文章编号:1671-4598(2020)07-0059-05 DOI:10.16526/j.cnki.11—4762/tp.2020.07.013 中图分类号:TP3 文献标识码:A

基于文本挖掘的高速铁路动车组故障多级分类研究

高 凡[1],李 樊[2],张 铭[2],王志飞[2],赵俊华[3]
(1.中国铁道科学研究院 研究生部,北京 100081;
2.中国铁道科学研究院集团有限公司,北京 100081;3.北京经纬信息技术有限公司,北京 100081)

摘要:针对高速铁路信号设备故障发生后记录的文本数据,提出基于文本挖掘方式的高速铁路信号设备故障多级分类模型研究;提出TF—IDF词汇权重与词汇字典结合的特征表示方法实现信号设备故障文本数据的特征提取;多级分类模型中,基于Stacking集成学习思想设计单层分类模型,将循环神经网络BiGRU和BiLSTM作为初级学习器,设计权重组合计算方法作为次级学习器,将多级分类任务分解为各层单分类任务,并采用K折交叉验证训练Stacking模型;采用高速铁路自开通至十年的信号转辙机故障数据,通过对故障原因文本数据的分析,实现故障部位和故障原因的二级分类,经过$K=5$次训练,BiGRU较BiLSTM各评价指标都较高,经实验BiGRU分配权重为0.7,BiLSTM权重为0.3,组合加权对两个网络的输出计算,准确率提高为0.881 4,召回率提高为0.864 2;实验表明多级分类模型能够有效提升信号设备故障多级分类任务的分类评价指标,并能够保证分类结果隶属关系的正确性。

关键词:高速铁路信号设备;多级分类;Stacking集成学习;循环神经网络;多任务协作投票决策树

Research on Multi—level Classification of High—speed Railway Signal Equipment Fault based on Text Mining

Gao Fan[1], Li Fan[2], Zhang Ming[2], Wang Zhifei[2], Zhao Junhua[3]
(1. Postgraduate Department, China Academy of Railway Science, Beijing 100081, China;
2. China Academy of Railway Sciences Corporation Limited, Beijing 100081, China;
3. Beijing Jingwei Information Technologies Co., Ltd., Beijing 100081, China)

Abstract: Aiming at the text data recorded after the failure of high—speed railway signal equipment, a multi—level classification model of high—speed railway signal equipment failure based on text mining is proposed. A feature representation method combining Term Frequency—Inverse Document Frequency (TF—IDF) word weight and word dictionary is proposed to extract the feature of signal equipment fault text data. In the multi—level classification model, the single—layer classification model was designed based on Stacking Integrated learning idea, the recurrent neural network Bidirection Gated Recurrent Unit (BiGRU) and Bidirection Long Short Term Memory (BiLSTM) were used as primary learners, and the weight combination calculation method was designed as secondary learners, multi—level classification tasks were decomposed into single classification tasks of each layer, and K—fold cross—verification was used to train Stacking model. After $k=5$ training, the evaluation indexes of bigru are higher than those of bilstm. The weight of bigru and bilstm was 0.7 and 0.3 respectively. The output of the two networks is calculated by combination weighting, the accuracy is improved to 0.881 4, and the recall rate is increased to 0.864 2. High—speed railway from the opening to a decade of signal switch machine failure data, the secondary classification of fault location and fault cause is realized by analyzing the text data of fault cause, experiment show that multi—level classification model can effectively improve the classification of signal equipment failure multi—level classification task evaluation index, and can ensure the correctness of the subordinate relations classification results.

Keywords: high—speed railway signal equipment; multilevel classification; stacking integrated learning; recurrent neural network; multi—task collaborative voting decision tree

0 引言

高速铁路信号设备是保障高速铁路行车安全的重要基础设施[1],随着高速铁路运营里程的积累,产生了海量的信号设备故障数据,这些故障数据大多以非结构化文本的形式存储,该数据蕴含了高速铁路安全的重要信息,长期由业务人员根据经验对数据进行故障设备诊断与分类,由于高速铁路中存在众多不同类型的信号设备,设备故障类型多,且设备与故障原因的隶属关系严谨,为深入开展高速铁路故障数据分析工作,需要对故障数据进行多级分类,而人工进行多级分类工作容易造成分类的不准确性,在智慧铁路和铁路大数据的建设下,亟需研究基于文本挖掘的

