计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(4):288−295 [doi: 10.15888/j.cnki.csa.009480]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向销售数据的多项缺失值关联性的增量填补①

刘　智[1], 李　涛[2], 袁　冲[3]

[1](武汉科技大学 计算机科学与技术学院, 武汉 430081)
[2](武汉科技大学 智能信息处理与实时工业系统湖北省重点实验室, 武汉 430081)
[3](武汉海云健康科技股份有限公司 技术管理部门, 武汉 430081)
通信作者: 刘　智, E-mail: 1343693123@qq.com

摘　要: 数据缺失会影响数据的质量, 可能导致分析结果的不准确和降低模型的可靠性, 缺失值填补能减低偏差方便后续分析. 大多数的缺失值填补算法, 都是假设多项缺失值之间是弱相关甚至无相关, 很少考虑缺失值之间的相关性以及填补顺序. 在销售领域中对缺失值进行独立填补, 会减少缺失值信息的利用, 从而对缺失值填补的准确度造成较大的影响. 针对以上问题, 本文以销售领域为研究目标, 根据销售行为的多维度特征, 利用不同模型输出值的空间分布特征特性, 探索多项缺失值的填补更新机制, 研究面向销售数据多项缺失值增量填补方法, 根据特征相关性, 对缺失特征排序并用已填补的数据作为信息要素融合对后面的缺失值进行增量填补. 该算法同时考虑了模型的泛化性和缺失数据之间的信息相关问题, 并结合多模型融合, 对多项缺失值进行有效填补. 最后基于真实连锁药店销售数据集通过大量实验对比验证了所提算法的有效性.

关键词: 缺失值处理; 增量填补; 多模型混合; Stacking 算法; 药店销售

引用格式: 刘智,李涛,袁冲.面向销售数据的多项缺失值关联性的增量填补.计算机系统应用,2024,33(4):288–295. http://www.c-s-a.org.cn/1003-3254/9480.html

Incremental Filling of Multiple Missing Value Correlations for Sales Data

LIU Zhi[1], LI Tao[2], YUAN Chong[3]

[1](School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China)
[2](Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430081, China)
[3](Technical Management Department, Wuhan Haiyun Health Science & Technology Co. Ltd., Wuhan 430081, China)

Abstract: Missing data affects the quality of the data, which may lead to inaccurate results and reduce the reliability of the model. Missing value filling reduces the bias and facilitates subsequent analysis. Most missing value filling algorithms assume a weak correlation or even no correlation between multiple missing values, with little consideration of the correlation between missing values and the order of filling. Independent filling of missing values in the sales domain reduces the utilization of missing value information, which has a greater impact on the accuracy of missing value filling. To address the above problems, this study takes the sales field as the research objective and explores the updating mechanism of multiple missing values based on the multidimensional characteristics of sales behavior and the spatial distribution characteristics of output values of different models. In addition, the work studies the incremental filling method of multiple missing values of sales data, which is based on the correlation of features, orders the missing features, and fuses the already-filled data as an information element to incrementally fill in the following missing values. The algorithm also takes into account the generalization of the model. The algorithm takes into account the generalization of the model and the information correlation between the missing data and combines with multi-model fusion to effectively

① 基金项目: 武汉市重点研发计划 (2022012202015070); 武汉东湖新技术开发区“揭榜挂帅”项目 (2022KJB126)
收稿时间: 2023-10-18; 修改时间: 2023-11-15, 2023-12-15; 采用时间: 2023-12-20; csa 在线出版时间: 2024-03-04
CNKI 网络首发时间: 2024-03-08

fill multiple missing values. Finally, the effectiveness of the proposed algorithm is verified by a large number of experimental comparisons based on a real-chain drugstore sales dataset.

Key words: missing value handling; incremental filling; multi-model hybrid; Stacking algorithm; drugstore sales

在数字时代, 在对数据进行分析和挖掘时, 如何有效地处理缺失值是一个重要的挑战. 对于类似销售领域方面的缺失值填补, 销售数据反映出个体行为信息, 以往的方法往往简单地忽略缺失值或者仅对可观测数据进行插补, 这可能会导致信息的损失和准确性的降低. 而通过挖掘缺失值之间的相关性, 我们可以更准确地估计缺失值, 提高数据的质量和完整性. 在销售领域, 如何合理地进行缺失值的填补, 是研究这个课题的首要任务.

