文章编号：2095-2163(2019)06-0124-08 中图分类号：TP181 文献标志码：A

# 基于特定领域知识的医疗问答系统信息质量预测

胡 泽，张 展，左德承
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

**摘 要：** 伴随着智能手机以及移动互联网的高速普及，健康消费者越来越倾向于随时随地地在线咨询疾病、健康信息。其中最流行的方式便是医疗问答系统，因为其作为一种典型的在线问诊平台，可以为广大健康消费者提供足不出户、高效率以及高性价比的专业医生诊断体验。然而由于缺乏有效的信息质量管控机制，当前的医疗问答系统仍然会出现医生回答质量参差不齐的状况，这不仅会误导健康消费者，而且会造成医生的重复努力，同时也导致了积累的医疗问答知识库无法被有效复用。因而，对医疗问答系统的信息质量进行自动化预测就显得迫在眉睫。为此，本文提出了一种基于特定领域知识视角、协同训练以及集成学习的医疗问答系统信息质量预测算法。通过俘获不同特定领域知识视角间的高度非线性关系，有效地挖掘出了嵌入在大量未标记医疗问答数据中的特定领域语义知识，显著地提升了信息质量的预测性能。

**关键词：** 特定领域时序特征；特定领域表面语言特征；特定领域社会特征；协同训练；集成学习；医疗问答系统

# Information quality prediction of medical question-answering systems based on domain-specific knowledge

HU Ze, ZHANG Zhan, ZUO Decheng
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

**【Abstract】** With the rapid adoption of smartphones and mobile internet, health consumers are increasingly inclined to consult disease and health information online anytime and anywhere. The most popular way is medical question-answering systems. As typical online inquiry platforms, they can provide health consumers with high efficiency and cost-effective professional doctor diagnosis experience without leaving home. However, due to the lack of effective information quality control mechanism, the current medical question-answering systems still present a situation in which the quality of the physicians' answers varies greatly, which will not only mislead the health consumers, but also cause the repeated efforts of the doctors, and also lead to the accumulated medical question-answering knowledge base cannot be effectively reused. Therefore, it is extremely urgent to automatically predict the information quality of the medical question-answering systems. To this end, we propose a medical question-answering systems information quality prediction algorithm based on domain-specific knowledge views, co-training, and ensemble learning. By capturing the highly non-linear relationship between the different domain-specific knowledge views, we effectively mine domain-specific semantic knowledge embedded in a large amount of unlabeled medical question-answering data, which significantly improves the prediction performance of information quality.

**【Key words】** domain-specific surface linguistic features; domain-specific social features; domain-specific temporal features; co-training; ensemble learning; medical question-answering systems

## 0 引 言

在世界范围内，医疗资源分布的不均衡以及医疗资源的短缺长期困扰着广大健康消费者。特别是伴随着人口老龄化的加剧以及慢性疾病的频发，医疗资源匮乏的形势变得更加严峻，这不仅给医院造成了严峻的运营压力，而且也导致了医患关系的紧张[1]。得益于移动互联网和智能手机的普及，各种形式的医疗问答系统（如：好大夫在线、春雨医生以及平安好医生等）如雨后春笋般迅速兴起[2]。通过线上整合不同地域的医疗资源，优化导诊分诊服务流程，可以有效地缓解医生的工作压力，同时也可以提升健康消费者获得高质量的医疗服务的效率。这将很大程度上改善日益紧张的医患关系，缓解人们日益增长的高质量医疗服务需求和不平衡不充分的医疗水平发展之间的矛盾。

医疗问答系统中为健康消费者提供在线疾病和健康咨询服务的均为经过资格认证的专业医师，因

**基金项目：** 国家自然科学基金(61370085)。

**作者简介：** 胡 泽(1989-)，男，博士研究生，主要研究方向：医学信息学、自然语言处理、人工智能；张 展(1978-)，男，博士，副教授，硕士生导师，主要研究方向：容错计算、移动计算、系统可用性；左德承(1972-)，男，博士，教授，博士生导师，主要研究方向：容错计算、可穿戴计算、移动计算等。

**收稿日期：** 2019-02-26

而和传统的搜索引擎(如:百度搜索、搜狗搜索等)以及社区问答系统(如:百度知道、新浪爱问等)相比,可以提供更加权威、更值得信赖的回答[3-5]。借助于医疗问答系统的移动应用,老年人、看护者以及慢病患者可以在某种程度上实现足不出户的健康自我管理[5]。

尽管医疗问答系统为健康消费者们带来了极大的便利,但是由于缺乏有效的信息质量管控机制,其中仍旧充斥着一些低质量的回答[1,5]。例如:个别医生为了推销自己及其医院,往往提供一些答非所问的广告信息[5]。而一些高级别的医生由于工作繁忙,往往让专业水平不足的助理或者实习生代为解答健康消费者的问题[1]等等。低质量的回答通常导致健康消费者不得不再次咨询其他医生,浪费了本就短缺的医疗资源,同时还对积累的医疗问答知识库的二次开发使用造成了巨大干扰。高质量的医疗问答知识库是构建虚拟健康助理以及智能医疗问答系统的关键[1,6-8]。因而,对医疗问答系统中的信息质量采用自动化评估技术取代传统的人工评估就显得尤为重要。

