基于广义马氏距离的缺损数据补值算法

陈 欢 黄德才
(浙江工业大学计算机学院 杭州 310023)

摘 要 在数据收集过程中数据缺损是不可避免的。如何还原这些缺损数据,成为数据挖掘研究的热点问题之一。与许多现有算法一样,基于马氏距离的缺损数据补值算法充分利用了实际数据之间的相关性,具有较好的补值效果,但它要求数据的相关性协方差矩阵可逆,使其应用范围受到了极大的限制。在改进传统主成分分析方法的基础上,利用矩阵的奇异值分解理论和Moore-Penrose广义逆性质,提出了广义马氏距离的概念,并运用于SOFM神经网络,结合信息熵理论设计了基于广义马氏距离的缺损数据补值算法——GS算法。理论分析和数值仿真结果表明,广义马氏距离完全继承了马氏距离在处理相关性数据上的性能优势,新算法不仅在补值的精确度和稳定性上有很好的效果,而且适用于任意数据集合。

关键词 主成分分析,Moore-Penrose伪逆,广义马氏距离,SOFM神经网络,信息熵
中图法分类号 TP18 **文献标识码** A

Missing Data Imputation Based on Generalized Mahalanobis Distance

CHEN Huan HUANG De-cai
(College of Computer Science & Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract Missing data are inevitable in data-collection, how to restore these data has become one of the hottest issues in data mining. Just like most algorithms, missing data imputation algorithms based on Mahalanobis Distance make full use of relationships between data. Though the results are acceptable, the covariance matrixes are not always reversible, which limit the algorithms greatly. This paper improved a traditional principal component analysis(PCA) method, proposed a new distance named Generalized Mahalanobis Distance according to SVD and Moore-Penrose pseudoinverse. Combining with SOFM neural network and entropy, we designed GS missing data imputation algorithms. After academic analysis and simulation, it was proved that Generalized Mahalanobis Distance inherits the advantages of Mahalanobis Distance wonderfully in dealing with relatived data. Not only the new algorithm has good accuracy and stability, but also suits for any datasets.

Keywords PCA, Moore-penrose pseudoinverse, Generalized mahalanobis distance, SOFM neural network, Entropy

1 引言

在数据挖掘过程中,如果采集到的数据集是一个虚假的、劣质数据泛滥的数据集,很难想象所做的数据挖掘的结果会是怎样,因此数据的预处理便成为了整个数据挖掘至关重要的一步。其主要工作有:数据清理、数据集成、数据转换和数据归约等。在数据预处理过程中,经常会发现某些个体缺少属性值,这将使系统丢失大量有用信息,而且系统的不确定性也会显著增加。这些都将严重影响数据挖掘的可靠性。

因此,各种缺损数据补值方法应运而生。根据相关性原则,其主要可分为两类:基于半相关性补值方法[1,2,4,10]和基于相关性补值方法[5,6,8,11,12]。前者主要是通过欧式距离、范数等绝对度量进行补值,虽然它们平均计算复杂度较后者好,但是鉴于实际数据之间大多具有相关性,因此这些方法的精确度普遍存在缺陷。后者主要是通过马氏距离、灰色理论等相关性度量进行补值,它们较充分地考虑了数据间的相关性,具有较高的准确性,在中小数据端有较好的应用,即使是在大数据量的情况下,也有学者通过蒙特卡罗等统计算法将数据简化后再运用相关性补值方法对缺损值进行填补。同时实际数据中的量纲也对补值方法的准确度产生了较大影响。

文献[1]认为分类是补值技术中关键的一步。Pedro等学者改进了KNN算法,通过熵权矩阵来确定与缺损值最近邻的K个相关数据,并且采用了互信息来确定最终估值的加权系数,从而避免了一般的拟合所带来的残差控制或阈值的问题。熵权矩阵不仅很好地解决了传统KNN算法K值的确定问题,而且由于熵能较好地反映各数据相关性,因此所确定的K个最近邻数据也较准确,实验结果也证明了这点。但是在处理高维海量数据时,由于没有对数据集进行精简,因此算法的复杂度将成为一大问题。

