DOI: 10.13382/j.jemi.B2104511

基于 LDA-CLCBA 组合模型的高速铁路道岔故障诊断*

林海香 卢 冉 陆人杰 许 丽 赵正祥 白万胜
(兰州交通大学自动化与电气工程学院 兰州 730070)

摘 要:ZY(J)7 电液道岔转换设备已在高速铁路大量投入使用,对其进行精确的故障诊断有助于高速铁路道岔的日常维护作业。以 ZY(J)7 道岔故障文本数据作为研究对象,提出一种基于 LDA(latent dirichlet allocation)主题模型与关联规则分类技术相结合的高速铁路道岔故障诊断模型。该模型首先采用 LDA 主题模型实现 ZY(J)7 道岔故障文本数据的特征提取;其次,由于道岔各故障类别数据的不均衡性,将原有的关联规则分类算法引入类支持度相关概念进行不平衡数据的处理,最终实现 ZY(J)7 道岔的故障诊断。通过对某铁路局 2017~2019 年的 ZY(J)7 道岔故障文本数据进行实验分析,实验结果表明提出的故障诊断方法分类精确率和召回率分别达到 95.08%和 90.24%,既保证了整体分类的准确率又有较好的小类别分类性能。

关键词: ZY(J)7 道岔;故障诊断;LDA 主题模型;关联规则分类;类支持度;类支持度阈值
中图分类号: U216.42 **文献标识码**: A **国家标准学科分类代码**: 520.20

Fault diagnosis for turnout of high-speed railway based on LDA-CLCBA hybrid model

Lin Haixiang Lu Ran Lu Renjie Xu Li Zhao Zhengxiang Bai Wansheng
(School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract: ZY(J)7 electrohydraulic turnout switch equipment has been widely used in high-speed railway, and accurate fault diagnosis is helpful to the daily maintenance of high-speed railway turnout. Taking the fault text data of ZY(J)7 turnout as the research object, a fault diagnosis model for high-speed railway turnout was proposed, which combined LDA topic model with association rules classification technology. Firstly, this model adopted LDA topic model to extract the feature of ZY(J)7 turnout fault text data. Secondly, due to the unbalanced data of each fault type of turnout, the original association rule classification algorithm was introduced into the concept of class support to deal with unbalanced data, and finally the fault diagnosis of ZY(J)7 switch was realized. Through the experimental analysis of ZY(J)7 turnout fault text data of a railway bureau from 2017 to 2019, the experimental results indicate that the classification precision and recall rate of the proposed fault diagnosis method are 95.08% and 90.24% respectively, which not only guarantees the accuracy of the whole classification, but also gets better classification performance of minority class.

Keywords: ZY(J)7 turnout; fault diagnosis; LDA topic model; association classification; class support; class support threshold

0 引 言

ZYJ(7)电液道岔作为高速铁路信号设备中的重要部分,可用来实现转换锁闭道岔尖轨基本轨或心轨,经过多年来的运营积累,产生了大量的道岔维护故障文本数据,而现阶段高速道岔的故障主要依靠铁路工作人员及专家的经验进行排查,并没有充分利用文本数据,致使维修效率相对较低。因此,对道岔故障文本数据进行深度挖掘,实现高速道岔故障自动化诊断是当前急需解决的问题。

近阶段,有学者在铁路领域内实现了以文本数据为基础的自动化故障诊断方法,诊断主要包括两个方面:文本特征提取和文本分类模型的构建。在文本特征提取方面,文献[1-2]采用 PLSA(probability latent semantic analysis)主题模型对车载故障文本数据进行特征选择,

收稿日期:2021-07-09 Received Date:2021-07-09
*基金项目:甘肃省高等学校创新基金项目(2020B-104)、2021 年度甘肃省优秀研究生“创新之星”项目(2021CXZX-606)资助

但 PLSA 主题模型的概率模型不一致,对于待预测的未知样本,无法选取合适的先验参数,且若训练数据过大,会导致过拟合。文献[3-4]利用 TF-IDF(term frequency-inverse document frequency)实现铁路信号系统的车载设备故障文本特征提取。文献[5]通过向量空间模型(vector space model,VSM)将车载日志转化为向量形式,在已向量化的数据上利用粗糙集进行特征选择。但 TF-IDF 和 VSM 均未考虑词与词之间位置顺序和相关性,且存在严重数据稀疏问题,造成维数灾难,致使分类效果不佳。而文献[6]提出的 LDA 主题模型可有效避免这些问题,其作为全概率模型,具有清晰的内在结构,参数空间规模与训练的文档数量无关,适合处理大规模文本数据。在文本分类模型的构建上,当前采用的 NB[7]、SVM[8]、KNN[9]、关联规则[10]等分类模型已取得了不错的效果,如文献[11]利用 SVM 模型实现对铁路道岔故障文本的分类;文献[12]将关联规则应用于铁路信号设备的故障分类。但这些分类器对于平衡样本数据分类效果良好,当应用到不均衡训练样本时,效果并不明显。

文中对高速 ZY(J)7 道岔故障文本数据进行分析,结合专家经验,将高速 ZY(J)7 道岔故障分成 12 种类别;将 LDA 主题模型与关联规则分类技术相结合运用到道岔故障诊断中。其中,通过 LDA 主题模型构建特征向量空间,实现高速 ZY(J)7 道岔故障文本数据的特征选择;针对各类别故障的不均衡性,在关联规则分类算法(classification based on associations,CBA)中引入类支持度[13]相关概念以提高故障分类的效果。为证明模型的可靠性,通过对某铁路局 2017~2019 年高速 ZY(J)7 道岔的故障文本数据进行实验验证。

