

基于单输出子网迭代学习的缺失值填补方法

关李晶, 何洁帆, 张立勇*, 闫晓明

(大连理工大学 控制科学与工程学院, 辽宁 大连 116024)

摘要: 现实世界中不完整数据广泛存在,通常会增加数据分析与挖掘的难度,影响分析结果的可靠性.合理填补不完整数据的缺失值已经成为当前数据分析和挖掘中一个非常重要的环节.采用不完整数据属性关联建模的方法填补缺失值,鉴于不完整数据属性关联关系的复杂性,使用具有强大学习能力的单输出子网模型对不完整数据的缺失值进行填补,并针对由于缺失值的存在所导致的模型输入不完整问题,从缺失值的对待与描述切入,提出一种基于单输出子网迭代学习的缺失值填补方法.实验结果表明,通过单输出子网迭代学习能够取得更精确的填补结果,验证了所提方法的有效性.

关键词: 不完整数据;缺失值填补;单输出子网;基于模型的填补;迭代学习

中图分类号: TP181

文献标识码: A

doi: 10.7511/dllgxb202204012

0 引言

近年来,随着计算机技术和信息技术快速发展,各行业各领域每天产生海量数据,大到军事、科技研究,小到日常生活,数据无处不在.但是现实世界中,数据缺失问题非常普遍.在采集、传输以及存储等环节可能出现的数据丢失,都会导致最终数据集的不完整.数据缺失增加了数据分析与挖掘的难度,并可能影响分析结果的准确性.因此,合理填补缺失值是数据处理过程中非常重要的一个环节.

不完整数据的处理可以分为删除法和填补法两类.删除法,就是直接删除含有缺失值的样本或属性,从而形成一个完整数据集^[1].删除法简单、易于操作,但是破坏了数据的规模,降低了可用数据的数量.填补法利用不完整数据集中存在的属性值信息,通过相关的技术方法或算法为缺失值寻求一个尽可能合理的近似值填补其空缺.根据技术方法的差异,填补法常被分为基于统计学的填补方法和基于机器学习的填补方法^[2-3].

常用的统计学填补方法有均值填补^[4]、热卡填补^[5]、回归填补^[6]、多重填补^[1,7-8]以及 EM 填补^[9-10]等.基于机器学习的填补方法则是利用适当的算法,构造数学模型或者模型组,从数据集已

知数据中寻找样本之间、属性之间的关联关系,通过模型输出填补缺失值.常用的机器学习填补方法有 KNN 算法^[11]、支持向量机^[12-13]、神经网络^[14-16]、遗传算法^[17]以及聚类算法^[18-19]等.为了追求更高的填补精度,也有很多学者尝试将多种算法结合来填补缺失值^[20].Aydi 等^[21]将支持向量机、遗传算法与 FCM 算法相结合应用于缺失值填补;Abdella 等^[22]采用遗传算法训练神经网络以提升模型的逼近能力,使缺失值的填补精度得到了提高.

目前,基于模型的填补方法大多采用数据集中的完整样本来训练模型,或是使用预填补方法将不完整数据完整化后来训练模型.当不完整数据缺失率较高时,完整样本训练方案往往会出现训练样本不足导致模型精度低甚至无法建模的情况;而预填补方法中模型的构建过分依赖预填补的结果.为此,本文从缺失值的对待与描述切入,将缺失值视作变量,使得不完整数据集中所有样本均能参与模型训练.鉴于单输出子网结构简洁并具有强大的学习能力,逐一构建子网模型挖掘单个属性与其他属性之间的关联关系,进而实现对缺失值的填补.

1 基于神经网络的缺失值填补方法

神经网络是机器学习中的热点方向. 神经网络具有很强的非线性学习能力, 可以有效发掘数据集中的隐含信息, 从而降低数据集的不完备性^[23].

基于神经网络的填补方法通常先根据缺失数据的分布, 针对每种缺失属性组合构建模型, 并基于完整样本求解网络参数, 挖掘数据属性间的关联关系, 然后将不完整样本作为模型输入, 得到的模型输出作为缺失值的填补结果. 上述方法在数据缺失率较高或缺失数据分布比较复杂的情况下存在一定的局限性.

本文从简化模型结构和缺失值的对待与描述切入, 利用单输出子网结构降低模型数量、提高建模效率, 并引入迭代学习方案提高网络训练集的规模以及存在属性值信息的利用率, 进而改善神经网络对缺失数据的填补效果.

