**Computer Science and Application** 计算机科学与应用, 2022, 12(2), 280-289
Published Online February 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.122028

Hans汉斯

# 基于知识图谱的儿童疾病推理模型

**罗柏涛，王　勇，王　瑛**

广东工业大学计算机学院，广东　广州

收稿日期：2022年1月10日；录用日期：2022年2月7日；发布日期：2022年2月14日

## 摘　要

**为提高儿童多种疾病推理的准确率，提出以知识图谱为数据基础并结合推理机的儿童疾病推理算法。通过Neo4j构建知识图谱对数据进行储存和应用，而推理机以TransE推理机为基础，并结合朴素贝叶斯分类器来提升推理机处理不同疾病含有同种症状的问题的能力，再通过建立的自适应机制来降低不同疾病的症状数不同对推理机的影响。实验结果表明，所提出算法在疾病推理的精确率，召回率和*F*1值上均有所提升，说明该方法提高了儿童多种疾病推理的准确率。**

## 关键词

**知识图谱，疾病推理，自适应机制**

# Children's Disease Reasoning Model Based on Knowledge Graph

**Baitao Luo, Yong Wang, Ying Wang**

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong

Received: Jan. 10th, 2022; accepted: Feb. 7th, 2022; published: Feb. 14th, 2022

## Abstract

**To improve the accuracy of children's multiple diseases reasoning, a children's disease reasoning algorithm based on the knowledge graph and the reasoning machine is proposed. A knowledge graph is constructed by Neo4j to store and apply the data, and the inference engine is based on the TransE inference engine, and the Naive Bayes Classifier is combined to improve the ability to deal with problems with the same symptoms in different diseases, and an adaptive mechanism is established to reduce the impact of different symptoms of different diseases on the reasoning ma-**

文章引用: 罗柏涛, 王勇, 王瑛. 基于知识图谱的儿童疾病推理模型[J]. 计算机科学与应用, 2022, 12(2): 280-289.
DOI: 10.12677/csa.2022.122028

**chine. The experimental results show that the proposed algorithm improves the precision, recall and *F*1 value of disease reasoning, indicating that this method improves the accuracy of children's multiple disease reasoning.**

## Keywords

**Knowledge Graph, Disease Reasoning, Adaptive Mechanism**



## 1. 引言

伴随医学信息化的提升，现已经产生和积累了大量医学数据，且在健康网上咨询的情况也越来越常见。而回答咨询的基础为疾病推理，所以提高疾病推理的准确率有利于更好地推荐宜忌食物或推荐就诊科室等推荐服务。而儿童时期为人生的关键时期且儿童疾病的症状明显，所以儿童疾病推理具有好的应用前景。

疾病推理即根据病人的症状等信息，推理出病人可能有的疾病。Wang 等人[1]结合专家系统和人工智能的推理模型正确率高，但需要设定大量规则。Qiu 等人[2]基于贝叶斯的方法能够模拟人脑的学习，但所学的结构可能不太准确。龚乐君等人[3]基于决策树方法准确率高，但是微调数据可能会使模型失去稳定性。Jia [4]结合了知识图和深度强化学习提出模型 DKDR，该模型诊断准确率较高，但是特征少的时候对训练结果的影响较大。刘勘等人[5]提出的 CDR (CNN-DNN-TransR)模型准确率提升高，但是该模型需要在较为完备的知识图谱上进行且由于医疗记录的复杂性，用语的口语化和多样性，将影响症状实体准确识别。Chai [6]提出了 BLSTM (Bi-directional Long Short-Term Memory)用于甲状腺疾病的诊断，该方法实验结果稳定，对甲状腺疾病的认知率在 80%以上，但是参与的特征对结果影响较大。

上述方法的问题是在多种疾病推理上准确率低，或未考虑输入不平衡性的问题。因此受混合推理[7]的启发，本文结合了 Neo4j 构建的知识图谱储存数据可用性好，TransE 推理速度快，朴素贝叶斯在多分类上复杂度低和自适应机制抗输入不平衡性干扰效果好的优点，从而提高了在儿童多种疾病推理上的准确率。

