DOI:10.20033/j.1003-7241.(2026)01-0118-05

融合 CSS-IMP 的医院麻醉数据挖掘与动态调整研究

张 奥[1], 张 文[2]

(1.湖北科技学院,湖北 咸宁 437000; 2.湖北科技学院附属二医院,湖北 咸宁 437000)

摘 要:为了提升医院麻醉数据的利用率,设计了一种融合卡方统计量和改进多层感知器的(chi square statistics and improved multilayer perceptron,CSS-IMP)麻醉数据处理模型。模型通过调取卡方统计量对输入数据进行分类,再利用基于语义插值改进的多层感知器完成数据挖掘模型的建立。结果显示,所构建的模型在2 000次迭代后对DEMSET和REPSET数据集的有效挖掘量分别达到9 805条和8 750条,在实际应用中将不同科室的关联比例提升到0.79。以上结果说明融合卡方统计量和改进多层感知器的麻醉数据处理模型具有较强的数据挖掘能力,能输出普适性高的处理数据,对医院整体麻醉数据的共享能起到促进作用。

关键词:麻醉数据;卡方统计;多层感知;数据挖掘;关联比例;共享

中图分类号: TN614 **文献标志码**: A **文章编号**: 1003-7241(2026)01-0118-05

Research on hospital anesthesia data mining and dynamic adjustment integratingchi CSS-IMP

ZHANG Ao[1], ZHANG Wen[2]

(1. Hubei Institute of Science and Technology, Xianning, 437000, Hubei, China;
2. The Second Affiliated Hospital of Hubei Institute of Technology Xianning 437000, Hubei, China)

Abstract: In order to improve the utilization of anesthesia data in hospitals, anesthesia data processing model integrating is designed that integrates chi square statistics and improved multilayer perceptron (CSS-IMP) by combining chi square statistics and improved multilayer perceptron. The model classifies input data by calling chi square statistics, and then uses a semantic interpolation based improved multilayer perceptron to establish a data mining model. The results show that the constructed model achieved an effective mining volume of 9 805 and 8 750 on the DEMSET and REPSET datasets, respectively, after 2 000 iterations. In practical applications, the correlation ratio between different departments is increased to 0.79. The above results indicate that the anesthesia data processing model that integrates chi square statistics and improved multilayer perceptron has strong data mining capabilities and can output highly universal processing data, which can promote the sharing of overall anesthesia data in hospitals.

Keywords: anesthesia data; chi square statistics; multi layer perception; data mining; correlation ratio; share

随着计算机技术的发展,医院管理平台所能存储的数据量愈发庞大。同时社会医疗水平逐年提升,对所积累的庞大数据进行充分挖掘利用成为医院进一步提升治疗能力的一大需求[1]。在医院的诊疗数据中,麻醉数据是一种共享性高、可用性强的数据类别,对多个科室的临床工作开展均能提供帮助。目前医院管理系统中,对麻醉数据的调取和利用通常限于人工的查找、翻阅,且不同科室之间难以流通,存在很大的不便[2]。针对这一问题,学术界也提出了许多解决办法。常见的数据处理方式包括时间序列分析法、主次因素分析法等,这些方法均对庞大的数据进行了一定程度的分类整理,但存在处理效率低下、信息调用烦琐等问题[3-4]。近年来,国内外的研究学者们也对这一问题进行了广泛的探索。凌佳君等为了提升医疗文本的处理效果,提出了一种基于词向量的医疗文本结构化处理方法。过程中对数据进行清洗、集成与转换以及规约,使得数据具有一致性,再通过n维空间词向量的转换提取文本中的关键信息。结果证明该方法有效地提升了文本的整合程度[5]。Edara D C等为了解决医疗时序数据缺失的问题,设计了一种双向多任务循环神经网络模型。模型以同一时间序列的缺失值插补为主任务,基于时序数据进行顺序、逆序学习,进而完成数据提取。实验结果表明,与传统的模型相比,该模型具有更优的学习性能[6]。从上述的国内外研究现状可以看出,研究者们围绕模型的创新性设计和数据预处理进行了深入的研究,但其底层逻