2 基于子网迭代学习的填补方法

基于多层感知机的不完整数据填补方法分别为每种缺失属性的组合建立一个网络, 而单输出子网迭代学习的方法通过简化整体网络架构, 解决模型数量过多导致的建模复杂低效问题.

2.1 单输出子网结构

2001 年, 王秀坤等^[24]提出了单输出子网结构, 并通过理论和实验分析表明相较于多输出网络, 单输出子网结构具有更强的学习能力. 多输出网络的输入信息通过隐含层同时作用到所有输出神经元, 相比而言, 单输出子网结构更能清晰获得各输入对任一输出的作用. 因此, 本文采用单输出子网结构为不完整数据建立任一属性与其他属性间的关联关系.

单输出子网的构造方法如图 1 所示, 将 n 维输出的神经网络拆分成 n 个单输出子网, 形成

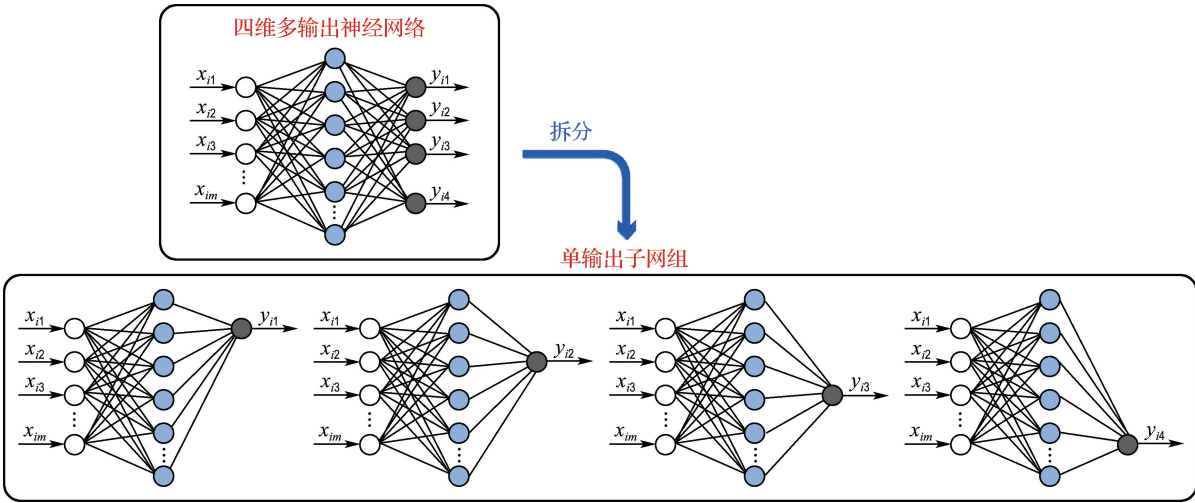


图 1 单输出子网结构

Fig. 1 The structure of single output sub-network

一个由 n 个子网组成的单输出子网组. 子网的输入层、隐含层神经元个数与多输出网络保持一致, 不同之处在于子网仅有一个输出层神经元, 子网的输出与各输入单独通过连接权重相互关联.

神经网络的参数训练如下:

$$\frac{\partial E_t}{\partial w_{ij}^l}=0 \tag{1}$$

式中: E_t 表示总误差, w_{ij}^l 表示第 $l-1$ 层第 i 个神经元与第 l 层第 j 个神经元间的权重. 式(1)在多输出神经网络上可展开为

$$\frac{\partial E_t}{\partial w_{ij}^l}=\frac{1}{n}\left(\frac{\partial E_1}{\partial w_{ij}^l}+\cdots+\frac{\partial E_p}{\partial w_{ij}^l}+\cdots+\frac{\partial E_n}{\partial w_{ij}^l}\right)=0 \tag{2}$$

其中 E_p 表示网络第 p 个输出的误差.

多输出神经网络在满足式(2)时停止训练, 尽管各个输出误差对权重偏导数之和为 0, 但无法确保每一个偏导项均为 0. 而单输出子网组中的每一个子网都使用梯度下降法单独进行训练, 需满足式(3)所示的所有目标才能停止训练:

$$\frac{\partial E^1}{\partial w_{ij}^{pl}}=0, \cdots, \frac{\partial E^p}{\partial w_{ij}^{pl}}=0, \cdots, \frac{\partial E^n}{\partial w_{ij}^{pl}}=0 \tag{3}$$

式中: E^p 表示第 p 个子网的输出误差, w_{ij}^{pl} 表示第 p 个子网第 $l-1$ 层第 i 个神经元与第 l 层第 j 个神经元间的权重.

