doi:10.16088/j.issn.1001-6600.2015.04.010

基于LPP和$l_{2,1}$的KNN填充算法

苏毅娟[1],孙 可[2,3],邓振云[2,3],尹科军[2,3]

(1.广西师范学院 计算机与信息工程学院,广西 南宁 530023;2.广西师范大学 计算机科学与信息工程学院,
广西 桂林 541004;3.广西师范大学 广西多源信息挖掘与安全重点实验室,广西 桂林 541004)

摘 要:传统的KNN缺失值填充算法存在没有利用样本间属性的相关性,也没有考虑到保持样本数据本身的结构和去除噪声样本的问题。本文提出利用训练样本重构测试样本从而进行最近邻缺失值填充的方法,该方法重构过程充分利用样本间的相关性,也用到LPP(保局投影)保持数据结构在重构过程中不变,同时引入$l_{2,1}$范式用于去除噪声样本。在UCI数据集上的仿真实验结果表明,该方法比传统的KNN填充算法以及基于属性信息熵的Entropy-KNN算法有更高的预测准确度。

关键词:缺失值填充;K最近邻;保局投影;重构

中图分类号:TP181 **文献标志码**:A **文章编号**:1001-6600(2015)04-0055-08

0 引言

数据缺失在实际的数据应用过程中十分常见,因为实际的数据存在着大量噪声、冗余、缺失的问题[1],这些不完整、不准确的数据属性值,会影响模式识别和数据挖掘形成模型的正确性、导出规则的准确性,所以在机器学习、数据挖掘、模式识别等领域的应用过程中,很大一部分时间是用在数据的预处理上的[2]。

处理缺失值的方法大致可分为4种:删除存在缺失属性值的对象;不作处理直接使用;部分缺失值填充[3]和缺失值填充。通常认为缺失值填充是比较合理,且常用的处理缺失值的方法。这是因为它最大程度地利用样本自身信息。而且,无论是数据挖掘领域还是统计领域,基于KNN的填充方法是一种应用比较广泛的方法。

近年来,许多学者在缺失值填充领域做了大量的研究工作,使用机器学习的方法占据很大比重。其中有:基于多层感知器的人工神经网络方法[4],这是一种自动填充缺失值的方法;IHDIFC方法[5],其是基于属性选择和聚类分析的方法,可以解决在高维数据中缺失数据的填充;还有一些工作考虑样本属性之间的关系,如文献[6]利用数据间属性值存在的模式来进行缺失值填充;另外还有基于统计学方法的研究,如文献[2]解决有关混合属性的缺失值数据,对于离散和连续的目标缺失值提出两个相容估计量,并引入了mixture-kernel-based的迭代估计量。

传统的KNN缺失值填充方法只是单一地使用各个样本数据,并没有考虑到样本之间的相关性。事实上,样本之间具有相关性,如果模型能利用这种相关性,就能够获得更好的学习效果[7]。另外,在学习过程中,清除噪声样本,避免他们对学习的影响是必要的[8-9]。LPP(保局投影)[10]认为,保持算法前后数据自身具有的结构,能保留样本更多的信息,通常可以取得更显著的学习效果。另外,KNN算法K值的确定是个公开性问题,传统的K值由用户自定义给定或者通过十折交叉法得到[1],但这种方法麻烦,并且没

收稿日期:2015-03-16

基金项目:国家自然科学基金资助项目(61170131,61263035,61363009);国家863计划资助项目(2012AA011005);国家973计划资助项目(2013CB329404);广西自然科学基金资助项目(2012GXNSFGA060004,2015GXNSFAA139306);广西八桂创新团队和广西百人计划资助项目

通信联系人:苏毅娟(1976—),女,广西灵川人,广西师范学院副教授。E-mail:syj76@163.com

有考虑到数据的结构。为此,本文首先通过重构,并在重构过程中利用样本间的相关性,考虑保持数据结构,并利用LPP,应用$l_{2,1}$范式去除噪声样本的方法确定K值。然后对每一个缺失样本,利用得到的K值,对缺失属性值进行填充。本文定义这种方法为Reconstruction KNN,简称R-KNN方法。

1 相关背景知识

1.1 KNN算法描述

K最近邻方法(KNN)是一种基于非参数的多元变量判别方法,最早由Fix和Hodge以及Cover和Hart等人提出,是一种基于实例的惰性学习方法,广泛应用于数据挖掘、机器学习和模式识别等领域。K最近邻通过将给定的测试元组与和它相似的训练元组进行比较来学习,找出最接近未知元组的K个训练元组[11-12]。

KNN算法步骤介绍如下:

