# 基于标准欧氏距离的燃油流量缺失数据填补算法

陈静杰 车 洁
(中国民航大学电子信息与自动化学院 天津 300300)

**摘 要** 为减小数据缺失对飞机油耗统计推断精度带来的负面影响,针对基于传统欧氏距离、马氏距离以及精简关联度的最近邻填补算法的不足,提出了一种基于标准欧氏距离的填补算法来估计 QAR(Quick Access Recorder)数据中部分燃油流量数值的缺失。该算法通过 QAR 数据样本之间的标准欧氏距离选择最近邻样本,并利用熵值赋权法计算最近邻的加权系数,基于最近邻样本中燃油流量的加权平均即可得到缺失燃油流量的估计值。实验结果表明,标准欧氏距离能够有效度量样本相似性,所提出的算法优于常规填补算法,是处理飞机油耗数据缺失的一种有效方法。

**关键词** 标准欧氏距离,燃油流量缺失数据估计,K 近邻填补算法,熵值赋权法,RKNN 算法
**中图法分类号** TP391.9 **文献标识码** A

## Fuel Flow Missing-value Imputation Method Based on Standardized Euclidean Distance

CHEN Jing-jie CHE Jie
(College of Electronic Information and Automation, Civil Aviation University of China, Tianjin 300300, China)

**Abstract** To reduce the negative impact of aircraft fuel consumption statistical inference accuracy caused by the data missing, an estimated method based on standardized Euclidean distance was proposed to solve the fuel flow data missing problems. The nearest neighbors were chosen by the standardized Euclidean distance between QAR data samples, and then entropy was utilized to obtain the weight of the nearest neighbors. The missing value was estimated by the weighted average fuel flow of the nearest neighbors. Experiments prove that this method is valid to process fuel consumption data missing problems, and its performance is higher than the other imputation methods based on normal Euclidean distance, Mahalanobis distance or reduced relational grade.

**Keywords** Standardized euclidean distance, Fuel flow missing value estimation, KNN imputation method, Entropy, RKNN

## 1 引言

我国航空业碳市场将于 2017 年全面启动。碳核查中关于数据缺失而致的碳总量核算精度在 EU-ETS 和 ISO14064 系列标准中均有明确要求。我国航空业 98% 以上的碳排放量由航空燃油的消耗而产生,这要求更加精确地预估航空燃油的消耗。通过 QAR(Quick Access Recorder)数据来准确测度燃油消耗、控制碳排放量已成业内共识。然而,QAR 记录的包括燃油流量参量和影响其值的变量因素众多,而且由于运行环境复杂多变,几乎每一个航班的 QAR 数据中的不同参量都会出现不同程度的缺失,这已成为业界和学界准确统计分析和推断燃油消耗及碳排放量的热点和难点问题。

目前,关于数据缺失问题的处理方法各异,主要有删除和填补。在数据缺失不大于 5% 时,可以删除含有缺失值的样本从而得到完全样本,后期的统计分析误差也将在允许范围内[1]。当数据缺失大于 10% 时,为避免缺失而致的各类统计误差,需要利用已观测样本对缺失数据进行填补。基于数据属性和数据之间的相关性,提出利用基于贝叶斯[2]和 K 近邻(KNN)的缺失值填补算法[3-7]来处理基因序列和飞行过程的数据缺失问题。KNN 填补算法较常规均值插补算法有更好的健壮性和准确性。但是,传统 KNN 填补算法[3]用欧氏距离度量样本相似性,会受到样本属性量纲以及相关性的影响。为了改进 KNN 算法的这些不足,文献[6-7]又提出了基于马氏距离的填补算法 MKNN。马氏距离虽然考虑了各个观测变量之间的相关性,避免了量纲带来的影响,但是其计算过程繁琐,并在一定程度上夸大了变动微小的变量的作用,导致其结果不稳定。文献[4]中,精简关联度的使用避免了欧氏和马氏距离的缺点,在一定程度上提升了 MKNN 算法的计算速度,但是估计准确率并没有提高。

