文章编号：2095-2163(2020)02-0034-06 中图分类号：TP181 文献标志码：A

一种基于树搜索的层次多标签乳腺疾病分类诊断方法

金程笑[1]，潘 乔[1]，张敬谊[2]，俞春儒[1]
（1 东华大学 计算机科学与技术学院，上海 201620；2 万达信息股份有限公司，上海 201112）

摘 要：随着医疗信息化的快速发展，医疗机构在临床诊断的过程中产生了大量的原始电子病历数据，存在着大量的可挖掘信息，作为临床的辅助诊断。由于乳腺疾病患者的患病情况较为复杂，同一位患者可能会患有多种相关疾病，每个大类疾病分类下可能会存在很多的小类疾病，而小类疾病分类下又可能存在更细粒度的疾病类别。传统的分类问题（如二分类和多标签分类）往往会忽略各标签之间存在的依赖关系并且分类算法输出数目呈指数级，占用空间过大，造成预测性能不佳。因此本文提出了一种基于树搜索的层次多标签乳腺疾病分类诊断方法，利用树结构可以充分考虑到标签集之间的层次结构的依赖关系，规范化诊断结论。按诊断结果之间的层次关系构建了层次多标签树，通过对标签树的路径搜索，最终实现乳腺疾病的多标签分类。

关键词：乳腺疾病；树搜索；层次多标签；规范化

Classification and prediction method of hierarchical multi-label breast disease based on tree search

JIN Chengxiao[1], PAN Qiao[1], ZHANG Jingyi[2], YU Chunru[1]
(1 School of Computer Science and Technology, Donghua University, Shanghai 201620, China;
2 Wonders Information Co., Ltd., Shanghai 201112, China)

【Abstract】 With the rapid development of medical informatization, medical institutions generate a large amount of original electronic medical record data during the clinical diagnosis process, and there is a large amount of information that can be mined for clinical auxiliary diagnosis. Because the prevalence of patients with breast disease is more complicated, the same patient may suffer from multiple related diseases. There may be many small diseases under each major disease classification, and there may be more detailed diseases categories under the small disease classification. In the traditional classification problems (such as binary classification and multi-label classification), the dependencies between the labels are tended to be ignored and the number of classification algorithm outputs is exponential, taking up too much space, resulting in poor prediction performance. Therefore, a hierarchical multi-label breast disease classification diagnosis method based on tree search is proposed in this paper. The tree structure can fully consider the hierarchical relationship between label sets and standardize the diagnosis conclusion. According to the hierarchical relationship between the diagnosis results, a hierarchical multi-label tree is constructed. By searching the path of the label tree, the multi-label classification of breast diseases is finally realized.

【Key words】 breast disease; tree search; hierarchical multi-label; normalize

0 引 言

近年来，乳腺疾病的发病率正在逐渐上升，严重影响了妇女和少数男性的生命安全和生活质量，据统计，全球每年查出患乳腺癌的人数约有 120 万，其中 50 万人死于乳腺癌[1-2]。所以，积极寻找有效的乳腺疾病诊断方法，尽早对诊断结果作出预防，提高乳腺病患的治愈率在目前的研究中尤为重要。随着现代化临床医疗信息系统的快速发展，电子病历系统中积累了越来越多的医疗数据，其中乳腺疾病数据占据了一定的比例，对乳腺疾病的诊断、预测和治疗等有着重要的研究价值[3]。

人工智能中常用的预测方法一般都归结为二分类或多分类问题，对于疾病的预测方法有例如甲状腺良恶性的预测方法，阿尔兹海默症的多分类诊断方法等[4]。但是在实际的临床上，患者的患病情况较为复杂，同一名患者可能会有 3～4 种疾病，例如患有乳腺肿瘤疾病的患者，可能还伴随转移、高血压以及骨质疏松等疾病，各个大类疾病分类下会存在很多的小类疾病分类，小类疾病分类下可能还会有更细粒度的疾病类别标签存在，例如，乳腺良性肿瘤

基金项目：上海市经信委人工智能创新发展专项资金（RX-RJJC-08-16-0483，2017-RGZN-01004）。

作者简介：金程笑（1996-），女，硕士研究生，主要研究方向：医疗大数据、人工智能；潘 乔（1977-），男，博士，副教授，硕士生导师，主要研究方向：数据挖掘、网络性能分析；张敬谊（1974- ），女，博士，教授级高级工程师，主要研究方向：级联式体系架构、异构异质数据采集、大数据分析。

