

最大化广义瑞利熵在列车空调制冷剂 充注量诊断中应用^{*}

张梦源¹ 王 钊² 鲍 超³ 陈焕新⁴ 张鉴心¹ 程亨达⁴

(1. 华中科技大学中欧清洁与可再生能源学院, 430074, 武汉; 2. 广州鼎汉轨道交通车辆装备有限公司, 510260, 广州;
3. 广州地铁集团有限公司, 510330, 广州; 4. 华中科技大学能源与动力学院, 430074, 武汉)

摘 要 [目的] 利用机器学习方法对列车空调制冷剂充注量的诊断中, 经常出现许多特征, 如何选择特征成为诊断中两难的抉择。特征选择过多, 算法资源开销巨大; 特征选择过少, 则不能很好地学习故障信息, 导致诊断模型的效果不好。目前, 在特征选择中, 有一个比较常用的方法, 即 PCA (主成分分析), 还有一个引用次数比较少的为最大化瑞利熵, 2 个皆存在特征过多导致的算法资源开销过大问题。为减少资源开销及比较二者优劣, 特开展本研究。[方法] 选用比较经典的支持向量机和均值聚类模型作比较, 同时利用小样本历史数据构造降维算法而不是利用全体样本构造最大化瑞利熵的降维算法, 同样是为了减少资源开销, 属于对算法的改进。比较基于小样本历史数据的最大化广义瑞利熵降维和常用的 PCA 降维方式在均值聚类和支持向量机模型的 F1 得分、准确率和时间开销上的优劣。[结果及结论] 结果表明, 最大化广义瑞利熵的支持向量机的故障诊断和探测时间资源开销只有原始数据训练支持向量机模型的 3%, 无论是均值聚类还是支持向量机在使用最大化瑞利熵的投影后的数据, 测试数据准确率比其他模型更接近 100%。

关键词 列车空调; 制冷剂充注量; 小样本; 广义瑞利熵

中图分类号 U260.39

DOI:10.16037/j.1007-869x.2024.10.045

Application of Maximum Generalized Rayleigh Entropy in Diagnosing Train Air-conditioning Refrigerant Charge

ZHANG Mengyuan¹, WANG Zhao², BAO Chao³, CHEN Huanxin^{4*}, ZHANG Jianxin¹, CHENG Henda⁴

(1. China-EU Institute for Clean and Renewable Energy at Huazhong University of Science & Technology, 430074, Wuhan, China; 2. Guangzhou Dinghan Railway Vehicles Equipment Co., Ltd., 510260, Guangzhou, China; 3. Guangzhou Metro Group Co., Ltd., 510330, Guangzhou, China; 4. School of Energy and Power Engineering, Huazhong University of Science and Technology, 430074, Wuhan, China)

^{*} 国家自然科学基金项目(51876070)

ty of Science and Technology, 430074, Wuhan, China)

Abstract [Objective] In diagnosing train air-conditioning refrigerant charge using machine learning methods, numerous features often appear, posing a dilemma in feature selection for diagnosis. Excessive feature selection leads to high algorithm resource costs, while insufficient feature selection results in poor learning of fault information, negatively affecting the diagnostic model performance. Currently, a commonly used method for feature selection is PCA (principal component analysis), and another less frequently cited method is MRE (Maximum Rayleigh Entropy). Both methods, however, face the problem of high resource costs due to excessive features. The research is specially carried out aiming to reduce resource costs and the effectiveness of these two methods is compared.

[Method] For this purpose, the typical SVM (support vector machine) and K-means clustering model are selected for comparison, and a small sample of historical data is used to construct dimensionality reduction algorithm, instead of using the entire sample to construct MRE dimensionality reduction algorithm, representing an improvement to the algorithm under the same goal of reducing resource costs. MRE dimensionality reduction based on small historical data sample and the commonly used PCA dimensionality reduction are compared in terms of their performance on F1 score, accuracy rate, and time cost in K-means clustering and SVM models. [Result & Conclusion] The results indicate that the fault diagnosis and detection time resource costs of SVM machine using MRE is only 3% that of SVM model trained on original data. Whether using K-means clustering or SVM, the test data accuracy rate after using the projected data from MRE is closer to 100% compared to other models.

