面向公众健康问句分类数据挖掘算法评测研究*

徐晓巍 郭海红 李 姣

（中国医学科学院/北京协和医学院医学信息研究所 北京 100020）

〔摘要〕 详细阐述公众健康问句分类数据挖掘算法评测任务开展及参赛情况，包括任务设计、评测数据采集与标注、评测指标设定、评测结果总结，为相关研究提供参考。此次评测所用语料库、评测指标、自动评测脚本工具都将公开，用于科学研究。

〔关键词〕 公众健康；问句分类；算法评测

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2021.10.004

Study on the Evaluation of Data Mining Algorithms for Classification of Public Health Questions *XU Xiaowei, GUO Haihong, LI Jiao, Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China*

〔**Abstract**〕 The paper elaborates the evaluation task of data mining algorithms for classification of public health questions and the competition situation, including task design, evaluation data collection and annotation, evaluation index setting and summary of evaluation results, and provides references for related study. The corpus, evaluation index and automatic evaluation script tools used in this evaluation task will be made public for scientific research.

〔**Keywords**〕 public health; question classification; algorithm evaluation

1 引言

1.1 国内外生物医学文本评测任务发展情况

基于真实文本数据进行评测任务设计可以有效提升自然语言处理（Natural Language Processing，NLP）技术在医疗健康领域的应用水平。随着医疗健康领域数据越来越多以自由文本形式存在，如临床电子病历、影像报告、健康网站问答数据等，如何有效应用并发挥数据价值成为当前研究热点[1-2]。将 NLP 技术应用于医疗健康类文本研究可以有效提升研究效率。2002 年国际知识发现和数据挖掘竞赛（KDD CUP）首次设置针对生物医学文本的评测任务[3]，有效推动了医疗健康与 NLP 技术结合。国内外科研组织纷纷针对不同类型健康医疗数据设计合适的评测任务，构建共享评测数据集，吸引研究人员探索不同机器学习算法在特定任务中的应用方法。从 NLP 技术分类角度看，英文评测任务命题已

〔修回日期〕 2021－04－25

〔作者简介〕 徐晓巍，硕士，助理研究员，发表论文 4 篇；通讯作者：李姣，博士，研究员，硕士生导师。

〔基金项目〕 中国医学科学院创新工程“医学人工智能算法评价标准库构建”（项目编号：2018－I2M－AI－016）；中国医学科学院基本科研项目“医学人工智能技术与人机交互关键问题研究”（项目编号：2018PT33024）。

经涵盖 NLP 技术多个方面，包括信息检索、命名实体识别、信息抽取到自动问答与摘要生成[4]。上述评测任务的开展丰富了医疗健康领域开放数据，通过汇聚不同领域人才拓宽了特定问题解决思路。随着近年来我国医疗信息化技术发展，大量医疗健康数据得到积累。为有效利用数据、探索中文医疗健康领域的 NLP 技术应用、激发技术人才潜力，国内科研机构组织多种形式的评测比赛。如全国知识图谱与语义计算大会（China Conference on Knowledge Graph and Semantic Computing，CCKS）与中国健康信息处理会议（China Health Information Processing Conference，CHIP）多次发布医疗健康领域相关评测任务，包括面向电子病历等医学文本的命名实体识别[5]、患者健康咨询问句匹配[6]等。上海瑞金医院于 2018 年与阿里云联合启动基于糖尿病文献的实体识别和关系抽取评测任务[7]，希望参赛选手通过设计高效、高准确率的算法构建全面的糖尿病知识图谱以助力糖尿病医学科研。综上可知，当前中文医疗健康领域自然语言类评测主要集中于面向临床文本的命名实体识别、关系抽取方面，评测任务类型有限。

