max_{\forall j}\max_{\forall k}|x_0(k)-x_j(k)|}{|x_0(p)-x_i(p)|+\rho\max_{\forall j}\max_{\forall k}|x_0(k)-x_j(k)|} \tag{2}$$

其中，$i,j=1,2,\cdots,n$；$k,p=1,2,\cdots,m$；p 表示具体的一个特征；$x_0(p)$表示实例 x_0 中特征 p 处的值. 就公式而言，$|x_0(p)-x_i(p)|$项考虑两个实例 x_0 和 x_i 在特征 p 处取值间的差值，其余三项通过求最值过程，遍历所有的实例($\forall j$)和所有的特征($\forall k$)求其和参考实例 x_0 间的差值的最小值和最大值，这一过程相比于一般的距离公式的优点在于它不仅考虑了 x_0 和 x_i 间的局部的差距，还引入了一个全局性的距离考量. 显然，这也是 GRC 在描述相关性时比欧式距离更为有效的原因. ρ 是一个区别系数，$\rho\in[0,1]$，在本文中，根据文献[3]的实验结果，取 $\rho=0.5$.

对于字符型属性，GRC 值的计算公式如下.

$$GRC(x_0(p),x_i(p))=\begin{cases}1, & \text{if } x_0(p) \text{ and } x_i(p) \text{ are the same.}\\ 0, & \text{if } x_0(p) \text{ and } x_i(p) \text{ are different.}\end{cases} \tag{3}$$

显然，由上述公式，GRC 的取值范围是[0,1]，它衡量了 $x_0(p)$和 $x_i(p)$之间的相似度或相关性，灰色关联度(GRG)是 GRC 的平均值，表示为

$$GRG(x_0,x_i)=\frac{1}{m}\sum_{k=1}^{m}GRC(x_0(k),x_i(k)),\quad i=1,2,\cdots,n \tag{4}$$

然而，它与欧式距离的差别在于 GRG 值越大，相关性越大，在本文中，我们会用两个实例间的 GRG 值来衡量他们之间的相关性.

2.3　缺失值填补方法

均值填补(Mean Imputation, MI)是一个广泛应用的统计技术，它用相同特征的观测值的平均值或众数来代替缺失值，它是一个直接且有效的简单的单值填补方法.

回归填补(Regression Imputation, RI). 例如，Sefidian 等人[4]提出了一个新的缺失值填补方法结合了灰色关联分析，模糊 c 均值，互信息和回归模型，实验结果表明，在 7 个 UCI 数据集上，所提出的方法优于其他 5 个填补方法.

K 近邻填补(K-Nearest Neighbors Imputation, KNNI). 最简单的 K 近邻填补用缺失实例的 K 个最近邻的平均值或众数来代替缺失值，基于此，一些改进的 K 近邻填补方法被提出，他们引入了灰色关联分析和互信息来更进一步地提高分类精度和填补效果[2-3,6]. 此外，核函数，区间值聚合函数和加权的填补也被应用于 K 近邻填补[7-8].

人工神经网络(Artificial Neural Network, ANN). Yan 等人[9]提出了一个可选择的神经网络集成分类方法来处理不完整数据，它是一个改进的神经网络，基于一个阈值和一个优化过程来完成完整数据子集的选择，然后用完整子集来训练神经网络，实验结果表明，该方法优于一般的神经网络算法.

多重填补(Multiple Imputation, MIs). 在多重填补过程中，会产生若干个完整的数据子集，然后用多个估计值来代替缺失值. 与单值填补的区别是，多重填补中，一个缺失值的估计值不只一个，显然，多重填补一般比单值填补更加有效[10-12].

基于学习的填补(Learning-based Imputation, LBI). 此时，缺失值的填补在一个学习过程中实现，缺失值视为目标输出变量，而其他的观测值则视为输入变量，学习模型的预测结果被用来填补缺失值，例如，支持向量机(Support Vector Machine, SVM)，多元线性回归(Multiple Linear Regression, MLR)，朴素贝叶斯分类器(Naive Bayes Classifier, NBC)，决策树(Decision Tree, DT)，随机森林(Random Forest, RF)和多层感知机(Multi-Layer Perceptron, MLP)[13-15]被广泛应用于缺失值填补中.

