稀疏编码的最近邻填充算法*

苏毅娟[1]，程德波[2a,2b†]，宗 鸣[2a,2b]，李 凌[2a,2b]，朱永华[3]

(1. 广西师范学院 计算机与信息工程学院，南宁 530023；2. 广西师范大学 a. 计算机科学与信息工程学院；b. 广西多源信息挖掘与安全重点实验室，广西 桂林 541004；3. 广西大学 计算机与电子信息学院，南宁 530004)

摘 要： 针对 K 最近邻填充算法(K-nearest neighbor imputation，KNNI)的参数 K 值固定问题进行了研究，发现对缺失值填充时，参数 K 值固定很大程度上影响了填充效果。为此，提出了基于稀疏编码的最近邻填充算法来解决这一问题。该算法是用训练样本重构每一缺失样本，在重构过程中充分考虑了样本之间的相关性；并用 ℓ_1 范数来学习确保每个缺失样本用不同数目的训练样本填充，以此解决 KNNI 算法参数 K 值选取问题。基于数据性能分析指标 RMSE 和相关系数的实验比较结果表明，该算法比 KNNI 算法的效果要好。该算法能很好地避免了 KNNI 算法存在的缺陷，适用于数据预处理环节需要对缺失值进行填充的应用领域。

关键词： 缺失值填充；稀疏编码；重构；均方根误差；相关系数；数据预处理

中图分类号： TP181；TP301.6 **文献标志码：** A **文章编号：** 1001-3695(2015)07-1942-04

doi:10.3969/j.issn.1001-3695.2015.07.005

K-nearest neighbor imputation based on sparse coding

Su Yijuan[1], Cheng Debo[2a,2b†], Zong Ming[2a,2b], Li Ling[2a,2b], Zhu Yonghua[3]

(1. *College of Computer & Information Engineering, Guangxi Teachers Education University, Nanning* 530023, *China*; 2. *a. School of Computer Science & Information Engineering, b. Guangxi Key Laboratory of Multi-source Information Mining & Security, Guangxi Normal University, Guilin Guangxi* 541004, *China*; 3. *School of Computer & Electronics Information, Guangxi University, Nanning* 530004, *China*)

Abstract: Aimed at the parameter K fixed issues of K-nearest neighbor imputation (KNNI) algorithm, it was found that when impute the missing values, the fixed value of the parameter K resulted in a large extent influence to the imputation effect. Therefore, this paper proposed the K-nearest neighbor based on sparse coding (KNNI-SC) algorithm to solve this problem. This method reconstructed each missing sample with the training samples, fully considering the correlation between samples in the reconstruction process. And it used an ℓ_1 norm to learn to ensure each missing sample was imputed by different number of training samples, so it solved the parameter K selection problem of KNNI algorithm. Performance comparison based on the data analysis of the experimental results indicators RMSE and correlation coefficients show that the algorithm is better than KNNI algorithm. The algorithm can well avoid the defects of KNNI algorithm, it is available to data preprocessing step that needs missing values imputation's applications.

Key words: missing value imputation; sparse coding; reconstruct; RMSE; correlation coefficient; data preprocessing

0 引言

在数据挖掘和机器学习应用中，经常因一些原因使数据不能完整无缺地获取，造成数据缺失。如有些信息暂时无法获取、数据被遗漏、不能正常收集的信息、获取某些信息的代价太大等都可能导致数据缺失[1,2]，而且在一些工业领域缺失值的比率高达 80% 以上[3,4]。数据缺失会影响到从中抽取规则的正确性和运行性能，甚至导致建立错误的数据挖掘模型。因此，缺失值填充[5]是数据挖掘和机器学习领域中一个实际而富有挑战性的问题。研究者已经提出了各种处理缺失值的方法，如最近似值替换缺失值方法，以及贝叶斯网络、神经网络、粗糙集理论方法等。其中，K 最近邻[6]填充算法(KNNI)因其原理简单、易于实现得到了极大的推崇。例如，在美国人口普查部和加拿大统计署就采用 KNNI 来处理缺失数据。KNNI 算法是一种基于实例的不需要先验知识、无师学习的懒散算法。其通常用缺失实例的最近 K 个无缺失的训练实例的均值或者中位数来填充缺失值。然而，KNNI 存在两个问题：a) K 比较难以取定，文献[7]建议 $K=\sqrt{n}\ (n>100)$，n 是数据集缺失数据的个数，但这种取法通常不能得到满意结果，而且该想法也没有理论保证；b) 固定 K 值，每一个缺失样本用同一个 K 值来填充使得填充效果低效。结合以下一个例子来分析 KNNI 存在的缺陷。