Key words train air-conditioning; refrigerant charge amount; small sample; maximum Rayleigh Entropy

空调系统的制冷剂充注量对于整个系统的寿命期限影响很大,制冷剂过充导致冷凝器充满液体,进而造成制冷量不足和排气温度过高;充注量不足,冷凝器出口出现气液两相流且蒸发器没有完全充满,蒸发压力与蒸发温度降低,压缩机排气温度过高。其都会导致电机寿命降低和能耗居高不下。而列车等车辆体积有限不会携带储液器等制冷剂冗余设备^[1],空调系统内的充注量波动频繁,对列车空调系统的运行经济性与能耗影响巨大,因此,对列车空调的制冷剂充注量诊断与预警十分必要。

针对空调系统的充注量故障诊断,文献[2]提出超参优化的 RelifeF 特征选择方法,结合支持向量机模型用于制冷剂充注量的诊断,可以获得 90% 以上的诊断效果;文献[3]利用融合集成特征选择的方法,结合轻型梯度提升器可以快速精准地实现充注量的诊断;文献[4]比较支持向量机与后向传播算法支持的人工神经网络对于故障诊断的效果,在 PCA(主成分分析)的加持下,支持向量机获得优于后向传播算法支持的神经网络模型的效果;瑞利熵在天文、高能物理、信号分析以及数据分析等领域中应用广泛;文献[5]基于瑞利熵做了进一步的改进,使之在数据处理方面更加强大;文献[6]将瑞利熵最大化应用于计算较高信噪比的信号抗干扰抑制矩阵中,证明瑞利熵在信号分析中的应用;文献[7]将瑞利熵应用于快速估算样本协方差矩阵的最大特征值,将其与 Spike 模型中的动态阈值比较,然后判定电网分区是否异常。

空调故障诊断中鲜有将最大化瑞利熵与故障诊断结合,本文将其用于数据的预处理,最终送入诊断器中。利用机器学习方法对列车空调制冷剂充注量的诊断中,经常出现许多特征。在特征选择中,一个比较常用的方法是 PCA,还有一个是引用次数比较少的最大化瑞利熵,这 2 个方法皆存在特征过多导致的算法资源开销过大问题。为减少资源开销及比较二者优劣,选用比较经典的支持向量机和均值聚类模型做比较。

本文采用相同的试验机组产生的小样本历史数据获得投影矩阵,避免使用原始训练数据集降维的巨大资源开销。同时,对单一指标问题使用 3 个指标(训练时间消耗,召回率与精确率的加权得分 $f1$ 和准确率)来衡量模型的好坏。

1 试验方案与数据处理

1.1 瑞利熵与广义瑞利熵

关于 Hermitian 矩阵^[8] A 与向量 x 的多项式,形如公式(1)的称为瑞利熵^[9],将 x 施密特标准正交化^[10],则瑞利熵就会退化为 $R(x, A) = x^H A_1 x$,这是人们熟知的二次型问题,很容易得出 $\lambda_{\min} < R(x, A) = \frac{x^H A x}{x^H x} < \lambda_{\max}$,其中 λ 是矩阵 A_1 的特征值。

广义瑞利熵的形式如下:

$$R(x, A, B) = \frac{x^H A x}{x^H B x} \quad (1)$$

其中: A 、 B 均是 Hermitian 矩阵, B 是正定矩阵,令 $x = B^{-1/2} x_1$,广义瑞利熵退化为瑞利熵,处理起来变得更加简单,变为公式(2)所示,后面为表达方便,一律称广义瑞利熵为瑞利熵。

$$R(x, A, B) = R(x_1, A, B) = x_1^H (B^{-\frac{1}{2}})^H A B^{-\frac{1}{2}} x_1 \quad (2)$$

