基于文本分析的自动化疾病编码方法①

鲍庆升, 程绍银, 蒋 凡

(中国科学技术大学 计算机科学与技术学院, 合肥 230027)

摘 要: 对疾病进行编码是将疾病诊断名称转化为标准 ICD(国际疾病分类)编码的过程. 鉴于编码量庞大和人工编码效率低等原因, 有必要实现疾病编码的自动化. 提出一种自动化的疾病编码方法, 使用一种文本建模方法将 ICD 表示为文本集, 然后借助文本相关性度量, 获取与待编码疾病诊断名称最相关的 ICD 编码. 经实验验证, 本文提出的自动化疾病编码方法准确率较高、效率优秀、分类层次变换灵活, 可广泛应用于各种类型的数据分析场景.

关键词: 疾病编码; 文本分析; ICD; 相关性分析

Automated Disease Coding Method Based on Text Analysis

BAO Qing-Sheng, CHENG Shao-Yin, JIANG Fan

(School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: Disease coding is the process of transforming diagnosis names to standard ICD(International Classification of Diseases) codes. Due to the huge amount of workload and the low efficiency of manual coding, it's necessary to achieve the automation of disease coding. This paper presents a method for automated coding of disease. Specifically, we present ICD as documents with a text modeling approach, and then get the most relevant ICD code of a diagnosis name with the text correlation measure. The experiments proves that our method has a good accuracy, it's very efficient and easy to switched among classification levels, could be widely applied to various types of data analysis scenarios.

Key words: disease coding; text analysis; ICD; correlation analysis

随着医疗信息化的逐步推进和数据分析技术的快速发展, 人们逐渐意识到医疗数据中潜在着巨大的价值待以挖掘[1-3]. 疾病种类信息, 作为诊疗信息重要的组成部分, 在医院科学精细化管理、医疗费用控制、服务质量改善等方面都具有非常重要的作用[4]. 疾病编码, 是编码人员按照国家规定的疾病分类标准 ICD(国际疾病分类), 将临床医生记录的非结构化的文本"疾病诊断名称"转化为分类编码的过程. 通过疾病编码过程, 疾病种类数据能以标准化结构化的形式存储, 进而应用于各种数据分析场合.

在实际研究中我们发现, 部分电子诊疗记录中并没有记录疾病编码. 究其原因有以下两点: 1)国内普遍采用的人工编码形式效率低下, 而且准确率受到编码员业务水平等因素的影响[5]. 2)编码量庞大. 一方面实时产生的诊疗记录数量多, 如安徽省某三甲医院日诊疗人次已达六千以上; 另一方面不同医疗机构在不同历史时期的信息化程度不一、编码标准不一致, 导致部分历史病案未予编码或需重新编码[6].

针对上述问题, 本文提出基于文本分析的自动化疾病编码方法. 首先建立 ICD 的文本模型将其表示为文本集, 每个 ICD 疾病类别对应一个反应其特征的文本. 对于待编码的疾病诊断名称, 取与其关联度最高

① 基金项目:高等学校博士学科点专项科研基金新教师类资助课题(20113402120026);安徽省自然科学基金(1208085QF112);安徽省高等学校优秀青年人才基金(2012SQRL001ZD);中央高校基本科研业务费专项资金(WK2101020004,WK0110000007)

收稿时间:2015-03-19;收到修改稿时间:2015-05-12

的 ICD 疾病类别文本所对应的编码作为 ICD 编码.

1 相关工作

因为疾病编码工作量大、人工编码效率低, 所以实现编码自动化是有必要的. 然而鲜有相关的研究成果被发表. 原因在于医疗数据难以获取, 主要涉及:信息化的不完善导致数据录入不规范不完整; 医疗数据的使用因涉及到患者个人隐私受到隐私保护相关法规的限制; 自动化疾病编码因其具有统计学特征主要用于大数据分析的预处理过程, 不能完全取代主要用于医疗业务的人工编码. 有学者研究了“计算机辅助编码系统”[7,8], 编码人员借助计算机辅助编码系统通过检索、匹配疾病关键词等进行疾病编码. 这种编码方式效率要高于纯人工编码方式, 但其本质上依然没有脱离人工的参与, 仍然存在效率低、准确率依赖于编码员业务水平等缺点. 本文所提出的自动化疾病编码方法, 借助文本分析技术, 实现了疾病编码的完全自动化.