基于以上考虑以及航空业的特殊性,本文提出了一种基于标准欧氏距离的填补算法 Stand-KNN,该算法采用标准欧氏距离来度量样本相似性,选择 $K$ 个最近邻样本;同时采用信息熵赋权法对最近邻进行赋权,基于最近邻对应燃油流量的加权平均即可得到缺失燃油流量的估计值。所提算法有效解决了欧氏距离的缺点,避免了马氏距离计算中协方差逆矩阵不存在的问题,并有效提高了基于近邻插补算法的准确率。

## 2 燃油流量数据样本选择

燃油流量缺失数据估计中,近邻样本的选择至关重要。为提升填补算法的速度和准确度,要求参与距离计算的各个

本文受科技支撑项目(2012BAC20B03),民航局节能减排专项计划项目(DPDSR0010)资助。
陈静杰(1967-),男,博士,教授,主要研究方向为决策分析、民航运输过程能效管理与碳排放控制,E-mail:jjchen@cauc.edu.cn;车 洁(1991-),女,硕士生,主要研究方向为民航运输过程能效管理与碳排放控制。

属性与燃油流量的消耗密切相关。根据 QAR 数据特性，参考已有的飞机燃油估计模型[8-10]，选择 QAR 数据中的地速(GS)、纵向加速度(LONG)、垂直加速度(VRTG)、空速(IAS)、总重(GW)、高度(ALT)、总温(TAT)、风速(WIN-SPDR)、风向(WIN-DIR)以及倾斜角(PITCH)这 10 个影响飞机燃油消耗的主要因素作为样本属性，选择燃油流量(FF)作为样本的目标变量。

为直观验证对比实验结果，文中有目的地人为构造燃油流量数据缺失实例，并将其归类为间隔型数据缺失，样本如表 1 所列(NULL 表示该数据缺失)。

表 1 QAR 缺失数据样本

| 时间 | ALT | IAS | GS | WIN-SPDR | WIN-DIR | VRTG | LONG | TAT | PITCH | GW | FF |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1:17:32 | 6324 | 45 | 0 | 0 | 0 | 1.005 | −0.011 | 6 | −0.7 | 140240 | 880 |
| 1:17:33 | 6234 | 45 | 0 | 0 | 0 | 1.009 | −0.013 | 6 | −0.7 | 140240 | 928 |
| ... | | | | | | | | | | | |
| 1:24:09 | 7153 | 165 | 196 | 17 | 228.52 | 1.060 | 0.292 | 13 | 13.36 | 140080 | NULL |
| 1:24:10 | 7187 | 166 | 196 | 17.5 | 228.52 | 1.034 | 0.286 | 13 | 13.54 | 140080 | NULL |
| ... | | | | | | | | | | | |
| 1:24:19 | 7598 | 170 | 198 | 19.5 | 234.84 | 1.025 | 0.274 | 13.25 | 15.82 | 140080 | 17088 |

为方便后续使用，给出如下定义。

**定义 1** 从 QAR 缺失数据样本中提取含有缺失值的样本，记作缺失样本：$g_t=[g_{t1},g_{t2},\cdots,g_{tm}]$，$t=1,\cdots,N$ 为缺失样本的大小，$m$ 为样本维度，本文中 $m=11$，$g_{t1}-g_{tm}$ 为选择的 10 个影响飞机燃油消耗的主要因素以及燃油流量。

**定义 2** 将提取缺失样本后的 QAR 缺失数据样本记作完全样本，$z_i=[z_{i1},z_{i2},\cdots,z_{im}]$，$i=1,\cdots,M$ 为完全样本的大小，$N+M$ 为 QAR 数据样本的大小。

## 3 基于标准欧氏距离的燃油流量填补算法

根据 QAR 数据样本特性，燃油流量数据缺失样本的各属性间具有相关性，并且各属性的度量量纲不同。为了克服传统欧氏距离和马氏距离的缺点，采用标准欧氏距离来度量飞行数据各样本间的相似程度，选择最近邻，其在考虑样本整体特性的同时提高了估计准确率。通过缺失样本的 $K$ 个近邻样本中燃油流量的加权平均得到燃油流量估计值。K 近邻样本的加权系数采用信息熵赋权法得到。