随机森林(RF). 随机森林是多棵决策树的集成，它是广泛应用的集成学习技术之一，最近一些基于随机森林的缺失数据处理方法被提出，例如，Hapfelmeier 等人[16]提出了在不完整数据集上，用随机森林来进行变量选择，目的在于提高预测和解释数据的能力. 此外，Xia 等人[17]提出了一个调整的加权的随机森林算法来处理缺失值，该算法通过估计缺失值对树的决策的影响来调整树的投票权重，从而提高了算法在处理缺失数据时的有效性，这也是本文对比实验中的一个基准算法.

基于聚类的填补(Clustering based Imputation, CBI). 一个实例中的缺失值将通过位于相同聚类中最近的实例点来填补，例如，Tian 等人[5]提出了一个综合的多重填补算法，该算法用灰色系统理论和基于熵的聚类来填补缺失值. Tran 等人[18]提出了一个新的算法，用特征选择和聚类来处理不完整数据. 实验结果表明，这些算法显著提高了分

类效果.

除了上述缺失数据处理方法,最近一个改进的基于统计技术的算法[1]被用来处理缺失数据,名为基于类中心的缺失值填补(Class Center Missing Value Imputation,CCMVI)方法,利用类中心和标准差这些简单的统计量来填补缺失值,先用类中心来代替缺失值,然后基于一个阈值来判断是否加上(减去)标准差. 本文所提出的方法可以视为一个改进的类中心缺失值填补方法,名为灰色类中心缺失值填补方法 GCCMVI,灰色关联度代替文献[1]中的欧式距离,用来计算实例间的相关性. 此外,我们改进了类中心,标准差和阈值的计算,具体的介绍在下文给出.

3 灰色类中心缺失值填补方法

本文所提出的灰色类中心缺失值填补(GCCMVI)方法包括两个模块,名为模块 A 和模块 B,模块 A 是通过计算类中心与其他观测数据间的相关性来确定阈值,模块 B 是利用上述得到的阈值来填补缺失值,具体介绍如下.

3.1 模块 A:识别阈值

模块 A 的目的在于识别阈值,该阈值会用于后续的填补过程,它包含 6 步,算法的具体步骤如算法 1.

算法 1 识别阈值

输入:不完整数据集 D 包含 M 个特征,N 个类和 Num 个数据实例

输出:N 个阈值

---

1) 根据式(5),用最小最大标准化预处理数据集 D.

2) 根据已知的类标签将原始数据集 D 分为 N 个子集,表示为 $D_i, i=1,2,\cdots,N$.

3) 遍历数据集的所有实例,根据是否含有缺失值,将 D_i 分为完整子集 $D_{i\_complete}$ 和不完整子集 $D_{i\_incomplete}$,后者含有缺失值.

4) 利用数据子集 $D_i, i=1,2,\cdots,N$ 中所有的观测值来计算特征平均值和标准差,分别记为:Avg$[i,j]$和 Std$[i,j]$,然后由它们得到类中心和类标准差,即 $cent_i$=Avg$[i,]$和 Std$[i,]$.

5) 根据 $cent_i$ 的值,利用均值填补法预填补数据集 D_i 从而得到完整数据集,记作 $T_i, i=1,2,\cdots,N$.

6) 计算类中心 $cent_i$ 与 T_i 中其他实例间的相关性,根据方程(4)中的灰色关联度来计算相关性,相关性的中值就是类别 i 的阈值.

遍历所有的 i 值,就得到了每个类别的阈值.

