

基于特征融合的产科多标记辅助诊断研究

马鸿超^{1, 2}, 张坤丽¹, 赵悦淑³, 咎红英¹, 庄雷¹

(1. 郑州大学信息工程学院 郑州 450001;

2. 郑州大学产业技术研究院 郑州 450001;

3. 郑州大学第三附属医院 郑州 450052)

摘要: 中文产科电子病历中蕴含着大量的医疗知识和健康信息, 针对电子病历的信息抽取及辅助诊断对提高人口的生育健康水平具有重要意义。电子病历中首次病程记录的入院诊断是根据主诉、辅助检查、查体等信息得出的, 通常情况下诊断中包含正常诊断、病理诊断及并发症, 非单一结果, 因此该文将辅助诊断问题转化为多标记分类任务。在对产科电子病历首次病程记录进行数据清洗和结构化的基础上, 规范化诊断结论, 将 LDA 所抽取的文本特征与病历中的数字特征采用向量拼接的方法融合为新的特征, 按诊断结果出现的频次不同形成不同的多标记集, 根据首次病程中部分信息进行辅助诊断, 采用 RAKEL、MLkNN、CC 和 BP-MLL 方法进行多标记分类。实验结果表明采用融合特征的多标记分类方法能够提升中文产科电子病历辅助诊断的效果。

关键字: 中文产科电子病历; 数据清洗; 辅助诊断; 特征融合; 多标记分类

The Study of Multi-label Auxiliary Diagnosis of Obstetrics Based on Feature Fusion

MA Hong-chao^{1, 2}, ZHANG Kun-li¹, ZHAO Yue-shu³, ZAN Hong-ying¹, Zhuang Lei¹

(1. Information Engineering School, Zhengzhou University, Zhengzhou 450000, China;

2. Industrial Technology Research Institute, Zhengzhou University, Zhengzhou 450000, China;

3. The Third Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, China)

Abstract: The Chinese obstetric EMRs contain massive amounts of medical knowledge and health information, and the information extraction and assistant diagnosis of obstetric EMRs is of great significance in improving the fertility level of the population. The admitting diagnosis in first course record of EMR is reasoned from the information which includes chief complaints, auxiliary examinations, physical examinations etc. Inspired by this, we transform the diagnostic process into multi-label classification problem. The diagnostic conclusion is standardized on the basis of data cleaning and structuring, while the features of LDA extraction and the digital features of medical records are fused into new features by vector merging. According to the diagnosis results of different frequency to form different multi-label set, we make the auxiliary diagnosis more effective with the first part of the course of information. RAKEL, MLkNN, CC and BP-MLL are used for multi-label classification. The experimental results show that the method of multi-label classification with fusion feature can be used to improve the auxiliary diagnosis of Chinese obstetric electronic medical records.

Key words: Chinese obstetric electronic medical record; data cleaning; assistant diagnosis; features fusion; multi-label classification

基金项目: 本文承 973 课题 (2014CB340504), 国家自然科学基金项目 (61402419, 60970083), 国家社会科学基金项目 (14BY096), 计算语言学教育部重点实验室开放课题项目, 河南省科技厅基础研究项目 (142300410231, 142300410308), 河南省科技厅科技攻关项目 (172102210478) 资助。

作者简介: 马鸿超 (1990—), 男, 河南开封人, 主要研究领域为自然语言处理, E-mail: ma-hc@foxmail.com; 通讯作者: 张坤丽 (1977—), 女, 河南巩义人, 讲师, 博士生, 主要研究领域为自然语言处理、语言资源构建等, E-mail: ieklzhang@zzu.edu.cn

1 引言

自计划生育写入我国基本国策之后,晚婚晚育政策带来诸多益处的同时,也导致年龄超过 35 岁的高龄孕妇所占比例逐年增加^[1]。2016 年“全面二胎”政策实施之后,预计我国将迎来一个新的生育高潮,而高龄孕妇所占比重会更大。针对高龄产妇,难产、胎儿畸形的发生率及并发症风险都有所增加,这对各医疗机构的产科是一个巨大挑战。自 2010 年国家卫生计生委医政医管局出台《电子病历基本规范(试行)》^[2]之后,各诊疗机构累积了丰富的产科电子病历资源,海量电子病历数据是医疗领域的大数据,蕴含着大量医疗知识和健康信息。如何利用这些资源实现临床信息决策支持,从而改善临床治疗效果,是迫在眉睫的研究任务。