收稿日期:2024-08-17;**录用日期**:2024-09-09

基金项目:教育部产学合作协同育人项目2024年第一批次立项项目(230904094234148)

作者简介:张 奥(2001—),男,硕士在读,研究方向:临床医学、数据挖掘。

通信作者:张 文(1978—),男,硕士研究生,主治医师,研究方向:医院外科麻醉。

引用本文:张奥, 张文. 融合CCS-IMP的医院麻醉数据挖掘与动态调整研究[J]. 自动化技术与应用, 2026,45(1):118-122. (ZHANG Ao, ZHANG Wen. Research on hospital anesthesia data mining and dynamic adjustment integratingchi CSS-IMP[J]. Techniques of Automation and Applications, 2026, 45(1):118-122.)

辑限制于数据或者模型的单方面改进,缺少更多的融合探究。为了提升医院麻醉数据的挖掘效率和动态调整能力,研究提出一种融合卡方统计量和改进多层感知器的麻醉数据处理模型。模型通过对所输入数据进行卡方统计,获取到清晰的特征分类,再通过词义参数的插入,对多层感知器的输入层进行改造,实现更广泛的词义匹配,从而完成录入数据的挖掘提取,并保证其可以随输入量的变化而进行动态调整。模型在构建过程中创新性地用到了语义插值的方式对感知网络进行改造,并将其与特征数据进行融合。期待通过研究所提出的方法,为医院麻醉数据的共享提供一种高效的调用工具。

1 数据处理模型建立

1.1 数据特征提取方法

对医院麻醉数据进行挖掘和利用的前提是获取到数据中的特征信息,利用高效的特征提取方法对数据进行处理是提升信息准确性的关键所在。医院后台的麻醉数据主要涉及文本数据,卡方统计是一种高效的文本数据提取方法[7]。卡方统计可以对未整理的数据进行评估验证,并通过合理预测来构建变量的关联,其基本计算公式为

$$\mathrm{CHI}=\sum_{j=1}^{s}\sum_{i=1}^{r}\frac{(n_{ij}-N\hat{p}_{ij})^2}{N\hat{p}_{ij}} \tag{1}$$

式中,CHI 代表卡方统计量;N 为样本总数;n_{ij} 为读取的样本数量;p_{ij} 为预测风险值;r 和 s 代表参考量的上限。假设一个指定词语的数据量为 w_i,其所属类别包含的数据总量为 C_j,则可以对全部数据中关于该词语的数据量进行卡方统计为

$$\begin{cases}\mathrm{CHI}(w_i,C_j)=\dfrac{N\times(A\times D-B\times C)^2}{(A+B)(A+C)(D+B)(D+C)}\\ \mathrm{CHI}(w_i)=\max_j\{\mathrm{CHI}(w_i,C_j)\}\end{cases} \tag{2}$$

式中,A 为全部数据中包含 w_i 的类别数量;B 为其余类别数量;C 为不属于 C_j 的数据范围内包含 w_i 的类别数量;D 为不属于 C_j 的数据范围内的其余类别数量。通过式(2)的计算,可以得到一个特定词汇的次数概率,由此体现该词汇与文本的相关性。卡方统计可以在杂乱的数据之间寻找到相关性联系,以构建一条清晰的信息传递路径,其基本逻辑可以由图 1 表达。

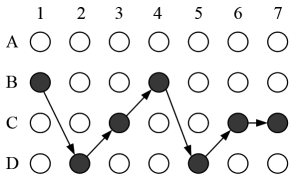


图 1 卡方统计的基本逻辑

Fig. 1 Basic logic of chi square statistics

如图 1 所示,原点代表所处理数据的特征标记,其相互之间的间隔可视为关联概率。卡方统计通过预测连接,实现统计方向的变迁,从而将相关性最强的数据相互牵引,完成数据内部关联路径的规划[8]。衡量不同数据之间的相关度大小需要用到交互率的概念,其表达式为

$$\mathrm{MI}(x_i,y_i)=\log\left(\frac{P(x_i\mid y_i)}{P(x_i)}\right) \tag{3}$$

式中,MI 表示指定词汇之间的交互率;P 为指定词汇的出现概率;(x_i,y_i) 代表词汇矢量坐标值。词语与类别之间的关系评估同样可以用到交互率,其计算公式为

$$\begin{aligned}\mathrm{MI}(w_i,C_j)&=\log\left(\frac{P(w_i\mid C_j)}{P(w_i)}\right)\\&\approx\log\left(\frac{A\times N}{(A+B)\times(A+C)}\right)\end{aligned} \tag{4}$$

