CONTEXTCARE

HAO zhaojun

2017 年 12 月 18 日

目录

1 简介

CONTEXTCare 利用上下文信息来提高医学论坛上数据的表示学习 (ContextCare incorporating contextual information networks to representation learning on medical forum data), 有效地利用医学论坛上表达不正式, 聊天性强的上下文信息, 一定程度上克服了医学论坛上文本提取的症状-疾病关系稀疏的问题.

2 用途

CONTEXTCare 允许利用

- 搜索引擎上的搜索和对应的疾病文件
- 医学论坛上问答中浩如烟海的信息

来提取症状 (Symptoms) 和疾病 (Diseaseas) 的特征, 进而根据症状做

- 疾病预测
- 疾病种类预测
- 疾病聚类

3 现存方法问题

一般我们是通过症状和疾病的二分图来表示症状-疾病连接关系, 但是这个二分图十分稀疏. 没有足够的数据来保证表示的质量. 同时, 医学论坛上有大量的对话信息.

4 CONTEXTCare 的创新

CONTEXTCARE 定义了三个图, 在传统的**症状-疾病网络**的基础上, 增加了**症状共发网络**和**疾病演进网络**, **有效缓解了症状-疾病二分图的连接稀疏性问题**.
并定义**能量函数**, 能量函数越小, 代表症状-疾病联系越强.
能量函数定义为

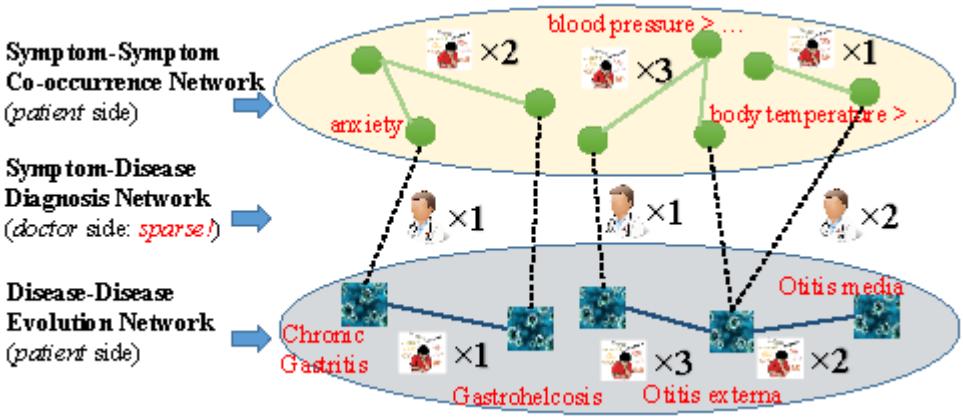
$$f(S^p, d) = \left\| \frac{1}{|S^p|} \sum_{s \in S^p} \mathbf{s} + \mathbf{t} - \mathbf{d} \right\|_1 \quad (1)$$

其中 s,t,d 分别是症状和疾病的表示,t 为症状-疾病转移向量.S^p 代表病人的症状.

5 模型结构

5.1 三个网络

CONTEXTCARE 定义了三个网络: 1. 症状-疾病网络 2. 症状共发网络 3. 疾病演进网络



5.1.1 症状-疾病网络

症状和疾病的网络定义如下损失函数

$$L(G^{SD}) = \sum_{\substack{(S^p, d) \in \mathcal{M}^+ \\ (S^{p'}, d') \in \mathcal{M}^-}} [\gamma + f(S^p, d) - f(S^{p'}, d')]_+ \quad (2)$$

其中

- $[x]_+ = max(0, x)$.
- M^+ 是从医学论坛问答中获得的正确的症状-疾病对,M^- 是通过替换症状或者疾病来构造的错误的症状-疾病对.

背后的思想是, 学习了 S^p 和 d 的表示后, 给了 S^p 后我们可以预测到正确的 d.

5.1.2 症状共发网络

症状的共发关系可以帮助获得更加准确的症状表示.

$$R_1(G^{SS}) = \sum_{(s_i,s_j)\in E^{ss}} w_{ij} \|\mathbf{s}_i - \mathbf{s}_j\|_1 \tag{3}$$

通过最小化**症状正则项**
其中

- $w_{i,j} = \frac{|\Gamma(s_i) \bigcap \Gamma(s_j)|}{|\Gamma(s_i) \bigcup \Gamma(s_j)|}$
- $\Gamma(s_i)$ 表示 s_i 的临近节点的集合. 一旦两个症状的共生频率超过临界点, 这两个症状就成为临近节点.