文献[2]介绍了最小二乘回归算法,并提出了基于前馈神

到稿日期:2010-03-22 返修日期:2010-06-19 本文受浙江省自然科学基金项目(Y105118)资助。
陈 欢(1987—),男,硕士生,主要研究方向为数据挖掘、云计算等,E-mail:godchenhuan@163.com;黄德才(1958—),男,博士,教授,博士生导师,主要研究方向为数据挖掘、网格调度、供应链管理等(通信作者)。

经网络模型模糊感知器回归分析法、基于 BP 算法神经网络的回归分析法、基于遗传算法的回归分析法。该文从理论和实践两方面证明了基于神经网络的回归在某些领域优于传统的回归方法。随着神经网络学习时间的延长，神经网络回归系数不断提高，剩余标准差不断减小。而且对于非线性回归问题，神经网络的回归优势将体现得更明了。但是正如文献[3]所说的，普通的 BP 神经网络在误差函数和适应度函数上均存在问题。同时，BP 神经网络采用的是有导师的学习方法，受操作人员的影响较大。

文献[4]采用了双阈值的自适应聚类算法，它通过阈值 C 控制每个类的数据个数都大于 C，使算法受孤立点的影响减小且保证了每个类的数据体个数；通过阈值 r 使得每个类内部数据体的相似度达到精度要求。作者联系类中心和新加入数据的贡献，充分利用数据间的相关性，通过熵值理论确定双阈值的值，从而较好地实现了算法的自适应。但是在类中心和新数据点的关联程度上还有待提高，因为阈值自适应程度差将出现慢收敛等问题，将影响算法的复杂度，这也是自适应聚类算法有待解决的一个重要问题。

鉴于欧氏距离在量纲与变异性上的缺陷，杨涛等学者在文献[5]中采用了马氏距离选择最近邻。他们充分考虑到了不同数据体之间的相关性及不同属性之间的相关性，并且运用信息论中熵的概念确定各属性关于缺损值的加权系数，从而估计缺损值。虽然实验结果证明了该算法优于 KNN 算法及 SKNN 算法，但作者的仿真仅仅建立在数据集极小的情况下，同时马氏距离的存在性也没得到研究，KNN 算法中出现的很多问题，比如 K 的确定、对噪声的抗性等，没有得到根本解决。

文献[6]针对高维变量及预测因子间的相关性给预测带来的困难，首先对具有分散特性的样本进行谱系法聚类，缩减数据样本，然后通过贡献率和累计贡献率[7]进行主成分分析降维，减小神经网络的运行压力。为了解决因子间的相关性，作者又使用马氏距离辨别法判断预测因子属于哪个类，最后通过神经网络进行预测。虽然作者认识到了影响预测的主要因素，但是其没有深入研究，只是通过一些效率低下的算法进行拼接，试图达到高效的预测目的。整个算法的各个部分都有待深入研究。

文献[8]认为，自组织特征映射网络族(SOFMF)可以通过族内成员间互相协作、互相纠正，可以更好地实现聚类。但是，为了避免现实数据属性间的相关性，作者企图使用基于欧式距离的拓扑相似阵来解决。从欧式距离的定义来看，由于其量纲问题，在处理高维数据上的问题也显而易见；同时，使用网络族也会使原本时间复杂度较高的 SOFM 网络背上更沉重的包袱。

由于粗糙集理论能有效地处理不精确、不完整等各种不完备信息，并能从中揭示潜在的规律，因此它在数据挖掘等众多领域得到了广泛的应用[9]。文献[10]利用下近似集的性质对不完备信息进行初次填补，然后根据各属性取值的概率分布构造灰数，根据概率取值的大小进行白化处理，即进行二次填补，从而完成了一种基于粗糙集不完备信息数据填补方法。众多基于粗糙集的不完备信息填补方法都主要集中在关联规则的提取上[11,12]，但文献[10]着重利用了粗糙集对不完备信息的处理能力，并根据得到的聚类信息进行数据填补。虽然作者证明了在系统数据和缺损数据分布均比较均匀的情况下，该方法的效果还是可以接受的，但是鉴于现实中的系统数据和缺损数据分布通常并不是均匀的，各属性间和属性内部的数据差异也并不是相近的，所以没有充分利用关联规则，使得该方法很难具有普遍的应用性。

