Hans Journal of Data Mining 数据挖掘, 2020, 10(2), 129-135
Published Online April 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.102013

Hans汉斯

Analysis of Pump Data Based on Association and Kmeans Algorithm

Ziyi Liu[1], Zhe Zhang[2], Tian Gao[2], Qiang Wu[3]

[1]College of Computer Science, Nankai University, Tianjin
[2]Nankai University Binhai College, Tianjin
[3]Tianjin Pump Industry Machinery Group Co. Ltd., TianJin
Email: 2120180525@mail.nankai.edu.cn, 289208171@qq.com, gaotian1002@163.com

Received: Mar. 18th, 2020; accepted: Apr. 9th, 2020; published: Apr. 16th, 2020

Abstract

Aiming at the performance test and unstable operation of the pump before leaving the factory, this paper uses multiple linear regression and Apriori algorithm to analyze the correlation between the characteristics of the pump, and removes the redundant attributes, and then uses the kmeans algorithm to carry out cluster analysis on the basis of removing the redundant attributes, to find out the relationship between the effluent pressure and the flow rate, and the relationship between the voltage of the two lines as well as the relationship between three-phase current and effluent pressure.

Keywords

Multiple Regression, Kmeans, Clustering, Correlation Analysis

基于关联及Kmeans算法对泵数据的分析

刘子熠[1]，张　喆[2]，高　天[2]，武　强[3]

[1]南开大学计算机学院，天津
[2]南开大学滨海学院，天津
[3]天津市泵业机械集团有限公司，天津
Email: 2120180525@mail.nankai.edu.cn, 289208171@qq.com, gaotian1002@163.com

收稿日期：2020年3月18日；录用日期：2020年4月9日；发布日期：2020年4月16日

摘　要

针对泵企业在泵出厂前的性能检测以及运行不稳定等问题，本文应用多元线性回归、Apriori算法分析有

文章引用: 刘子熠, 张喆, 高天, 武强. 基于关联及 Kmeans 算法对泵数据的分析[J]. 数据挖掘, 2020, 10(2): 129-135.
DOI: 10.12677/hjdm.2020.102013

关泵的特征之间的相关性，去除冗余属性，然后在去除了冗余属性的特征基础上应用Kmeans算法进行聚类分析，找出了出水压力与流量间的关系，两根线的电压间的关系，三相电流和出水压力间的关系等结论。

关键词

多元回归，Kmeans，聚类，关联分析

1. 引言

近年来，国内外的许多研究机构及专家学者对泵的特性进行了许多研究：轴向柱塞泵是液压传动中使用做广泛的液压元件之一，由于其工作原理和结构特点等，柱塞泵内部各因素之间存在着一些关系，由于管路及负载阻抗的影响，流量脉动在传动过程中又必然会引起压力脉动，进而对其性能产生影响[1]。还有一些学者对异常进行了分析，对输注泵进行质量检测，提高输注泵的使用效率，延长其使用寿命[2]。也有学者对其他产品连续时间产生的数据进行分析，张洋洋对油田异常的大量检测数据进行深入挖掘分析，定位油田异常的隐性故障[3]。刘凯针对采油厂潜油电泵生产现状，为提高潜油电泵井系统效率，节能降耗，分析影响系统效率的主要因素，提出了相应的治理方案[4]。

然而目前尚缺乏针对泵的各项参数一起分析的研究。如何正确合理地判断泵的运行状态则是现阶段需要解决的问题，本文利用多元回归分析，基于 Apriori 算法的关联规则挖掘，基于 Kmeans 的聚类算法对泵运行的数据进行了分析与挖掘。

2. 数据选取

本文分析的数据源为某泵企业在 2019 年 9 月的某天收集的正常运行的泵的共计 95 条数据。数据源包含如下字段：

1) 日期，泵号；

2) 和泵产品综合质量指标相关的机油温度，噪声；

3) 该泵的进水压力，出水压力，机油压力，流量，总有功功率，正向有功电能；

4) 体现电机质量与稳定性的参数：A 相电流，B 相电流，C 相电流，AB 线电压，AC 线电压，BC 线电压。

这些数据的含义如表 1 所示。

3. 基于多元回归与 Apriori 算法的影响分析

3.1. 基于多元回归的各特征对机油温度的影响分析

有关信息显示，机油温度是产品综合质量指标考核的要点，其会受到多个因素的影响，为了观察是否存在某几种因素对最终的机油温度结果影响较大，本文给出基于多元回归分析的各因素影响因子分析。