电子病历是医务人员对医疗活动进行的详细记录,其中非常重要的形式是自由文本(半结构或无结构)数据^[3],如何采用自然语言处理技术对电子病历进行结构化和信息抽取是充分利用电子病历所蕴藏知识的重要一步。随着人工智能技术的发展,让医疗辅助诊断成为了可能。在电子病历中,首次病程记录以文本形式存储,包括患者的主诉、查体以及辅助检查等信息,而通常情况下,产科的入院诊断包括正常诊断、病理诊断及并发症等,即并非单一结果。如一份电子病历中的诊断中可能有“羊水过多”和“妊娠期高血压”,若将一个电子病历看作一个实例,诊断结果看作标记,则可认为每个实例都可能属于多个标记。可以将针对产科电子病历的辅助诊断问题转化为机器学习中的多标记分类问题,而同一份病历中的多个诊断结果即为不同的标记。

本文在分析中文产科电子病历自由文本结构及内容的基础上,对首次病程记录进行数据清洗,并规范化诊断结论,根据已收集到的电子病历中首次病程记录的主诉、入院查体、产科检查和辅助检查等信息,将 LDA(Latent Dirichlet Allocation)主题模型抽取的特征与生理参数数值型特征融合,采用多种多标记分类的方法对产科电子病历进

行自动诊断。

2 相关工作

多标记分类问题即实例空间中的每一个实例可以属于多个标记,即设实例空间为 X ,实例集合为 D , L 为标记集合, $|L|=n$ 。在单标记中只有一个标记 l_i 与实例 x_i 相对应,其中 $l_i \in L$, x_i 是集合 D 的第 i 个实例,集合 D 可表示为 $\{(x_1, l_1), (x_2, l_2), \dots, (x_n, l_n)\}$ 。而在多标记中,每一个实例可以属于多个标记,即实例集合 D 可以表示为 $\{(x_1, L_1), (x_2, L_2), \dots, (x_n, L_n)\}$,其中 $L_i \subseteq L$, L_i 是 x_i 关联的标记集合。不少学者对多标记分类进行研究,分别应用于文本分类^[4],情感分类^[5],图像和视频分类^[6],生物信息^[7]和医学^[8]等领域。

目前针对多标记学习的研究主要集中在三个方面:第一是改进或者提出新分类或排序的模型,杨小健等^[9]提出了一种结合类别权重及多示例的多标记学习改进算法(CWMI-INSDF),在对单一对象拆分时,充分考虑数据的内部特性,加入权重函数和自适应惩罚策略;李哲等^[10]改进了分类器链,形成了有序分类器链(Ordered Classifier Chain, OCC)方法,OCC 可以有效利用各个标记之间的依赖关系。第二,改进、提出新的特征选择模型或者联合模型,李峰等^[11]提出了一种局域互信息的粒化特征加权多标签学习 k 近邻算法 GFWML-kNN(Granular feature weighed k-nearest neighbors algorithm for multi-label leaning),该算法计算了每个特征的权重系数,并把标记间的相关性融合进行特征的权重系数,解决了特征相关性和组合爆炸问题,实验表明,该算法在整体上取得了较好的效果;Teisseyre, P^[12]把分类器链和强行网络正则化相结合,提出了 CCnet 算法,CCnet 最重要的优点是在学习过程中能选择相关特征。第三,在新的领域应用多标记学习,Liu G P 等^[13]把技术应用到中文冠心病数据集上,进行多症状选择。

在医学领域,也有学者采用多标记分类方法进行数据的处理。Ira Goldstein 等^[8]利

用 I2B2 2008 的数据, 为每一个类别训练一个分类器, 对肥胖症及其他 15 种并发症进行多标记分类; Shao H 等^[14]提出了一种混合优化的特征选择算法 (Hybrid Optimization based Multi-Label, HOML) 用于特征选择, HOML 把模拟退火算法、遗传算法和爬山贪婪算法进行结合, 在中医冠心病数据集上显著提高了分类的效果; Li Y 等^[15]把中药方剂与中医证候关系转化为一个多实例学习和多标记学习问题, 提出了 WSSH-MIML (Weighted Sampling based on Similar Herbs MIML) 的方法, 用于预测基于多实例多标记框架处方的主要症状; 徐玮斐等^[16]把随机森林与多标记学习算法相结合对慢性胃炎的实证症状进行选择 and 模型构建; 程波等^[17]提出一种基于多模态特征数据的多标记迁移学习模型, 并将其应用于早期阿尔茨海默病诊断: 首先对图像提取特征, 再选择最优特征子集, 最后用多标记迁移学习模型分类回归器进行分类。

目前在医学领域的研究多在公开数据集或类别较少的真实数据集上展开, 如文献 [8] 中采用的是公开的评测数据。产科诊断种类较为复杂, 且一些特征不易直接提取, 这也为开展中文产科电子病历研究带来了一定困难, 到目前为止, 还鲜有人对较复杂的中文产科电子病历进行辅助诊断的研究。