一般来说,数据标准化方法包括最小最大标准化和 z 得分标准化等方法,为了便于分析,本文采用最小最大标准化将数据值转换为[0,1]区间内的值,计算公式如下.

$$x'_p(j)=\frac{\max_{\forall i}x_i(j)-x_p(j)}{\max_{\forall i}x_i(j)-\min_{\forall i}x_i(j)} \tag{5}$$

本文所提出的灰色类中心缺失值填补方法(GCCMVI)是一个改进的算法,相比于一个名为基于类中心的缺失值填补(CCMVI)方法[1],两者间的主要不同点反映在类中心,标准差和阈值的计算中. CCMVI 算法将原始的不完整数据集分为完整子集和不完整子集,然后仅依赖于完整子集中的观测值来计算类中心和标准差. 显然,当缺失比例增加时,相应的完整子集中的实例数目会减少,类中心和标准差的计算就会变得困难,或者依赖于极少数的实例点得到的类中心和标准差就不足以代表整个数据集的信息.

然而,我们所提出的方法解决了这一问题,如算法 1 的步骤 4)所描述的,根据 D_i 中的所有观测值来计算类中心和标准差,这一改进,不仅考虑了 $D_{i\_complete}$ 中的观测值,也考虑了 $D_{i\_incomplete}$ 中的观测值,这使得计算得到的结果更能有效的反映整个数据集的信息. 另一个不同是阈值的计算不同,CCMVI 算法通过衡量类中心与 $D_{i\_complete}$ 中实例间的欧式距离来得到阈值,当缺失比例增加,阈值的计算也会产生上述所提到的问题:计算困难或计算得到的结果不够有效. GCCMVI 通过一个预填补过程解决了这一问题,如算法 1 的步骤 5)和步骤 6)所述. 采用均值填补来完成预填补过程的原因是,模块 B 会先用类中心来代替缺失值,若采用其他的填补方法会对最终结果产生较大的影响. 本文中相关性的计算都依赖于灰色关联度的计算,具体计算在式(2)~(4)中给出,就公式而言,它区别考虑了数值型属性和字符型属性,这比欧式距离更适合于计算字符型和混合型数据间的相关性,且很多参考文献中的实验结果都表明:灰色关联度比欧式距离或其变体在衡量相关性时更具有效性.

3.2 模块 B:缺失值填补

模块 B 是利用模块 A 得到的阈值来填补缺失值,主要包括三步,下面所有的相关性的计算都是

根据灰色关联度(方程(4))得到，具体算法的步骤如算法 2.

算法 2　缺失值填补算法

输入：不完整数据集 $D_{i\_incomplete}$，$i=1,2,\cdots,N$，总共包含 M 个特征，N 个类和 Num 个数据实例

输出：填补后的完整数据集

1) 遍历所有的实例，若实例含有一个缺失值，则执行单值填补，若实例含有多个缺失值，则执行多值填补.

2) 单值填补. 先用类中心 $cent_i$ 代替缺失值，然后计算类中心与当前实例间的相关性，比较相关性与阈值的大小关系，若相关性大于等于阈值，则保持填补的值不变，若相关性小于阈值，则用填补的值加上或减去标准差来代替缺失值.

3) 多值填补. 首先计算实例中缺失值的数目，且记录每个缺失值存在的位置指标. 然后用类中心 $cent_i$ 代替缺失值，且计算类中心与当前填补的实例间的相关性，比较相关性与阈值的大小关系，若相关性大于等于阈值，则保持填补的值不变，否则，遍历所有的缺失位置，在每个缺失位置处依次用填补的值加上(减去)标准差来代替缺失值. 根据位置的不同会得到与缺失值数目相同的填补实例，并且计算类中心与这些填补实例间的相关性，最终的填补实例是使得相关性最大的那个填补实例.

4)重复上述步骤，直到不存在缺失值为止，此时就得到了填补后的完整数据集.

4　实验设置及结果分析

4.1　实验设置

本文的实验用了三种不同的数据集，包括：数值型，字符型和混合型数据集，其中有 8 个数值型数据集，6 个字符型数据集和 3 个混合型数据集，这些数据集都来自 UCI 机器学习数据库，实例数目在 101～28 056 之间，特征数目在 4～60 之间. 此外，一些数据集中存在着严重的不平衡问题，为了去除不平衡对实验结果的影响，当少数类的实例数目少于 5 时，采用了一个简单的过采样技术，简单地复制少数类实例，以此来增加少数类的实例数目，从而在缺失比例较高时，保证依然有足够的实例来计算少数类的类中心和标准差，数据集的基本信息在表 1 中给出.