与许多现有算法一样，基于马氏距离的缺损数据补值算法充分利用了实际数据之间的相关性，具有较好的补值效果，但它要求数据的相关性协方差矩阵可逆，使其应用范围受到了极大的限制。本文在改进传统主成分分析方法的基础上，利用矩阵的奇异值分解理论和 Moore-Penrose 广义逆性质，提出了广义马氏距离的概念并对其公式进行了构造，利用数学工具严格地证明了它满足距离定义的 3 条性质；将广义马氏距离运用于传统 SOFM 神经网络，结合熵论知识设计了基于广义马氏距离的 GS 缺损数据补值算法。仿真实验证明了改进的复相关系数倒数赋权法的准确性；GS 算法在补值的精确度和稳定性能上都有很好的效果；广义马氏距离则继承了马氏距离在处理相关性数据上的性能优势，解决了量纲对数据的影响及变量间的变异性，并且它存在于任一数据集，可以很好地替代传统马氏距离在神经网络及数据挖掘等方面的作用。

2 基于广义马氏距离的 GS 补值算法

一个原始采集的数据集通常是非常庞大的，而且它可能存在着大量的无用信息。本文提出的基于广义马氏距离的 GS 补值算法，首先对数据集进行主成分分析，即列约简。改进了复相关系数倒数赋权法，通过它忽略一些与缺损值关联较小的属性。其次，本文又将提出的广义马氏距离与 SOFM 神经网络结合进行聚类，即行约简，达到了在不影响补值效果的前提下对数据集进行约简的目的。最后利用信息论中熵值的概念计算得到最简数据集，从而填补缺损值。

2.1 改进的复相关系数倒数赋权法

在对某一数据集进行实证研究过程中，为了更全面地反映数据集的特征及其内部数据间的关联性，一方面我们为了避免遗漏重要的信息而考虑尽可能多的属性，另一方面所考虑属性的增加不仅增加了问题的复杂性，而且造成了大量信息的冗余。主成分分析是解决上述矛盾的很好工具，它对数据集各属性进行辨析，得出对所研究问题关联度最大的属性集，在保证精确度的同时起到了降维和简化问题的作用。

现在的主成分分析方法主要基于对协方差阵、相关系数阵等关联度度量。本文改进了“复相关系数倒数赋权法”进行主成分分析，它在计算复杂度和准确性上都将优于改进前的复相关系数倒数赋权法。

假设现有数据集

$$G=(G_1,G_2,\cdots,G_n)=\begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1j} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2j} & \cdots & g_{2n} \\ \vdots & \vdots & \ddots & & & \vdots \\ g_{i1} & g_{i2} & & g_{ij} & & g_{in} \\ \vdots & \vdots & & & \ddots & \vdots \\ g_{m1} & g_{m2} & \cdots & g_{mj} & \cdots & g_{mn} \end{bmatrix}$$

$G_j=(g_{1j},g_{2j},\cdots,g_{mj})'$，$G_j$ 的属性名为 k_j。对于属性值 g_{ij}，若有 $g_{ij}=a_1g_{i1}+a_2g_{i2}+\cdots+a_tg_{it}+\cdots+a_mg_{im}$，$t\neq j$，则系数 a_t 越大，说明 g_{it} 表征量的变化越能引起 g_{ij} 的变化，也就说明

了属性 k_t 对属性 k_j 越发重要。因为 g_{ti} 是相互关联的量，所以可将 a_i 看成是因素变量 g_{ti} 组合的权重，根据“复相关系数的倒数赋权法”来确定哪些属性相对缺损值来说是主要属性。