收稿日期：2019-12-18

这一大类会分类为纤维瘤、脂肪瘤、乳头状瘤这三个小类。但是，传统的分类问题往往会忽略各标签之间存在的依赖关系，并且分类算法输出数目呈指数级，占用空间过大，造成预测性能不佳，因此，多标签分类成为解决该类问题的主要方法[5-6]。

多标签分类指的是一个样本可能同时属于多个类别(即有多个标签)，并且这些类别之间可能存在一定的相关性[7]。针对同一个样本进行多标签分类相较于单标签分类要复杂得多，而在实际生活中存在较多的多标签分类的问题[8]，例如电影分类、图书分类和疾病分类等。

多标签分类算法通常分为 2 个类别。一类是通过数据集分解，将多标签分类问题分解为多个单标签分类问题处理。给定 n 个元素的标签集合 $L=(L_1,L_2,\cdots,L_n)$，将 L 中的任意 2 个标签 L_n，L_m 组合病构建一个分类器，该分类器中只含有对应标签 L_n，L_m 的类别的数据。如果将 L 中所有标签进行组合会有 $n*(n-1)/2$ 个分类器。因此，多标签分类问题可以转化为通过构建 $n*(n-1)/2$ 个二分类问题进行处理，如 Goldstein 等人[9]在 i2b2 2008 数据上实验，使用一对一策略将肥胖症及其他 15 种并发症进行多标记分类问题转换为多个二元分类问题；另一类是通过基于单个优化的多标签分类算法，如耿丽娟[10]提出基于域数的加权 KNN 算法，针对 9 980篇的医疗相关文本进行多分类，构建内、外层体系结构分别通过 KNN 算法进行分类，该算法的优点是不需要更改数据集的结构，根据近邻域数进行选择性文本加权，保留了标签之间的依赖，有效地提高了分类精度。

上述两类算法的问题在于基于数据集分解的算法无法保证类别之间存在的依赖性，而基于单个优化的算法虽然保留了标签之间的依赖，但是又因为多标签分类问题的输出空间过大会出现计算效率较低的问题。因此，一些研究者根据多标签问题的 2 个主要缺点提出了层次多标签分类算法，如 Clare 等人[11]利用分层多标签分析微生物突变表型生长实验的数据，以预测新的基因功能，使得准确率超过 80%。该算法通过将数据集分层可以保证类别间的依赖关系，通过将标签分层在训练时可以将数据集进行分类，减少输出空间，很好地提高计算性能。

本文提出了一种基于树搜索的层次多标签乳腺疾病分类预测方法。按诊断结果之间的层次关系构建了层次多标签树，通过对标签树的路径搜索，最终实现乳腺疾病的多标签分类。该方法的特点是利用树结构可以充分考虑到标签集之间的层次结构的依赖关系，达到规范化诊断结论的目的。

1　具体方法

本文提出的基于树搜索的层次多标签分类诊断方法的总体流程如图 1 所示。首先，通过对所要预测的诊断疾病进行层次标签树构建。然后，对每个层次标签树的非叶子节点进行基分类器的训练。最后，对层次标签树的路径进行打分，选取高于某设定阈值的路径进行反馈，实现对乳腺电子病历的层次多标签分类诊断。

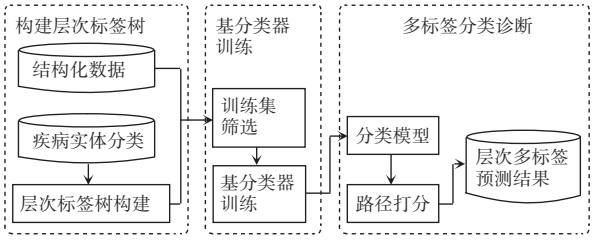

图 1　基于树搜索的层次多标签分类诊断算法的流程图

Fig. 1　Flow chart of hierarchical multi-label classification and diagnosis algorithm based on tree search