表 1　数据集的基本信息

Tab. 1　The basic information of the datasets

	Datasets	#instances	#features	#classes
Numerical data	ecoli	336	7	8
	ionosphere	351	34	2
	Magic	19 020	10	2
	pima	768	8	2
	segment	2 310	19	7
	sonar	208	60	2
	waveform	5 000	40	3
	yeast	1 484	8	10
Categorical data	chess_m	28 056	6	18
	lymphography	148	18	4
	mushroom	5 644	22	2
	nursery	12 960	8	5
	promoters	106	57	2
	SPECT	267	22	2
Mixed data	card	653	15	2
	liver	345	6	2
	zoo	101	16	7

首先，将原始数据集按照 10 折交叉验证法，分为 90%的训练集和 10%的测试集，其中训练集的类标签是已知的，假设测试集的类标签未知，通过训练集得到的分类器来预测测试实例的类标签，比较预测得到的类标签与真实的类标签就可以得到该分类器的分类精度. 本文要研究的是不完整数据问题，因此会人为的在原始的完整的训练集中引入缺失值，由上文所述，有三种缺失机制，名为完全随机缺失，随机缺失和非随机缺失. 一般来说，完全随机缺失是一种最广泛存在的缺失机制. 因此，本文仅考虑了完全随机缺失情况. 实验所用的缺失比例为 10%～50%，增加步长为 10%，为了避免产生有偏的结果，每个缺失比例会做 10 次实验，每个缺失比例的最终的结果是 10 次实验的平均值.

我们将所提出的灰色类中心缺失值填补方法(GCCMVI)方法与 6 个基准算法进行比较，分别是 Mean，KNNI，SVM，WRF[17]，FKNNI[2]，和 CCMVI[1]，其中，WRF(Weighted Random Forest)是加权的随机森林的缩写，FKNNI(Feature K-Nearest Neighbors Imputation)是特征加权的灰色 K 近邻填补的缩写. 在不完整训练集中的缺失值用上述不同的填补方法分别填补后，得到的完整的训练集来训练支持向量机(SVM)分类器，测试集用来测试分类器的分类精度，显然分类精度越高，填补效果就越好.

除了分类精度，填补效果是衡量填补方法的另

一个主要指标,对于数值型属性,均方根误差(RMSE)用来衡量真实值与填补值之间的差距,显然 RMSE 值越小则表示填补效果越好. 对于字符型属性,命中率(Hit ratio)用来衡量真实值与填补值之间的差距,相反 Hit ratio 值越大,填补效果越好. 计算公式如下.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(e_i - \tilde{e}_i)^2} \tag{6}$$

$$Hit\ ratio = \frac{1}{m}\sum_{i=1}^{m} I(e_i, \tilde{e}_i) \tag{7}$$

其中,e_i 是真实值;$\tilde{e}_i$ 是填补的值;m 是缺失值的数目;函数 I 是一个指示函数,有 $I(e_i, \tilde{e}_i)=1$, $e_i=\tilde{e}_i$,否则等于 0.

4.2 数值型数据集实验结果及分析

表 2 和表 3 分别给出了在数值型数据集上,由不同填补方法得到的完整数据集训练得到的 SVM 的平均分类精度,以及平均的均方根误差(RMSE)值,表 2 中,CCMVI(+),CCMVI(−)和 GCCMVI(+),GCCMVI(−)分别表示在 CCMVI 和 GCCMVI 中加上和减去标准差.

每个数据集的最好的结果用黑体标记出来了,显然,根据表 2 给出的结果可知,所提出的 GCCMVI 方法在 5 个数值型数据集上取得了最高的分类精度,且相比于其他方法,分类精度增加了 10%~24%,而均值填补法(Mean)在 ionosphere 和 waveform 两个数据集上效果最好,且与 GCCMVI 间的差距较大. 此外 CCMVI 在 ecoli 数据集上取得了最好的分类效果,与 GCCMVI 间的差距较小. 这意味着对于不同的数据集,加上或减去标准差对结果的影响可能与数据集自身的分布相关. 且总的来说,加上标准差(+)比减去标准差(−)的精度高一些,但两者间没有显著的差距. GCCMVI 的填补误差普遍都低,是因为在数据集预处理时采用了最小最大标准化,去除了维度的影响,从而提高了填补效果.

