面向深度自动问答的糖尿病饮食问题分类*

张芳芳 马敬东 王小贤 卢乃吉 夏晨曦

（华中科技大学同济医学院医药卫生管理学院 武汉 430030）

〔摘要〕 以糖尿病患者饮食问题为例，从用户视角出发，提出基于功能的问题分类体系，利用支持向量机算法对患者提出的问题进行分类，为深度自动问答系统的构建提供重要支持。

〔关键词〕 问题分类；支持向量机；糖尿病；饮食

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2017.03.003

Classification of Diabetes Diet Problems Oriented by Deep Automatic Question Answering *ZHANG Fang－fang, MA Jing－dong, WANG Xiao－xian, LU Nai－ji, XIA Chen－xi. School of Medical and Health Management, Tongji Medical College, HUST, Wuhan 430030, China*

〔**Abstract**〕 Taking the diet problem of diabetic patients as an example, the paper puts forward the problems classification system based on functions in the view of users, classifies the problems put forward by patients through the Support Vector Machine (SVM) algorithm, and provides important support for the construction of the deep automatic Question Answering (QA) system.

〔**Keywords**〕 Question classification; Support Vector Machine (SVM); Diabetes; Diet

1 引言

随着科技的飞速前进，互联网的不断发展及信息的不断增长，如何从海量信息中快速、准确地获取有用信息逐渐演变成一个更加重要的课题。在医学领域，在线健康网站是公众获取健康信息的主要渠道之一。面对海量质量参差不齐的健康信息，一方面，即便是专业人员仍需要耗费大量时间来实现健康信息的搜索、浏览及获取，搜索成功与否取决于用户的搜索技巧[1]；另一方面，由于医学专业的特殊性[2]，非专业人员在查找、理解及获取医学信息方面存在诸多困难和障碍，难以满足用户健康需求[3]。自动问答是一种智能化的信息服务方式，是特殊的搜索引擎，能够理解用户以口语化表达的问题，从后台知识库中直接返回答案。自动问答系统主要分为问题分类、问题理解、答案的抽取消歧等步骤，其中问题分类占据着关键步骤中的首要位置，其准确性直接影响到自动问答系统的问题解析、答案生成和整体性能[4]。

开放领域自动问答是当前自动问答研究的主流，其问题分类大都面向事实类问题，分类体系主要依据问题涉及的主题构建[5]，重点关注问题的内

〔收稿日期〕 2017－02－27

〔作者简介〕 张芳芳，硕士研究生，发表论文 1 篇；通讯作者：夏晨曦，讲师。

〔基金项目〕 教育部网络时代的科技论文快速共享专项课题“基于互联数据的论文共享方法研究”（项目编号：0214516155）。

容，对于具体领域并不适用。在医学信息领域，患者提问信息以功能性问题为主，但目前还没有研究对患者提问信息进行系统的分类[6]，中文健康领域的问题分类研究也相对匮乏，尤其是面向深度自动问答的问题分类研究。鉴于此，本文以糖尿病饮食问题为例，开展中文患者问题分类研究，阐述问题分类体系和语料库的构建，问题与处理、特征提取和分类方法，以及性能评测的结果，重点分析分类方法的特点和不足，讨论不同特征选取对分类效果的影响，同时展望了未来问题分类性能的改进措施，以期为医学领域自动问答研究的发展提供参考。

2 研究方法

2.1 患者问题语料库构建

本文选择专业健康网站及综合问答社区中糖尿病患者饮食的相关问题作为研究对象，这些问题由用户提出。为保证数据采集的完整性，采用网络蜘蛛抓取用户提问的问题数据，其中专业健康网站抓取糖尿病问答版块所有问题，而综合问答社区因其搜索引擎发展较为成熟，可直接抓取糖尿病饮食相关问题。对数据进行初步筛选，以“食”、“吃”、“喝”作为关键词，剔除药物相关问题及专业健康网站中同一用户、同一时间提问的重复问题，同时人工剔除与糖尿病无关的问题以及糖尿病患者提出的与饮食无关的无效问题。一些用户提出的真实问题比较复杂，由多个问题组成，本文将此类问题进行拆分，分解为多个问题。问题类型由手工标注完成，每个问题由 3 人标注，标注不一致的问题进行讨论并重新进行标注，最终选取 3 707 个有效问题作为问题集。