例如图 1 为一含有缺失的数据集，当 $K=1$ 时，缺失样本 1 和 2 直接用最靠近它们的实心圆的样本来填充即可。当 $K=3$ 时，缺失样本 1 用靠近它的三个实心圆的样本值均值来填充，而缺失样本 2 要用离它很远的两个完全样本和最近的一个样本取均值来填充，明显会造成一定的偏差。当 $K=7$ 时，缺失

收稿日期：2014-05-20；修回日期：2014-07-16 基金项目：国家自然科学基金资助项目(61170131，61263035 和 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)；广西八桂创新团队和广西百人计划资助项目；广西高校科学技术研究重点资助项目(2013ZD041)；广西研究生教育创新计划项目(YCSZ2015095，YCSZ2015096)

作者简介：苏毅娟(1976-)，女，广西桂林人，副教授，主要研究方向为机器学习和数据挖掘；程德波(1990-)，男(通信作者)，江西丰城人，硕士，主要研究方向为数据挖掘、机器学习(7294835098@qq.com)；宗鸣(1990-)，男，江苏泰州人，硕士，主要研究方向为机器学习、数据挖掘；李凌(1988-)，男，湖南衡阳人，硕士，主要研究方向为数据库、数据库安全；朱永华(1994-)男，广西桂林人，本科，主要研究方向为数据挖掘.

样本 1、2 都将会用离它们很远的样本来填充,这样偏差将会更明显。因此,如果 K 值选取得不合适,填充缺失样本时将会直接导致填充的低效。

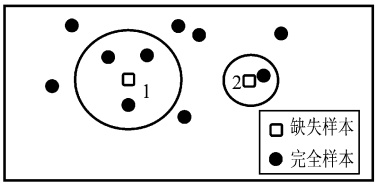

图 1　缺失样本数据集

对于上述存在的问题,本文借用稀疏编码理论和重构的相关知识运用于 KNNI 算法,以此来避免 KNNI 算法的缺陷。

1　KNNI-SC 算法

1.1　相关理论

稀疏学习理论[8,9] 将样本之间的系数权重作为一种自然鉴别信息引入模型,以此考虑样本数据分布结构信息和模型的鲁棒性,而且稀疏学习的正则化因子 ℓ_1 范数优化问题是一个凸优化,能保证得到唯一的全局最优解。因此稀疏编码理论[10,11] 在机器学习领域已经得到了快速的发展。

重构[12,13] 是指用一组线性无关向量的线性组合来近似地表达一个给定的向量。例如,设 $X=\{x_i\}_{i=1}^{n}$ 矩阵由一组列向量组成,其中 $x_i \in \mathrm{R}^d$,表示位于矩阵 X 的第 i 列的实向量; w_j 是数据点 x_j 对 x_i 点的重构系数(W 的第 j 个列向量) ,则重构可以用如下表达式来求解最优化问题。

$$W^* = \arg\min \| x_i - \sum_j x_j w_j^{\mathrm{T}} \| \tag{1}$$

通过重构,可以找到 x_i 一定半径内的数据点或者 K 个最近的数据点。根据以上理论,本文提出一种最优化方法——稀疏编码的最近邻缺失值填充算法(K-nearest neighbor imputation based on sparse coding,KNNI-SC) 。该算法: a) 将样本数据分成训练样本和测试样本; b) 用训练样本对测试样本进行重构获得系数矩阵; c) 借助稀疏编码理论,采用 ℓ_1 范数正则化因子惩罚目标函数,使得重构系数矩阵稀疏。若稀疏使得重构系数为 0,则表示该训练样本不参与此测试样本的重构,而且重构系数子数的稀疏位置不同使得每个测试样本用不同的训练样本进行重构,保留了样本之间的相关性。