表 2 数值型数据集的 SVM 的平均分类精度

Tab. 2 Average classification accuracies of SVM over numerical datasets

Dataset	Mean	SVM	KNNI	WRF	FKNNI	CCMVI(−)	CCMVI(+)	GCCMVI(−)	GCCMVI(+)
ecoli	0.643	0.669	0.69	0.667	0.667	**0.767**	**0.767**	0.725	0.730
ionosphere	**0.923**	0.918	0.510	0.871	0.870	0.894	0.895	0.825	0.829
Magic	0.650	0.651	0.651	0.609	0.611	0.650	0.650	0.829	**0.833**
pima	0.649	0.649	0.649	0.599	0.599	0.651	0.651	0.745	**0.746**
segment	0.292	0.357	0.390	0.576	0.569	0.404	0.404	0.615	**0.619**
sonar	0.564	0.565	0.530	0.563	0.552	0.603	0.603	0.787	**0.788**
waveform	**0.860**	0.858	0.556	0.824	0.823	0.828	0.828	0.672	0.678
yeast	0.396	0.398	0.389	0.372	0.372	0.436	0.440	0.672	**0.678**

表 3 数值型数据集的 RMSE

Tab. 3 RMSE of numerical datasets

Dataset	Mean	SVM	KNNI	WRF	FKNNI	CCMVI	GCCMVI
ecoli	0.16	0.14	0.16	0.18	0.17	0.11	**0.09**
ionosphere	0.61	**0.55**	0.61	0.82	0.74	0.63	2.08
Magic	38.90	36.30	38.82	22.30	23.11	17.55	**3.45**
pima	44.60	50.70	49.20	47.33	47.88	45.32	**0.97**
segment	35.10	36.19	30.22	36.20	26.21	29.48	**0.25**
sonar	0.22	0.21	0.25	0.79	0.64	0.24	**0.12**
waveform	1.43	**1.36**	1.52	1.92	1.68	1.54	9.74
yeast	0.10	0.11	0.10	0.17	0.18	**0.09**	0.23

4.3 字符型数据集的实验结果及分析

字符型数据集的实验结果在表格 4 和表 5 中给出,Hit ratio 为填补方法的命中率,显然命中率越高,填补效果就越好.

综上所述,GCCMVI 在 4 个字符型数据集上效果最好,且相比于其他方法,分类精度提高了 0.7%~13%,CCMVI 和 WRF 分别在 chess_m 和 nursery 数据集上获得了最高的分类精度. CCMVI 和 GCCMVI 在加上标准差(+)和减去标准差(−)两种情况下的分类精度分别相同,这表明加上或减去标准差对字符型数据集无显著的影响. 此外,GCCMVI 在 5 个数据集上取得了最高的 Hit ratio

值. 总体来说，所提出的 GCCMVI 方法在字符型　　数据集上效果最好.

表 4　字符型数据集的 SVM 的平均分类精度

Tab. 4　Average classification accuracies of SVM over categorical datasets

Dataset	Mean	SVM	KNNI	WRF	FKNNI	CCMVI(−)	CCMVI(+)	GCCMVI(−)	GCCMVI(+)
chess_m	0.549	0.542	0.527	0.604	0.596	**0.524**	**0.524**	0.335	0.335
lyphography	0.745	0.735	0.722	0.701	0.709	0.716	0.716	**0.882**	**0.882**
mushroom	0.931	0.919	0.853	0.956	0.960	0.874	0.874	**0.967**	**0.967**
nursery	0.932	0.934	0.935	**0.939**	0.936	0.932	0.932	0.891	0.891
promoters	0.651	0.668	0.682	0.678	0.668	0.669	0.669	**0.777**	**0.777**
SPECT	0.735	0.735	0.736	0.703	0.705	0.729	0.729	**0.826**	**0.826**