① 每个训练样本表示为向量:$(x_1,x_2,\cdots,x_r,\cdots,x_n)^{\mathrm{T}}$,其中$[x]_r$是样本的第$r$个属性值,$n$是样本的维数。

② 给定一个测试实例x_q,一般用欧氏距离衡量,即:

$$d(x_i,x_j)=(x_i,x_j)=\sqrt{\sum_{r=1}^{n}(x_i-x_j)^2}。\tag{1}$$

找出与实例x_q距离最近的K个实例,记为$x'_1,x'_2,\cdots,x'_K$。

③ 计算上述K个最近邻目标值的均值,即$f(x_q)$的估计为:

$$\hat{f}(x_q)\leftarrow\frac{\sum_{i=1}^{K}f(x'_i)}{K}。\tag{2}$$

1.2 LPP算法描述

LPP(locality preserving projections)又叫保局投影,是非线性方法,且为拉普拉斯特征映射的线性近似,它的本质特点是既能获得新的数据样本点在较低维度的投影,又能保持原始数据非线性的流形特征,它的目标是保证原始数据空间上相邻的数据点,在投影后的空间上也保持相应的相邻关系。所以说,它是一种能较好地保持非线性流形中局部数据特征的线性流形学习算法[7]。

假设输入样本为:$\boldsymbol{X}=\{\boldsymbol{x}_i\}_{i=1}^{n}\in\mathbf{R}^m$,$\boldsymbol{x}_i$代表$n$维空间里的一个点。投影矩阵为:$\boldsymbol{W}=\{\boldsymbol{w}_i\}_{i=1}^{m}$,其中$\boldsymbol{w}_i$表示低维空间,彼此之间独立且不为0。

通常使用最小化下面的目标函数来确定变换阵:

$$\min\sum_{ij}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_j)^2s_{ij}。\tag{3}$$

其中$\boldsymbol{S}$为权值矩阵,可以简单地取值为1或者0,$\boldsymbol{S}$中每个元素为$s_{ij}=\exp\left(\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{t}\right)$,其中$t$是一个大于0的常量。对式(3)进行变换,其中令$\boldsymbol{y}_i=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i$,则有:

$$\sum_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2s_{ij}=\sum_{ij}\boldsymbol{y}_i^2s_{ij}-2\sum_{ij}\boldsymbol{y}_i\boldsymbol{y}_is_{ij}+\sum_{ij}\boldsymbol{y}_i^2s_{ij}=$$
$$2\sum_{i}\boldsymbol{y}_i^2d_{ii}-2\sum_{ij}\boldsymbol{y}_i\boldsymbol{y}_js_{ij}=2\,\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{S})\boldsymbol{y}=2\,\boldsymbol{y}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{y}。\tag{4}$$

其中$\boldsymbol{D}=[d_{ij}]$是一个对角阵,$d_{ii}=\sum_j s_{ij}$叫做拉普拉斯矩阵。

2 R-KNN算法描述

2.1 算法描述

假设给定训练集$\boldsymbol{X}\in\mathbf{R}^{d\times n}$和测试集$\boldsymbol{Y}\in\mathbf{R}^{d\times m}$,其中$d$是样本维数,$n$、$m$是样本数量。本文希望寻找

到一个投影变换矩阵 $\boldsymbol{W} \in \mathbf{R}^{m\times n}$，使得经过 $\boldsymbol{W}$ 变换后的 $\boldsymbol{X}$ 可以尽可能接近 $\boldsymbol{Y}$，显然这种关系使用最小二乘损失函数模型表示更加合理，即：

$$\min_{\boldsymbol{W}} \| \boldsymbol{Y}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X} \|_F^2。 \tag{5}$$

考虑到式(5)是一个凸函数，它的解可表示为：$\boldsymbol{W}^* =(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$。但是$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$存在不可逆的问题，通常情况下是考虑利用岭回归引入一个 l_2 范数来解决[13]：

$$\min_{\boldsymbol{W}} \| \boldsymbol{Y}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X} \|_F^2+\rho \| \boldsymbol{W} \|_2^2。 \tag{6}$$

此时公式(6)的解为：$\boldsymbol{W}^* =(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\rho\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$，然而 $\boldsymbol{W}$ 是实数矩阵，应用到 KNN 中，K 通常取 n。但是正如本文引言提到的数据中通常存在的噪声样本的问题，$K=n$ 就是选取了所有的样本数据，这显然是不合理的。为此本文将基于岭回归的 l_2 范数改为 $l_{2,1}$ 范数，去除非相关性噪声样本，同时引入 LPP(保局投影)算法保持样本自身的流形学结构。利用测试样本和训练样本的条件属性重构测试样本缺失的决策属性，这样就利用了样本之间的相关性。故我们所用的模型如下：

$$\min_{\boldsymbol{W}} \| \boldsymbol{Y}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X} \|_F^2+\rho_1 \| \boldsymbol{W} \|_{2,1}+\rho_2\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W})。 \tag{7}$$