1.2 医学数据挖掘算法评测大赛背景

自动问答系统作为一种智能化的信息服务方式，因其易用性与便捷性被广泛应用于医疗健康问答领域。据《第 46 次中国互联网络发展状况统计报告》[8]统计，截至 2020 年 6 月我国有 17.9% 的网民曾使用网上挂号、问诊等在线医疗服务。作为自动问答系统核心模块之一，问句分类任务准确性直接影响系统性能。研究表明在进行线上健康信息检索时问句类别的确定可以大幅提高检索结果权威性与准确性[9]。基于此，中华医学会医学信息学分会于 2020 年举办以公众健康问句分类任务为主题的医学数据挖掘算法评测大赛。该项研究从多个公众健康类网站收集健康相关问句，基于前期构建的标注体系对所收集健康相关问句进行人工标注，构建中文健康问句语料库，随机切分为训练集与测试集以确定评价指标。本文针对此次评测任务的开展情况进行阐述。

2 任务描述

2.1 公众健康问句分类

给定公众健康相关中文问句并对问句主题进行分类。基于前期研发的公众健康问句分类体系[10]，本次评测任务所用问句主题定义为 6 类：A 诊断、B 治疗、C 解剖学/生理学、D 流行病学、E 健康生活方式、F 择医。由于 1 个中文健康问句往往归属于多个主题类别，因此该自动分类任务是 1 个多标签分类问题。形式化定义如下：

χ 代表由健康问句组成的样本空间；$\iota = \{\lambda_1, \lambda_2, \cdots\cdots, \lambda_n\}$ 代表类别标签集合；$y = \{y_1, y_2, y_3, \cdots\cdots, y_n\}$ 表示标签集合 ι 的向量，其中 $y_i = 1 \Leftrightarrow \lambda_i \in \iota$，$y_i \in \{0, 1\}$

本次评测任务需构建 1 个多标签分类器 ξ 实现 $\chi \rightarrow y$ 的映射。

2.2 评测任务流程

评测任务组织者确定评测任务所用数据来源及测试集和训练集数据量，开发算法评测开放平台用于评测任务发布、数据发布、选手报名、结果提交和实时反馈等；专业标注人员对评测数据进行标注；参评团队通过大赛发布平台报名参赛，获取数据、构建模型、上传结果，根据分数调整模型。具体流程，见图 1。

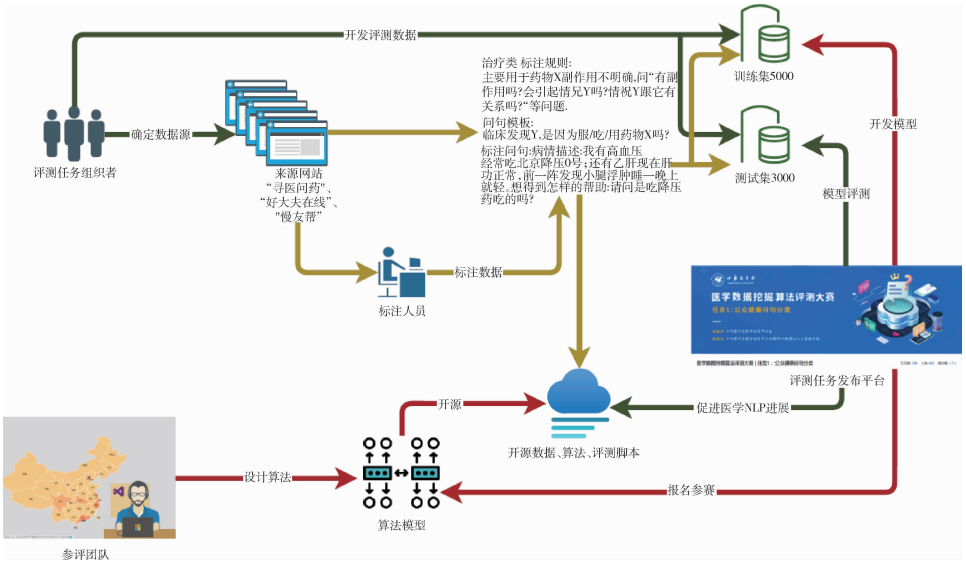

图 1 评测任务组织流程

3 评测数据

3.1 数据组成

本次评测所用数据由 3 部分组成。数据集 1：共 2 000 条与高血压相关问句，来自寻医问药网；数据集 2：共 3 000 条问句，分别来自好大夫在线、寻医问药、慢友帮，覆盖 6 类疾病，涉及内科、外科、妇产科、儿科、感染科和中医科；数据集 3：共 3 000 条问句，网站来源与疾病覆盖情况同数据集 2。数据集 1 和数据集 2 共 5 000 条问句作为本次算法评测大赛的训练集，数据集 3 作为测试集。