1.1　构建层次标签树

1.1.1　获取实体标签集

本文的实验数据均来自上海某三甲医院提供的真实的乳腺电子病历数据，主要采用电子病历中的出院小结和首次病程记录作为研究对象。根据 i2b2(2010)电子病历标注规范中 5 类实体的描述对乳腺电子病历进行标注，实体名称及其标注见表 1[12]。采用了乳腺电子病历的实体和关系联合抽取模型，对乳腺电子病历进行建模，同时完成乳腺电子病历实体识别与关系抽取，获得了最终的实体标签集[13-15]。

表 1　实体名称及其标注

Tab. 1　Entity name and its label

实体名称	实体标注
疾病实体	DIS
症状实体	SYN
检查实体	TES
治疗实体	TRE
其他	O

1.1.2　疾病实体分类

在乳腺电子病历中，患者的患病情况较为复杂，同一名患者可能会有 3~4 种疾病，例如患有乳腺肿瘤疾病的患者，可能还伴随转移、高血压以及骨质疏松等疾病，各个大类疾病分类下会存在很多的小类疾病分类，小类疾病分类下可能还会有更细粒度的

疾病类别标签存在,例如,乳腺良性肿瘤这一大类会分类为纤维瘤、脂肪瘤、乳头状瘤这三个小类。疾病的划分见表 2。

表 2 疾病类别

Tab. 2 Disease category

疾病类别标签	疾病从属标签
疾病总类别	乳腺肿瘤、其他疾病
乳腺肿瘤	乳腺良性肿瘤、乳腺恶性肿瘤
乳腺良性肿瘤	纤维瘤、脂肪瘤、乳头状瘤
乳腺恶性肿瘤	浸润性导管瘤、血管肉瘤
浸润性导管、血管肉瘤	淋巴转移、肺转移、肝转移、腹腔转移、骨转移
其他疾病	高血压、糖尿病、其他乳腺疾病
其他乳腺疾病	乳腺增生、乳腺结节

1.1.3 构建层次标签树

通过表 2 可以发现这些标签(疾病)之间存在树形层次结构关系,将上表划分的疾病构建为层次标签树,疾病关系层次结构映射如图 2 所示,标签树包括非叶子节点和叶子节点两类。非叶子节点作为疾病大类一般包含多个子类标签,即疾病子类,在标签树上从根节点至叶子节点,也表示了从大的疾病分类逐渐缩小到疾病小类的过程。

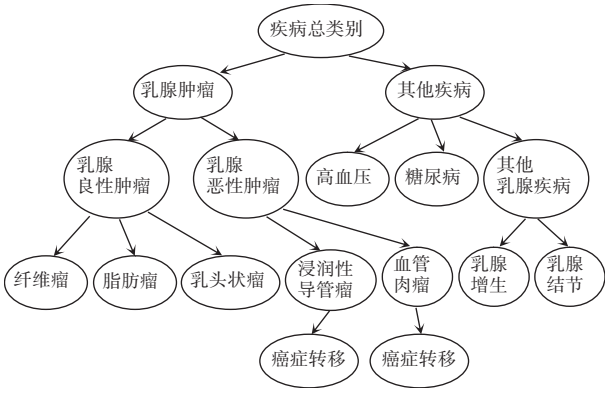

图 2 疾病层次多标签树结构体

Fig. 2 Disease level multi-tag tree structure

1.2 基分类器训练

1.2.1 训练集筛选

由于在层次多标签树中,每个非叶子节点对应作为一个分类器对其所对应的孩子节点进行分类。每一个分类器 c_i 的训练集分为 2 个部分。一部分由对应距离非叶子节点 c_i 最近一层的子节点 $sub+(c_i)$ 组成,记为 $train+(c_i)$,用于训练属于节点 c_i 的分类器;另一部分由不含有 c_i 子节点的所有标签组成,用于训练完全不属于 c_i 节点的分类结果,记为 $train-(c_i)$。若 c_i 没有兄弟节点,则在层次标签树中向上搜索,找到离 c_i 最近的含有兄弟节点的非叶子节点 $bro(parent(c_i))$,并且将这个节点不包含 c_i 的样本加入 $train-(c_i)$。

例如,当前节点 y 为乳腺良性肿瘤,则 c_i 这个节点的训练集由正样本训练集 c_i 下所有包含子节点的样本组成,同时,负样本由乳腺恶性肿瘤这个节点的样本组成且负样本中不含有 c_i 的节点和 c_i 的子节点。