其中：$\boldsymbol{X}$ 为训练样本；$\boldsymbol{Y}$ 为测试样本；ρ_1 和 ρ_2 是两个调整参数。模型中 ρ_1 控制 $\boldsymbol{W}$ 阵行的稀疏性，即它的值越大时行为 0 的数量增加，值越小时行为 0 的数量减少，通过合适的 ρ_1 产生合适的 $\boldsymbol{W}$ 阵，为零的行用来去除噪声样本。ρ_2 是 LPP(保局投影) 部分用来保持样本流行结构不变的，具体来说 ρ_2 是用来控制 LPP 部分的数量级和 $l_{2,1}$ 范式部分的数量级保持一致的。另外，考虑到 $\boldsymbol{W}$ 中的元素值大小一般是不相同的，并且相应的训练样本和测试样本之间的相关度大小也是不同的，本文利用加权平均算法[14-15]计算 $\boldsymbol{W}$。由 $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$ 知 $\boldsymbol{W}^{\mathrm{T}}$ 中元素表示第 i 个训练样本与第 j 个测试样本之间的相关性大小，$\boldsymbol{W}_{ij}^{\mathrm{T}}>0$ 时代表正相关，$\boldsymbol{W}_{ij}^{\mathrm{T}}<0$ 时代表负相关，$\boldsymbol{W}_{ij}^{\mathrm{T}}=0$ 时代表不相关。矩阵 $\boldsymbol{W}^{\mathrm{T}}$ 表示所有测试样本与全部训练样本的相关性，此时 $\boldsymbol{W}^{\mathrm{T}}$ 的列代表训练样本，行代表测试样本。列向量 $\boldsymbol{W}_{(:,j)}^{\mathrm{T}}$ 表示第 j 个测试样本与全部训练样本的相关性，若列向量中有 r 个元素值不为0，即不为0的行数为 r，则预测时 K 应取 r，并用对应的这 r 个训练样本来预测测试样本。

举例说明算法的过程。假设输入一个样本，程序使用十折交叉法将其分为 10 组，9 组做训练样本，1 组做测试样本。利用训练样本重构测试样本寻找关系矩阵 $\boldsymbol{W}$，对 $\boldsymbol{W}$ 进行加权平均计算，将重构得到的值和测试样本对应的值进行相关性计算，找出效果最佳的 $\boldsymbol{W}$，假设通过模型得到一个 5×4 的矩阵：

$$\boldsymbol{W}^{\mathrm{T}}=\begin{pmatrix} 0.2 & -0.1 & 0.3 & -0.8 \\ 0 & 0 & 0 & 0 \\ 0.1 & 0.5 & 0.7 & 0.6 \\ 0.6 & 0.4 & 0.8 & 0.9 \\ 0 & 0 & 0 & 0 \end{pmatrix}。$$

这里为 0 的行对应的训练样本是噪声样本，重构时，利用 $\boldsymbol{W}^{\mathrm{T}}$ 为 0 的行与之相乘就忽略掉了这些样本，利用余下的样本进行重构，此时可以确定不为 0 的行数 3 便是 K 的值。

程序的主要步骤归纳如下：

① 首先采用十折交叉法将样本进行分组，每组有 9 份样本做训练样本集，1 份做测试样本集。

② 利用每一个训练样本和测试样本对应的属性关系寻找 $\boldsymbol{W}$。

③ 对得到的 $\boldsymbol{W}$ 进行加权平均计算。

④ 将重构得到的值和测试样本对应的值进行相关性计算，找出效果最佳的 $\boldsymbol{W}$。

⑤ 利用得到的加权后的 $\boldsymbol{W}$ 对测试样本属性进行重构，得到相应的缺失值。

2.2　优化问题

公式(7)是一个凸但非光滑的函数，本文通过设计一种近似加速梯度法来解决这个问题[16]。首先，本文通过如下变化在公式(7)上利用近似加速梯度方法：

$$f(\boldsymbol{W})=\frac{1}{2} \| \boldsymbol{Y}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X} \|_F^2+\rho_2\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}), \tag{8}$$

$$\vartheta(\boldsymbol{W})=f(\boldsymbol{W})+\rho_1\|\boldsymbol{W}\|_{2,1}。\tag{9}$$