缺失值填补按填补方法可分为统计填补和机器学习填补[1]. 统计填补是针对数据本身分布信息进行分析和似然估计等, 常见的统计学填补方法有均值填补[2]、热卡填补[3]、回归填补[4] 以及多重填补等. 基于机器学习的缺失填补方法是利用适当的算法, 构造相关模型, 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. 通常关于机器学习的填补方法有 K 近邻填补[5]、K-means 填补[6]、决策树[7]等. 统计填补和机器学习填补使用的场景不同. 面临海量数据, 很难使用传统的学习方法来处理缺失数据[8−13], 因为大部分统计学习是基于线性关系并且需要大量标签数据进行有监督学习, 依赖于预定义的模型和假设, 缺乏灵活性. 而机器学习可以进行交叉验证与调参进行模型优化, 其模型具有更强的泛化能力, 能够识别出异常数据和冗余特征[14]. 它可以对大部分新输入进行理解性识别. 所以本文在基于海量数据的基础上, 是采用基于机器学习方式的缺失值填补.

在运用机器学习进行缺失值填补研究中, 文献[15]研究一种基于支持向量机的缺失值填补方法, 该方法与均值填补法、基于决策树回归的填补法相比较, 其准确率更高, 具有更强的拟合能力, 并具有良好的抗噪声能力. 文献[16]提出了 K 近邻算法和用于后插补的期望最大化算法结合 (kEMI) 方法, 在运用 K 近邻找到最佳值 K 的同时, 利用最佳值 K 的邻近数据, 通过学习全局相似性来估计缺失分数. 文献[17]研究出基于 Stacking 集成学习方法构建付费意向预测模型, 通过对比不同基模型组合预测效果确定基模型组合方案, 借助游戏玩家行为数据集验证模型优越性、稳定性和验证可移植性. 文献[18]为了对地面电磁学的缺失数据插补, 比较传统统计方法和监督机器学习方法之间的各种插补技术. 结果证明多种机器学习如 SVR 能大幅度地提升填补性能. 文献[19]研究在处理缺失值的同时进行特征选择的新方法, 以提高模型的学习性能并减少插补的负面影响. 文献[20]研究用决策树和模糊聚类集成的模型处理缺失值的同时进行特征选择的新方法, 以提高模型的学习性能并减少插补的负面影响.

上述众多基于机器学习的缺失值填补方法, 考虑了从特征工程上对特征进行处理、不同缺失率的填补方法和算法模型的优化, 大部分都是假设多项缺失值之间是弱相关甚至无相关, 很少考虑缺失值之间的相关性. 在很多领域中如销售领域, 一些行为特征和物品特征有着强关联性. 而缺失值之间的关联信息很少用于信息填补. 因此, 本文以销售领域为研究对象, 基于 Stacking 融合策略的融合模型充分利用多个经典机器学习模型的优点, 侧重考虑多项缺失值之间的相关性, 用已预测的缺失值作为已知特征, 对剩下缺失值进行增量预测, 研究面对多项缺失值相关性的增量填补方法, 并应用于药店销售的进行实验, 证明了该方法能有效地提升总体的填补缺失率.

1 基于模型混合的增量填补方法

1.1 增量填补算法思路

本文研究在销售领域中, 销售数据和人员基本信息, 经过数据预处理、特征选择与构建得到最终数据集, 在此基础上提出的针对多项缺失值如何填补问题. 而本文提出的填补算法着重研究不同缺失特征之间关系, 缺失值之间存在的信息关系, 会对其自身特征的填补准确度造成一定的影响. 面向多项缺失值的增量填补算法整体架构如图 1 所示.

首先, 将会员基本信息数据和药店销售数据进行数据预处理, 特征统计和聚类等方法获得特征 $A_1, A_2, \cdots, A_m$. 其中 $A_1, A_2, \cdots, A_n$ 特征是含有缺失值待填补的特征. 针对需要填补的特征 $A_1, A_2, \cdots, A_n$, 研究与 A_{n+1},

A_{n+2}, …, A_m 的特征相关度, 并按照相关度从降序排序作为填补顺序. 假设排序后顺序是 A_1, A_2, …, A_n. 然后针对 A_1 特征拆分数据集运用基于 Stacking 融合策略的多模型进行训练, 并用测试集来进行验证. 后面把已填补的 A_1 的特征作为已知特征, 针对 A_2, …, A_n 特征, 研究与 A_1, A_{n+1}, …, A_m 的特征相关度, 后面执行和 A_1 特征一样的步骤, 依次对 A_2, …, A_n 进行填补. 此算法为基于模型混合的增量填补方法.