3 产科电子病历特点及数据集

3.1 结构与特点

产科电子病历多为自由文本形式或半结构化文本形式, 结合具体的电子病历, 对其结构及特点进行了分析。重点对首次病程记录的内容及特点进行了分析。产科电子病历主要包括病程记录和出院小结两部分内容, 具体如表 1 所示。在电子病历文本中, 所有内容混排在一起, 为了便于数据分析, 将首次病程记录按主诉、入院查体、产科检查、辅助检查、入院诊断、诊断依据、鉴别诊断和诊疗计划等进行结构化, 形成了本文

进行实验的首次病程记录集合 (图 1 则是按小节内容整理后的形式)。

在电子病历中由于医院管理信息 (Hospital Information System, HIS) 系统的原因, 存在首次病程记录冗余 (如首次病程记录和出院小结的重复)、缺失 (如首次病程记录和出院小结的缺失)、时序错乱 (时间出现逻辑错误) 等问题, 因此也对数据进行了清洗。

表 1 病程记录和首次病程记录所包含的主要内容

名称	主要包含内容
病程记录	首次病程记录、日常病程记录 (也称查房记录)、上级医师查房记录、出院小结
首次病程记录	记录时间、主诉、入院查体、产科检查、辅助检查、入院诊断、诊断依据、鉴别诊断和诊疗计划

3.2 实验数据集

本文以 15 家医院随机抽取的 1 万余份产科电子病历为研究对象, 在抽取电子病历的同时进行去隐私化预处理, 隐去了包括病人的姓名、身份证号码、家庭住址、医生姓名等涉及病人及医生的隐私性信息。

针对电子病历中存在的不同问题, 采取不同的方式进行数据清洗。针对首次病程记录冗余, 采用自动比对的方式进行筛选, 当检测到同一个病历中有多个首次病程记录时, 根据信息的完整性以及记录时间, 甄选出正确的首次病程记录; 针对首次病程记录的缺失, 则直接删除; 而针对首次病程中的时序错误, 根据产科治疗的时序逻辑, 设计出了时序错误检测方案, 对这一类首次病程记录也进行了删除。最终得到包含 10, 886 份首次病程记录的数据集。

在首次病程记录中, 入院诊断是医生依据病人的各项生理指标对病人的情况综合分析后作出的诊断, 然后给出诊疗计划。因此, 首次病程记录的入院诊断可以看作是依据显式或隐式的特征进行多个诊断标记的分类。本文的辅助诊断模拟这一过程, 利用

主诉: NAME, 女, 34 岁, 以“停经 8 月余, 发现血压偏高一天。”为主诉入院。该孕妇平素月经规律, LMP2014.5.1, EDC2015.2.8.停经 35 天自测尿 HCG 阳性。停经 1 月余行 B 超检查诊断为宫内早孕。停经 60 天出现恶心、呕吐等早孕反应。……

入院查体: T:36.0℃, P:76 次/分, R:19 次/分, BP:154/110mmHg 发育正常, 营养中等, 神志清, 精神可, 步入病房, 自主体位, 查体合作。全身皮肤粘膜红润无黄染、皮疹、出血点, 未触及肿大的浅表淋巴结。……

产科检查:骨盆外测量 IS:25.0cmIC:27.0cmEC:20.0cmTO:9.0cm。宫高 32.0cm 腹围 100.0cm 先露头, 胎位枕左, 未衔接胎心 146 次/分胎儿估重 3400g。无宫缩。肛诊: 未查。

辅助检查: 胎儿彩超(本院 2014.12.30): BPD:86.0mmFL:60.0mmAFI:133.0mmEFW:1793g 胎方位: ……。

入院诊断: 1.重度子痫前期 2.宫内孕 34+5 周 3.孕 2 产 14.胎儿宫内生长受限 5.头位 6.脐绕颈二周 7.妊娠合并子宫肌瘤。

诊断依据: 收缩压 ≥ 160 mmHg 和/或舒张压 ≥ 110 mmHg, 尿蛋白(+++)以上, 或者 $\geq 2g/24h$, 或者出现系统并发症。

鉴别诊断: 1.慢性高血压合并妊娠: 妊娠前或者妊娠 20 周前发现血压升高, 血压 $\geq 140/90$ mmHg, 尿蛋白(-)。2.……

诊疗计划: 1、左侧卧位, 间断吸氧, 解痉、镇静、控制血压的治疗。2、……

图 1 首次病程记录示例

主诉、入院查体、产科检查和辅助检查来预测入院诊断。抽取了其中的主诉、入院查体、产科检查、辅助检查和入院诊断作为实验数据集, 把入院诊断看作标记, 其余四部分看作特征。在分词工具 ICTCLAS¹中增加来自互联网及《妇产科学》^[18]中的妇产科医疗术语和药物名称对实验特征文本进行分词, 对实验文本进行分词作为实验数据集。入院诊断中的每一条看作是一个标记, 其中的“孕 X+Y 周”及“孕 Z 产 U”是计算或主诉的结果, 不作为类标记, 将其他诊断作为多标记