1　高速铁路道岔故障数据分析

1.1　高速道岔故障数据的不均衡性

当前我国使用的高速道岔主要类型有:ZDJ9 型、ZY(J)7 型以及 S700K-C 型道岔。本文以 ZY(J)7 电液道岔转换设备的故障文本数据作为分析对象,再依据专业领域相关知识,将高速 ZY(J)7 道岔故障类型进行细致划分为 12 类,结果见表 1。某铁路局 2017~2019 年高速铁路 ZY(J)7 道岔设备故障数据的分布情况如图 1 所示。

由图 1 可知,道岔故障主要以电机油泵组、接点组和油缸组为主,底壳、锁闭表示杆和断路器故障等故障相对较少,其中最大类别与最小类别的故障数据比例高达 28∶1,存在着严重的数据不均衡性。若直接采用相关的故障诊断模型进行诊断,会造成小类别故障分类不明显的后果,这是故障诊断方法研究过程中必须重视的问题。

表 1　高速 ZY(J)7 道岔故障类别

Table 1　Category of high speed ZY(J)7 turnout fault

故障标号	故障类别
C1	电机油泵组
C2	油缸组
C3	接点组
C4	油管接头
C5	底壳
C6	空动缸
C7	动作杆
C8	锁闭表示杆
C9	工务病害
C10	继电器故障
C11	断路器故障
C12	电缆线路相关故障

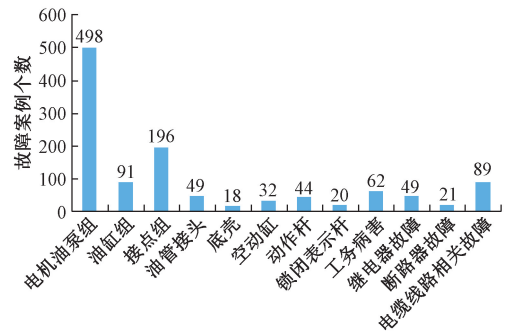


图 1　高速 ZY(J)7 道岔故障数据分布情况

Fig. 1　Distribution of fault data of high speed ZY(J)7 turnout

1.2　高速道岔故障文本数据

高速 ZY(J)7 道岔故障文本数据是由电务相关人员以自然语言方式记录的,主要包括故障现象和故障分类等。表 2 给出部分高速 ZY(J)7 道岔故障文本数据实例。

表 2　部分高速 ZY(J)7 道岔故障文本数据实例

Table 2　Fault text data examples of some high speed ZY (J) 7 turnout

序号	故障现象	故障类型
1	2017 年 8 月 24 日,朔黄至神池南,主机 121177# 空动缸紫铜垫密封不良导致漏油。	空动缸故障
2	2017 年 9 月 13 日,郑州至郑州东,心一 11992# 启动油缸压力传感器处渗油,无法拆卸。	油缸组故障
3	2018 年 12 月 18 日,济南至牛王村,尖一 173046 #转辙机内滚轮不解锁,造成道岔无法动作。	接点组故障
4	2019 年 6 月 17 日,成都至乐山,尖二 174602#动作杆解锁压力大,现场反映解锁曲线高。	动作杆故障

2　基于组合模型的高速铁路道岔故障诊断

2.1　诊断模型的建立

基于 LDA-CLCBA 组合模型的高速 ZY(J)7 道岔故障诊断实现过程如图 2 所示。其中,数据处理层是对高速 ZY(J)7 道岔故障文本数据进行分词处理,生成原始故障词库;特征提取层是实现原始故障词库的结构化处理,得到词项-文档矩阵,接着通过 LDA 主题模型将其转化为主题-文档矩阵,实现语义级故障特征提取,最后对主题-文档矩阵进行离散化处理,得到故障特征矩阵。故障诊断层是将故障特征矩阵送入类支持度关联分类规则(class support classification based on associations,CLCBA)分类器中,实现高速 ZY(J)7 道岔故障分类。

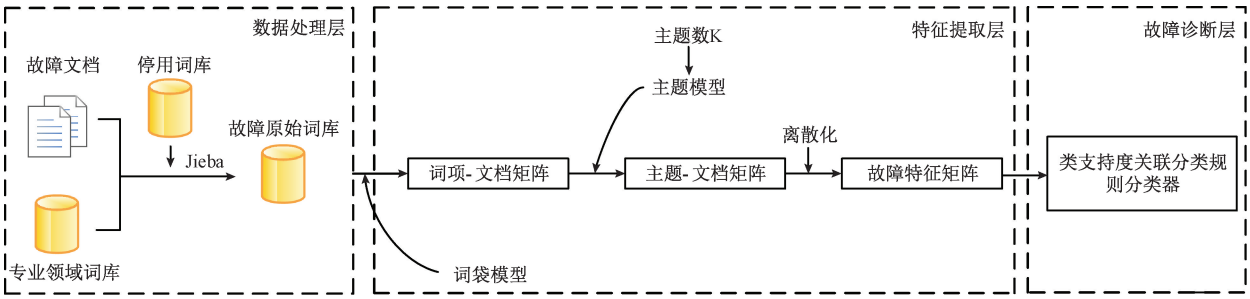


图 2　高速 ZY(J)7 道岔故障诊断实现过程

Fig. 2　Realization process of fault diagnosis for high speed ZY (J) 7 turnout

2.2　高速道岔故障数据处理

对故障文本进行结构化处理首先要实现故障文本的分词,本文采用 Jieba 分词工具实现该任务。由于中文通用词库在高速道岔设备方面不包含相应的专业词汇,因此需要建立专业领域词库,在分词过程中把领域词库中的词项进行保留,同时剔除故障发生的时间、地名等无意义的词项,得到高速道岔故障原始词库,图 3 为道岔专业领域词库中的部分词项。