收稿日期:2020-04-06; **修回日期**:2020-05-12。
基金项目:国家自然科学基金(51967010);铁科院集团公司重点课题(2019YJ115);铁科院集团公司青年课题(2019YJ125);中国国家铁路集团有限公司科研专项课题(J2019X005)。
作者简介:高 凡(1987-),女,河北石家庄人,副研究员,博士研究生,主要从事交通信息工程及计算机控制方向的研究。

机器学习算法，实现高速铁路信号故障设备的多级分类。

多级别分类方法包括自上而下分类、全局分类和收缩分类方法[2-3]，高速铁路信号设备故障多级分类，采用自上而下分类方法中分而治之的策略，将设备故障多级分类问题分解为单层分类问题，通过设计单级分类模型得出每一级别的分类结果，最后通过多任务协作决策树投票策略，将各级的分类结果进行汇集与隶属关系矫正，实现信号设备故障的多级分类。

采用铁路安全文本特征提取和单层分类模型的研究方法，设计高速铁路信号设备故障多级分类模型[4]。首先针对高速铁路信号设备故障数据特点，提出基于词频－逆向文件频率（term frequency－inverse document frequency，TF－IDF）改良的特征提取方法[5]。针对故障类别样例数量差异较大，为避免防止单一分类器造成过拟合的问题，采用 K 折交叉验证＋ Stacking 分类模型实现单层分类模型[6]，Stacking 模型中提出将相似网路结构的变体双向门控循环单元（bidirection gated recurrent unit，BiGRU）与双向长短时记忆网络（bidirection long short term memory，BiLSTM）初级学习器[7]，设计整体与类别权重相结合的权重分配机制作为次学习器，提升单层分类模型的分类性能。通过 Staking 模型对各级别任务进行分类[8]，设计多任务协作投票决策树，实现多个级别分类结果的隶属关系矫正，同时提升整个多级分类模型的分类性能。最后应用高速铁路 2009 年到 2018 年信号设备故障数据进行实验，验证多级分类模型的有效性与正确性。

1 高速铁路信号故障文本特征提取

高速铁路信号设备故障数据来源于铁路牵引供电管理信息系统（EMIS），故障数据以结构化的形式记载了故障的详细信息，如表 1 所示，记录故障发生的原因信息以自然语言文本的形式存储。本文基于高速铁路故障原因分析文本数据，对故障的信息进行故障原因和部位二级分类。

表 1 高铁信号道岔故障部分样例数据

序号	原因分析	故障部位	故障原因
1	6 号道岔 J3 隐患排查。	密贴检查器	杆件调整
2	276＃－1 动道岔安全接点开路，经调整后恢复。	转辙机	其他
3	13＃道岔 J3 反位勾头与锁闭杆凸台处结冰（道岔类型：S700K）	外锁闭及安装装置	自然灾害
4	147＃－X1 反位密检器故障，后续要点更换后恢复	密贴检查器	接点组
5	1 号(1/42、S700K)道岔尖 1 滑床板夹冰造成卡阻。	外锁闭及安装装置	自然灾害

高速铁路信号设备故障原因分析文本数据中包含道岔、红光带、密贴器等具有特征的关键词，采用 TF－IDF 对故障文本数据进行特征提取[9]，TF－IDF 方法的原理是若某个词在样本中出现的频率越高，而有该词的样本在全文档中越少，说明该词对这个样本有着越高的辨识度，具有很好的区分能力。由于高速铁路信号设备故障文档数量较大，但是每个故障文档都是短文本，直接采用 TF－IDF 方法抽取特征，易造成特征向量冗余性和稀疏性，缺失了数据的特异性，所以本文针对高速铁路信号设备故障数据特征提取方法在 TF－IDF 方法基础上进行了改良。