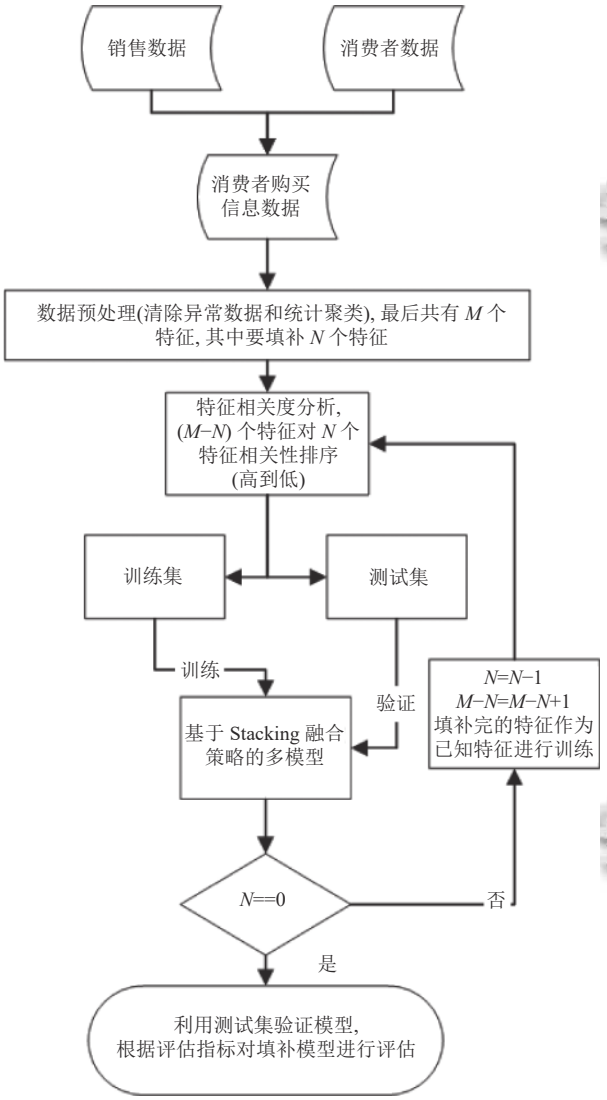

图 1 增量填补流程图

1.2 增量填补关联度算法

文献[21]介绍了多种特征相关性分析方法, 着重提到 Spearman 相关性和 Pearson 相关性的对比. Spearman 适用于非正态分布数据, 相关性基于秩矩阵, 能很好地反映数据的非线性相关性. 当数据中存在离群值时, Spearman 相关性可以提供更稳健的关联度量. 在针对需进行缺失值填补的特征 A_1, A_2, …, A_n 情况下, 要根据其他已知特征与之相关性来决定填补顺序. 结合预处理后药店数据特征信息, 本文运用的特征相关度算法是 Spearman 相关系数, 其计算方法如下.

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{1}$$

其中, r 为 Spearman 相关系数, n 为样本数量, d_i 为数据序列中第 i 个数据对的位次之差.

按式 (1) 分别计算待填补 A_1, A_2, …, A_n 与已知特征 A_{n+1}, A_{n+2}, …, A_m 的相关度. 例如 A_1 的特征相关系数如下:

$$p_v = \sum_{z=n}^{m} \left| r_z^v \right| \tag{2}$$

其中, p_v 代表 v 与其他特征的总体相关度, r_z^v 代表 v 与第 z 个特征 A_z 的相关度.

1.3 基于 Stacking 的融合模型的增量填补

Stacking 算法是一种分层集成的方法. 不同于 Bagging, Boosting 整合同类型模型, Stacking 算法能够集成异质模型. 融合模型通过结合多个模型, 先完成个体学习器训练, 再按照一定的融合策略或投票策略等方式得出最后的结果. 基于 Stacking 的多模型利用融合了场景信息和单模型信息, 一定程度上结合了多个模型的优点, 提高了数据的泛化性和实验结果的准确性. 本文所研究的基于药店销售数据结合会员数据的缺失值填补问题, 选用 Stacking 策略进行融合. 在基于 Stacking 的模型设计上, 在选择基学习器时应当遵循好而不同的原则, 即考虑个体学习器性能好坏的同时, 也要考虑个体学习器的两两不相似性. 本研究先挑选 KNN、SVM、Bagging、AdaBoost、XGBoost、随机森林、GBDT 等模型作为基学习器的候选模型, 分别用来对单独的特征进行填补, 然后挑选出预测性能较好的模型作为基学习器. 经过初级筛选, 发现 KNN、SVM、random forest、GBDT 这 4 个模型在这 8 个模型中表现突出, 本文选取这 4 个表现较优的模型作为基学习器. 结合预测的数据特征, 元学习器为朴素贝叶斯模型 (NBM) 时可以取得较好的填补效果, 故本文选用 NB 朴素贝叶斯模型作为元学习器. 图 2 为多模型 Stacking 融合框架.

图 2 中的混合模型由第 1 层的基模型: K 近邻 (K-nearest neighbor, KNN)、支持向量机 (support vector

machine, SVM)、梯度提升回归树算法 (gradient boosting regression tree, GBRT) 和第 2 层的元模型的朴素贝叶斯模型 (naive Bayesian model, NBM), 具体算法步骤如算法 1.