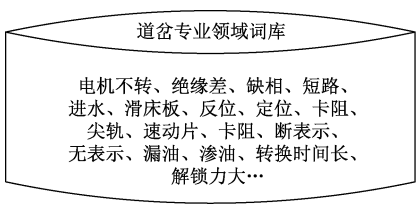


图 3　高速 ZY(J)7 道岔专业领域词汇

Fig. 3　Professional field vocabulary of high speed ZY(J)7 turnout

2.3　基于 LDA 主题模型的文本特征提取

数据处理阶段生成的故障原始词库,在采用“词袋模型”假设下,将经过分词后的词项转化为向量形式,进而生成词项-文档矩阵。但是生成的词项-文档矩阵未将不同词项的次序和关系进行考虑,并且无法解决同义词和一词多义的问题。文中采取的 LDA 主题模型作为潜在语义信息挖掘的工具,能够有效解决故障特征提取中语义缺失的问题。模型将词项-文档矩阵进一步转化为主题-文档矩阵,进而实现语义级故障特征的提取。LDA 主题模型的生成过程如图 4 所示,图中参数含义如表 3 所示。

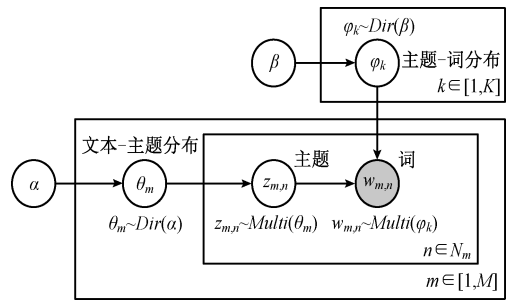


图 4　LDA 主题模型的生成过程

Fig. 4　Generation process of LDA topic model

表 3　生成过程的参数含义

Table 3　The meaning for the parameters of the generation process

参数	含义
K	文本主题个数
M	文本的总数
N_m	文本 m 的长度
α、β	先验参数
$w_{m,n}$	文本 m 的第 n 个词
$z_{m,n}$	文本 m 的第 n 个主题
θ_m	文本 m 的主题分布
φ_k	主题 k 的词项分布

根据图 4 可知,由先验参数 α 得到第 m 个文档的主题分布 θ_m,在 θ_m 中选择一个主题,并由 β 得到这个主题的词项分布 φ_k,随后从此词项分布中选择一个词,重复上述步骤,直至生成整个文档。

Gibbs 抽样算法[14-15]。隐含变量 θ 和 φ 可采用 Gibbs

抽样算法进行计算。经过推导,可以得出主题的条件分布概率公式为:

$$p(z_i = k \mid z_{\neg i}, w) = \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_{m,\neg i}^{(k)} + \alpha_k} \cdot \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t} \tag{1}$$

式中:$n_{m,\neg i}^{(k)}$ 代表的是第 m 个文档的词项 i 所对应的主题 K 下的出现次数;$n_{k,\neg i}^{(t)}$ 代表的是第 k 个主题对应词项 t 出现的概率。最后,得出的 θ 和 φ 为:

$$\theta_{m,k} = \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_{m,\neg i}^{(k)} + \alpha_k} \tag{2}$$

$$\varphi_{k,t} = \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t} \tag{3}$$

1)主题数判定

进行 LDA 主题挖掘时,首先确定故障数据中主题 K 的个数,通过预先设定主题 K 个数的大致区间进行算法寻求最优解。采取最大似然估计值进行计算。似然函数如下:

$$p(X \mid \theta) = \prod_{x_1}^{x_n} p(x_i \mid \theta) \tag{4}$$

对似然函数两边取对数,生成新的似然函数如下:

$$p(X \mid \theta) = \sum_{x_1}^{x_n} \log(x_i \mid \theta) \tag{5}$$

2)离散化处理

主题-文档矩阵离散化可以实现诊断效果提高。依据各主题在各文档中所占的概率,对矩阵进行离散化处理,首先设定离散的区间数,接着根据概率的大小设定重要度范围,进行量化处理。

2.4　基于 CLCBA 的高速道岔故障分类

Liu 等[16]提出了关联规则分类算法 CBA,该算法能够将关联规则应用到数据分类当中。CBA 算法未考虑样本中各类别的不均衡性,生成的部分规则置信度不高,特别是难以生成高置信度的小类规则。为解决上述问题,设计了一种类支持度关联规则分类算法,实现高速 ZY(J)7 道岔的故障分类。

1)关联规则分类的基本定义

由 Apriori 算法[17-18]挖掘所有的分类关联规则 CARs(class associations rules)。对于事务集 I,其中的 CAR 可表示为 $A \Rightarrow B(A \in I,\ B \in I,\ A \cap B = \varnothing)$,$A$ 和 B 分别为前提和后果。在 CARs 中,B 被限制为分类任务中的类别属性,A 为若干个特征属性组成的集合。