改良的 TF－IDF 高速铁路信号设备故障文本数据特征抽取方法：首先要将中文文本内容进行分词，本文采用基于专业语料库和常用语料库的 Jieba 分词工具对信号故障文本分词[10]，并对助词如“的”，“了”等不能表示文档特征的词语进行清理，然后将分词后的词汇集合进行 TF－IDF 权重计算形成词汇权重矩阵，以及对每个词汇的数量统计形成词汇字典。TF－IDF 权重矩阵中 m 为文档数，n 为所有文档的词汇，所以 n 的维度很大，TF－IDF 权重矩阵具有严重的稀疏性。根据 TF－IDF 值为每个样本中的词进行排序，允许词汇重复，选取前 100 个最具有样本特征的词语，降低特征向量维度，并将词频替换为对应的词汇 ID，形成特征字典矩阵，将文本特征向量以及经过 one－hot 编码后的各级别的标签输入到文本分类模型中。

2 基于 Stacking 的信号设备故障单层文本分类模型计

高速铁路信号设备故障文本特征数据集分为训练集、验证集以及测试集输入到 Stacking 单层分类模型中。基于 Stacking 的高速铁路信号设备故障单层文本分类，通过将循环神经网络 BiGRU 与 BiLSTM 作为 Stacking 的初次学习器，将两个神经网络预测的结果作为特征来训练组合加权次级学习器，通过次级学习器整合初级学习器的预测结果。为了避免训练集训练出来的模型反过来预测训练集造成过拟合问题，以及训练多个单层分类模型，达到相同测试集产生多个预测结果的目的，采用了 K 折交叉验证方法，如图 1 所示。

2.1 软件设计思路和编程方法 BiLSTM 与 BiGRU 初级学习器原理

循环神经网络（RNN）是一种处理序列信息的神经网络，由于在结构上存在前后依赖关系，在自然语言应用上得到了广泛的应用。RNN 特殊性在于其在 t 时刻的输出 s_t，不仅取决于输入层的 x_t，而且还取决于上一节点的输出 s_{t-1}，其学习过程是一个预测下一个词的过程，例如，x_{t-1}，x_t，x_{t+1} 是一个输入“道岔定位无”，那么 o_{t-1} 和 o_t 对应的是“定位”和“无“这两项”，预测下一个次最有可能是什么，通过信号故障语料训练，o_{t+1} 最有可能的是“表示”。h_t 表示 t 时刻隐藏层的状态，x_t 表示 t 时刻的输入，o_t 表示 t 时刻的输出，s_t 表示 t 时刻的记忆单元，U，W 模型的线性参数矩阵。双向 RNN 同时考虑预测词的上文信息和下文信息，由前向后、由后向前均保留该词的重要信息，能够更加有效的进行预测。

$$\sigma(z) = y = \frac{1}{1+e^{-z}} \tag{1}$$

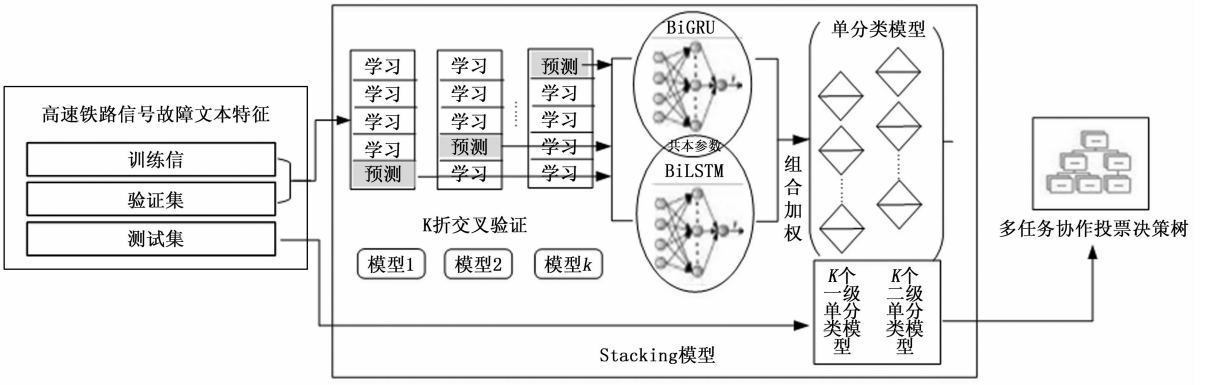

图 1　Stacking 信号设备故障文本分类模型

$$\tanh(z) = y = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \tag{2}$$

$$s_t = f(U * x_t + W * s_{t-1}) \tag{3}$$

$$o_t = softmax(Vs_t) \tag{4}$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{5}$$