因此, 从误差函数取极值的角度, 单输出子网

组的必要条件比多输出神经网络更加充分,与多输出神经网络相比,单输出子网组应具有更强的学习能力.

2.2 迭代学习训练方案

依次以不完整数据集每一个存在缺失值的属性作为输出、其他属性作为输入,来构建单输出子网组模型,其中每个子网用来拟合相应输出属性与其他属性之间的关联关系.最终需要构造的子网个数不会超过数据集的属性个数.

在单输出子网组模型的基础上,针对模型输入的不完整性,从缺失值的对待和描述切入,提出一种基于单输出子网迭代学习的不完整数据建模及填补方法.在单输出子网组训练过程中,将缺失值当作系统级变量,网络参数与缺失值变量交替更新,即缺失值变量根据每次迭代时与之对应的子网模型输出进行动态更新.随着迭代学习的深入,缺失值的填补与模型参数的训练相互作用、协同提升,当模型训练结束时随之完成对缺失值的填补.基于迭代学习的填补方法使数据集中的所有完整和不完整样本都能参与模型训练,不仅增加了训练样本的数量、提高了存在属性值信息的利用率,还能有效解决由于缺失值的存在导致模型输入不完整问题.

设不完整数据集 $\mathbf{X}$ 的样本个数为 n ,属性个数为 m ,令 $\mathbf{X}_M$ 表示所有缺失值构成的集合,基于单输出子网迭代学习的属性关联建模和缺失值填补流程如下:

- (1)依次以数据集每一维属性作为输出,其他属性作为输入,构建 m 个单输出子网,对各子网的参数随机初始化,采用属性均值初始化缺失值变量 $\mathbf{X}_M^{(0)}$.给定收敛阈值 $\epsilon>0$,置子网序号 $i=1$,迭代次数 $r=0$.
- (2)更新后的数据集的所有样本参与第 i 个子网的训练,使用随机梯度下降法更新网络参数.
- (3)利用上一步子网 i 训练后的输出填补第 i 维属性上的缺失值,更新数据集,令 $i=i+1$.
- (4)若 $i\leq m$,返回步骤(2),训练下一个单输出子网;若 $i>m$,则子网组中的所有子网模型已完成一轮训练,数据集中所有缺失值完成一次更新,得到 $\mathbf{X}_M^{(r+1)}$.
- (5)判断迭代收敛条件 $\|\mathbf{X}_M^{(r+1)}-\mathbf{X}_M^{(r)}\|_2\leq\epsilon$,若成立,退出迭代循环;若不成立,令 $r=r+1,i=1$,返回步骤(2).

3 实验及结果分析

3.1 实验数据集

从机器学习数据库 UCI 中选取如表 1 所示的 6 个数据集,在完整数据集的基础上通过人工缺失来验证所提方法的填补效果.将缺失率 θ 分别设为 0.05、0.10、0.15、0.20、0.25、0.30,按照给定缺失率随机去除属性值,从而得到不完整数据集.

表 1 实验数据集
Tab.1 Experimental datasets

名称	样本个数	属性个数
Iris	150	4
Seed	210	7
Blood	748	4
Wine	178	13
Glass	214	9
Abalone	4 177	9

3.2 评价指标

以平均绝对百分比误差 E_{map} 作为衡量填补精度的性能指标,如下:

$$E_{\text{map}}=\frac{1}{s}\sum_{i=1}^s\frac{|\hat{x}_i-x_i|}{x_i}$$
 (4)

其中 s 表示缺失值的个数, $\hat{x}_i$ 表示填补值, x_i 表示实际值.

3.3 实验方案

为研究不同网络结构和模型训练方案对不完整数据填补效果的影响,本文开展 4 种不同神经网络填补方法的对比实验,具体如下:

- (1)单输出子网填补方法(SONN).该方法由建模和填补两阶段组成.建模阶段,将各属性依次作为输出,其他属性作为输入,建立 m 个子网,基于完整样本求解子网模型参数.填补阶段,将不完整样本根据缺失情况输入对应的子网模型,利用模型输出填补缺失值.
- (2)子网迭代学习填补方法(SONN+IL).构造 m 个单输出子网模型,利用均值填补初始化缺失值,让所有样本参与子网训练.子网训练和填补流程在 2.2 节中已有具体描述,将缺失值视为系统级变量,使其在迭代过程中动态更新,子网模型训练结束的同时完成对缺失值的填补.
- (3)自编码器填补方法(AE).该方法也由建模和填补两阶段组成.建模阶段,构造一个输入节点和输出节点个数均为 m 的自编码器模型,并基于完整样本训练网络.填补阶段,将不完整样本输

入模型,以模型对应输出填补缺失值.