为此,本文在深入剖析医疗问答系统的特点后,提取出了 2 种类别的特定领域非文本特征:特定领域表面语言特征[1]和特定领域时序特征[5]。将这 2 个特征作为特定领域非文本视角输入到结合了集成学习技术的协同训练框架中。由此获得了一个可以挖掘嵌入在大量未标注医疗问答数据中的特定领域隐藏语义知识的算法。通过俘获不同特定领域非文本视角间的高度非线性关系,该算法获得了比基线方法更好的信息质量预测性能。

## 1　相关工作

本研究主要涉及到回答质量预测的相关工作。由于主流的问答场景主要分为开放领域的社区问答系统以及垂直领域的专家问答服务,本研究将从社区问答系统的回答质量预测以及医疗问答系统的回答质量预测 2 个方面来做介绍。

### 1.1　社区问答系统质量预测

社区问答系统是一个任何用户都可以提出问题或者解答问题的平台,用户可以轻松地获得问题的解决方案,并且自由地交流和共享知识[4,9-10]。在社区问答系统的回答质量预测任务中,已有的研究工作通常假设由用户选择的最佳答案或者获得投票数最多的答案为高质量答案,其余的候选答案则被认为是低质量答案,如此该任务被归约为一个经典的二分类问题[10]。通过使用特征工程提取一系列的特征,并且将这些特征输入到机器学习模型进行训练,一个用于社区问答系统回答质量预测的分类模型被建立[9]。

Jeon 等人使用从 Naver 提取的非文本特征以及最大熵模型进行了社区问答系统的回答质量预测的初始研究[11]。Cai and Chakravarty 发现社区问答系统在本质上是动态的,提取具有时序属性的特征可以有效地提升回答质量预测性能,同时认为社区问答系统的数据集是不平衡的,使用精确率和召回率来衡量回答质量高低是不恰当的。为此,提出了学习排名方法来对所有的候选答案的质量进行评估[12]。Shah and Pomerantz 提出了 13 个不同的回答质量评价标准,对挑选的问答对进行人工标注,并且和真实用户的评分进行了比对,发现所提出的质量评价标准可以真实地反映提问者的看法。此外,还从问题文本、回答文本以及用户的个人资料中提取了一系列特征,并且使用逻辑回归分类模型探究了哪些特征可以有效地辨别最佳答案,因而发现诸如用户的个人资料之类的上下文信息对于评估和预测社区问答系统的回答质量是至关重要的[10]。Agichtein 以及 Bian 等人使用基于内容的特征、基于使用性统计的特征以及开发的反映贡献者关系的基于图的模型对社交媒体的信息质量进行了评估,并且获得了接近人类认知水平的预测性能[13-14]。Harper 等人探究了收费与否对于回答质量的影响,发现收费服务可以显著地调动回答者的积极性,使得提问者可以获得比免费咨询服务质量、效率更高的回答[15]。

### 1.2　医疗问答系统质量预测

与社区问答系统的回答质量预测任务相比,医疗问答系统的回答质量预测任务近年来才引起研究者的关注。Hu 等人使用可扩展多模深度信念网络以及特定领域非文本特征对医疗问答系统的回答质量进行了初始研究。通过挖掘回答短文本中的隐藏语义表示,以及与特定领域非文本特征进行特征融合,提出的深度学习框架获得了当时最好的预测性能[1]。稍后,Hu 等人又提出了特定领域时序特征、协同决策策略以及首个可以为医疗问答系统上下文提供特定领域知识的特定领域词嵌入。在此基础上,提出了一个新颖的名为“协同决策卷积神经网络”的深度学习框架,该框架不仅可以俘获不同类别特征间的高度非线性关系,而且可以俘获同一类别不同特征间的非独立交互关系,同时还可以从特

定领域词嵌入中引入额外的语义知识。如此,Hu 等人提出的深度学习框架有效地扩充了回答短文本的语义空间,克服了回答短文本所面临的严峻的特征稀疏问题,获得了医疗问答系统上下文中当前最好的回答质量预测性能[5]。

## 2 材料和方法

### 2.1 问题定义

与社区问答系统上下文中的回答质量预测任务相似,本研究将医疗问答系统上下文中的回答质量自动化预测任务定义为一个二分类问题[1]。由于文中引入了结合集成学习的协同训练框架,因而该研究也可以被看作一个多视角学习问题。即使用多视角学习从 2 个条件独立而充分冗余的特定领域非文本视角学习出一个可以鉴别医生回答质量高低的融合分类器,使用该融合分类器对新产生的医生回答的质量进行量化计算。

### 2.2 数据集准备以及性能评价指标

本研究采用了先前研究中所采集的好大夫在线数据集。其中包含用作协同训练初始标注数据集和监督学习训练集随机抽取的 2 800 个已标注问答对,用作协同训练和监督学习的标准测试集的 400 个已标注问答对,以及用作协同训练未标注数据集的5 000个未标注问答对[5]。值得注意的是,在已标注数据集中,高低质量的问答对数量是相等的。由先前的研究得知,在不平衡数据集上将会训练出一个预测偏差较大的糟糕的分类模型,特别是在协同训练初始标注数据集往往较小的情形下[1]。