5.1.3 疾病演进网络

疾病的演进关系可以帮助获得更加准确得疾病表示, 并且给出疾病的发展轨迹和变种信息.
通过最小化**疾病正则项**

$$R_2(G^{DD}) = \sum_{(d_n,d_m)\in E^{ev}} v_{mn} \|\mathbf{d}_n - \mathbf{d}_m\|_2 \tag{4}$$

其中

- $v_{m,n} = \frac{|\Gamma(d_m) \bigcap \Gamma(d_n)|}{|\Gamma(d_m) \bigcup \Gamma(d_n)|}$
- $\Gamma(d_m)$ 表示 d_m 的临近节点的集合. 一旦两个疾病存在演进关系, 这两个疾病就成为临近节点.

这里使用 L2 正则是为了

1. 和症状共发网络区别开来
2. 强调疾病演进关系

5.1.4 表示学习

通过最小化**总目标函数**

$$\min_{\{s\},\{d\},t} L(G^{SD}) + \alpha R_1(G^{SS}) + \beta R_2(G^{DD}) \qquad (5)$$

可以帮助找到正确的 (S^p, d) 对.

6 其他模型及其理论缺点

1. 分类模型 (Classification models)
2. 主题模型 (Topic modeling method) : LDA(Latent Dirichlet Allocation)
3. 连接预测模型 (Link prediction models) : P-PageRank, SimRank 和HeteSim
4. 网络嵌入模型 (Network embedding methods) : LSHM 模型

6.1 LDA

LDA 模型通过三层贝叶斯概率来得到文章的主题.

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^{N} p(z_n | \theta) p(w_n | z_n, \beta), \qquad (2)$$

图 1: LDA_loss

其中

- $p(\theta|\alpha)$ 描述了文本各个主题的概率, 符合参数为 α 的 dirichlet 分布
- $p(z_n|\theta)$ 描述了字的各个主题的概率, 符合参数为 θ 的多项式分布
- $p(w_n|z_n, \beta)$ 描述了给定主题 z_n, 字 w_n 的概率, 符合参数为 β 的分布, 也就是 $p(w_n^j = 1|z^i = 1) = \beta_(i, j)$

在疾病预测中, 把症状看做词, 疾病看做文本, 但是症状-疾病连接很稀疏, 所以 ***LDA* 模型的数据很稀疏 *(文本的词数很少, 并且词比较口语化)***.

6.2 HeteSim

用来计算不同质物品的相似度.

DEFINITION 3: **HeteSim**: Given a relevance path $\mathcal{P} = R_1 \circ R_2 \circ \cdots \circ R_l$, the HeteSi objects s and t ($s \in R_1.S$ and $t \in R_l.T$) is:

$$HeteSim(s,t|R_1 \circ R_2 \circ \cdots \circ R_l) = \frac{1}{|O(s|R_1)||I(t|R_l)|} \sum_{i=1}^{|O(s|R_1)|} \sum_{j=1}^{|I(t|R_l)|} HeteSim(O_i(s|R_1), I_j$$

where $O(s|R_1)$ is the out-neighbors of s based on relation R_1, and $I(t|R_l)$ is the in-neighbors R_l.

如果 $|O(s|R_1)| * |I(t|R_l)| = 0$, 则 $heteSim(s,t) = 0$, 所以**症状-疾病的关系是稀疏的**.

用来计算同质物品的相似度

DEFINITION 4: **HeteSim based on self-relation**: HeteSim between two same-typed objects on the self-relation I is:

$$HeteSim(s,t|I) = \delta(s,t)$$

其中 $HeteSim(o,i)$ 为不同质元素的距离.

即使是同质物品之间, 只有对自己为 1, 其他为 0. 在疾病预测中, 我们可以用来推断症状到疾病的相似度, 所以 HeteSim 模型**不能很好得刻画症状之间或者疾病之间的关系**.

6.3 LSHM

计算不同质物品之间的相似度并考虑度量的平滑性 (相似物品之间的潜在向量相近).

