文章编号:1001 - 9081(2009)09 - 2502 - 03

基于马氏距离和灰色分析的缺失值填充算法

刘星毅
(钦州学院 数学与计算机科学系,广西 钦州 535000)
(qznc@163.com)

摘 要:针对 kNN 算法中欧氏距离具有密度相关性敏感的缺点,提出综合马氏距离和灰色分析方法代替 kNN 算法中欧式距离的新算法,应用到缺失数据填充方面。其中马氏距离能解决密度相关明显的数据集,灰色分析方法能处理密度相关不明显的情况。因此,该算法能很好处理任何数据集,实验结果显示,算法在填充结果上明显优于现有的其他算法。

关键词:数据预处理;缺失数据;最近邻算法;灰色分析;马氏距离
中图分类号: TP331 **文献标志码**:A

Improved kNN algorithm based on Mahalanobis distance and gray analysis

LIU Xing-yi
(*Department of Mathematics and Computer Science, Qinzhou University, Qinzhou Guangxi* 535000, *China*)

Abstract: The Euclidean-based k-Nearest Neighbor (kNN) algorithm is restricted to the dataset without correlation-sensitive on density. The author proposed an improved kNN algorithm based on Mahalanobis distance and gray analysis for imputing missing data to replace the existing Euclidean distance. The Mahalanobis distances can deal with the issue of correlation-sensitive on density, and the gray-analysis method can deal with the opposite case. Hence, the proposed method can deal with any kind of datasets, and the experimental results show the proposed method outperforms the existing algorithms.

Key words: data preprocessing; missing data; Nearest Neighbor (NN) algorithm; gray analysis; Mahalanobis distance

0 引言

最近邻(Nearest Neighbor, NN)算法由于容易理解,操作简单,效果明显,在科研和实际生活中都具有广泛的应用。在填充缺失数据方面,T. M. Cover 在 1967 年首次提出最近邻方法[1],它是热门的冷卡方法,最近邻算法已经被嵌入一些常见的软件中,例如,SAS,WEKA 等。

最近邻算法原理非常简单:两个具有最近距离事例的关系是最紧密。从原理可以看出,此算法中最重要的部分是如何有效地计算两个事例之间的距离。实际应用的最近邻算法中,两个事例的距离通常使用 Minkowski 距离来计算,它的计算式为:

$$d(i,j) = (|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \cdots + |x_{ip} - x_{jp}|^q)^{\frac{1}{q}}$$

其中 q 为一个正整数,称为 Minkowski 参数,当 $q = 1$ 时,代表曼哈顿距离;$q = 2$ 就是欧氏距离。通常,不同的数据集,应该选取不同的 q 值;不同的 q 值,能产生不同的填充效果。其中,欧氏距离(Euclidean distance)是一个常用的距离定义,它表示在 p 维空间中两个点之间的真实距离,两个向量之间的欧氏距离计算公式如下:

$$d(i,j) = (|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2)^{\frac{1}{2}} \quad (1)$$

常见的基于欧氏距离的最近邻算法[2-3] 可以描述为如下。

1) 对每一个含有缺失数据的事例,根据式(1) 计算它与数据集中其他没有缺失事例的距离。

2) 对得到的所有距离,进行排序,选取 K 个最小距离。

3) 如是离散属性,取这 K 个距离中的最大类为此事例的值。如此事例是连续属性,取这 K 个距离中的中位数为此事例的值;否则,取这 K 个完全事例最大类为缺失事例的填充值;

因此,要填充缺失数据,寻找两事例相似的方法是非常重要的。近来很多研究已经证明:欧氏距离在计算事例间的相似性方面适用范围有限,例如文献[4 - 5]指出,基于欧氏距离的缺失数据最近邻算法对一些密度基础的数据集处理效果很不理想,即对数据集密度相关性很敏感。所谓数据集敏感性,就是最近邻算法因为密度不同的数据集而产生不同的填充效果,如果数据集的点被投影到二维平面而显得比较集中在一个区域的时候,基于欧氏距离的最近邻填充算法的效果不好,如果数据集近似于均匀分布时候,填充效果却很好。显然,实际应用中,用户既不知道自己要处理的数据集是否均匀分布,甚至对数据集的分布没有任何先验知识,此时,强行贯彻基于欧氏距离的最近邻填充方法肯定不适合。本文提出使用马氏距离和灰色度方法替换欧氏距离的计算两事例间的相似度,然后综合它们提出了一种新的最近邻缺失数据填充算法。将在第 1 章论证马氏距离适合密度集中的数据集,而灰色分析是专门为在对数据集情况很模糊的情况下提出的。本文改进存在的最近邻填充算法 MGNN (Mahalanobis-Gray and k-Nearest Neighbor algorithm),使用马氏距离和灰色分析综合的方法能适用于任何类型的数据集,而且针对最近邻的另一个缺点(即易出现维灾难)也进行了改进。实验证明本文算法 MGNN 能有效地解决使用欧氏距离遇上的难题。