## 2. 医学知识图谱分析

知识图谱是图 $G = (V, E)$的某种扩展形式，$V$ 是顶点集合，表示实体，$E$ 是边集合，表示实体间的联系。知识图谱是图数据模型的继承和发展，其在一般图模型的顶点和边上附加更多的属性信息，用于描述现实世界中事物的广泛联系[8]。其常用类型为 RDF 图。

定义 1. RDF 图。设 $U$、$B$ 和 $L$ 为互不相交的无限集合，分别代表 *URI*、空顶点和字面量。一个三元组$(s, p, o) \in (U \cup B) \times U \times (U \cup B \cup L)$称为 RDF 三元组，其中 $s$ 是主语，$p$ 是谓语，$o$ 是宾语。RDF 图 $G$ 是三元组$(s, p, o)$的有限集合。

定义 2. 医学知识图谱 *MedKG* (Medical Knowledge Graph)。*MedKG* 如图 1 所示，可用式(1)表示图谱中的每一条医学知识[9]。$h$、$t$ 分别为头实体和尾实体，$r$ 为实体间的关系。

$$MedKG = (h, r, t) \tag{1}$$

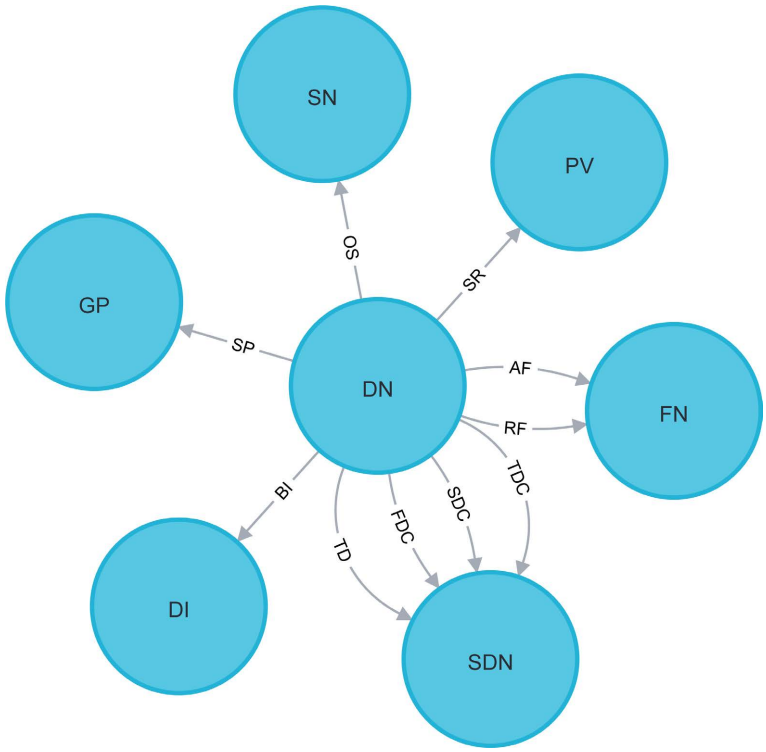


**Figure 1.** Structure diagram of *MedKG*
**图 1.** 医学知识图谱的结构图

定义 3. *MedKG* 的实体(结点)集合。*MedKG* 的实体集合是医学领域中具体事物的集合，主要由疾病、症状、科室等构成，部分结点集合如表 1 所示。*MedKG* 中的 $h,t \in E(\text{Entities})$。

**Table 1.** Node set
**表 1.** 结点集合

| 名称 | 英文全称 | 中文说明 |
|---|---|---|
| DN | Disease Name | 疾病名称 |
| SDN | Specific Department Name | 具体科室名称 |
| DI | Disease Introduction | 疾病的介绍 |
| …… | …… | …… |

定义 4. *MedKG* 的关系(边)集合。*MedKG* 的关系集合是医学领域中具体事物间联系的集合，主要由类别、病因等构成，部分关系(边)集合如表 2 所示。*MedKG* 中的 $r \in R(\text{Relation})$。

**Table 2.** Edge set
**表 2.** 边集合

| 名称 | 英文全称 | 中文说明 |
|---|---|---|
| TD | Treatment Department | 疾病属于哪个科室治疗 |
| BI | Brief Introduction | 疾病与疾病介绍的解说关系 |
| OS | Occurrence Site | 疾病产生的症状发生的位置 |
| …… | …… | …… |