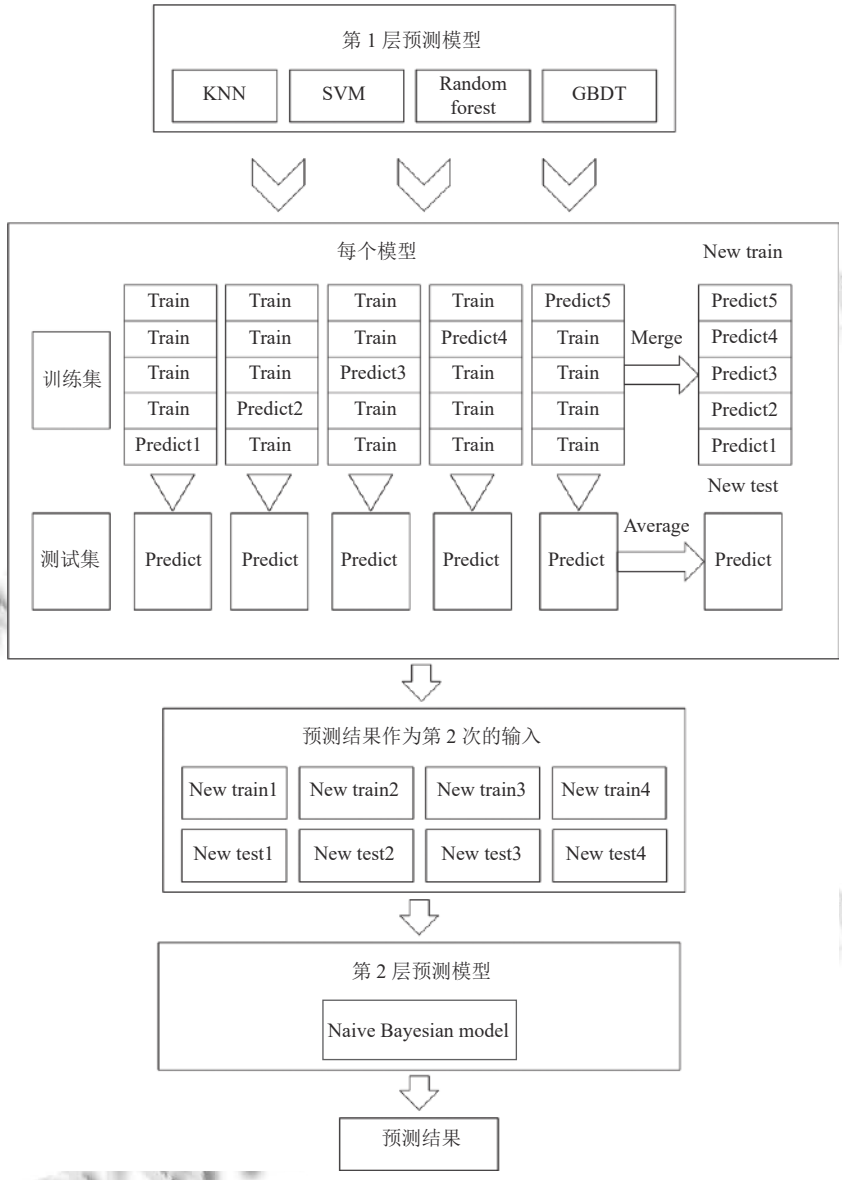

图 2 混合模型整体架构图

算法 1. 面向多项缺失值的增量算法

1) 选择特征相关度最高的作为第 1 个填补的特征, 划分数据集为训练集 T1 与测试集 T2;

2) 将训练集分别输入到基学习器中, 在每个基学习器中采用五折交叉验证, 分别进行 5 次实验, 每次将训练集 T1 分割成 5 等分, 用 4 份训练出模型并对一份训练集预测, 预测结果记为 p*i*, 并且每次训练出的模型对测试集 T2 进行一次预测, 预测结尾记为 z*i*. 把每次预测结果 p*i* 拼接作为最后的训练集 P. 如果填补是数据是离散值, 对预测的 z1, z2, z3, z4, z5 进行投票作为最终的测试集预测结果, 否则, 对预测的 z1, z2, z3, z4, z5 进行结果平均作为最终测试集结果 T;

3) 每个基学习器重复步骤 2), 并把训练集 P 和测试集 T, 作为第 2 层元学习器的数据来源, 得到最终预测结果.