(4)自编码器迭代学习填补方法(AE+IL).
首先采用均值填补初始化缺失值,让所有样本参与自编码器的训练,其中缺失值作为变量随着训练动态更新.

训练算法选用随机梯度下降法.在实验中,对于学习率、动量因子、隐含层个数、隐含层神经元个数等超参数,选取验证集重构误差最小的超参数组合.

3.4 实验结果

针对实验数据集,在每个缺失率下随机生成 5 个不同的不完整数据集,分别开展实验,取 5 次实验的 E_{map} 平均值,结果如图 2~7 所示.

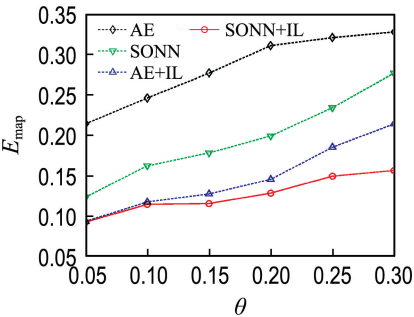


图 2 不完整 Iris 数据集的实验结果
Fig. 2 The experimental results on incomplete Iris datasets

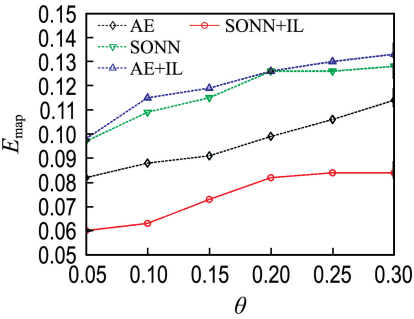


图 3 不完整 Seed 数据集的实验结果
Fig. 3 The experimental results on incomplete Seed datasets

对于 Wine 数据集,缺失率为 0.30 时随机生成的不完整数据集中几乎无完整样本,AE 和 SONN 方法无法实施.

3.5 实验结果分析

由图 2~7 可知,除 Wine 数据集在缺失率为 0.05 的情况外,最优填补结果都来源于本文提出的 SONN+IL 填补方法,表明了单输出子网迭代学习在填补性能方面的优越性.

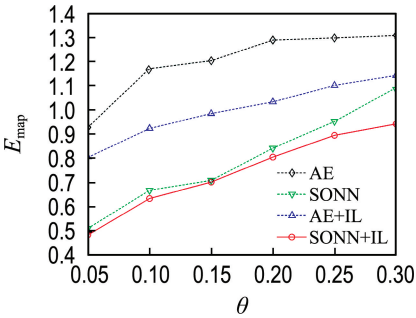


图 4 不完整 Blood 数据集的实验结果
Fig. 4 The experimental results on incomplete Blood datasets

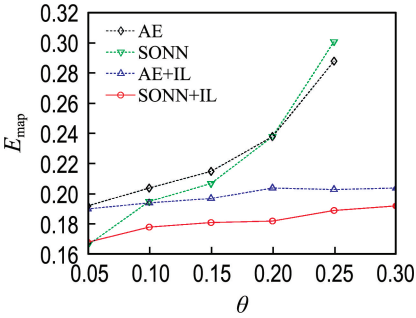


图 5 不完整 Wine 数据集的实验结果
Fig. 5 The experimental results on incomplete Wine datasets

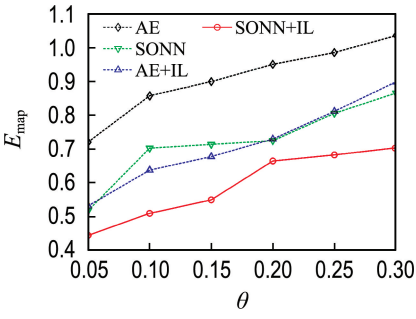


图 6 不完整 Glass 数据集的实验结果
Fig. 6 The experimental results on incomplete Glass datasets