## 3. 医学知识图谱分析

### 3.1. 解决思路和整体框架

在多种疾病推理上准确率低问题的主要特点：大规模数据，不同疾病含有同种症状和不同疾病的症状数不同。针对大规模数据，过去常用 CiteSpace 知识图谱[10]，但其可用性差和计算性差。Neo4j 具有成熟数据库的所有特性[11]，因此本文用 Neo4j 构建知识图谱。

推理机部分，由于以知识图谱为数据结构，因此使用 TransE 推理机方便且效率高。但是 TransE 在多对多上准确率低，为此提出了许多改进的模型，但文献[12]用实验表明了在生物医学数据集上，TransE 在链接预测上依旧表现出了最好的性能和效率。所以针对不同疾病含有同种症状的问题，引入在多分类中有较好效果的朴素贝叶斯算法，本文建立 TransE-NBC 模型。同时针对不同疾病的症状数不同的问题，即输入不平衡性的问题，建立了自适应机制，即模型会根据数据集的输入数目的特点进行拆分，拆分后的每个模型将处理输入数目近似的数据集，且拆分后的每个模型的参数会根据对应的数据集训练后自主调节得出。通过以上的操作，本文最终建立了推理模型 Adapt-TransE-NBC。整体框架如图 2 所示，整体框架由知识图谱的构建、推理机学习阶段对推理机参数的训练和推理机的应用(测试)阶段的说明组成。