### 3.1 K 近邻样本选择

在传统的 KNN 和 MKNN 算法的基础上，本文利用标准欧氏距离代替欧氏距离、马氏距离以及精简关联度来度量燃油流量数据样本间的相似性。标准欧氏距离利用完全样本的标准差，考虑了各观测指标取值的差异程度，抑制了传统欧氏距离计算中观测指标量纲带来的影响，也避免了马氏距离协方差矩阵的逆矩阵不存在的问题，降低了算法的计算复杂度。其计算公式为：

$$d(z_i,g_t)=\sqrt{(z_i-g_t)^{\mathrm{T}}V^{-1}(z_i-g_t)} \tag{1}$$

其中，$V$ 为一次完整飞行过程中不含缺失值的完全数据样本的标准差。$d(z_i,g_t)$ 表示缺失样本 $g_t$ 和完全样本 $z_i$ 之间的标准欧氏距离，距离越小，两个样本越相似。根据计算的标准欧氏距离，选择距离最小的 $K$ 个完全样本作为最近邻。

### 3.2 燃油流量缺失值估计

采用信息论中熵值的概念，确定最近邻的加权系数[5-6]，最近邻对应燃油流量的加权平均即可得到缺失燃油流量的估计值。具体步骤如下：

1)单位化 $K$ 个最近邻的标准欧氏距离。

$$p_i=\frac{d(z_i,g_t)}{\sum_{i=1}^{K}d(z_i,g_t)},i=1,2,\cdots,K \tag{2}$$

其中，$\sum_{i=1}^{K}p_i=1$。如果 $d(z_i,g_t)=0$，则令其等于 0.0005。

2)计算第 $i$ 个最近邻的熵值 $h_i$。

$$h_i=-(\ln K)^{-1}\times p_i\times\ln p_i \tag{3}$$

其中，$K$ 为选择的近邻个数，ln 为自然对数。

3)计算第 $i$ 个最近邻的变异程度系数 $v_i$。

由于 $0\leqslant h_i\leqslant 1$，熵值的大小与其变异程度相反，则定义第 $i$ 个最近邻的变异程度系数为：

$$v_i=1-h_i \tag{4}$$

4)计算第 $i$ 个最近邻的加权系数 $w_i$。

$$w_i=\frac{1}{K-1}\left(1-\frac{v_i}{\sum_{i=1}^{K}v_i}\right) \tag{5}$$

根据信息熵原理，相似样本的变异程度越小，其包含的确定性信息就越多，对应的加权系数就越大，反之亦然。加权系数满足：$\sum_{i=1}^{K}w_i=1$。

5)估计燃油流量的缺失值。

缺失样本 $g_t$ 中燃油流量的估计值由其对应的 $K$ 个近邻样本的燃油流量数据加权平均得到，计算公式如下：

$$\widetilde{FF}_t=\sum_{i=1}^{K}w_i*FF_i \tag{6}$$

其中，$FF_i$ 为第 $i$ 个最近邻数据样本中的燃油流量。$\widetilde{FF}_t$ 即为缺失样本 $g_t$ 中燃油流量的估计值。

重复式(1)—式(6)，即可完成燃油流量数据中缺失样本的填补。

## 4 结果分析

实验选取实际的 QAR 数据为测试样本，人为构造如表 1 所列的不同缺失率下的燃油流量数据间隔型缺失，并保留原有 QAR 数据进行填充效果验证。为了验证方法的有效性和普适性，分别选取了两个兰州飞往北京和一个首尔仁川飞往青岛的航班 QAR 数据作为实验数据集，并人为构造燃油流量数据缺失。