作为一个二分类问题,本研究将高低质量回答分别看作为正负类,并且报告了算法在正类上的预测性能。所有的实验都经过 5 轮重复,并且报告了平均性能以及对应的预测偏差,以此来确保实验结果的稳定性和可靠性。本研究所使用的预测性能评价指标包括精确率($P$)、召回率($R$)、$F1$ 和 $AUC$。

### 2.3 特定领域非文本视角

本文采用了先前研究中所提出的 2 种类型的特定领域非文本特征来作为特定领域非文本视角,包括特定领域表面语言特征[1] 和特定领域时序特征[5]。

• 特定领域表面语言特征(slf):特定领域表面语言特征主要反映了医生回答的 3 方面的属性:

(1)医生的写作风格,例如医生回答中的不重复词语的数量在某种程度上反映了医生回答的流畅度。

(2)问答对之间的关系,例如一个高质量的回答往往和患者所提问题有着高度的相关性,而低质量的回答往往是不相关的广告及垃圾信息。

(3)医生的专业水平,例如一个受过高等教育、临床经验丰富的医生往往会给出一个包含较多医疗专业术语以及权威参考文献的高质量回答。而一个资质一般的医生给出的回答通常比较通俗简单,缺乏医疗专业性。

主要的特定领域表面语言特征包括患者总数目以及总的患者总访问数量等。经过预处理以及正则化,每个医生的回答被间接表示为一个 34 维度的实数值向量。

特定领域时序特征(tf):特定领域时序特征主要反映了一个医生在特定时间周期内的动态表现。因为医疗问答系统本质上是动态的,而一个医生在不同时间的状态和表现也截然不同。例如一个通常给出高质量回答的医生在工作繁忙或者心情不好的时候也可能给出低质量的回答。而一个通常给出低质量回答的医生在遇到自己擅长的问题或者心情足够舒畅的时候也可能给出详细而高质量的回答。主要的特定领域时序特征包含特定时间周期内患者访问数量以及特定时间周期内医生被患者推荐的水平等。经过预处理以及正则化,每个医生的回答被间接表示为一个 17 维度的实数值向量。

### 2.4 算法

协同训练框架作为一种经典的半监督学习方法。在指代消解、词性标记[16]、垃圾邮件分类[17] 以及词义消歧[18] 等方面有着重要的应用[19]。在存在大量未标注数据集的场景,仅需要数量较小的初始标注训练集,协同训练框架便可以利用大量未标注数据来提升预测性能,有效地避免了昂贵的人工标注,提升了效率。标准的协同训练框架最早是由 Blum 和 Mitchell 提出,通过在真实网页分类任务中使用协同训练框架,证实了该框架可以有效地利用大量未标注的廉价、易获取的数据来提升分类性能。标准的协同训练框架要求 2 个视角满足条件独立和充分冗余的假设[20]。Wang 和 Zhou 对协同训练框架进行了全新的理论分析,并且将协同训练抽象为一种双视角间的组合标签传播过程[21]。Yu 等人对协同训练框架所基于的理论假设的适用场景进行了深入探究,并且提出了一种全新的适用于多视角学习的基于无向图模型的贝叶斯协同训练框架,有效地解决了存在视角缺失情形数据的利用问题[22]。Sun 等人则提出了一种基于实体的协同训练算法,

该算法无需依赖类别分布先验知识便可以获得接近监督学习的预测性能[23]。

以 2.3 节的 2 个特定领域非文本视角为基础，提出了一个基于协同训练和集成学习的医疗问答系统质量预测算法。该算法的工作流程如下：

(1)来自于特定领域表面语言特征视角的已标注回答的特征映射被用于训练基级分类器 $C_1$。

(2) $C_1$ 用于预测来自于特定领域表面语言特征视角的未标注回答的特征映射。

(3)获得最置信特定数量新的已标注回答被传递到特定领域时序视角的已标注回答数据集作为训练集，以此来提升预测性能。特定领域时序视角对于已标注回答和未标注回答的特征映射的处理方式与特定领域表面语言视角相似。

经过特定的迭代次数后，可为每个特定领域视角分别学习出一个最优的基级分类器。随后，使用集成学习的和规则融合 2 个最优基级分类器的结果，因为在 4 个结果融合规则(和规则、最小值规则、最大值规则和乘法规则)[19,24-25]中，和规则表现出最佳性能。如此，不仅俘获了特定领域表面语言视角和特定领域时序视角间的高度非线性关系，而且挖掘出了嵌入在大量未标注问答数据集中的高度非线性语义知识。详细的算法流程见表 1。

**表 1　基于协同训练和集成学习的医疗问答系统质量预测算法**

**Tab. 1　Medical question-answering systems quality prediction algorithm based on co-training and ensemble learning**

Algorithm　ST-CoT (Surface linguistic features and temporal features (ST)-based co-training (CoT))