这里指出:$f(\boldsymbol{W})$ 是凸的可微的,$\rho_2\|\boldsymbol{W}\|_{2,1}$ 是凸的但非光滑的。为了利用近似加速梯度法来优化 $\boldsymbol{W}$,我们利用如下的优化准则来迭代更新 $\boldsymbol{W}$:

$$\boldsymbol{W}(t+1)=\arg\min_{\boldsymbol{W}}\boldsymbol{G}_{\eta(t)}(\boldsymbol{W},\boldsymbol{W}(t)),\tag{10}$$

$$\boldsymbol{G}_{\eta(t)}(\boldsymbol{W},\boldsymbol{W}(t))=f(\boldsymbol{W}(t))+\langle\nabla f(\boldsymbol{W}(t)),\boldsymbol{W}-\boldsymbol{W}(t)\rangle+\frac{\eta(t)}{2}\|\boldsymbol{W}-\boldsymbol{W}(t)\|_F^2+\rho_1\|\boldsymbol{W}\|_1,\tag{11}$$

$$\nabla f(\boldsymbol{W}(t))=(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+\rho_2\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}})\boldsymbol{W}(t)-\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}。\tag{12}$$

这里的 $\eta(t)$ 和 $\boldsymbol{W}(t)$ 分别是调优参数和 t 次迭代后的 $\boldsymbol{W}$ 值。

通过忽略公式(10)中 $\boldsymbol{W}$ 的条件独立性,可以把公式(10)重新改写为:

$$\boldsymbol{W}(t+1)=\pi_{\eta(t)}(\boldsymbol{W}(t))=\arg\min_{\boldsymbol{W}}\frac{1}{2}\|\boldsymbol{W}-\boldsymbol{U}(t)\|_2^2+\frac{\rho_1}{\eta(t)}\|\boldsymbol{W}\|_{2,1}。\tag{13}$$

其中 $\boldsymbol{U}(t)=\boldsymbol{W}(t)-\frac{1}{\eta(t)}\nabla f(\boldsymbol{W}(t))$ 和 $\pi_{\eta(t)}(\boldsymbol{W}(t))$ 是 $\boldsymbol{W}(t)$ 在凸集 $\eta(t)$ 上的欧几里德投影。由于 $\boldsymbol{W}(t+1)$ 在每一行的可分性,可以分别为每一行更新权重:

$$\boldsymbol{w}^i(t+1)=\arg\min_{\boldsymbol{w}^i}\frac{1}{2}\|\boldsymbol{w}^i-\boldsymbol{u}^i(t)\|_2^2+\frac{\rho_1}{\eta(t)}\|\boldsymbol{w}^i\|_2,\tag{14}$$

其中 $\boldsymbol{u}^i(t)=\boldsymbol{w}^i(t)-\frac{1}{\eta(t)}\nabla f(\boldsymbol{w}^i(t))$。$\boldsymbol{u}^i(t)$ 和 $\boldsymbol{w}^i(t)$ 分别是 $\boldsymbol{U}(t)$ 和 $\boldsymbol{W}(t)$ 的第 i 行,根据公式(12),$\boldsymbol{w}^i(t+1)$ 的解决方法为:当 $\|\boldsymbol{u}^i(t)\|_2^2>\frac{\rho_1}{\eta(t)}$ 时,$\boldsymbol{w}^{i*}=\left(1-\frac{\rho_1}{\eta(t)\|\boldsymbol{u}(t)^i\|_2^2}\right)\boldsymbol{u}^i(t)$,其他情况下值为0。

同时,为了加快公式(10)的近似梯度法,我们引入一个辅助变量 $\boldsymbol{V}(t+1)$:

$$\boldsymbol{V}(t+1)=\boldsymbol{W}(t)+\frac{\alpha(t)-1}{\alpha(t+1)}(\boldsymbol{W}(t+1)-\boldsymbol{W}(t)),\tag{15}$$

这里系数 $\alpha(t+1)$ 通常取 $\frac{1+\sqrt{1+4\alpha(t)^2}}{2}$。

优化算法公式(7)的伪代码如下:

Input:$\eta(0),\alpha(1)=0,\gamma$;

Output:$\boldsymbol{W}$;

1. Initialize $t=1$;
2. Initialize $\boldsymbol{W}(1)$ as a random diagonal matrix;
3. Repeat
4. While $\boldsymbol{L}(\boldsymbol{W}(t))>\boldsymbol{G}_{\eta(t-1)}(\pi_{\eta(t-1)}(\boldsymbol{W}(t)),\boldsymbol{W}(t))$ do
5. Set $\eta(t-1)=\gamma\eta(t-1)$;
6. end
7. Set $\eta(t)=\eta(t-1)$;
8. Compute $\boldsymbol{W}(t+1)=\arg\min\boldsymbol{G}_{\eta(t)}(\boldsymbol{W},\boldsymbol{V}(t))$;
9. Compute $\alpha(t+1)=\frac{1+\sqrt{1+4\alpha(t)^2}}{2}$;
10. Compute Eq.(13);
11. Until Eq.(5) converges。