3.2 分类体系

根据前期研究设计的分类体系[11]对本次评测数据使用 one - hot 形式进行标注[12]，0 表示不属于该类别，1 表示属于该类别。例如用户咨询“总胆固醇 7.38，属于什么程度？”为诊断问题，属于诊断类（类别 A）标注为 1；用户同时咨询“在服瑞舒伐他汀，隔日一片，行吗？”为药物选择和如何用药问题，属于治疗类（类别 B）标注为 1。未涉及其他主题类别的问题均标注为 0。

3.3 评测数据标注流程

共设有 6 名标注人员。首先，根据标注人员数量将数据分成 3 份，每份由两人独立标注并计算标注一致率，即标注结果相同的问句数量除以标注问句总量；其次，收集和讨论在标注过程中存在的问题；再次，将打乱数据分配顺序，由两人再次独立标注并计算标注一致率，再计算该批数据平均标注一致率作为最终标注一致率。数据集 1 的平均标注一致率为 0.885，数据集 2 的平均标注一致率为 0.92，数据集 3 的平均标注一致率为 0.955 8，标注一致性较高。最后对标注不一致的问句通过讨论达成一致。最终数据集与测试集问句主题人工标注情况，见图 2。

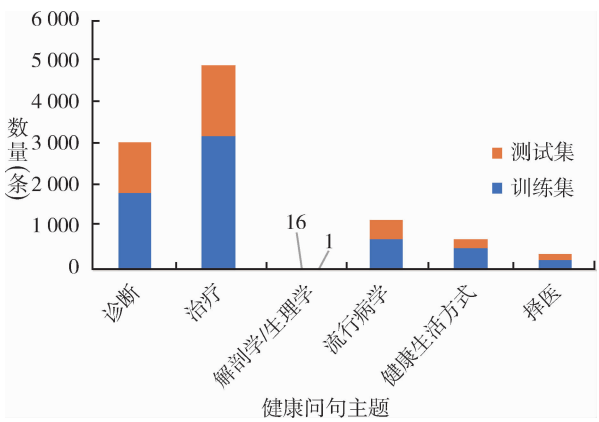

图 2 数据集分布情况

4 评测指标

本评测任务使用宏平均 $F1$ 值（macro – $F1$ score）对参赛选手构造的公众健康问句分类模型生成的分类结果进行评测。各项指标计算方法如下：

$$\text{Preision} = \frac{\text{预测为类目 X 且实际为类别 X 的问题数量}}{\text{预测为类别 X 的问题数量}}$$

$$\text{Recall} = \frac{\text{预测为类目 X 且实际为类别 X 的问题数量}}{\text{实际为类别 X 的问题数量}}$$

$$F1 - \text{score} = \frac{2}{\frac{1}{\text{preision}} + \frac{2}{\text{preision}}} = \frac{2 * (\text{preision} * \text{recall})}{\text{preision} + \text{recall}}$$

$$macre\_\ F1\ score = \frac{\sum_{p=1}^{|C|} F1(C_p)}{|C|}$$

选手通过线上平台提交算法在测试集上的运行结果，使用自动化脚本对结果进行评测，生成相应 macro – $F1$ 值。

5 评测结果

5.1 概述

本次线上评测自 2020 年 9 月 21 日开启线上报名，共有 488 人组成 396 支队伍参加。至 2020 年 9 月 30 日线上评测截止，共有 149 支队伍完成 1 300 次提交，参与线上评测。大多数队伍的宏平均 $F1$ 值在 0.46 ~ 0.65 区间，成绩最高团队的宏平均 $F1$ 值为 0.755。所有参赛队伍线上评测结果的平均精确率、召回率、宏平均 $F1$ 值分别为 0.562、0.585、0.546。最高精确率为 0.810，最高召回率为 0.872，见图 3。