文本分析是从文本中获取高质量信息的过程, 是信息检索、数据挖掘、机器学习等学科中的重要研究领域, 现已有许多成熟的文本分析方法成功应用于各种应用场景[9]. 然而鉴于对文本表示模型的需要, 文本分析不能直接应用于 ICD 编码, 本文提出一种转换方法将 ICD 建模为结构化的文本模型, 使得现有的文本分析方法能够应用到 ICD 编码过程.

2 将ICD表示为文本集

ICD 是国际疾病分类(International Classification of Diseases)的简称, 是 WHO(世界卫生组织)发布的疾病分类标准. 同时它也是我国有关疾病与代码的国家标准[10]. 根据卫生部要求, 自 2002 年 1 月起, 国内各医院开始采用 ICD 第 10 版进行疾病分类工作.

ICD-10 编码采用“字母数字编码”形式的 3 位代码、4 位代码、6 位代码表示. ICD 包含多个章. 每个章包含多个块, 如块 K35-K38(阑尾疾病), 每个块细分为用 3 位编码表示的类目, 如K35(急性阑尾炎). 类目细分为4 位编码表示的亚目, 如 K35.1(急性阑尾炎伴腹膜脓肿). 类目以下还划分为6 位扩展码. 类目及其以上的分类层次亚目、块、章都具有统计分类意义[10].

2.1 文本建模方法

将 ICD 表示为文本集, 是对于 ICD 划分的各个疾病类别分别构建一个反映其特征的文本. 具体地, 本文选择 ICD 在亚目层次(4 位编码)的疾病分类标准, 并使用向量空间模型来表示文本. 向量空间模型由 G.Salton 等人于上世纪 60 年代末提出, 是目前最为成熟和应用最为广泛的文本表示模型之一[11]. 以下为具体的文本建模方法:

将亚目的集合记作 D. 将一个亚目包含的 6 位码对应的词集的并集称作该亚目的词集. 例如, 亚目 J35.1 包含 6 位码 J35.101(“扁桃体肥大”)和 J35.102(“扁桃体增生”), 它们分别对应的词集[“扁桃体”, “肥大”]和[“扁桃体”, “增生”]的并集[“扁桃体”, “肥大”, “增生”]即为亚目 J35.1 的词集. 所有的亚目的词集的并集记作 T. 用记号 DT 表示 D 为行、T 为列构成的矩阵, 矩阵单元 DT_{ij} 的值为词项 T_j 在亚目 D_i 中的权重.

若将 $d \in D$ 视作一个文本, D 视作文本集, 则矩阵 DT 可被视为一个文本-词项矩阵(document-term matrix). 文本-词项矩阵是文本分析领域常用的文本结构化表示方法, 它描述了词项在文本集中的出现频率, 许多现有的文本分析方法都以其为分析对象[9]. 因此除了本文的应用场景, 文本-词项矩阵的构建也使得众多的成熟文本分析方法能够应用在疾病文本数据上.

以向量空间模型表示文本, 由矩阵 DT 的第 i 行构成的向量, 即为文本 D_i 的特征向量:

$$\overrightarrow{D_i} = (DT_{i1}, DT_{i2}, DT_{i3}...)$$

其中权重 DT_{ij} 的大小表示 T_j 词项在亚目 D_i 表示的疾病种类中的重要性. 权重常用的计算方法为 TF-IDF, 后文将详细介绍. 至此完成对 ICD 各个亚目的文本

3 自动化疾病编码

自动化疾病编码主要包含 ICD 文本建模及初始化和文本相关性度量两个模块, 具体流程见图 1. 以下分别介绍各个模块的功能及实现.

3.1 ICD 文本建模及初始化

对 ICD 的文本建模及初始化共包括中文分词、构建文本-词项矩阵、去停用词三个部分.

3.1.1 中文分词

对于词项非天然划分的中文文本, 首先要经历一个分词的过程, 即将汉字序列切分成一个个单独的词. 现有一些成熟的分词器能够完成对中文文本的自动分词. 我们采用中国科学院技术研究所开发的汉语词法分析系统 ICTCLAS, ICTCLAS 包括中文分词、词性标注等功能, 是当前世界上最好的汉语词法分析器. 分词的对象包括所有的 6 位码的标准疾病名称和待分类的疾病诊断名称. 分词后根据词性过滤掉符号、字母等无意义词.