2.2 分类体系

分类体系是问题分类的依据。对于英文问题分类的研究，以 UIUC 问题集最具代表性。该问题集已成为当前英文问题分类研究的公用数据集[7]。分类体系并未有统一的标准，绝大多数机构都采用自己的分类体系[8]。针对中文问题分类，研究人员大多参照哈尔滨工业大学提出的中文问题分类体系；但该分类体系过多地关注于问题的内容，对于具体领域并不适用。针对以上问题，Bu 等[9]提出基于功能的问题分类体系，包括 fact、list、reason、solution、definition 及 navigation 共 6 个类别。董才正等[5]提出面向社区的中文问题分类，将问题类型分为定义、事实、过程、原因、观点、是非、描述 7 种类型。本文借鉴其分类方法，考虑到医疗健康领域的特殊性，从答案生成类型角度，提出一种面向在线健康问答社区的粗粒度分类体系，将在线健康问答社区中的问题分为 6 大类，即解释类、是非类、列举类、事实类、行为方式类及其他类。

2.3 分类方法

自动问答中问题分类的方法主要有两种，即手工编写规则和机器学习算法。早期的问题分类研究大多基于手工编写规则来实现，而且针对特定领域和特定问题分类体系，这种问题分类方法效果较好；但是，手工编写分类规则可扩展性较差。对于机器学习分类算法，通过提取能表达各种问题类型的特征规则，建立学习模型，可实现各种问题的类型识别。目前，基于统计的机器学习算法，如支持向量机（Support Vector Machine，SVM），是问题分类的主流方法。SVM 是基于统计的机器学习模型，在解决小样本、非线性及高维模式识别问题中具有许多特有的优势[10]，在小样本分类问题上的效果已经在文本分类、自然语言处理等方面得到了验证[11]，在查询分类领域使用 SVM 算法的正确率明显高于其他机器学习算法[12]。SVM 的主要工作原理是通过事先选择的非线性映射（核函数）将输入向量 X 映射到一个高维特征空间，以将原始空间的非线性可变问题分为高维空间的中线性可分问题[5]。本文采用 LibSVM 软件包中的 RBF 核函数进行问题分类试验，其基本模型如下：

$$K(x, x_i) = \exp\left(-\frac{\| x - x_i \|^2}{\sigma 2}\right)$$

3 试验及结果分析

3.1 试验流程

3.1.1 概述 对问题进行分类试验，首先，需要利用网络蜘蛛在健康问答网站抓取相关数据，对数据进行预处理，包括分词、去停用词、词性标注等工作；其次，对经过预处理的问题进行特征提取，以利用 SVM 分类器对问题进行分类。本文的问题分类试验主要包括数据抓取、数据预处理、特征提取、分类及分类效果评测等步骤，见图 1。

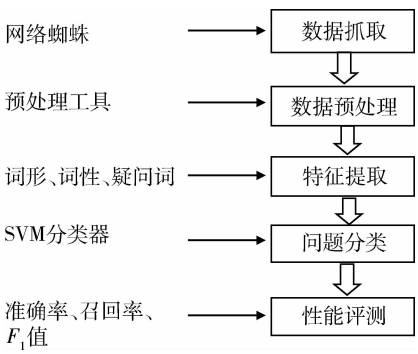

图 1 问题分类流程

3.1.2 试验数据抓取 由于目前关于问题分类的公开数据集没有对患者提问信息进行系统的分类[6]，因此本文抽取专业健康网站及综合问答社区中 3 707 个糖尿病饮食问题作为试验数据，其中训练集为 2 965 个，测试集为 742 个。手工标注问题类型，标注完成后对问题类型分布进行统计，见表 1。

表 1 问题类型分布情况

分类结果	问题数量（个）	百分比（%）
是非类	1 451	39.14
列举类	1 387	37.42
行为方式类	161	4.34
事实类	36	0.98
解释类	72	1.94
其他类	600	16.18

3.1.3 数据预处理 首先将语料集中的问题使用 jieba 分词开源工具进行分词及词性标注，根据停用词表去除对分类不具有实际意义的停用词，如标点符号、“的”、“啊”等，以减少影响分类的噪音。目前，文本分词已经有很多比较成熟的算法和工具，在处理中文文本时更多使用专门的分词工具，本文采用 jieba 分词工具，加入人工构建的外部词表，进行分词及词性标注以及去停用词等文本预处理。

3.1.4 特征提取和特征值计算 本文采用问题的词形特征、词形加词性特征及疑问词特征作为问题分类的特征。对于特征的取值，采用布尔编码方式，该特征出现在问题中即为 1，否则为 0。训练和分类时，将每个问题转换为分类器可识别的特征向量格式：