1 算法描述

1.1 马氏距离

欧氏距离虽然常被用于计算两事例的距离,但也有明显

收稿日期:2009 - 03 - 23;**修回日期**:2009 - 05 - 12。
基金项目:广西自然科学基金资助项目(桂科自 0899018);广西教育厅科研项目(200808MS062)。
作者简介:刘星毅(1972 -),男,广西钦州人,副教授,硕士,CCF 会员,主要研究方向:数据库、计算机网络。

的缺点。例如,它将样本的不同属性(即各指标或各变量)之间的差别等同看待,这一点经常不能满足实际要求,因为实际中各个属性的存在对决策起到不同的作用。在一些研究中,例如教育研究,经常遇到对人的分析和判别,个体的不同属性对于区分个体有着不同的重要性。因此,此时欧氏距离明显不足,需要采用不同的距离函数,马氏距离是一个很好的替代者。

马氏距离(Mahalanobis distance)是由印度统计学家马哈拉诺比斯(P. C. Mahalanobis)提出,用来表示数据的协方差距离。它是一种有效计算两个未知样本集的相似度方法。与欧氏距离不同的是它考虑到各种特性之间的联系(例如:一条关于身高的信息会带来一条关于体重的信息,因为两者是有关联的)并且是尺度无关的(scale-invariant),即独立于测量尺度。对于一个均值为$\mu = (u_1,\cdots,u_p)$协方差矩阵为Σ的多变量向量$x = (x_1,x_2,\cdots,x_p)$,其马氏距离为:

$$D_M(x) = \sqrt{(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)}$$

马氏距离也可以定义为两个服从同一分布并且其协方差矩阵为Σ的随机变量$\vec{x}$与$\vec{y}$的差异程度:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^{\mathrm{T}}\Sigma^{-1}(\vec{x}-\vec{y})}$$

如果协方差矩阵为单位矩阵,那么马氏距离就简化为欧氏距离,如果协方差矩阵为对角阵,则其也可称为正规化的欧氏距离,即:

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{p}\frac{(x_i-y_i)^2}{\sigma_i^2}}$$

其中σ_i是x_i的标准差。因此,可以说欧氏距离是马氏距离的一种特殊情况。

本文我们可以定义目标事例(有缺失事例)x_0与无缺失事例x_i之间的马氏距离为:

$$Mahal(x_0,x_i) = \sqrt{(x_0-x_i)^{\mathrm{T}}\Sigma^{-1}(x_0-x_i)};\quad i = 1,2,\cdots,n \tag{2}$$

马氏距离还有很多优点,例如,它不受量纲的影响,两点之间的马氏距离与原始数据的测量单位无关,而这种情况下运用欧氏距离,经常会出现由于数量级不同的偏置现象,而此时由标准化数据和中心化数据(即原始数据与均值之差)计算出的两点之间的马氏距离是相同的。此外,马氏距离还可以排除变量之间的相关性的干扰。下面,以一个具体实例显示马氏距离的优越性。

图 1 是一个二维坐标上的点图,从图上显然可以看出,线段AB的欧氏距离小于线段AC的欧氏距离。但是,从图的整体结构可以看出,AC的距离应该短于AB之间的距离。这是因为,点A和点C都在一个类中,而点B是点A和点C所属类的一个孤立点,B到A的距离应该长于点C到A的距离。但是欧氏距离不能体现出这个性质,马氏距离却能体现。