从数据集中共得到诊断 737 个, 形成集合 L_1 。由于本文中所使用的电子病历来源于多家诊疗结构, 医生书写习惯并不相同, 如在 L_1 中出现的“胎盘前置状态”及“前置胎盘”两种写法, 以及“上感”与“上呼吸道感染”这样的写法, 即存在同一类别标记的多种不同表现形式, 以 ICD10 疾病命名规范为依据, 对诊断结果进行分词之后, 采用基于语义的方法对标记的相似度进行计算², 计算方法如式 (1) 所示, 其中, S_s 是两个诊断标记的语义向量表示。

$$S_s = \frac{S_1 \times S_2}{\|S_1\| \times \|S_2\|} \quad (1)$$

根据相似度计算的结果, 由医学专业人员对类标记进行规范化, 合并了同一诊断结果的不同表述形式, 得到类标记集合 L_2 , 包含标记 249 个, 出现频次统计结果如表 2 所

示。其中仅出现 1 次的诊断标记有 84 个, 占标记总数的 34%, 在数据集中共出现诊断标记 25,881 次, 首次病程中最少标记数为 1, 最大标记数为 9, 平均每份电子病历出现 2.38 次。规范化之后把标记转化为 0, 1 向量, 出现为 1, 否则为 0。

表 2 入院诊断频率分布情况

出现的频次	数目	所占比例
1	84	34%
2-10	96	39%
11-50	33	13%
50-100	10	4%
>100	26	10%

4 多标记辅助诊断方法

4.1 特征融合

LDA 可以提取文本特征, 而不能有效的表达数值特征, 所以把 LDA 提取的文本特征与数值型特征融合。以下分别介绍 LDA 特征、数值型特征及相应的融合方式。

4.1.1 LDA 特征

LDA 由 Blei 等^[19]在 2003 年提出, 已经被广泛的应用到特征提取上。LDA 是一个三层的贝叶斯模型, 第一层是文档集合层 D,

¹<https://github.com/NLPIR-team/NLPIR>

²<https://my.oschina.net/twosnail/blog/370744#comment-list>

每个文档 d ($d \in D$) 由 $K(k_1, k_2, \dots, k_m)$ 个主题构成, 每个主题 k_i 由第三层的 N 个特征词构成, 文本中的词是可观察的, 而主题则是隐含的, LDA 模型如图 2 所示。模型由参数 (α, β) 确定, α 反映了隐含主题之间的相对强弱, β 反映所有隐含主题自身的分布。首先, LDA 从参数为 β 的 Dirichlet 分布中抽取主题与单词的关系 ϕ , 而后 LDA 生成一个文本: 首先从参数为 α 的 Dirichlet 分布中抽样出该文本 d 与各个主题之间的关系 θ_d ; 其次, 从参数为 θ_d 的多项式分布中抽样出当前所属的主题 z ; 最后从参数为 ϕ 的多项式分布中抽取具体的单词 w 。一个文本中所有单词与其所属主题联合概率分布如式 (2) 所示。

$$P(w, z | \alpha, \beta) = P(w|z, \beta)P(z|\alpha) \int P(z|\theta)P(\theta|\alpha)d\theta \int P(w|z, \phi)P(\phi|\beta)d\phi \quad (2)$$

0.00530035335689046	0.0017667844522968198	...	0.00530035335689046
0.021551724137931036	0.004310344827586207	...	0.0021551724137931034
0.008880994671403197	0.0053285968028419185	...	0.014209591474245116

图 3 M 矩阵示意图

4.1.2 数值型特征

数值是电子病历中很重要的指标, 不能忽略, 需单独考虑。在首次病程记录中, 孕妇年龄、停经月数、宫高腹围等 16 个生理指标是影响判断的重要因素, 因此需要提取这些生理参数。文中有一些数值单位不统一的情况, 以停经月数为例, 一般描述为“停经 X 月余”, 但也“停经 Y 周”等的出现, 在抽取时, 我们把周转换为月, “四周”近似为“一个月”, 在抽取时需要充分考虑这些问题, 并换算单位。同时也要兼顾数据的真实有效性, 在抽取孕妇体重增加生理参数中, 其中一份, 孕妇体重增加为“158kg”, 可以推测这一数据可能是错误数据, 抽取时需要设定一个阈值来判定数据的准确性。对于可能错误的的数据, 直接删除不用, 因为可能出现错误的的数据会影响实验的结果。表 3