所谓复相关系数倒数赋权法，就是将每一个被选属性 k_i，用其余的属性对它的相关程度——复相关系数 $\rho_{k_i,k_1k_2,\cdots,k_n}$ $(n\neq i)$ 来计算它的权重，简记为 ρ_i，它反映了非 k_i 的哪些属性替代 k_i 的能力。当 $\rho_i=1$ 时，k_i 可以去掉，即可用非 k_i 属性替代 k_i；当 ρ_i 很小时，k_i 不可以去掉，它对缺损值是个重要的属性，即非 k_i 的属性并不能代替它。由此定义权重系数 a_i：

$$a_i=\frac{|\rho_i|^{-1}}{\sum_{j=1}^{k}|\rho_j|^{-1}},k=1,2,3,\cdots$$

式中，$\rho_i=\frac{SSR_i}{SST}=1-\frac{SSE_i}{SST}$，$SST=\sum_{i=1}^{n}(g_{ji}-\overline{g_{ji}})^2$，$SSE_i=\sum_{i=1}^{n}(g_{ji}-\hat{g}_{ji})^2$，$\overline{g_{ji}}=\frac{1}{n}\sum_{i=1}^{n}g_{ji}$，$\hat{g}_{ji}$ 是 g_{ji} 的估计值。

显然，ρ_i 的计算需要一定的时间复杂度，而且 $\hat{g}_{ji}$ 的未知性使得 ρ_i 无法求解或增加了补值的误差。为了解决上述问题，令 w_{ij} 表示属性 k_i 与属性 k_j 的相关系数，它能准确地确定各属性间的相关性，也弥补了复相关系数的缺陷。

$$w_{ij}=\frac{\sum_{t=1}^{m}(g_{ti}-\overline{g_i})(g_{tj}-\overline{g_j})}{\sqrt{\sum_{t=1}^{m}(g_{ti}-\overline{g_i})^2\sum_{t=1}^{m}(g_{tj}-\overline{g_j})^2}}$$

式中，$\overline{g_i}$ 表示第 i 个属性的属性值的平均值。取 $a_i=\frac{|w_{ti}|}{\sum_{j=1}^{k}|w_{ij}|}$ 表示 g_{ti} 的权重。a_i 越大，说明 k_i 与自身的关系相对较大，也就是说它与其他因素的关系越小，当然也就说明该因素越重要了。通过设置阈值 ε，当 $a_i<\varepsilon$ 时判定 k_i 为次要因素，不予考虑。我们得到了属性约简后的新数据集 $R=(R_1,R_2,\cdots,R_p)$，$\{R_i\}$ 为 $\{G_i\}$ 的子集。

2.2 SOFM 神经网络聚类

2.2.1 聚类

聚类是指将物理或抽象对象的集合中相似的对象聚集成一个类，同一类内的对象具有高度相似性，不同类间的对象差距较大。

较常用的聚类算法有：

1）基于分割的聚类算法，如 K-means 算法、PAM 算法等。

2）基于层次的聚类算法，如 AGNES 算法、DIANA 算法等。

3）基于密度的聚类算法，如 DBSCAN 算法、OPTICS 算法等。

4）基于网格的聚类算法，如 STING 算法、WaveCluster[13]算法等。

5）基于模型的聚类算法，如 COBWEB 算法、SOFM 算法等。

它们均通过某种相似性度量来判断不同对象的相似性。同时，学者们经常通过如下指标评价聚类效果：伸缩性、算法效率、适合的数据类型、处理的聚类形状、领域知识的依赖性、对噪声的敏感性、对数据输入顺序的敏感性、处理高维数据能力。

2.2.2 SOFM 神经网络

人工神经网络是人类在对其大脑神经网络的认识和理解的基础上建立的、能够完成某种特定功能的神经网络。它是人脑神经网络的一个数学模型，是人脑神经网络的一个计算机仿真。其主要特点有：

1）信息处理的并行性、信息存储的分布性、信息处理单元的互连性、结构的可能性。

2）高度的非线性、良好的容错性和计算的非精确性。

3）自学习、自组织与自适应性。

1982 年芬兰人 Kohonen 提出了自组织的神经网络（Self-Organizing Feature Mapping），即 SOFM 神经网络，它是一种无导师的、竞争学习神经网络。它通过自身训练，自动对输入模式进行聚类。自组织的过程实际上就是一种无导师的学习，该学习算法称为 Kohonen 学习算法。