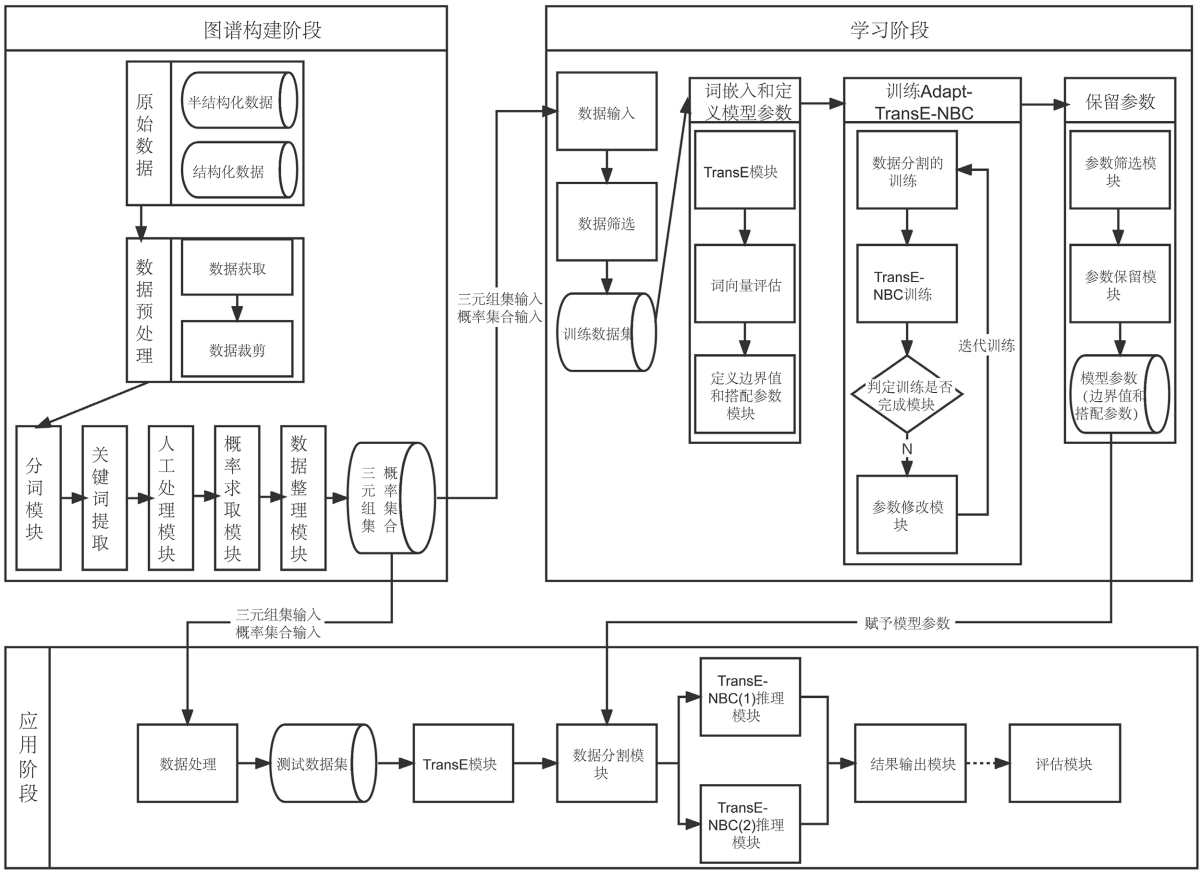


**Figure 2.** Overall structure of model
**图 2.** 模型的整体框架

### 3.2. 知识图谱的构建

本文采用 Neo4j 构建知识图谱，具体过程如下所示：

1) 利用 Python 的 urllib 库获取网页上的需要字段。

2) 基于 Pkuseg 分词包进行分词操作，将所需要的关键信息和其他无用信息进行分离。

3) 基于 THUOCL_medical 词库做关键词提取，从分词后的数据，提取关键信息。

4) 由于 THUOCL_medical 词库不一定涵盖了所有的词，因此结合人工提取和补充关键词，得到以三元组<症状，发生部位，疾病>为结构的数据集。

5) 将上述获得的三元组可视化后，构建出知识图谱。

构建完成后，根据式(2)计算出<症状，发生部位>随机出现在疾病 $c$ 中的先验概率 $P(c)$，其中 $N$ 是三元组集中所有的疾病数，$N_c$ 是三元组集中疾病 $c$ 的数目。

$$P(c)=\frac{N_c}{N} \tag{2}$$

再根据式(3)计算条件概率 $P(w|c)$，其中 $N_{d,c}$ 为疾病 $c$ 包含<症状，发生部位>$d$ 的疾病数。

$$P(w\mid c)=\frac{N_{d,c}}{N_c} \tag{3}$$

获得以<症状，发生部位，疾病>为结构的三元组集和概率集合(先验概率和条件概率)。

### 3.3. 推理机学习阶段

首先选取所有<症状，发生部位，疾病>三元组，送入翻译模型 TransE 里训练和评估后得到映射向量。定义输入参数为 $x$，$x$ 代表的是预测疾病时输入的<症状，发生部位>的数目。定义边界值 $M$，其作用是根据输入数目与边界值的大小关系将数据集分割成两份。$M$ 的取值范围如式(4)所示，具体的 $M$ 的值的确定通过训练得出。

$$1\le M\le \frac{\max(x)}{2},(M\in Z) \tag{4}$$

分割后，找合适的搭配参数。定义当 $x$ 小于 $M$ 时的参数为 $\alpha_1$ 和 $\beta_1$，当 $x$ 大于等于 $M$ 时的参数为 $\alpha_2$ 和 $\beta_2$。其取值范围如式(5)所示。

$$\begin{cases}0\le \alpha_1,\beta_1\le 1,(\alpha_1,\beta_1\in Z) & x<M\\ 0\le \alpha_2,\beta_2\le 1,(\alpha_2,\beta_2\in Z) & x\ge M\end{cases} \tag{5}$$

将以三元组<症状，发生部位，疾病>为结构的数据集作为训练数据集。假设头向量 $\boldsymbol{h}$ 为症状，关系向量 $\boldsymbol{r}$ 为发生部位和尾向量 $\boldsymbol{t}$ 为疾病，且 $L$ 类疾病分别设为 $\{b_1,b_2,\cdots,b_L\}$。将数据集中的单个疾病的三元组<症状，发生部位，疾病>输入。判断输入的<症状，发生部位>的数目，让不同输入数目的三元组进入不同的模型，算法如式(6)所示。

$$F=\begin{cases}\max\left(\sum_{i(1,x)}\left(Gap_i^{\alpha_1}Chance_i^{\beta_1}\right)\right) & x<M\\ \max\left(\sum_{i(1,x)}\left(Gap_i^{\alpha_2}Chance_i^{\beta_2}\right)\right) & x\ge M\end{cases} \tag{6}$$