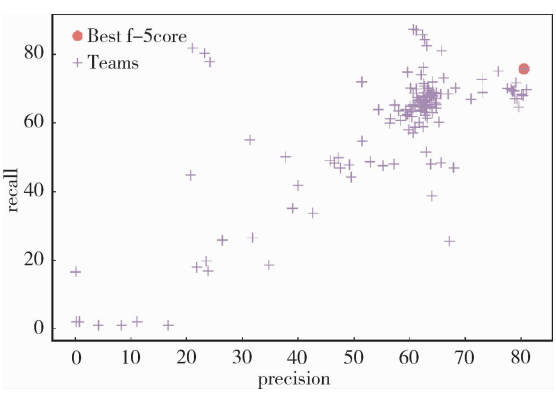

图 3 所有队伍提交成绩分布

5.2 典型方法及模型总结

5.2.1 预训练语言模型 宏平均 $F1$ 值排名前 10 位的参赛队伍均使用预训练语言模型，其中 8 支队伍使用哈尔滨工业大学发布的 ROBERTa Large 模型，1 支队伍使用 ROBERT 模型，1 支队伍使用百度 ERNIE 模型。

5.2.2 数据增强 本次评测任务发布数据量有限，用于深度模型训练获得效果有限，因此参赛选手大多进行数据增强操作，包括对医学实体的同义词替换，随机插入、交换、删除，使用翻译软件进行回译等。

5.2.3 数据不均衡的处理 训练集各类别中正负例比例差异较大，尤其是 C 类，在训练集中只有 1 个正例。在使用宏平均 $F1$ 值对模型结果进行评估时，单个类别的宏平均 $F1$ 值对总体结果影响较大，因此各参赛队对 C 类均进行单独处理，包括采用基于关键词的标签筛选方法对给定问句做 C 类标签的预测、对 C 类问题构建规则模板、构建基于规则与解剖学名词词典等。

5.3 最终评测结果

最终评测结果分布情况（排名前 10 位），见图 4。右侧是各队伍最终宏平均 $F1$ 值分数的倒序排列，左侧是不同队伍使用的预训练模型及数据增强策略热图，热图颜色越偏蓝色则表明该队伍使用的数据增强策略越多。

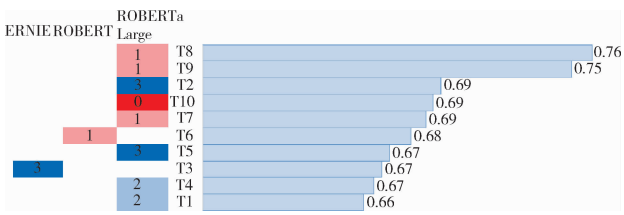

图 4 评测结果分布情况（排名前 10 位）

6 结语

本文对公众健康问句分类评测任务开展及参赛情况进行阐述，包括任务设计、语料库准备、评测性能指标设定以及参赛团队评测结果总结。从参赛结果看，即使是成绩最好的团队的宏平均 $F1$ 值也

仅达到 0.755，主要原因之一在于数据集数量较小且各类别分布不均，各参赛队模型创新不足。NLP 技术在医疗健康问句分类应用上尚有较大提升空间。通过这次评测推动该领域研究进展和技术实用化，促进不同学科背景研究人员的交流与合作。此外本次评测所用的语料库将开放共享，相关语料及评测脚本可通过网址获取[12]。

参考文献

1 Koleck T A, Dreisbach C, Bourne P E, et al. Natural Language Processing of Symptoms Documented in Free - text Narratives of Electronic Health Records: a systematic review [J]. J Am Med Inform Assoc, 2019, 26 (4): 364 - 379.

2 Iroju O G , Olaleke J O . A Systematic Review of Natural Language Processing in Healthcare [J]. International Journal of Information Technology and Computer Science, 2015, 8 (8): 44 - 50.