1.2 试验和数据处理流程

试验是在如图 1 列车空调系统原理的机组上进行的,试验的步骤是按照过充 110% 制冷剂,以及 100%、80% 和 60% 的充注量依次运行试验机组。首先,超量充注 110% 运行,运行一段时间后,获得超量充注运行的原始数据;然后泄漏掉一部分获得充注量为 100% 的充注量,依次泄漏制冷剂为 80%、60% 的充注量运行,获得运行的原始数据。各原始数据的特征一致,原始数据集各特征的物理含义如图 1 所示。标签为 110%、100%、80% 和 60% 充注量的有效数据量分别为:17 129 个,22 169 个,20 207 个和 22 930 个,共计 82 435 条样本。获得不同充注量的数据后,把数据打上标签,将样本数据打乱形成原始数据集。数据处理与研究流程如图 2 所示,将所有数据视为总体,按照 6:3:1 拆分为训练集 1 与验证集 1 以及小样本数据集;利用最大化小样本数据集的瑞利熵获取投影矩阵 M_i (i 是降维数据的维度,降维的最大维度不超过原始数据集的所有类别标签数减去 1),选取不同的降维数据维度,获取投影矩阵 M_i 。使用投影矩阵 M_i 对测试集 1 与训练集 1 矩阵变换,获得测试集 2 与训练集 2,变换后的数据集获得均值聚类模型 2 和支持向量机模型 2,原始数据的模型记为均值聚类模型 1 和支持向量机模型 1。

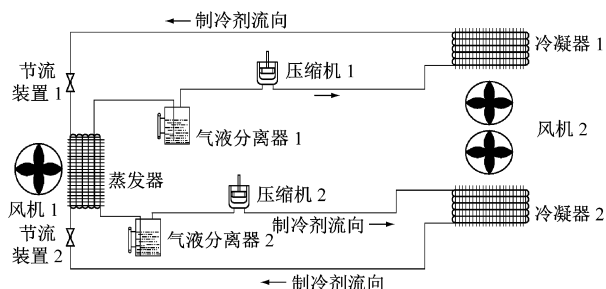


图 1 列车空调系统原理图

Fig. 1 Diagram of train air-conditioning system

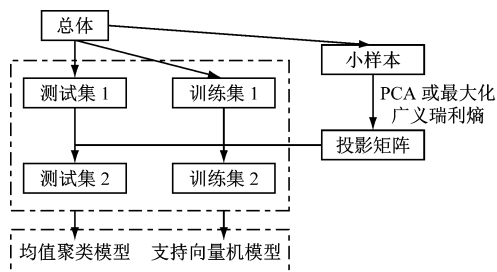


图 2 数据处理步骤图

Fig. 2 Diagram of data processing steps

同时也使用 PCA 分别获得相应的降维测试集与训练集,以及训练均值聚类模型和支持向量机模型。

2 故障诊断和探测模型的训练与选择

为了进一步了解均值聚类和支持向量机两种模型使用不同数据集的表现和算法开销(本文以时间为准),分别使用原始数据集和降维数据集训练均值聚类和支持向量机。

2.1 原始数据下训练的故障诊断和探测模型

2.1.1 均值聚类训练结果

由于均值聚类模型的训练不稳定,可使用多轮迭代训练以避免单次聚类结果不稳定。图 3 所示为均值聚类模型的训练流程图,将聚类中心改为随机选取,重复多轮训练以获得最佳表现模型。如果训练次数 i 大于 200 轮,直接跳出循环获得最优 f_1 得分的模型。由多轮训练出的 f_1 得分和准确率变化的结果可知,多轮训练下的准确率上限在 0.35 左右,因此将 0.35 作为该模型在当前数据集下的训练极限准确率。为了进一步验证是否存在最优的上限,将循环的条件改为 f_1 得分大于 0.35,循环次数大于 500 轮跳出循环,获得结果其准确率始终在 0.35 以下。

综上,均值聚类模型在原始数据下的训练效果仍然不佳,实际的应用意义不大。

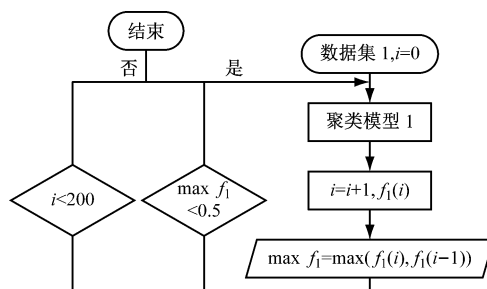


图 3 均值聚类诊断模型训练流程图

Fig. 3 Training procedure of K-means clustering diagnostic model

2.1.2 支持向量机训练结果

在 39 个原始数据特征全部参与的条件,由于支持向量的计算复杂度随着维度的增加呈现爆炸式增加,花费数倍均值聚类所用的时间才能训练出期望的支持向量机(如表 1 所示)。