采用均方根误差(Root Mean Squared error，$\mathrm{RMS}_{\mathrm{error}}$)对各种填补算法的性能进行评价：

$$RMS_{error}=\sqrt{\frac{\sum_{t=1}^{N}(FF_t-\widetilde{FF}_t)^2}{N}} \tag{10}$$

其中，$FF_t$ 是缺失样本 $g_t$ 对应的实际燃油流量，$\widetilde{FF}_t$ 为缺失样本 $g_t$ 对应的燃油流量估计值。$RMS_{error}$ 的值越小，估计值越准确，反之结果就越差。

### 4.1 有效性验证

首先利用兰州-北京(1)的 QAR 数据训练模型，选择合适

的最近邻个数 $K=4$，再利用其余数据集和不同的缺失率来测试基于标准欧氏距离的填充算法的有效性，并与迭代 KNN（IKNN，在普通 KNN 的基础上添加了迭代过程）、MKNN 以及基于精简关联度的近邻填补算法（RKNN）进行比较。

通过 MATLAB 仿真，对于不同缺失率的两个数据样本，不同填充算法的填补效果如图 1 和图 2 所示。

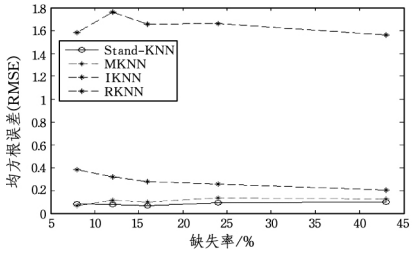


图 1　兰州-北京(2)填充效果

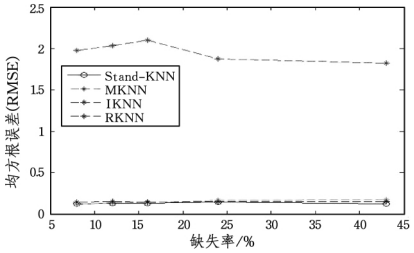


图 2　首尔仁川-青岛填充效果

图 1 利用了与训练样本同一航段（兰州-北京）的 QAR 数据，数据样本大小为 7761。图 2 使用了首尔仁川-青岛的 QAR 数据，数据样本大小为 5397。从图中可以看出，在缺失率小于 10% 时，MKNN 算法的均方根误差略小于 Stand-KNN 算法，但在缺失率增大时，它并没有保持这种优越性。而 RKNN 算法的均方根误差始终最高。从图 1 和图 2 中可以看出，在两种不同的机型和不同的环境因素下，标准欧氏距离填充算法始终保持着最低的填充误差，预测准确率明显优于 RKNN 和 IKNN 算法，填补效果略优于 MKNN 算法，从而验证了基于标准欧氏距离的燃油流量填补算法的有效性。

表 2 列出了进行上述对比实验时 4 个算法的耗时情况。对比表中数据可以发现，MKNN 算法耗时最多，Stand-KNN 算法和 RKNN 算法相对快捷。根据 EU-ETS 和 ISO14064 系列标准的核算精度要求，Stand-KNN 算法更适用于填补 QAR 数据中的缺失样本。

表 2　算法耗时/s

| 缺失率/% | 兰州-北京(2) | | | | | 首尔仁川-青岛 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 8 | 12 | 16 | 24 | 43 | 8 | 12 | 16 | 24 | 43 |
| Stand-KNN | 29.5 | 38.9 | 52.4 | 67.3 | 87.4 | 12.8 | 18.9 | 24.5 | 31.8 | 41.9 |
| MKNN | 295.8 | 410.5 | 518.2 | 718.5 | 943.8 | 110.5 | 198.1 | 206.5 | 338.3 | 448.0 |
| IKNN | 37.0 | 55.4 | 66.8 | 92.1 | 119.8 | 20.2 | 24.9 | 32.4 | 43.1 | 56.7 |
| RKNN | 23.8 | 32.3 | 41.9 | 57.3 | 75.0 | 11.0 | 15.8 | 20.6 | 27.1 | 35.6 |