以之前处理过得到出机油温度以外的各时间点的其他因素作为自变量，各时间点机油温度为因变量，建立如公式(1)所示的多元线性回归模型进行多元线性回归分析。

Table 1. Corresponding table of the meaning of valid data items
表 1. 有效数据项含义对应表

编号	因素名称	具体含义
a	机油压力	产品性能参数
b	A 相电流	通过 A 相的电流
c	B 相电流	通过 B 相的电流
d	C 相电流	通过 C 相的电流
e	AB 线电压	体现电机质量
f	AC 线电压	体现电机质量
g	BC 线电压	体现电机质量
h	总有功功率	产品性能参数，做工的功率
i	出水压力	产品性能参数，出水处检测的压力
j	噪声	产品综合质量指标，泵运行过程中发出的噪音
k	正向有功电能	显示用电量
l	流量	产品性能参数
m	机油温度	产品综合质量指标

$$y = \sum_{i=1}^{13} \theta_i x_i + b \tag{1}$$

y 表示机油温度，$x_i\ (i = 1 \sim 13)$ 表示特征在对应某时间点中的参数。

使用最小二乘法来确定模型中的各特征对应系数 θ_i 和回归方程的截距 b，即使得SSE最小。为了实现多元回归，本文使用了 Python-Sklearn 中的 Linear Regression 函数进行拟合：使用包中的 Train_test_split 函数，将数据集分为训练集和测试集，测试集占总数据量的 20% (Train_size = 0.80)，即随机选择 76 个时间点的数据进行训练，使用 19 个时间点的数据进行测试。

根据多元回归的拟合系数及拟合效果，初步判断泵的机油温度与其他元素成非线性关系，机油温度受到任意元素同种程度的影响。在测试后，机油温度多元回归模型很好。同时也测试了其他和综合性能有关的数据例如噪音的多元回归，如图 1 所示，其中横坐标为各因素产生的影响，纵坐标为噪声：