式中可以看出,当一个词汇在一个类别里面出现概率较高时,有可能导致其交互率被忽略,由此需要引入信息熵的概念,以对高频词汇的特征信息进行权重评估[9]。信息熵可以赋予词汇在类别中的重要程度估值,其定义为

$$\mathrm{Entropy}(X)=-\sum_{i=1}^{n}p(x_i)\log(p(x_i)) \tag{5}$$

式中,Entropy 代表指定词汇的信息熵;p 代表词汇出现的频率。一个类别的数据不规则程度越大,则指定词汇在其中的重要程度受出现频率的影响越大。当该词汇出现的次数较多时,说明其对类别的规整性影响较大,从而可以对其赋予更多的重要性标记,在随后的搜索和提取中参与更多的特征表达[10]。由以上步骤所确立的基于卡方统计的数据特征提取规则,可以对庞杂且无规律的医院麻醉数据进行整理归纳,获取到显著的特征分类特性,以便对其进行建模分析处理。

1.2 数据挖掘建模

在完成数据的有效提取整理之后,需要一种合适的感知模型来对分类后的数据进行挖掘分析。在医疗数据的分析中,常用到一种多层感知结构,其可以对分类特征的误差值、聚合度等指标进行修正,再根据目标指令的序列关系对数据特征进行边缘划线,以实现模糊搜索和广义匹配的功能[11]。多层感知器的运用首先涉及特征的矢量化调整,将词汇在类别中的索引范围转变为矢量长度,其过程可由式(6)表示为

$$\mathrm{Vector}_i=(0,0,\cdots,\mathrm{Val}_i,\cdots,0) \tag{6}$$

式中,Vector 代表矢量长度;Vector 为单一索引时间内的特征容量。式(6)所表达的矢量关系在于,单一索引所呈现的特征一旦确定,则将其余索引的特征量降为 0,由此实现词汇数据在序列坐标上的凸显。由此,一个词汇的序列定义可由式(7)表示为

$$\mathrm{VT}=\sum_{i=0}^{n}\mathrm{Vector}_i \tag{7}$$

式中,VT 代表词汇在检索中所用的序列值。多层感知器对矢量数据的调取主要依靠一种隐藏交互网络,其基本结构如图 2 所示。

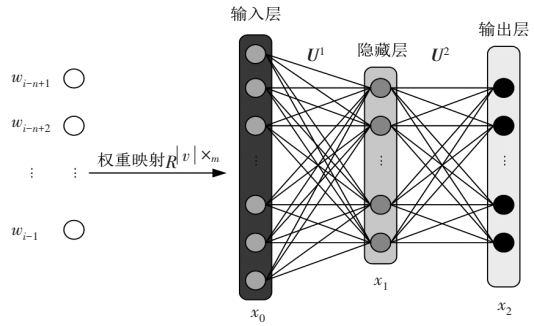
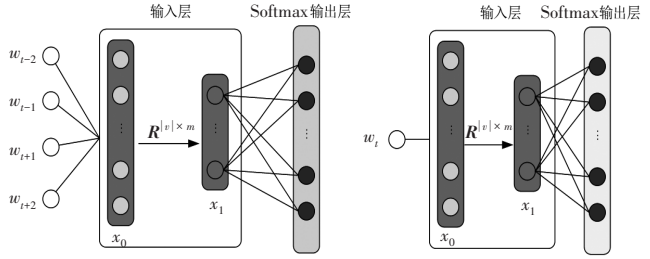