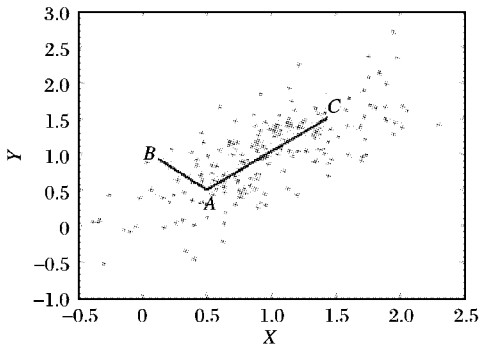

图 1　马氏距离与欧氏距离对比

根据图 1 中所示的坐标,A:(0.5,0.5),B:(0,1),C:(1.5,1.5),计算两线段AC和AB的协方差为:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

根据马氏距离计算式(2)可得到:$Mahal(A,B) = 5$;$Mahal(A,C) = 4$。由此可以得出结论:$|AB| \geqslant |AC|$,这个结果与上面分析一致。因此,从事例说明,在这种情况下,马氏距离比欧氏距离效果要明显。而且欧氏距离是马氏距离的一种特殊情况。因此,使用马氏距离取代欧氏距离是合理的。

马氏距离对密度相关(密度相关[5]:数据集点的分布是聚集到有限个部分得形状而非均匀分布形状)的数据集的优势是明显的,但是也存在一定的缺点。文献[5-6]认为,一般的数据集可能是密度相关也可能不是密度相关的。在密度相关不明显的情况下,马氏距离的填充效果有时比欧氏距离要差。因此,本文接下来提出密度相关不明显情况下使用灰色分析来计算两事例间的相似度。

1.2　灰色分析

灰色系统理论(Grey System Theory, GST)首先由邓聚龙教授在 1982 年提出[6-7]。灰色分析法是一种对含有不确定因素的系统进行预测的方法。灰色系统是介于白色系统和黑色系统之间的一种系统。白色系统是指一个系统的内部特征是完全已知的,即系统的信息是完全充分的。而黑色系统是指一个系统的内部信息对外界来说是一无所知的,只能通过它同外界的联系来加以观测研究。灰色系统的一部分信息是已知的,另一部分是未知的,系统内各因素间具有不确定关系。现在,灰色分析方法已经被广泛应用到图像处理(image processing),机器视图检测(machine vision inspection),决策推理(decision making),股票价格预测(stock price prediction)和系统控制(system control)方面。在本算法中,由于有时对所处理的数据集密度相关与否不明确(类似于对此密度相关与否的认识是灰色的),灰色分析方法能比欧氏距离或者马氏距离的效果更加明显[6]。

考虑$n+1$个事例$\{x_0,x_1,x_2,\cdots,x_n\}$,$x_0$有缺失数据的事例,$x_1,x_2,\cdots,x_n$没有缺失的事例。每个事例$x_i$有$m$个条件属性记为$x_i = (x_i(1),\cdots,x_i(m))$,$i = 1,2,\cdots,n$和一个类标签$D_i$,则灰色相关程度式为:

$$GRG(x_0,x_i) = \frac{1}{m}\sum_{k=1}^{m}GRC(x_0(k),x_i(k));\ i = 1,2,\cdots,n \tag{3}$$

$GRG(x_0, x_1)$大于$GRG(x_0, x_2)$说明x_0与x_1的相似程度小于x_0和x_2的相似程度,反之亦然。$GRG(x_0, x_1) = 1$说明这两个事例没有任何关系,$GRG(x_0, x_1) = 0$,说明这两个事例几乎一样。

1.3　缺失数据填充算法 MGNN

1.1 节和 1.2 节我们提出了两种计算两事例间相似度的方法,并且也指出马氏距离适合于密度相关性已知情况,灰色分析方法适用于对数据集情况不甚了解情况。实际应用中,用户对数据集的密度相关性通常没有什么先验知识,此时,综合这两种度量应该是最好的方式,即综合式(2)~(3)可得:

$$Dist(x_0,x_i) = \lambda GRG(x_0,x_i) + (1-\lambda)Mahal(x_0,x_i);\quad i = 1,2,\cdots,n \tag{4}$$

其中λ是调整参数,如果用户事先知道数据集的密度相关性,则$\lambda = 0$;如果完全不知道,则$\lambda = 1$。一般情况下,令$\lambda = 0.5$,也可以由专家指定具体的值。因此,本文的填充算法应该同

