基于多示例多标记的抽油机故障诊断①

陈 妍[1], 许少华[1,2]

[1](东北石油大学 计算机与信息技术学院, 大庆 163318)
[2](山东科技大学 信息科学与工程学院, 青岛 266000)

摘 要: 针对抽油机工况数据可从位移、载荷、电流等多个方面进行描述, 若仅仅使用一个特征向量来描述抽油机工况数据会使其信息过于简化, 丢失一部分有效信息的问题, 以及工况数据具有多义性的特征, 提出基于多示例多标记的抽油机故障诊断. 该学习方法中, 用抽油机的位移、载荷、电流数据作为抽油机工况样本包的多个示例, 使用 k-medoids 聚类算法对样本包进行聚类, 将多个样本包转换为若干示例, 新示例的每一维表示样本包到样本各聚类中心的距离, 再利用 MLSVM 算法对转换后的多标记问题进行求解. 实验结果表明, 多示例多标记学习能够及时、准确地诊断出抽油机故障问题.

关键词: 多示例多标记; 抽油机; 故障诊断

Pumping Unit Diagnose Based on Muli-Instance and Multi-Label

CHEN Yan[1], XU Shao-Hua[1,2]

[1](School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)
[2](The College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao 266000, China)

Abstract: The operating condition data of pumping unit can be described from the aspects of displacement, load and electric current. If only one feature vector is used to describe the operating condition of the pumping unit, the information will be too simplified, and it will lost some effective information. In view of the above problems and polysemy which is the essential characteristics of operating condition data, the fault diagnosis of pumping unit based on multi-instance and multi-label is presented. In this study, the displacement, load and current data of the pumping unit are used as multiple instances of pumping unit working condition data bags. Using k-medoids clustering algorithm cluster the bags and convert bags into several instances. Each dimension of the new instance indicates the distance from the bags to each cluster center, and then the MLSVM algorithm is used to solve the multi label problem. Experimental results show that multi-instance and multi-label learning can diagnose the trouble of oil pumping machine timely and accurately.

Key words: muli-instance multi-label; pumping unit; fault diagnosis

1 引言

在现代采油工艺技术中, 油井停止自喷后通常使用机械采油方式进行采油, 螺杆泵便是众多机械采油方式中应用最为普遍的一种[1]. 而油井所处地理环境恶劣、井下条件复杂、不明因素众多, 这些不利因素导致无法实时判断抽油机运行状态, 从而不能及时进行故障诊断和处理, 严重影响了抽油机的抽油效率、增加了采油成本. 因此及时准确地了解有杆抽油系统的工作情况并进行故障诊断, 对提高油田生产效率和紧急效益具有中要的意义[2].

示功图是分析抽油机工作状况的重要依据, 技术人员主要通过观察示功图的上载荷、下载和、加载带、卸载带来判断示功图是否处于正常工作状态, 这样的分析更多的依赖于技术人员的工作经验和技术水平, 人为影响因素较大. 也有研究人员利用机器学习进行抽油机故障诊断操作.

① 收稿时间:2015-11-18;收到修改稿时间:2016-01-04 [doi: 10.15888/j.cnki.csa.005255]

文献[2]采用了半监督竞争过程元网络, 将离散 Fréchet 距离与欧氏距离相结合利用了示功图的时间细节特征对其进行分类识别; 文献[3]将示功图识别看作动态系统连续曲线(位移-时间曲线和载荷-时间曲线)的模式识别问题, 将一个周期内的位移-时间曲线和载荷-时间曲线直接作为模型输入; 文献[4]采用一种矩特征和傅里叶描述子相结合的方式进行的示功图故障诊断; 文献[5]通过两个分类支持向量机的组合来实现支持向量机的多分类算法, 应用支持向量机的多分类算法来实现示功图诊断操作; 文献[6]利用最小二乘法对示功图进行自动分类识别.

上述识别方法都取得了不错的识别效果, 但都是针对单示例或是单标记的学习. 而抽油机工况数据是具有多义性的, 只使用一个特征向量来进行描述会丢失很多有用信息, 因此本文提出利用多示例多标记方法对抽油机故障进行诊断.

2　各模块的算法设计与实现

多示例多标记学习主要用于对多义性对象进行学习, 需要给予对象适合的类别标记, 这里的类别标记不再是单一的类别标记了, 而是一个类别标记子集. 同样, 对多义性对象的描述也不再是采用单一示例进行表达, 而是使用示例集合表示[7,8].