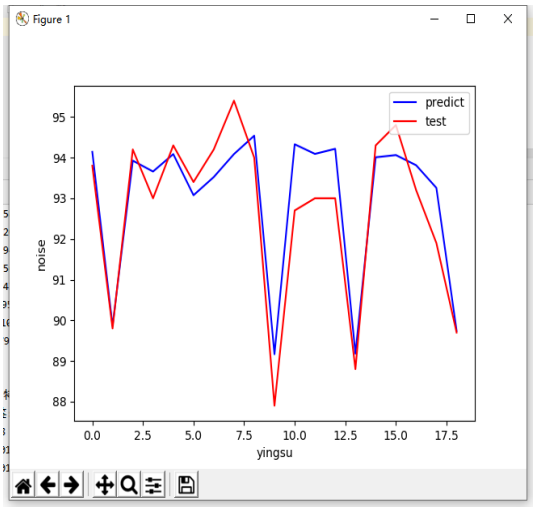

Figure 1. Noise multiple regression fitting
图 1. 噪声多元回归拟合

测试后发现，表现最好的仍旧是机油温度，拟合度平均达到了 99%。如图 2 所示，其中横坐标为各因素产生的影响，纵坐标为机油温度：

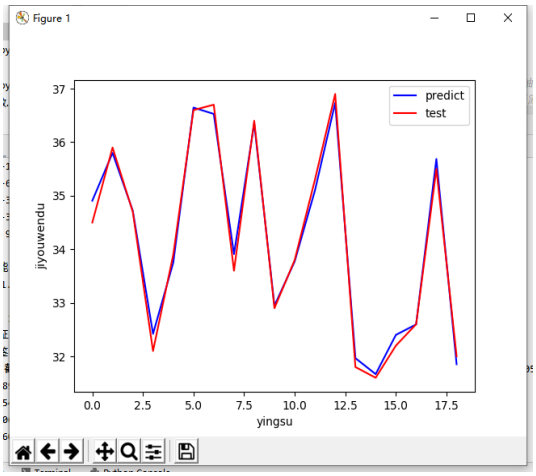

Figure 2. Multiple regression fitting of oil temperature
图 2. 机油温度多元回归拟合

由此可知，机油温度的结果确实会被其他参数影响，机油温度受到任意元素同种程度的影响。

3.2. 基于 Apriori 算法对各参数之间关联的分析

泵产生的各生产参数看似独立却又相辅相成，为了找出其中的关系，则对其进行相关性分析。对参数相关性的分析的目的是找到各参数间的强关联关系，获取潜在的信息，更好地指导生产。本文使用 Apriori 算法对各元素进行了关联规则分析。

Apriori 算法是一种最有影响的挖掘布尔关联规则频繁项集的算法[5]。该算法的基本思想是：首先利用递归的方法找出所有的频集[6]，然后由频集产生强关联规则，将大于给定最小可信度的规则保留下来。在本数据集中，项集(流量 0)出现了 20 次，部分批次数据组中共 95 条记录，那么该项集的支持度就是 20/95。

3.2.1. 基于一维聚类的数据离散化

应用 Apriori 算法要求使用离散的数据格式，而本文数据是连续数值型数据，故先将各批次瑕疵数据进行离散化。

离散化就是在数据的取值范围内设定若干个离散的划分点，将取值范围划分为一些离散化的区间，最后用不同的符号代表落在每个区间中的数据值。所以离散化涉及两个过程：确定分类数，将连续属性值映射到 N 个分类值。

常用的离散化方法有：等宽离散、等频离散、一维聚类离散。本文所分析的数据采用基于 Kmeans 的一维聚类离散化方法，按照瑕疵属性值之间的差值大小，将连续属性值划分为 K 个区间，让簇内的点差值小，而让簇间点差值大。

对数据进行离散化处理后，除流量被离散为从高到低的三类，其余特征都被离散为了四类，定性化离散后的数据如下：

$C_{i,j}=0$ (第 j 个时间点的 i 因素离散化数值小)；

1 (第 j 个时间点的 i 因素离散化数值较小)；

2 (第 j 个时间点的 i 因素离散化数值较大)；

3 (第 j 个时间点的 i 因素离散化数值大)。

3.2.2. 关联规则结果分析

使用 Apriori 算法进行关联规则分析。最小支持度 0.3 最小支持度 0.9，除流量 K 设置为了 3 类，其余 K 均在测试后，先设置为了被离散为从高到低的 4 类。得到关联关系 9 条，单属性对的关联关系共 6 条，多属性组之间的关联关系共 3 条。其中单属性对的关联关系如下：

属性	关联属性	可信度
{'b3'}	{'h3'}	1.0
{'d3'}	{'h3'}	1.0
{'c3'}	{'h3'}	1.0
{'f1'}	{'g1'}	0.9666666666666668
{'g1'}	{'f1'}	0.9666666666666668
{'i0'}	{'l1'}	1.0

其中多属性对的三条关联关系如下：

属性对	关联属性对	可信度
({'c3', 'l1'})	{'h3'}	1.0
({'d3', 'l1'})	{'h3'}	1.0
({'c3', 'd3'})	{'h3'}	1.0

观察得知多属性组之间的关联关系都是依托于单属性对之间的关联关系的，接着对上述单属性对关联规则进行分析，得出一下规则：

规则一：b3 指代 A 相电流离散化数值大，c3 指代 B 相电流离散化数值大，d3 指 C 相电流离散化数值大，h3 指代总有功功率离散化数值大。三相电流稳定在一个位置时，根据物理知识，总有功功率也应该维持在一个水平，这个关系与我们的常识相符。在于企业人员咨询后，得知该种数据体现出电机质量较好。