1.2　KNNI-SC 算法

缺失填充研究中[15-17] ,缺失可能发生在一般属性,也可能发生在决策属性。本文算法主要讨论的是决策属性缺失时进行填充。当一般属性缺失时,可以采用将缺失属性归结为决策属性来考虑,即将缺失属性作为决策属性,其他无缺失属性作为一般属性,再运用 KNNI-SC 算法同样可以达到近似填充的效果。假设有样本 $X=\{x_i\}_{i=1}^{n} \in \mathrm{R}^{n\times d}$,$n$ 为行数,d 为维数; 有测试样本 $Y=\{y_i\}_{i=1}^{m} \in \mathrm{R}^{m\times d}$,$m$ 为行数,d 为维数。通常使用最小二乘损失函数求解线性问题,即

$$\arg\min_W \| W^{\mathrm{T}}X - Y \|_F^2 \tag{2}$$

其中: $\| A \|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2$(矩阵 A 全部元素的平方和) 为 Frobenius 范数; $\| W^{\mathrm{T}}X - Y \|_F^2$ 为重构误差。$W \in \mathrm{R}^{n\times m}$ 为 X 对 Y 进行重构后的重构系数矩阵。式(2) 表达的目标函数是凸的,易知其解为 $W^* = (XX^{\mathrm{T}})^{-1}XY$。在实际应用中 XX^{T} 不一定可逆,为此加上一正则化因子 $\| W \|_2^2$ 使得函数可逆。所以函数转变为如下目标函数:

$$\arg\min_W \| W^{\mathrm{T}}X - Y \|_F^2 + \varphi \| W \|_2^2 \tag{3}$$

其中: $\| W \|_2^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} |w_{ij}|^2$; φ 为 ℓ_2 范数正则化因子参数。此目标函数的解为 $W^* = (XX^{\mathrm{T}} + \varphi I)^{-1}XY$,但由于目标函数式(3) 得到的 W 不稀疏,不能解决本文问题,为此本文采取将函数的正则化因子 $\| W \|_2^2$ 换成惩罚函数 ℓ_1 范数 $\| W \|_1$ 得到如下目标函数:

$$\arg\min_W \| W^{\mathrm{T}}X - Y \|_F^2 + \lambda \| W \|_1 \tag{4}$$

其中: $\| W \|_1 = \sum_{i=1}^{n}\sum_{j=1}^{m} |w_{ij}|$; λ 为 ℓ_1 范数正则化因子参数用来惩罚函数,λ 值越大表示受到的惩罚越重,从而得到的 W 越稀疏。目标函数式(4) 即为 the least absolute shrinkage and select operator(Lasso) [18,19] 。Lasso 用 ℓ_1 范数代替目标函数式(2) 是通过 ℓ_1 范数惩罚目标函数使得绝对值较小的系数自动压缩为 0,从而产生稀疏模型。并且参数 λ 越大得到的稀疏密度越大,即 W 为 0 的元素越多,反之亦然。而参数 λ 的确定一般通过 10 折交叉验证方法获得。

在得到的 W 中有 m 个列向量,每个列向量有 n 个元素,因此,每列就是表示某个测试样本跟训练样本的关系。例如,w_{ij} 的值表示第 i 个测试样本跟第 j 个训练样本的相关性。若 $w_{ij} > 0$,说明第 i 个测试样本跟第 j 个训练样本正相关,且相关值越大,说明第 j 个测试样本对第 i 个训练样本越重要; 若 $w_{ij} = 0$,表示它们不相关; 若 $w_{ij} < 0$,说明它们负相关。从缺失值填充方面考虑,如果 $w_{ij} \neq 0$,表明第 i 个测试样本跟第 j 个训练样本相关。因此,第 i 个测试样本的缺失数据由这些测试样本来填充(或者成为第 i 个样本的最近邻) 是合理的。例如

$$W = \begin{pmatrix} 0 & 0.1 & -0.6 \\ 0.5 & 0 & 0.1 \\ 0.1 & 0 & 0 \end{pmatrix}$$

W 包含 3 个测试样本和 3 个训练样本。第一列的数字表明第一个测试样本跟第二个和第三个训练样本相关,因此,填充第一个测试样本的时候可以考虑设置 $K=2$; 同理,填充第二个和第三个样本的 K 值依次应为 1 和 2。