### 4.2　普适性验证

为了验证 Stand-KNN 算法对不同类型燃油流量数据缺失的普遍适用性，人为将兰州-北京(2)的 QAR 数据燃油流量在某一段时间间隔内全部缺失，以与训练数据集使用的间隔型数据缺失形成对比。根据时间间隔的大小，决定数据缺失率的大小。4 种算法的填补效果如图 3 所示。由图 3 可以看出，在缺失率小于 12% 时，标准欧氏距离填充算法的均方根误差与 IKNN 和 MKNN 填补算法比较接近；当缺失率增大时，标准欧氏距离填补算法的优势得到了突显，而 RKNN 算法对这类缺失情况的填补误差依旧保持在相对较大的范围。实验结果表明，基于标准欧氏距离的燃油流量填补算法对各种类型的燃油流量数据缺失情况普遍适用。

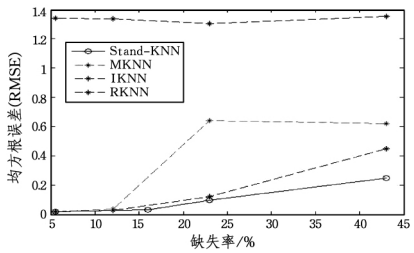


图 3　兰州-北京(2)大间隔缺失的填补效果

结束语　本文参考了目前已有的 K 近邻填补算法，通过改变相似度度量指标，提出了一种适用于航班 QAR 数据中燃油流量数据缺失问题的标准欧氏距离填补算法。实验证明，该算法能够有效填补飞行过程中的燃油流量数据缺失，这为航空公司进行燃油消耗的预测和监控提供了有力保障。该方法具有简单、稳定并能广泛适用于各种类型的数据缺失情况的特点，同时还能保证填充之后的数据满足航空业的误差要求。当缺失率大于 50% 时，已观测数据已不能完全体现数据特征，后续将探讨利用同一机型、同一航段的一定规模航班量的 QAR 数据进行大间隔、高缺失率数据的填补。

## 参 考 文 献

[1] 庞新生. 缺失数据处理方法的比较[J]. 统计与决策，2010(24)：152-155.

[2] SHIGEYUKI O B A，MASA-AKI S，ICHIRO T，et al. A Bayesian Missing Value Estimation Method for Gene Expression Profile Data[J]. Bioinformatics，2003，19(16)：510-514.

[3] TROYANSKAYA O，CANTOR M，SHERLOCK G，et al. Missing Value Estimation Methods for DNA Microarrays[J]. Bioinformatics，2001，17(4)：520-525.

[4] 何云，皮德常. 基于精简关联度的基因表达数据迭代填补算法[J]. 计算机科学，2015，42(11)：251-255，283.

[5] STOICA P，LI J，LING J. Missing Data Recovery Via a Nonparametric Iterative Adaptive Approach[J]. IEEE Signal Processing Letters，2009，16(4)：241-244.

[6] 杨涛，骆嘉伟，王艳，等. 基于马氏距离的缺失值填充算法[J]. 计算机应用，2002，25(12)：2868-2871.

[7] 王雪飘，张宏亭，李学仁. 基于马氏距离的飞行缺失数据估计方法[J]. 火力与指挥控制，2009，34(8)：113-115.

[8] 刘婧. 基于飞行数据分析的飞机燃油估计模型[D]. 南京：南京航空航天大学，2010.

[9] 陈静杰，肖冠平. 飞机油耗分析工具设计[J]. 计算机工程与设计，2014，35(11)：4012-4016.