时考虑这两种情况。所以,算法 MGNN 可以描述为:

1)对所有数据进行数据规范;

2)数据降维;

3)对每一个要找最近邻的事例,根据式(4)计算它与数据集中其他事例的距离;

4)对得到的所有距离,进行排序,选取 K 个最小距离;

5)如是离散属性,取这 K 个距离中的最大类为此事例的值;如此事例是连续属性,取这 K 个距离中的中位数为此事例的值。

在实际应用中,如果数据不在一个数量级,计算的时候通常结果容易偏向数量级大的属性,这就是常说的偏置问题。为了解决这个偏置问题,可以在计算前把各个数据规范化到0和1之间,这样,所有属性的数据都在一个数量级就可以有效避免偏置的问题了。常见的数据规范方法有最小—最大规范化,z-score 规范化等。只要使用数据规范方法均可有效避免属性间不同数量级的偏置问题,因此,本文选用最小—最大规范化对连续数据进行线性变换。假定 min_A 和 max_A 分别为属性 A 的最小和最大值,最小—最大规范化通过计算:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

将 A 的值 v 映射到区间 $[new\_min_A, new\_max_A]$ 中的 v'。

上面数据规范化是本文与引言部分列出的经典最近邻算法的第一个不同之处,也是文献[8]没有注意的一个细节问题。

规范化的数据可以降低算法偏置问题,但是由于普遍数据集含有较多属性,一些基因数据集通常有上百个属性,而文献[2]认为,一般属性超过5个左右,使用最近邻算法就容易出现维灾难问题。因此,MGNN 算法的第2)步就是对规范化的数据集进行降维处理。常用的降维处理非常多,有主成分分析法(Principal Component Analysis, PCA),奇异值分解法(Sigular Value Decomposition, SVD),随机映射法(Random Projection, RP)等,本文采用 SVD。事实上,本文算法的第一和第二步是可以交换的,不会影响最终效果。

与经典最近邻算法的第二个不同之处是算法 MGNN 的第三步,本文使用综合计算事例相似度的方法计算两事例的距离,而文献[8]也采用了马氏距离方法处理基因数据,但是没有分析欧氏距离的不足,也没有分析马氏距离代替欧氏距离的必要性和可能性,更没有考虑到对两个事例进行灰色分析。因此,在实验部分,本算法将于此两种算法进行比较。

2 实验分析

文献[8]已经证明他们的算法优于常见的最近邻算法,算法基本步骤为:首先对数据进行降维处理,他们采用的是 PCA 方法(本文采用 SVD 算法),然后对每两事例之间的距离用马氏距离取代常用的欧氏距离。我们把此算法记为 MKNN(Mahalanobis and K-Nearest Neighbor algorithm)。而本文算法有数据规范化预处理,然后 SVD 方法降维,最后采用式(4)计算两事例间的距离。实验数据采用基因表达数据(可从开放的公共基因数据库获取):识别酵母中能调节细胞周期的基因的研究[9]、酵母从发酵到氧化过程中新陈代谢变化对应的临时基因表达的探索研究[10]、在酵母中环境变化引起的基因表达变化的研究[11]。前两个数据集是时间序列数据,其中一个包含的噪声较小,称其为时间序列(记为 D1);另一个则有较大的噪声,称为噪声时间序列(记为 D2);最后一个为非时间序列数据集(记为 D3)。

从数据库中获取的数据本身可能会包含有缺失值,如果直接作为实验数据,则得到的结果无法进行评价,因此需要将其中包含有缺失数据的行和列删除,人为地获得完整的数据集。在获得的完整的数据集中,根据算法的需要随机删除一定比例的数据产生测试数据,然后再使用各种算法来恢复测试数据中的缺失值,并将估计值与真实值进行比较。