表 5　字符型数据集的 Hit ratio

Tab. 5　Hit ratio of categorical datasets

Dataset	Mean	SVM	KNNI	WRF	FKNNI	CCMVI	GCCMVI
chess_m	**0.81**	0.76	0.74	0.78	0.76	0.74	0.78
lymphography	0.44	0.47	0.41	0.46	0.45	0.43	**0.87**
mushroom	0.47	0.44	0.40	0.43	0.42	0.43	**0.79**
nursery	0.73	0.68	0.68	0.68	0.68	0.68	**0.83**
promoters	0.71	0.73	0.74	0.75	0.73	0.71	**0.79**
SPECT	0.39	0.36	0.33	0.35	0.35	0.39	**0.62**

4.4　混合型数据集实验结果及分析

表 6 和表 7 分别给出了混合型数据集的平均分类精度，填补误差(RMSE)和填补命中率(Hit).

显然，GCCMVI 在 card 和 liver 数据集上表现都很好，在 zoo 数据集上效果不够好，但差距较小.

表 6　混合型数据集的 SVM 的平均分类精度

Tab. 6　Average classification accuracies of SVM over mixed datasets

Dataset	Mean	SVM	KNNI	WRF	FKNNI	CCMVI(−)	CCMVI(+)	GCCMVI(−)	GCCMVI(+)
card(6:9)	0.543	0.537	0.541	0.542	0.539	0.547	0.546	0.834	**0.836**
liver(5:1)	0.579	0.578	0.579	0.575	0.574	0.585	0.585	**0.647**	**0.647**
zoo(1:15)	0.816	0.810	0.811	0.849	0.853	**0.876**	**0.876**	0.821	0.821

表 7　混合型数据集的 RMSE 和 Hit ratio

Tab. 7　RMSE and Hit ratio of mixed datasets

Dataset	Mean/mode		SVM		KNNI		WRF		FKNNI		CCMVI		GCCMVI	
	RMSE	Hit	RMSE	Hit	RMSE	Hit	RMSE	Hit	RMSE	Hit	RMSE	Hit	RMSE	Hit
card	1204	0.25	1204	0.18	2117	0.18	1151	0.20	1178	0.18	1185	0.20	**0.14**	**0.81**
liver	22	0.08	23	0.12	23	0.17	23.63	0.11	21.26	0.19	21.41	0.13	**0.06**	**0.70**
zoo	0.50	0.30	0.50	0.30	0.64	0.23	**0.44**	**0.30**	0.79	0.24	0.44	0.12	1.28	0.80

4.5　填补时间评估

不同填补方法的填补时间在表 8 中给出，GCCMVI 方法快于除了 Mean 和 CCMVI 的其他的方法. 比 Mean 慢是因为在 GCCMVI 中，均值填补是其中的一个预填补过程；比 CCMVI 慢的主要原因是当数据实例数目过大时，执行最小最大标准化过程，需要遍历所有值找最值，这会增加计算量. 但显然这一增加量是可以接受的，因为最小最大标准化过程导致了较少的填补时间的增加量，然而却换来了更小的填补误差，这使得填补效果更好.

表 8　不同方法的填补时间(单位：s)

Tab. 8　Imputation time of different methods(in seconds)

Mean	SVM	KNNI	WRF	FKNNI	CCMVI	GCCMVI
0.1	18 284	1 000	16 783	13 412	1.3	33

5　结　论

目前已有的缺失值填补方法都存在着填补效果不够好或者不能适用于缺失比例较高等缺陷. 因

此,我们提出了一个有效的缺失值填补方法名为灰色类中心缺失值填补方法(GCCMVI),它包括两个模块,模块 A 通过类中心与其他实例间的相关性来得到阈值,相关性通过灰色关联分析来计算,模块 B 用得到的阈值来填补缺失值. 实验使用了 3 种不同类型的数据集:数值型,字符型和混合型数据集,此外,所提出的方法与 6 种基准方法进行了比较,它们是 Mean,SVM,WRF, KNNI,FKNNI, CCMVI,实验结果表明,本文方法显著提高了分类精度和填补效果,且 GCCMVI(+)稍微好于 GCCMVI(-),但它们之间没有显著差异.