0.00530035335689046	0.0017667844522968198	...	0.00530035335689046	35	...	100
0.021551724137931036	0.004310344827586207	...	0.0021551724137931034	27	...	92
0.008880994671403197	0.0053285968028419185	...	0.014209591474245116	24	...	98

图 4 M 矩阵示意图

LDA 的输出包含文档-主题概率分布矩阵 $M[m \times n]$, m 表示文档总数 D , 一行代表一个文档 (在本文中为一个病历实例), n 是主题数, 每一列表示一个主题。LDA 的主题数目 $K=n$ 表示一个实例的 K 个特征, 每一个元素代表文档属于该主题的概率。假设 $D=3, K=120$, 图 3 是 LDA 的输出的一个实例 $M[3 \times 120]$ 。但 LDA 训练时是以词语为单位, 并且高频的全局性词对主题的贡献概率大于低频词的贡献概率, 而对于每一份病例中的检查指标都是自己独有的, 频率较低, LDA 无法有效表达数值特征。所以引入了数值型特征, 以弥补 LDA 的不足。

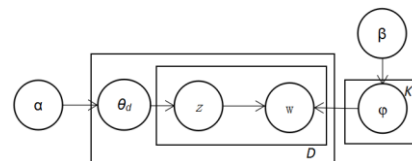


图 2 LDA 模型示意图

是抽取的生理参数的数值型特征。

表 3 数值型特征示例

年龄: 35	停经月数: 8	体温: 36	IS: 25
早孕反应: 60	孕体增加: 17	脉搏: 76	心率: 19
收缩压: 154	舒张压: 110	T0: 9	IC: 27
EC: 20	HCG 阳性: 35	宫高: 32	腹围: 100

4.1.3 特征融合

由上文可知, 产科电子病历中既有描述性的文本特征 (主诉和入院查体的描述性文本) 也有生理参数的数值型特征 (年龄, 体温和宫高等)。只用数值型特征又无法表达描述性的文本特征, 只用文本特征又无法表达数值型特征。因此我们把数值型特征单独抽取, 与 LDA 特征向量进行拼接, 进行融合, 即为分类的最终特征。图 4 是拼接之后的示意图, 矩阵 $M[3 \times 136]$ 。

4.2 计算方法

本文采用 BP-MLL^[7] 进行多标记分类, 这种方法是在传统的多层反向传播神经的基础上, 改变原有的误差函数。为了将适用于单标记示例的传统前馈神经网络应用到多标记示例, 必须设计一些特定的误差函数, 而不是简单的平方和函数来捕捉多标记学习的特征, 并且使误差函数最小化。公式(3)所示误差函数认为每个类标记是独立的, 没有考虑每一个标记之间的关系, 公式(4)所示误差函数侧重于网络在属于一个实例的标记上的输出与不属于其标记之间的差异。公式(4)中误差函数的最小化, 可以使系统, 对于属于训练实例的那些标记, 输出较大的值, 对于不属于训练实例的标记, 输出较小的值。

$$E_i = \sum_{j=1}^Q (c_j^i - d_j^i)^2 \quad (3)$$

$$E = \sum_{i=1}^n E_i = \sum_{i=1}^n \frac{1}{|Y_i|} \sum_{(k,l) \in Y_i \times \bar{Y}_i} \exp(-(c_k^i - d_l^i)) \quad (4)$$

4.3 评价指标

实验采用多标记常用的 Hamming loss, One Error, Coverage, Ranking loss, Average precision 作为评价指标^[20]。

Hamming loss: 该指标用于评估样本的真实标记与系统预测所得标记之间的误差率, 即示例具有标记 Y_i 但未被识别出, 或不具有标记 Y_i 却被误判的可能性。