对各种算法的缺失值估计算的性能采用均方根误差(Root Mean Squared, RMS)来评价:

$$RMS = \sqrt{\frac{\sum_{i=1}^{N}(R_i - I_i)^2}{N}}$$

其中,R_i 为真实值,I_i 是估计值,N 为缺失值个数。计算得到的 RMS 的值越小,其估计值就越准确,反之结果就越差。

由于篇幅有限,本文测试每个算法的各种不同的 k 值,从 $k = 1$ 到 $k < N$,然后取最优的 k,对三个不同的数据集得到的结果如图2所示。

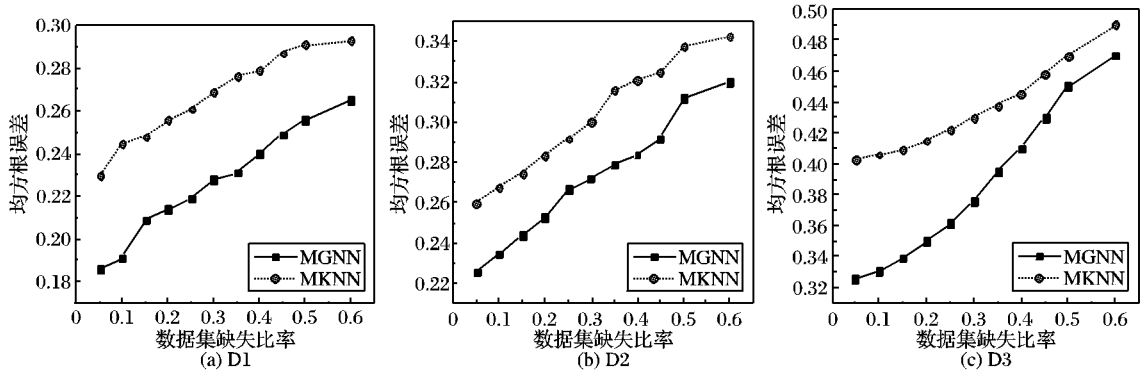

图2 真实数据实验结果

从图2可以看出:对于不同类型的数据集,例如,有噪音时间序列数据集,无噪音时间序列数据集以及非时间序列数据集,本文算法 MGNN 均优于文献[8]的算法 MKNN。而且,可以知道,数据缺失率越大,填充结果越差,这是显然的,因为缺失越多,能利用的有效信息就越少,填充结果就越差。由于文献[8]已经证明了他们的算法优于欧氏距离的最近邻算法,因此,本文算法 MGNN 显然优于传统的最近邻填充算法。

3 结语

本文在常用算法—最近邻算法使用马氏距离和灰色分析综合方法代替常用的欧氏距离,改进后的算法在实际例子和真实数据集上的实验上都显示,本文提出的算法优于存在的算法。

(下转第2536页)

空间中。利用 toolhelp.dll 的 THCreateSnapshot 函数可以枚举所有正在运行的进程,然后利用 PerformCallBack4 请求指定进程通过调用 LoadLibraryW 来载入 API 挂钩模块这个 DLL。最后挂钩 CreateProcess,在 CreateProcess 的钩子函数中,监控有没有新进程产生。一旦有新进程生成,就把 API 挂钩模块这个 DLL 装载进新进程的进程空间中。

卸载是安装的逆操作。通过调用 PerformCallBack4 可以请求所有进程调用 FreeLibrary,从而把 API 挂钩模块这个 DLL 从它们的进程空间中安全地卸载出来。

5.2 API 挂钩模块

API 挂钩模块包括挂钩指定 API 以及撤销挂钩指定 API 两个功能,这两个功能互为逆操作。

挂钩指定 API 的就是利用上文所述的新的钩子技术,为指定 API 建立钩子函数。这样一来,当用户调用这个 API 时,实际上执行的是钩子函数的代码而不是原 API 的代码,这也就实现了截获 API 的功能。

在给指定 API 进行挂钩时,我们必须知道该 API 所属的 SystemAPISet 在 SystemAPISsts 表中的下标,以及该 API 所对应的 Win32Method 在 Win32Methods 表中的下标。本文件监控系统所挂钩的 API 如表 1 所示。

表 1 挂钩 API 列表

API 名称	在 SystemAPISets 表的下标	在 Win32Methods 表的下标
LoadLibraryW	0	8
FreeLibrary	0	9
CreateProcessW	0	53
OpenProcess	0	119
MoveFile	20	4
DeleteFile	20	6
CreateFile	20	9