1.2.2 基分类器训练

模型的训练算法描述如算法 1 所示。

算法 1 乳腺电子病历层次多标签分类训练算法

输入:乳腺电子病历未标注数据集 U,疾病分类标签 Y

输出:学习模型 L

initialize: $U' \in U$

/* 训练集初始化,进行标注 */

$LabelTree = createTree(UJ')$

/* 创建层次多标签树 */

For c_i in $LabelTree$:

If c_i is not leafnode

/* 判断节点 c_i 是否为叶子节点 */

$train+ = train+.add(sub+(c_i))$

/* 把 c_i 最近的子节点加入 $train+$ 集合中 */

If c_i has brother node /* 如果 c_i 有兄弟节点 */

$train- = train-.add(sub-(c_i))$

/* 把 c_i 兄弟节点最近的子节点加入 $train-$ 集合中 */

Else

$train- = train-.add(bro(parent(c_i)))$

/* 找到离 c_i 最近的含有兄弟节点的非叶子节点 $bro(parent(c_i))$,并且将这个节点不包含 c_i 的样本加入 $train-(c_i)$ */

End If

$modelL = train(train+ \cup train-)$

/* 训练学习器 */

End If

End For

return L

算法 1 是根据乳腺电子病历的特点,先通过表 2 的分类构建层次多标签分类树,再将训练集按照树中的每一个非叶子节点的标签进行分类,最后形成乳腺电子病历层次多标签分类训练算法框架。该框架也可根据数据的实际需要更换合理的基分类器

进行训练、分类。

1.3　多标签分类诊断

层次标签树中一条路径的得分是通过每个非叶子节点上基分类器的预测结果进行加权求和获得的。层次标签树中的权值如式(1)所示：

$$w(c_i)=\frac{maxlayer-layer(c_i)+1}{\max layer+1},\qquad(1)$$

式(1)的主要作用是反映路径中层次对于节点的影响，即越靠近根节点的非叶子节点的分类准确性对整个分类起到的影响更大。如果高层的节点出现分类错误，则对整个路径上的分类会出现较大的影响，产生的错误损失也会越大。

非叶子节点标签 y_i 的高度由 $level(y)$ 表示，层次标签树中树的最大高度通过 $maxlayer$ 表示。层次标签树中的每条路径的得分通过式(2) 来计算：

$$s_i=\sum_{i=1}^{m}w(c_i)*p(c_i\mid x).\qquad(2)$$

路径得分 s_i 计算流程是：给定第 i 条路径，节点个数为 m，首先计算每一个基分类器所预测概率 p，然后再与每一层的权值 $w(c_i)$ 进行加权，最后通过计算预测概率的加权和。计算路径得分如算法 2 所示。

算法 2　乳腺电子病历层次多标签分类算法

输入：乳腺电子病历测试数据集 U，疾病分类标签 Y，基分类器 Classifier 阈值 σ

输出：预测标签集 *Labels*

initialize：$U'\in U$

/* 训练集初始化，进行标注 */

$LabelTree=createTree(UJ')$

/* 创建层次多标签树 */

For c_i in *LabelTree*：

　　If c_i is not *leafnode* /* 判断节点 c_i 是否为叶子节点 */

　　$p(c_i)=classifier(c_i)$ /* 计算非叶子节点 c_i 的预测概率 */

$$w(c_i)=\frac{maxlayer-layer(c_i)+1}{maxlayer+1}$$

/* 计算 c_i 所在层次节点的权值 */

$$s_i=\sum_{i=1}^{m}w(c_i)*p(c_i\mid x)$$

/* 计算 c_i 节点的得分 */

　　End If

End For

For s in *score*：

　　$scoreTree=SumTree(s)$

　　If $s>=\sigma$：

　　$Labels.add(c_i)$

　　End If

End For

return *Labels*

算法 2 首先计算了每一个节点的概率和节点所在层的权重，再通过式(2)计算该路径的得分，比较选取不同的阈值σ对结果的验证、比较，将得分大于阈值的路径中的节点加入分类结果的集合中，作为最终结果返回，每个返回节点对应的标签则为最终的预测标签集合。