2 实验与分析

2.1 数据来源和数据处理

本文数据来源于某连锁药店近 4 个月的会员消费记录, 结合了会员自身信息、销售信息和药品信息. 去除退货和赠送等噪音数据, 最后有效数据共为 150 594 条. 首先利用 Elastic-net 算法进行特征选择, 该算法是一种联合 Lasso 的 L_1 正则化项和 Ridge 的 L_2 正则化作为惩罚项的线性回归模型, 起到了平衡模型稀疏性和非稀疏性的作用, 减少了由于稀疏性所带来的模型泛化能力不足与信息丢失的情况, 同时缓解了非稀疏

性模型的解释性差和信息冗余的情况.

$$W=\min\left\{\frac{1}{2N}\sum_{i=1}^{N}\left\|X\cdot w-y\right\|\frac{2}{2}+\lambda P_a(w)\right\} \tag{3}$$

其中, $P_a(w)$ 是 Elastic-net 惩罚项, 并且:

$$P_a(w)=\frac{1-a}{2}\|w\|\frac{2}{2}+a\|w\|_1 \tag{4}$$

其中, a 和 λ 均是非负正则化参数, 当 a 为 0 时, 式 (3) 是 Ridge 回归; 当 a 为 1 时, 式 (3) 是 Lasso 回归. 图 3 是特征相关分析系数热力图.

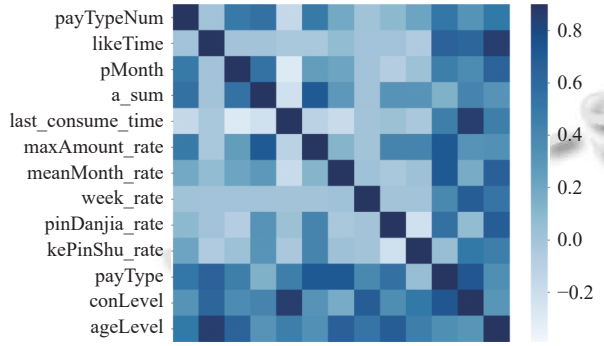

图 3 特性相关性分析热力图

在数据样本存在一些数据特征缺失, 并且一些数据特征在后面进行数据分析, 人物画像是必不可少的. 于是针对样本数据中的会员基本信息包括会员的一些敏感信息如薪资水平和年龄期间和消费行为中的是否使用医保卡数据进行缺失预测填补. 特征选择后的部分特征如表 1 所示, 表 2 为待填补输出特征的缺失率.

2.2 评价标准

因为本文针对的是对薪资水平 (月薪是否大于 4 000)、年龄期间 (18–34, 35–60, 60 以上)、是否用医保卡支付进行缺失填补属于离散值分类问题, 并且数据分布较为均匀, 所以本文选用准确率作为薪资水平和是否用医保卡支付的评价指标和平均准确率增量作为年龄期间的评价指标. 准确率表示在预测数据中, 预测正确数据占预测总数的比例, 作为对比实验的评估标准, 公式如下:

$$accuracy(y,\hat{y})=\frac{\sum_{i=0}^{n_{\text{test}}}1(y_i=\hat{y}_i)}{n_{\text{test}}} \tag{5}$$

平均准确率增量表示在混合模型下, n 个待需填补的特征个数, 分别计算排除第 1 次填补的特征的 n–1 个特征. 在后续实验中增量填补下的准确率记为 $accz$, 在独立模型算法填补下的准确率记为 $accd$, 然后求平均值. 用于衡量不同填补顺序情况下的总体提升的填补效果, 作为消融实验的评估标准, 公式如下:

$$avg\_acc=\frac{\sum(accz-accd)}{n-1} \tag{6}$$

表 1 医药销售缺失值填补实验部分特征集

特征含义	特征名称
总消费额	aNum
消费次数	paySum
会员日消费偏好	weekRate
购买平均单价	meanPDJ
购买最高单价	maxPDJ
购买最低单价	minPDJ
月均消费次数	pMonth
慢病药订单数	MPnum
保健食品订单数	BJnum
医疗器械订单数	Qxnum
中药饮片订单数	ZYnum
会员的年龄期间	ageLevel
会员的薪资水平	salaryLevel
是否医保卡支付	payType

表 2 待填补缺失特征缺失率 (%)

特征名称	缺失率
会员的年龄期间	19
会员的薪资水平	27
是否医保卡支付	21

2.3 实验对比及结果分析

2.3.1 对比实验

为了验证本文融合模型以及增量填补在面对多项缺失值填补的有效性, 本文进行如下对比实验.

采用经典模型和混合增量更新填补模型的对比实验来证明混合增量更新填补带来的效果提升.

主要采用的经典模型有通过构建多个决策树并取其输出的平均值来进行预测的算法随机森林、基于特征之间相互独立的前提假设的贝叶斯算法、通过将数据划分为 K 个簇来达到将数据分组的目标聚类的 K-means 算法、通过迭代地添加决策树来提高预测精度的 GBRT 算法、基于间隔最大化的算法, 通过找到一个最优超平面将数据分隔开, 以达到最小化误分类的目标的 SVM 和基于最近邻距离的 KNN 算法.