公式中用到的收敛定理如下:设$\{\boldsymbol{W}(t)\}$是由优化算法所生成的序列,然后对 $\forall t\geqslant 1$,式(16)成立:

$$\vartheta(\boldsymbol{W}(t))-\vartheta(\boldsymbol{W}^*)\leqslant\frac{2\gamma\boldsymbol{L}\|\boldsymbol{W}(1)-\boldsymbol{W}^*\|_F^2}{(t+1)^2}。\tag{16}$$

其中:γ 是一个正的预定义常数;$\boldsymbol{L}$ 是 $f(\boldsymbol{W})$ 的一个渐变 Lipschitz 常数;$\boldsymbol{W}^*=\arg\min_{\boldsymbol{W}}\vartheta(\boldsymbol{W})$。

此收敛定理表明，该加速近似梯度方法的收敛速度是 $O\left(\frac{1}{t^2}\right)$，$t$ 是迭代的次数。

3　实验结果分析

3.1　实验、数据集和评价指标介绍

本实验所用的 3 个数据集均来自 UCI 有关回归的数据集，因为需要对比填充值的效果，本文采用的是完备的实验数据。实验采用 Matlab2010b 软件，在 PC 机上进行编程操作。数据集信息统计如表 1。

表 1　数据集信息统计

Tab.1　Statistical dataset

数据集	数据类型	属性类型	样本个数	属性个数
housing	多变量	实数	500	13
bodyfat	多变量	实数	250	15
slump	多变量	实数	100	10

我们对 3 个数据集同时使用传统 KNN、基于属性信息熵的 Entropy-KNN[17] 和本文的 R-KNN 算法进行比较，比较的性能指标主要是统计学中常用的两个指标：相关系数和均方根误差(RMSE)。其中相关系数是用来衡量填充的属性值与测试集中对应的实际属性值之间相关关系的性能指标，其值越大，代表相关性越紧密。均方根误差(RMSE)用来衡量填充值的准确性，其值越大代表填充值与实际值越接近。

另外，R-KNN 算法中 ρ_1 和 ρ_2 是两个调整参数，它们的取值控制着算法结果的准确性，其中 ρ_1 控制 $\boldsymbol{W}$ 阵行的稀疏性，产生合适的 $\boldsymbol{W}$ 阵；ρ_2 是用来控制正则项 $\|\boldsymbol{Y}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\|_{\mathrm{F}}^2$ 与损失函数 $\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W})$ 的数量级保持一致的。3 个数据集参数设置过程大致相同，现以数据集 housing 实验时参数设定为例加以说明。

通过初步实验，我们发现参数 ρ_1、ρ_2 的值大致分布在$[10^{-5},10^{-1}]$时 $\boldsymbol{W}$ 的稀疏性比较合适，同时能保证正则项和损失函数数量级大致相同。当 ρ_1 大于 10^{-1}时行稀疏性过大，也就是说去除了太多本来不是噪声的样本，使得填充的效果变差；ρ_1 小于 10^{-5}时行稀疏性过小，使得许多噪声样本参与重构过程。所以我们在这个区间内分别令 ρ_1、ρ_2 以 10^{-1}为步长比较每个组合(共计 25 种)中的填充效果，每个组合做 10 次实验取平均值绘制成图 1。

从图 1 中可以看出参数 ρ_1、ρ_2 在 10^{-3}附近时相关系数最大，RMSE 最小。实验在$[8\times10^{-4},1.2\times10^{-3}]$区间中随机地选取一个参数，另外两个数据集参数的选取方法和 housing 数据集类似。经实验发现，数据集 bodyfat 和 slump 的参数 ρ_1、ρ_2 范围分别在 10^{-3}和 5×10^{-3}内。