2　实验

2.1　实验数据

为了对本文提出的层次多标签方法进行有效性评估，首先将电子病历原始数据经过上述的实体识别与关系抽取，得到同时含有 TeAS(因症状而采取检查)和 TeRD(检查发现某种疾病)这两种关系的乳腺电子病历数据作为训练数据集，然后将含有症状的电子病历语句筛选出作为输入数据，将疾病作为对应的结果集。数据集中部分数据见表 3。

除了乳腺电子病历入院简要病史数据外，还额外加入体检摘要和生命体征指标共同作为特征，作为基分类器的输入。症状为乳腺电子病历实体识别后提取的结构化数据，体检摘要为患者进行 B 超、MRI 等检查的报告，生命体征为患者检查过程中各项指标的记录，对疾病的诊断同样有重要的参考意义，所以把体检摘要、生命体征和症状集合的数据一并加入作为特征。在此基础上，可得设计研发内容分述如下。

表 3　乳腺电子病历数据

Tab. 3　Breast electronic medical record data

数据	分项内容		
症状	肿块、发热、乳头溢液、淋巴结、橘皮样变		
体检摘要	双乳对称，乳头无内陷及歪斜，双侧乳头位于同一水平面。左乳内上触及 4 cm 肿块，质韧，边界欠清，活动度一般。右乳未见明显异常，双腋下及锁骨上淋巴结阴性。 腹部 B 超：肝内脂肪浸润，胆囊胰体脾肾未见明显异常。 双乳 MR：左乳内上多发团块		
生命体征	血红蛋白：119	前白蛋白：249	尿素：4.2
	肌酐：64	谷丙转氨酶：12	碱性磷酸酶：88
诊断		纤维瘤、高血压	

(1)词语编码。首先，将症状中的词语映射为

一个数字,当输入至基分类器时,这个词语对应的数字和词语所在词向量表 lookup table 中对应的向量将会共同作为基分类器的输入,词语的编号由词语在句子中的起始位置决定。

(2)标签编码。给定大小为 n 的疾病的多标签分类集合为 $L=(L_1,L_2,\ldots,L_n)$,当某样本 x 含有 L_i 时,则 $L_i=1$,否则 $L_i=0$。在本文中,标签的总数为 19,标签集合编码顺序则按从上到下,从左到右进行排列。如果样本 x 包含乳腺纤维瘤和高血压这两类疾病,则 x 在 L 中会对应 5 个标签,对应的多标签分类的集合对应的诊断属性见表 4。

表 4 乳腺电子病历结构化数据编码

Tab. 4 Structured data encoding for electronic breast medical records

数据分项	对应数据编码
症状	(12,16,20,29,35,42)
生命体征	(119,249,4.2,64,12,88)
诊断	(1,1,1,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0)

2.2 实验评价标准

给定大小为 n 的标签集合 $Y=(y_1,y_2,\cdots,y_n)$,集合中 y_i 表示某样本含有第 i 个标签,分别用 1 和 0 表示样本含有标签 y_i 和不含有 y_i。本文中,y_i 表示目标集合,y_i' 表示预测集合。这里,对研究中选用的设计评价指标将做阐释表述如下。

(1)预测标签在子集中的准确率(*subset accuracy*)。表示测试集中预测的标签集合完全正确的样本占全部样本的比例,如式(3)所示:

$$subset\ accuracy=\frac{1}{N}\sum_{i=1}^{N}(y_i=y_i'), \tag{3}$$

(2)准确率(*accuracy*)。如式(4)所示:

$$accuracy=\frac{1}{p}\sum_{i=1}^{p}\frac{|y_i\cap y_i'|}{|y_i\cup y_i'|}, \tag{4}$$

(3)精度(*precision*)。如式(5)所示:

$$precision=\frac{1}{p}\sum_{i=1}^{p}\frac{|y_i\cap y_i'|}{|y_i'|}, \tag{5}$$

(4)召回率(*recall*)。如式(6)所示:

$$recall=\frac{1}{p}\sum_{i=1}^{p}\frac{|y_i\cap y_i'|}{|y_i'|}, \tag{6}$$

(5) F_1 值。如式(7)所示:

$$F_1=2*\frac{precision*recall}{precision+recall}. \tag{7}$$

2.3 多模型实验对比

子集准确率(*subset accuracy*)判断为真需要满足算法预测的标签集合等于目标集合。由于多标签分类的输出空间较大,完全准确地预测每一个集合中的标签并不容易,所以子集准确率通常提升不明显。首先,通过本文层次标签树多分类方法,这项指标提升至 70.3%。同时,由标签树的分层结构,保留标签之间的依赖关系,可以对训练数据集进行有效划分,从而减少计算性能。通过对比逻辑回归模型和 KNN 模型分别提高了 16%和 8%,所以使用标签树有效避免了传统多标签分类样本空间过大导致分类效果欠佳的问题。准确率、精度、召回率和 F_1 四类指标同样也作为算法常规的评价标准,层次多标签与其他模型进行比较的结果见表 5。