表 3 说明, 对同一缺失值进行预测时, 不同模型填补准确率都存在差异, 体现出不同模型的差异化特征, 同时说明不同缺失数据特征对一些经典算法运算产生一定的影响. 本文提出混合增量更新填补模型算法对

各项缺失值的填补准确率均高于单模型的最高填补效率. 单模型考虑到单方面的特征信息, 如距离特征, 统计方差等特征, 而混合增量更新填补模型在一定程度上, 结合了多种模型的特性和特征之间的相关信息, 有效降低单一模型的泛化能力缺陷和特征有效信息缺漏. 此外, 在对单一模型进行训练优化时, 往往会陷入局部最小点, 而局部最小并不代表模型的泛化能力好. 通过多种基学习器进行结合, 可有效避免模型陷入局部最小点. 在全局上, 提高模型泛化能力和稳定性, 达到更好的整体性能.

表 3 经典算法和增量算法对比实验结果 (%)

算法	ageLevel	salaryLevel	payType
KNN	79.34	82.27	82.32
SVM	74.83	84.39	84.47
GBRT	73.21	83.06	81.29
K-means	70.98	80.74	79.69
Naive Bayes	73.66	82.87	84.46
Random forest	78.95	82.39	83.91
Ours	85.73	91.16	89.84

2.3.2 消融实验

为了验证该混合增量填补算法的有效性, 分别进行对基模型和增量更新算法消融的 3 组对照实验.

1) 对混合模型分别移除单个基模型, KNN, SVM, random forest, GBDT 和原混合增量填补模型进行实验对比.

如图 4–图 6 所示, 图 4 代表移除单个基模型对缺失特征 payType 的填补准确比较图. EX_KNN 代表移除了 KNN 基模型的混合增量填补算法. EX_SVM、EX_RF 和 EX_GB 分别代表移除了 SVM、random forest 和 GBDT 基模型的混合增量填补模型. 图 5 和图 6 代表分别对缺失特征 ageLevel 和 salaryLevel 的填补对比实验.

图 4–图 6 体现出单个基模型对混合模型有一定程度的提升, 但影响程度较小. 结合表 3 整个混合模型对比与单独模型还是一定的提升的, 进一步证明了混合模型的优点, 结合了多个模型各自维度特点, 提高了多项缺失值的填补准确率, 侧面反映主要的填补提升来自于增量填补算法.

2) 在混合模型基础下, 移除增量更新算法, 分别对多项缺失值进行独立填补和原混合增量填补进行实验对比.

图 7 代表各缺失特征独立预测和增量填补预测的实验对比. 由于一个缺失填补只用到了已知特征进行填补, 因此增量特征个数为 0, 特征 payType 填补准确率和独立准确率相似, 而对于缺失特征 ageLevel 和 salaryLevel, 结合之前已填补特征信息进行增量填补, 该实验表明增量填补算法的在一定程度上提升其他缺失特征的填补准确性.