本文通过 Lasso 来研究 KNNI 的 K 值固定问题。本文算法 KNNI-SC 虽然使用跟 Lasso 一样的目标函数,但是与 Lasso 有两个区别: a) Lasso 一般用于分类或者回归等应用,本文创新地利用 Lasso 解决 KNNI 算法的 K 值固定值问题; b) 不同于使用 least angle regression (LARS) 方法求解 Lasso 的目标函数,本文采用 alternating direction method of multipliers(ADMM) [20] 算法来优化目标函数。这是因为 ADMM 显示收敛速度比 LARS 更快,收敛效果更好。

KNNI-SC 跟常见的 KNNI 算法比较,a) KNNI 是懒惰学习方法,对每一个测试样本单独地求 KNN 中的 K 值,KNNI-SC 一次重构所有测试样本,既考虑了测试样本的相关性也可以考虑训练样本的相关性; b) KNNI 算法采用用户自定义或者 10 折交叉验证方法获取 K 值,且 K 值通常对所有测试例子一样,而本文的 KNNI-SC 通过 Lasso 重构方法首先得到测试样本和训练样本之间的相关性矩阵 W。学习 W 的过程是数据驱动(data-driven) ,因此能确保关系是最优的,通过 Lasso 的 ℓ_1 范数使得每个测试样本需要的 K 值不同。实验也验证,本文算法是优于 KNNI 方法的。

1.3　填充算法

与一般 KNNI 算法类似,通过最近的 K 个 y_i 的决策属性

$f(x_i)$ 的均值来对缺失值进行填充。因稀疏矩阵 W 系数子数 w_{ij} 代表着 x_i 对 y_i 的相关度，利用决策属性 $f(x_i)$ 来填充缺失值时可用 W 对其加权，从而使得填充时更精准。

稀疏权重矩阵 ω 是由 W 中每列不为 0 的系数子数之和与所有系数子数之和的比值。由如下公式计算:

$$\omega = \frac{\sum_{i=1}^{n} w_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}} \tag{5}$$

因此由权重系数矩阵 ω 可以构造线性填充模型如下:

$$\hat{f}(x) = \omega * f(x) \tag{6}$$

算法 1 KNNI-SC 算法

输入: 读取数据，并作规范化处理。

输出: 填充值。

1 通过 ADMM 算法求得最优 W;

2 根据 W 来确定每个测试样本的 K 值;

3 根据 K 值通过式(5)加权，用式(6)对缺失值进行填充;

4 返回填充值。

1.4 算法优化

因目标函数式(4)是凸非平滑的，用 ADMM 算法来优化目标函数。首先将式(4)分解成如下 N 个独立的子问题:

$$\arg\min_{w_i} \| y_i - x_i w_i \|_F^2 + \lambda \| w_i \|_1 \tag{7}$$

其中: w_i 向量是 W 的第 i 个列子向量且 $\mathrm{vec}(W) = [w_1, \cdots, w_n]^{\mathrm{T}}$。

式(4)为 Lasso 回归，使用了约束的参数矢量 ℓ_1 范数，因为 W 中的每个元素不能独立地进行处理，故在公式中应用的软阈值 λ 是不平凡且低效的。然而可以将其应用在优化问题，这时称其为乘数交替方向法。此方法的一般形式可以采用虚变量使得目标成为

$$\arg\min \| Y - XW \|_F^2 + \lambda \| C \|_1 + \frac{\rho}{2} \| W - C \|_F^2 \quad \text{s.t. } W = C \tag{8}$$

此目标函数较复杂，使用扩展拉格朗日函数代替式(8)得到如下模型:

$$L(W, C, \Lambda) = \| Y - XW \|_F^2 + \lambda \| C \|_1 + \frac{\rho}{2} \| W - C \|_F^2 + \mathrm{vec}(\Lambda)\,\mathrm{vec}(W - C) \tag{9}$$

ADMM 算法采用迭代法的基本思想，包括如下迭代步骤:

1 $W^{k+1} = \arg\min_W L(W, C^k, \Lambda^k)$

2 $C^{k+1} = \arg\min_C L(W^{k+1}, C, \Lambda)$

3 $\Lambda^{k+1} = \Lambda^k + \rho(W + C)$

通过 ADMM 算法使得能够将式(4)的问题拆分成两个子问题。

首先分析第一个子问题。如果只最小化式(8)中的 W，那么当 ℓ_1 惩罚 $\| C \|_1$ 使得 W 从目标函数里面消失时，这将子问题 1 转换成为一个非常有效、简单的最小二乘回归问题。然后分析第二个子问题。如果只最小化式(8)中的 C，那么当 $\| Y - XW \|_F^2$ 消失时，允许 C 通过每个元素独自求解。通过这两个子问题，本文能够有效地使用软阈值 λ。W 和 C 的当前估计与 ADMM 算法的第 3 步相结合，以此来更新当前估计的拉格朗日乘子矩阵 Λ。此处引入的目标罚参数 ρ 在这里起着特殊的作用，即允许建立一个有缺的估计 Λ 来求解 W 和 C。