表 5 多模型实验对比结果

Tab. 5 Multi-model experiment comparison results

模型	*subset accuracy*	*accuracy*	*precision*	*recall*	F_1
LR	53.8	61.3	65.3	62.7	63.9
KNN	62.7	68.4	72.2	70.9	71.4
层次多标签	70.3	73.7	78.6	77.8	78.1

2.4 多分类器实验对比

本节通过实验来对比多种基分类器对基于层次多标签分类算法的效果,并选择性能最优的基分类器来测试本文的方法。根据 4 种不同的分类方法来比较不同的分类器对该层次多标签分类算法的性能。使用 4 种常见的方法作为基分类器。见表 6。

表 6 多分类器结果对比

Tab. 6 Comparison of multiple classifier results

模型	*subset accuracy*	*accuracy*	*precision*	*recall*	F_1
LR	60.7	63.7	68.4	66.2	67.3
LSTM+LR	65.6	69.3	73.7	71.6	72.6
KNN	63.6	68.9	71.4	70.2	70.7
层次 Bi-LSTM[16]	70.3	73.7	78.6	77.8	78.1

通过实验的对比,将 4 种基分类器应用于层次多标签分类方法,在训练数据的维度都为 300 维时,LSTM+LR 对比 KNN 算法在相同输入的情况下,效果略好于 KNN,各项指标普遍提升约 2%。层次 Bi-LSTM 模型对比 LSTM+LR 与 KNN 模型的精确度提升明显,准确度高出约 5%。层次 Bi-LSTM 算法的分层提取特征的特点,将一个较长维度的输入进行分层特征提取,有效地降维。通过实验对比,本文选择层次 Bi-LSTM 作为最终的基分类器。

2.5 多阈值 σ 实验对比

本节中通过模型简化测试方法对多阈值 σ 条件下模型的训练效果和性能进行定量分析。采用逐步增加阈值的数量检验模型的分类能力通过汉明损失进行对比,如式(8)所示:

$$HamLoss = \frac{1}{p} \mid \sum_{i=1}^{p} y_i' \Delta y_i \mid . \qquad (8)$$

对比各算法输出的汉明损失，汉明损失表示多标签分类模型精度，首先计算每个样本中标签对预测错误的个数，计算 $y_i \Delta y_i'$，Δ 为异或操作，再与对应预测标签 y_i' 的预测概率相乘，最后计算每个预测标签乘积和的均值。

如图 3 所示，当选取词向量作为网络输入时，通过逐步增加阈值使得各算法最终输出的路径得分超过所设定的阈值。通过实验对比发现，当阈值设定在 0.50~0.70 之间汉明损失趋势总体呈现逐步下降，而当阈值大于 0.70 时，*accuracy* 的趋势稳定或者呈现出略微上升的趋势。接下来，通过十折交叉验证进行试验，经过 LSTM 进行初步语义特征提取的逻辑回归算法比 LR 逻辑回归算法的汉明损失降低0.2，而 LSTM+LR 算法与树层次标签算法普遍相差0.1。

当阈值为 0.65 时，LSTM+LR 算法的汉明损失最小，多标签分类的效果较好。当阈值为 0.7 时，层次多标签算法的分类效果最为显著，整体汉明损失低于前述 2 种算法。

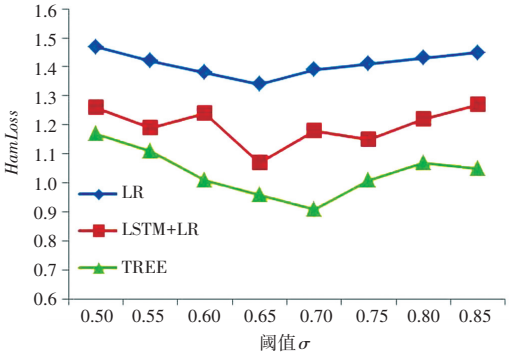

图 3　不同大小阈值在层次多标签算法的汉明损失对比

Fig. 3　Comparison of Hamming loss of hierarchical multi-label algorithm with different thresholds