图 2 多层感知器的基本结构

Fig. 2 Basic structure of multilayer perceptron

图 2 中可以看到,在数据输入之前会进行权重映射,通过概率矩阵 $\boldsymbol{R}$ 的设计,将数据调整为同一方向上的矢量参数。输入层与隐藏层、隐藏层与输出层之间均会进行矩阵堆叠的处理,其过程依靠两个理论矩阵 $\boldsymbol{U}^1$ 和 $\boldsymbol{U}^2$,将矢量化的数据与目标指令进行迭代交互,以达到学习分析的效果。研究在此引用到一种词组拓展工具,通过语义插值的方式对输入层进行改造。由于权重映射后的输入数据常常存在矢量模糊的问题,向其中插入部分近义值可以起到很好的填充作用[12]。拓展改造主要针对输入层,其基本逻辑可以由图 3 表示。



(a) 近义矢量输入 (b) 目标矢量输入

图 3 输入层改造过程

Fig. 3 Input layer transformation process

如图 3 所示,输入层改造分为两个步骤,首先向输入层插入多个与选定矢量相关的近义矢量,在输入层内部进行二次权重投射,并得到简化的结果。此步骤完成之后,再将需要迭代的目标词汇的矢量数据进行正式输入,同样在输入层内部重复权重投射的步骤。经过改进结构的处理,词汇量的矢量特征会得到充分地放大,在后续的学习交互中会带来更快的计算速度和更精确的输出结果。此过程中还需对精简程度进行约束,约束计算公式为

$$x_0 = \frac{1}{2C_{k=i-C,k\neq i}} \sum^{k=i+C} \mathrm{VT}(w_k) \tag{8}$$

式中,w 为词汇的矢量标记值;C 为精简程度的参数;x_0 为精简尺度。确定精简尺度之后即可确定语义插入的范围,从而避免过于宽泛的词组拓展。模型性能的优化还涉及数据损失值的缩减,缩减方式为

$$\mathrm{Loss} = -\sum_{w_i \in V} \mathrm{VT}(w_i)\log(P(w_i \mid w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2})) \tag{9}$$

如式(9)所示,各个词矢量将分别进行损失值的计算,通过序列值的乘积可以进行总损失值的统计。模型在迭代过程中会根据总损失值的大小不断调整训练的次数和方向,以将损失值降到最低[13]。通过以上的构建和改造,基于改进多层感知器的医院麻醉数据挖掘模型得以建立。模型可以对前期预处理后的规整数据进行进一步分析,实现指令目标的匹配和相关性的动态调整,帮助医院工作人员精准地查询调取、观察分析所需数据,并在产生新录入数据时实现整体系统的动态调整。

2 性能测试与应用效果

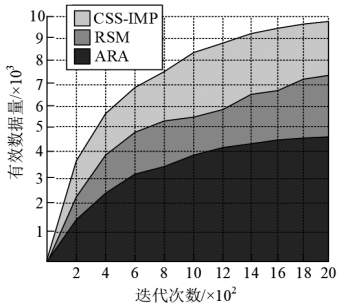
为了验证 CSS-IMP 模型在实际使用中的有效性和先进性,研究选取了基于粗糙集的医疗数据挖掘模型(medical data mining model based on rough set, RSM)、基于关联规则算法的医疗数据挖掘模型(medical data mining model based on association rule algorithm, ARA)与所构建的模型展开对比[14-15]。实验所用到的相关技术栈版本信息见表 1。

表 1 技术栈版本信息

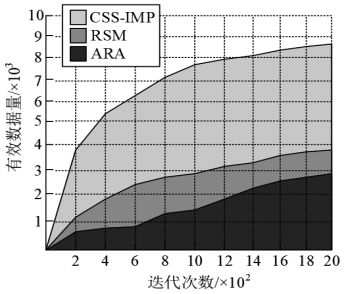
Tab. 1 Technical stack version information

名称	版本号	名称	版本号
FLUT	2.2.1	LANGE	5.6.0
TYMANTE	2.4.1	KATEL	4.67.5
NANDAE	6.1.4	ANTEL	4.5.1
SIRADO-FIL	1.0.4	BEWOG	6.3.2
RUMP	1.3.3	OWENY	7.3.4
WODA	6.22.4	AGONT	1.0.1