Input：

基级分类器 $C_1$，$C_2$

特定领域表面语言视角 $v_{slf}$

特定领域时序视角 $v_{tf}$

已标记数据集 $\int$ 和未标记数据集 $u$

Output：使用集成学习规则生成的融合分类器 $C_{fusion}$

1：for $i$ = 1 to $M$ do // $M$ 为最大迭代次数

2：　在已标记数据集 $\int$ 上使用 $\int$：$v_{slf}$ 训练基级分类器 $C_1$，并且使用 $C_1$ 来标记 u：$v_{slf}$

3：　在已标记数据集 $\int$ 上使用 $\int$：$v_{tf}$ 训练基级分类器 $C_2$，并且使用 $C_2$ 来标记 $u$ $v_{tf}$

4：　从基级分类器 $C_1$ 的预测结果中挑选置信度大于等于特定阈值的 $p$ 个正例和 $n$ 个负例，并且将其加入 $\varepsilon_{slf}$

5：　从基级分类器 $C_2$ 的预测结果中挑选置信度大于等于特定阈值的 $p$ 个正例和 $n$ 个负例，并且将其加入 $\varepsilon_{tf}$

6：　从未标记数据集 $u$ 中移除 $\varepsilon_{slf} \cup \varepsilon_{tf}$，并且将其加入到已标记数据集 $\int$ 中

### 2.5　基线视角以及基线半监督学习方法

基线视角包括 3 种社区问答系统，回答质量预测上下文中流行的文本特征提取方法(BOW_binary、BOW_CHI 和 LDA)，以及一种特定领域非文本特征(特定领域社会化特征)。

(1)BOW_binary：文本特征提取方法是基于高频词构建的二进制加权的词袋模型。经过分词、词性标记以及去停用词的标准预处理后，每个医生的回答被表示为一个二进制加权的2 812维度的 0/1 向量[5]。

(2)BOW_CHI：文本特征提取方法使用卡方统计[26]构建词袋模型。经过标准的预处理之后，每个医生的回答被表示为一个文档逆文档频率[27]加权的2 812维度的实数值向量[5]。

(3)LDA：文本特征提取方法是一个经典的主题模型[28]。通过对医生的回答文本进行粗粒度建模，每个医生的回答被表示为一个 25 维度的实数值主题向量[5]。

(4)特定领域社会化特征(sf)：主要从统计角度反映了医生的历史表现。通过对医生的个人资料进行统计分析，可以获得医生的受欢迎程度、受教育水平、好评率以及专业等级水平等。先前的研究也证实一个受教育程度高(例如博士)、专业等级水平高(例如主任医师)的医生更可能给出一个详细而高质量的回答。主要的特定领域社会化特征包括问答对之间的重复词语数量以及问答对之间的相似度等。经过预处理以及正则化，每个医生的回答被间接表示为一个 26 维度的实数值向量[5]。

基线半监督学习方法包括基于随机子空间切割的 RSS-CoT 算法[19]、基于内容和社交的 CS-CoT 算法[19]以及经典的直推式向量机 TSVM[29]。

### 2.6　超参数调优

对于文中提出的 ST-CoT 算法，使用大小为 300 的初始已标注训练集，大小为 400 的已标注测试集。每次迭代过程中挑选的最置信的正负样本数目均为 2，获得最佳实验结果的迭代次数为 20，结果的融合规则使用集成学习中的和规则。参数调优的详细细节请参见 3.3 节。

## 3　结果和讨论

### 3.1　监督学习预测性能分析

对不同视角及其组合在来自于社区问答系统回答质量预测上下文的常用的 3 种分类器逻辑回归(LR)[30]、支持向量机(SVM)[31]和朴素贝叶斯

(NB)[32]的预测性能进行了综合的比较与分析。

为了公平的比较,采用了与提出的 ST-CoT 算法获得最佳预测性能时相同数目的从已标注数据集中随机抽取的训练集和测试集来进行监督学习预测性能分析,即 460 (300+2 * (2+2) * 20)个训练集和 400 个测试集。见表 2 和表 3,可以观察到如下结论:

(1)从最能体现分类模型整体性能的 AUC 指标来看,特定领域表面语言视角和特定领域时序视角的组合在 LR、SVM 以及 NB 分类器上均获得了最佳性能。使用这两个特定领域非文本视角来作为本研究所提出的 ST-CoT 算法的两个视角。

(2)见表 2 和表 3 的前三行,三种基线文本视角在 SVM 和 NB 分类器上均表现出了十分糟糕的性能。这表明三种基线文本视角不适用于在数量特别小的已标注数据集上构建有效的监督分类器。因而在视角的两两组合实验以及小节 3.2 中的半监督学习方法对比实验中,三种基线文本视角不再被考虑。

(3)见表 2 和表 3 的最后三行,从最能反映模型整体性能的 AUC 指标来看,在三种特定领域非文本视角的两两组合实验中,LR 分类器始终可以获得最佳的性能。受此启发将 LR 分类器作为本研究所提出的 ST-CoT 算法的基级分类器,以此获得更好、更稳定的预测性能。