式(6)中的 $Gap_i$ 为 TransE 推理机，其构成过程如下所示：由 $x$ 组症状及发生部位预测疾病，则将 $x$ 组输入与输出间的距离值累乘得到最终的距离值，如式(7)所示。其中 $L_1$ 为曼哈顿距离， $L_2$ 为欧氏距离。

$$D(b_j)=\prod_{i(1,x)}\left\|h_i+r_i-t_{b_j}\right\|L_1/L_2 \tag{7}$$

其中单个症状及发生部位与疾病的距离值的计算方法为 $Gap_i$，如式(8)所示。

$$Gap_i = \left\| h_i + r_i - t_{b_j} \right\| L_1 / L_2 \tag{8}$$

式(6)中的 $Chance_i$ 为修改后的朴素贝叶斯分类器，其构成过程如下所示：首先朴素贝叶斯分类器假设各个特征相互独立，其算法如式(9)所示。

$$P\left(t_{b_j} \mid h_1 r_1, h_2 r_2, \cdots, h_i r_i\right) = P\left(t_{b_j}\right) \prod_{i(1,x)} P\left(h_i r_i \mid t_{b_j}\right) \tag{9}$$

经过式子(9)后再通过病症寻找最大概率疾病输出最终结果。因式(9)输出的是最大值，因此修改后 TransE 推理机的算法如式(10)所示。

$$Gap_i = \left( \left\| h_i + r_i - t_{b_j} \right\| L_1 / L_2 \right)^{-1} \tag{10}$$

其中单个症状及发生部位与疾病间概率值的算法为 $Chance_i$，其表达式如式(11)所示。

$$Chance_i = P\left(t_{b_j}\right) P\left(h_i r_i \mid t_{b_j}\right) \tag{11}$$

获得了 $Gap_i$ 和 $Chance_i$ 后，进行累乘得出结果，如式(12)所示。

$$F\left(b_j\right) = \prod_{i(1,x)} \left( Gap_i * \left( Chance_i \right) \right) \tag{12}$$

但两个较小数乘积后累乘，可能会产生极小数而溢出下界造成错误，因此一般加入 *log* 函数解决该问题，从而避免下界溢出的问题，但为了防止 0 值对模型的影响，因此将原本累乘得到的 $F(b_j)$修改为累加得到，具体的 $F(b_j)$的计算方式如式(13)所示。

$$F\left(b_j\right) = \sum_{i(1,x)} \left( Gap_i * \left( Chance_i \right) \right) \tag{13}$$

由于式(8)和式(10)都求最大值且都为正值，所以当输入为 $x$ 组 $h_i$ 和 $r_i$ 输入时，最终输出的疾病为最大的 $F(b_j)$值的所对应的疾病 $b_j$，具体的 $F_{finall}$ 值的算法如式(14)所示。

$$F_{finall} = \max\left( \sum_{i(1,x)} \left( Gap_i * \left( Chance_i \right) \right) \right) \tag{14}$$

式(6)在式(14)的基础上增加自适应机制，分别作用于输入和输出上。输出上的自适应机制：在推理训练时将根据数据集的输入的特性，调节式(14)中 $Gap_i$ 和 $Chance_i$ 所占的权重比。输入上的自适应机制：考虑到算法复杂度的问题，因此在训练时根据输入的数目将输入数据集进行拆分并且仅将其拆分成两份。

输入数据需要经过式(6)后进行训练，将数据集经过所有的 $M$，$\alpha_1$，$\beta_1$，$\alpha_2$ 和 $\beta_2$ 后训练得出总准确率最高时的 $M_f$，$\alpha_{1(f)}$，$\beta_{1(f)}$，$\alpha_{2(f)}$和 $\beta_{2(f)}$。

### 3.4. 推理机应用阶段

将总准确率最高时的 $M_f$，$\alpha_{1(f)}$，$\beta_{1(f)}$，$\alpha_{2(f)}$和 $\beta_{2(f)}$带回式(6)后如式(15)所示。

$$F = \begin{cases} \max\left( \sum_{i(1,x)} \left( Gap_i^{\alpha_{1(f)}} Chance_i^{\beta_{1(f)}} \right) \right) & x < M_f \\ \max\left( \sum_{i(1,x)} \left( Gap_i^{\alpha_{2(f)}} Chance_i^{\beta_{2(f)}} \right) \right) & x \geq M_f \end{cases} \tag{15}$$