One-error: 该评价指标用于考察在样本的类别标记排序序列中, 排名最高的标记不是样本真实标记的可能性, 在单标记学习中, 就演化成一般的分类错误率。

Coverage: 该评价指标考察了在样本的类别标记排序队列中, 平均需要多少搜索深度才能覆盖样本所有相关标记。

Ranking loss: 该指标用于考察在样本的类别标记排序序列中出现排序错误的情况, 即样本对其所具有标记的排名低于对其所不具有标记的排名的可能性。

Average precision: 该评价指标考察了在样本的类别标记排序队列中, 隶属度值大的标记仍为其相关标记的情况, 即反映了预测

类标的平均精确度。

5 实验及结果分析

5.1 实验设置

本文选用的 BP-MLL 与常用的 RAKEL^[21]、MLkNN^[20] 和 Classifier Chain^[22] (CC) 分类方法进行比较, 分别考虑三个因素对实验结果的影响。首先, 如表 2 所示, 诊断标记出现频次不均衡, 且低频标记出现个数较多, 因此考察不同频次的类别标记集合对实验结果的影响; 其次, 实验中利用 LDA 提取特征时, 不同特征数目对实验结果也有影响, 本文也予以考察。因此本节设置了三组实验, 第一组设置 LDA 的主题数目 K 为 120, 比较不同规模标记数目对于分类性能的影响, 第二组是标记数目为 69 时, LDA 的不同主题数目对 average precision 的影响, 第三组是数值型特征对实验结果的影响。

5.2 类标记数目对实验结果的影响

首先选定 $L=249$, LDA 的主题数 $K=120$, 结果如表 4 所示, “↓”表示值越小效果越好, “↑”表示值越大效果越好(下同)。MLkNN 所选用的五个评价指标中结果最好, 而 BP-MLL 除 hamming loss 表现较差外, 在其余四个评价指标中也都有排到了第二位, 从整体来看 MLkNN 和 BP-MLL 都有绝对的优势。从表一中可以看出, 即使是表现最好的 BP-MLL 和 MLkNN 算法, 其 average precision 才为 0.6916 ± 0.0059 和 0.6259 ± 0.0092 。根据表 2 统计结果, 仅出现 1 次的诊断标记有 84 个, 频次在 2-10 的 96 个, 两者共计 180 个, 对这部分标记进行分析, 主要是有三种情况:

第一种情况是由于对电子病历并未做分类, 因此所抽取的标记是所有在产科办理住院的患者, 而有一些针对产科来讲是非典型的诊断结果, 如过敏性皮炎、腹泻、左上肢骨折、左下肢陈旧性血栓、类风湿关节炎、

抑郁症等。

第二种情况是由于医生对入院诊断的书写习惯不同造成的，个别医生会在正常诊断中写出在类标记中出现频率较低的标记，如“单胎妊娠胎”仅出现一次。

第三部分是较为少见的诊断结果，在抽取的 10,886 份电子病历中，仅出现一次的，如胎儿鼻骨缺失、胎儿十二指肠梗阻和抗磷脂抗体综合征等。这些标记在规模为 10,886 的数据集中仅出现一次，在一定程度上造成了数据的稀疏性。

将只出现一次的类标记删除，形成标记集 L_3 ，包含 165 个类标记，实验结果如表 5

表 4 L=249, K=120 时的实验结果

	RAKEL	MLkNN	CC	BP-MLL	
Hamming Loss	0.0075±0.0002	0.0069±0.0001	0.0082±0.0001	0.0334±0.0086	↓
Coverage	129.9043±3.9508	15.4746±0.9800	114.3751±4.4837	15.8939±1.0859	↓
One-Error	0.2767±0.0098	0.2148±0.0076	0.4472±0.0125	0.2354±0.0073	↓
Ranking Loss	0.2805±0.0100	0.0260±0.0019	0.2334±0.0127	0.0280±0.0017	↓
Average Precision	0.6088±0.0099	0.6916±0.0059	0.5415±0.0101	0.6259±0.0092	↑

表 5 L=165, K=120 时的实验结果

	RAKEL	MLkNN	CC	BP-MLL	
Hamming Loss	0.0106±0.0002	0.0104±0.0002	0.0125±0.0002	0.0374±0.0047	↓
Coverage	73.3000±1.6293	13.1090±0.6338	77.1609±2.7463	12.4803±0.7289	↓
One-Error	0.2548±0.0123	0.2151±0.0106	0.4512±0.0122	0.2353±0.0123	↓
Ranking Loss	0.2194±0.0043	0.0338±0.0021	0.2423±0.0116	0.0318±0.0023	↓
Average Precision	0.6513±0.0069	0.6932±0.0065	0.5396±0.0111	0.6497±0.0118	↑

表 6 L=69, K=120 时的实验结果

	RAKEL	MLkNN	CC	BP-MLL	
Hamming Loss	0.0243±0.0005	0.0243±0.0004	0.0294±0.0005	0.0440±0.0035	↓
Coverage	24.7648±0.8647	9.7269±0.3153	33.3615±1.2292	8.2251±0.6049	↓
One-Error	0.2257±0.0134	0.2149±0.0106	0.4524±0.0205	0.2342±0.0100	↓
Ranking Loss	0.1699±0.0062	0.0629±0.0028	0.2586±0.0084	0.0493±0.0043	↓
Average Precision	0.6907±0.0097	0.6984±0.0078	0.5504±0.0102	0.7136±0.0128	↑