RNN 的变体神经网络 LSTM 和 GRU 是通过设计门的结构来选择通过神经元的信息，由于 sigmoid 的输出是在 0 到 1 之间的取值，有助于信息的选择与忘记，0 表示全部舍弃，1 表示全部保留，通常选择 sigmoid 函数作为激活函数，tan$h$ 函数作为输出函数。

LSTM 分为三个门：输入门、遗忘门与输出门。遗忘门 f_t 决定丢弃信息，输入门 i_t 有两层组成，首先通过 sigmoid 层作为输入层，决定要更新的值，然后由 tan$h$ 层产生一个新向量 $\overline{C_t}$ 到细胞状态中，C_t 将新输入的信息代替需要忘记的信息。最后输出层 o_t 由 sigmoid 层确定细胞状态哪些需要输出去。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{6}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{7}$$

$$\overline{C_t} = \tanh(W_C . [h_{t-1}, x_t] + b_C) \tag{8}$$

$$C_t = f_t * C_{t-1} + i * \overline{C_t} \tag{9}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{10}$$

$$h_t = o_t * \tanh(C_t) \tag{11}$$

GRU 是将 LSTM 的遗忘门、输入门和输出门变为更新门 z_t 与重置门 r_t，并将单元状态与输出合并为一个状态 h_t。

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{12}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{13}$$

$$\overline{h_t} = \tanh(W \cdot [r_t * h_{t-1}, x_t]) \tag{14}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \overline{h_t} \tag{15}$$

2.2　组合加权次级学习器原理

组合加权次级学习器不仅考虑神经网络的整体学习能力，同时也考虑神经网络在不同类别上的表现。根据单个神经网络对相同输入的学习结果，给单个神经网络分配权重，准确度越高的神经网络权重越大，这种方法可以有效抑制神经网络学习过程中少数值，极端值的影响。神经网络在各类别上的权重根据公式（16）、（17）计算，通过计算分类神经网络在某个类别上的错误比例对数计算在该类别上的权重，表现好的，错误比例小的权重越大，当错误比例超过 0.5 时，权重计为 0。最后按公式（18）将神经网络的整体权重与类别权重相加，重新计算模型对标签的预测值。

$$\varepsilon_{ij} = \frac{errorNum_{ij}}{textNum_i} \tag{16}$$

$$\begin{cases} \alpha_{ij} = In(\dfrac{1 - \varepsilon_{ij}}{\varepsilon_{ij}}) & \varepsilon_{ij} < 0.5 \\ 0 & \varepsilon_{ij} \geqslant 0.5 \end{cases} \tag{17}$$

$$p_i = \sum_{j=1}^{n} (w_j + \alpha_{ij}) \cdot p_{ij} \tag{18}$$

式中，ε_{ij} 表示第 j 个神经网络在第 i 类标签的预测错误比例，α_{ij} 表示第 j 个神经网络在第 i 类标签的权重，w_j 为第 j 个神经网络整体分配的权重值，并且 $\sum_{j=1}^{n} w_j = 1$。

2.3　Stacking 单层分类模型

Stacking 模型采用两个神经网络作为初级学习器，在数据预处理层将字符特征向量进行降维，转换为神经网络嵌入层（Embedding），分别输入到两个双向神经网络 BiGRU 和 BiLSTM 中，两个神经网络经过学习后分别在 Softmax 层输出对各分类标签的预测概率，经过组合权重分类器对两个初级学习器的预测结果整合计算，最后输出 Stacking 模型对输入数据的分类情况，如图 2 所示。

3　实验结果与分析

本文以高速铁路信号基础设备中的道岔转辙设备 2009～2018 年 10 年数据进行验证，其中，70%作为训练集样本，20%作为验证集样本，10%作为测试集样本。数据包括 7 类一级分类标签，64 类二级分类标签，采用准确度（Precision）和召回率（Recall）构建 F1 值综合评价模型。

其中，Precision 计算公式为：

$$Precision = \frac{1}{|C|} \sum_{i \in c} \frac{(TP_i + TN_i) \times TP_i}{TP_i + FP_i} \tag{19}$$