**表 2 不同视角在监督学习方法 LR 和 SVM 上的预测结果(平均值±标准偏差)**

**Tab. 2 Prediction results from different views on supervised learning methods LR and SVM (mean value±standard deviation)**

| Feature sets | LR | | | | SVM | | | |
|---|---|---|---|---|---|---|---|---|
| | *P*/ % | *R*/% | *F*1/ % | *AUC*/% | *P*/ % | *R*/% | *F*1/ % | *AUC*/% |
| BOW_binary | 71.70±0.83 | 76.90±5.10 | 74.10±2.18 | 73.90±2.41 | 47.80±14.70 | 80.20±31.80 | 58.80±20.00 | 51.00±15.20 |
| BOW_CHI | 76.10±1.96 | 71.80±3.14 | 73.80±1.29 | 75.00±0.80 | 44.10±22.50 | 64.50±35.10 | 50.70±24.50 | 44.60±20.00 |
| LDA | 70.10±1.63 | 70.10±1.63 | 77.70±1.20 | 75.80±1.69 | 73.80±11.00 | 58.10±33.70 | 55.40±26.60 | 65.20±10.20 |
| slf | 79.40±2.26 | 81.90±3.18 | 80.60±1.60 | 80.50±1.49 | 81.10±1.66 | 80.80±3.12 | 80.90±1.30 | 81.10±0.99 |
| sf | 72.90±2.86 | 74.80±5.23 | 73.60±1.04 | 73.70±0.64 | 70.70±3.46 | 72.70±3.89 | 71.50±1.06 | 71.50±0.90 |
| tf | 72.40±1.76 | 77.70±4.21 | 74.90±1.53 | 74.30±1.60 | 69.80±10.30 | 69.80±10.30 | 66.10±6.47 | 67.10±4.52 |
| slf+sf | 84.80±2.46 | 83.80±2.98 | 84.20±1.20 | 84.50±1.17 | 77.40±3.02 | 79.20±3.02 | 78.20±0.46 | 78.20±0.61 |
| slf+tf | **85.40±2.11** | **87.50±3.45** | **86.30±1.34** | **86.40±1.22** | **84.10±1.28** | **82.60±2.69** | **83.30±1.18** | **83.50±1.04** |
| sf+tf | 76.00±4.33 | 77.10±4.81 | 76.30±1.19 | 76.40±1.03 | 71.20±4.11 | 71.60±4.40 | 71.20±1.17 | 71.50±1.04 |

**表 3 不同视角在监督学习方法 LR 和 NB 上的预测结果(平均值±标准偏差)**

**Tab. 3 Prediction results from different views on supervised learning methods LR and NB (mean value±standard deviation)**

| Feature sets | LR | | | | NB | | | |
|---|---|---|---|---|---|---|---|---|
| | *P*/ % | *R*/% | *F*1/ % | *AUC*/% | *P*/ % | *R*/% | *F*1/ % | *AUC*/% |
| BOW_binary | 71.70±0.83 | 76.90±5.10 | 74.10±2.18 | 73.90±2.41 | 50.90±22.00 | 74.00±36.40 | 59.20±28.50 | 62.30±12.70 |
| BOW_CHI | 76.10±1.96 | 71.80±3.14 | 73.80±1.29 | 75.00±0.80 | 55.90±21.20 | 66.30±29.30 | 59.30±24.40 | 60.60±17.00 |
| LDA | 70.10±1.63 | 70.10±1.63 | 77.70±1.20 | 75.80±1.69 | 67.90±6.01 | **88.30±5.98** | 76.30±2.38 | 73.70±3.40 |
| slf | 79.40±2.26 | 81.90±3.18 | 80.60±1.60 | 80.50±1.49 | 72.30±8.99 | 75.30±17.20 | 71.40±5.28 | 71.80±2.11 |
| sf | 72.90±2.86 | 74.80±5.23 | 73.60±1.04 | 73.70±0.64 | 65.00±3.63 | 85.20±5.28 | 73.50±0.71 | 69.80±1.14 |
| tf | 72.40±1.76 | 77.70±4.21 | 74.90±1.53 | 74.30±1.60 | 72.60±9.12 | 74.00±18.30 | 70.70±5.43 | 71.50±0.97 |
| slf+sf | 84.80±2.46 | 83.80±2.98 | 84.20±1.20 | 84.50±1.17 | 71.50±5.09 | 87.40±3.68 | **78.40±1.76** | 76.30±2.36 |
| slf+tf | **85.40±2.11** | **87.50±3.45** | **86.30±1.34** | **86.40±1.22** | **81.50±6.06** | 76.50±14.00 | 77.60±5.98 | **79.20±3.28** |
| sf+tf | 76.00±4.33 | 77.10±4.81 | 76.30±1.19 | 76.40±1.03 | 71.90±6.68 | 77.70±10.90 | 73.60±2.94 | 73.10±1.13 |

### 3.2 半监督学习预测性能分析

表 4 展示了本文提出的 ST-CoT 算法与基线方法的对比结果,为了更直观地表示本文提出的算法的具体构成,在表 4 中使用 LR_slf_tf_CoT 代替 ST-CoT 来表示算法。Supervised( )代表了 3.1 节中监督学习分类器 LR 所取得的结果。RSS 前缀表示该方法的视角经过了随机子空间切割算法的处理[33]。LR_slf_tf_CoT (slf)代表 ST-CoT 算法的 $C_1$, LR_slf_tf_CoT (tf)代表 ST-CoT 算法的 $C_2$, LR_slf_tf_CoT (slf+tf)代表 ST-CoT 算法的 $C_{fusion}$。从表 4 可以观察到如下结论:

（1）本文算法在最能体现模型整体性能的 AUC 指标上获得了最佳的性能，并且明显优于社区问答系统回答质量预测上下文中的最新半监督学习方法 RSS-CoT 和 CS-CoT。同时该算法也优于典型的半监督学习方法 TSVM。这是因为本文提出的算法不仅可以从两个特定领域非文本视角引入额外的领域知识，而且可以挖掘隐藏在大量未标注问答数据集中的高度非线性语义知识。此外，算法还可以借助于集成学习俘获两个特定领域非文本视角间的高度非线性关系。