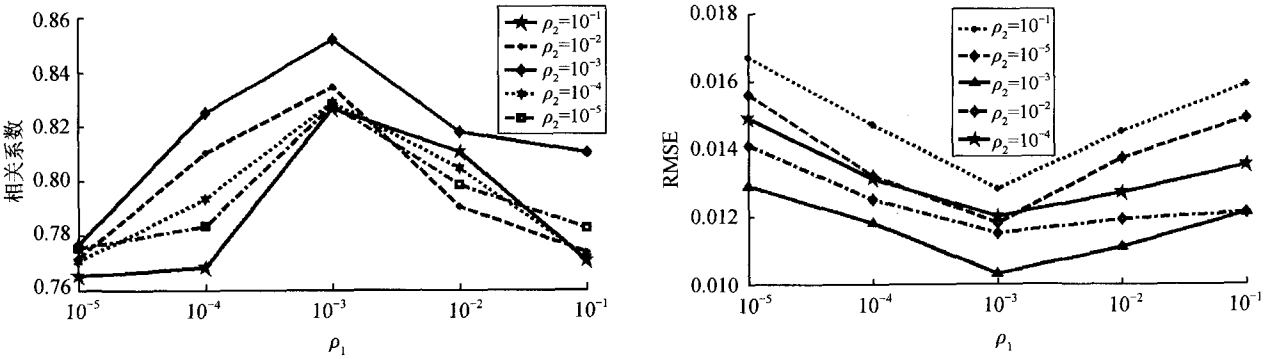

图 1　不同参数下的相关系数和 RMSE 统计图

Fig.1　Statistical correlation and RMSE results of different parameters

3.2 实验结果统计与分析

我们将每组数据使用十折交叉法,即将数据集分为10份,其中1份作为测试样本,9份作为训练样本,并将KNN算法以及Entropy-KNN算法的K值取定为5,在相同的实验条件下,进行10次缺失值填充,对比传统KNN算法、Entropy-KNN算法和R-KNN算法,得到3种算法相关系数和均方根误差(RMSE)的实验结果,将结果绘制成折线图,并统计两者的均值和方差,将统计结果制成表。

3个数据集3种算法所得的评价指标的均值和方差统计结果如表2、3所示。

表2 3种数据两种算法的相关系数统计结果

Tab.2 Statistical correlation results of three algorithms

算法	slump	bodyfat	housing
R-KNN	$0.811\ 3\pm3.6\times10^{-3}$	$0.996\ 8\pm2.137\ 9\times10^{-6}$	$0.852\pm5.121\ 3\times10^{-4}$
KNN	$0.601\ 3\pm1.87\times10^{-3}$	$0.983\ 2\pm5.282\ 1\times10^{-5}$	$0.776\ 7\pm2.7\times10^{-3}$
Entropy-KNN	$0.559\ 6\pm1.07\times10^{-3}$	$0.990\ 9\pm1.997\ 8\times10^{-5}$	$0.806\ 7\pm9.643\ 9\times10^{-4}$

表3 3种数据两种算法的RMSE统计结果

Tab.3 Statistical RMSE results of three algorithms

算法	slump	bodyfat	housing
R-KNN	$0.013\ 5\pm2.027\ 1\times10^{-5}$	$1.726\ 5\times10^{-4}\pm2.402\ 3\times10^{-10}$	$0.010\ 3\pm2.173\ 3\times10^{-7}$
KNN	$0.015\ 7\pm1.629\ 8\times10^{-5}$	$2.594\ 4\times10^{-4}\pm5.836\times10^{-9}$	$0.011\ 4\pm1.327\ 2\times10^{-6}$
Entropy-KNN	$0.017\ 1\pm2.937\ 3\times10^{-5}$	$2.105\ 6\times10^{-4}\pm1.409\ 8\times10^{-9}$	$0.012\ 1\pm8.622\ 8\times10^{-6}$

从表2、3可以看出,R-KNN算法、Entropy-KNN算法与传统的KNN算法相比,不仅性能指标更好,稳定性也相当好。

3个数据集上的3种算法所得到的相关系数折线图如图2所示,图中横坐标的实验序次表示的是我们利用十折交叉法做的10次对比实验,每次实验都进行一次缺失值的填充。从图2中可以看出,R-KNN算法所得的相关系数比传统的KNN算法所得的系数要大,这就说明R-KNN算法所填充的结果和实际值之间有更好的相关性。

3个数据集的3种算法所得到的均方根误差(RMSE)折线图如图3所示。从图3结果可以看出,R-KNN算法所得的均方根误差比传统的KNN算法以及Entropy-KNN算法所得结果要小,这就说明R-KNN算法所得的填充结果具有更好的准确性。