5.3 主题特征数目对结果的影响

由于上文实验中选择的主题数目 K 为 120，所以本实验中的 K 在 120 上下选取，分别选择 100, 110, 130, 140，并与 K 等于 120 时进行对比。本实验旨在研究 LDA 的主题数目 K 对于分类结果的影响，以 average precision 为例，横坐标是不同的主题数目，

所示。可以看出，除了 RAKEL 在 average precision 上高于 BP-MLL，在整体性能上 MLkNN 和 BP-MLL 依然是表现最好的两个算法，且 average precision 值有所提高，且 BP-MLL 提高了两个百分点还多。

尝试进一步降低数据稀疏性，去除出现频率不大于 10 的标记，形成标记集 L_4 ，包含 69 个类标记，结果如表 6 所示。可以看出 BP-MLL 在 coverage、ranking loss 和 average precision 都排到了第一位，而 MLkNN 在整体性能上也有不错的表现，且 BP-MLL 的 Average precision 高达 0.7136。

纵坐标是各个方法在不同主题下的 average precision。从图 5 中可以看出，随着 LDA 的主题数目 K 不断增长然后到达最高点，而后下降，在主题数目为 110 左右或者 120 左右时，达到最高点。MLkNN 和 BP-MLL 整体效果均优于其他两种算法，而 BP-MLL 算法在整体上优于其他三种算法。

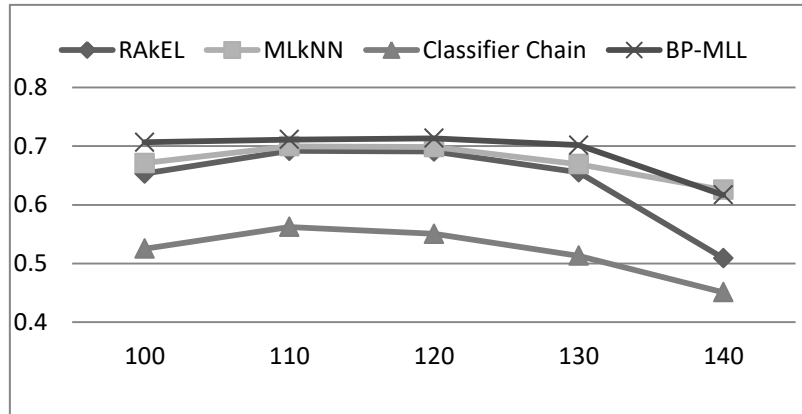


图 5 主题数目 K 的大小对四种算法的 Average precision 影响

5.4 数值型特征对结果的影响

为了研究数值型生理参数对实验的影响，提取了 16 个生理参数作为数值特征，把 16 个生理参数加入特征之中，与单独用 LDA 提取的特征作对比，在标记数目为 69 时，K 为 100, 110, 120, 130, 140 时实验，具体如图 6，表 7 所示。其中图 6 是上加入数值特征前后，在 BP-MLL 方法上 average precision 的对比，表 7 是另三种方法的对比。

加入生理参数的数值特征之后，average precision 均有不同程度的提高。在 BP-MLL 算法上，当 K 取 140，average precision 提高了将近十个百分点，当 K 取 120 时达到了最好的分类效果，其值为 0.7280。可以表明加入数值型生理参数之后，在一定程度上弥补了 LDA 不能表达数值特征的缺陷，有效提高了个分类结果。

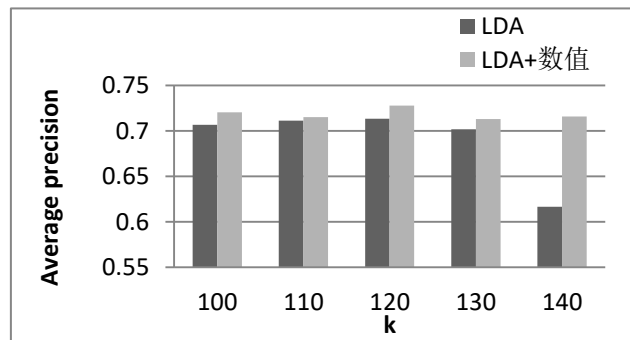


图 6 数值型特征对 BP-MLL 的 Average precision 影响

表 7 数值型特征对 Average precision 影响

		K=100	K=110	K=120	K=130	K=140
RAkEL	LDA	0.6532	0.6918	0.6907	0.6554	0.5089
	LDA+数值	0.6890	0.7220	0.7222	0.6900	0.6984
MLkNN	LDA	0.6712	0.6998	0.6984	0.6692	0.6255
	LDA+数值	0.6837	0.7093	0.7127	0.6789	0.6774
CC	LDA	0.5250	0.5622	0.5504	0.5131	0.4507
	LDA+数值	0.5513	0.5882	0.5818	0.5445	0.5475