应用阶段，将 $x$ 组<症状，发生部位>经翻译模型后输入到推理模型中，模型首先判断输入数目 $x$ 与边界值 $M_f$ 的大小关系，然后将 $x$ 组<症状，发生部位>输入到对应的拆分模型中，对应的拆分模型中会输出值最大的。最大值对应的疾病，即为输出的预测疾病。

## 4. 实验

### 4.1. 数据集及预处理

数据来源于“快速问医生”健康网上儿童相关疾病数据。“发生部位”来源于“夏禾健康”网。按图谱构建预处理后 Neo4j 可视化。图 3 是疾病“百日咳”在 Neo4j 知识图谱上的部分图像，如果正确推理出疾病，有利于更好地推荐宜忌食物或推荐就诊科室等推荐服务。

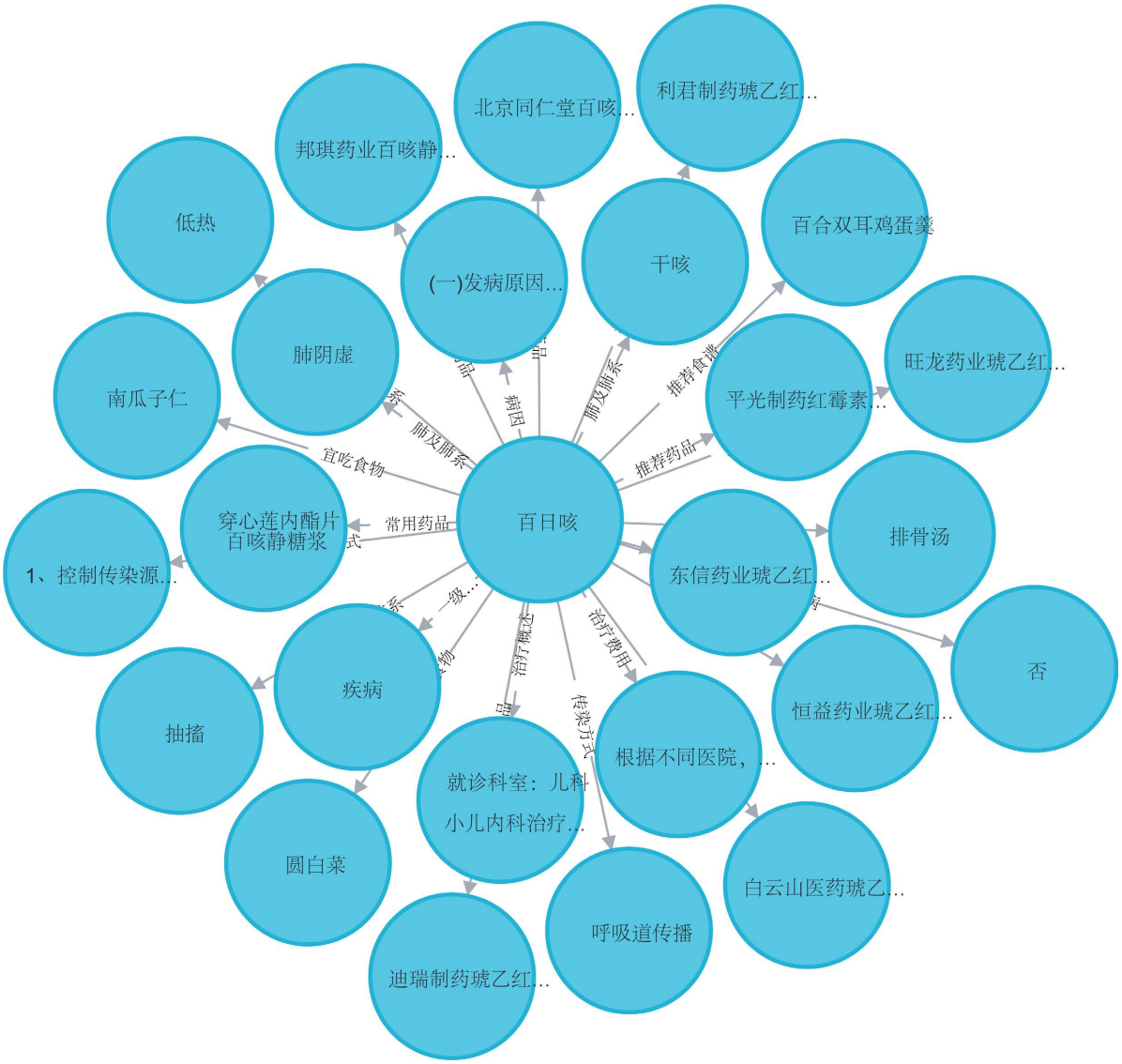


**Figure 3.** Some images of the disease “whooping cough” on Neo4j
**图 3.** 疾病“百日咳”在 Neo4j 上的部分图像

构建完成后，以每个疾病对应的多个<症状，发生部位>随机产生输入，形成 20 万条左右的数据集，

测试数据集的信息如表 3 所示。负样本为当预测单个疾病时其他疾病名作为错误样本，将正负样本合并成为实验数据。将 20 万条左右的数据集分成 10 组，每次模型都分别测试 10 组，并计算每个模型的测试标准值。

**Table 3.** Test data set
**表 3.** 测试数据集

| 三元组数目/个 | 疾病种类/种 | 总数目/条 | 测试组数/组 | 每组数目/条 |
|---|---|---|---|---|
| 7031 | 990 | 198,200 | 10 | 19,820 |

### 4.2. 推理机部分的实验设置

Adapt-TransE-NBC 推理机是能够根据症状和发生部位推理出疾病。由于实际情况是由随机数目的症状和发生部位输入，因此实验设置将每个疾病的随机数目的症状和症状对应的发生部位作为输入预测疾病。对比算法为 Text-CNN 和 Bi-LSTM/BLSTM。

Text-CNN：基于卷积神经网络分类的推理机。

Bi-LSTM/BLSTM：基于双向 LSTM 的推理机。

实验还对比了不同的 TransE 词向量维度时的 Adapt-TransE-NBC，目的是为了体现模型的灵活性，并为进一步的改进提供方向。

### 4.3. 测试标准

标准为精确率，召回率和 *F*1 值，分别用 *Precision*，*Recall* 和 *F*1 表示，如式(16)~式(18)所示。当做了 *N* 次时，需要计算平均值作为测试标准，*Precision*，*Recall* 和 *F*1 的平均值分别用 *M_Precision*，*M_Recall* 和 *M_F*1 表示，如式(19)~式(21)所示。