多示例多标记学习主要是通过已知的数据集 $\{(X_1,Y_1),(X_2,Y_2),\Lambda,(X_m,Y_m)\}$ 来学习映射 $f:2^x\to 2^y$, 其中, $X_i\subset X$ 是一个含有 n_i 个示例的包 $\{x_{i1},x_{i2},\Lambda,x_{in_i}\}$, $x_{ij}\in X(j=1,2,\Lambda,n_i)$, $Y_i\subset Y$ 是包 X_i 对应的拥有 l_i 个类别的类别集合 $\{y_{i1},y_{i2},\Lambda,y_{il_i}\}$, $y_{ik}\in Y(k=1,2,\Lambda,l_i)$. 图 1 所示即为多示例多标记学习框架示意图.

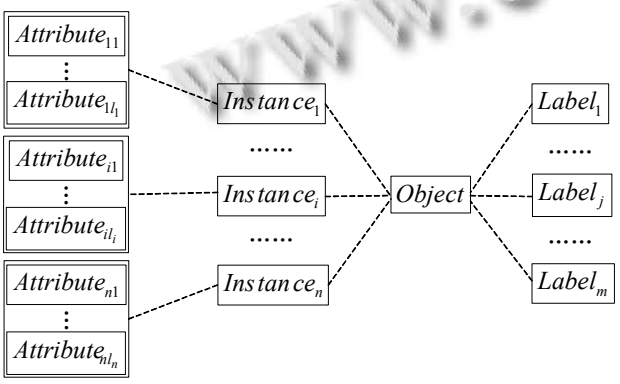


图 1　系统总体框图

3　多示例多标记算法描述

3.1 MIML 框架学习策略

传统监督学习可以看做是多示例学习或者多标记学习的一种特殊情况, 而多示例学习或者多标记学习有可以看成是多示例多标记学习的单标记情况或者单示例情况, 因此, 多示例多标记问题可以先转换成为多示例问题或者多标记问题, 再转化为单示例单标记问题, 也就是传统监督问题, 基于这种思想, Zhou Z-H 等提出了 MIMLBOOST 算法和 MIMLSVM 算法[7,9]. 基于最大间隔策略以及正则化机制, 提出了 D-MIMLSVM 算法[8]和 M^{3}MIML [10]算法. 在本文中, 使用的是 MIMLSVM 算法.

3.2 MIMLSVM 学习算法

MIMLSVM 算法针对每个多示例多标记样本 (X_i,Y_i) 都会给出一个中间变量 $z_i=\phi(X_i)$, 函数 ϕ 将每一个多示例子集 X_i 转化成为一个示例 z_i, 即 $2^x\to Z$, $Z\times Y\to\{-1,+1\}$, 其中, 对于任意的 $y\in Y$, 若 $y\in Y_i$, 则令 $\phi(z_i,y)=+1$, 否则, $\phi(z_i,y)=-1$. 将 X_i 转换成为 z_i 后, 再利用 MLSVM[11]算法对通过转换获得的多标记问题进行学习. MIMLSVM 算法的整体思想如图 2 所示.

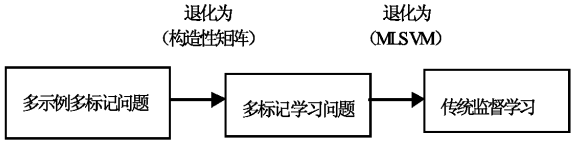


图 2　MIMLSVM 算法整体思想

3.3 算法描述

步骤 1. 针对所有的多示例多标记样本 (X_i,Y_i) $(i=1,2,\Lambda,m)$, 建立包含所有示例的数据集 Γ, $\Gamma=\{X_i|i=1,2,\Lambda,m\}$.

步骤 2. 在数据集 Γ 的基础上运用 k-medoids 聚类算法, 得到初始化的聚类中心点 $M_t(t=1,2,\Lambda,k)$. 对于每一个示例 $X_i\in(\Gamma-\{M_t|t=1,2,\Lambda,k\})$, 都有:

$$j=\arg\min_{t\in\{1,2,\Lambda,k\}}d_H(X_i,M_t) \tag{1}$$

$$\Gamma_i=\Gamma_i\cup\{X_i\} \tag{2}$$

$$M_t=\arg\min_{A\in\Gamma_t}\sum_{B\in\Gamma_t}d_H(A,B)(t=1,2,\Lambda,k) \tag{3}$$