< label > < index 1 > : < value 1 > < index 2 > : < value 2 > . . . . . < index n > : < value n >

3.1.5 评价指标 本文使用常用的准确率、召回率和 F_1 值来评价问题分类结果的质量，具体定义如下：

$$\text{准确率} = \frac{\text{被正确分类问题数}}{\text{总的问题数}}$$

$$\text{召回率} = \frac{\text{该类被正确的问题数}}{\text{该类总的问题数}}$$

$$F_1 = \frac{\text{准确率} \times \text{召回率} \times 2}{(\text{正确率} + \text{召回率})}$$

3.2 试验结果

3.2.1 整体分类结果 为保证试验结果的可靠性，本文分别通过 5 - fold 交叉验证及测试集来测试问题的分类效果，见表 2。

表 2 分类准确率

项目	词形（%）	词形 + 词性（%）	词形 + 词性 + 疑问词（%）
5 - fold 交叉验证	93.22	93.25	93.89
20% 测试集	93.13	92.99	93.13

采用 5 - fold 交叉验证测试分类效果时，加入词性及疑问词特征后，分类效果有一定程度的提高；使用随机抽取的测试集测试分类效果时，加入词性及疑问词特征后，分类效果并无明显提升。本文以下试验均采用 5 - fold 交叉验证完成。整体来看，采用 SVM 分类器进行问题分类可以取得优秀的性能。

3.2.2 各类别分类结果 由于问题来自于真实的患者提问，每类问题的样本数量存在差异，因此为

验证各问题类别的分类效果，本文对各类别问题进行分类试验，见表 3。

表 3 各细粒度类别分类结果

问题类别	特征集 1（%）			特征集 2（%）			特征集 3（%）		
	P（%）	R（%）	F_1	P（%）	R（%）	F_1	P（%）	R（%）	F_1
分类 1（1 451）	97.84	93.47	0.95	97.83	92.78	0.95	99.63	92.44	0.96
分类 2（161）	88.89	75.00	0.81	85.71	75.00	0.80	72.97	84.38	0.78
分类 3（1 387）	92.06	96.39	0.94	92.39	96.39	0.95	93.66	96.03	0.95
分类 4（36）	66.67	28.57	0.40	66.67	28.57	0.40	80.80	57.14	0.67
分类 5（72）	88.89	53.33	0.67	88.89	53.33	0.67	66.67	40.00	0.50
分类 6（600）	87.40	98.33	0.93	86.86	99.17	0.93	86.86	99.17	0.93

注：特征集 1：词特征；特征集 2：词 + 词性特征；特征集 3：词 + 词性 + 疑问词特征

由表 3 可知，第 1、3、6 类的准确率、召回率及 F_1 值都较高，分类效果较好，而第 2、4、5 类的分类效果较差，可见样本数量不均衡影响着分类的准确率及召回率。从加入疑问词（特征集 3）后的试验结果发现，对于每一分类效果而言，提升并不明显。由此表明对于英文分类效果影响显著的特征，在中文问题分类中并不一定适用。总体来说分类效果较好。

3.3 结果分析

3.3.1 *分类性能提升* 在中文问题分类研究中，大都基于描述类问题[13]，而较少针对用户提出的真实性问题进行分类研究；且基于中文问题分类目前尚无统一的标准，中文问题发展的基础语料库仍不完善，其分类的准确度有待于进一步提高。由于涉及中文语言的特殊性，国外一些相关成熟的技术和研究成果不能利用；基础语料资源的缺乏，也严重制约着问题分类研究的进一步发展。与传统基于规则的问题分类研究相比，本试验可应用于医学领域用户提出的问题分类中。同时，鉴于对问题大都根据主题词进行分类，对开放领域分类效果较好，但对于特殊领域，如糖尿病饮食这一小类中，则根据用户的提问方式进行分类，该分类方法效果较好，且具有可移植性。与其他问题分类试验的效果比较，见表 4。

表 4 问题分类相关实验研究

相关研究	数据集	分类器	特征集	分类准确率（%）
文勖等（2006）	训练集 5 265，测试集 1 300	Bayes	主干词、疑问词及附属	71.92
杨思春等（2014）	训练集 4 966，测试集 1 300	SVM	基本特征及词袋绑定特征	84.70
董才正等（2016）	训练集 3 404，测试集 699	SVM	疑问词、基于焦点词	86.80
Loni，et al.（2012）	UIUC 数据集	SVM	词形、中心词、上位词等	93.60
本研究试验	训练集 2 695，测试集 742	SVM	词及词性绑定特征、疑问词特征	93.13