规则二：f1 指代 AC 线电压离散化数值较小，g1 指代 BC 线电压离散化数值较小。二者之间本应无关系，都是人为设置在某一点，在于企业人员咨询后，得知这里可能是操作过程中产生了一定影响。

规则三：i0 为出水压力离散化数值小，l1 指代流量离散化数值较小。出水压力处在一个低值时，流量必定也处在一个低值。

4. 基于 Kmeans 的数据聚类

4.1. Kmeans 聚类及特征选择

为了找出连续的时间点泵的运行特点，本文对其数据进行 Kmeans 聚类划分。

聚类是指将数据集划分为若干簇，使得同一个簇中的数据最为相似，而不同簇之间的数据相似度差别尽可能大。Kmeans 算法作为一种基于划分的经典聚类算法，主要采用迭代的方式对数据进行处理分析，主要目的是将 N 个样本对象划分到 K 个类中，使每个类中的样本对象间的属性特征最相近。同时由前文所述多元回归分析结果可知：各参数对机油温度的影响程度相同，度量标准相同，且数据量少、维数较低，不需要弱化离群点的影响。故采用欧几里得度量方式进行距离计算。

根据前文关联规则分析和相关系数的分析结果，b (A 相电流)，c (B 相电流)，d (C 相电流)和 h (总有

功功率)，以及实际情况，可将三者合并为 1 个特征 h。出水压力和流量有关，但是流量基本是恒定值，所以二者特征通用出水压力来代替。

本文最终选择 9 个特征为聚类分析的数据，分别是：机油压力量，AB 线电压量，AC 线电压量，BC 线电压量，总有功功率量，出水压力，噪声量，正向有功电能和机油温度。

为了找出合适的聚类数，接着通过肘部法则(一种计算不同数目的类别对整体数据类别畸变程度的方法)绘制 K 值与平均畸变程度折线图(横轴为 K 值，纵轴为对应畸变程度)，在 K = 2 时平均畸变程度最大，且再继续增大 K 值得到的平均畸变程度变化不大，但因为 K = 2 是划分的类比较小，比较不利于找出问题，因此聚类数目 K 选为平均畸变程度变化稍大的 4。如图 3 所示：

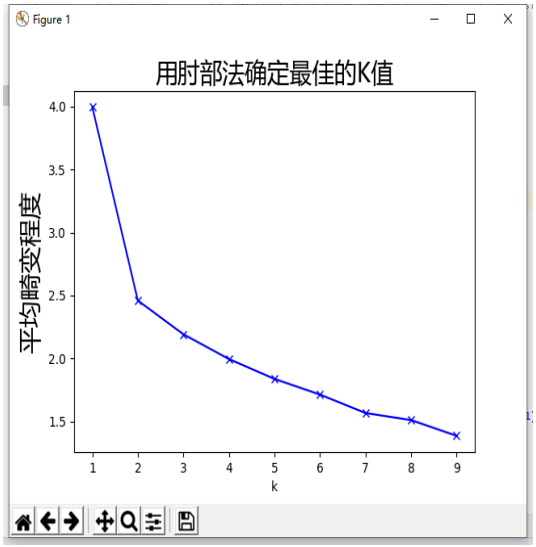

Figure 3. Best K value determined by elbow method
图 3. 用肘部法确定的最佳 K 值 https://fanyi.baidu.com/ - zh/en/javascript:void(0);

4.2. 聚类结果及分析

对数据进行处理后，类型为 0 的数据共 5 条数据，约占总数据 5%，类型为 1 的数据共 4 条数据，约占总数据 4%，类型为 2 的数据共 25 条数据，约占总数据 26%，类型为 3 的数据共 61 条数据，约占总数据 64%。聚类质心如下：

序号	出水压力	机油压力	机油温度	噪声	AB 线电压	AC 线电压	BC 线电压	总有功功率	正向有功电能
1	19.71212121	2.97272727	32.17575758	94.06969697	395.33939394	395.35454545	395.0969697	26.24545455	19.07878788
2	2.12	2.62	35.86	89.512	397.608	397.932	397.916	5.612	25.104
3	2.05	2.575	35.775	90.975	394. 3	398	396.	27.	26
4	19.34848485	2.66969697	34.44848485	93.66969697	395.29393939	395.28787879	395.37575758	26.57878788	23.23333333