通过以上分析可以将式(4)拆分成 N 个独立的子问题，再使用 ADMM 算法求解最优化的 W。

算法 2 ADMM 算法。

输入: 数据集、惩罚参数 ρ。

输出: W、Λ。

1 initialize W^0, C^0, Λ^0

2 repeat

3 $W^{k+1} \leftarrow W^k$; $C^{k+1} \leftarrow C^k$

4 $\Lambda^{k+1} \leftarrow \Lambda^k + \rho(W + C)$

5 $k = k + 1$

6 until W 最优

2 实验与分析

2.1 实验评价指标

实验通过 KNNI-SC 算法构造模型填充每个缺失样本，同时用 KNNI 算法填充每个缺失样本，用计算出的相应的评价指标值比较 KNNI-SC 算法和 KNNI 算法。

评价指标是填充准确度，采用均方根误差(root mean square error，RMSE)来衡量[8]，RMSE 值越小填充的准确度就越高。均方根误差是填充值与真实值偏差的平方和缺失值个数 n 比值的平方根:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

同时，为了衡量填充数据和测试数据的相关密切程度，本文还进行了相关性分析。相关系数(correlation coefficent)的大小一般在(-1,1)，在(-1,0)时表示负相关，0 表示不相关，(0,1)时表示正相关，且相关系数越接近 1 表示填充值越接近于测试值。如果显著性越小，则偏离真值越小。

实验部分使用的数据集来源于文献[21,22]，因为需要测试算法的填充正确性及有效性，所以选取的都是无缺失的完备数据集。实验数据集如表 1 所示。

表 1 用于实验的数据集

数据集	样本个数	属性维数	数据集	样本个数	属性维数
Housing	506	13	Mpg	329	8
Abalone	4 177	8	Pyrim	74	28

实验用 10 折交叉验证法，从样本数据中取出十分之一作为测试样本，余下的作为训练样本。然后调置好参数 λ，并对 KNNI 算法参数 K 置值为 5，用训练样本对每一测试样本进行重构。在同一索引号下分别使用 KNNI 算法和 KNNI-SC 算法对缺失样本进行填充，且计算出两个评价指标值。

2.2 实验结果分析

本文算法通过 MATLAB 语言编程，并在 Windows7 系统下的 MATLAB 7.1 软件上运行测试。通过采用表 1 数据集进行实验，每个数据集实验时重复十次，并取评价指标值的均值可以得到实验结果如表 2、3 所示。

表 2 KNNI-SC、KNNI 均方根误差比较(RMSE)

数据集	KNNI-SC	KNNI
Housing	3.6637 ± 0.2996	4.9460 ± 0.5861
Abalone	1.9633 ± 0.0776	2.2618 ± 0.1210
Mpg	2.8408 ± 0.2532	3.3663 ± 0.3747
Pyrim	0.0478 ± 0.0003	0.0718 ± 0.0002

表 3 KNNI-SC、KNNI 相关系数对比

数据集	KNNI-SC	KNNI
Housing	0.9206 ± 0.0007	0.8455 ± 0.0028
Abalone	0.7438 ± 0.0045	0.6473 ± 0.0092
Mpg	0.9337 ± 0.0004	0.8938 ± 0.0009
Pyrim	0.9068 ± 0.0037	0.7674 ± 0.0124