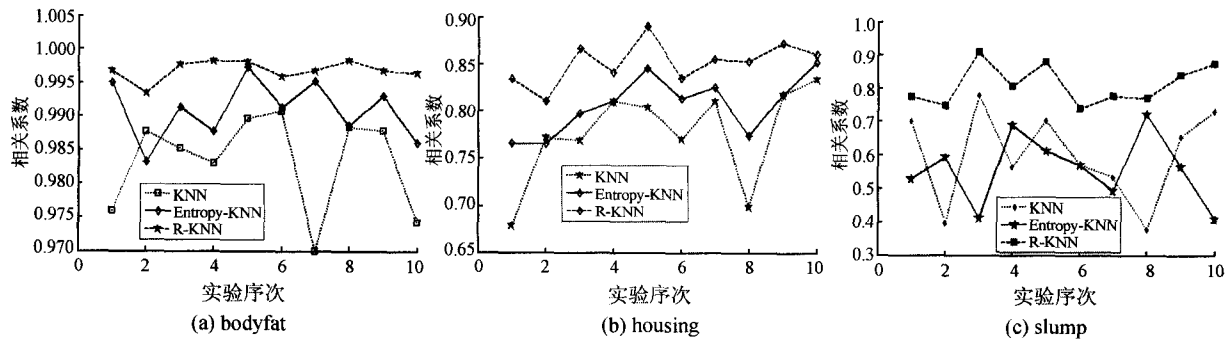

图2 3种数据两种算法得出的相关系数折线图

Fig.2 Correlation line chart of three algorithms

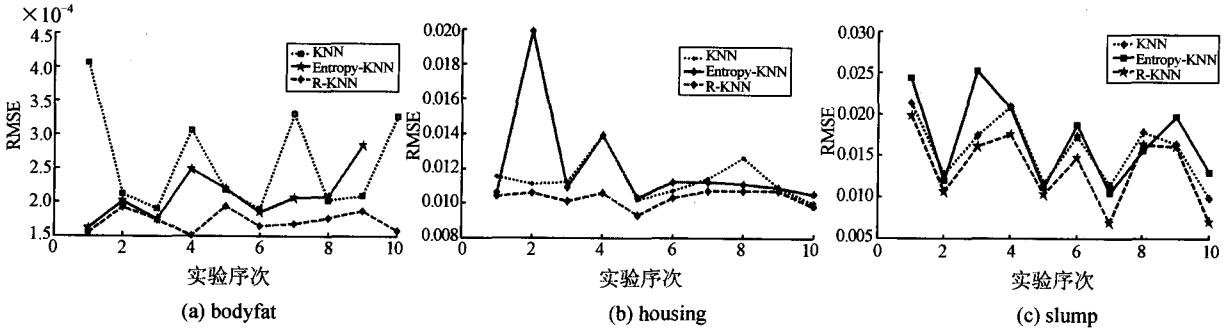

图 3　3 种数据两种算法得出的 RMSE 折线图

Fig.3　RMSE line chart of three algorithms

4　结语

本文提出了一种基于 LPP 和 $l_{2,1}$ 范式正则化因子，并考虑样本数据之间相关性的改进的 KNN 缺失值填充算法 R-KNN。该算法的特点是能够去除样本数据中的不相关的噪声样本，并保持数据本身的流行学结构不变，充分利用样本数据间的相关性重构出测试样本缺失值。仿真实验结果表明，该算法比传统的 KNN 填充算法以及 Entropy-KNN 算法有更好的性能。

参　考　文　献：

[1]　ZHANG Shi-chao, JIN Zhi, ZHU Xiao-feng. Missing data imputation by utilizing information within incomplete instances[J].Journal of Systems and Software,2011,84(3):452-459.

[2]　ZHU Xiao-feng, ZHANG Shi-chao, JIN Zhi, et al. Missing value estimation for mixed-attribute data sets[J]. IEEE Trans Knowl Data Eng,2011,23(1):110-121.

[3]　ZHANG Shi-chao, ZHANG Cheng-qi. Propagating temporal relations of intervals by matrix[J]. Applied Artificial Intelligence,2002,16(1):1-27.

[4]　SILVA-RAMIREZ E L, PINO-MEJIAS R, LOPEZ-COELLO M, et al. Missing value imputation on missing completely at random data using multilayer perceptrons[J].Neural Networks,2011,24(1):121-129.

[5]　BU Fan-yu, CHEN Zhi-kui , ZHANG Qing-chen, et al. Incomplete high-dimensional data imputation algorithm using feature selection and clustering analysis on cloud[EB/OL]. (2015-05-06) [2015-06-22]. http://link.springer.com/article/10.1007/s11227-015-1433-9.

[6]　RAHMAN M G, ISLAM M Z. FIMUS: a framework for imputing missing values using co-appearance, correlation and similarity analysis[J].Knowl Based Syst,2014,56:311-327.

[7]　ZHU Xiao-feng, HUANG Zi, SHEN Heng-tao, et al. Dimensionality reduction by mixed kernel canonical correlation analysis[J].Pattern Recognition,2012,45(8):3003-3016.