关联规则分类的几个重要指标如下:

(1)支持度表达式为:

$$\text{Support}(A) = \frac{\text{Count}(A)}{|I|} \tag{6}$$

式中:Count(A)为包含项集 A 的所有事务集,$|I|$ 为总事务集数。支持度可以衡量项集 A 在事务集中出现的频繁程度。

(2)置信度表达式为:

$$\text{Confident}(A \Rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)} \tag{7}$$

置信度可以度量在包含项集 A 的事务数中类别 B 出现的频繁程度。

(3)频繁项集

支持度阈值,是用户设置的衡量标准。若 A 项集的支持度不小于支持度阈值,则称 A 项集为频繁项集。

(4)提升度表达式为:

$$\text{Lift}(A \Rightarrow B) = \frac{\text{Support}(A \Rightarrow B)}{\text{Support}(A) \cdot \text{Support}(B)} \tag{8}$$

若 Lift>1,则 A 与类别 B 为正关联,反之则为负关联,需去除负关联规则。提升度越高,A 对类别 B 的影响越大。

(5)针对不平衡数据,引入类支持度(class support),其表达式为:

$$\text{CLsup}(A \Rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(B)} \tag{9}$$

式中:若某类别 B 的数据较少,则关联规则 $A=>B$ 的类支持度有可能较高。

(6)类频繁项集

若 A 项集在类别 B 的类支持度不小于类别 B 的类支持度阈值,则称 A 项集为类别 B 的类频繁项集。

2)CLCBA 算法分类原理

(1)规则生成。与 CBA 算法设置全部故障数据集的支持度阈值不同,CLCBA 算法是对每个故障类别设置各自的类支持度阈值,然后对各类故障的类频繁项集进行挖掘,最后根据类频繁项集生成满足置信度阈值的各类故障的规则。

(2)规则排序。与 CBA 算法相同,CLCBA 算法是根据置信度大小排序,而不同的是,若置信度大小相同,则其以类支持度大小排序。由式(9)可知,小类别故障的规则也可具有较高的类支持度,使得小类别故障的规则也可能优先排列。

3)CLCBA 算法的挖掘流程

(1)计算各故障类别的类支持度阈值。式(10)是综合最大类别故障样本、其他各类别故障样本和最大类类支持度阈值总结出的各类故障的类支持度阈值公式。

$$\text{minCLsup}(B_n) = \text{minCLsup}(B_m) \cdot \frac{\lg(\text{Count}(B_m)/|I|)}{\lg(\text{Count}(B_n)/|I|)} \tag{10}$$

式中：B_m 为故障数据最多的电机油泵组类别，minCLsup(B_m)为电机油泵组的类支持度阈值，即最大类类支持度阈值。

(2)独立挖掘各类故障的关联分类规则。

(3)合并所有故障类别的关联规则。

(4)将合并的关联规则排序；剔除掉提升度小于 1 和分类错误的规则；同时，若产生规则形如 $A=>B_i$ 和 $A=>B_j$(B_i 与 B_j 为某一类别)，将置信度较小的规则也要剔除，综合以上形成最终的规则集。CLCBA 算法的挖掘流程如图 5 所示。

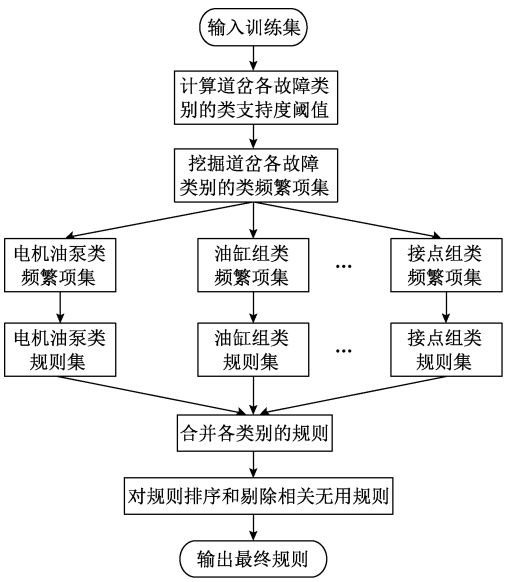


图 5　CLCBA 算法挖掘过程

Fig. 5　Mining process of CLCBA

3　实验分析

3.1　模型评价指标

选取了某铁路局 2017~2019 年 1 169 条高速 ZY(J)7 道岔设备故障数据进行实验验证，其中用于训练的部分占 85%，测试部分占 15%。实验采用 1)精确率、2)召回率、3)F_1 值 3 种指标对所提模型进行评价和对比。

精确率的计算公式为：

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

召回率的计算公式为：

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

以上两个指标是分别通过查准与查全两个方面对分类结果进行评价；

F_1 值则是以上两个指标的调和平均值，公式为：

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{13}$$

上述指标由混淆矩阵中的参数决定，各参数的含义如下：

TP：正确预测到的此类的样本个数；

FN：属于此类却误被预测至其他类的样本个数；

FP：错误预测至此类的样本个数。

3.2　LDA 主题模型训练

对于 LDA-CLCBA 模型中的 LDA 部分，通过对数似然函数来求解最优主题数 K。根据道岔故障文本数据，假设主题范围在 4~20 个，经过逐次的代码实验，不断缩小步长，由图 6 可知最佳主题个数为 14。