表 1 原始数据、降维数据训练的支持向量机表现

Tab. 1 SVM performance trained on original data and dimensionality reduction data

制冷剂充注量/%	精确率		召回率	
	原始数据	降维数据	原始数据	降维数据
110	0.85	1.00	1.00	0.99
100	1.00	0.99	0.81	1.00
80	0.94	1.00	0.99	1.00
60	0.99	1.00	1.00	1.00

注:原始数据的准确率为 0.95;降维数据的准确率为 1.00。

从表 1 可知,不仅充注量的分类召回率在 80% 以上,而且各标签精确率都在 85% 以上,支持向量机模型整体的充注量诊断准确率高达 95%。

2.2 降维数据的故障诊断和探测

使用最大化瑞利熵和 PCA 获得降维数据,均降至三维,然后比较两个模型的表现。

2.2.1 基于瑞利熵投影数据的故障诊断和探测

表 1 所示为支持向量机降维数据训练模型验证报告,同时均值聚类在使用降维数据后,其能力与支持向量机接近。可以看到几乎所有标签的召回率、精确率和准确率都接近 100%,证明无论支持向量机还是均值聚类,最大化瑞利熵的训练集 2 构建模型的能力要优于一般模型。

2.2.2 基于 PCA 的故障诊断和探测

PCA 也是常用的数据降维方法,均值聚类模型在该方法降维数据中的表现如图 4 所示。由图 4 可知, f_1 得分有峰值,此时主成分个数为 11,但 f_1 得

分小于 0.5,故实际的应用意义不大。根据图 5 所示,支持向量机的 f1 得分会随着选择主成分个数的增加而增加。但为了最快的训练和获得最佳 f1 得分,没有必要增加很多主成分以消耗计算资源,因此选择主成分个数为 13。

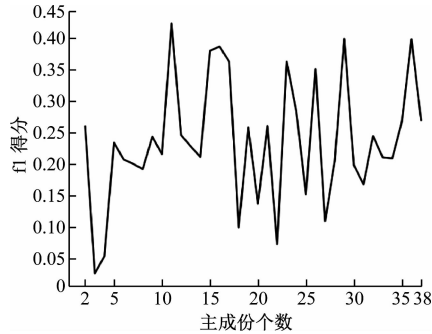


图 4 不同数量主成分均值聚类的 f1 得分

Fig. 4 K-means clustering f1 score with different numbers of PCA components resource costs

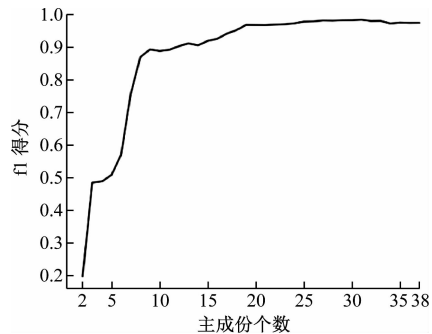


图 5 不同数量主成分支持向量机的 f1 得分

Fig. 5 SVM f1 score with different numbers of PCA components

2.3 最大化瑞利熵与 PCA 降维表现对比

本文的表现对比分别是数据降维方式对比以及不同模型资源开销对比。

1) 数据降维方式对比:根据图 6 和图 7 所示的情况,考虑算法时间资源开销的情况下,PCA 降维的数据处理方式训练的模型更节省资源。两种模型的时间开销都小于 1 s,但是 PCA 降维模型的准确率和 f1 得分均不如最大化瑞利熵的结果(见图 8)。图 9 显示在经过最大化瑞利熵之后,不同类别的样本之间已有了明显的区别度,而 PCA 处理后的样本之间区别度不是很大。

2) 不同模型之间的资源开销对比(如图 8 所示),不经过处理的原始数据训练的支持向量机时间资源开销巨大,不能为控制系统提供及时的决策输入,但支持向量机在原始数据下模型表现优于均值聚类。