但本文只考虑了支持向量机一种分类器,没有一个不同分类器间的综合的比较实验. 未来研究中可以考虑其他的分类器. 除此之外,仅考虑了完全随机缺失的情况,其他两种缺失机制:随机缺失和非随机缺失未考虑,未来研究可加入三种缺失机制间的比较实验.

参考文献:

[1] Tsai C F, Li M L, Lin W C. A class center based approach for missing value imputation [J]. Knowl-Based Syst, 2018, 151: 124.

[2] Pan R, Yang T S, Cao J H, *et al*. Missing data imputation by K nearest neighbours based on grey relational structure and mutual information [J]. Appl Intell, 2015, 43: 614.

[3] Huang C C, Lee H M. A grey-based nearest neighbor approach for missing attribute value prediction [J]. Appl Intell , 2004, 20: 239.

[4] Sefidian A M, Daneshpour N. Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model [J]. Expert Syst Appl, 2019, 115: 68.

[5] Tian J, Yu B, Yu D, *et al*. Missing data analysis:a hybrid multiple imputation algorithm using gray system theory and entropy based on clustering [J]. Appl Intell, 2014, 40: 376.

[6] García-Laencina P J, Sancho-Gómez J L, Figueiras-Vidal A R, *et al*. K nearest neighbours with mutual information for simultaneous classification and missing data imputation [J]. Neurocomputing, 2009, 72: 1483.

[7] Gerhard T, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods [J]. Comput Stat Data An, 2015, 90 : 84.

[8] Bentkowska U, Bazan J G, Rzasa W, *et al*. Application of interval-valued aggregation to optimization problem of k-NN classifiers for missing values case [J]. Inform Sciences, 2019, 486: 434.

[9] Yan Y T, Zhang Y P, Zhang Y W, *et al*. A selective neural network ensemble classification for incomplete data [J]. Int J Mach Learn Cyb, 2017, 8: 1513.

[10] Tsai C F, Chang F Y. Combining instance selection for better missing value imputation [J]. J Syst Software, 2016, 122: 63.

[11] Zhang S C, Jin Z, Zhu X F. Missing data imputation by utilizing information within incompete instances [J]. J Syst Software, 2011, 84: 452.

[12] Gao H, Jian S L, Peng Y X, *et al*. A subspace ensemble framework for classification with high dimensional missing data [J]. Multidim Syst Sign P, 2017, 28: 1309.

[13] Kang P. Locally linear reconstruction based missing value imputation for supervised learning [J]. Neurocomputing, 2013, 118: 65.

[14] Zhang X N, Song S J, Wu C. Robust bayesian classification with incomplete data [J]. Cogn Comput, 2013, 5: 170.

[15] Luengo J, García S, Herrera F. On the choice of the best imputation methods for missing values considering three groups of classification methods [J]. Knowl Inf Syst, 2012, 32: 77.

[16] Hapfelmeier A, Ulm K. Variable selection by random forests using data with missing values [J]. Comput Stat Data An, 2014, 80: 129.

[17] Xia J, Zhang S Y, Cai G L, *et al*. Adjusted weight voting algorithm for random forests in handling missing values [J]. Pattern Recogn, 2017, 69: 52.

[18] Tran C T, Zhang M J, Andreae P, *et al*. Improving performance of classification on incomplete data using feature selection and clustering [J]. Appl Soft Comput, 2018, 73: 848.

引用本文格式:

中 文: 刘莎, 杨有龙. 基于灰色关联分析的类中心缺失值填补方法 [J]. 四川大学学报: 自然科学版, 2020, 57: 871.

英 文: Liu S, Yang Y L. Imputing missing value by class center based on grey relational analysis [J]. J Sichuan Univ: Nat Sci Ed, 2020, 57: 871.