我们把 N_1 个症状和 N_2 个疾病, 看做节点 $x_i, \forall i \in [1, N_1+N_2]$, 一共 N_1+N_2 个节点, 其中 L 个打了标签的.

$$\sum_{i=1}^{\ell} \Delta(f_\theta^{y_i}(z_i), y_i) \tag{3}$$

where $\Delta(f_\theta^{y_i}(z_i), y_i)$ is the loss of predicting labels $f_\theta^{y_i}(z_i)$ instead of observed labels y_i. The final objective loss of our model combines the classification and regularization losses 3 and 2:

$$L(z, \theta) = \sum_{i=1}^{\ell} \Delta(f_\theta^{y_i}(z_i), y_i) + \lambda \sum_{i,j/w_{i,j}\neq 0} w_{i,j} ||z_i - z_j||^2 \tag{4}$$

图 2: LSHM_loss

其中,

1. z_i 为 x_i 的潜在向量,
2. f_θ^i 为对应于 z_i 的线性分类函数,
3. $w_{i,j}$ 为 x_i, x_j 节点间的权重. 如果两个节点没连接则为 0.
4. λ 为 L2 正则化项系数.

在我们的场景中可以通过每个节点的线性分类函数来给症状分类和计算症状-疾病的相似度. 可以看出,LSHM 模型将症状和疾病一视同仁, 症状共发和疾病进化关系都用 L2 度量, **无法区分两者的重要性 *(疾病进化关系更重要)***. 此外,LSHM 计算的是一对一的相似度, 而不像 CONTEXTCARE 模型, **体现不了症状群和疾病的相似度**.

7 实验结果

7.1 疾病预测

在疾病预测方面的准确率对比

Method	Accuracy				Precision@5
	Diagnosis Network	*+Co*	*+Ev*	*+Co,Ev*	
SVM (linear)	16.02	-	-	-	-
SVM (RBF)	16.79	-	-	-	-
Decision Tree (C4.5)	17.31	-	-	-	-
MaxEnt	18.98	-	-	-	-
LDA	14.73	-	-	-	23.46
P-PageRank	17.22	19.71	17.25	19.74	43.17
SimRank	19.36	21.97	19.38	21.98	46.52
HeteSim	20.62	23.03	20.69	23.32	55.31
LSHM	21.38	25.87	22.55	25.77	65.74
ContextCare$_\times$	22.35	28.09	24.74	30.66	69.38
CONTEXTCARE	**23.57**	**30.32**	**27.26**	**31.73**	**73.21**

Table 1: Comparison with baseline methods in accuracy and precision@N

图 3: comparison

7.2 疾病种类预测

ICD-10 作为疾病分类标准.

8 CONTEXTCARE 模型的优势

从其他模型理论缺点和 实验对比可以看出模型在**利用医学论坛数据的优势**和**预测疾病发展的能力**. 医学论坛上的数据症状-疾病连接很稀疏.

9 结论

我们可以试着分析中文医学论坛的数据特征, 使用 CONTEXTCARE 来通过症状描述来预测疾病和疾病分类.

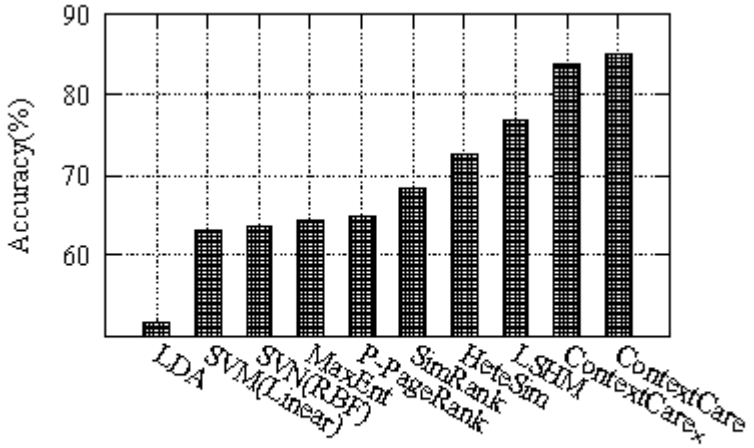


Figure 4: Comparison with baseline methods on disease category prediction in accuracy (%).

图 4: comparison in category prediction