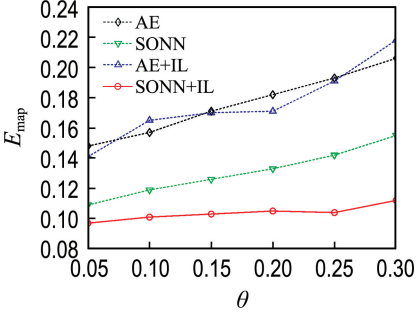


图 7 不完整 Abalone 数据集的实验结果
Fig. 7 The experimental results on incomplete Abalone datasets

(1) 网络结构对比

比较 SONN 和 AE 的填补结果可见,除 Seed 和 Wine 数据集外,在其他 4 个数据集上都是 SONN 方法的填补误差较小.对比 SONN+IL 和 AE+IL 两种方法的实验结果,前者的 E_{map} 更小.说明相比于自编码器,单输出子网通常能更精确地拟合不完整数据属性间的关联关系.单输出子网依次拟合某一属性与其他属性之间的关联关系,而自编码器一次性构建所有属性之间的相互关系,子网组模型更具针对性.另外,自编码器具有高度自跟踪特性,填补效果容易受到预填补结果的影响.

(2) 训练方案对比

比较 SONN 和 SONN+IL 的填补结果可见,除 Wine 数据集在缺失率为 0.05 的情况下,使用迭代学习训练方案的填补方法在各不完整数据集上都取得了小的填补误差;对比 AE 和 AE+IL 的填补结果,除 Seed 和 Abalone 数据集外,AE+IL 的填补误差都小于 AE.上述结果表明了迭代学习训练方案的优越性.迭代学习训练方案将缺失值作为变量,能使不完整样本也参与网络训练,通过增加训练样本数量提高模型拟合精度;另一方面,缺失值在迭代过程中动态优化,从而能降低预填补对建模所产生的影响.

4 结 语

本文提出了一种基于单输出子网迭代学习的不完整数据缺失值填补方法.采用单输出子网结构刻画不完整数据复杂的属性关联关系,并将缺失值视作系统级变量,采用迭代学习的训练方案,使缺失值与网络参数协同训练,这样,在解决由于缺失值的存在所导致的模型输入不完整问题的同时,还能使存在缺失值的不完整样本也参与模型训练,信息充分、客观的利用提高了神经网络的拟合精度和模型的填补性能.

参考文献:

- [1] HUGHES R A, HERON J, STERNE J A C, *et al.* Accounting for missing data in statistical analyses: Multiple imputation is not always the answer [J]. *International Journal of Epidemiology*, 2019, **48**(4): 1294-1304.
- [2] LIN Weichao, TSAI C F. Missing value imputation: A review and analysis of the literature (2006-2017) [J]. *Artificial Intelligence Review*, 2020, **53**(2): 1487-1509.
- [3] GARCIAREN U, SANTANA R. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers [J]. *Expert Systems with Applications*, 2017, **89**: 52-65.
- [4] MEHROTRA D V, LIU Fang, PERMUTT T. Missing data in clinical trials: Control-based mean imputation and sensitivity analysis [J]. *Pharmaceutical Statistics*, 2017, **16**(5): 378-392.
- [5] MASCONI K L, MATSHA T E, ECHOUFFO-TCHEUGUI J B, *et al.* Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: A systematic review [J]. *The EPMA Journal*, 2015, **6**(1): 7.
- [6] LIU Tianhong, WEI Haikun, ZHANG Kanjian. Wind power prediction with missing data using Gaussian process regression and multiple imputation [J]. *Applied Soft Computing Journal*, 2018, **71**: 905-916.
- [7] HAREL O, MITCHELL E M, PERKINS N J, *et al.* Multiple imputation for incomplete data in epidemiologic studies [J]. *American Journal of Epidemiology*, 2018, **187**(3): 576-584.
- [8] PEDERSEN A B, MIKKELSEN E M, CRONIN-FENTON D, *et al.* Missing data and multiple imputation in clinical epidemiological research [J]. *Clinical Epidemiology*, 2017, **9**: 157-166.
- [9] BASHIR F, WEI H L. Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm [J]. *Neurocomputing*, 2018, **276**: 23-30.
- [10] SUN Huayan, LI Yeli, ZI Yunfei, *et al.* Accelerating EM missing data filling algorithm based on the K-means [C]// *Proceedings - 2018 4th Annual International Conference on Network and Information Systems for Computers, ICNISC 2018*. Wuhan: IEEE, 2018: 401-406.
- [11] KIM M, PARK S, LEE J, *et al.* Learning-based adaptive imputation method with kNN algorithm for missing power data [J]. *Energies*, 2017, **10**(10): 1668.
- [12] SANTOS M S, SOARES J P, ABREY P H, *et al.* Influence of data distribution in missing data imputation [C]// *Artificial Intelligence in Medicine - 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Proceedings*. Vienna: Springer Verlag, 2017: 285-294.
- [13] KIM T, KO W, KIM J. Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting [J]. *Applied Sciences*, 2019, **9**(1): 204.