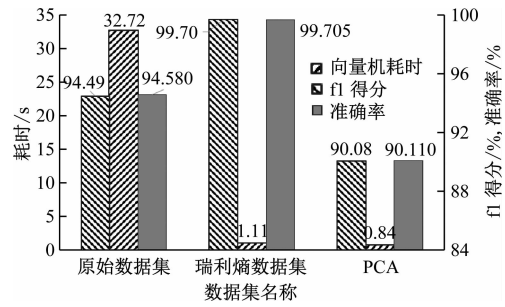


图 6 支持向量机的不同数据降维方式下的表现

Fig. 6 SVM performance under different data dimensionality reduction approaches

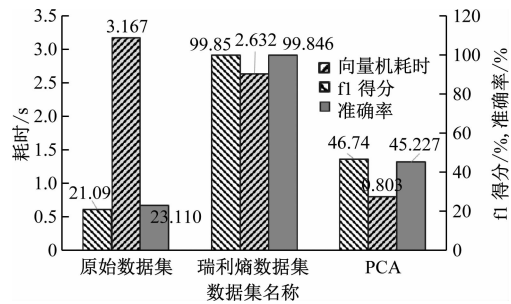


图 7 均值聚类的不同数据降维方式下的表现

Fig. 7 K-means clustering FDD performance under different data dimensionality reduction approaches

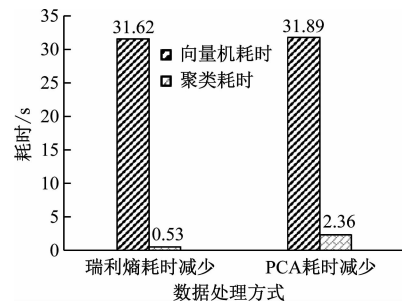


图 8 不同降维方式下的资源开销

Fig. 8 Resource costs under different dimensionality reduction approaches

3) 不同模型资源开销对比:本文以算法训练到诊断结果为止的时间消耗为准进行分析,从图 8 中可明显看出,支持向量机使用小样本最大化瑞利熵降维模型资源开销和均值聚类开销均有一定的缩减,意味着资源开销进一步的缩小,但最大化瑞利熵+支持向量机的准确性和 f1 得分较高。

综上所述,该机组的诊断模型为最大化瑞利熵降维+支持向量机的组合,既能满足快速运算的需要,又能满足诊断所需要的精确和准确程度。

3 结语

本文对利用最大化瑞利熵和 PCA 数据降维,获

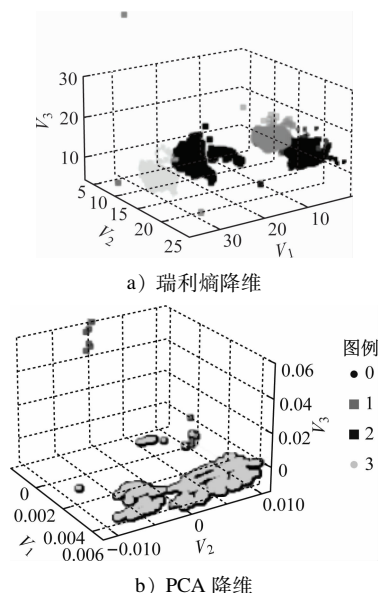


图 9 不同处理方式下的数据散点图

Fig. 9 Data scatter plots under different approaches

得不同的支持向量机与均值聚类模型,最大化瑞利熵结合支持向量机训练时间上相对于原始数据减小 97%,其诊断准确率为 98% 以上。主要结论如下:

1) 均值聚类模型运算速度快,但是结果不稳定,使用最大化瑞利熵数据降维和 PCA 降维,可缓解解决均值聚类的不可靠问题,但仍然没有彻底解决均值聚类本身采用迭代方式获取局部最优解的固有缺陷。