3 Yeh A S, Hirschman L, Morgan A A. Evaluation of Text Data Mining for Database Curation: lessons learned from the KDD Challenge Cup [J]. Bioinformatics, 2003, 19 (Suppl 1): i331 - i339.

4 Huang C C, Lu Z. Community Challenges in Biomedical Text Mining over 10 Years: success, failure and the future [J]. Brief Bioinform, 2016 , 17 (1): 132 - 144.

5 Zhang J, Li J, Wen Z, et al. Overview of CCKS 2018 Task 1: named entity recognition in Chinese electronic medical records [C]. Hangzhou: Proceedings of China Conference on Knowledge Graph and Semantic Computing, 2019.

6 中国中文信息学会医疗健康与生物信息处理专业委员会. 第四届中国健康信息处理大会 (CHIP - 2018) [EB/OL]. [2020 - 11 - 06]. http://icrc.hitsz.edu.cn/chip2018/Task.html.

7 上海市医药卫生发展基金会. 瑞金医院 MMC 人工智能辅助构建知识图谱大赛 [EB/OL]. [2020 - 11 - 06]. https://tianchi.aliyun.com/competition/entrance/231687/introduction.

8 中共中央网络安全和信息化委员会办公室. 第 46 次《中国互联网络发展状况统计报告》[EB/OL]. [2020 - 11 - 06]. http://www.cac.gov.cn/2020 - 09/29/c_1602939918747816.htm.

9 Deardorff A, Masterton K, Roberts K, et al. A Protocol - driven Approach to Automatically Finding Authoritative Answers to Consumer Health Questions in Online Resources [J]. J Assoc Inf Sci Technol, 2017, 68 (7): 1724 - 1736.

10 Guo H, Na X, Hou L, et al. Classifying Chinese Questions Related to Health Care Posted by Consumers via the Internet [J]. J Med Internet Res, 2017, 19 (6): e220.

11 Guo H, Na X, Li J. Qcorp: an annotated classification corpus of Chinese health questions [J]. BMC Medical Informatics and Deciesion Making, 2018, 18 (Suppl1): 16.

12 中华医学会医学信息学分会. 评测任务发布页面 [EB/OL]. [2020 - 11 - 06]. https://www.kesci.com/home/competition/5f2d0ea1b4ac2e002c164d82.

(上接第 16 页)

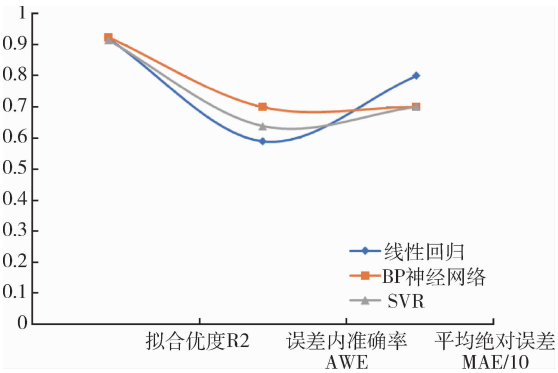

图 4 不同机器学习回归算法训练结果对比

5 结语

本研究基于 BP 神经网络方法构建手足口病预测预警模型并与其他预测预警模型进行比较。实验结果表明基于 BP 神经网络方法构建的手足口病预测预警模型具有较好拟合和泛化能力，相较而言具有更高准确性。

参考文献

1 刘金志. 小儿手足口病的流行病学特征及预防 [J]. 中国农村卫生, 2020, 12 (3): 34 - 35.

2 李小敏. 基于 BP 神经网络的心血管病预测系统的研究与实现 [D]. 银川: 宁夏大学, 2018.

3 王芳芳. 基于 BP 神经网络算法机理及应用探究 [J]. 科技创新导报, 2020, 17 (13): 150 - 151.

4 吴燎, 陈思珏. BP 神经网络在中医胃病病型诊断中的实现 [J]. 科技风, 2019 (20): 244.

5 张晗. BP 神经网络算法在人脸识别中的应用研究 [J]. 软件, 2020, 41 (5): 193 - 197.