[8]　ZHU Xiao-feng, HUANG Zi, CHENG Hong, et al. Sparse hashing for fast multimedia search[J]. ACM Trans Inf Syst, 2013,31(2):9.

[9]　ZHU Xiao-feng, HUANG Zi, YANG Yang, et al. Self-taught dimensionality reduction on the high-dimensional small-sized data[J].Pattern Recognition,2013,46(1):215-229.

[10]　HE Xiao-fei, NIYOGI P. Locality preserving projections[C]//THRUN S, SAUL L K, SCHOLKOPF B. Advances in Neural Information Processing Systems 16. Cambridge, MA: MIT Press, 2004:153-160.

[11]　QIN Zhen-xing, WANG A T, ZHANG Cheng-qi, et al. Cost-sensitive classification with k-nearest neighbors[C]// Knowledge Science, Engineering and Management: LNCS Volume 8041. Berlin: Springer, 2013:112-131.

[12]　HAN Jia-wei, KAMBER M, PEI Jian. 数据挖掘概念与技术[M]. 范明，孟小峰，译. 北京：机械工业出版社，2012:

275-276.

[13] HASTIE T, TIBSHIRANI R, FRIEDMAN J.统计学习基础:数据挖掘、推理与预测[M].范明,柴玉梅,咎红英,译.北京:电子工业出版社 2004:28-32.

[14] 王超学,潘正茂,马春森,等. 改进型加权 KNN 算法的不平衡数据集分类[J].计算机工程,2012,38(20):160-163, 168.

[15] ZHANG Shi-chao,ZHU Man-long.Weighting imputation methods and their evaluation under shell-neighbor machine [C]//Proceeding of 9th IEEE International Conference on Cognitive Informatics. Piscataway, NJ: IEEE Press, 2010:874-879.

[16] ZHU Xiao-feng,HUANG Zi, CUI Jiang-tao , et al.Video-to-shot tag propagation by graph sparse group lasso[J]. IEEE Transactions on Multimedia,2013,15(3): 633-646.

[17] 童先群,周忠眉.基于属性值信息熵的 KNN 改进算法[J].计算机工程与应用,2010,46(3):115-117.

KNN Imputation Algorithm Based on LPP and $l_{2,1}$

SU Yi-juan[1] ,SUN Ke[2,3] ,DENG Zhen-yun[2,3] ,YIN Ke-jun[2,3]

(1.College of Computer and Information Engineering,Guangxi Teachers Education University, Nanning Guangxi 530023,China;2. College of Computer Science and Information Technology, Guangxi Normal University,Guilin Guangxi 541004, China; 3.Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University,Guilin Guangxi 541004, China)

Abstract: Traditional KNN missing data filling algorithm does not utilize the correlation between the properties of samples, Neither considers but also does not consider to maintain the sample structures and removes noise samples. In this paper, a method of using training samples to reconstruct the test sample is proposed, which is used for the nearest neighbor missing data imputation. The method makes full use of the correlation between samples, uses the LPP (locality preserving projection) to maintain the data structure in the process of reconstruction, and uses $l_{2,1}$ norm to remove noise samples. Simulation experiments on UCI data sets show that the proposed method has higher prediction accuracy than the traditional KNN algorithm and Entropy-KNN algorithm based on attribute information entropy.

Keywords: missing data imputation;KNN;LPP;reconstruction

(责任编辑 黄 勇)

阅读此文的还阅读了：

26. 基于局部加权的Citation-kNN算法

27. 基于改进KNN的话题跟踪算法

28. 基于MCL与KNN的混合聚类算法

29. 基于KNN算法的MRR数据分类研究

30. 基于KNN模型的增量学习算法

31. kNN填充算法的分析和改进研究

32. 基于LPP和l2,1的KNN填充算法

33. 一种新颖的基于马氏距离的KNN分类算法

34. 基于KNN算法的OCR系统的设计

35. KNN-均值算法

36. KNN-均值算法

37. 基于kNN的聚类算法研究

38. 一种基于KNN的文本分类算法

39. 基于ELM特征映射的kNN算法

40. 基于改进kNN算法的人脸识别研究

41. KNN-均值算法

42. 基于Curvelet及LPP的人脸识别算法

43. 基于KNN算法的公路施工风险判别

44. 基于kNN算法实现中文手写数字识别

45. 基于CUDA的数据挖掘KNN算法的改进

46. 基于局部保持的KNN算法

47. 基于LPP-kNN方法的间歇过程故障监视

48. 基于局部相关性的kNN分类算法

49. 基于KNN算法的学生贷款风险模型

50. KNN-均值算法