2) 支持向量机具有高可靠性、高运算成本,基于小样本的最大化瑞利熵降维,训练模型所需要的运算量小,可在数秒内完成,故完全可以使用低端小成本计算单元。

本文验证了最大化瑞利熵在数据降维中可靠性和优势,不足之处是未在实车上运行。未来的研究方向是使用实车运行数据验证模型的可靠性,以提高模型的泛化性能等。

参考文献

- [1] 秋雨豪,乐清洁. 制冷剂充注量对冷藏车用制冷机组性能的影响[J]. 制冷与空调, 2020, 20(4): 51.
QIU Yuhao, LE Qingjie. Impact of refrigerant charging amount on performance of refrigeration unit for refrigerated van[J]. Refrigeration and Air-Conditioning, 2020, 20(4): 51.
- [2] 徐畅,李绍青,李正飞,等. 超参优化的 ReliefF-SVM 在制冷剂充注量故障诊断的应用[J]. 制冷技术, 2021, 41(1): 17.
XU Chang, LI Shaoqing, LI Zhengfei, et al. Application of ReliefF-SVM with hyperparameter optimization in fault diagnosis of

refrigerant charge[J]. Chinese Journal of Refrigeration Technology, 2021, 41(1): 17.

- [3] 胡宽,李正飞,陈焕新. 集成特征选择在 VRF 系统制冷剂充注量故障诊断中的应用[J]. 制冷与空调, 2020, 20(9): 90.
HU Kuan, LI Zhengfei, CHEN Huanxin. Forward search ensemble feature selection strategy for multi-line system fault diagnosis[J]. Refrigeration and Air-Conditioning, 2020, 20(9): 90.
- [4] 韩华,谷波,任能. 基于主元分析与支持向量机的制冷系统故障诊断方法[J]. 上海交通大学学报, 2011, 45(9): 1355.
HAN Hua, GU Bo, REN Neng. Fault diagnosis for refrigeration systems based on principal component analysis and support vector machine[J]. Journal of Shanghai Jiaotong University, 2011, 45(9): 1355.
- [5] FAN J, KE Z T, LIU H, et al. Quadro: a supervised dimension reduction method via Rayleigh quotient optimization[J]. Annals of Statistics, 2015, 43(4): 1498.
- [6] 陈艳,宋云超,陆洋. 最大化瑞利熵的干扰对齐优化算法[J]. 电讯技术, 2019, 59(3): 260.
CHEN Yan, SONG Yunchao, LU Yang. An optimized interference alignment algorithm based on maximizing Rayleigh entropy[J]. Telecommunication Engineering, 2019, 59(3): 260.
- [7] 李洪乾,韩松,周忠强. 利用 Rayleigh 熵和并行计算的大规模电网异常负荷快速识别[J]. 电力系统保护与控制, 2019, 47(23): 37.
LI Hongqian, HAN Song, ZHOU Zhongqiang. Efficient abnormal load identification in large-scale power system employing Rayleigh quotient and parallel computing technology[J]. Power System Protection and Control, 2019, 47(23): 37.
- [8] 李学良,王建峰,黄琼湘. 具有不同特征值的 Hermitian 矩阵及其应用[J]. 数学学报(中文版), 2015, 58(5): 731.
LI Xueliang, WANG Jianfeng, HUANG Qiongxiang. On the Hermitian matrix with distinct eigenvalues and its applications[J]. Acta Mathematica Sinica (Chinese Series), 2015, 58(5): 731.
- [9] 刘长安. Rayleigh 商极值定理的推广及其应用[J]. 西安工业大学学报, 1997, 17(4): 79.
LIU Chang'an. Generalization and application of Rayleigh quotient extremum theorem[J]. Journal of Xi'an Technological University, 1997, 17(4): 79.
- [10] 丁照银. 两种标准正交化方法的比较[J]. 科学技术创新, 2021(27): 62.
DING Zhaoyin. Comparison of two standard orthogonalization methods[J]. Scientific and Technological Innovation, 2021(27): 62.

• 收稿日期:2022-07-14 修回日期:2022-12-14 出版日期:2024-10-10
Received:2022-07-14 Revised:2022-12-14 Published:2024-10-10
• 通信作者:张梦源,硕士研究生,155965462@qq.com
• ©《城市轨道交通研究》杂志社,开放获取 CC BY-NC-ND 协议
© Urban Mass Transit Magazine Press. This is an open access article under the CC BY-NC-ND license