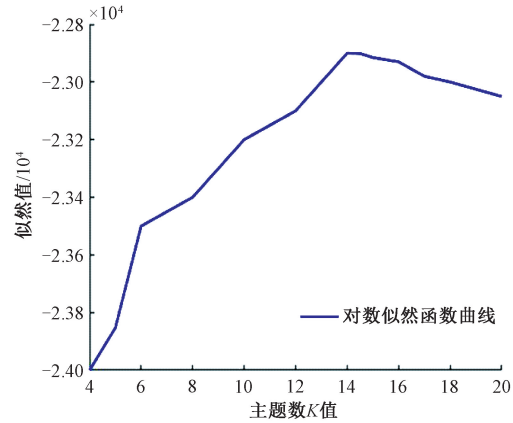


图 6　最优主题数的确定

Fig. 6　Determination for the optimal number of topics

确定了故障主题数为 14 个之后，便可通过 LDA 主题模型训练出主题特征结果。表 4 为高速 ZY(J)7 道岔故障文本数据主题特征的训练结果。

表 4　高速 ZY(J)7 道岔故障文本的主题特征训练结果

Table 4　Thematic features training results of high speed ZY(J)7 turnout fault text

主题编号	主题词项
Topic1	尖轨、基本轨、卡异物、无表示、挤、冰雪…
Topic2	漏油、渗油、溢流压力、压力、缺油…
Topic3	电缆、端子、分线盘、引接线、启动…
Topic4	电机、不转、转子、绝缘差、缺相、短路、进水…
Topic5	断路器、打落、短路、分线盘、磨损…
Topic6	继电器、线圈、2DQJ、1DQJ、不转极、不动作…
Topic7	锁钩、锁轴、解锁、锁闭框、解锁力大、锁闭杆…
Topic8	尖轨、滑床板、爬行、心轨、不密贴、反弹…
Topic9	动接点、静接点、接触不良、长花键轴、螺丝…
Topic10	定位、定操、反位、反操、控制台、压力、缺油…
Topic11	卡缺口、表示缺口、偏大、检查柱、密检器、卡口…
Topic12	松动、脱落、断裂、螺栓、螺丝…
Topic13	解锁、压力大、动作、不同步、曲线、报警…
Topic14	油泵、手摇、不起压、异响…

通过表4可知,Topic1是有关“道岔卡异物”的相关内容,Topic2为“转辙机漏油渗油”的相关内容,Topic3为“电缆芯线”的相关内容等。

最后对主题-文档矩阵离散化处理。设定离散区间数为3,当主题在文本中的概率为[0, 0.1)内时,将其量化为0(主题在文本中并不重要);当主题概率在文本中的概率在[0.1, 0.5)时,将其量化为1;当主题概率在文本中的概率在[0.5, 1)时,将其量化为2,生成最终的故障特征矩阵。图7为故障文本特征提取的表达总过程。

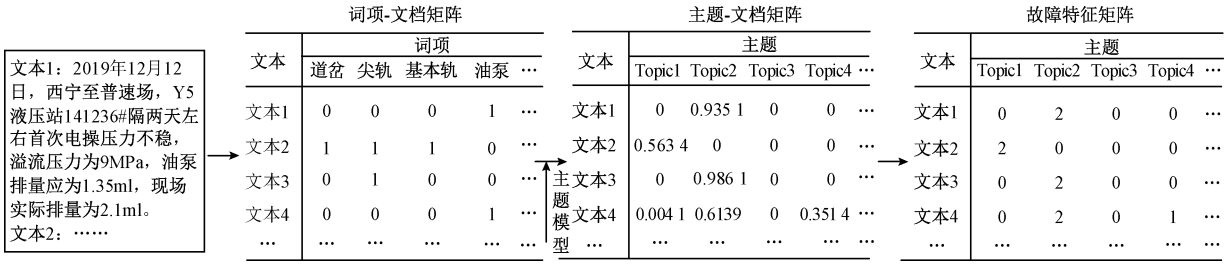


图7　故障文本特征提取的表达总过程

Fig. 7　The general process of fault text feature extraction

3.3　CLCBA 模型训练

对于LDA-CLCBA组合模型中的CLCBA部分,其分类准确率决定因素之一为最大类类支持度阈值。将阈值设定在区间范围[0.1,0.7]内,通过测试集的预测精度作为模型最终的诊断精度,获得能使模型最优化的阈值。图8为不同最大类类支持度阈值下的分类准确率,从图中可知,若阈值设置过小时,生成大量质量不好的规则,导致部分样本数据错误分类;若阈值设置过大时,生成的规则过少,导致部分规则丢失。当阈值设置在[0.25, 0.5]时,模型的整体准确率最高,为95%左右。因此,本文设置最大类类支持度阈值为0.35。