3　结束语

本文提出了一种基于树搜索的层次多标签乳腺疾病分类诊断方法，实验数据是来自上海某三甲医院提供的真实的乳腺电子病历数据，通过引用实体和关系联合抽取方法提取出的疾病实体作为实体标签集。首先介绍了层次多标签分类总体流程，然后对疾病的类别进行详尽分类，并阐述了根据分类结果构建层次标签树的过程，提出了基于树搜索的多标签分类诊断的计算方法，最后进行实验对比。

通过在真实数据集上进行对比实验，使用准确率、召回率等多组评价指标对模型结果进行评估，证明了对已有模型的改进并且有效地提高了电子病历实体识别以及关系抽取的准确性。通过多模型和多个基分类器进行对比，证明了基于树搜索的层次多标签乳腺疾病分类诊断方法的有效性。

接下来的研究工作可以从这 2 个方面展开。首先，使用网络上的公开训练集作为实验数据，为后续多标签预测提供更准确的训练集做模型训练。其次，将电子病历中的其他因素作为特征进行多标签分类，从而提高辅助诊断的真实性与全面性。

参考文献

[1] JOCHEN K, JÖRG D, MIENA A, et al. Cognitive performance and psychological distress in breast cancer patients at disease onset [J]. Frontiers in psychology, 2019.

[2] 李玉阳. 山东省乳腺疾病调查报告与乳腺癌危险因素分析[D]. 济南：山东大学，2011.

[3] 张晓雅，肖宝菊. 电子病历的现状与发展趋势[J]. 电子技术与软件工程，2018(8)：176.

[4] LIU Jin, LI Min, LAN Wei, et al. Classification of Alzheimer´s disease using whole brain hierarchical network [J]. IEEE/ACM transactions on computational biology and bioinformatics, 2018, 15(2)：624.

[5] 李思男，李宁，李战怀. 多标签数据挖掘技术：研究综述[J]. 计算机科学，2013，40(4)：14.

[6] CAI Zhiling , ZHU W. Feature selection for multi-label classification using neighborhood preservation [J]. IEEE/CAA Journal of Automatica Sinica, 2018, 5(1)：320.

[7] 冯雪东. 多标签分类问题综述[J]. 信息系统工程，2016(3)：137.

[8] 马鸿超，张坤丽，赵悦淑，等. 基于特征融合的产科多标记辅助诊断研究[J]. 中文信息学报，2018，32(5)：128.

[9] GOLDSTEIN I, UZUNOR Ö. Specializing for predicting obesity and its co-morbidities [J]. Journal of Biomedical Informatics, 2009, 42(5)：873.

[10] 耿丽娟. 基于健康医疗大数据的 KNN 分类算法研究[J]. 通讯世界，2017(20)：265.

[11] CLARE A , KING R D . Knowledge discovery in multi-label phenotype data [J]. Lecture Notes in Computer Science, 2001, 2168(2168)：42.

[12] De BRUIJN B, CHERRY C, KIRITCHENKO S, et al. Machine-learned solutions for three stages of clinical information extraction: The state of the art at i2b2 2010 [J]. Journal of the American Medical Informatics Association Jamia, 2011, 18(5)：557.

[13] PAN Qiao, YU Chunru, CHEN Dehua, et al. Joint extraction of entities and relations of breast ultrasound report based on deep learning [C]//The 20th IEEE International Conference on High Performance Computing and Communications (HPCC). Guangzhou：IEEE Society, 2018.

[14] QAMAS G K S，尹继泽，潘丽敏，等. 基于深度神经网络的命名实体识别方法研究[J]. 信息网络安全，2017(10)：29.

[15] GRIDACH M. Character-level neural network for biomedical named entity recognition [J]. Journal of Biomedical Informatics, 2017, 70：85.

[16] BAKER S , KORHONEN A. Initializing neural networks for hierarchical multi-label text classification [C]// BioNLP 2017. Vancouver, Canada：Association for Computational Linguistics, 2017：307.