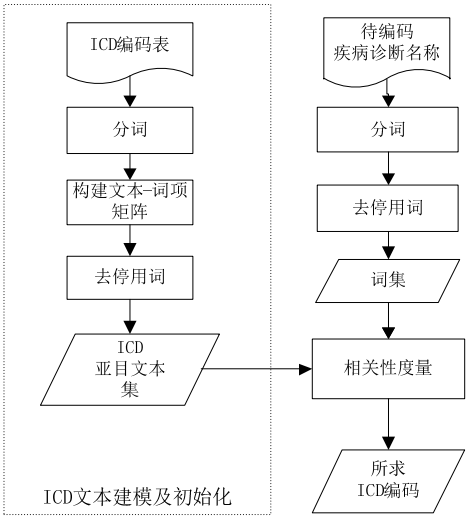

图 1 自动化疾病编码流程

3.1.2 构建文本-词项矩阵

采用上文阐述的文本建模方法, 经过对标准疾病名称的分词, 完成文本-词项矩阵 *DT* 的构建. 接下来是对矩阵单元 DT_{ij} 值的计算, 亦即文本 D_i 中词项 T_j 权值的计算.

TF-IDF(term frequency–inverse document frequency)是一种用于信息检索与文本挖掘的常用加权技术, 用以评估一个词项对于一个文本的重要程度[12]. TF-IDF 的主要思想是: 如果某个词在一篇文章中出现的频率高, 并且在其它文章中很少出现, 则认为此词具有很好的类别区分能力. 词项 T_j 在文本 D_i 中的 TF-IDF 值的计算方法为:

$$TF-IDF_{i,j}=TF_{i,j}\times IDF_j$$

其中, TF 为词频(Term Frequency), 词在文章中出现的频率; IDF 为逆文档频率(Inverse Document Frequency), 是一个词语在整个文本集中提供的信息量的度量. $TF_{i,j}$ 为词项 T_j 在文本 D_i 中的 TF 值, IDF_j 为词项 T_j 在文本集中的 IDF 值, 计算方法如下:

$$TF_{i,j}=\frac{n_{i,j}}{m_i}$$

$$IDF_j=\log\frac{|D|}{\left|\left\{i:T_j\in D_i\right\}\right|}$$

以上式子中, $n_{i,j}$ 为词项 T_j 在文档 D_i 中出现的次数, m_i 为 D_i 所表示亚目包含的六位码个数; $|D|$ 为文本总数, $\left|\left\{i:T_j\in D_i\right\}\right|$ 为包含词语 T_j 的文本数目.

3.1.3 去停用词

文本中存在一些出现频率很高的没有实际意义的词语, 如汉语中的“了”“的”“呢”“是”等, 它们称为停用词(stop words), 这些词应被视为噪声予以去除[13]. 目前已有一些通用的中文停用词表, 但考虑到疾病名称包含的词数较少而且医疗文本具有一定的专业性, 故选择直接从文本集 *D* 中筛选停用词. 具体地, 使用前面提到的IDF度量来识别停用词. IDF度量反映了各个词项在文本集中的普遍性, 一个词的IDF值较低表明其在大量文本中出现倾向于作为停用词. 表 1 列出了筛选出的部分停用词:

表 1 部分停用词列表

停用词	IDF	停用词	IDF
性	2.10	状	5.59
病	2.72	中	5.60
的	3.59	外	5.61
和	4.20	伤	5.66
炎	4.45	小	5.82

从表 1 中看出, 停用词大致分为两类, 一类是汉语中的通用停用词如“的”“和”等, 另一类为医疗术语中的专用停用词, 它们对疾病种类的划分区分度较小, 如: “炎”“病”等.

经过以上处理过程, 构建出一个带权的文本-词项矩阵 DT, 矩阵的每一行都对应一个亚目的文本, 文本 D_i 的特征向量为

$$\overrightarrow{D_i}=(DT_{i1},DT_{i2},DT_{i3}...)$$

至此形成了包含每个亚目对应文本的亚目文本集.

3.2 相关性度量

假设一个待分类的疾病诊断名称 *diag*, 分词和去停用词后生成词集 *W*:

$$\mathrm{W}=\{\mathrm{w}_1,w_2,\cdots\}.$$

在上文生成的 *DT* 矩阵中, 列即词集 *T* 为:

$$\mathrm{T}=\{\mathrm{T}_1,T_2,\cdots\}.$$

文本 D_i 的特征向量为:

$$\overrightarrow{D_i}=(DT_{i1},DT_{i2},...)$$

那么词语集合 *W* 与文本 D_i 的相关性的计算方法如下: 假设词集 *T* 中的词在词集 *W* 中有, 则它们在文本 D_i 中对应的权值的和,即为 *W* 与 D_i 的相关性:

$$rel(W,D_i)=\sum_{1\le j\le n}a_jDT_{ij}$$

其中: $a_j=\begin{cases}1 & if\ T_j\in W\\ 0 & otherwise\end{cases}$

根据上式计算 *W* 与文本集 *D* 中各个文本的相关性, 取与其相关性最高的文本 $d\in D$ 所对应的疾病编码作为疾病诊断名称 *diag* 的编码.