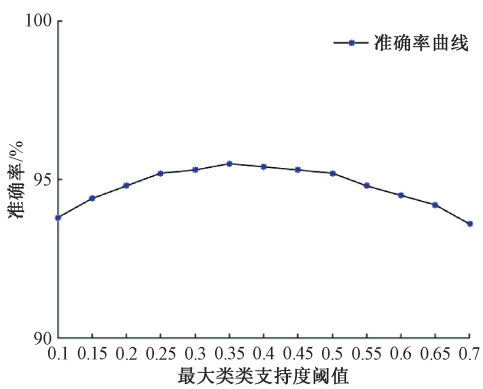


图8　不同最大类类支持度阈值下的准确率

Fig. 8　Accuracy under different maximum class support thresholds

根据式(10)计算出各故障类别的类支持度阈值,置信度阈值为0.5。其中,置信度阈值的选取方法与最大类类支持度阈值一致。表5筛选出最终规则集中部分规则,用部分代表性规则解释本模型的分类效果。以序号2中的规则为例进行分析,由于项集{Topic5}只在小类别C11中出现,且由式(10)得出C11的类支持度阈值为0.074,所以{Topic5}为C11的类频繁项集且{Topic5}到{C11}的置信度为1。因此,CLCBA算法可以生成{Topic5} => {C11}的分类规则。而采用CBA算法时,即使在支持度阈值取到很小的0.02时,{Topic5}也不是训练集中的频繁项集,所以CBA算法无法生成{Topic5} => {C11}这条小类别C11的规则。

表5　最终规则集中的部分规则

Table 5　A partial rule of the final rule set

序号	规则	类支持度	置信度	提升度
1	{Topic13} => {C7}	0.931 8	1	26.568 0
2	{Topic5} => {C11}	0.857 1	1	55.666 7
3	{Topic9} => {C3}	0.688 7	1	18.853 7
4	{Topic8} => {C9}	0.660 4	1	18.867 9
5	{Topic7,Topic11} => {C8}	0.300 0	1	58.450 0
6	{Topic11} => {C8}	0.750 0	0.714 3	41.750 8
7	{Topic8,Topic10} => {C7}	0.113 6	0.625 0	18.977 3
8	{Topic2,Topic12} => {C2}	0.692 3	0.508 1	6.530 8
…	…	…	…	…

3.4　LDA-CLCBA 组合模型的性能验证

本文通过混淆矩阵可视化直观的展示了故障诊断模型对各故障类别在测试集上的判别结果。LDA-CBA模型和LDA-CLCBA模型在测试集上的混淆矩阵分别如图9和10所示。图中的横纵坐标的1~12代表高速ZY(J)7道岔的故障类别C1-C12。其中,在LDA-CBA模型中,也获取了最优化的支持度阈值与置信度阈值分别为0.1和0.5,使此模型也可达到最高的准确率,让所提出的LDA-CLCBA模型在与之对比后更具说服力。

从混淆矩阵中可以发现,由于各故障类别数据的不均衡性,高速道岔故障诊断模型中的小类别故障样本易被错误分类至大类别故障样本。通过两幅图的对比,得

出 LDA-CLCBA 模型在小类别故障上的识别率比 LDA-CBA 的识别率高。

将两个模型在全部测试集样本上的性能进行比较，如表 6 所示，LDA-CBA 模型精确率较低，查全率低下，表明此模型不利于小类别故障识别，影响了整体的分类效果。对于 LDA-CLCBA 模型而言，本模型相较于 LDA-CBA 模型召回率提升了 10.81%，查全效果提升。因此，LDA-CLCBA 模型可以较好的适应不均衡性的样本数据。

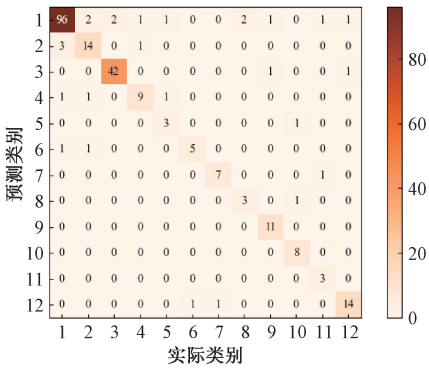


图 9　LDA-CBA 模型的混淆矩阵

Fig. 9　Confusion matrix of LDA-CBA model

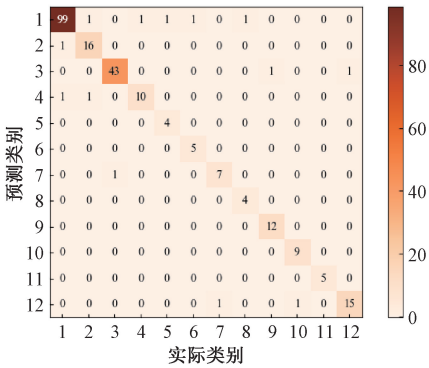


图 10　LDA-CLCBA 模型的混淆矩阵

Fig. 10　Confusion matrix of LDA-CLCBA model

表 6　LDA-CBA 和 LDA-CLCBA 模型性能比较

Table 6　Performance comparison between LDA-CBA and LDA-CLCBA models

模型	*Precision*	*Recall*	F_1
LDA-CBA	0.860 6	0.794 3	0.826 1
LDA-CLCBA	0.950 8	0.902 4	0.926 0

3.5　对比实验分析

在文本特征提取部分，将文中使用的 LDA 主题模型与文献[3-4]中的 TF-IDF 算法、文献[5]中的 VSM 模型、文献[1-2]中的 PLSA 主题模型进行对比实验，上述 4 种文本表示模型均与 CLCBA 结合，应用于高速 ZY(J)7 道岔故障文本分类时所取得的分类效果如图 11 所示。