重复计算公式(1)(2)(3), 直至中心点 M_t 不再改变. 其中, $d_H(A,B)$ 表示包 $A=\{a_1,a_2,\Lambda,a_{n_A}\}$ 和包 $B=\{b_1,b_2,\Lambda,b_{n_B}\}$ 之间的距离, 采用 Hausdroff[12]距离进行度量:

$$d_H(A,B)=\max\{\max_{a\in A}\min_{b\in B}\|a-b\|,\max_{b\in B}\min_{a\in A}\|b-a\|\} \tag{4}$$

其中$\|a-b\|$表示的就是使用 Hausdroff 距离计算出的 a 与 b 之间的距离.

步骤3. 将多示例多标记样本(X_i,Y_i)转换成为多标记样本(z_i,Y_i) $(i=1,2,\Lambda,m)$, 其中:

$$z_i=(z_{i1},z_{i2},\Lambda,z_{ik})=(d_H(X_i,M_1),d_H(X_i,M_2),\Lambda,d_H(X_i,M_k)) \quad (5)$$

步骤 4. 建立数据集D_y:

$$D_y=\{(z_i,\Phi(z_i,y))|u=1,2,\Lambda,m\} \quad (6)$$

针对数据集D_y使用 MLSVM 算法来进行 SVM 训练:

$$h_y=SVMTrain(D_y) \quad (7)$$

对于任意的$y\in Y$, 与标记集Y有关的示例都被认为是正例, 而对任意的$y\notin Y$, 与标记集Y无关, 被认为是反例.

步骤 5. 当训练后的 SVM 得分中含有正分时, 测试用例被标记为拥有最高正得分的类别, 若SVM训练后所有类别的得分都是负分, 则测试用例被标记为拥有最少负分的类别.

$$Y^*=\left\{\arg\max_{y\cup Y}h_y(z^*)\right\}\cup\{y|h_y(z^*)\geq 0,y\in Y\} \quad (8)$$

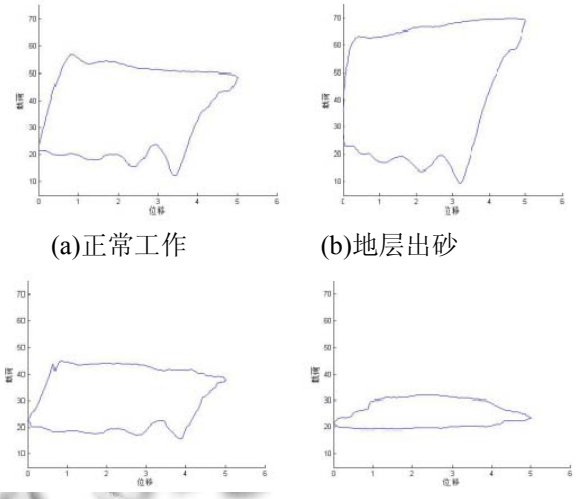
$$z^*=(d_H(X^*,M_1),d_H(X^*,M_2),\Lambda,d_H(X^*,M_k)) \quad (9)$$

3.4 算法表现评价

算法在实验中表现的好坏主要由 5 个指标进行评价, 分别是 HammingLoss、RankingLoss、OneError、Coverage、Average_Precision[13]. 其中, HammingLoss 表示的是对象分类错误的次数, 其数值越小, 算法的表现越好; RankingLoss 表示对象错乱标记平均值, 越小表示学习效果越好; OneError 表示排名第一的标记并不是该对象正确的标记的次数, 同样, 该值越小越好; Coverage 代表覆盖对象所有标记的距离, 数值越小表示覆盖精度越高; Average_Precision 代表标记排名平均分高于一个特定的标记$y\in Y_i$, Average_Precision 为 1 时, 是算法表现最好的时刻.

4 抽油机故障诊断应用举例

本文中对抽油机正常工作、地层出砂、泵漏失前期、泵漏失共四种工作状态进行诊断, 四种工作状态对应的示功图曲线如图 3 所示.