其二维阵列拓扑结构如图 1 所示。

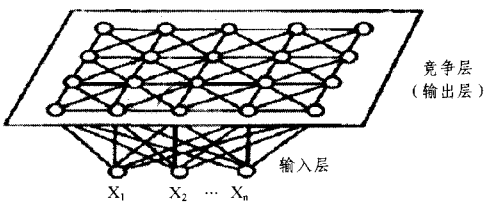

图 1　二维阵列 SOFM 网络模型

它是由输入层神经元数为 n、输出层有 m 个神经元组成的二维平面阵列，并且这两层之间是全互相连的，即每个输出节点与所有输入节点都存在着有权连接，权值表示这种连接的强度。样本进入输入层，经过两层之间连接加权后，最后在输出层得到一个输出值集合。获胜神经元周围的神经元因侧向相互作用也会产生较大响应，于是获胜神经元及其邻域内的所有神经元所连接的权向量根据其距获胜神经元远近对输入向量做不同程度的调整。网络通过自组织的方式不断调整权向量，最终使其成为各输入神经元的聚类中心。它是一种可伸缩性高、领域知识依赖性小、对噪声及输入顺序不敏感、处理高维数据能力强、可对任意聚类形状进行聚类的数值型聚类算法。

2.3 广义马氏距离

在前面分析的 SOFM 神经网络神经元的竞争过程中，获胜神经元是通过距离来判断的，涉及距离最近原则。也正因为传统的 SOFM 神经网络采用的距离不恰当，使得 SOFM 出现了最大的缺陷——算法效率较低，即对复杂数据的处理能力较弱。

欧氏距离有如下缺点：1）没有考虑总体的变异对“距离”远近的影响。显然，一个变异大的总体可能与更多样品近些，即使它们的欧氏距离不一定最近。2）欧氏距离受变量的量纲影响，尤其是任何一个计量单位的改变都会使此距离发生变化，这对多元数据的处理是不利的。而 Chebyshev 距离、Minkowski 距离等也都与量纲有关。相比而言，马氏距离考虑了各变量之间的相关性，可以忽略冗余的数据，同时解决了量纲对数据的影响及变量间的变异性。但是马氏距离在许多情况下是不存在的，因为一个数据集的协方差矩阵并不一定可逆，如 $\begin{bmatrix}1&1\\1&1\end{bmatrix}$，其协方差阵为 $\begin{bmatrix}0&0\\0&0\end{bmatrix}$，不可逆。为了解决上述距离的缺点，本文提出了广义马氏距离的概念，并将其运用于 SOFM 神经网络中进行聚类分析，对数据集进行数据约简。

定义 1 对于任意矩阵 A，必存在唯一矩阵 B，其阶数与 A' 相同，使得 B 同时满足如下方程：(1) $ABA=A$，(2) $BAB=B$，(3)$(AB)'=AB$，(4)$(BA)'=BA$，称 B 为 A 的 Moore-Penrose 伪逆矩阵[14]，记为 A^+。

命题 1 A^+ 很好地保留了 A^{-1} 最基本的几条性质：

1)$(A^+)^+=A$，

2) $AA^+A=A$，

3) $A^+AA^+=A^+$，

4) $AA^+=I$，

5) $A^+A=I$，

6)$(A')^+=(A^+)'$，

7) $A^+=(A'A)^+A'=A'(AA')^+$，

8)$(A'A)^+=A^+(A')^+$。

命题 2(A^+ 的构造) 若 A 是 $m*n$ 复矩阵，且其奇异值分解形式为 $A=UMV'$，则 $A^+=VTU'$，其中 $M=\mathrm{diag}(a_1,a_2,\cdots,a_r)$，$a_i>0$，$r$ 是矩阵 A 的秩，U,V 为正交阵。若 $M(i,j)\neq 0$，则 $T(i,j)=1/M(i,j)$；若 $M(i,j)=0$，则 $T(i,j)=0$。

