变量选择以及疾病打分策略

1.相关概念定义

对于我们现在的逻辑规则，都是形如$s_1 \wedge s_2 \wedge d_1 \wedge d_2 \rightarrow s_3 \vee s_4 \vee d_3 \vee d_4$这样的公式, 我们需要计算一个症状s对一个疾病的贡献度，记为$F(s_i, d)$,我们通过如下公式来计算:

$$ContributeScore(s_i, d) = \frac{TF(s_i, d)}{\sum_{s_j \in S} TF(s_j, d))} \times IDF(s_i)$$

其中

- $TF(s_i, d)$的计算仿照IF-IDF指数中的计算方式，对于$s_1 \wedge s_2 \wedge d_1 \wedge d_2 \rightarrow s_3 \vee s_4 \vee d_3 \vee d_4$，我们认为$s_3, s_4$在$d_1, d_2$中各出现0.5次，$s_1, s_2$在$d_3, d_4$中各出现0.5次（$s_1, s_2$与$d_1, d_2$没有联系，$s_3, s_4$在$d_3, d_4$没有联系，即他们之间的TF=0),即
$$TF(s_1, d_3) = TF(s_1, d_4) = TF(s_2, d_3) = TF(s_2, d_4) = \\ TF(s_3, d_1) = TF(s_3, d_2) = TF(s_4, d_1) = TF(s_4, d_2) = 0.5$$
- $IDF(s_i)$同样和IF-IDF指数中的计算方式相同
$$IDF(s_i) = \frac{1}{与\ s_i 相关的 d_i 的数量} = \frac{1}{|D_R|}(s.t. \forall d_i \in D_R, TF(s_i, d_i) > 0)$$

对于多个公式，比如:

$$d_1 \wedge d_2 \wedge d_3 \rightarrow s_1 \\ d_1 \wedge d_2 \rightarrow s_1 \vee s_2 \vee s_3 \\ s_1 \wedge d_1 \wedge d_2 \rightarrow d_3 \vee s_4$$

则TF计算如下:

	s_1	s_2	s_3	s_4
d_1	1/3+1/2	1/2	1/2	1/2
d_2	1/3+1/2	1/2	1/2	1/2
d_3	1/3+1	0	0	0

IDF计算如下:

	s_1	s_2	s_3	s_4
IDF	1/3	1/2	1/2	1/2

2.变量选择

2.1 最大熵

在推理过程中，假设当前的候选疾病集合是D,待确定症状集合是S,我们要从S中选出一个变量去询问患者，我们通过计算一个变量的熵，选择熵最大的去问。因为问题答案只有yes或no,令$p(s=true|d_1 \vee d_2 \vee ... \vee d_n)=x$

$$H(s)=-(xlogx+(1-x)log(1-x))$$

x越接近1/2，H(s)越大。

不过我们没有$p(s=true|d_1 \vee d_2 \vee ... \vee d_n)$的数据，我们通过

$$p(s|D)=p(s=true|d_1 \vee d_2 \vee ... \vee d_n) \approx \sum_{d_i \in D} p(s \wedge d_i)=\sum_{d_i \in D} p(s|d_i) \times p(d_i)$$

来进行估计，其中$p(s|d_i)$我们可以用TF来估计，即

$$p(s|d_i)=\frac{TF(s,d_i)}{\sum_{s_i \in S} TF(s_i|d_i)}$$
$$p(d_i)=\frac{Score(d_i)}{\sum_{d_i \in D} Score(d_i)}$$

2.2 最大期望疾病贡献度

一个症状的取值会使得相关的疾病得分发生变化,我们可以根据如下公式计算症状对所有疾病的贡献度, ，假设当前的候选疾病集合是D,待确定症状集合是S

$$E(s)=\sum_{d_i \in D} ContributeScore(s,d_i) \times p(s|D)$$

然后选择$E(s)$最高的症状进行询问

3.候选疾病打分

对于最终的推理结果，逻辑引擎肯定无法排除掉所有无关的疾病，所以最终的候选疾病列表可能很大，我们通过患者的主述来给候选疾病打分，选择分高的返回。

$$Score(d)=\sum_{s_i \in main\_info\_list} ContributeScore(s_i,d)$$