测试过程首先对各模型的有效数据挖掘量进行了分析。分析采用了 DEMSET 和 REPSET 两个数据集来进行对照,以观察各模型处理不同数据量的能力,结果见图 4。



(a) DEMSET



(b) REPSET

图 4 有效数据挖掘量结果对比

Fig. 4 Comparison of effective data mining results

图 4(a)为三种模型在 DEMSET 数据集上的有效数据挖掘量,可以看到,CSS-IMP 模型的有效数据挖掘量最大。到第 2 000 次迭代时,其累计所挖掘的有效数据量已达到了 9 805 条。RSM 模型和 ARA 模型在 2 000 次迭代后的有效数据量分别为 7 496 条和 4 659 条。图 4(b)为三种模型在 REPSET 数据集上的有效数据挖掘量,由于 REPSET 数据集具有更加混乱的数据特征,各模型的有效数据挖掘量均有所下降。但 CSS-IMP 模型依然保持了较高的挖掘有效性,下降幅度较另外两种模型更小。CSS-IMP 模型、RSM 模型、ARA 模型在 REPSET 数据集上经过 2 000 次迭代后,累计有效数据挖掘量分别为 8 750 条、3 896 条、2 960 条。CSS-IMP 模型由于经过了输入层的改造,获得了更强的数据提取能力,能够对更广范围内的有效数据进行精准提取。接下来,研究调取了某市三家医院患者麻醉数据进行实际应用测试,相关数据所涉科室信息见表 2。

表 2　麻醉数据来源信息

Tab. 2　Information source of anesthesia data

医院名称	科室	数据条数
某省人民医院	神经内科	55
	泌尿外科	106
	肿瘤内科	98
	普通外科	35
	肾内科	46
	耳鼻咽喉科	78
某大学附属医院	麻醉科	205
	消化内科	332
	普通外科	104
某红十字医院	心血管内科	52
	急救中心	19
	妇科	24
	普通外科	33
	眼科	85
	烧伤整形科	67
	皮肤科	44
	颌面外科	95
	骨科	35

研究主要对不同模型数据挖掘分析后在科室之间建立关联的程度进行了比对,科室关联的紧密程度可以直接反映模型所输出数据的宽泛性和精确性。紧密度的评价指标由各科室医生调取其他科室数据的次数占自身科室数据次数的比例来确定,比例越大代表紧密性越高,对比结果见图 5。

图 5 展示了三种模型对所输入麻醉数据进行挖掘后呈现的科室紧密度结果,可以看出,CSS-IMP 模型得到了最高的关联比例。RSM 模型在迭代次数较少时所能输出的关联比例较低,在 1 000 次迭代后有显著的攀升。在经过 6 000 次迭代后,CSS-IMP 模型、RSM 模型、ARA 模型输出的关联比例分别为 0.79、0.65、0.67。CSS-IMP 模型无论是计算速度或计算精度,均优于其他两种模型。对麻醉数据进行挖掘分析的目的在于实现相关数据的共享和连接,方便不同医院、不同科室之间进行数据调取,同时对新产生的数据进行动态调整。CSS-IMP 模型通过对卡方统计量对数据特征进行了清晰的分类,再利用改进的卷积网络结构对输入信息进行了精准匹配,显著提高了数据挖掘的清晰度和适用性,从而提升了医生调取数据的频率。