（下转第 125 页）

表 1　$a=0.5$ 时两种算法实验结果的对比

| 函数 | 算法 | $d=10$ | $d=20$ | $d=30$ |
|---|---|---|---|---|
| $f_1$ | 原算法 | $6.1698*10^{-5}$ | $7.4302*10^{-5}$ | $1.2596*10^{-4}$ |
| | 改进算法 | $3.8524*10^{-5}$ | $3.9227*10^{-5}$ | $3.9542*10^{-5}$ |
| $f_2$ | 原算法 | $8.9059*10^{-11}$ | $1.7910*10^{-10}$ | $3.3602*10^{-10}$ |
| | 改进算法 | $4.2239*10^{-12}$ | $4.7687*10^{-11}$ | $8.1150*10^{-11}$ |
| $f_3$ | 原算法 | 0.2318 | 0.7239 | 0.8273 |
| | 改进算法 | 0.8843 | 0.9775 | 8.4607 |
| $f_4$ | 原算法 | 0.0172 | $1.0259*10^{-8}$ | $2.7066*10^{-9}$ |
| | 改进算法 | 0.0271 | $1.3551*10^{-9}$ | $1.2529810^{-9}$ |
| $f_5$ | 原算法 | $1.1096*10^{-8}$ | $5.1128*10^{-8}$ | $8.7898*10^{-8}$ |
| | 改进算法 | $1.5285*10^{-9}$ | $1.7319*10^{-9}$ | $2.0827*10^{-8}$ |
| $f_6$ | 原算法 | −9.0019 | −18.8389 | −28.9458 |
| | 改进算法 | −9.6176 | −19.4823 | −29.4008 |
| $f_7$ | 原算法 | $2.0557*10^{-4}$ | $2.5456*10^{-4}$ | $3.8185*10^{-4}$ |
| | 改进算法 | 2170.2 | 2783.3 | 4540.2 |

从表 1 可以看出，对于函数 $f_1$，$f_2$，$f_4$，$f_5$，$f_6$ 改进后的算法表现出较好的寻优性能，尤其是对于高维函数。对于函数 $f_3$，$f_7$ 改进后的算法寻优效果不理想，$f_7$ 的搜索区间很大，改进后的算法出现不收敛的情况。

图 2－图 7 示出了优化函数在 30 维时的收敛曲线（图中虚线表示原算法，实线表示改进算法）。从图 2－图 7 可以看出，相比于原算法，改进后的算法的收敛速度更快，收敛精度更高；且改进后的算法克服了标准萤火虫算法对于高维函数容易陷入局部最优的缺点，整体表现出较理想的效果。