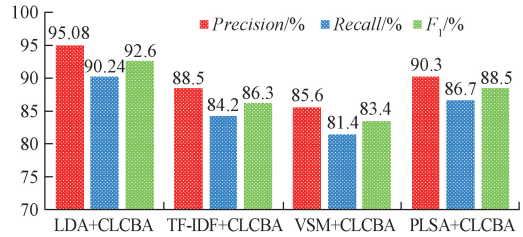


图 11　不同文本表示模型的实验结果

Fig. 11　Test results of different text representation models

图 11 表明 LDA 主题模型的特征表示方法在 3 项指标上均取得了良好效果，证明该方法可为后续模型提供较为理想的特征表示形式，以期在此后分类模型的训练中通过不断调整优化，能更全面地表示高速 ZY(J)7 道岔故障文本特征。

为进一步说明本文模型的有效性，在文本分类模型的构建上，将其与 NB[19]、LR[20]、SVM[21] 和 RF[22] 4 种分类模型进行对比实验，且上述模型均通过 LDA 主题模型完成特征提取。模型性能对比情况如表 7 所示。

表 7　各模型的性能对比

Table 7　Performance comparison of each model

模型	*Precision*	*Recall*	F_1
LDA-CLCBA	0.950 8	0.902 4	0.926 0
LDA-NB	0.746 1	0.685 8	0.714 7
LDA-LR	0.764 7	0.654 3	0.705 2
LDA-SVM	0.863 5	0.782 5	0.821 0
LDA-RF	0.865 1	0.800 3	0.831 4

由表 7 的诊断效果可知，对于 NB 和 LR 模型，二者的召回率只有 60%多，查全效果并不理想，无法有效地对不平衡样本进行分类；SVM 本身的泛化能力较强，较为适合不均衡样本的分类，分类的各指标有所提升，但其对参数选择过于敏感，若选择的参数并不理想，便会影响分类的效果；RF 在 4 种模型中分类最有效，因为 RF 采取了集成学习的方法，使其的泛化能力提高，但当样本数据噪声较大时，该算法容易出现过拟合。LDA-CLCBA 模型在 3 种指标上均优于以上 4 种模型，这是由于本模型中的 CLCBA 算法引入了类支持度的概念，采用每一类故障各自生成规则，对于小类别故障，它可以生成更多的具有高置信度的规则，并且对在置信度一致情况下的规则通过类支持度大小排列，使得小类规则也可能优先排列，因此故障诊断效果更好。

4　结　论

本文通过 LDA 主题模型对预处理后的高速 ZY(J)7

道岔故障文本数据进行特征表示,实现文本的结构化转换。同时,对关联规则分类算法进行优化,将其引入类支持度的相关概念,并对每一故障类别独立进行关联挖掘,使优化的算法在不均衡的道岔故障数据上更具有优势,提高诊断的效果。通过实验对比分析,所提出的基于LDA主题模型与CLCBA算法相结合的高速ZY(J)7道岔故障诊断方法在精确率、召回率和F_1值3种评价指标上均优于其他模型,验证了其能够有效提升高速ZY(J)7道岔故障文本的分类效果。

本文只是针对非结构化的文本数据进行特征提取,但在铁路信号系统中微机监测数据作为结构化数据也记录了重要的道岔故障信息。因此,未来还需研究非结构与结构化数据相融合的特征提取方法,以便充分利用各种数据,为道岔的维修提供指导。

参考文献

[1] 赵阳,徐田华,周玉平,等. 基于文本挖掘的高铁信号系统车载设备故障诊断方法[J]. 铁道学报,2015,37(8):53-59.
ZHAO Y,XU T H,ZHOU Y P,et al. Text mining based fault diagnosis for vehicle on-board equipment of high-speed railway signal system[J]. Journal of the China Railway Society,2015,37(8):53-59.

[2] 钟志旺,唐涛,王峰. 基于PLSA和SVM的道岔故障特征提取与诊断方法研究[J]. 铁道学报,2018,40(7):80-87.
ZHONG ZH W,TANG T,WANG F. Research on fault extraction and diagnosis of railway based on PLSA and SVM[J]. Journal of the China Railway Society,2018,40(7):80-87.

[3] 梁潇,王海峰,郭进,等. 基于贝叶斯网络的列控车载设备故障诊断方法[J]. 铁道学报,2017,39(8):93-100.
LIANG X,WANG H F,GUO J,et al. Bayesian network based fault diagnosis method for on-board equipment of train control system[J]. Journal of the China Railway Society,2017,39(8):93-100.

[4] 刘浩. 基于关联规则的高铁列控车载设备故障诊断方法研究[D]. 北京:北京交通大学,2018.
LIU H. Research on fault diagnosis method for CTCS on-board equipment of high-speed railway based on association rules[D]. Beijing: Beijing Jiaotong University,2018.

[5] 周璐婕,董昱. 基于GA-BP神经网络的列控车载设备故障诊断方法研究[J]. 铁道科学与工程学报,2018,15(12):3257-3265.
ZHOU L J,DONG Y. Research on fault diagnosis method for on-board equipment of train control system based on GA-BP neural network[J]. Journal of Railway Science and Engineering,2018,15(12):3257-3265.

[6] BLEI D M,NG A Y,JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research,2003,3:993-1022.

[7] ZHAO W, LUO Z. Web text data mining method based on Bayesian network with fuzzy algorithms[J]. Journal of Intelligent & Fuzzy Systems, 2020, 38(4): 3727-3735.