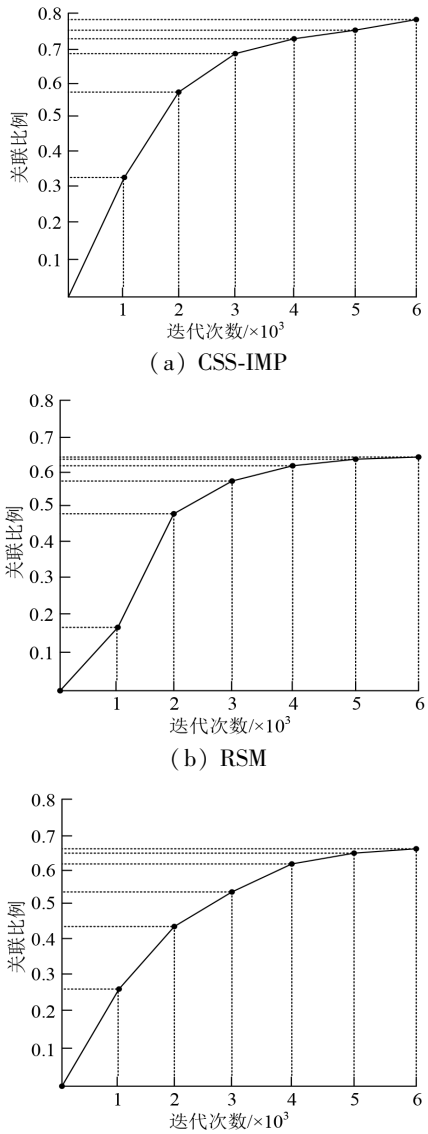


图 5　科室联系紧密度对比

Fig. 5　Comparison of departmental connectivity

3　结论

医院麻醉数据的共享对于提升整体医疗水平有很大的帮助,随着医疗技术的不断发展,后台存储的麻醉数据愈发庞大,对其进行有效化挖掘处理具有重要意义。为了提升医院麻醉数据的利用率,研究提出一种融合卡方统计量和改进多层感知器的麻醉数据处理模型。模型通过对卡方统计量的计算,完成后台数据的归纳整理及特征分类,随后通过语义插值的方式对卷积网络进行改进,使之能够完成更复杂的数据挖掘任务,最终实现医院麻醉数据的精准调用和动态调整。模型在搭建的过程中创新性地

用到了向卷积网络输入层填充近义参数的改进方式,起到了很好的词汇拓展效果,显著提升了数据匹配的效率。对所构建模型进行的性能测试结果显示,在 DEMSET 数据集上,CSS-IMP 模型的有效数据挖掘量可达到 9 805 条;在 REPSET 数据集上的有效数据挖掘量可达到 8 750 条,证明了模型极强的数据挖掘性能。在实际应用中,CSS-IMP 模型所输出的数据对三所医院共计 18 个科室间的关联比例可达到 0.79,表明模型具有很好的适用性。以上结果说明,CSS-IMP 模型具有优秀的数据挖掘分析及动态调整能力,能够很好地应用于实际的医院麻醉数据处理工作中,为医生的工作提供帮助。由于实验条件有限,所构建的模型尚未应用于更多的医院麻醉数据的处理中,期待未来在更大的数据集中进一步验证模型的使用性能。

参考文献

[1]尚龙飞, 王华杰, 徐露. 基于条件代理重加密的区块链医疗数据共享模型[J]. 现代电子技术, 2024, 47(1):78-83.

[2]曾梦, 邹北骥, 张文生, 等. 多模态医疗数据中海量小文件存储优化方法[J]. 软件学报, 2023, 34(3):1451-1469.

[3]苏强, 季荔. 基于随机演化博弈的医疗数据共享协调机制研究[J]. 情报科学, 2023, 41(9):37-47.

[4]杨健, 王开选. 区块链架构下医疗数据共享的三方演化博弈研究[J]. 计算机科学, 2023, 50(S01):545-551.

[5]凌佳君, 刘宇, 顾进广. 面向互联网问诊文本的医疗事件时序关系抽取[J]. 计算机应用与软件, 2023, 40(11):186-193.

[6]EDARA D C, VANUKURI L P, SISTLA V, et al. Sentiment analysis and text categ-orization of cancer medical records with LSTM[J]. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(5): 5309-5325.

[7]卢建璋. 大数据时代医学数据挖掘分析平台构建[J]. 情报科学, 2023, 41(8): 89-94.

[8]雷松泽, 刘博, 王瑜菲, 等. 结合多特征嵌入和多网络融合的中文医疗命名实体识别[J]. 电子与信息学报, 2023, 45(8): 3032-3039.

[9]喻芳宇, 高胜哲. 融合随机统计规律与优化思想的成分数据预测方法研究[J]. 现代电子技术, 2024, 47(2):171-175.

[10]吴欢欢, 谢瑞麟, 乔塬心, 等. 基于可解释性分析的深度神经网络优化方法[J]. 计算机研究与发展, 2024, 61(1):209-220.