证明：(1) $AA^+A=(UMV')(VTU')(UMV')=(UMV'VTU')(UMV')=(U(M(V'V)T)U')A=(U(MT)U')A=(UU')A=A$。

(2) 同理，根据矩阵的结合律可以验证其正确性。

(3)$(AA^+)=((UMV')(VTU'))=(U(M(V'V)T)U')$，为一个对称阵，所以$(AA^+)'=(AA^+)$。

(4) 同(3)，$(A^+A)'=(A^+A)$。

所以 $A^+=VTU'$ 满足定义 1 中的 4 个方程，即命题 2 成立。

广义马氏距离：$d_{gm}(X_i,X_j)=[(X_i-X_j)'S^+(X_i-X_j)]^{1/2}$，其中 S 为数据总体 X 的协方差阵。

广义马氏距离的实质就是通过奇异值分解，用协方差阵的伪逆阵代替逆矩阵，避免了马氏距离不存在的情况。它继承了马氏距离的相关性，也解决了协方差阵不可逆的问题。同时，它满足距离的 3 条基本性质，说明它是一个距离。

1) 对称性。显然，$d_{gm}(X_i,X_j)=d_{gm}(X_j,X_i)$成立。

2) 非负性。S 是对称矩阵$=>U=V=>S^+=VTU'=VTV'=>S$ 与 T 合同，又由于 T 为对角线上元素$\geqslant 0$ 的对角阵$=>T$ 半正定$=>S^+$ 半正定$=>d^2_{gm}(X_i,X_j)=(X_i-X_j)^TS^+(X_i-X_j)\geqslant 0$，广义马氏距离非负，当且仅当 $X_i=X_j$ 时 $d_{gm}(X_i,X_j)=0$。

3) 三角不等式。

$$d_{gm}(x,y)+d_{gm}(y,z)\geqslant d_{gm}(x,z)<=>[(x-y)'S^+(x-y)]^{1/2}+[(y-z)'S^+(y-z)]^{1/2}\geqslant[(x-z)'S^+(x-z)]^{1/2} \quad (*)$$

因为 $[(x-z)'S^+(x-z)]^{1/2}=[(x-y+y-z)'S^+(x-y+y-z)]^{1/2}=[(x-y)'S^+(x-y)]^{1/2}+(y-z)'S^+(y-z)]^{1/2}+[(x-y)'S^+(y-z)]^{1/2}+[(y-z)'S^+(x-y)]^{1/2}$

式(*)两边平方后，有

$$2\sqrt{(x-y)'S^+(x-y)(y-z)'S^+(y-z)}\geqslant (x-y)'S^+(y-z)+(y-z)'S^+(x-y)=2(x-y)'S^+(y-z)<=>(x-y)^TS^+(x-y)(y-z)^TS^+(y-z)\geqslant (x-y)^TS^+(y-z)(x-y)^TS^+(y-z) \quad (**)$$

因为 S^+ 与 T 合同，且 T 为对角线上元素$\geqslant 0$ 的对角阵，所以式(**)显然成立。

2.4 信息熵

在信息论中，信息熵是一个信源发出某一消息所含信息量的度量。当某一信源发出的消息越确定时，该信源的信息熵就越小[15]。它是系统无序程度或混乱程度的度量，表示了系统的平均不确定度。而熵值法是一种通过属性数值所提供信息的大小来确定权重系数的一种方法，具有客观性强、评价过程透明性和可再现性好的特点。

对于确定的属性 j，各数据第 j 个属性之间的差异越大，则说明该项指标的相对作用就越大，即其信息量就越大、熵值越小。

数据集

$$T=(T_1,T_2,\cdots,T_s)'=\begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1j} & \cdots & t_{1p} \\ t_{21} & t_{22} & \cdots & t_{2j} & \cdots & t_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ t_{i1} & t_{i2} & \cdots & t_{ij} & \cdots & t_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ t_{s1} & t_{s2} & \cdots & t_{sj} & \cdots & t_{sp} \end{bmatrix}$$

若数据 T_i 的第 j 个属性值 t_{ij} 缺损，则本文运用熵值理论确定 $t_{ij}=w_1t_{1j}+w_2t_{2j}+\cdots+w_qt_{qj}+\cdots+w_st_{sj}$，$(q\neq i)$ 中 w_q 的值，从而达到补值的效果。步骤如下：