6 总结与展望

本文在对中文产科电子病历分析的基础上,把辅助诊断问题转化为多标记分类问题任务。比较了四种算法的分类性能,并分别讨论了 LDA 主题数目,类标记集合的大小和不同的分类方法对实验的结果的影响。最后加入了提取的生理参数的数值型特征提高了各个算法的性能,当标记数目为 69, LDA 主题数目为 120 时, BP-MLL 算法取得了最好效果, average precision 达到了 0.7280 ± 0.0135 。而这样的结果可为医学院学生的学习提供辅助手段。以往的工作一般是利用公开的数据集或者比较规范、标记数目相对较少的数据集,本文是在真实的病历中抽出特征并进行诊断,不仅适用于产科电子病历,同样也适用于其他的病历。本文的工作为以后电子病历尤其是产科方面电子病历的辅助诊断研究,提供了参考和思路。

从实验结果可以看到,分类的 average precision 结果还有一定的提升空间。下一步的工作准备将多标记分类方法与产科医疗知识图谱进行融合,进行辅助诊断,进一步提高辅助诊断的精度及可靠性。

参考文献

- [1] 杨慧丽, 杨孜. 高龄妊娠对母胎结局的影响[J]. 中华产科急救电子杂志, 2016, 5(3):129-135.
- [2] 卫生部印发《电子病历基本规范(试行)》[J]. 中国病案, 2010, 11(3):64-65.
- [3] 杨锦锋, 于秋滨, 关毅, 等. 电子病历命名实体识别和实体关系抽取研究综述[J]. 自动化学报, 2014, 08:1537-1562
- [4] Schapire R E, Singer Y. BoosTexter: A boosting-based system for text categorization[J]. Machine learning, 2000, 39(2-3): 135-168.
- [5] Liu S M, Chen J H. A multi-label classification based approach for sentiment classification[J]. Expert Systems with Applications, 2015, 42(3): 1083-1093.
- [6] Wang C, Yan S, Zhang L, et al. Multi-label sparse coding for automatic image annotation[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 1643-1650.
- [7] Zhang M L, Zhou Z H. Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization[J]. IEEE Transactions on Knowledge & Data Engineering, 2006, 18(10):1338-1351.
- [8] Goldstein I, Uzuner Ö. Specializing for predicting obesity and its co-morbidities[J]. Journal of Biomedical Informatics, 2009, 42(5):873-886.
- [9] 杨小健, 王杉杉, 李荣雨. 一种结合类别权重及多示例的多标记学习改进算法[J]. 小型微型计算机系统, 2017, 38(4):857-862.
- [10] 李哲, 王志海, 何颖婧, 付彬. 一种启发式多标记分类器选择与排序策略[J]. 中文信息学报, 2013, (04):119-126.
- [11] 李峰, 苗夺谦, 张志飞, 等. 基于互信息的粒化特征加权多标签学习 k 近邻算法[J]. 计算机研究与发展, 2017, 54(5):1024-1035.
- [12] Teisseyre P. CCnet: Joint multi-label classification and feature selection using classifier chains and elastic net regularization[J]. Neurocomputing, 2017, 235:98-111.
- [13] Liu G P, Li G Z, Wang Y L, et al. Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning[J]. BMC complementary and alternative medicine, 2010, 10(1): 37.
- [14] Shao H, Li G Z, Liu G P, et al. Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine[J]. Science China Information Sciences, 2013, 56(5): 1-13.
- [15] Li Y, Li H, Wang Q, et al. Traditional Chinese Medicine formula evaluation using multi-instance multi-label framework[C]// IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2016:484-488.

- [16] 徐玮斐, 顾巍杰, 刘国萍, 等. 基于随机森林和多标记学习算法的慢性胃炎实证特征选择和证候分类识别研究[J]. 中国中医药信息杂志, 2016, 23(8):18-23.
- [17] 程波, 朱丙丽, 熊江. 基于多模态多标记迁移学习的早期阿尔茨海默病诊断[J]. 计算机应用, 2016, 36(8):2282-2286.
- [18] 谢幸, 苟文丽. 妇产科学. 第8版[M]. 北京: 人民卫生出版社, 2013.
- [19] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.
- [20] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern recognition, 2007, 40(7): 2038-2048.
- [21] Tsoumakas G, Katakis I, Vlahavas I. Random k-labelsets for multilabel classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(7): 1079-1089.
- [22] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification[J].