Recall 计算公式为：

$$Recall = \frac{1}{|C|} \sum_{i \in c} \frac{(TP_i + TN_i) \times TP_i}{TP_i + FN_i} \tag{20}$$

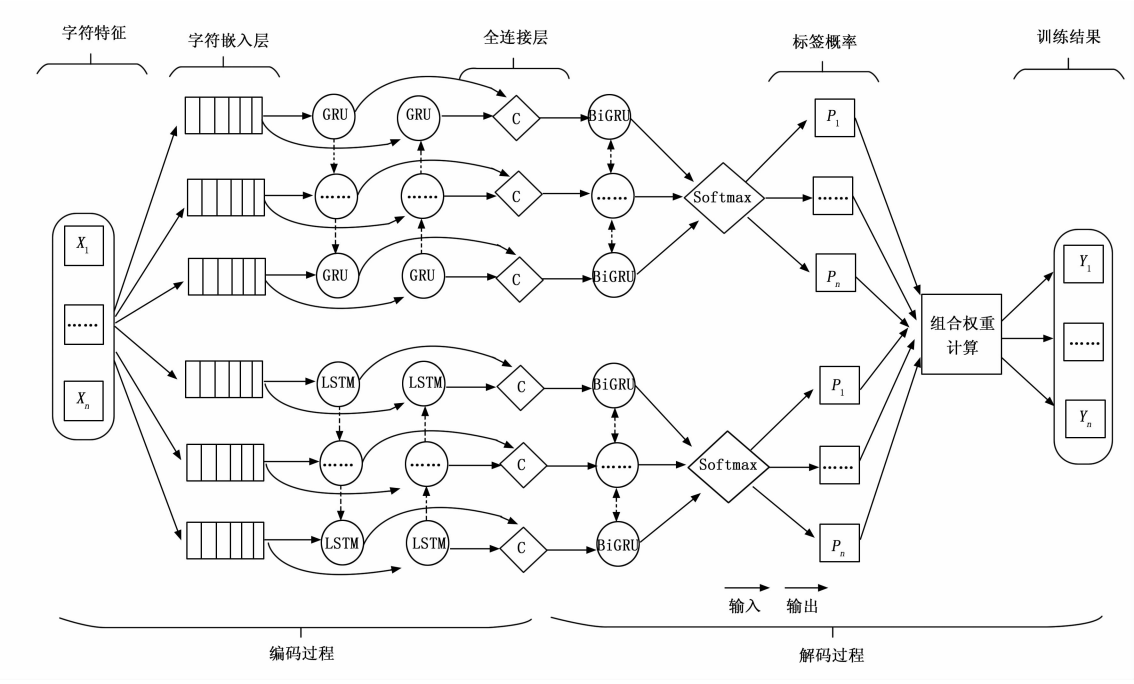

图 2 Stacking 单层分类模型网络结构

F－score 计算公式为：

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{21}$$

C 为所有样本的总数，c 为所有类别总数，TP_i 为被正确分到此类的样本个数，TN_i 为被正确识别不在此类的样本个数，FP_i 表示被误分到此类的样本个数，FN_i 表示属于此类但被误分到其它类的样本个数。

3.1 BiGRU 和 BiLSTM 整体权重分配

BiGRU 和 BiLSTM 整体权重大小根据单个神经网络对相同特征向量的学习结果。本文设计 BiGRU 和 BiLSTM 具有相同的网络参数，设定 K 折交叉验证 $K=5$，神经网络的迭代轮数为 50，网络输入批处理大小为 256，嵌入层维度为 100，隐藏层维度为 512。BiGRU 和 BiLSTM 一级和二级训练过程的 loss 函数值如图 3 所示，从图中可以看出，迭代轮数为 30～50 之间，损失函数 loss 值接近最小，并且基本稳定，BiGRU 相比于 BiLSTM 损失函数小，分类性能较优，二级分类较一级分类 loss 函数变化幅度较小，随着 K 值的增加，每一次迭代过程中，loss 值逐渐变小且趋于平稳。