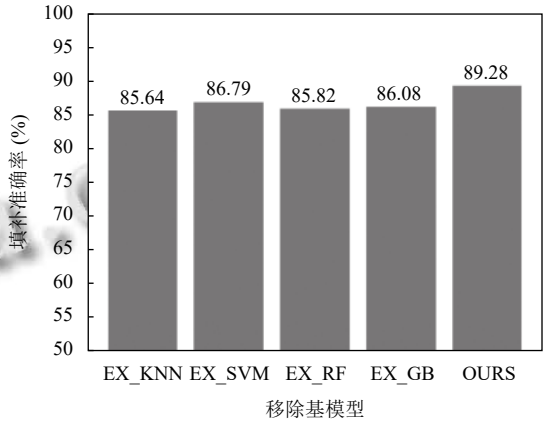

图 4 缺失特征 payType 填补图

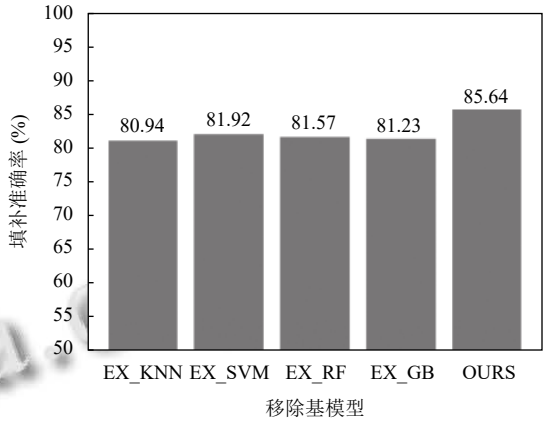

图 5 缺失特征 ageLevel 填补图

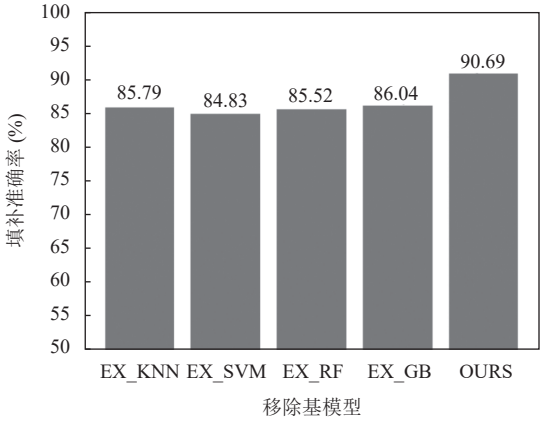

图 6 缺失特征 salaryLevel 填补图

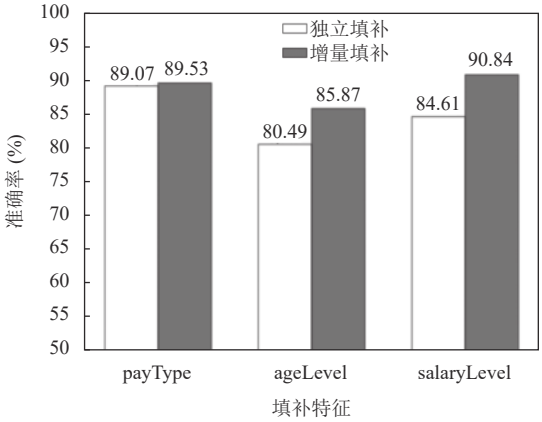

图 7 独立预测和增量预测实验对比图

3) 在混合模型基础下, 随机次序多项缺失特征的增量填补和原混合增量填补进行实验对比.

在混合模型下, 随意排列组合特征的增量预测和根据相关性分析的增量预测实验对比. 随意排列组合方案如 a_s_p 代表 先预测 ageLevel (年龄分类) 再增量预测 salaryLevel (薪资水平), 最后增量预测 payType (是否为医保卡支付). 根据特征相关性分析得到第 1 次填补的顺序, 第 1 次填补 payType, 第 2 次把 payType 作为已知特征再加入输入特征, 再进行特征相关性分析, 获取填补的特征是 ageLevel, 后续再把 ageLevel 作为输入特征. 最后填补的特征是 salaryLevel, 所以填补的特征顺序记为 p_a_s. 此次实验都是在混合模型上的对比, 因此, 用上述第 2 个消融实验中的独立模型作为基准来校验随机组合的增量填补对多项缺失值的整体填补提升率.

图 8 是随意排列组合特征的增量预测和根据相关性分析的增量预测的实验对比结果. 其中 p_a_s 是根据 Spearman 相关性和增量填补过程中得到的填补顺序. 该实验有力地证明了本文提出增量填补算法的有效性, 能够在混合模型对多项缺失值特征预测的基础上再提升总体填补 5%–6% 的准确率.

3 结论与展望

本文研究面向销售数据的多项缺失值关联性的增量填补方法并对国内某药店销售数据和会员基础信息数据中多项缺失值进行实际应用填补. 通过单一模型和融合模型对比发现, 融合模型能够结合多个模型的优势, 尤其在单一模型准确率较高的情况下, 能在一定程度提升填补准确率. 在融合模型的基础上, 通过依据特征相关增量填补和独立填补的实验对比, 证明了增量填补的可行性, 通过依据特征相关增量填补和其他顺序进行填补, 表明了特征相关增量填补的效果显著. 上述研究成果可以丰富对医药等各行业销售数据的多项缺失填补方法, 支撑行为分析和用户画像等工作. 今后的工作中将针对以下问题开展深入探讨. 对面不同缺失率的数据, 是否能结合缺失值可提供信息的重要性能对增量算法再进一步优化. 基于融合模型结构复杂, 对比单一模型, 即使数据集较少的情况下, 也存在训练时间过长问题, 后续研究可进一步对融合模型优化进行深入分析.

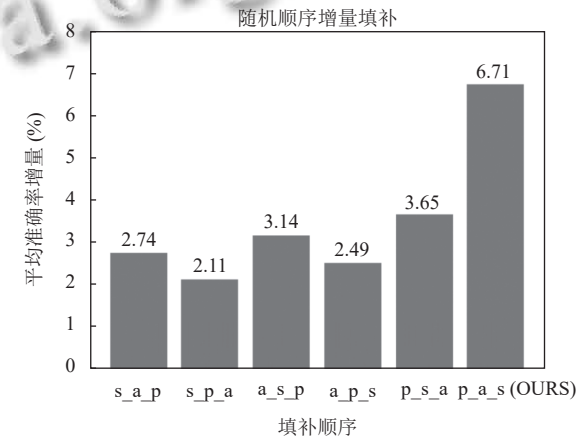

图 8 随机次序的增量填补对比图

参考文献

1 陈娟, 王献雨, 罗玲玲, 等. 缺失值填补效果: 机器学习与统计学习的比较. 统计与决策, 2020, 36(17): 28–32.