- [14] ZHANG Qiang, YUAN Qiangqiang, ZENG Chao, *et al.* Missing data reconstruction in remote sensing image with a unified spatial-temporal-spectral deep convolutional neural network [J]. **IEEE Transactions on Geoscience and Remote Sensing**, 2018, **56**(8): 4274-4288.
- [15] CHE Z, PURUSHOTHAM S, CHO K, *et al.* Recurrent neural networks for multivariate time series with missing values [J]. **Scientific Reports**, 2018, **8**(1): 6085.
- [16] 张楠, 程理, 王鹏. 利用支持向量机和人工神经网络填补缺失数据 [J]. 应用数学进展, 2017, **6**(5): 677-684.
- ZHANG Nan, CHENG Li, WANG Peng. The use of support vector machines and artificial neural networks to fill missing data [J]. **Advances in Applied Mathematics**, 2017, **6**(5): 677-684. (in Chinese)
- [17] ORDÓÑEZ GALÁN C, SÁNCHEZ LASHERAS F, DE COS JUEZ F J, *et al.* Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions [J]. **Journal of Computational and Applied Mathematics**, 2017, **311**: 704-717.
- [18] SHANG Qiang, YANG Zhaosheng, GAO Song, *et al.* An imputation method for missing traffic data based on FCM optimized by PSO-SVR [J]. **Journal of Advanced Transportation**, 2018(1879): 1-21.
- [19] HOU Ping, CAI Jiarui, QU Shen, *et al.* Estimating missing unit process data in life cycle assessment using a similarity-based approach [J]. **Environmental Science and Technology**, 2018, **52**(9): 5259-5267.
- [20] DEB R, LIEW A W C. Missing value imputation for the analysis of incomplete traffic accident data [J]. **Information Sciences**, 2016, **339**: 274-289.
- [21] AYDILEK I B, ARSLAN A. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm [J]. **Information Sciences**, 2013, **233**: 25-35.
- [22] ABDELLA M, MARWALA T. The use of genetic algorithms and neural networks to approximate missing data in database [C]// **ICCC 2005 - IEEE 3rd International Conference on Computational Cybernetics - Proceedings**. Mauritius: IEEE Computer Society, 2005: 1511574.
- [23] CHOUDHURY S J, PAL N R. Imputation of missing data with neural networks for classification [J]. **Knowledge-Based Systems**, 2019, **182**: 104838.
- [24] 王秀坤, 张晓峰. 用一组单输出的子网络代替多输出的 BP 网络 [J]. 计算机科学, 2001, **28**(10): 61-63, 18.
- WANG Xiukun, ZHANG Xiaofeng. A new architecture of neural networks array replacing of a multiple outputs BP model networks [J]. **Computer Science**, 2001, **28**(10): 61-63, 18. (in Chinese)

Missing value imputation method based on single output sub-network with iterative learning

GUAN Lijing, HE Jiefan, ZHANG Liyong*, YAN Xiaoming

(School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: Incomplete data exist widely in the real world, which usually increases the difficulty of data analysis and mining, and affects the reliability of analysis results. Therefore, it has become an important issue in data analysis and mining that missing values could be reasonably imputed. The method of attribute association modeling of incomplete data is used for missing value imputation. In view of the complexity of the relations among attributes, a kind of single output sub-network model with strong learning ability is adopted for missing value imputation. Aiming at the problem of incomplete model input caused by the existence of missing values, focusing on the treatment and description of missing values, a missing value imputation method based on single output sub-network with iterative learning is proposed. The experimental results show that the single output sub-network assisted by iterative learning can achieve more accurate imputation results, which verifies the effectiveness of the proposed method.

Key words: incomplete data; missing value imputation; single output sub-network; model-based imputation; iterative learning