经过 $K=5$ 次训练，将每次的训练结果求和平均，得到两个神经网络各自的训练结果，如表 2 所示。

表 2 BiGRU 和 BiLSTM 神经网络训练结果

方法	级别	准确率	召回率	F1 值
BiGRU	一级分类	0.8745	0.8579	0.8661
	二级分类	0.7456	0.6809	0.7117
BiLSTM	一级分类	0.8671	0.8368	0.8517
	二级分类	0.6713	0.6025	0.6350

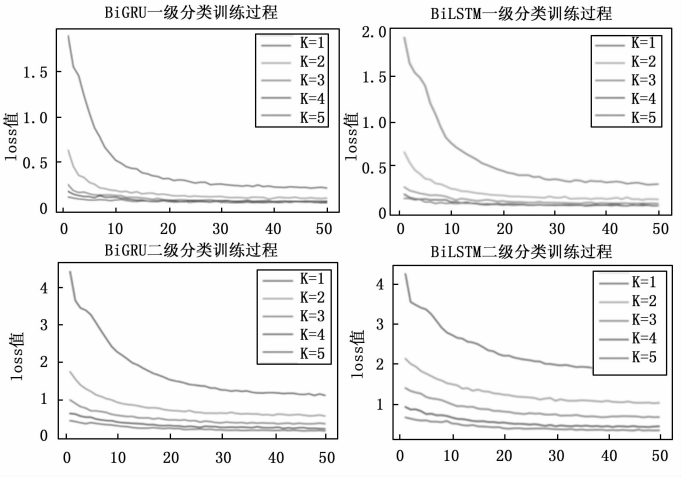

图 3 BiGRU 和 BiLSTM 网络 K 交叉 loss 函数

由表 2 可以看出，两个神经网络在相同参数下，BiGRU 较 BiLSTM 各评价指标都较高，经实验，给 BiGRU 分配权重为 0.7，BiLSTM 权重为 0.3。

3.2 BiGRU 和 BiLSTM 类别权重计算

高速铁路信号设备故障一级类别中各类别数量以及类别权重如表 3 所示，由于二级类别数量较大，考虑文章篇幅的原因，只列出一级类别的类别权重分析结果。从表 3 中可以看出密贴检查器、外锁闭及安装装置和转辙机的故障数据较少，在数据量基数小，错误数量稍大时类别权重就小，相反，在配套器材和原因不明故障数量基础较大时，网络学习效果好，类别权重也较大。

3.3 Stacking 模型分类

通过组合加权对两个网络的输出重新计算，得出共同

的分类预测结果，最终分类结果如表 4 所示，可以看出，各分类评价指标值都有所提升，实验证明，Stacking 单层分类模型是一种能够有效提升高速铁路信号设备故障文本分类指标的模型。

表 3　信号故障一级分类类别权重计算结果

类别	分类网络	分类错误数/类别总数	ε	类别权重
转辙机	BiGRU	5/28	0.1786	1.5259
	BiLSTM	8/28	0.2857	0.9164
外锁闭及安装装置	BiGRU	7/13	0.5386	0
	BiLSTM	8/13	0.6154	0
密贴检查器	BiGRU	2/3	0.6667	0
	BiLSTM	1/3	0.3333	0.6933
道岔控制电路器材	BiGRU	23/66	0.3485	0.6256
	BiLSTM	27/66	0.4091	0.3679
工务设备	BiGRU	24/100	0.2400	1.1527
	BiLSTM	31/100	0.3100	0.8001
配套器材	BiGRU	193/1038	0.1859	1.4769
	BiLSTM	216/1038	0.2081	1.3364
原因不明	BiGRU	87/436	0.1995	1.3894
	BiLSTM	93/436	0.2133	1.3051

表 4　Satcking 模型单层文本分类结果

方法	级别	准确率	召回率	F1 值
Stacking 单层分类模型	一级分类	0.8814	0.8642	0.8727
	二级分类	0.7691	0.6747	0.7188

3.4　实现总结

根据以上实验分析，各模型的分类指标按各级相应评价指标的平均值计算，BiGRU 模型、BiLSTM 模型、Stacking 模型，以及最后通过多任务协作投票的多级分类模型，各模型的分类性能如图 4 所示。