1) 计算第 f 个数据的熵值：

$$I_f=-k*p_f*\ln(p_f)$$

式中，$k=\frac{1}{\ln(S)}$，$p_f=\frac{d_f}{\sum_{i=1}^{s}d_f}$，$d_f=(\sum_{l=1}^{p}(d_{fl}-d_{il})^2)^{1/2}$，$(l\neq j)$，$d_f$ 为第 f 个数据 T_f 与缺损数据 T_i 的距离，由计算非缺损属性值得到。

2) 计算第 f 个数据的差异系数：

$r_f=1-I_f$，$f=1,2,\cdots,s$

差异系数是反映数据作用大小的量，其值越大，数据体的作用越大，反之亦然。

3) 计算第 f 个数据的权重系数：

$$w_f=\frac{r_f}{\sum_{f=1}^{s}r_f},\ f=1,2,\cdots,s$$

2.5 GS 补值算法的伪代码描述

步骤 1 计算各属性相对于缺损值属性的权重系数 a_i。

步骤 2 确定阈值 ε，当 $a_i<\varepsilon$ 时，第 i 个属性退出数据集；否则保留。

步骤 3 初始化网络权值 w_i。初始化学习效率参数，定义拓扑领域函数，初始化参数，设置 $k=0$。

步骤 4 检查停止条件。如果成功，执行步骤 8；如果失败，则继续。

步骤 5 对于每个训练样本 x，执行步骤 6。

步骤 6

6.1 计算和输入最佳匹配权值向量：

$$q(x)=\min_{\forall i}\{d_{gm}(x,w_i)\}$$

6.2 对于给定领域 $i\in N_q(k)$ 的所有单元(q 是获胜神经元)，按下式更新权值向量：

$$w_i(k+1)=\begin{cases} w_i(k)+a(k)[x(k)-w_i(k)], & \text{if } i\in N_q(k) \\ w_i(k), & \text{if } i\notin N_q(k) \end{cases}$$

式中，$0<a(k)<1$ (学习率参数)

6.3 调整学习率参数。

步骤 7 设置 $k \leftarrow k+1$，然后转到步骤 4。

步骤 8 对类中心集计算第 f 个数据的熵值 I_f 和差异系数 r_f。

步骤 9 计算第 f 个数据的权重系数 w_i'。

步骤 10 计算缺损值。

算法流程说明：步骤 1 和步骤 2 通过改进的复相关系数倒数赋权法进行主成分分析（即属性约简）；步骤 3－步骤 7 将广义马氏距离运用于 SOFM 神经网络进行聚类，进行数据约简；步骤 8－步骤 10 通过熵论对缺损值进行填补。

3 仿真实验

本文对 uci 数据库中 wine 数据库进行模拟。仿真结果证明了通过改进的复相关系数倒数赋权法确定主要属性的准确率是相当高的。我们在对 wine 进行计算时，算得任何一个属性缺损时，其他属性的权重系数 α_i 均大于 $0.14>>1/14$（14 为属性个数）。考虑到 wine 数据库是专家们长期对意大利的 3 个酒种进行化学分析的结果，因此每个属性均为主要属性，非常符合现实。

为了证明基于广义马氏距离的 GS 补值算法优于基于欧式 SOFM 神经网络的补值算法，我们对 wine 数据集进行了如下模拟（神经网络模型一致，竞争层采用 7 * 4 结构，拓扑函数采用 hextop，训练步数为 100，且初始权值为一组（0，1）内的随机数）：

相对误差率 η＝|真实值－预测值|/真实值

平均相对误差率＝$(\eta_1+\eta_2+\cdots+\eta_n)/n$

则有相对误差率如表 1 所列，标准差如表 2 所列。

表 1 相对误差率

缺损数据数	5 个	10 个	15 个	20 个
GS 补值算法的平均相对误差率 a	0.109445497	0.076923304	0.108560169	0.12242091
基于 SOFM 补值算法的平均相对误差率 b	0.144363009	0.129142907	0.152851885	0.17285536
b－a	0.034917512	0.052219603	0.044291716	0.05043457

表 2 标准差

缺损数据数	5 个	10 个	15 个	20 个
GS 补值算法的标准差 c	0.061538746	0.068792387	0.129293104	0.151219861
基于 SOFM 补值算法的标准差 d	0.100614256	0.111919471	0.155634531	0.169128083
d－c	0.039075510	0.043127084	0.026341427	0.017908222