为了更直观地比较 KNNI-SC 与 KNNI 两个算法，对比 RMSE 实验结果分析如图 2 ~ 5 所示，对比相关系数实验结果分析如图 6 ~ 9 所示。

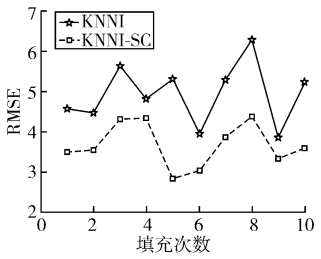

图 2　Housing 数据集实验 RMSE 对比

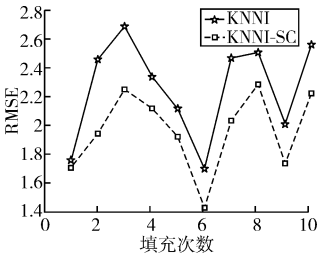

图 3　Abalone 数据集实验 RMSE 对比

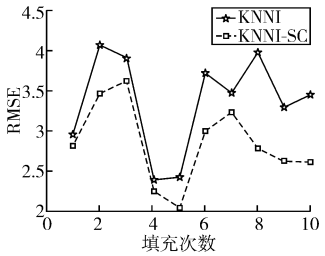

图 4　Mpg 数据集实验 RMSE 对比

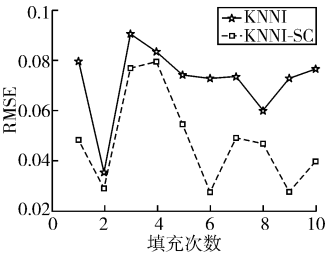

图 5　Pyrim 数据集实验 RMSE 对比

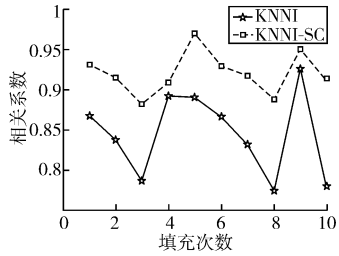

图 6　Housing 数据集实验相关系数对比

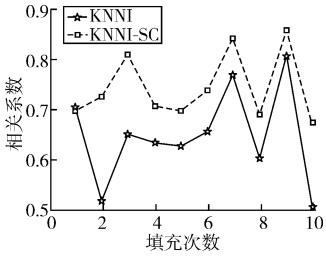

图 7　Abalone 数据集实验相关系数对比

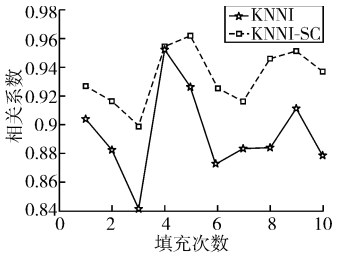

图 8　Mpg 数据集实验相关系数对比

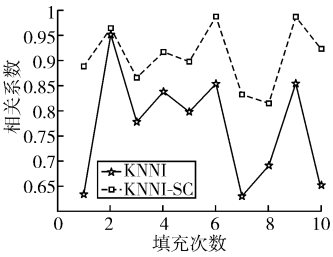

图 9　Pyrim 数据集实验相关系数对比

通过表 2 可以发现四个数据集在分别使用 KNNI-SC 算法填充时所得 RMSE 比一般 KNNI 算法填充时的 RMSE 小很多，表明 KNNI-SC 算法填充缺失值时比一般 KNNI 算法准确度高；由图 2 ~ 5 所示的四个数据集用两个算法填充后计算出的 RMSE 对比图可直观看出 KNNI-SC 折线在 KNNI 折线的下方，Housing 数据集的填充效果最明显（如图 2 所示，其中填充次数代表实验进行的次数，下同）。通过表 3 可以发现 KNNI-SC 算法得到的填充缺失值与测试样本的相关度比一般 KNNI 得到的填充缺失值相关度高，也就是说 KNNI-SC 算法填充的值呈测试值线性关系；由图 6 ~ 9 所示的四个数据集实验得到的相关系数对比图也可以明显看出 KNNI-SC 折线在 KNNI 折线的上方，Housing 数据集提高的相关性最为明显（图 6）。结合两个评价指标，表明 KNNI-SC 算法的填充效果比 KNNI 算法的填充效果要好。

3　结束语

本文提出的 KNNI-SC 算法，利用稀疏编码理论和重构技术解决了 KNNI 算法关于参数 K 的选取问题，通过以上技术充分考虑了样本之间的相关性。然而由于目标函数（4）是凸非光滑且不易求解，故本文采用 ADMM 算法将目标函数分解成若干个子问题来求解。本文通过采用 4 个数据集对 KNNI-SC 算法进行验证，并将 KNNI-SC 算法与 KNNI 算法进行比对，鉴于均方根误差和相关性分析的评价标准，实验结果表明，KNNI-SC 算法比 KNNI 算法的效果要好。

参考文献：

[1] Zhang Shichao, Jin Zhi, Zhu Xiaofeng. Missing data imputation by utilizing information within incomplete instances [J]. Journal of Systems and Software, 2011, 84(3): 452-459.