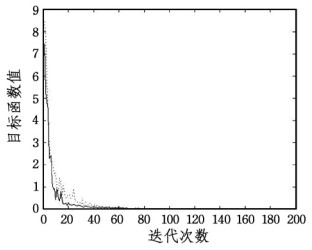


图 2　Ackley 函数的收敛曲线

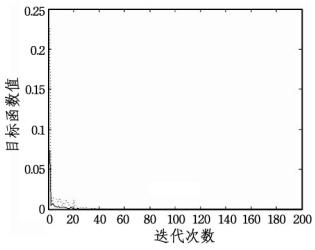


图 3　De Jong 函数的收敛曲线

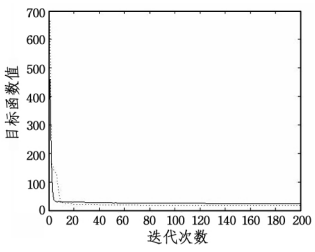


图 4　Rosenbrock 函数的收敛曲线

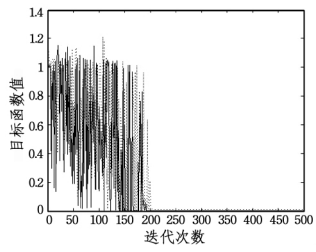


图 5　Griewank 函数的收敛曲线

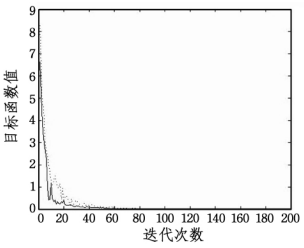


图 6　Rastrigin 函数的收敛曲线

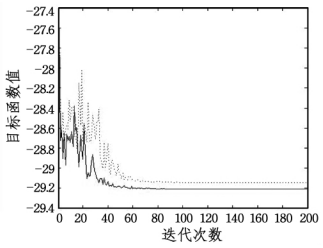


图 7　Michalewicz 函数的收敛曲线

结束语　通过比较高维优化函数的仿真结果，可以看出，改进后的算法对于函数 $f_1$，$f_2$，$f_4$，$f_5$，$f_6$ 都有较好的效果，即可使高维优化函数更好地收敛到全局最优且收敛速度更快，说明本文对算法的改进有一定的可行性和有效性。

本文对文献[1]提出的基于对偶和维度的萤火虫算法做出了进一步改进，即在算法迭代的寻优过程中的替换部分引入了加权，取部分萤火虫信息和部分全局最优萤火虫信息，保证替换过程不丢失标准萤火虫算法产生的当前最优萤火虫的信息。最后，用测试函数证明了本文改进的有效性及可行性。

## 参 考 文 献

[1] VERMA O P，AGGARWAL D，PATODI T. Opposition and dimensional based modified firefly algorithm[J]. Expert Systems with Applications，2016(44)：168-176.

[2] YANG X S. Nature-Inspired Metaheuristic Algorithms [M]. Frome：Luniver Press，2008：83-96.

[3] 赵玉新，YANG X S，刘利强. 新兴元启发式优化方法[M]. 北京：科学出版社，2013：148-170.

[4] 李瑞青. 改进的萤火虫算法及应用[D]. 长春：吉林大学，2015.

[5] 王沈娟，高晓智. 萤火虫算法研究综述[J]. 微型机与应用，2015，34(8)：8-11.

[6] SENTHILNATH J，OMKAR S N，MANI V. Clustering Using-firefly Algorithm：Performance Study[J]. Swarm & volutionary EComputation，2011，1(3)：164-171.

[7] ZAMAN MA，MATIN A. Nonuniformly Spaced Linear Antenna Array Design Using Firefly Algorithm[J]. International Journal of Microwave Science and Technology，2012，8(36)：40-48.

[8] 王吉权，王福林. 萤火虫算法的改进分析及应用[J]. 计算机应用，2014，34(9)：2552-2556.

[9] YU S H，SU S B，LU Q P，et al. A novel wise step strategy for firefly Algorithm[J]. International Journal of Computer Mathematics，2014，91(12)：2507-2513.

[10] 李一玄. 萤火虫算法参数研究[J]. 物流工程与管理，2015，(9)：195-197.

（上接第 111 页）

[10] 陈静杰，邹迎欢. 油耗预测中显著影响因素参数提取方法仿真[J]. 计算机仿真，2013，30(6)：55-58.

[11] 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.

[12] 肖辉辉，段艳明. 基于属性值相关距离的 KNN 算法的改进研究[J]. 计算机科学，2013，40(11A)：157-159，187.

[13] 王凤梅，胡丽霞. 一种基于近邻规则的缺失数据填补方法[J]. 计算机工程，2012，38(21)：53-55，62.

[14] HUANG C C，LEE H M. A grey-based nearest neighbor approach for missing attribute value prediction[J]. Applied Intelligence，2004，20(3)：239-252.

[15] MEESAD P，HENGPRAPROHM K. Combination of KNN-Based Feature Selection and KNN-Based Missing-Value Imputation of Microarry Data[C]// The 3rd International Conference on Innovative Computing Information and Control. Washington，USA：IEEE Computer Society，2008：18-20.

[16] MOORTHY K，MOHAMAD M S，DERIS S. A Review on Missing Value Imputation Algorithms for Microarray Gene Expression Data[J]. Current Bioinformatics，2014，9(1)：18-22.