可以看出，本文提出的基于广义马氏距离的 GS 缺损值补值算法的精确度是相当高的，同时稳定性也相当优越。由于刘泉风等学者在文献[16]中已经证明了基于 SOFM 神经网络的聚类算法在综合性能上明显优于其他各类聚类算法中的优秀算法，如 Clarans 法（基于分割）、CURE 法（基于层次）、DENCLUE 法（基于密度）、STRING 法（基于网格）、CLIQUE 法（混合聚类），而实验结果告诉我们 GS 算法优于基于 SOFM 网络的补值算法，因此 GS 算法性能也超过了当前大部分基于聚类的补值算法。同时，由于使用广义马氏距离及改进的“复相关系数倒数赋权法”，将使得基于广义马氏距离的 GS 补值算法受噪声数据的影响更小，其优势将更能体现。

结束语 根据实际应用中的数据大多具有相关性的特点及传统马氏距离不一定存在的局限性，本文从数据集内部的相关性出发，利用矩阵的奇异值分解理论和 Moore-Penrose 广义逆性质，提出了广义马氏距离的概念，且利用数学工具严格地证明了它满足距离定义的 3 条性质。本文改进了复相关系数倒数赋权法并进行主成分分析（属性约简），将广义马氏距离运用于 SOFM 网络进行聚类，最终得到一个行列都约简过的最大化保留原数据集特征的新数据集，最后通过熵值法对缺损值填补。仿真实验证明了广义马氏距离在处理相关性及量纲数据时的优势，它继承了马氏距离的优点，也克服了马氏距离不存在的局限性，可以很好地替代传统马氏距离在神经网络及数据挖掘等方面的作用；改进的复相关系数倒数赋权法可以很好地运用于主成分分析；GS 补值算法在补值的精确度和稳定性能上较其他算法有更好的效果。

参 考 文 献

[1] Garcia-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(7-9): 1483-1493

[2] 冯勤. 基于回归数据挖掘预测系统的分析与研究[D]. 天津：天津大学，2005

[3] 林香，姜青山，熊腾科. 一种基于遗传 BP 神经网络的预测模型[J]. 计算机研究与发展，2006，43(Suppl.)：338-343

[4] Cheng Ching-hsue, Wei Liang-ying. Data spread-based entropy clustering method using adaptive learning[J]. Expert Systems with Applications, 2009, 36: 12357-12361

[5] 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

[6] 林树宽，张冬岩，李文贤，等. 基于聚类和主成分分析的神经网络预测模型[J]. 小型微型计算机系统，2005，26(12)：2160-2163

[7] 李文贤. 基于数据挖掘的高炉煤气流分布模型研究[D]. 沈阳：东北大学，2003

[8] 江波，张黎. 基于多维自组织特征映射的聚类算法研究[J]. 计算机科学，2008，35(6)：181-185

[9] Pawlak Z, Grzymal-Busse J, Slowins R, et al. Rough sets[J]. Communications of the ACM, 1995, 38(11): 89-95

[10] 张忠林. 基于粗糙集的不完备信息系统统计批判填补方法[J]. 计算机应用，2007，27(6)：1385-1387

[11] 王国胤，于洪，杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报，2002，25(7)：759-766

[12] 潘巍，王阳生，杨宏戟. 粗糙集理论中新的针对不完备信息系统的处理方法研究[J]. 计算机科学，2007，34(6)：158-161

[13] Sheikholesland G, Chatterjee S, Zhang A. WaveCluster: A MultiResolution Cluster Approach for Very Large Spatial Databases [C]// Proc. 1998 Int. Conf. Very Large Data Bases(VLDB'98). 1998: 428-439

[14] 陈公宁. 矩阵理论与应用（第二版）[M]. 北京：科学出版社，2007：192-206

[15] 王洪春，彭宏. 一种基于熵的聚类算法[J]. 计算机科学，2007，34(11)：178-200

[16] 刘泉凤，陆蓓. 数据挖掘中聚类算法的比较研究[J]. 浙江水利水电学报，2005，17(2)：55-58