2 Batra S, Khurana R, Khan MZ, *et al*. A pragmatic ensemble strategy for missing values imputation in health records. Entropy, 2022, 24(4): 533. [doi: 10.3390/e24040533]

3 余嘉茵, 何玉林, 崔来中, 等. 针对大规模数据的分布一致缺失值插补算法. 清华大学学报 (自然科学版), 2023, 63(5): 740–753.

4 王一棠, 庞勇, 张立勇, 等. 面向盾构机不完整数据的模糊聚类与非线性回归填补. 机械工程学报, 2023, 59(12): 28–37.

5 周敏. 多分类等级量表数据缺失填补方法的比较研究 [硕士学位论文]. 沈阳: 中国医科大学, 2022. [doi: 10.27652/d.cnki.gzyku.2022.000407]

6 Ahmed M, Seraj R, Islam SMS. The K-means algorithm: A comprehensive survey and performance evaluation. Electronics, 2020, 9(8): 1295. [doi: 10.3390/electronics9081295]

7 黄发明, 曹中山, 姚池, 等. 基于决策树和有效降雨强度的滑坡危险性预警. 浙江大学学报 (工学版), 2021, 55(3): 472–482. [doi: 10.3785/j.issn.1008-973X.2021.03.007]
8 Nti IK, Quarcoo JA, Aning J, *et al*. A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. Big Data Mining and Analytics, 2022, 5(2): 81–97. [doi: 10.26599/BDMA.2021.9020028]
9 Blazek K, Van Zwieten A, Saglimbene V, *et al*. A practical guide to multiple imputation of missing data in nephrology. Kidney International, 2021, 99(1): 68–74. [doi: 10.1016/j.kint.2020.07.035]
10 Le TT. Practical hybrid machine learning approach for estimation of ultimate load of elliptical concrete-filled steel tubular columns under axial loading. Advances in Civil Engineering, 2020, 2020: 8832522.
11 Láruson ÁJ, Fitzpatrick MC, Keller SR, *et al*. Seeing the forest for the trees: Assessing genetic offset predictions from gradient forest. Evolutionary Applications, 2022, 15(3): 403–416. [doi: 10.1111/eva.13354]
12 Buo I, Sagris V, Jaagus J. Gap-filling satellite land surface temperature over heatwave periods with machine learning. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 7001105.
13 Ma J, Cheng JCP, Jiang FF, *et al*. A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. Energy and Buildings, 2020, 216: 109941. [doi: 10.1016/j.enbuild.2020.109941]
14 卢冰洁, 李炜卓, 那崇宁, 等. 机器学习模型在车险欺诈检测的研究进展. 计算机工程与应用, 2022, 58(5): 34–49. [doi: 10.3778/j.issn.1002-8331.2109-0312]
15 Leong WC, Kelani RO, Ahmad Z. Prediction of air pollution index (API) using support vector machine (SVM). Journal of Environmental Chemical Engineering, 2020, 8(3): 103208. [doi: 10.1016/j.jece.2019.103208]
16 Razavi-Far R, Cheng BY, Saif M, *et al*. Similarity-learning information-fusion schemes for missing data imputation. Knowledge-based Systems, 2020, 187: 104805. [doi: 10.1016/j.knosys.2019.06.013]
17 李美玉, 刘洋, 王艺璇, 等. 基于 Stacking 集成学习的用户付费转化意向预测方法研究——以免费增值游戏为例. 数据分析与知识发现. http://kns.cnki.net/kcms/detail/10.1478.g2.20230317.1235.004.html. (在线出版)(2023-03-20).
18 Asraf H M, Dalila K A N, Tahir NM, *et al*. Missing data imputation of MAGDAS-9's ground electromagnetism with supervised machine learning and conventional statistical analysis models. Alexandria Engineering Journal, 2022, 61(1): 937–947. [doi: 10.1016/j.aej.2021.04.096]
19 Awawdeh S, Faris H, Hiary H. EvoImputer: An evolutionary approach for missing data imputation and feature selection in the context of supervised learning. Knowledge-based Systems, 2022, 236: 107734. [doi: 10.1016/j.knosys.2021.107734]
20 Nikfalazar S, Yeh CH, Bedingfield S, *et al*. Missing data imputation using decision trees and fuzzy clustering with iterative learning. Knowledge and Information Systems, 2020, 62(6): 2419–2437. [doi: 10.1007/s10115-019-01427-1]
21 Rovetta A. Raiders of the lost correlation: A guide on using Pearson and Spearman coefficients to detect hidden correlations in medical sciences. Cureus, 2020, 12(11): e11794.

(校对责编: 孙君艳)