(a)正常工作 (b)地层出砂

(c)泵漏失前期 (d)泵漏失

图 3 抽油机示功图曲线

本次试验共使用了 173 条抽油机工作数据作为样本数据, 其中 123 个样本最为训练数据集(trainbags), 50个样本作为测试数据集(testbags). 173 个样本数据中 49 个样本数据的上载荷稳定在 55KN 左右, 且功图曲线光滑平稳, 满足正常工况数据特征; 35 个样本数据的上载荷下降, 重复度降低, 为泵漏失前期的特征; 42 个样本数据几乎不见上载荷, 为泵漏失数据的特征; 35 个样本数据加载线的斜率明显增大, 为地层出砂的典型特征. 173 个样本数据对应的功图见图 4, 与图 3 进行比较, 可看出样本数据很好的代表了本文中进行诊断的 4 种抽油机工作状态.

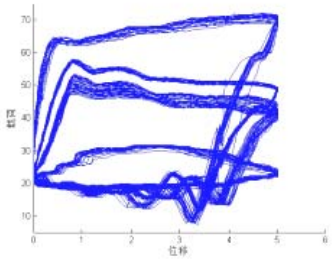


图 4 样本数据叠加示功图

训练中使用的SVM核函数为RBF核函数, γ值分别取 0.25、0.5、3. 另外, 输入参数 ratio*样本包个数即为聚类数目, 在本次试验中 ratio 取值为 0.0325; 另一个输入参数 cost 设置为 1. 具体实验结果见表 1.

表 1 MIMLSVM 实验表现

γ	evaluation metric				
	Hamming-Loss$^{\downarrow}$	Ranking-Loss$^{\downarrow}$	One-Error$^{\downarrow}$	Coverage$^{\downarrow}$	Average-Precision$^{\uparrow}$
0.25	0.2050	0.2115	0.3311	0.3000	0.8747
0.5	0.2182	0.2375	0.3946	0.5028	0.7129
3	0.2430	0.2402	0.4285	0.4964	0.6690

从表 1 中可以看出, 当 γ 值为 0.25 时, 各项算法评价指标体现出最好的结果, 此时聚类数目为 4, 即当 $\gamma = 1/k$ (k 为聚类数目)时, 算法 MIMLSVM 能够取得最佳效果.

5 结语

本文利用多示例多标记学习模型来解决抽油机故障诊断问题, 诊断结果能够使人满意. 由于抽油机是一个连续工作的机器, 其工况数据同样也是随时间连续的, 因此, 使用连续曲线之间距离度量方法代替 Hausdroff 距离计算两包之间的距离一个今后值得研究的方面.

参考文献

1 张楠.基于示功图分析的抽油机故障诊断系统[硕士学位论文].大连:大连理工大学,2009.

2 王兵,许少华,孟耀华.基于半监督竞争过程神经网络的抽油机故障诊断.信息与控制,2014,43(2):235–240.

3 张强,许少华,李盼池.对传过程神经网络在油井故障诊断中的应用.计算机工程与应用,2013,49(2):9–12.

4 付光杰,周昕奇,王磊,牟海维.基于矩特征傅里叶描述的示功图故障诊断研究.化工自动化及仪表,2015,42(4):401–405.

5 魏军.基于支持向量机的抽油机故障诊断模型研究.计算机与数字工程,2014,42(11):2094–2098.

6 檀朝东,曾霞光,檀革勤,张杰.基于最小二乘法的抽油机示功图自动分类及故障诊断.数据采集与处理,2010,25(增刊):157–159.

7 Zhou ZH, Zhang ML. Multi-instance multi-label learning with application to scene classification. In: Schölkopf B, Platt J, Hofmann T, eds. Advances in Neural Information Processing Systems 19 (NIPS'06), Cambridge, MA: MITPress, 2007: 1609–1616.

8 Zhou ZH, Zhang ML, Huang SJ, Li YF. MIML: A framework for learning with ambiguous objects. CORRabs/ 0808.3231, 2008.

9 Chang CC, Lin CJ. Libsvm: A library for support vector machines [Technical Report]. Department of Computer Science and Information Engineering, Taiwan University, Taipei, 2001.

10 Zhang ML, Zhou ZH. A maximum margin method for multi-instance multi-label learning. Proc. of the 8th IEEE International Conference on Data Mining (ICDM'08). Pisa, Italy. 2008. 688–697.

11 Boutell MR, Luo J, Shen X, Brown CM. Learning multi-label scene classification. Pattern Recognition, 2004, 37(9): 1757–1771.

12 Edgar GA. Measure, Topology, and Fractal Geometry. Springer, Berlin, 1990.

13 Zhou ZH, Zhang ML, Huang SJ, Li YF. Multi-instance multi-label learning. Artificial Intelligence, 2008, 176(1): 2291–2320.