[8] 朱芳鹏,王晓峰. 面向船舶工业新闻的文本分类[J]. 电子测量与仪器学报,2020,34(1):149-155.
ZHU F P,WANG X F. Text classification for ship industry news[J]. Journal of Electronic Measurement and Instrumentation,2020,34(1):149-155.

[9] CHEN Z, ZHOU L J, LI X D, et al. The lao text classification method based on KNN[J]. Procedia Computer Science,2020,166.

[10] GE S, ZHUANG Y, HU Y, et al. Research on enterprise hidden danger association rules based on text analysis[C]. IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2019, 252(3): 032170.

[11] 杨连报,沈翔,李新琴,等. 基于文本分析的高速铁路道岔故障分类模型研究[J]. 中国铁路,2020(8):13-18.
YANG L B,SHENG X,LIN X Q,et al. Classification model of high speed railway turnout failures based on text analysis[J]. China Railway,2020(8):13-18.

[12] 路雅云,梁玉琦. 基于关联规则的铁路信号设备故障诊断技术研究[J]. 铁道技术监督,2019,47(2):46-49.
LU Y Y,LIANG Y Q. Research on fault diagnosis technology of railway signal equipment based on association rules[J]. Railway Quality Control,2019,47(2):46-49.

[13] 周忠眉,李家辉. 基于各类支持度阈值独立挖掘的关联改进算法[J]. 计算机工程与科学,2019,41(11):2088-2094.
ZHOU ZH M,LI J H. An associative classification algorithm based on various class-support thresholds and independent mining rules[J]. Computer Engineering & Science,2019,41(11):2088-2094.

[14] PARK H,PARK T,LEE Y S. Partially collapsed Gibbs sampling for latent Dirichlet allocation[J]. Expert Systems with Applications,2019,131(OCT.):208-218.

[15] 王瑞,龙华,邵玉斌,等. 基于Labeled-LDA模型的文本特征提取方法[J]. 电子测量技术,2020,43(1):141-146.
WANG R,LONG H,SHAO Y B,et al. Text feature

extract method based on Labeled-LDA model [J]. Electronic Measurement Technology, 2020, 43 (1): 141-146.

[16] LIU B, HSU W, MA Y. Integrating classification and association rule mining[C]. KDD' 98 Proceedings of the Rourth International Conference on Knowledge Discovery and Data Mining, 1998: 80-86.

[17] XIE H Y. Research and case analysis of apriori algorithm based on mining frequent item-sets [J]. Open Journal of Social Sciences, 2021, 9(4): 458.

[18] JI H P, WANG T Y, LIU J, et al. An efficient parallel association rules mining algorithm for fault diagnosis[C]. Key Engineering Materials. Trans Tech Publications Ltd, 2016, 693: 1326-1330.

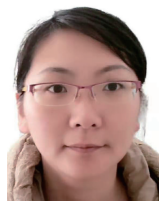
[19] 谢明军,何剑峰,胡小溪,等. 基于故障日志的城轨地面信号故障诊断 [J]. 北京交通大学学报, 2020, 44(5): 27-35.
XIE M J, HE J F, HU X X, et al. Fault diagnosis for urban rail transit trackside signaling equipment based on fault logs [J]. Journal of Beijing Jiaotong University, 2020, 44(5): 27-35.

[20] 张绍荣,朱志斌,冯宝,等. 基于组稀疏贝叶斯逻辑回归运动想象脑电信号分类模型的通道选择与分类新算法 [J]. 仪器仪表学报, 2019, 40(10): 179-191.
ZHANG SH R, ZHU ZH B, FENG B, et al. New channel selection and classification algorithm based on group sparse Bayesian logistic regression motor imagery EEG signal classification model [J]. Chinese Journal of Scientific Instrument, 2019, 40(10): 179-191.

[21] 上官伟,袁亚辉,王剑,等. 基于 Labeled-LDA 的列控车载设备故障特征提取与诊断方法研究 [J]. 铁道学报, 2019, 41(8): 56-66.
SHANGGUAN W, YUAN Y H, WANG J, et al. Research of fault feature extraction and diagnosis method for CTCS on-board equipment based on Labeled-LDA [J]. Journal of the China Railway Society, 2019, 41(8): 56-66.

[22] 周璐婕,党建武,王瑜鑫,等. 基于 CNN-CSRF 组合模型的列控车载设备故障诊断 [J]. 铁道学报, 2020, 42(11): 94-101.
ZHOU L J, DANG J W, WANG Y X, et al. Fault diagnosis for on-board equipment of train control system based on cnn-csrf hybrid model [J]. Journal of the China Railway Society, 2020, 42 (11): 94-101.

作者简介



林海香(通信作者),分别在 2000 年和 2007 年于兰州交通大学获得学士学位和硕士学位,2020 年于同济大学获得博士学位,现为兰州交通大学副教授,主要研究方向为交通信息数据挖掘。
E-mail: linhaixiang@ mail. lzjtu. cn

Lin Haixiang (Corresponding author) received her B. Sc. and M. Sc. degrees from Lanzhou Jiaotong University in 2000 and 2007, and Ph. D. degree from Tongji University in 2020, respectively. Now she is an associate professor at Lanzhou Jiaotong University. Her main research interest includes traffic information data mining.

卢冉,现为兰州交通大学硕士研究生,主要研究方向为自然语言处理。
E-mail: 191724705@ qq. com

Lu Ran is a M. Sc. candidate at Lanzhou Jiaotong University. His main research interest includes natural language processing.