[2] Zhang Chengqi, Qin Yongsong, Zhu Xiaofeng, *et al*. Clustering-based missing value imputation for data preprocessing [C] //IEEE International Conference on Industrial Informatics. 2006: 1081-1086.

[3] Zhang Shichao, Jin Zhi, Zhu Xiaofeng, *et al*. Missing data analysis: a kernel-based multi-imputation approach [M] //Transactions on Computational Science III. Berlin: Springer, 2009: 122-142.

[4] Zhang Shichao, Qin Yongsong, Zhu Xiaofeng, *et al*. Optimized parameters for missing data imputation [C] //Proc of the 9th Pacific Rim International Conference on Articial Intelligence. 2006: 1010-1016.

[5] Zhang Chengqi, Zhu Xiaofeng, Zhang Jilian, *et al*. GBKII: an imputation method for missing values [C] //Proc of the 11th Pacific-Asia Conference. 2007: 1080-1087.

[6] Cover T, Hart P. Nearest neighbor pattern classification [J]. IEEE Trans on Information Theory, 1967, 13(1): 21-27.

[7] Lall U, Sharma A. A nearest neighbor bootstrap for resampling hydrologic time series [J]. Water Resources Research, 1996, 32(3): 679-693.

[8] Zhu Xiaofeng, Huang Zi, Cheng Hong, *et al*. Sparse hashing for fast multimedia search [J]. ACM Trans on Information System, 2013, 31(2): 1-24.

[9] Zhu Xiaofeng, Huang Zi, Shen Hengtao, *et al*. Dimensionality reduction by mixed kernel canonical correlation analysis [J]. Pattern Recognition, 2012, 45(8): 3003-3016.

[10] Jenatton R, Gramfort A, Michel V, *et al*. Mutial-scale mining of fMRI data with hierarchical structured sparsity [J]. SIAM Journal on Imaging Sciences, 2012, 5(3): 835-856.

[11] 邓战涛，胡谷雨，潘志松，等. 基于核稀疏表示的特征选择算法 [J]. 计算机应用研究，2012，29(4)：1282-1284.

[12] Kang P, Cho S. Locally linear reconstruction for instance-based learning [J]. Pattern Recognition, 2008, 41(11): 3507-3518.

[13] Zhu Xiaofeng, Huang Zi, Shen Hengtao, *et al*. Linear cross-modal hashing for efficient multi-mediasearch [C] //Proc of the 21st ACM International Conference on Multimedia. 2013: 143-152.

[14] Zhang Shichao. Nearest neighbor selection for iteratively KNN imputation [J]. Journal of Systemsand Software, 2012, 85(11): 2541-2552.

[15] Zhang Shichao. Shell-neighbor method and its application in missing data imputation [J]. Applied Intelligence, 2011, 35(1): 123-133.

[16] Zhang Shichao, Zhang Jilian, Zhu Xiaofeng, *et al*. Missing value imputation based on data clustering [M] //Transactions on Computational Science I. 2008: 128-138.

[17] Zhang Shichao, Huang Zi, Zhang Jilian, *et al*. Mining follow-up correlation patterns from time related databases [J]. Knowledge and Information Systems, 2008, 14(1): 81-100.

[18] Tibshirani R. Regression shrinkage and selection via the Lasso [J]. Journal of the Royal Statistical Society Series B: Methodological, 1996, 58(1): 267-288.

[19] Zhu Xiaofeng, Huang Zi, Cui Jiangtao, *et al*. Video-to-shot tag propagation by graph sparse group Lasso [J]. IEEE Trans on Multimedia, 2013, 15(3): 633-646.

[20] Boyd S. Alternating direction method of multipliers [C] //Proc of NIPS Workshop on Optimization and Machine Learning. 2011.

[21] Codes and data [EB/OL]. [2014-04-20]. http://www.cc.gatech.edu/~lsong/code.html.

[22] LIBSVM data [EB/OL]. [2014-04-20]. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.