（2）见表 4 最后三行所示，算法在最能反映模型整体性能的 AUC 指标上获得了比监督学习方法更好的性能。这是因为与监督学习方法，简单的线性组合 2 个特定领域非文本视角相比，本文提出的算法可以俘获隐藏在 2 个特定领域非文本视角间的高度非线性关系。同时发现 RSS_LR_slf_tf_CoT 算法获得了最差的性能，这是因为 RSS 算法破坏了 2 个特定领域非文本视角的原有特征空间，造成原有的视角间的高度非线性关系以及同一视角内的特征间的非独立交互关系丢失。

**表 4　本文提出的算法和基线方法的比较结果（平均值±标准偏差）**

**Tab. 4　Comparison results of our proposed algorithm and baseline methods (mean value±standard deviation)**

| Methods | *P* /% | *R*/ % | *F*1/ % | *AUC*/ % |
|---|---|---|---|---|
| TSVM | **94.50±1.39** | 75.90±3.80 | 84.20±2.66 | 85.70±2.33 |
| RSS-CoT | 72.50±0.74 | 69.80±0.87 | 71.10±0.46 | 71.70±0.46 |
| CS-CoT | 88.90±1.55 | 76.60±0.58 | 82.30±0.38 | 83.50±0.52 |
| Supervised (slf) | 79.40±2.26 | 81.90±3.18 | 80.60±1.60 | 80.50±1.49 |
| RSS_LR_slf_tf_CoT ($slf_{RSS}$) | 79.60±0.49 | 87.90±1.02 | 83.50±0.38 | 82.70±0.34 |
| LR_slf_tf_CoT (slf) | 79.40±1.21 | 82.20±1.12 | 80.80±0.54 | 80.50±0.64 |
| Supervised (tf) | 72.40±1.76 | 77.70±4.21 | 74.90±1.53 | 74.30±1.60 |
| RSS_LR_slf_tf_CoT ($tf_{RSS}$) | 68.40±1.34 | 75.70±2.71 | 71.80±1.01 | 70.40±0.90 |
| LR_slf_tf_CoT (tf) | 77.50±0.74 | 75.00±2.86 | 76.20±1.57 | 76.60±1.18 |
| Supervised (slf+tf) | 85.40±2.11 | 87.50±3.45 | 86.30±1.34 | 86.40±1.22 |
| RSS_LR_slf_tf_CoT ($slf_{RSS}$+$tf_{RSS}$) | 76.70±0.25 | **88.10±1.39** | 82.00±0.50 | 80.70±0.37 |
| LR_slf_tf_CoT (slf+tf) | 87.90±1.37 | 86.80±0.81 | **87.30±0.70** | **87.40±0.77** |

### 3.3　算法核心超参数对于预测性能的影响

本小节对本文提出的 ST-CoT 算法的三个核心参数对于最终预测性能的影响进行了讨论，并且使用最能反映模型整体性能的 AUC 指标来进行可视化。此外，不同的集成学习结果组合规则对于结果的影响也被探究。

图 1 展示了初始标注数据集大小对于算法性能的影响。由此发现算法的性能伴随着初始数据集大小的增加先迅速增加，随后保持平稳。这是因为伴随着初始标注数据集大小的增加，模型的泛化能力不断提高，从而抵抗数据集中的噪声干扰的能力也不断变强。本研究提出的算法致力于降低人工标注成本的宗旨，选择了模型可以获得稳定泛化能力的最小初始标注数据集数目 300 来进行实验。

图 2 展示了迭代次数对于算法性能的影响。伴随着迭代次数的增加，算法的性能先快速上升，然后开始波动式下降。这是因为在算法训练的初始阶段，伴随着迭代次数的增加，已标注数据集的大小开始逐渐变大，模型的泛化能力也随之升高。但是在迭代次数达到特定阈值后，算法积累的来源于未标注数据集的噪声干扰开始显现负面作用。为了取得最佳性能，本文选择 20 作为迭代次数。