[11]张蕾, 徐叶青. 基于改进 K-Medoids 聚类算法的医院 HRP 系统设计[J]. 自动化技术与应用, 2025, 44(12):142-146,188.

[12]李腾, 方保坤, 马卓, 等. 基于同态加密的医疗数据密文异常检测方法[J]. 中国科学:信息科学, 2023, 53(7):1368-1391.

[13]陈玥丹, 肖国庆, 阳王东, 等. 基于异构系统的多级并行稀疏张量向量乘算法[J]. 计算机学报, 2024, 47(2):441-455.

[14]马文胜, 侯锡林, 王宏波, 等. 基于粒度树和使用关系的大数据价值计算研究[J]. 计算机科学, 2023, 50(S02):658-665.

[15]王婷, 王娜, 崔运鹏, 等. 基于深度学习的医疗电子数据特征学习方法[J]. 应用科学学报, 2023, 41(1):41-54.

(上接第 48 页)

均值,作为模型的最终性能评估指标,采用正则匹配和无 copy 机制作为对比实验,来展示算法模型的性能。准确率对比如表 1 所示。

表 1 准确率对比

Tab. 1 Accuracy comparison

算法类型	正则匹配	无 copy 机制	有 copy 机制
时间	0.72	0.96	0.98
设备	0.68	0.92	0.96
动作	0.63	0.92	0.95

由表 1 可知,基于深度学习的模型解析算法准确率远超正则匹配算法,无 copy 机制的深度学习算法提取实体的平均准确率 93.33%,远高于正则匹配算法的 67.67%。加入 copy 机制准确率 96.33%,比没有 copy 机制的准确率提升了 3%。

基于构建好的工业设备调令管理系统,从调令的输入、摘录入库、解析下发到指定工厂的终端,测试数据集在全链路的系统响应时间。时间对比如表 2 所示。

表 2 系统响应时间对比

Tab. 2 Comparison of system response time

算法类型	正则匹配	无 copy 机制	有 copy 机制
时间/s	2.522	2.872	2.919

由表 2 可知,算法的不同带来的系统响应时间差异较小,深度学习算法比正则匹配算法消耗时间多 0.3 s 左右,其他主要时间是消耗在数据流转的传输过程中。

4 结论

针对工业设备调令管理存在的成本较高和效率低下的问题,本文构建了基于深度学习的工业设备调令管理系统。通过设计文本语义解码盒子,配合算法的优化,采取具有 copy 机制的深度模型,提高了系统解析文本的准确率,且全链路的系统响应时间较短。构建的系统能有效完成调令的解析、存储和下发,能够为企业降低成本和提高效率。在后续研究中,可以通过融入多模态解析的模块,使系统的功能更加全面,适用性更加广泛。

参考文献

[1]汪海. 工业设备维护管理系统研究与设计[J]. 河南科技, 2023, 42(8):22-26.

[2]田野, 陈大卫, 付安民. 一种基于加密二维码的工业设备管理系统[J]. 工业信息安全, 2022(2):39-50.

[3]刘志华. 工业设备信息和数据管理系统[J]. 机械工程与自动化, 2022(4):168-170, 173.

[4]康华夏, 周正宇, 刘文军, 等. 工业设备运维云平台的设计与实现[J]. 物联网技术, 2022, 12(9):108-110, 113.

[5]袁枫, 戴琳琳, 景辉, 等. 基于生成式摘要模型和知识蒸馏算法的铁路调度命令解析算法研究[J]. 铁路计算机应用, 2023, 32(3): 11-16.

[6]张锡然, 张保林, 苏适, 等. 调度指令自动解析与防误系统衔接示范技术的应用与研究[J]. 云南电力技术, 2023, 51(4):25-31.

[7]周燕. 基于 GloVe 模型和注意力机制 Bi-LSTM 的文本分类方法[J]. 电子测量技术, 2022, 45(7):42-47.

[8]焦凯楠, 李欣, 朱容辰. 中文领域命名实体识别综述[J]. 计算机工程与应用, 2021, 57(16):1-15.