5.3 用户控制模块

用户控制模块提供了用户与文件监控系统进行交互的操作界面。当某文件访问操作(即调用相关 API)发生时,API 挂钩模块会截获该操作并通知用户控制模块进行处理;用户控制模块则根据用户设置的监控策略分析判断该文件访问操作是否合法,如果合法,则允许该文件访问操作正常执行,否则弹出警报提示用户进行处理。通过用户控制模块,用户还可以查看监控记录等。

6 实验及结果

本文件监控系统在多普达 S1(操作系统是 Windows Mobile 6 Professional)上进行测试。测试结果表明,对于一般的文件访问操作,如创建、打开、读、写、执行、修改及删除等,本文件监控系统都能成功进行截获,实现文件监控的功能。

为了进一步验证该系统的有效性,实验中将监控策略设置为监控特定的 PE 文件(如含有某特征码)。结果表明,系统能够达到预期目标,实现预定功能。可见,该系统是有效的,且具有很大的实用性。

7 结语

本文根据 Windows Mobile 系统本身的特点,提出并设计了一种新的钩子技术,并以此设计和实现了基于 Windows Mobile 的文件监控系统。实验证明了该系统的有效性及实用性。此外,作为该系统的延伸,该系统还可以用来分析判断某试图访问的文件是否具有特定的标志,如数字签名等,也可以对文件访问进行某些特殊的处理,如数据加密、压缩或使用统计数据进行监视等。

参考文献:

[1] 刘彦博,胡砚,马骐. Windows Mobile 平台应用与开发[M]. 北京:人民邮电出版社,2006.

[2] DAGON D, MARTIN T, STARNER T. Mobile phones as computing devices: The viruses are coming! [J]. IEEE Pervasive Computing, 2004, 3(4): 11 - 15.

[3] LEAVITT N. Mobile phones: The next frontier for hackers [J]. Computer, 2005, 38(4): 20 - 23.

[4] 石京民,陈道敏. 钩子及其应用[J]. 计算机应用, 2001, 21(4): 83 - 84.

[5] 微软. Windows Mobile 6 SDK 帮助文档[EB/OL]. [2009 - 02 - 20]. http://www.microsoft.com/downloads/details.aspx?FamilyID=06111A3A-A651-4745-88EF-3D48091A390B&displaylang=en.

[6] 李蒙,舒云星. Windows CE 驱动程序开发[J]. 计算机工程与设计, 2004, 25(6): 963 - 981.

[7] MURRAY J. Inside Microsoft Windows CE [M]. Santa Clarita, CA: Microsoft Press, 1998.

[8] 姜波. Windows CE.Net 程序设计[M]. 北京:机械工业出版社,2007.

[9] 戴春达,符红光. Win32 中钩子的实现技术及其应用[J]. 计算机应用, 2002, 22(8): 72 - 74.

[10] LEMAN D. Spy: A Windows CE API interceptor [J]. Doctor Dobbs Journal, 2003, 28(10): 54 - 59.

(上接第 2504 页)

参考文献:

[1] COVER T M, HART P E. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 1967, 13(1): 21 - 27.

[2] HAN J, KAMBER M. Data mining concepts and techniques [M]. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006.

[3] SCHAFER J, GRAHAM J. Missing data: Our view of the state of the art [J]. Psychological Methods, 2002, 7(2): 147 - 177.

[4] LAKSHMINARAYAN K, HARP S A, SAMAD T. Imputation of missing data in industrial databases [J]. Applied Intelligence, 1999, 11(3): 259 - 275.

[5] LITTLE R, RUBIN D. Statistical analysis with missing data [M]. 2nd ed. New York: John Wiley and Sons, 2002.

[6] 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.

[7] 邓聚龙. 灰色系统理论[M]. 武汉:华中理工大学出版社, 1984.

[8] 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, 2005, 25 (12): 2868 - 2871.

[9] SPELLMAN P T, SHERLOCK G, ZHANG M Q, *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by micro array hybridization [J]. Molecular Biology of the Cell, 1998, 9(12): 3273 - 3297.

[10] DERISI J L, IYER V R, BROWN P O. Exploring the metabolic and genetic control of gene xpression on a genomic scale [J]. Science, 1997, 278(5338): 680 - 686.

[11] GASCH A P, SPELLMAN P T, KAO C M, *et al.* Genomic expression programs in the response of yeast cells to environmental changes [J]. Molecular Biology of the Cell, 2000, 11(12): 4241 - 4257.