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{18}$$

$$M\_Precision = \frac{\sum_{i(1,N)} Precision_i}{N} \tag{19}$$

$$M\_Recall = \frac{\sum_{i(1,N)} Recall_i}{N} \tag{20}$$

$$M\_F1 = \frac{\sum_{i(1,N)} (F1)_i}{N} \tag{21}$$

### 4.4. 实验结果和分析

测试结果如表 4 所示。Adapt-TransE-NBC(20)是 TransE 的词向量维度为 20 时的 Adapt-TransE-NBC 模型，而 Adapt-TransE-NBC(50)是 TransE 的词向量维度为 50 时的 Adapt-TransE-NBC 模型。表中的 *Precision*，*Recall* 和 *F*1 是 10 次实验中最好效果时的值，而 *M_Precision*，*M_Recall* 和 *M_F*1 为实验 10

次的平均值。

**Table 4.** Test Results (%)
**表 4.** 测试结果(%)

| 方法 | *Precision* | *Recall* | *F*1 | *M_Precision* | *M_Recall* | *M_F*1 |
|---|---|---|---|---|---|---|
| Text-CNN | 76.77 | 76.75 | 76.77 | 76.04 | 75.53 | 75.78 |
| Bi-LSTM | 79.71 | 78.54 | 79.12 | 78.45 | 78.38 | 78.56 |
| Adapt-TransE-NBC(20) | 81.15 | 81.23 | 81.19 | 80.89 | 81.08 | 80.99 |
| Adapt-TransE-NBC(50) | **81.22** | **81.61** | **81.41** | **81.13** | **81.21** | **81.17** |

Adapt-TransE-NBC 在 *Precision*，*Recall*，*F*1，*M_Precision*，*M_Recall* 和 *M_F*1 上相对于 Text-CNN 和 Bi-LSTM/BLSTM 均有所提升，且随着词向量维度提升，所有的指标仍有小幅度提升。