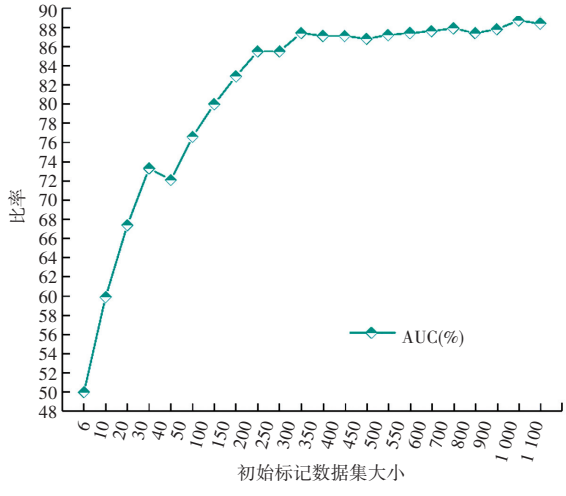


**图 1　初始标记数据集大小对算法性能的影响**

**Fig. 1　The influence of initial labeled dataset size on the performance of our proposed algorithm**

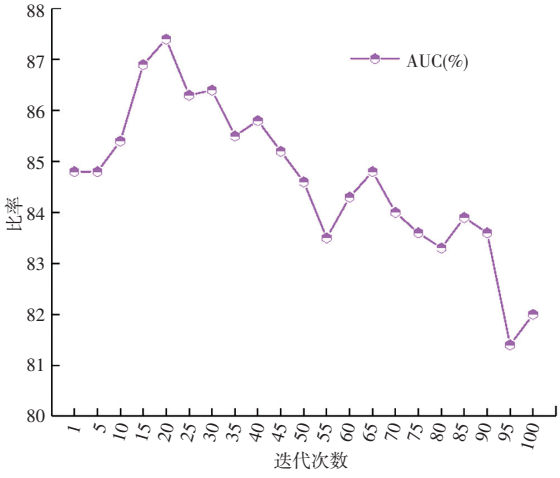


图 2 迭代次数对算法性能的影响

**Fig. 2 The influence of iterations on the performance of our proposed algorithm**

图 3 展示了算法在每次迭代中选择的最置信正负样本数目对于算法性能的影响。发现正负样本数目相等时,算法可以获得一个稳定的性能,而当样本数目不均衡时,特别是正样本数目低于负样本数目时,算法性能出现了严重的下降。这是因为相等数量的正负样本可以训练出更加稳定的基级分类器,避免预测偏差。此外还发现当正负样本数目相等时,算法的性能伴随着数目的增加出现波动式下降。这是因为单次迭代中选择的正负样本数目越大,越容易引入未标注数据集中的噪声干扰。因此,本文使用数目为 2 的正负样本来进行实验。

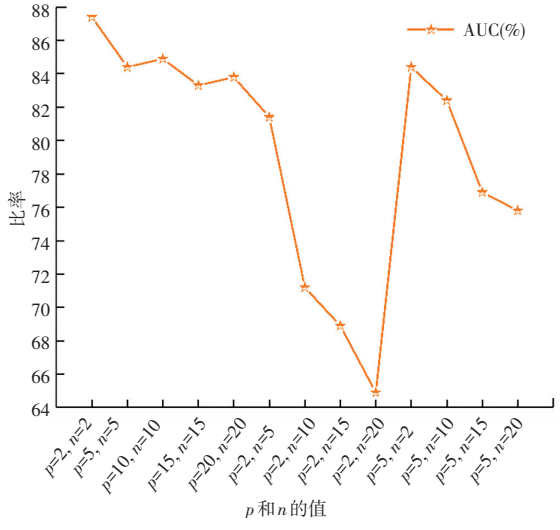


图 3 $p$ 和 $n$ 对算法性能的影响

**Fig. 3 The influence of $p$ and $n$ on the performance of our proposed algorithm**

见表 5 所示,集成学习的和规则在除了召回率之外的所有性能指标上均获得了最佳性能。是因为和规则可以更好地拟合 2 个特定领域非文本视角间的高度非线性关系。因而,本文将和规则用于结果融合。

表 5 不同结果融合规则对算法性能影响(平均值±标准偏差)

**Tab. 5 The influence of different result fusion rules on the performance of our proposed algorithm (mean value ± standard deviation)**

| Rules | *P* /% | *R*/ % | *F*1 /% | *AUC*/ % |
|---|---|---|---|---|
| Max | 84.60±1.31 | 87.80±1.03 | 86.20±0.84 | 85.90±0.92 |
| Product | 84.50±0.59 | 89.60±0.37 | 86.90±0.24 | 86.60±0.29 |
| Min | 82.90±1.95 | **90.70±0.24** | 86.60±0.98 | 86.00±1.22 |
| Sum | **87.90±1.37** | 86.80±0.81 | **87.30±0.70** | **87.40±0.77** |

## 4 结束语

本文使用特定领域非文本视角以及结合了集成学习的协同训练框架,对医疗问答系统上下文中的回答质量预测任务进行了研究。通过使用特定领域非文本视角,引入了额外的特定领域统计知识。通过使用协同训练框架,有效地挖掘出了隐藏在大量未标注问答数据集中的特定领域语义知识。通过使用集成学习,俘获了 2 个不同的特定领域视角间的高度非线性关系。如此提出的算法相较于已有方法获得了显著的性能提升。

在下一步研究中,将探索新的特定领域视角来进一步提升模型的预测性能。同时准备对协同训练框架的基分类器进行改进,使其可以更好地建模面临特征稀疏问题的短文本视角以及可以在更小的初始标注数据集上稳定工作。

## 参考文献

[1] HU Z, ZHANG Z, YANG H, et al. A deep learning approach for predicting the quality of online health expert question-answering services[J]. Journal of Biomedical Informatics, 2017, 71(C): 241-253.

[2] SILVA B M, RODRIGUES J J, DE LA TORRE DÍEZ I, et al. Mobile-health: A review of current state in 2015[J]. Journal of biomedical informatics, 2015, 56:265-272.

[3] SHAH C, OH S, OH J S. Research agenda for social Q&A[J]. Library & Information Science Research, 2009, 31(4):205-209.

[4] LIU Y, BIAN J, AGICHTEIN E. Predicting information seeker satisfaction in community question answering[C]. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008:483-490.