4 实验结果与分析

本文实验环境的配置为64位Windows 7操作系统, Intel Core i5-4440处理器, 8G内存, Oracle数据库. 在Eclipse开发平台上使用Java语言实现的上述基于文本相关性的自动化编码方法.

实验选取了安徽省某市城镇居民医疗保险记录的从2007至2014年的住院数据. 经数据预处理, 共筛选出264012条住院记录. 其中有68823条(26%)已完成人工编码, 剩余195189条(74%)待编码.

对已编码的68823条住院记录进行疾病编码, 将其结果与人工编码进行比对, 准确率如表2:

表2 编码准确率

分类层次	亚目	类目	块	章
准确率	79.23%	84.90%	85.14%	85.31%

在亚目层次的准确率为79.23%, 随着将编码结果转换到更抽象的分层层次, 准确率也随之升高, 尤其是从亚目提升到类目层次, 准确率提高了6.67%. 主要原因在于属于同个类目的亚目的疾病名称都非常接近. 在涉及疾病数据的数据分析中, 可根据具体应用的需求, 在疾病种类抽象层次和准确率中折衷选择适宜的分类层次.

随后将算法应用到未分类的264012条住院记录. 平均耗时209秒, 速率约为1250例/秒, 全安徽省每天新增入院病例在16秒内即可完成分类. 所以可应用于某些实时分析应用.

将得出的结果与《2013 中国卫生统计年鉴》[2]公布的疾病分布进行了比较, 详见图2. 从中看出, 实验结果与全国卫生局公布的统计信息基本一致. 两组数据的相关系数为0.953, 在0.01水平(单侧)上显着相关. 验证了本文自动化疾病编码方法的统计有效性.

5 结语

本文提出一种基于文本分析的自动化疾病编码方法. 具体地, 我们提出一种文本建模方法将ICD表示为文本集; 然后借助文本相关性度量, 获取与待编码疾病诊断名称最相关的ICD编码. 经实验验证, 本文提出的自动化疾病编码方法是有效的, 在准确率、效率等方面都较为优秀, 可广泛应用于各种类型的数据分析场景. 本文采用的是基于词频向量的文本建模方法, 没有考虑文本中词项的语义信息, 下一步研究中可考虑加入语义分析, 提高编码准确率.

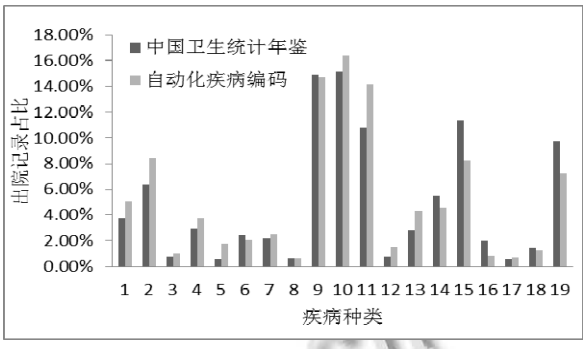

图2 疾病种类分布对比

参考文献

1 Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity. 2011.

2 World health statistics. WHO Library Cataloguing-in- Publication Data, 2011.

3 Agresti A. Categorical Data Analysis, chapter 5. Wiley-Interscience, Hoboken, New Jersey, second edition, 2002.

4 丰玉荣,陈俐,冯洁,邹文,王辉.对国际疾病分类编码工作重要性的再认识.中国病案,2013,14(4):4–5.

5 耿连华.疾病编码的质量分析与对策.中国病案,2012,13(8); 31–32.

6 王晓丹.当前医疗信息化存在的问题及对策研究.医学信息学杂志,2011,32(1):44–47.

7 陈子星,罗丽妮,李开祥.计算机辅助疾病分类编码可以提高编码效率.中国卫生信息管理杂志,2010(3):69–71.

8 温赟.疾病分类系统的研究与应用[硕士学位论文].北京:清华大学,2012.

9 Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge: Cambridge university press, 2008.

10 2001 G B T. 疾病分类与代码.

11 Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613–620.

12 Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information processing & management, 1988, 24(5): 513–523.

13 Van Rijsbergen CJ. Information retrieval. Butterworths Scientific Publication, London, 1975.