3.3.2 *样本分布对分类性能的影响* 由表 3 可知，样本数量较少的类别，分类效果相对较差。为验证样本数量与分类性能的关系，以词特征为例进行样本均衡性试验，各类别随机抽取 36 个样本，见表 5。

表 5 样本均衡性试验

问题类别	P（%）	R（%）	F_1
分类 1（36）	77.78	87.5	0.82
分类 2（36）	1	1	1.00
分类 3（36）	1	1	1.00
分类 4（36）	77.78	87.5	0.82
分类 5（36）	1	1	1.00
分类 6（36）	1	1	1.00

与表 3 结果相比，原样本相对较少的类别，如分类 2、4、5，在训练样本总量减少的条件下，F_1 值显著提高，可见原分类 2、4、5 的结果不理想，不是因为样本数量较少，而是因为训练集各分类样本数量分布不均。究其原因为：问题分类中样本数量过少，则可提取的特征相对较少，与样本量相差较大的类别相比，分类效果一定程度地下降。

3.3.3 *各种特征组合对性能的影响* 由试验结果可以看出，加入疑问词特征后分类精度有所提高，但效果并不明显。这是因为本试验所抓取的问题由用户提出的真实问题构成，大都是陈述问题，而疑问词特征并不明显，因此加入疑问词特征后分类精度的提升也并不明显。由此可见，适用于英文的分类特征，在中文环境下并不一定适用。

4 结语

本文针对在线专业健康网站及综合问答社区中的问题分类提出了基于功能的问题分类体系，从在线问答社区中抓取数据，构建语料库，通过试验考察了医学领域中文问题分类的效果及各类特征对问题分类性能的影响，结果表明样本分布不均对问题分类效果有较大影响，而某些在英文环境中重要的特征，在中文领域并不适用。本研究不局限于事实类问题，而是强调依据用户提问的方式或目的进行问题分类，因而更适合深层自动问答研究的需要。另外，以提取在线社区真实问题的方式构建语料库及具有可移植性的问题分类体系，针对中文问题的试验结果，对相关研究具有一定的参考价值。从试验结果来看，用户提问的问题中疑问词特征并不明显，同时词汇特征过于口语化，且存在用词错误等问题，鉴于此在后续研究中，应构建用户健康词表，将用户词汇与医学专业词汇进行匹配，进而提高分类效果。

参考文献

1 Wren JD. Question Answering Systems in Biology and Medicine—the time is now [J]. Bioinformatics, 2011, 27(14): 2025 – 2026.

2 李迎娟. 网络医学信息资源高效利用的绿色评价研究[J]. 医学信息学杂志, 2015, 36(5): 75 – 78.

3 王若佳, 李月琳. 基于用户体验的健康类搜索引擎可用性评估[J]. 图书情报工作, 2016(7): 1 – 11.

4 Bae K, Ko Y. How to Combine Translation Probabilities and Question Expansion for Question Classification in cQA Services [J]. IEICE Transactions on Information & Systems, 2016, 99(4): 1019 – 1022.

5 董才正, 刘柏嵩. 面向问答社区的中文问题分类[J]. 计算机应用, 2016, 36(4): 1060 – 1065.

6 吴东东, 刘锋, 于鸿飞. 利用决策树的患者咨询问题分类[J]. 中国数字医学, 2016, 11(2): 101 – 103.

7 Hovy E, Gerber L, Hermjakob U, et al. Toward Semantics – based Answer Pinpointing [C]. San Diego, CA: International Conference on Human Language Technology Research, 2001: 1 – 7.

8 杨思春, 戴新宇, 陈家骏. 面向开放域问答的问题分类技术研究进展[J]. 电子学报, 2015, 43(8): 1627 – 1636.

9 Bu F, Zhu X, Hao Y, et al. Function – based Question Classification for General QA [C]. Massachusetts, USA: Proc of the 2010 Conference on Emmpirical Methods in Natural Language Processing, 2010: 1119 – 1128.

10 高超. 中文问题分类中特征选择研究[D]. 合肥: 安徽工业大学, 2011.

11 Zhang XG. Introduction to Statistical Learning Theory and Support Vector Machines [J]. Acta Automatica Sinica, 2000, 26(1): 32 – 42.

12 杨思春, 高超, 戴新宇, 等. 基于 SVM 的中文查询分类[J]. 情报学报, 2011, 30(9): 946 – 950.

13 施亦龙, 许鑫. 在线健康信息搜寻研究进展及其启示[J]. 图书情报工作, 2013, 57(24): 123 – 131.