[5] HU Z, ZHANG Z, YANG H, et al. Predicting the quality of online health expert question-answering services with temporal features in a deep learning framework[J]. Neurocomputing, 2018, 275:2769-2782.

[6] YANG PJ, FU WT. Mindbot: a social-based medical virtual assistant[C]. 2016 IEEE International Conference on Healthcare Informatics (ICHI), 2016:319-319.

[7] DO H J, FU W T. Empathic Virual Assistant for Healthcare Information with Positive Emotional Experience[C]. 2016 IEEE International Conference on Healthcare Informatics (ICHI), 2016: 318-318.
[8] KINCAID R, POLLOCK G. Nicky: Toward a Virtual Assistant for Test and Measurement Instrument Recommendations[C]. 2017 IEEE 11th International Conference on Semantic Computing (ICSC), 2017:196-203.
[9] TIAN Q, ZHANG P, LI B. Towards Predicting the Best Answers in Community - based Question - Answering Services [C]. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2013:725-728.
[10] SHAH C, POMERANTZ J. Evaluating and predicting answer quality in community QA[C]. Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010:411-418.
[11] JEON J, CROFT W B, LEE J H, et al. A framework to predict the quality of answers with non-textual features[C]. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006:228-235.
[12] CAI Y, CHAKRAVARTY S. Answer Quality Prediction in Q/A Social Networks by Leveraging Temporal Features[J]. International Journal of Next-Generation Computing, 2013, 4(1):1-27.
[13] AGICHTEIN E, CASTILLO C, DONATO D, et al. Finding high-quality content in social media[C]. Proceedings of the 2008 international conference on web search and data mining, 2008:183-194.
[14] BIAN J, LIU Y, ZHOU D, et al. Learning to recognize reliable users and content in social media with coupled mutual reinforcement[C]. Proceedings of the 18th international conference on World wide web, 2009:51-60.
[15] HARPER F M, RABAN D, RAFAELI S, et al. Predictors of answer quality in online Q&A sites [C]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2008:865-874.
[16] CLARK S, CURRAN J R, OSBORNE M. Bootstrapping POS taggers using unlabelled data [C]. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, 2003:49-55.
[17] KIRITCHENKO S, MATWIN S. Email classification with co-training[C]. Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative research, 2001:301-312.
[18] MIHALCEA R. Co-training and self-training for word sense disambiguation[C]. Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004, 2004:1-8.
[19] LIU B, FENG J, LIU M, et al. Predicting the quality of user-generated answers using co-training in community-based question answering portals[J]. Pattern Recognition Letters, 2015, 58:29-34.
[20] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co - training [C]. Proceedings of the eleventh annual conference on Computational learning theory, 1998:92-100.
[21] WANG W, ZHOU Z H. A New Analysis of Co-Training[C]. ICML, 2010:1135-1142.
[22] YU S, KRISHNAPURAM B, ROSALES R, et al. Bayesian co-training[J]. Journal of Machine Learning Research, 2011, 12 (Sep):2649-2680.
[23] SUN A, LIU Y, LIM E P. Web classification of conceptual entities using co-training[J]. Expert Systems with Applications, 2011, 38(12):14367-14375.
[24] KITTLER J, HATEF M, DUIN R P, et al. On combining classifiers[J]. IEEE transactions on pattern analysis and machine intelligence, 1998, 20(3):226-239.
[25] KUNCHEVA L I. A theoretical study on six classifier fusion strategies[J]. IEEE Transactions on pattern analysis and machine intelligence, 2002, 24(2):281-286.
[26] ZHAO X, MA J. Modify the Method of Feature´s Weight in Text Classfication[J]. Computer Knowledge and Technology, 2009, 5 (36):10626-10628.
[27] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5):513-523.
[28] WANG T, HUANG Z, GAN C. On mining latent topics from healthcare chat logs[J]. Journal of biomedical informatics, 2016, 61:247-259.
[29] JOACHIMS T. Transductive inference for text classification using support vector machines[C]. ICML, 1999:200-209.
[30] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification[J]. The Journal of Machine Learning Research, 2008, 9:1871-1874.
[31] CHANG C C, LIN C J. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3):1-27.
[32] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in Python [J]. The Journal of Machine Learning Research, 2011, 12:2825-2830.
[33] HO T K. The random subspace method for constructing decision forests[J]. IEEE transactions on pattern analysis and machine intelligence, 1998, 20(8):832-844.

(上接第 123 页)

一定的应用价值。在实验过程中,发现并行平台中计算节点通常较多,而且不同的计算任务需要重新配置环境,为了避免在部署的过程中做重复性工作,下一步研究的重点考虑使用 PXE 无盘启动技术来进行节点的维护。

## 参考文献

[1] 刘韬. SSH 协议公钥登录的配置与应用[J]. 现代工业经济和现代化,2016.98-100.
[2] DONG Yanhua, LI Xiaojia, LI Shuang. Research on security telecommunication of diskless parallel computing system [J]. Information Management, Innovation Management And Industrial Engineering,2012.
[3] BRUCE SCHNEIER.应用密码学[M]. 北京:机械工业出版社,2002.29-30.
[4] 王鹏,吕爽,聂治等编著. 并行计算应用及实战[M]. 北京:机械工业出版社,2008.15-16.