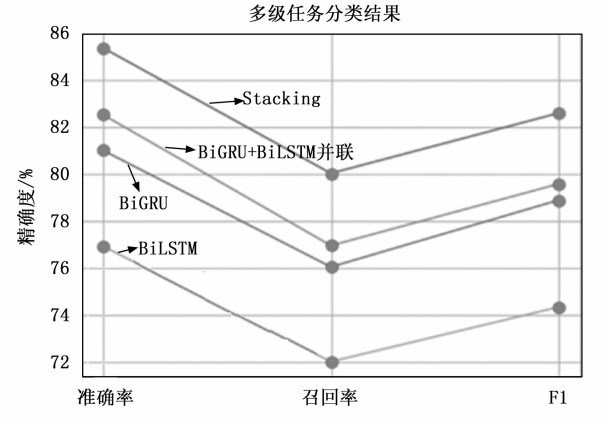

图 4　各模型分类性能对比

转辙机、外锁闭及安装装置、密贴检查器、道岔控制电路器材、工务设备等设备的一级、二级故障按相应评价指标的平均值计算精准度，BiGRU 模型为 81%、BiLSTM 模型为 75%、BiGRU+BiLSTM 并联模型为 82.3%，Stacking 模型为 85.5%，从图 4 中可以看出，针对高速铁路信号设备故障文本数据进行信号设备故障多级分类，本文设计的 Stacking 模型有效提高了各级别的分类指标，基于多任务协作投票的机制有效解决分类结果的从属关系，并提升了 Stacking 模型的整体分类性能，实验证明，本文提出的基于文本挖掘技术的 Stacking 模型在解决高速信号设备多级分类问题具有优势。

4　结束语

高速铁路设备故障文本数据是挖掘高速铁路运营安全状况与安全规律的重要数据，基于文本挖掘技术实现高速铁路设备故障多级分类是深入分析高速铁路设备故障数据的必要手段。本文就高速铁路信号设备故障文本数据设计多级分类模型，解决各级分类之间的隶属关系，并有效提升了分类评价指标。本文基于 Stacking 思想设计的 K 折交叉验证单层分类模型，保证了初级学习器的算法差异和多样性，有效降低分类过拟合的风险，并且分类指标相比单神经网络分类器有所提升，多任务协作投票机制保证了分类结果的隶属关系。本文中的 Stacking 单层分类模型和多级分类模型在铁路文本分类中都具有借鉴价值。本系统在试点工程中根据实际设备及用户的关注度需要进一步调整模型参数，使系统达到最优效果。

参考文献：

[1] 林瑜筠. 高速铁路信号技术（修订版）[M]. 北京：中国铁道出版社，2018.

[2] 何　力，贾　焰，韩伟红. 大规模层次分类问题研究及其进展[J]. 计算机学报，2012，35（10）：2101－2115.

[3] 李保利. 基于类别层次结构的多层文本分类样本扩展策略[J]. 北京大学学报（自然科学版），2015，51（2）：357－366.

[4] Wu Lei，Hoi，Steven C H，et al. Semantics－Preserving bag－of－words models and applications [J]. IEEE Transactions on Image Processing，2010，19（7）：1908－1920.

[5] Zhai Chengxiang，Lafferty J. A study of smoothing methods for language models applied to information retrieval [J]. Acm Transactions on Information Systems，2004，22（2）：179－214.

[6] Turney，Peter D，Pantel，et al. From frequency to meaning：vector space models of semantics [J]. Journal of Artificial Intelligence Research，2010，37（1）：141－188.

[7] 张笑铭，王志君，梁利平. 一种适用于卷积神经网络的 Stacking 算法 [J]. 计算机工程，2018，44（4）：243－247.

[8] Wang Chenglong，Jiang Feijun，Yang Hongxia. A Hybrid Framework for Text Modeling with Convolutional RNN [A]. 23th ACM SIGKDD International Conference [C]. Canada：ACM，2017.

[9] LEI Tao，Barzilay R，Jaakkola T. Molding CNNs for text：non－linear，non－consecutive convolutions [J]. Indiana University Mathematics Journal，2015，58（3）：1151－1186.

[10] 王　伟，孙玉霞，齐庆杰，等. 基于 BiGRU－Attention 神经网络的文本情感分类模型 [J]. 计算机应用研究，2019，27（9）：1－10.