因此当在选择<症状，发生部位>输入后，该模型可更可靠地预测出疾病。但词向量维度的提升需更长时间的 TransE 训练和整体模型训练，因此需要根据实际情况慎重考虑。

## 5. 结论

针对现有模型在儿童多种疾病推理上准确率低的问题，本文提出了以 Neo4j 构建知识图谱并建立了 Adapt-TransE-NBC 模型作为疾病推理机的机制。通过建立的自适应机制使每个模型处理的为输入数目近似的数据集，且使每个模型不再是简单的组合，而是根据数据集训练出合适的搭配参数，从而减少了输入不平衡性的影响。通过实验表明，提出的方法在儿童多种疾病推理上的精确率，召回率和 *F*1 值均有所提升。因此所提出的方法可为在健康网上回答咨询和推荐服务提供更可靠的基础。

下一步的任务是，探究如何更好地优化模型的算法结构和降低训练时间，使其能够适应更大型知识图谱的推理。

## 参考文献


[1] Wang, J., Liu, N., Hu, Q., *et al*. (2020) The Intelligent Diagnostic System for Common Diseases Using the Optimized Medical Knowledge Graph. 2020 *International Wireless Communications and Mobile Computing*, Limassol, Cyprus, 15-19 June 2020, 1504-1509. https://doi.org/10.1109/IWCMC48107.2020.9148406

[2] Qiu, W., Zhang, Y. and Cheng, B. (2018) Building Syndrome and Symptom Association Network by Bayesian Network. 4*th International Conference on Computer and Communications* (ICCC), Chengdu, China, 7-10 December 2018, 1762-1766. https://doi.org/10.1109/CompComm.2018.8780599

[3] 龚乐君, 张立鹏, 李宇茜, 吴向辉, 高志宏, 潘传迪, 杨庚. 基于决策树的乳腺癌病历文本的挖掘与决策[J]. 南京师大学报(自然科学版), 2019, 42(3): 42-51.

[4] Jia, Y., Tan, Z. and Zhang, J. (2019) DKDR: An Approach of Knowledge Graph and Deep Reinforcement Learning for Disease Diagnosis. 2019 *IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking* (*ISPA/BDCloud/SocialCom/SustainCom*), Xiamen, China, 16-18 December 2019,1303-1308. https://doi.org/10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00187

[5] 刘勘, 张雅荃. 基于医疗知识图谱的并发症辅助诊断[J]. 中文信息学报, 2020, 34(10): 85-93+104.

[6] Chai, X. (2020) Diagnosis Method of Thyroid Disease Combining Knowledge Graph and Deep Learning. *IEEE Access*, **8**, 14978-149795. https://doi.org/10.1109/ACCESS.2020.3016676

[7] 官赛萍, 靳小龙, 贾岩涛, 王元卓, 程学旗. 面向知识图谱的知识推理研究进展[J]. 软件学报, 2018, 29(10): 2966-2994.

[8] 王鑫, 陈蔚雪, 杨雅君, 张小旺, 冯志勇. 知识图谱划分算法研究综述[J]. 计算机学报, 2021, 44(1): 235-260.

[9] 董丽丽, 程炯, 张翔, 叶娜. 融合知识图谱与深度学习的疾病诊断方法研究[J]. 计算机科学与探索, 2020, 14(5): 815-824.

[10] 关媛媛, 郝阳, 田春颖, 孙璇, 王东军, 田之魁, 王泓午. 基于 CiteSpace 的舌诊诊断标准研究的可视化分析[J]. 世界科学技术-中医药现代化, 2021, 23(1): 263-270.

[11] Zou, Y. and Liu, Y. (2020) The Implementation Knowledge Graph of Air Crash Data Based on Neo4j. 4*th Information Technology, Networking, Electronic and Automation Control Conference*, Chongqing, China, 12-14 June 2020, 1699-1702. https://doi.org/10.1109/ITNEC48623.2020.9085182

[12] Choi, W. and Lee, H. (2019) Inference of Biomedical Relations among Chemicals, Genes, Diseases, and Symptoms Using Knowledge Representation Learning. *IEEE Access*, **7**, 179373-179384. https://doi.org/10.1109/ACCESS.2019.2957812