对归类后的数据进行分析，得到以下结论：

1) 对于 4 种分类，有着不同的特征：0 类：出水压力、总有功功率、正向有功电能较大；1 类：出水压力小、总有功功率小，正向有功电能较大；2 类：出水压力、总有功功率很小，正向有功电能较大；3 类：出水压力、总有功功率大，正向有功电能较大。同时，机器运行状态总是维持在 2 和 3 类。0 类 1 类很少，是异类。

2) 机油温度随着时间的推移而上升，噪声，电压稳定在区间内。

3) 第 46 条数据处出现电压突变，导致出水压力陡降至 14.7，其他参数也变小，最后将其归位了 0 类。此处被测试的泵可能出现了一定的问题，例如密封泄漏，阀组损坏造成压力不稳或下降，建议企业对此情况进行排查。

4) ABC 三相电流的参数对该机器各项数据有较大的影响。ABC 三相电流维持在高值时，总有功功率也维持在一个高值，依据企业给的信息，可推断此时运行的泵性能良好。

总之，通过对参数的分析以及与企业相关人员进行交流，证明和泵产品综合质量指标相关的机油温度，噪声都稳定，即说明被测试的泵综合性能较好。同时得知由于出水压力经常处于不稳定的状态，这也一定程度上造成了运行数据的分化：0 类可以被概括为功率异常低类；1 类可以被概括为出水压力异常低类；2 类可以被概括为出水压力&功率异常低类；3 类可被概括为正常运行类，不需要进行特别的处理。在此建议企业加强对泵的出水压力的控制，以助于有效减少异常数据的出现。

5. 结束语

本文将多种数据分析方法如多元回归、聚类、Apriori 关联规则挖掘应用到了泵运行参数分析中，通过对多个时间点泵运行数据的挖掘分析，发现了被分为 0 类和 1 类的异常数据。在这里对企业提出了一些建议：1) 出水压力小，流量也会比较小，这属于异常状态。二者同为性能指标参数，对此应该对泵进行进一步检测，检测流量和出水压力的关系，以保证产品性能良好。2) AC 线电压和 BC 线电压在数值较小时会有一定的联系，可检查是否有其他元素干扰。3) 在进水压力一定，出水压力变化可能是由于 ABC 三相电流维持在一个低水平导致的，维持在 62 以上稳定高值时出水压力正常化，此时功率维持在 26 左右。可以控制 ABC 三相电流，从而使出水压力稳定。4) 加强对泵的出水压力的控制，检查密封泄漏，阀组损坏等问题以减少异常数据的出现。

本文也有一定的局限性，首先是数据量有限，更多的数据可得到统计上更显著的结论，其次，泵具体出问题的位置，如何改善等问题，还有待企业和相关研究人员进一步的研究。

基金项目

天津市智能制造专项资金项目(201810602, 201907206, 201907210, 20191009)；天津市互联网先进制造专项资金项目 18ZXRHGX00110。

参考文献

[1] 苌飞霸, 潘克新, 刘伟, 等. 输注泵质量检测分析[J]. 医疗卫生装备, 2018, 39(11): 55-57.

[2] 张洋洋. 基于聚类算法和关联规则的油田异常井分析[D]: [硕士学位论文]. 兰州: 兰州理工大学, 2019.

[3] 刘凯. 优化机泵参数, 提高电泵井系统效率[J]. 内蒙古石油化工, 2012(1): 65-78.

[4] 韩家炜, Micheline Kamber. 数据挖掘概念与技术[M]. San Francisco: Margan Kaufmann, 2006: 157-164.

[5] 张良均, 王路, 谭立云, 等. Python 数据分析与挖掘实战[M]. 北京: 机械工业出版社, 2016: 113-114.

[6] 孟璇. 印品质量检测技术的发展[J]. 印刷质量与标准化, 2006(3): 47-49.